Machine learning approach to semantic augmentation of multimedia documents for efficient access and retrieval

Kosinov, Serhiy

UNIVERSITÉ DE GENÈVE
Département d'informatique

FACULTÉ DES SCIENCES
Docteur S. Marchand-Maillet
Professeur T. Pun

Machine learning approach to semantic augmentation of multimedia documents for efficient access and retrieval

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par

Serhiy Kosinov

de
Kirovograd (UKRAINE)

Thèse N° 3707

GENÈVE
2005

Preface

This thesis is dedicated to the theoretical development and practical application of machine learning methods to content-based multimedia analysis and retrieval. For clarity and in order to avoid confusion, a special attention must be given to the way this thesis interprets and uses the term *multimedia*, that has a great number of meanings and ways of being understood, and whose definitions continuously evolve and change keeping the pace with technological progress.

According to the literal meaning that comes from the latin *multus* = "many, multiple ..." and *medium* = "a channel or system of communication, information, or entertainment", multimedia relates to the use of computers to present text, graphics, video, animation, and sound in an integrated way. However, the perception of the term multimedia is distinctively different from the perspective of a machine learning system, or any multimedia information system in general. An information system deals with the same bit pattern abstraction of the digitized and possibly encoded information, irrespective of its sensory origins. As some authors put it, multimedia in this context refers to any visual information, audio information or textual information, taken either separately or in combination. Digital multimedia information is immediately visible, audible, readable and in most cases understandable to the user, but *not* to the system. This important discrepancy between the digital representation and semantic interpretation is known as *semantic gap* problem, which is the main focus of the machine learning techniques developed in this thesis.

List of Frequently Used Acronyms

2D	2-dimensional
ACM	Analytic center machine
ARD	Automatic relevance determination
BDA	Biased discriminant analysis, BiasMap
BPM	Bayes point machine
CBIR	Content-based image retrieval
CL-LSI	Cross-language latent semantic indexing
DANN	Discriminant adaptive nearest neighbor
DDA	Distance-based discriminant analysis
FPP	False positive projection
GM	Geometric mean
HMM	Hidden Markov Model
HSE	Hierarchical semantic ensemble
HSV	Hue saturation value
IDF	Inverse term frequency
IM	Iterative majorization
KCCA	Kernel canonical correlation analysis
KDDA	Kernel distance-based discriminant analysis
KFD	Kernel Fisher discriminant
LDA	Linear discriminant analysis
LSI	Latent semantic indexing
MH	Multiple-hyperplane
MHMM	Multi-resolution hidden Markov model
MML	Minimum message length
NDA	Non-parametric discriminant analysis
NN	Nearest neighbor
OPC	One per class, one-against-all
PCA	Principal component analysis
PSD	Positive semi-definite
RBF	Radial basis function

RKHS	Reproducing kernel Hilbert space
RKKS	Reproducing kernel Krein space
SDP	Semi-definite programming
SMACOF	Scaling by majorizing a complex function
SP	Specificity
SQP	Sequential quadratic programming
SVC	Support vector clustering
SVM	Support vector machine
TF	Term frequency
TRS	Trust region problem
XOR	Exclusive OR

Abstract

One of the major challenges in content-based multimedia retrieval is due to the problem of semantic gap encountered as a consequence of significant disparities between inherent representational characteristics of multimedia and its meaningful content sought by the user. This work concerns the advancement of the *semantic augmentation* techniques focused on bringing together low-level visual representation of multimedia and its semantics thus attempting to alleviate the above semantic gap problem by augmenting the information used by a multimedia database system in order to improve the efficiency of access and retrieval. The main emphasis and contributions of this work summarized below are in the domain of supervised discriminative learning methods and ensemble techniques.

Proposed is a non-parametric distance-based discriminant analysis method, DDA, focused on finding a discriminative linear transformation that enhances data conformance to the compactness hypothesis and its inverse. The sought transformation, in turn, is found as a solution to an optimization problem formulated in terms of inter-observation distances only, using the technique of *iterative majorization*. The proposed approach is suitable for both binary and multiple-class categorization problems, and can be applied as a dimensionality reduction technique. In the latter case, the number of discriminative features is determined automatically since the process of feature extraction is fully embedded in the optimization procedure.

In order to overcome the limiting assumption of linearity of the sought disciminative transformation, a kernel-based extension of the above discriminant analysis method, KDDA, is formulated whereby the optimization criterion is expressed in terms of distances projected from a feature space induced by a given kernel function. Additionally, an application of *indefinite kernels* rendered as unrestricted linear combinations of hyperkernels is considered in the KDDA framework. The proposed formulation entails a solution of a series of quadratic minimization problems, whose computationally advantageous property of being convex is guaranteed regardless of the definiteness of the selected kernel function. Finally, an adverse condition referred to as the *false positive projection effect* is studied and its elimination strategies are assessed.

Using the above DDA method as a basic building block classifer, a hierarchical ensemble learning approach is developed and applied in the context of multimedia semantic augmentation. In contrast to the standard multiple-category classification setting that assumes independent, non-overlapping and exhaustive set of semantic categories, the proposed approach

models explicitly the hierarchical relationships among target classes, and estimates their relevance to a query as a trade-off between the goodness of fit to a given category description and its inherent uncertainty.

Finally, a motivational analogy of the DDA optimization criterion formulation to that of the analytical center machine approach extended to a case with several separating hyperplanes is considered. An explicit multiple-hyperplane extension of the optimal separating hyperplane classifier is formulated and investigated from the point of view of optimization problem complexity and generalization performance guarantees. The latter properties are derived in terms of the associated fat-shattering dimension bound.

Contents

Chapter 1

Introduction

The vast increase of the amount of available multimedia content necessitates the development of new techniques and methods that not only are able to store and retrieve data effectively, but also can, with or without user's assistance, overcome the semantic gap problem. The said problem is encountered due to major disparities between inherent representational characteristics of multimedia, such as color, texture, shape or motion descriptors, and its meaningful content sought by the user. Exacerbated by the issue of perception subjectivity, i.e. the change of relevance judgments from one individual to another, the semantic gap problem represents a formidable challenge that has been shown to adversely affect the performance of many multimedia database retrieval systems [132]. Naturally, this area has been a prominent research direction addressed by a great number of approaches originating from such areas as machine learning, statistics, natural language processing, etc.

In the discussion that follows, we would like to identify such approaches as those of *semantic augmentation* since most of them are specifically focused on bringing together low-level visual representation of multimedia and its semantics thus augmenting the information used by a multimedia database system in order to improve the efficiency of access and retrieval. The choice of a referral term as general as semantic augmentation allows us to encompass and analyse holistically a great variety of techniques that are designed but with one common goal: to alleviate the semantic gap problem, solved in the case of each individual method via an extremely diverse range of paradigms and formulations related to indexing, learning, classification, categorization, prediction, etc. This choice of terminology is ultimately linked to the way the term *multimedia* is perceived throughout this thesis. Namely, multimedia is broadly understood as any combination of representation of human-perceptible information, whose automatically computable characteristics do not contain direct expression of semantics sought by a user. Thus, digital images, waveform audio signals, video shots are all seen as acceptable instances of multimedia to which semantic augmentation techniques may be applied.

In order to set a proper context for describing the contributions of this thesis, our further consideration will focus on two groups of methods for semantic augmentation: *interactive* (Section 1.1) - the adaptive approaches that are guided by relevance feedback supplied by

the user, and *automatic* (Section 1.2) - those attempting to derive useful correlations between representational characteristics of multimedia and its semantic aspects by applying techniques that do not involve the user.

1.1 Interactive semantic augmentation

The common trait of the methods belonging to the first group of techniques to be considered is the assumption of active user presence during the retrieval process. The user is regarded as the principal source of semantic knowledge in various possible scenarios of interaction. This knowledge may not be explicitly mapped onto a semantic concept (i.e., named) at the end of the process but we consider that the level of description or discrimination attained during the interactive process is high enough to be called semantic.

In this section, we therefore review the most common ways of capturing and exploiting *user interaction* in view of enhancing the retrieval process. In most systems, user interaction is taken as *relevance feedback* [127] from a search result to a subsequent search step. In this scheme, the user is offered a search result and (s)he should mark (some) items of this set as being *relevant* or *irrelevant* (possibly under a fine scale).

Primarily, this mechanism allows for having a direct computation of individual items' importance within the search context. This is exploited in the computation of Rocchio's formulation [125] for adapting term weight at search time (section 1.1.1). From a different viewpoint, when considering that relevance feedback creates inter-item relationships, one may then derive properties from their co-occurence. This is used in both association rule mining and collaborative filtering (section 1.1.2). Alternatively, the interaction may be exploited in a Bayesian framework. Relevance feedback is therefore the base for learning and classification (section 1.1.3). These techniques are originally essentially blind to the type of item under management. In section 1.1.5, we review some schemes that adapt these into the content of interactive image retrieval. More specifically, various tasks classically associated to CBIR are combined into an integrated framework for a collaborative interactive semantic description of the data.

1.1.1 Rocchio's algorithm

Many of the early methods for interactive semantic augmentation emerged from the efforts proven effective in the field of document retrieval, and were built according to the scheme illustrated in Figure 1.1. These approaches were closely tied to the underlying vector space model [131] inheriting the weight calculation rules based on the notions of term and inverse document frequencies, and processed relevance judgments supplied by the user via an additive adjustment formula known as Rocchio's algorithm:

$$Q_{new} = \alpha Q_{old} + \beta \sum D_i + \gamma \sum \bar{D}_i, \tag{1.1}$$

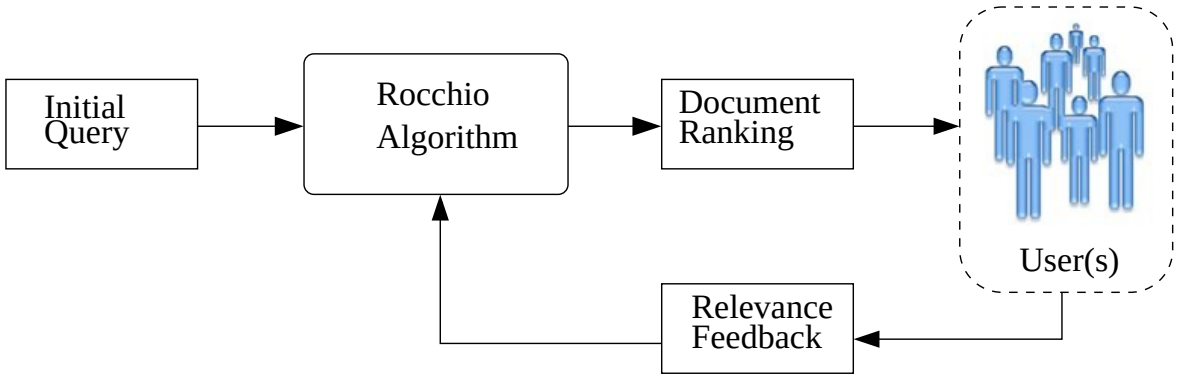


Figure 1.1: Typical scenario of retrieval with relevance feedback

where Q_{old} is a query feature vector from the previous relevance feedback iteration, D_i and $\bar{D}_i$ are given documents from relevant and non-relevant sets, respectively, and α, β and γ are tuning parameters. Using this strategy, integrated with the vector space model for retrieval, the notion of similarity is interactively adapted to the user profile by distorting components of the indexing space.

The *Viper* system [140] adapts classical document retrieval techniques to the context of CBIR. It includes Rocchio's algorithm as a way to handle feedback and to account for the sparsity of positive examples against negative examples. By tuning β and γ parameters in Equation (1.1), this system is robust against abundant negative examples that would normally make retrieval inconsistent [125].

1.1.2 Association rules and collaborative filtering

While the above feedback strategy is based on the features of the items within the collection, the principle here is to derive knowledge from the users themselves.

Assuming the same system is accessed by several users, one wishes to predict information in a given case based on a history of interaction. Let a dataset (collection) D be composed of items, regrouped in itemsets. We wish to create associations between itemsets X and Y ($X \cap Y = \emptyset$) within a particular transaction T (e.g. query process) of the form $X \Rightarrow Y$. That is, within a particular context of interaction, we wish to state that whenever itemset X is considered, itemset Y will also be considered. More formally, we wish to estimate the value of

$$P(Y \subseteq T | X \subseteq T) \tag{1.2}$$

Müller *et al.* [110], propose to use this technique to achieve long-term learning in the context of Content-based Image Retrieval in the *Viper* system based on the vector space model for retrieval. From usage log, relevance feedback is exploited to derive association rules between pairs of images marked relevant or otherwise. Rather than acting of the documents themselves, the authors propose to apply a long-term weight to the basic image features so as to set emphasis on discriminant features.

Collaborative filtering approach [122] uses aggregated subjective evaluations from a group of users to *recommend* items to an *active* user. Typically, from a history of choices made by a population of users on a number of items, one wishes to predict the choice of a user on a particular item. That is, propagating other user's choices onto a particular user, based on known correlations between that particular user and others who already made a decision on that item. More formally, let $v_{i,j}$ be the vote of user u_i on item o_j and I_i the set of items on which user u_i has made a decision. Then, in the simplest case, $\overline{v}_i = (\sum_{j \in I_i} v_{i,j})/|I_i|$ is the average vote of user u_i on I_i, which correlates with the profile of user u_i. Thus, the predicted feedback $\hat{v}_{a,j}$ of active user u_a on item $o_j \notin I_a$ is given by the "profile" of user u_a added with a weighted combination of personalised votes on item o_j ($v_{i,j}$ for all $u_i \neq_a$).

$$\hat{v}_{a,j} = \overline{v}_a + \kappa \sum_{i=1; i \neq a}^{n} w(a,i)(v_{i,j} - \overline{v}_i), \tag{1.3}$$

where the weight $w(a,i)$ represents the correlation between user u_a and user u_i. In early studies, this is simply taken as the Pearson correlation coefficient

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v}_a)(v_{i,j} - \overline{v}_i)}{\sqrt{\sum_j (v_{a,j} - \overline{v}_a)^2 \sum_j (v_{i,j} - \overline{v}_i)^2}}. \tag{1.4}$$

User choices are accumulated. After showing a certain profile ($\overline{v}_a$) by interacting with the system, user u_a then receives recommendations for subsequent searches.

It is important to note here that items are blindly considered as entities and that the complete recommendation procedure is done without any knowledge of the item features. The performance of the system is uniquely based on the quality of the correlation computed and the consistency of the information propagated. In [77], this system is used to create a *WebMuseum* able to distinguish between styles of painting, simply by accumulating user relevance feedback.

1.1.3 Bayesian handling of relevance feedback

Here, we still consider the classical relevance feedback protocol. Simply, positive and negative examples become the base for a Bayesian classification. In [142], positive and negative examples are treated separately. Positive examples are successively used to estimate the parameters of a Gaussian distribution of their features. Negative examples are used as center of penalty functions so that inferred results "stay away" from these examples (this process is referred to as a 'dibbling process' in [142]).

Similarly, Vasconcelos and Lippman [129] propose to use positive and negative examples in a learning cycle by first evaluating a classification based on positive examples x_t and, based on the N best positive classes, solve an equation of the type

$$S_i^* = \text{argmax}_i \left\{ \alpha \log \frac{P(x_t|S_i = 1)}{P(\bar{y}_t|S_i = 1} + (1-\alpha) \log \frac{P(S_i = 1|x_1 \cdots x_{t-1})}{P(S_i = 1|\bar{y}_1 \cdots \bar{y}_{t-1})} \right\}, \tag{1.5}$$

integrating $\bar{y}_t$ as negative examples over time t. Equation (1.5) simply states that the class best described by the set of positive and negative examples at hand is that maximizing the posterior odds ratio between hypothesis "class i is the target" and "class i is not the target".

A different setup, yet using similar techniques has been used in the classical Cox *et al.*'s PicHunter browsing system [32]. It is assumed that the user seeks a specific *target* within the collection. (S)he is then successively proposed samples of the collection containing the most probable targets infered by the system using a Bayesian *posterior* estimation.

1.1.4 Active learning

Still in the spirit of learning from feedback, Tong and Chang [144] among others [25, 71] propose to use Support Vector Machines (SVM) for achieving concept *active learning*. The essential trait of the active learning approach is its ability to proactively select the examples for which relevance feedback is solicited from the user, as opposed to simply asking to label a random subset thereof. The said selection, in turn, is based on focusing on the examples whose classification is difficult. The approach has been further enhanced by Gosselin and Cord [53] by making the selection process depend on the current ranking of samples, rather than on the less stable decision boundary of the classifier at a given feedback iteration. Among the earlier developed active learning methods are such techniques as *uncertainty sampling* [90] and *query by committee* [46]. In the latter approach the subsequent query for feedback is chosen by the principle of maximal disagreement among a committee of classifiers.

1.1.5 Image specific framework

Most of the previous techniques have been developed in the context of document retrieval and may be applied on multimedia in general, provided the right features are used. In the field of CBIR, a number of alternative usage of relevance feedback have been proposed. Here, we do not just aim at creating adaptive similarity associations between documents (ie feature vectors), we wish to derive further useful properties of the image themselves.

For example, in [70, 72], Jing *et al.* propose a strategy to discovering region importance in images handled by a CBIR system. The strategy is to pre-segment the images using the classical JSEG algorithm [39] and then to compute inter-region similarity. A region and an image are called similar if the image contains a region similar to the region the image is compared with. Among the set of positive examples, each region is weighted by a region frequency (RF) denoting its consistency with other regions within the positive set. Then, based on a scheme similar to the TF*IDF scheme for document retrieval, Inverse Image Frequency (IIF) is also computed as the importance of a given region within an image (ie its ability to characterize the image) Finally, the region importance (RI) of region i in a given image is computed as

$$RI_i = \frac{RF_i * IIF_i}{\sum_{j=1}^{n} RF_j * IIF_j}, \tag{1.6}$$

where n is the number of regions within the image. This region importance is finally accumulated in a linear scheme along the feedback steps.

The fact of deriving region importance within images is an appealing process since it forms a step towards identifying objects within the image. From there, several processes may be facilitated. This is true for retrieval, becoming region-based retrieval but also segmentation and compression. In [98], this is discussed, along with the presentation of a complete integrated framework for interactive document retrieval and description. The main observation is that any interaction with the data may be a valuable semantic input. The proposal is therefore to make the data accessible from a number of ways including retrieval, description, viewing and so on. An important aspect is that the data is accessed from several points by several people. In the proposed system, the authors aim at incrementally and collaboratively gathering and inferring high-level information on the data immersed within this system. Eventually, content description may be fixed into a knowledge base. It is shown that such an approach tightly relating content-based retrieval and content description poses new solvable challenges, as opposed to classical CBIR whose performances tend to saturate in current systems.

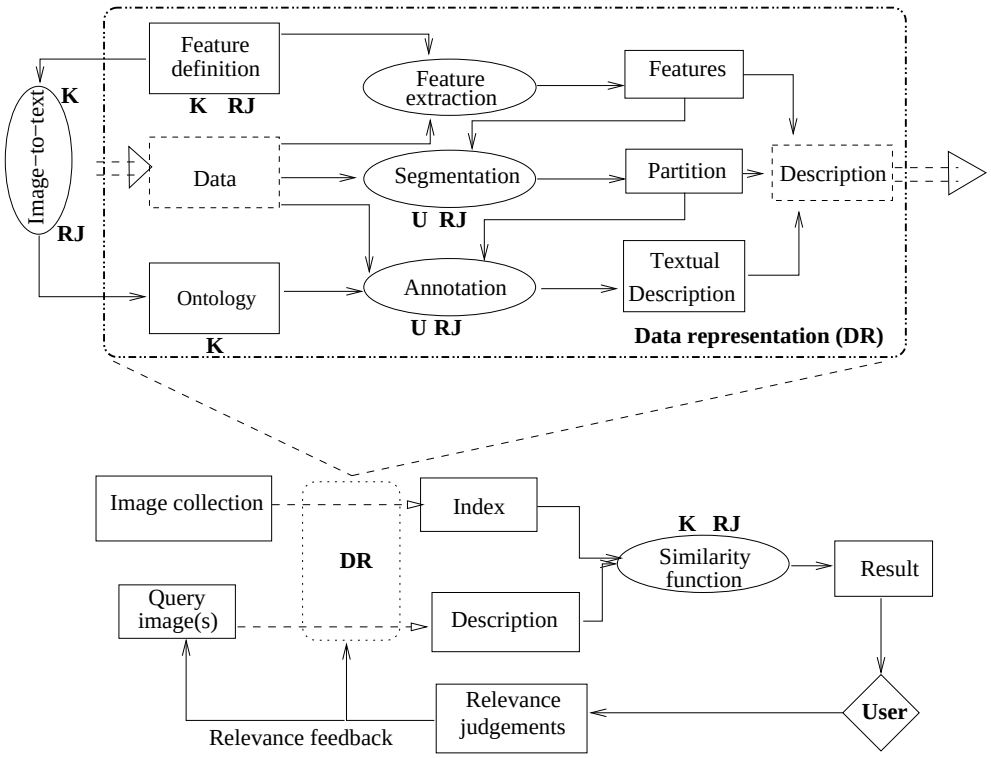


Figure 1.2: The functional schema of a CBIR system completed with possible acquisition of semantic knowledge. **K**, **RJ** and **U** mark places where *a priori* **K**nowledge, **R**elevance **J**udgments and **U**ser interaction may be inserted, respectively

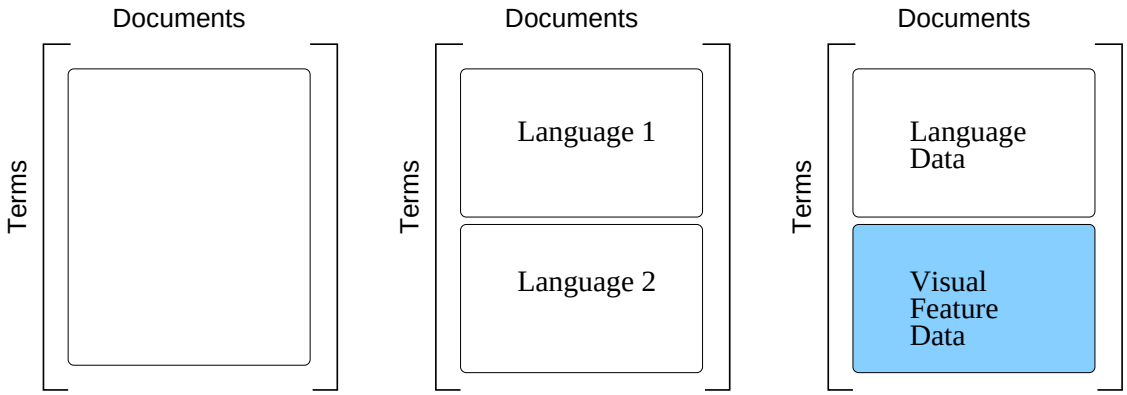
In the setup shown in Figure 1.2, user knowledge is captured at various locations of an integrated framework. Techniques such as that described in the previous sections may then be used to infer long-term semantic knowledge about the data.

1.2 Automatic semantic augmentation

Similarly to the methods described in the previous section, many of the automatic semantic augmentation techniques have their origins in the areas of textual document retrieval and statistical natural language processing.

1.2.1 Latent semantic indexing

One of the most influential text-based approaches whose main principles are still actively used in research to date is that of *latent semantic indexing* (LSI) [37]. Introduced as a means to tackle the problem of synonymy, i.e., the non-uniqueness of sets of words (or, terms) that can describe the same concept, the LSI method assumes existence of an underlying latent semantic structure in the textual data partially obscured by the randomness of word choice. In order to recover such latent structure, the method performs a truncated singular value decomposition of the original term-document co-occurrence matrix (see Figure 1.3(a)) transforming it into its reduced-rank approximation. Thus, the main idea behind this transformation is to capture the major term to document association relations ignoring minor differences in terminology. Finally, a cross-language variation of the LSI (CL-LSI), proposed by Landauer and Littman [86], that allows a query in one language to retrieve documents in another language can be considered a conceptual prototype of a whole family of automatic semantic augmentation methods [106, 121, 165, 166]. Indeed, as can be seen from a compar-



(a) LSI (b) CL-LSI (c) Multimedia-LSI

Figure 1.3: A comparison of vector space models in LSI-based methods

ison of term-document matrices for CL-LSI and Multimedia-LSI methods shown in Figures 1.3(b) and 1.3(c), one can easily replace the part that corresponds to the other language keyword information with multimedia feature data extracted from images, videos, etc. Thus, instead of retrieving documents in a language different from that of a query, it should be possible to find multimedia "documents" whose visual content corresponds to that of a query

specified by keywords, and vice versa. In other words, the same approach can be used to establish important associations between visual feature representation of multimedia and its corresponding semantics conveyed by the annotation keywords. And this is exactly the issue explored in detail by LSI-based automatic semantic augmentation methods such as those of Zhao and Grosky [166, 165] and others mentioned before.

1.2.2 Cross-language modeling

An analogous idea of treating the visual feature data as another language to translate keywords to and from is developed in a substantially different state of the art approach proposed by Barnard *et al.* [7, 6]. According to the adopted *translation model*, the authors consider the problem of object recognition as the one of machine translation. Given a representation in one form (image regions, or blobs, derived by clustering segmented images) they attempt to turn it into another form (keywords) using a developed model that acts in the capacity of a lexicon. Thus, the pairs of images and their respective annotation keyword sets are regarded as aligned bitexts, in which word-to-blob correspondence is to be established (see Figure 1.4 for an example). Finally, the sought correspondence is determined by optimizing

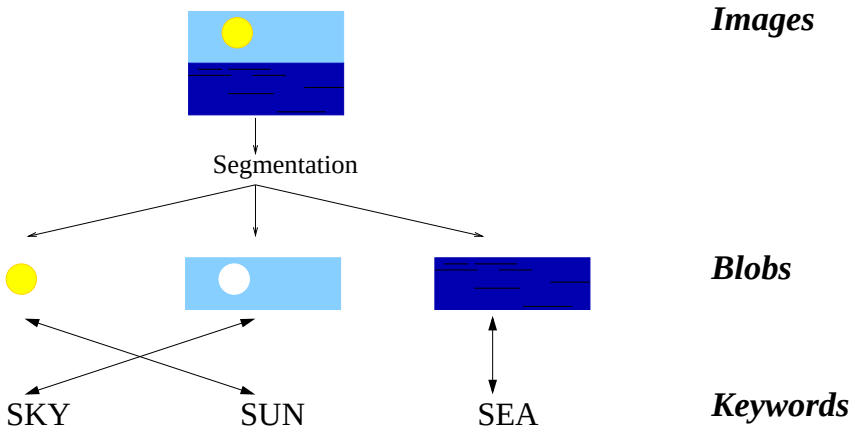


Figure 1.4: An example of correspondence between image regions (blobs) and annotation keywords sought by the translation model approach [7, 6]

the likelihood of word-to-blob association over all possible assignments, expressed as:

$$p(w|b) = \prod_{n=1}^{N} \prod_{j=1}^{M_n} \sum_{i=1}^{L_n} p(a_{nj} = i) t(w = w_{nj} | b = b_{ni}), \tag{1.7}$$

where N is the number of images, M_n is the number of keywords associated with the n-th image, L_n is the number of blobs that the n-th image is segmented into, $p(a_{nj} = i)$ is the probability of association of a particular blob b_i with a specific keyword w_j, and $t(w = w_{nj}|b = b_{ni})$ is the transition probability of word w given blob b. This likelihood is subsequently maximized via the EM algorithm [38]. A further development of these ideas by Jeon *et*

al. [69] lead to the *cross-media relevance model* for automatic image annotation and retrieval, while research efforts with a greater focus on various aspects of the underlying generative probabilistic models undertaken by Blei and Jordan [17] produced *correspondence latent Dirichlet allocation*, - a model that finds conditional relationships between latent variable representations of sets of image regions and sets of words. The latter method's properties were assessed through a comparison with two alternative hierarchical mixture models of image data and associated text (Gaussian-multinomial mixture model and Gaussian-multinomial latent Dirichlet allocation) demonstrating its superior performance on the applications of automatic image annotation, automatic image region annotation, and text-based image retrieval.

A method proposed by Vinokourov *et al.* [154] explores the similar cross-language paradigm for learning a semantic representation of web images and their associated text. In contrast to the above mentioned approaches [6, 7, 17, 69], these authors take a different route and choose not to model the latent semantic aspects via generative probabilistic schemes. Instead, they focus more on statistical techniques, namely, the kernel Canonical Correlation Analysis (KCCA) [155], originally developed for extracting the translation-invariant semantics from the aligned corpora in English and French, i.e., where every text in one language $x_i \in \mathcal{X}$ has a corresponding translation $y_i \in \mathcal{Y}$ in another language. The main hypothesis of such a technique is that having the corpus $\{x_i\}_{i=1}^N$ mapped to some high-dimensional feature space $\mathcal{F}_x$ as $\Phi(x_i)$ and corpus $\{y_i\}_{i=1}^N$ to $\mathcal{F}_y$ as $\Phi(y_i)$, it is possible to learn semantic directions $f_x \in \mathcal{F}_x$ and $f_y \in \mathcal{F}_y$ in those spaces so that the projections $(f_x, \Phi(x_i))_{i=1}^N$ and $(f_y, \Phi(y_i))_{i=1}^N$ of the original data in two different languages would be maximally correlated. This leads to a correlation coefficient maximization problem, formulated as given in (1.8):

$$\rho_{\mathcal{F}} = \max_{(f_x, f_y) \in \mathcal{F}} \frac{\sum_i (f_x, \Phi(x_i))\,(f_y, \Phi(y_i))}{\sqrt{\sum_i (f_x, \Phi(x_i))^2 \sum_j (f_y, \Phi(y_j))^2}}, \tag{1.8}$$

which, as the authors show, can be solved as a generalized eigenvalue problem. Of course, the same underlying formalism can be applied not only to extract translation-invariant semantics from aligned bilingual texts, but also to find correlations between, for instance, web images and their attached textual annotation [61, 154] and subsequently query an image database only by keywords.

1.2.3 Statistically motivated techniques

Considering the automatic semantic augmentation approaches we have mentioned so far, the powerful influence of the natural language processing and modeling perspectives is evident. However, there exist a number of methods derived from purely statistical premises, such as that of Mori *et al.* [109] based on dual clustering of visual and associated keyword information, also referred to as the *co-occurrence model* in some sources [7, 6, 69]. The authors propose to subdivide every image from the annotated collection into a number of non-overlapping segments, each of which inherits keywords associated with its corresponding image. Then, the visual feature representations of these segments are clustered by vector quantization and the

estimates of likelihood of each keyword for every cluster are derived by pooling the associated keyword frequencies. Having established such a clustering, the method then processes a query image with unknown annotation by subdividing it into segments and predicting its likely keywords from those of the clusters to which the query image segments are most similar.

A two-dimensional multiresolution hidden Markov model (2D MHMM) [92] is at the core of another statistical approach to automatic semantic augmentation proposed by Li and Wang [93]. The author's system for automatic linguistic indexing of pictures (ALIP) operates with a predefined number of semantic image categories, specified by sets of keywords according to the problem domain. For each category, the system profiles a 2D MHMM using as input the feature vectors extracted from training images at multiple resolutions and arranged on a pyramid grid. Once the training is complete, the set of the obtained 2D MHMM's together with the keywords of their corresponding categories are stored in a common dictionary of semantic concepts (see Figure 1.5). This dictionary can subsequently be used to annotate

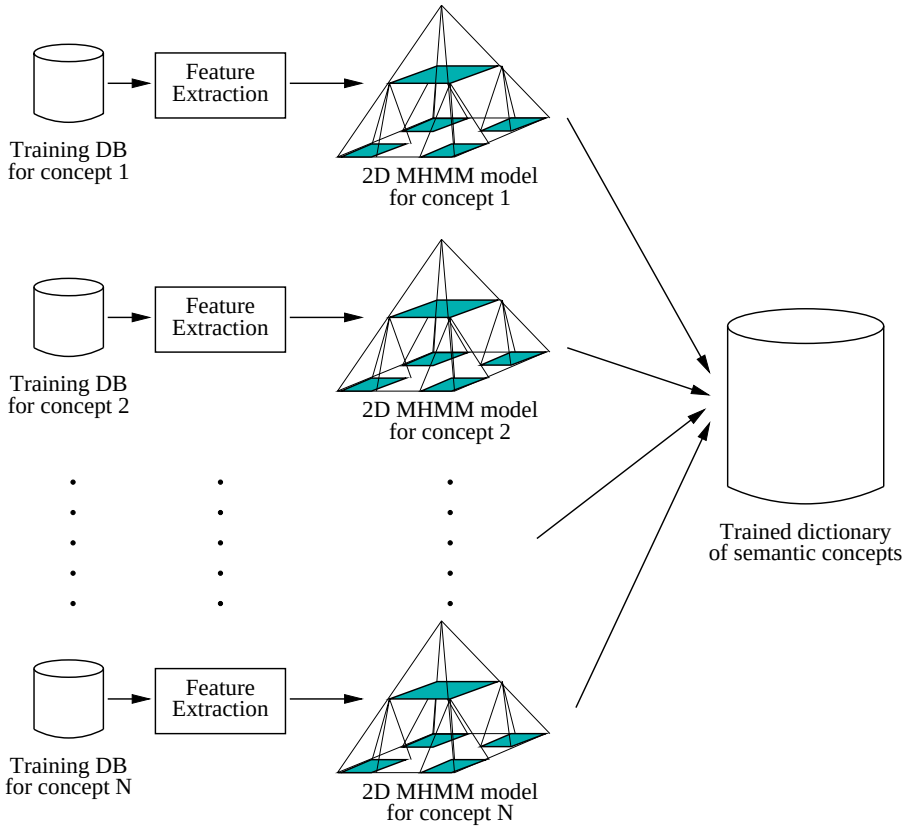


Figure 1.5: Structural design of the ALIP system [93]

new, i.e. not present in the training sets, images, which is done by selecting the keywords of the categories whose 2D MHMM's yield highest likelihoods computed from the features extracted from the images to be annotated.

The advent of supportvector machines [33, 152], whose excellent performance was backed up by strong results of the statistical learning theory [153], prompted the development of

a great number of classification-based methods. Starting with the direct application of support vector machines to color histogram data of digital images [24], there appeared more sophisticated techniques focused on introducing local and region-based visual feature sets [14] alongside the kernels suitable for such features [57], incorporating domain knowledge in terms of useful invariances [119], reformulating the problem in the multiple-instance learning context [2], and so on. Further improvements are continuously being made in this popular research area, which is evidenced by an ever-growing number of contributions [11, 34, 55, 120, 158].

1.3 Proposed research in context

While by no means complete from the point of view of providing a representative overview of all relevant methods, the above discourse serves the purpose of setting context for the research described in this thesis. The contributions presented in the chapters to follow belong to the group of automatic semantic augmentation methods, and are considered from the machine learning [103] perspective. From this vantage point, the proposed techniques are designed to be able to improve their performance based on previous experience and results in an autonomous fashion, in an attempt to eliminate the need for, or alleviate the burden placed on, human intuition in the analysis of a problem at hand. Even though the need for expert knowledge and human intuition may never be completely obviated due to the importance of clever engineering decisions in data representation and characterization, the machine learning approach provides a clear and incontestable advantage. Indeed, by solving a more general and, likely, far more difficult problem of machine learning, the same technology may be applied to a wide variety of particular problems without having to redesign the solution from scratch each time. That is, once solved, the same general method is applicable in the above mentioned scenarios of semantic augmentation for digital images, waveform audio signals, video shots, and in many other possible settings.

This thesis adopts the above machine learning perspective in approaching the problem of semantic augmentation of multimedia databases for efficient access and retrieval, and establishes a number of contributions in the areas outlined below.

- **Discriminant analysis**. Choosing the field of discriminant analysis as a starting point of this study reflects a deliberate bias and preference for a discriminative machine learning method over an alternative generative approach. This decision is motivated by a number of reasons, most important of which are the following [148]. First, a discriminative model has much more flexibility in the parts of input space where posterior probabilities differ significantly from 0 or 1, whereas generative approaches model details of input data dsitribution which may be irrelevant for determining posterior probabilities. Second, discriminative models are typically very fast at making predictions for test data, while generative models often require an iterative solution (albeit there definitely exist some pathological cases where generative models are fast and discriminative ones

are slow). Third, other things being equal, it is expected that discriminative methods would have better predictive performance since they are trained to predict the class label rather than the joint distribution of input data and labels. Thus, taking into account the specific properties of the intended application of semantic augmentation, we develop a distance-based transformational discriminant analysis technique, DDA. An extensive effort is undertaken to make the DDA formulation *non-parametric* asserting minimal assumptions on input data distribution, *asymmetric* to match the most popular deployment scenario in 1-againt-all classification, and *transformation-based* in order to allow for extensions, post-processing and the use of the derived transformation to provide a discriminative distance metric. This metric accounts for differences in the scales of different features, removes global correlations and redundancies among features to some extent, and adapts to the fact that some features may be much more informative about the class labels than others. In order to fulfill these conditions, the DDA's ability to extract discriminative features and reduce input data dimensionality, while determining the number of sufficient dimensions automatically, is of crucial importance. From the semantic augmentation point of view, DDA provides a binary learning machine to be used to discriminate between a certain high level semantic concept and its complement, e.g. to determine whether a digital image is showing an entity of interest or not.

- **Kernel-based methods**. Many linear machine learning algorithms have been enhanced through a special sort of functions known as *kernels*[1] in order to be able to deal with more complex learning problems requiring non-linear decision functions. Similarly, the DDA method is cast as a *non-linear discriminant analysis* technique, thereby overcoming the linearity assumption of the sought discriminative transformation, and naturally leading to the development of KDDA, a kernel extension of DDA. This is accomplished by reformulating the problem in terms of distances projected from a richer, possibly infinite-dimensional, feature space induced by a chosen kernel function. An important aspect distinguishing KDDA from other kernel-based methods is the convexity of the formulation that holds regardless of whether or not an underlying kernel is positive definite. This property is evaluated empirically by incorporating *indefinite kernels* in KDDA. In several application domains [163] these kernels are known to correspond to distance measures that better capture perceptual similarity. Additionally, an adverse condition referred to as the *false positive projection effect* is examined and its elimination strategies are assessed. With respect to semantic augmentation, the KDDA approach not only extends previous formulation to non-linear cases, but also lets the binary learning machine of DDA be effectively used with non-metric dissimilarity measures via indefinite kernels.

- **Classifer ensembles**. Rarely do we encounter a practical semantic augmentation prob-

[1] We defer the detailed discussion of kernels till Chapter 3.

lem that can be handled in its entirety by a single binary learning machine. A natural and, coincidentally, the most popular extension for the case with several classes, categories or target semantic concepts is to derive as many binary learning machines, one for each class label and, if needed, perform some arbitrage among their predictions. In the present study, it is argued that such construct and its variations are disadvantageous from the semantic augmentaion perspective, because of the limiting assumption of the set of target semantic concepts being *independent, non-overlapping, and exhaustive*. While providing a means to extend many existing techniques to a multiple-category case, this assumption may lead to inconsistent results, e.g. predicting fairly unlikely combinations of concepts such as "submarine" and "desert sand" for a test video shot, or estimating an error of misclassification of "river" as "lake" to be as important as misclassifying "river" as "fighter jet". In order to address these limitations, an approach is proposed to model explicitly the hierarchical semantic relationships among the target classes that are automatically derived and extended via a semantic lexicon. Practically, this method is implemented as a hierarchical semantic ensemble (HSE) of individual classifiers, realized as DDA binary learning machines that interact by influencing each other's decisions through the links mandated by the structure of relations among corresponding semantic concepts. Thus, HSE utilizes the earlier developed DDA classifiers as basic building blocks in a hierarchical structure of a more sophisticated learning machine, designed to be applied in the domain of semantic augmentaion with multiple inter-related target classes.

- **Theoretical issues**. In addition to the empirical evaluation of all of the proposed methods, an express effort is made to establish theoretical connections and analogies with some extant state-of-the-art machine learning methods. Through this anaylsis, an important detail of the DDA formulation reliance on several separating hyperplanes is highlighted and examined separately. The latter inquiry is performed in the context of margin-based classification, providing some new results on generalization performance of the classifier in question and thus revealing promising venues for future investigations.

1.4 Plan of the thesis

Chapter 2 covers in detail the linear formulation of the proposed distance-based discriminant analysis method. It accentuates some of its important properties alongside the implementational details, concluding with an empirical evaluation of the developed approach and a discussion with respect to a number of existing techniques. Chapter 3 presents an extension of the proposed distance-based discriminant analysis formulation to a kernel formulation, allowing more complex nonlinear machine learning problem to be solved within the same framework. The provided material also considers certain aspects specific to kernel-based methods, such as use of indefinite kernels and elimination of a false positive projection effect. Chapter 4 considers the distance-based discriminant analysis method from a different point of

view, deploying it as a basic building block classifier in a more sophisticated learning machine. The latter construct is a hierarchical semantic ensemble of classifiers that models explicitly the relationships among the target semantic classes in order to overcome the consequences of the limiting assumption of them being independent, non-overlapping and exhaustive. The experimental evaluation of the proposed method is carried out in comparison with alternative ensemble techniques, as well as considering different types of baseline classifiers. Chapter 5 is dedicated to a more detailed scrutiny of some theoretical issues linked to the formulation of the proposed distance-based discriminant analysis. It is followed by some closing remarks and perspectives for future work in Chapter 6.

Chapter 2

Distance-based discriminant analysis (DDA)

This chapter describes a discriminant analysis method whose development is driven by the motivation to create a technique with a range of properties both suitable and beneficial in the area of intended application, i.e. automatic semantic augmentation. The sought characteristics include:

- ability to perform discriminative *feature extraction* and *dimensionality reduction*, while possessing the means to determine how many dimensions are sufficient to distinguish among a given set of classes,
- *assymetry of formulation* suitable for the most popular deployment scenarios in 1-against-all classification, as well as in the case of data set imbalance,
- *transformational* and *non-parametric* specification that would allow for extensions, use as a discriminative data pre-processing technique, minimal assumptions on data distribution, and maximum utilization of the capabilities of the prospective classifier, such as nearest neighbor (NN) [44], to be used with the transformed data,
- ease of extension to a *multiple-category* case, together with the property of being *efficiently* approximated and computed.

The following sections provide a detailed account of the proposed method and the various aspects relating to its formulation, algorithmic specification, numerical implementaion, extensions, and experimental evaluation.

2.1 Problem formulation

Suppose that we seek to distinguish between two classes represented by data sets X and Y having N_X and N_Y m-dimensional observations, respectively. For this purpose, we are looking for such transformation matrix $T \in \mathbb{R}^{m \times k}$, $k \ll m$, such that $\{X \mapsto X', Y \mapsto Y'\}$,

that places instances of a given class near each other while relocating the instances of the other class sufficiently far away. In other words, we want to ensure that the compactness hypothesis [3] holds for either of the two classes in question, while its opposite is true for both.

While the above preamble may fit just about any class-separating discriminant analysis method profile (e.g., [21, 42, 49, 62, 159]), we must emphasize several important assertions that distinguish the presented method and naturally lead to the problem formulation that follows. First of all, we must reiterate that one of our primary goals is to improve the NN performance on the task of discriminant analysis. Therefore, the sought problem formulation must relate only to the factors that directly influence the decisions made by the NN classifier, namely - the distances among observations. Secondly, in order to benefit as much as possible from the non-parametric nature of the NN, the sought formulation must not rely on the traditional class separability and scatter measures that use class means, weighted centroids or their variants [48] which, in general, connote quite strong distributional assumptions. Finally, an asymmetric product form should be more preferable, justified as consistent with the properties of the data encountered in the target application area of multimedia retrieval and categorization [169], as well as beneficial from the viewpoint of insightful parallels to some margin-based state-of-the-art techniques considered in Chapter 5.

Let $d_{ij}^W(T)$ denote a Euclidean distance between observations i and j from transformed data set X' given a transformation matrix T, and, analogously, $d_{ij}^B(T)$ specify a distance between the i-th observation from data set X' and the j-th observation from data set Y', where superscripts "W" and "B" stand for within-class and between-class type of distance, respectively:

$$d_{ij}^W = \sqrt{(x_i - x_j)^T T T^T (x_i - x_j)}, \tag{2.1}$$

$$d_{ij}^B = \sqrt{(x_i - y_j)^T T T^T (x_i - y_j)}, \tag{2.2}$$

for $\{x_i\}_{i=1}^{N_X} \in \mathbb{R}^m$, $\{y_j\}_{j=1}^{N_Y} \in \mathbb{R}^m$. Using this notation, the sought discriminative data transformation can be obtained by minimizing the following criterion[1] :

$$J(T) = \frac{\left(\prod_{i<j}^{N_X} \Psi\left(d_{ij}^W(T)\right) \right)^{\frac{2}{N_X(N_X-1)}}}{\left(\prod_{i=1}^{N_X} \prod_{j=1}^{N_Y} d_{ij}^B(T) \right)^{\frac{1}{N_X N_Y}}}, \tag{2.3}$$

where the numerator and denominator of (2.3) represent the geometric means of corresponding distances, and $\Psi\left(d_{ij}^W(T)\right)$ denotes a Huber robust estimation function [67] parametrized

[1] Here and in several other places we will use shorthand $\prod_{i<j}^{N_X}$ to designate double product $\prod_{i=1}^{N_X} \prod_{j=i+1}^{N_X}$.

by a positive constant c and defined as:

$$\Psi(d_{ij}^W) = \begin{cases} \frac{1}{2}\left(d_{ij}^W\right)^2 & \text{if} \quad d_{ij}^W \leq c; \\ cd_{ij}^W - \frac{1}{2}c^2 & \text{if} \quad d_{ij}^W > c. \end{cases} \tag{2.4}$$

The choice of Huber function in (2.3) is motivated by the fact that at c the function switches from quadratic to linear penalty allowing to mitigate the consequences of an implicit unimodality assumption that the formulation of the numerator of (2.3) leads to. Additionally, Huber function has several attractive properties that greatly facilitate the derivation of the majorizing inequalities, as will be shown in section 2.2.2.

In the logarithmic form, criterion (2.3) is written as:

$$\begin{aligned} \log J(T) &= \frac{2}{N_X(N_X-1)} \sum_{i<j}^{N_X} \log \Psi\left(d_{ij}^W(T)\right) - \frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \log d_{ij}^B(T) \\ &= \alpha S_W(T) - \beta S_B(T), \end{aligned} \tag{2.5}$$

which highlights the theoretical underpinnings motivating the above formulation. Indeed, $\log J(T)$, being a weighted sum of log-barrier functions [111], may be viewed as an extended formulation of analytic center machine (ACM) method that finds a separating hyperplane as an analytic center of the classifier version space [146], discussed in greater detail in Chapter 5. For notational convenience, the first and the second summation terms of (R.1) are going to be referred to as $S_W(T)$ ("within" distances) and $S_B(T)$ ("between" distances) in the following discussion to allow for a more convenient notation and due to their apparent functional similarity with the notions of within- and between-class scatter measures used in a number of well-known discriminant analysis techniques [42, 43, 49, 62]. We will also shorten the notation by reassigning the normalizing quantities $\frac{2}{N_X(N_X-1)}$ and $\frac{1}{N_X N_Y}$ to α and β, respectively.

Although a straightforward differentiation of (R.1) might appear sufficient in order to proceed with a generic optimization search technique such as gradient descent, our preliminary experiments showed that the quality of the found solutions is severely impaired by the problems due to local minima and considerable degree of dependence on the initial starting value, as detailed in section 2.4. Moreover, the computational costs of such an endeavor very quickly become prohibitive[2] , especially if one adheres strictly to the main premise of this work, i.e., uses only pairwise distances among observations linked to quadratic complexity, as opposed to deviations from class means (linear complexity) of the customary class separability and scatter measures abundant in clustering literature. The computational cost situation will be further exacerbated if, in addition to the descent direction, a proper step length must be calculated, so that gradient descent does not overshoot and actually manages to improve the optimization criterion, while the latter outcome is guaranteed by the introduced below iterative majorization technique (and, hence its alternative name: "guaranteed

[2]Consider, for example, a small set of 500 images represented by 300-dimensional feature vectors. A brute force computation of the descent direction will entail $\approx 500^2$ pairwise distance calculations at each of the 300^2 elements of the sought matrix leading to $\approx 2 \cdot 10^{10}$ calculations.

descent"). Furthermore, some of the tested state-of-the-art optimization routines, such as SQP and Quasi-Newton with line search, did not scale well either and happened not to be able to converge, even on fairly simple data sets.

In order to avoid the above pitfalls, it was decided to derive some useful approximations of criterion (R.1) that would make the task of its optimization amenable to a simple iterative procedure based on the majorization method, which we discuss in the following section.

2.2 Iterative majorization

2.2.1 General overview of the method

As stated in [18, 149, 63], the central idea of the majorization method is to replace the task of optimizing a complicated objective function $f(x)$ by an iterative sequence of simpler minimization problems in terms of the members of the family of auxiliary functions $\mu(x, \bar{x})$, where x and $\bar{x}$ vary in the same domain Ω. In order for $\mu(x, \bar{x})$ to qualify as a *majorizing function* of $f(x)$, the auxiliary function $\mu(x, \bar{x})$ is required to fulfill the following conditions, for $x, \bar{x} \in \Omega$:

- the auxiliary function $\mu(x, \bar{x})$ should be simpler to minimize than $f(x)$,
- the original function must always be less or equal to the auxiliary function:

$$f(x) \leq \mu(x, \bar{x}), \tag{2.6}$$

- the auxiliary function should touch the surface of the original function at the *supporting point*[3] $\bar{x}$:

$$f(\bar{x}) = \mu(\bar{x}, \bar{x}). \tag{2.7}$$

To understand the principle of minimizing a function by majorization, consider the following observation [18]. Let the minimum of $\mu(x, \bar{x})$ over x be attained at x^*. Then, (2.6) and (2.7) imply the chain of inequalities

$$f(x^*) \leq \mu(x^*, \bar{x}) \leq \mu(\bar{x}, \bar{x}) = f(\bar{x}). \tag{2.8}$$

This chain of inequalities is named the *sandwich inequality* by De Leeuw [88], because the minimum of the majorizing function $\mu(x^*, \bar{x})$ is squeezed between $f(x^*)$ and $f(\bar{x})$. A graphic illustration of these inequalities is shown in Figure 2.1 for two subsequent iterations of iterative majorization of function $f(x)$. Thus, given an appropriate function $\mu(x, \bar{x})$, the iterative majorization (IM) algorithm proceeds as follows:

1. Assign an initial supporting point $\bar{x} = \bar{x}_0 \in \Omega$,
   choose tolerance ϵ;

[3]The similar notation will be used further on, where a dash over a variable name will signify that the variable either depends on or is itself a supporting point.

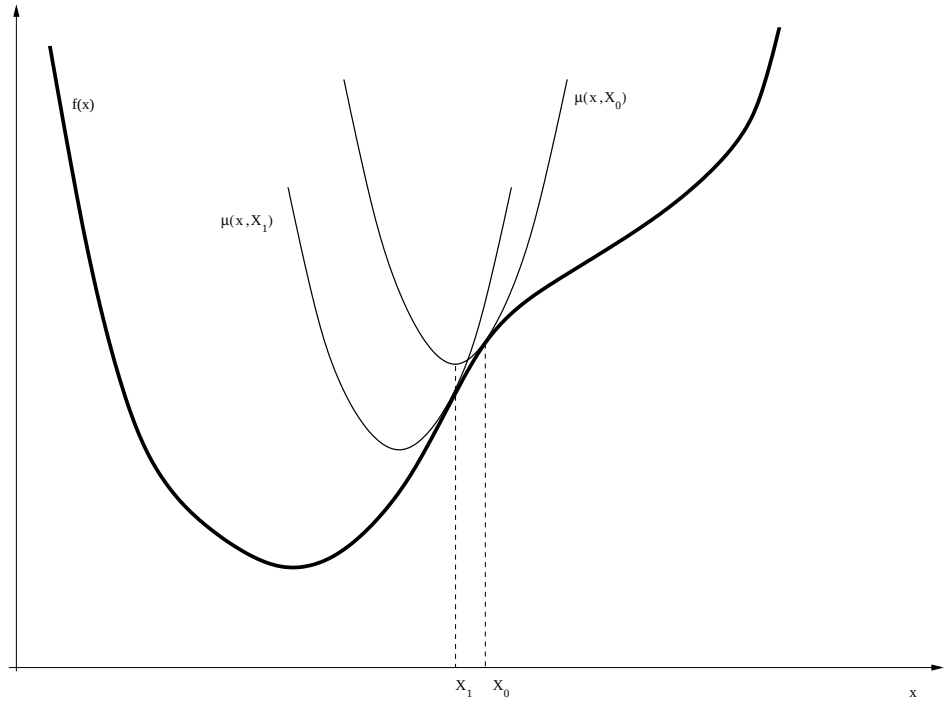


Figure 2.1: Illustration of two subsequent iterations of the iterative majorization method. The first iteration starts by finding the auxiliary function $\mu(x, x_0)$, which is located above the original function $f(x)$ and touches at the supporting point x_0. The minimum of the auxiliary function $\mu(x, x_0)$ is attained at x_1, where $f(x_1)$ can never be larger than $\mu(x_1, x_0)$. This completes one iteration. The second iteration proceeds analogously from supporting point x_1, and so on, until convergence.

2. Find a successor point $x_s : x_s = \arg\min_{x \in \Omega} \mu(x, \bar{x})$;

3. If $f(\bar{x}) - f(x_s) < \epsilon$, then stop;

4. Set $\bar{x} = x_s$, go to 2.

The essential property of the above procedure is that it generates a non-increasing sequence of function values, which converges to a stationary point whenever $f(x)$ is bounded from below and x is sufficiently restricted. As noted by Fletcher [45], the found point is in most cases a local minimizer. Furthermore, according to the results reported by Van Deun *et al.* [149], the majorization method has a valuable property of a low to negligible dependence on the initial value, compared to other applicable techniques. Another advantage of the majorization approach is due to the fact that there exist a number of specifically tailored global optimization techniques, such as objective function tunneling [18], that can be applied if the problem domain is abundant with low quality local minima. In the next section we will derive the majorizing expressions of (R.1) and show how they are used for optimizing the chosen criterion.

2.2.2 Majorizing the optimization criterion

It can be verified that majorization remains valid under additive decomposition. Therefore, a possible strategy for majorizing (R.1) is to deal with $S_W(T)$ and $-S_B(T)$ separately and subsequently recombine their respective majorizing expressions.

We begin by noting that the logarithm, as much as any other concave function, can always be majorized by a straight line $y = ax + b$ whose coefficients $a = 1/\bar{x}$ and $b = \log(\bar{x}) - 1$ are determined from the majorization requirements (2.6) and (2.7) rendering

$$\log(x) \leq \bar{x}^{-1}x + \log(\bar{x}) - 1. \tag{2.9}$$

Also, as previously reported in [29, 63], Huber distance (2.4) is convex and has a bounded second derivative, and hence can be majorized by a convex quadratic function:

$$\Psi(x) \leq \frac{1}{2}\bar{w}x^2 + \frac{1}{2}\left(\bar{v} + sign(\bar{x} - c)\bar{v}\right), \tag{2.10}$$

where $x > 0$, and coefficients $\bar{v}$ and $\bar{w}$ are defined as:

$$\bar{v} = \frac{1}{2}c\bar{x} - \frac{1}{2}c^2, \tag{2.11}$$

$$\bar{w} = \begin{cases} 1 & \text{if} \quad \bar{x} \leq c; \\ \frac{c}{\bar{x}} & \text{if} \quad \bar{x} > c. \end{cases} \tag{2.12}$$

Combining (2.9) and (2.10) together while substituting the result into the formulation of $S_W(T)$, we can obtain its majorizing expression $\mu_{S_W}(T, \bar{T})$:

$$\begin{aligned} S_W(T) &= \sum_{i<j}^{N_X} \log \Psi\left(d_{ij}^W(T)\right) \\ &\leq \sum_{i<j}^{N_X} \frac{\bar{w}_{ij} \cdot \left(d_{ij}^W(T)\right)^2}{2\Psi\left(d_{ij}^W(\bar{T})\right)} + K_1 \\ &= \mu_{S_W}(T, \bar{T}), \end{aligned} \tag{2.13}$$

where $T, \bar{T} \in \mathbb{R}^{m \times m}$, $\bar{T}$ is a supporting point for T, $\bar{w}_{ij}$ is a weight of the Huber function majorizer, that in this case is equal to 1 if $\Psi(d_{ij}^W(\bar{T})) < c$ or $c/\Psi(d_{ij}^W(\bar{T}))$ otherwise, and K_1 is a constant term that collects all of the other terms that are irrelevant from the point of view of minimization with respect to T. Switching to matrix notation (see Appendix A for derivation details), we define a square symmetric matrix R:

$$r_{ij} = \begin{cases} -\dfrac{\bar{w}_{ij}}{\Psi\left(d_{ij}^W(\bar{T})\right)} & \text{if} \quad i \neq j; \\ -\displaystyle\sum_{k=1, k\neq i}^{N_X} r_{ik} & \text{if} \quad i = j; \end{cases} \tag{2.14}$$

which lets us rewrite the majorizing expression of $S_W(T)$ in its final form, as follows:

$$\mu_{S_W}(T, \bar{T}) = \frac{1}{2}\mathbf{tr}\left(T^T X^T R X T\right) + K_1. \tag{2.15}$$

An attempt to majorize $-S_B(T)$ directly runs into problems due to the difficulties of finding a proper quadratic majorizing function of the negative logarithm. As a practical solution we consider two alternative replacements of $-\log(x)$ in $-S_B(T)$:

- a piece-wise linear approximation,
- a second order Taylor expansion.

According to the first alternative, we replace the neg-logarithm with its piece-wise linear approximation (see an illustration in Figure 2.2), which, in turn, can be represented as a sum

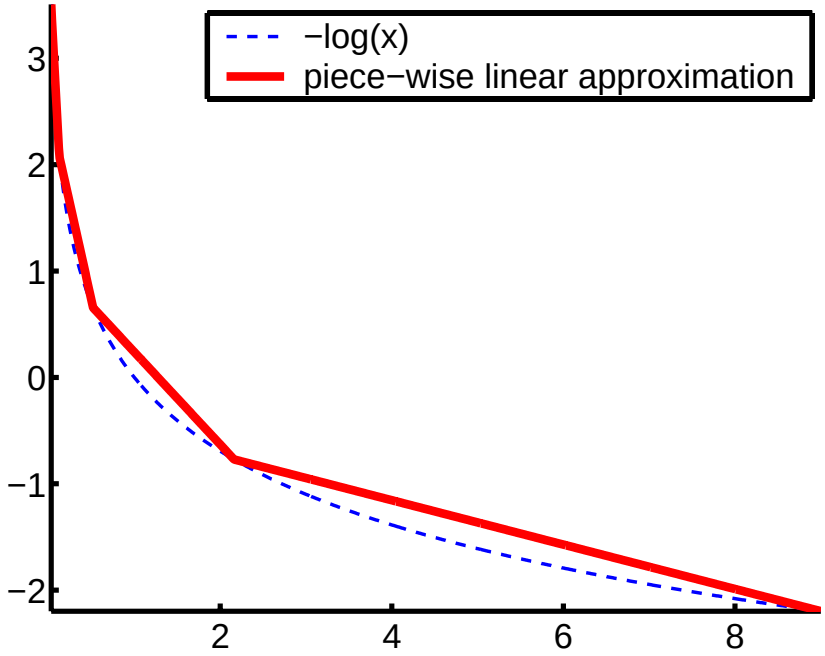


Figure 2.2: Piece-wise linear approximation of $-\log(x)$

of the functions defined as:

$$g(x; x_0, l, r) = \begin{cases} r(x - x_0) & \text{if} \quad x \geq x_0, \\ -l(x - x_0) & \text{if} \quad x < x_0; \end{cases} \tag{2.16}$$

where $l + r > 0$, to ensure convexity. It is easy to see that the family of functions defined in (2.16) is one of the many possible generalizations of the absolute value function $|x|$, the former being equivalent to the latter whenever $x_0 = 0$ and $l = r = 1$. Similarly to $|x|$, $g(x; x_0, l, r)$ can be majorized by a quadratic $ax^2 + bx + c$ with coefficients

$$a = \frac{r + l}{4|\bar{x} - x_0|}, \tag{2.17}$$

$$b = \frac{r - l}{2} - \frac{(r + l)x_0}{2|\bar{x} - x_0|}, \tag{2.18}$$

$$c = \frac{(r + l)x_0^2}{4|\bar{x} - x_0|} + \frac{(l - r)x_0}{2} + \frac{(r + l)|\bar{x} - x_0|}{4}, \tag{2.19}$$

for a supporting point $\bar{x}$ and $a > 0$, b and c determined directly from the majorization requirements (2.6) and (2.7). Figure 2.3 depicts an example of a function from $g(x; x_0, l, r)$ family alongside its majorizer. The final expression of the majorizer based on the piece-wise

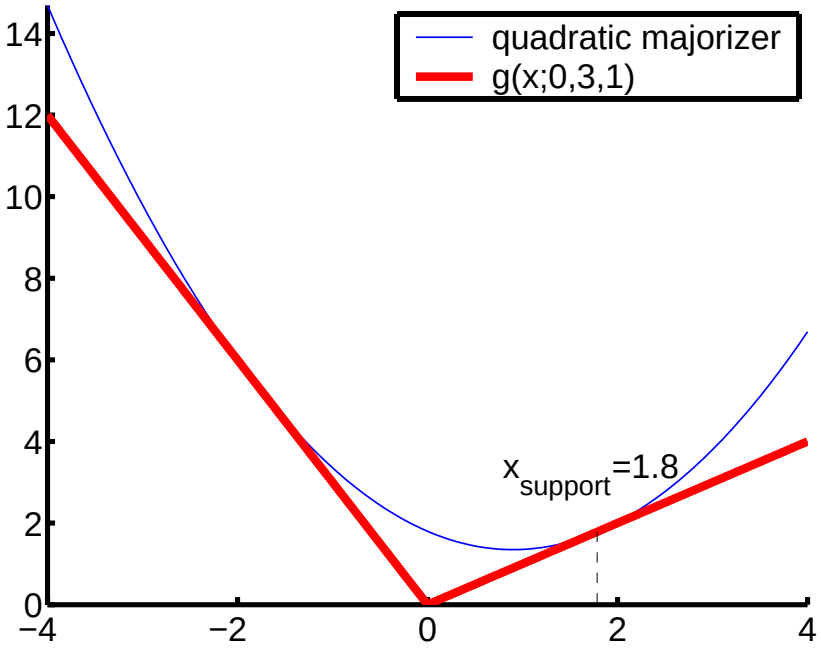


Figure 2.3: Example of a quadratic majorizer of $g(x; 0, 3, 1)$ around supporting point $\bar{x} = 1.8$

linear approximation, as shown in the Appendix A, is quite unwieldy and computationally costly even for a moderate number of g-family functions comprising the approximation. For this reason, we chose the other solution provided by a Taylor series expansion, as a faster and more stable alternative[4] .

Following the second approach, we express every term of $S_B(T)$ using a second order Taylor series expansion of the logarithm function around a supporting point $\bar{T}$:

$$\log\left(d_{ij}^B(T)\right) \approx -\frac{1}{2}\left(\frac{d_{ij}^B(T)}{d_{ij}^B(\bar{T})}\right)^2 + 2\frac{d_{ij}^B(T)}{d_{ij}^B(\bar{T})} + \log\left(d_{ij}^B(\bar{T})\right) - \frac{3}{2}. \tag{2.20}$$

Substituting (2.20) into the expression of $-S_B(T)$ leads to:

$$\begin{aligned} -S_B(T) &= -\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\log d_{ij}^B(T) \\ &\approx \frac{1}{2}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\left(\frac{d_{ij}^B(T)}{d_{ij}^B(\bar{T})}\right)^2 - 2\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\frac{d_{ij}^B(T)}{d_{ij}^B(\bar{T})} + K_2, \end{aligned} \tag{2.21}$$

[4] A more detailed analysis may demonstrate that resorting to the Taylor series approximation might break compliance with the majorization requirements in the strict sense. However, the empirical evidence proved otherwise (see section 2.4), confirming the technique as an alternative of preference.

where K_2 is a constant term that collects all of the other terms that are irrelevant from the point of view of minimization with respect to T. One may notice that in (2.21) only the second term, the sum of appropriately scaled negative Euclidean distances, requires majorization since the other two are either constant with respect to T or given as a quadratic which is simple enough to handle as is.

In order to find a majorizing expression of (2.21) we will make use of a well-known fact frequently mentioned in literature [18, 29, 149, 63], stating that the negative of a Euclidean distance is linearly majorizable:

$$-||x|| \leq -\frac{\bar{x}^T x}{||\bar{x}||} \tag{2.22}$$

which is a direct consequence of the Cauchy-Schwarz inequality $||x||||\bar{x}|| \geq \bar{x}^T x$. Switching to matrix notation (see Appendix A for derivation details), we define a square symmetric matrix G of size $N = N_X + N_Y$, such that[5] :

$$g_{ij} = \begin{cases} -\dfrac{1}{\left(d_{ij}^B(\bar{T})\right)^2} & \text{for} \quad i \in [1; N_X] \\ & \text{and} \quad j \in [N_X + 1; N], \\ -\dfrac{1}{\left(d_{ij}^B(\bar{T})\right)^2} & \text{if} \quad i \in [N_X + 1; N] \\ & \text{and} \quad j \in [1; N_X], \\ -\displaystyle\sum_{k=1, k\neq i}^{N_X+N_Y} g_{ik} & \text{if} \quad i = j, \end{cases} \tag{2.23}$$

which, combined with the result of (2.22) substituted into (2.21), lets us derive the majorizing expression for $-S_B(T)$ in its final form, as follows:

$$\mu_{-S_B}(T, \bar{T}) = \frac{1}{2}\mathbf{tr}(T^T Z^T G Z T) - 2\mathbf{tr}(T^T Z^T G Z \bar{T}) + K_2, \tag{2.24}$$

where Z is the matrix obtained by joining X and Y together, row-wise:

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}. \tag{2.25}$$

Finally, combining results (2.15) and (2.24), we obtain a majorizing function of the $\log J(T)$ optimization criterion:

$$\begin{aligned} \mu_{\log J}(T, \bar{T}) &= \alpha\mu_{S_W} + \beta\mu_{-S_B} \\ &= \frac{\alpha}{2}\mathbf{tr}\left(T^T X^T R X T\right) \\ &\quad + \frac{\beta}{2}\mathbf{tr}(T^T Z^T G Z T) - 2\beta\mathbf{tr}(T^T Z^T G Z \bar{T}) + K_3, \end{aligned} \tag{2.26}$$

[5]The elements g_{ij} of matrix G not affected by the first two rules of (2.23) are assumed to have been initially set to zero.

that can be used to find an optimal transformation T minimizing $\log J(T)$ criterion via the iterative procedure outlined in section 2.2.1. Similarly to the expressions shown in (2.13) and (2.21), K_3 is a constant term that collects all of the other terms that are irrelevant from the point of view of minimization with respect to T.

2.2.3 Minimization of the majorizer of $\log J(T)$

It is possible to minimize (R.2) with respect to T in a straightforward fashion by setting its derivative to zero and solving the resulting system of linear equations with any of the computationally efficient methods, such as QR decomposition [50]. However, it is often recommended [8, 83, 87] that a length-constrained (or, regularized, as usually referred to in the domains of signal processing, inverse problems [13] and regularized risk minimization [152]) solution be found by deploying such techniques as weight-limiting, weight decay, etc., especially in the case of classifiers capable of achieving zero training error, to prevent overfitting and thus improve generalization performance of the classifier. In order to find an optimal transformation T that satisfies the length constraint, we first form the Lagrangian function

$$\mathcal{L} = \mu_{\log J}(T, \bar{T}) + \lambda(\mathbf{tr}(T^T T) - \Delta), \tag{2.27}$$

where λ is a Langrangian multiplier and Δ is the value of the length constraint that is estimated from the classification performance on a validation data set [103]. It follows from (2.27) that an optimal solution T is:

$$T = (M + 2\lambda I)^{-1} L \tag{2.28}$$

where M is defined as $\frac{\alpha}{\beta} X^T R X + Z^T G Z$, L is equal to $2Z^T G Z \bar{T}$, and I is an identity matrix. Plugging (2.28) back into the expression of the length constraint, we obtain the following:

$$\begin{aligned} \Delta &= \mathbf{tr}\left(L^T (M + 2\lambda I)^{-1} (M + 2\lambda I)^{-1} L\right) \\ &= \mathbf{tr}\left(L^T U \frac{1}{(2\lambda I + D)^2} U^T L\right). \end{aligned} \tag{2.29}$$

where U and D are the respective matrices of eigenvectors and eigenvalues of M. Here, we have used the fact that symmetric matrices M and $M + 2\lambda I$ have the same eigenvectors, while the eigenvalues of $M + 2\lambda I$ are equal to those of M increased by 2λ. Also, to simplify the notation of (2.29), the reciprocal and squaring operations should be understood as applied to the diagonal matrix D on the element by element basis taking into account the magnitudes of each eigenvalue so as to avoid division by zero problems. Clearly, (2.29) is an equation of one variable λ with a computable derivative, that is easily solved by any suitable root-finding technique, such as Newton-Raphson method, or with a method specifically tailored to solving this type of problems, commonly referred to as a TRS, i.e. *trust region problem* [107, 126]. Once the constraint-satisfying value λ has been found, the optimal transformation T, i.e. the successor point in the iterative majorization algorithm is recovered as:

$$T_s = U\left(2\lambda I + D\right)^{-1} U^T L, \tag{2.30}$$

where the bracketed expression is a diagonal matrix whose inverse is easily computed through the reciprocal of the diagonal elements.

It should be mentioned that for the problems such as minimization of (R.2) the universally suggested approach [63, 75] is to decompose the design matrices of each quadratic component of the function being optimized into a sum of a diagonal positive definite and a negative definite matrices, and use the definiteness property to derive another majorizing inequality. This method, although theoretically sound and well-justified, in our experiments demonstrated a significantly slower rate of convergence induced by larger condition number of the matrices involved, and thus was subsequently replaced by the solution defined in (2.30), even though the latter method involves a costly eigendecomposition operation.

2.3 Putting it all together

2.3.1 Complete algorithm

Considering all of the derivations we have desribed so far, the complete distance-based discriminant analysis (DDA) algorithm for iterative majorization of $\log J(T)$ criterion (R.1) can be specified as follows:

Algorithm DDA.

1. Assign an initial supporting point $\bar{T} = \bar{T}_0 \in \mathbb{R}^{m \times m}$;
2. Find a successor point T_s using (2.30);
3. If $\log J(\bar{T}) - \log J(T_s) < \epsilon$, then stop;
4. Set $\bar{T} = T_s$, go to 2.

2.3.2 Dimensionality reduction

Observe that setting the column size of T to an arbitrary value $k \ll m$ renders the presented method of DDA a dimensionality reduction technique[6] that may be used in a variety of applications such as feature selection, low-dimensional data visualization, etc. Moreover, the value of k, i.e., the exact number of dimensions the data can be reduced to without loss of discriminatory power with respect to (R.1), is precisely determined by the number of non-zero singular values of T. Indeed, the distances between the transformed observations may be viewed as distances between the original observations in a different metric TT^T, that can

[6] A word of caution is in order as for the choice of $k = 1$, which corresponds to an ill-posed combinatorial problem [18].

be expressed as $TT^T = USV^TVSU^T = U_kS_k^2U_k^T$ using the singular value decomposition of T. The obtained expression reveals that the effect of the full-dimensional transformation T is captured by the first k left-singular vectors of T scaled by the corresponding non-zero singular values, whose number gives an answer to the question of how many dimensions are needed in the transformed space.

A summary of various other properties that distinguish DDA from existing dimensionality reduction methods is provided in section 2.5.

2.3.3 Multiple class discriminant analysis

While the above discussion is concentrated mostly on the two-class configuration, it is straightforward to generalize the presented formulation to a multiple-class discriminant analysis setting, for the number of classes $K \geq 2$:

$$\log J_K(T) = \sum_{i=1}^{K-1} \left(\alpha^{(i)} S_W(T)^{(i)} - \beta^{(i)} S_B(T)^{(i)} \right), \tag{2.31}$$

for per-class quantities of (R.1) indexed by superscript (i). Note that (R.4) becomes exactly (R.1) for the two-class formulation, when $K = 2$. Again, similarly to the latter case, the particular class to be left out may be determined using domain knowledge, or via statistical techniques, i.e., by maximum within-class variance in the original feature space, etc. In order to accommodate the changes required for adopting (R.4), the individual matrices R and G from (2.15) and (2.24) will be replaced with

$$R_K = \sum_{i=1}^{K-1} \frac{\alpha^{(i)}}{\beta^{(i)}} R^{(i)}, \text{ and} \tag{2.32}$$

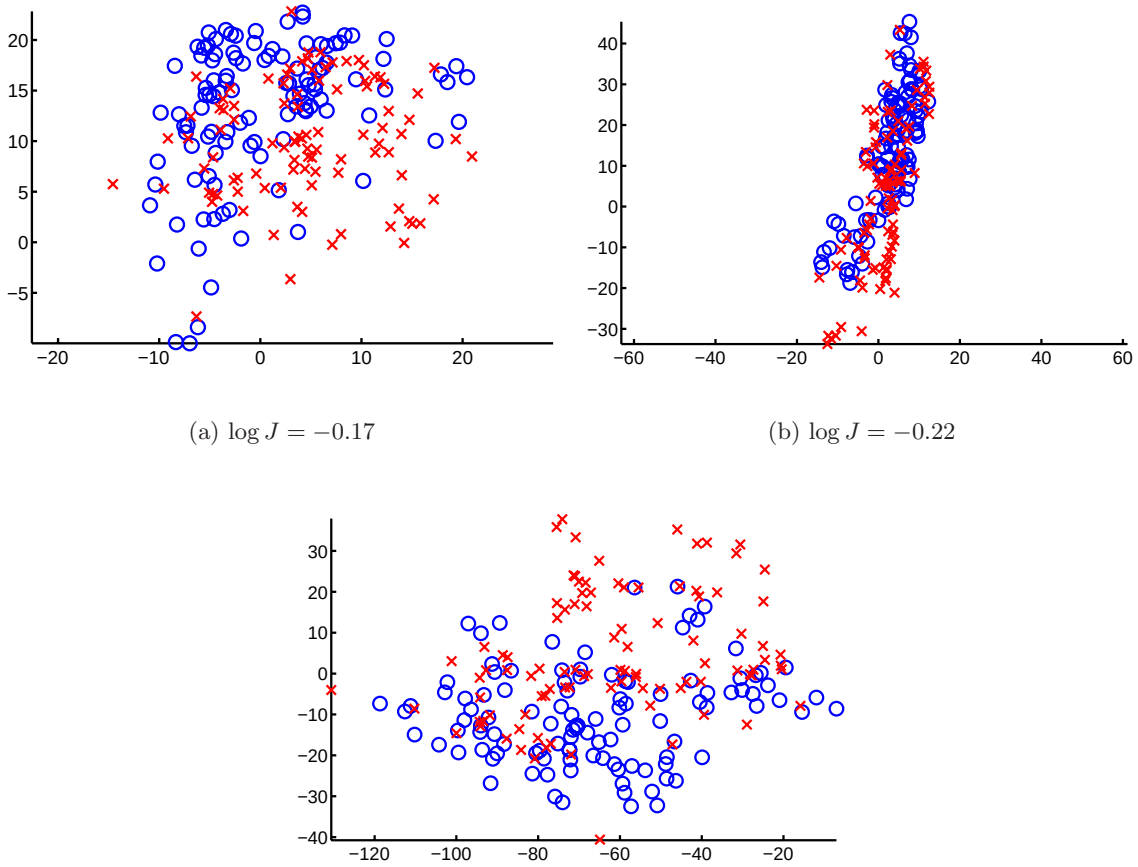
$$G_K = \sum_{i=1}^{K-1} G^{(i)}, \tag{2.33}$$

respectively, where each of the matrices $R^{(i)}$ is computed according to (2.14) using observations from class i, while matrices $G^{(i)}$ are calculated as indicated in (2.23) with proper index interval adjustment for computing distances between data points of a given class i and the rest of the data set.

2.4 Experimental results

2.4.1 UCI Benchmark data set performance

Our preliminary empirical analysis was based on data sets from the UCI Machine Learning Repository [16]. First of all, we verified that the solutions of the optimization problem formulated in section 2.1 found by the proposed method were of better quality and less dependent on the choice of the initial value compared to those of generic techniques, confirming the results reported by Van Deun [149] and Webb [161]. Indeed, numerous random initializations



(a) $\log J = -0.17$

(b) $\log J = -0.22$

(c) $\log J = -0.19$

Figure 2.4: Sonar data: local minima-prone solutions found by the gradient descent method. The target dimensionality of the sought discriminative subspace was set to $k = 2$

of the gradient search led to inferior as well as unstable results reflected in higher values of $\log J$ (see examples for the Sonar data set in Figure 2.4), while the IM-based method regularly reached far lower criterion values, as seen in Figure 2.5, and proved nearly insensitive to the choice of the initial supporting point. In addition to that, we thoroughly verified that the convergence property of the IM procedure was indeed preserved, as illustrated in Figure 2.5, despite the use of a Taylor series approximation in the derivation of (R.2). Finally, we validated the proposed dimensionality reduction technique by analysing how the classification performance varied with respect to k, the dimensionality of the transformed space, and how it was related to the number of non-zero singular values of the full-dimensional transformation, an example of which for the Sonar data set is depicted in Figure 2.6. Figure 2.6(b) plots 10 largest singular values of the full-dimensional transformation, in descending order, while Figure 2.6(a) documents the results of 10-fold cross-validation performance with respect to the transformed space dimensionality. It is easy to see that the singular values beyond the 7^{th} one are virtually zero, which corresponds to the point after which increasing

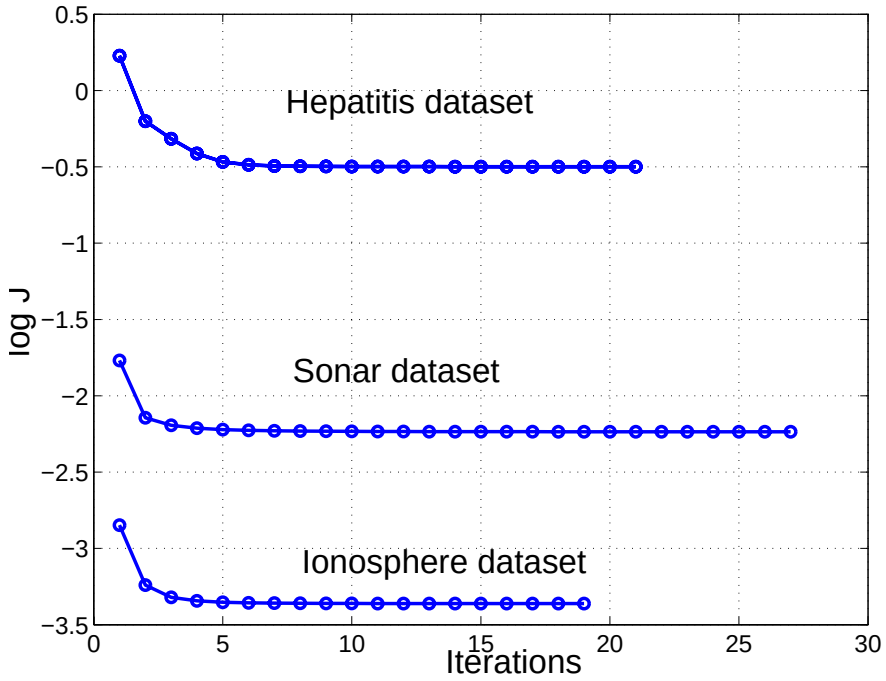
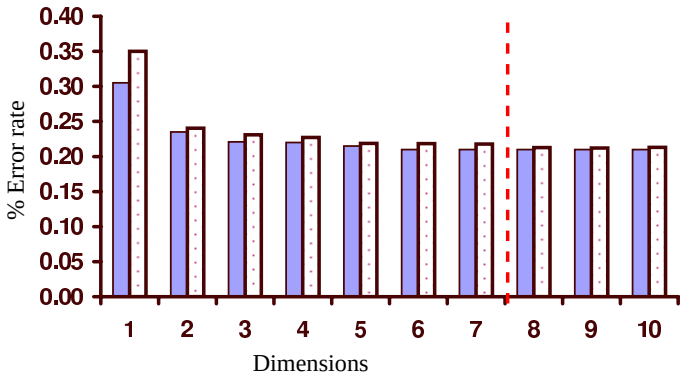


Figure 2.5: Convergence of the iterative majorization procedure in the DDA method. The horizontal and vertical axes correspond to the iteration number and optimization criterion value, respectively.

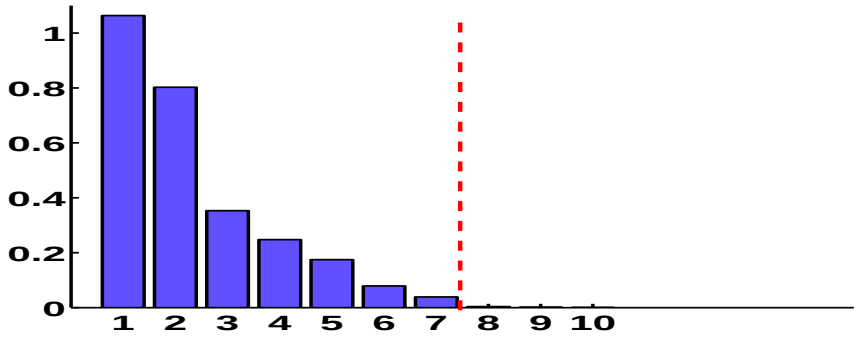
the transformed space dimensionality, by either setting k to a particular value (dot-filled bars) or using a larger number of appropriately scaled left-singular vectors (shaded bars), no longer significantly improves the classification performance, as confirmed by Chow test for structural change [26] at 99% confidence.

Further, the results of classification performance in terms of error rate of two types of experiments were compared. For the first type of experiments, which we will refer to as simply "NN" experiments, we measured classification error rate of the NN classifier using 10-fold cross-validation [162]. In the second type of experiments, that are going to be called "DDA+NN" experiments, an additional stage of applying a discriminating transformation T derived with the proposed DDA method prior to measuring the cross-validation performance of the NN classifier was introduced. Therefore, the goal of this analysis was to assess the effect of applying a DDA transformation on the accuracy of the NN classifier.

Several well-known data sets from the UCI Machine Learning Repository [16] were used in our experiments. All of the available data from each data set were utilized on the "as is" basis without performing any preprocessing, such as feature expansion for categorical, discrete or binary attributes. For some datasets, specific instructions were supplied as for partitioning the data into the training and testing portions, in which cases cross-validation procedure was not applied. The summary of important characteristics of the data sets used for testing is shown in Table 2.1. The error rates of NN and DDA+NN data classification experiments



(a) Classication error rate



(b) Singular values

Figure 2.6: Dimensionality reduction experiments: classification performance results and singular values of the transformation matrix. The dashed lines mark the boundary that determines the sufficient dimensionality of the transformed space.

averaged over twenty trial cross-validation runs are presented in Table 2.2. The obtained results confirm our conjecture about the positive effect of applying the DDA transformation on the accuracy of the NN classifier showing an improvement in performance (see Table 2.2).

2.4.2 Low-level feature representation

In order to assess the proposed DDA method in the context of the semantic augmentation domain, we perform a number of basic experiments of visual object recognition, categorization and semantic retrieval, where multimedia data is provided in the form of digital images and an algorithm is examined to determine how well it can learn the associated semantic information. Before detailing these experiments, however, we take a closer look at the low-level visual feature representation of the said image data, as extracted by the *Viper* system [141].

Table 2.1: Summary of data set characteristics

Data set	Classes	Attributes	Examples
Hepatitis	2	19	155
Ionosphere	2	34	200
Diabetes	2	8	768
Heart	2	13	270
Monk's Problem 1	2	6	432
Balance	3	4	625
Iris	3	4	150
DNA	3	180	2000
Vehicle	4	18	846

Table 2.2: Classification results for UCI data sets

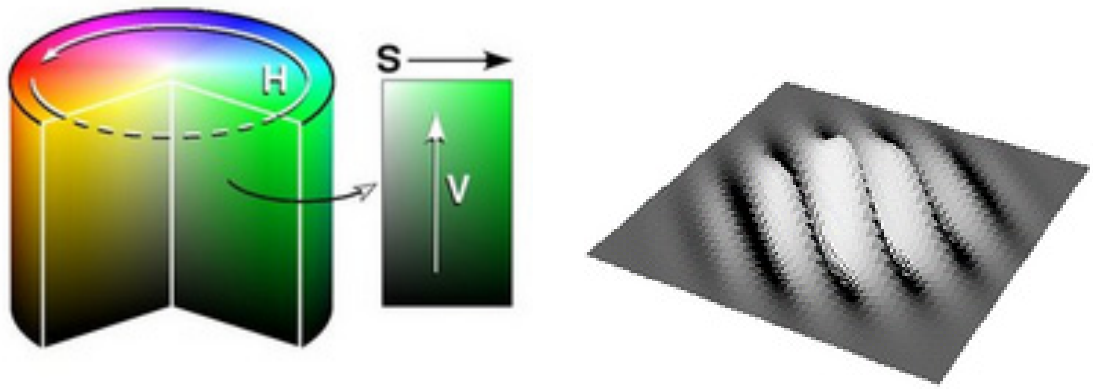
Data set	% Error of NN	% Error of DDA+NN
Hepatitis	29.57	0.00
Ionosphere	13.56	7.14
Diabetes	30.39	27.11
Heart	40.74	21.11
Monk's P1	14.58	0.69
Balance	21.45	3.06
Iris	4.00	3.33
DNA	23.86	6.07
Vehicle	35.58	24.70

Viper uses a palette of 166 colors, derived by uniformly quantizing the cylindrical *HSV* color space (see an illustration shown in Figure 2.7(a)) into 18 hues, 3 saturations, and 3 values. These are augmented by 4 gray levels. This choice of quantization means that more tolerance is given to changes in saturation and value, which is desirable since these channels can be affected by lighting conditions and viewpoint. The choice of the *HSV* color space is due to its perceptual uniformity and a relatively low complexity of computation and inversion in comparison to such alternatives as *CIE-LUV* and *CIE-LAB* [137].

As for the texture features, *Viper* employs a bank of real, circularly symmetric Gabor filters, see Figure 2.7(b), defined in the spatial domain by:

$$f_{mn}(x,y) = \frac{1}{2\pi\sigma_m^2} e^{-\frac{x^2+y^2}{2\sigma_m^2}} \cos[2\pi(u_{0_m} x\cos\theta_n + u_{0_m} y\sin\theta_n)], \tag{2.34}$$

where m indexes the scales of the filters, and n their orientations. The center frequency of



(a) HSV cylindrical color space: the hue angle is determined by pixels' color from the visual spectrum, the saturation is a measure of whiteness of the color, and the value is a measure of the pixel brightness

(b) Texture via Gabor filter: a set of image features is obtained by convolving an image with a number of Gabor filters providing a local energy estimate

Figure 2.7: Color and texture visual information

the filter is specified by u_{0_m}. The half-peak radial bandwidth is given by:

$$B_r = \log_2 \left(\frac{2\pi\sigma_m u_{0_m} + \sqrt{2\ln 2}}{2\pi\sigma_m u_{0_m} - \sqrt{2\ln 2}} \right), \tag{2.35}$$

where B_r is chosen to be 1, i.e. a bandwidth of one octave, which then allows us to compute σ_m:

$$\sigma_m = \frac{3\sqrt{2\ln 2}}{2\pi u_{0_m}}. \tag{2.36}$$

The highest center frequency is $u_{0_1} = \frac{0.5}{1+\tan(1/3)} \approx 0.5$, so that it is within the discrete frequency domain. The center frequency is halved at each change of scale, which implies that σ is doubled (2.36). The orientation of the filters varies in steps of $\pi/4$, and three scales are used. These choices result in a bank of 12 filters, which renders appropriate coverage of the frequency domain with little overlap between the filters. Given the 10 band energy quantization, this design provides 120 global texture characteristics of the image. Combining this information with the color data, we obtain a common 286-dimensional feature vector representation for every image.

2.4.3 Application to visual object recognition

For our object recognition experiments we chose a recently developed database ETHZ80 for object categorization and recognition composed of entities corresponding to the basic level of human knowledge organization [89]. The database contains high-resolution color images of

Figure 2.8: The 8 classes of objects of the ETHZ-80 database. Each class contains 10 objects with 41 views per object, for a total of 3280 images

80 objects from 8 different classes, for a total of 3280 images, an overview of which is shown in Figure 2.8.

The training set comprised images taken one per class object viewed from a fixed position, while the rest (3200 images) was allocated to the test set. An illustration of a training set image from class "car" and several test set images is provided in Figure 2.9. Again,

Figure 2.9: An illustration of images of the same class used in the training (leftmost) and test (the rest) sets

similarly to the setup described above (see section 2.4.1), we compared performance results for "NN"and "DDA+NN" experiments for each of the 8 classes, but this time, using a one-against-all classification configuration typically encountered in ensemble learning [40], and setting target dimensionality to 2D according to the magnitude of the transformation singular values as explained in section 2.3.2. The results are summarized in Table 2.3.

Table 2.3: Object recognition results for the ETHZ80 image database

Object class	% Error of NN	% Error of DDA+NN (unconstrained)	% Error of DDA+NN (constrained)
(1) Apple	4.47	18.66	0.75
(2) Car	14.47	18.72	5.78
(3) Cow	12.12	16.91	10.97
(4) Cup	3.09	16.94	2.22
(5) Dog	14.00	16.66	12.72
(6) Horse	14.47	14.84	13.16
(7) Pear	6.13	18.94	3.84
(8) Tomato	2.50	16.87	1.88

It is importnant to emphasize here that image representation for these experiments was

reduced via DDA to two dimensions only. Nevertheless, as shown in the last column of Table 2.3, the proposed technique still was able to descrease recognition error rate, which improved the overall performance average. The results in Table 2.3 also reveal the importance of the length constraint (or, regularization), introduced in (2.27), for the purpose of avoiding data over-fitting problems. Both unconstrained and length-constrained solutions found by the DDA procedure lead to zero error rate on the training data, but, as can be easily seen from Table 2.3, their performance turned out to be drastically different on the test data sets, demonstrating an adequate generalization capability induced by the length-constrained version of the proposed method. Consistent with the figures reported earlier for color- and texture-based feature sets [89], the error rates are highest for classes 3, 5 and 6. An example of the 2D representation of the training set for image class 2 obtained by DDA is shown in Figure 2.10. As can be easily seen from the figure, the target class images are well separated from those of all of the other classes seen to be freely mixed together in the derived 2D discriminative subspace, which is exactly the requirement one seeks to satisfy in one-against-all classification. Additionally, the separation margin visually noticeable in

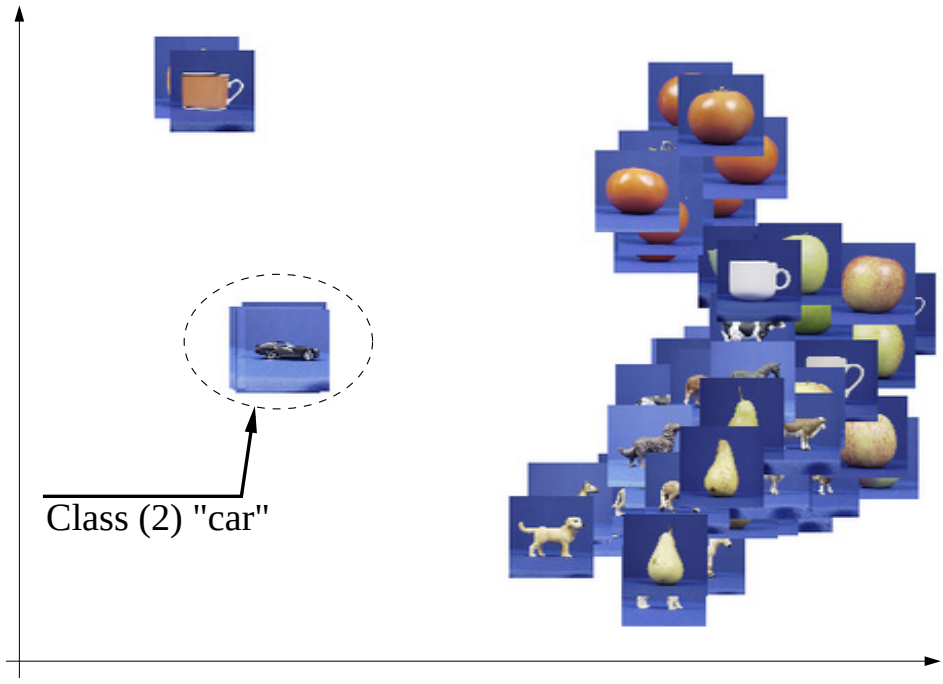


Figure 2.10: Result of applying a discriminative dimensionality-reducing (286 to 2) DDA transformation to the training set for recognition of objects from class (2) "car". Images from class 2 are projected close to each other while images belonging to the other classes are freely scattered maintaining a certain distance margin from class 2

the shown projection suggests that the proposed method may perform as well or better as margin-based techniques. We touch upon this observation in the following section and examine deeper theoretical connections in Chapter 5.

2.4.4 Application to semantic image retrieval

In addition to the tests mentioned above, we also explored empirically the influence of the DDA transformation on the performance of other classification methods, including NN as a baseline, on the task of semantic image retrieval. For these experiments, three potentially overlapping image sets were selected from the Washington University annotated image collection [94], based on the presence of keywords "trees", "cars" and "ocean" in their annotation. Every classifier was then tested by 10-fold cross-validation. The results of these experiments demonstrate that applying the DDA transformation not only consistenly improves NN classifier accuracy, but also provides a boost in performance to some more advanced non-linear classification methods, such as SVM [33], as shown in Table 2.4.

The latter finding emphasises the importance of the alternative interpretation we gave to the DDA method in section 2.3.2. That is, in addition to the explicitly sought transformation T, the technique may also be seen as providing a discriminative distance metric TT^T that accounts for differences in the scales of different features, removes global correlations and redundancies among features to some extent, and adapts to the fact that some features may be much more informative about the class labels than others. This observation is easily illustrated by the example of SVM classifier with a Gaussian kernel[7] :

$$k_\Sigma(x_i, x_j) = e^{-(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}, \tag{2.37}$$

for some covariance matrix Σ and observations x_i, x_j represented as column vectors. A typical choice of Σ here is an identity matrix multiplied by some constant factor. However, when the DDA technique is applied to preprocess the training data before the SVM learning occurs, the SVM classifier fully takes advantage of the discriminative features extracted by the DDA method since the kernel products can now be seen as evaluated in a new discriminative metric TT^T:

$$k_\Sigma(x_i, x_j) = e^{-(x_i - x_j)^T T T^T (x_i - x_j)}. \tag{2.38}$$

This eventually results in SVM being able to find a simpler solution involving fewer support vectors and better generalization properties, which naturally leads to an improvement in classification performance, as shown in Table 2.4.

From the empirical point of view, in order to verify that non-trivial collection-independent learning has occurred, we also examined the categorization performance of the derived above category-specific DDA transformations on a completely separate image set taken from the COREL database. The empirical evidence demonstrates that the application of the DDA transformation leads to robust categorization of unseen images producing semantically relevant matches that may (Figure 2.11, row one) or may not (Figure 2.11, row two) share the same vocabulary with the query category, as well as allowing images to be assigned to multiple relevant categories (Figure 2.11, the last two images in both rows).

[7] We consider kernel-based methods in more detail in the following chapter.

Table 2.4: Semantic image retrieval results

Classifier	% Error on image data set		
	Trees	Ocean	Cars
Fisher's LDA	43.89	45.56	17.72
SVM (linear)	31.11	21.11	1.58
DDA+SVM (linear)	17.78	11.11	1.40
SVM (gaussian)	23.89	16.67	1.58
DDA+SVM (gaussian)	17.78	11.11	1.40
NN	38.33	19.44	2.46
DDA+NN	18.89	18.33	1.23

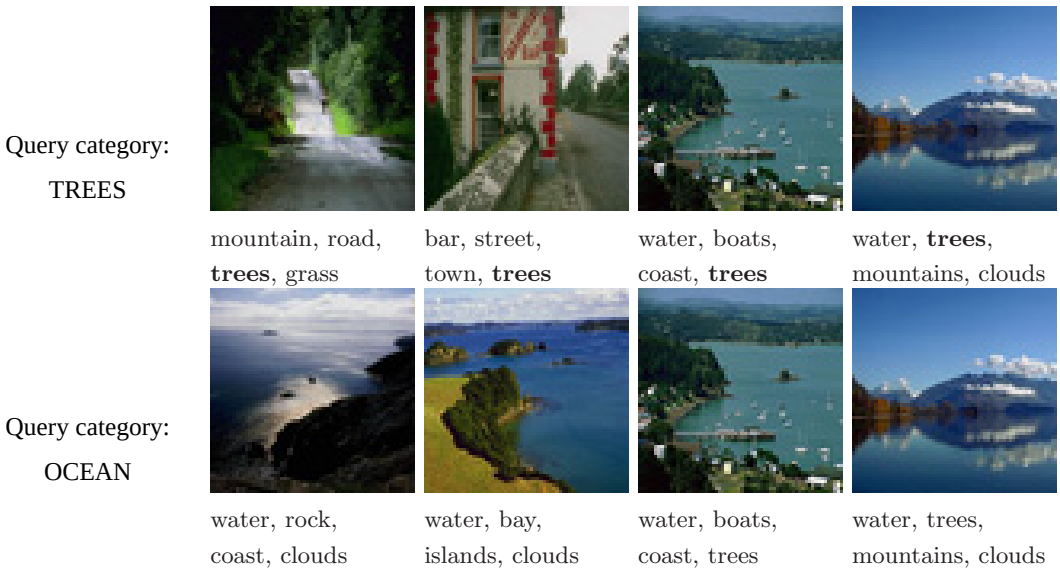


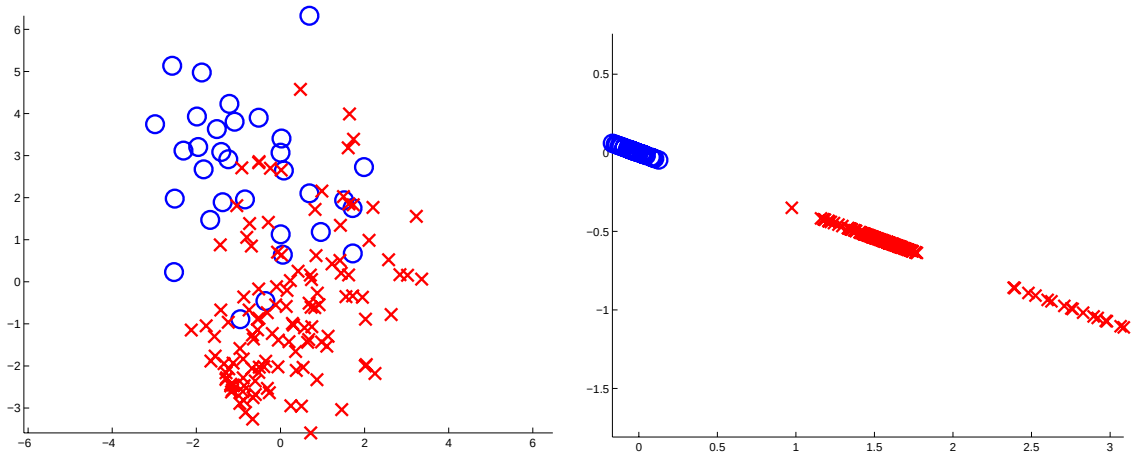
Figure 2.11: Examples of semantic image retrieval. The semantic query specified as a natural language keyword is shown on the left. The true (manually assigned) annotation keywords are listed underneath each image. The annotation keywords overlapping with the query are in bold font.

2.5 Discussion

In this section we briefly review some of the previously developed approaches of discriminant analysis and dimensionality reduction, demonstrating on simple examples the essential differences between existing techniques and the proposed DDA method.

First, we consider principal component analysis (PCA), a fundamental tool for dimensionality reduction that finds a set of orthogonal vectors that account for as much as possible of the data's variance. Apparently, PCA method disregards class membership information altogether and consequently is of limited use as a discriminatory transform. This conjecture

is easily confirmed by comparing 2D projections of the Hepatitis dataset by the PCA and DDA methods illustrated in Figure 2.12, which shows a perfect class separation for the latter approach explaining its 100% classification accuracy reported earlier (see Table 2.2). The singular value decomposition of the resulting transformation reveals that there is only one significantly different from zero singular value, meaning that in order to distinguish between the two classes one may use just one dimension, i.e., project the data set onto a line, as seen in Figure 2.12(b).



(a) PCA projection

(b) DDA projection

Figure 2.12: 2D projections of the Hepatitis dataset

Fisher's linear discriminant analysis (LDA) [42, 43, 48] projects original data into a smaller number of dimensions, while trying to preserve as much discriminatory information as possible by maximizing the ratio of between-class scatter over within-class scatter. Based on the second order statistical information, the method is proven to be optimal whenever data classes are represented by unimodal Gaussians with well-separated means. A violation of this assumption drastically deteriorates LDA's performance, as seen in Figure 2.13 that compares class separation achieved by the projections found by LDA and DDA methods for the classical XOR problem [143]. As for the DDA approach, Figure 2.13 illustrates that the proposed technique does not require data Gaussianity assumption. Furthermore, the method can determine discriminative projection transformations of up to as many dimensions as there are in the data, whereas LDA is limited by rank restrictions on the between-class scatter matrices to have no more than $K - 1$ dimensions, where K is the number of classes.

A biased discriminant analysis (BDA) approach [168, 169] developed with a goal in mind to improve efficiency of interactive multimedia retrieval applications, is based on an appealing idea of asymmetric treatment of positive and negative relevance feedback examples that is brilliantly conveyed by a famous citation: "All happy families are alike, each unhappy family is unhappy in its own fashion" (L. Tolstoy, *Anna Karenina*). According to this metaphor,

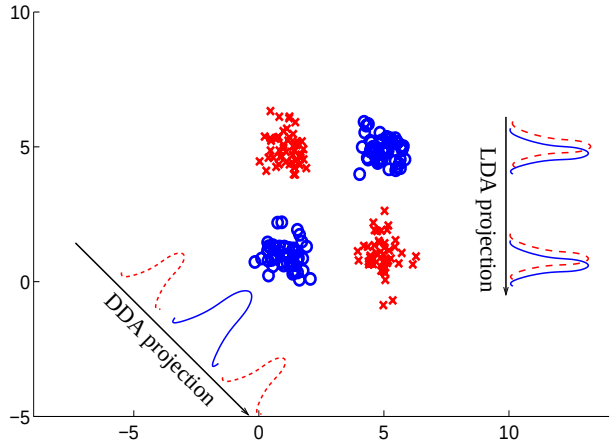


Figure 2.13: XOR problem solution obtained by the LDA and DDA methods

the approach seeks a compact representation of the class of positive examples, while the only constraint placed on negative examples is to stay away as far as possible from the positives. This technique excels in overcoming several important drawbacks of LDA induced by scatter matrix rank restrictions and Gaussianity assumptions and, conceptually, is closest to the two-class version of the proposed DDA method. However BDA's implementation is occasionally offset by suboptimal solutions whenever the observations from the two classes overlap considerably along the direction orthogonal to that of minimal variance of the positive examples. An illustration of this adverse condition is depicted in Figure 2.14).

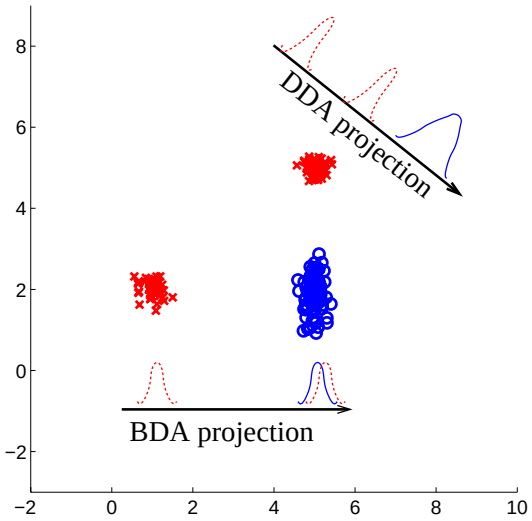


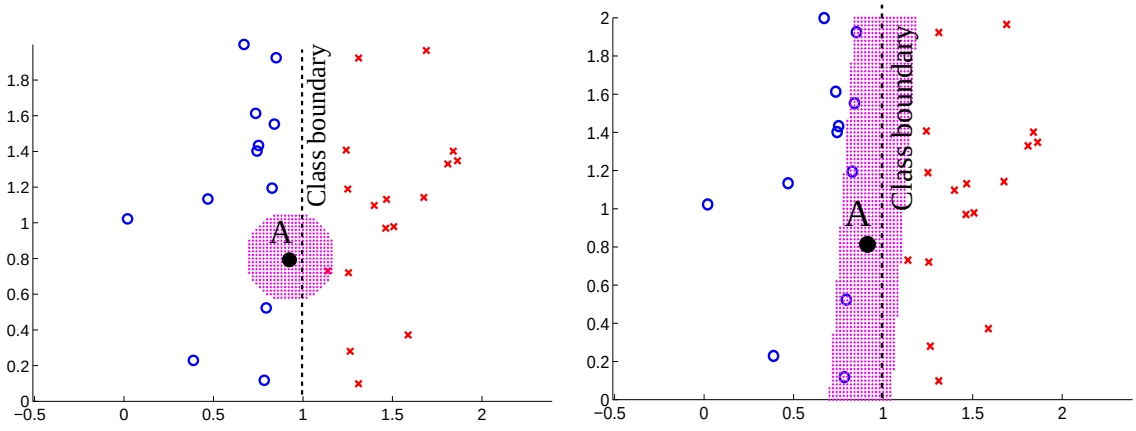
Figure 2.14: Solution of the "dominant variance direction" problem obtained by the BDA and DDA methods

Another advantage of relying exclusively on the distances among the observations lets us relax the sought transformation orthogonality condition often found necessary in other methods. For instance, feature transformation based on maximizing mutual information be-

tween transformed data and their corresponding class labels proposed by Torkkola *et al.* [145] parametrizes the transformation via planar rotations and hence is by design orthogonal, as are those of other methods, which operate on orthogonal subspaces.

There also exist other discriminant analysis methods that are specifically designed to work well for non-Gaussian data sets (e.g., NDA [49]) and target the nearest neighbor classifier performance (e.g., a recent enhancement of NDA proposed in [21]), whose main difference from DDA lies in the fact that these methods still rely on parametric within-class scatter matrices. This is likely to explain why these approaches are generally outperformed by the SVM techniques, while DDA demonstrates comparable results (see Table 2.4).

Among iterative techniques, DANN [62] and CDW [116] methods must be highlighted. Similarly to the proposed DDA, the class-dependent weighted (CDW) dissimilarity approach seeks to optimize a certain criterion for improving NN classification accuracy, which is done by deploying the Dinkelbach's algorithm [41] combined with gradient descent. Effectively, a transformation found by the CDW method may be considered a restricted case of the DDA transformation where no dimensionality reduction is allowed and T is required to be diagonal. As opposed to CDW, the discriminant adaptive nearest neighbor (DANN) approach does permit global dimensionality reduction. It operates according to an iterative scheme to obtain a metric modifying local neighborhoods, which makes it different from the DDA in the way that DANN does not optimize any global criterion or objective function. However, both DDA and DANN in many cases lead to similar results, as demonstrated in Figure 2.15. This illustration shows how DDA transformation corrects the decision of an NN classifier and, conceptually, is an exact reproduction of the motivational example used by the authors in [62] to describe the intuition behind their technique.



(a) NN region of **A** (shaded area) in the original space leads to an error

(b) NN region of **A** after applying DDA produces a correct classification decision

Figure 2.15: Effect of DDA on local neighborhoods - a comparison to DANN [62]

2.6 Summary

We have described a semantic augmentation method formulated in the discriminant analysis framework. The presented method focuses on finding a transformation of the original data that enhances its degree of conformance to the compactness hypothesis and its inverse, which has been shown to lead to a better recognition performance. The classification accuracy has been shown to improve not only with the classifier of choice, NN, but also with more advanced non-linear methods, such as SVM. The latter result underlines the important alternative use of the derived transformation in the capacity of a discriminative metric that accounts for differences in the scales of different features, removes to some extent global correlations and redundancies, and adapts to the fact that some features may be much more informative about the class labels than others.

The presented DDA formulation extends naturally from binary to multiple class discriminant analysis problems. The method can also serve as a discriminating dimensionality reduction technique with the ability to overcome the limitation of the classical parametric approaches that typically extract at most $K-1$ features for a K-class problem, while possessing the means to determine in a data-dependent fashion how many dimensions are sufficient to distinguish among a given set of classes.

We have verified the classification performance of the proposed DDA method and its extensions on a number of the benchmark data sets from UCI Machine Learning Repository [16] and on the real-world semantic image retrieval tasks. The encouraging results demonstrated that the method outperforms several popular methods, and improves classification accuracy, sometimes dramatically, making it an excellent candidate for the intended application of automatic semantic augmentation.

Chapter 3

Kernel distance-based discriminant analysis (KDDA)

The introduction of kernels made solving a variety of complex machine learning problems plausible not only by techniques specifically tailored for such settings, such as neural networks [15, 124], but also by many algorithms, originally designed as linear. In this chapter, we seek to overcome a linearity assumption of the transformation derived by the previously described DDA approach, leading to a formulation of its kernel extension, KDDA. Additionally we focus on two particular aspects of KDDA. The first, that opens up a possibility of using indefinite kernels, stems from a theoretical property of KDDA problem formulation convexity that holds irrespective of the definiteness of the kernel in question. And the second, observed through an empirical evaluation of KDDA as well as several other projective non-linear discriminant analysis methods, results in an adverse condition referred to as the false positive projection effect, often encountered in classification with data set imbalance. The following sections provide a detailed discussion of the topics outlined thus far.

3.1 Brief introduction to kernel methods

Kernel methods have been successfully applied and become prevalent in many areas of pattern recognition and machine learning [1, 22, 28, 35, 54, 56, 123], owing to the solid foundations of the underlying algorithms from the statistical learning theory [153], flexibility and existence of fast and efficient implementations [27, 73, 74]. For any kernel method, an essential component that serves to provide algorithm modularity and represent a problem-specific similarity measure for diverse structured and unstructured types of data is expressed through a *kernel function* [135]. The term *kernel* stems from the first use of this type of function in the area of integral operators studied by Hilbert and others [30, 100]. A function k which gives rise to an operator T_k via

$$T_k f(x) = \int_{\mathcal{X}} k(x, x') f(x') dx' \tag{3.1}$$

is called the kernel of T_k. The following theorem represents a basic functional analysis result that provides valuable insight and helps understand many important properties of kernels and kernel methods.

Mercer's theorem [100]. *Suppose* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is symmetric and satisfies* $\sup_{x,x'} k(x, x') < \infty$*, and define an operator*

$$T_k f(x) = \int_{\mathcal{X}} k(x, x') f(x') dx'; \tag{3.2}$$

suppose $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$ *is positive semi-definite*

$$\int_{\mathcal{X}} \int_{\mathcal{X}} k(x, x') f(x) f(x') dx dx' \geq 0 \tag{3.3}$$

for all $f \in L_2(\mathcal{X})$*. Let* λ_i, ψ_i *be the eigenfunctions and eigenvectors of* T_k*, with*

$$\int_{\mathcal{X}} k(x, x') \psi_i(x') dx' = \lambda_i \psi_i(x). \tag{3.4}$$

Then

1. $\sum_i \lambda_i < \infty$,
2. $\sup_x \psi_i(x) < \infty$,
3. $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \psi_i(x) \psi_i(x')$,

where the convergence is uniform in x, x'*.*

The above result indicates important properties of a kernel function of being continous, symmetric and positive semi-definite, specifies a mapping $\Phi : \mathcal{X} \to \mathcal{F}$

$$x \mapsto \Phi(x) = \left(\sqrt{\lambda_1} \psi_1(x), \sqrt{\lambda_2} \psi_2(x), \ \ldots \right) \tag{3.5}$$

and suggests that $k(x, x')$ corresponds to a inner product in this mapped normed space $\mathcal{F}$, since

$$k(x, x') = (\Phi(x))^T \Phi(x'). \tag{3.6}$$

Table 3.1 lists some of the most popular kernel functions used in practice, while more sophisticated kernels, including those not fully satisfying Mercer's theorem, are discussed at a later time in section 3.3. The identity (3.6) constitutes the basis for the so-called *"kernel trick"*: given an algorithm which is formulated in terms of inner products, one can construct an alternative algorithm by replacing inner products with a kernel function. The benefits of this technique are apparent: an existing algorithm can be extended by applying it to the data mapped into a rich and expressive feature space $\mathcal{F}$ where the solution may be far simpler than in the original input space $\mathcal{X}$, while the actual mapping does not need to be carried out explicitly[1] because of the computational shortcut provided by the kernel. By the same token, the method of distance-based discriminant analysis (DDA) discussed in Chapter 2

[1] or, as is in the case of Gaussian kernel, simply cannot be explicitly computed.

Table 3.1: Kernel functions in common use: Gaussian RBF parametrized either by $\sigma \in \mathbb{R}$ or distance metric Σ^{-1}, Polynomial with $a \in \mathbb{R}$, $d \in \mathbb{N}$, and Sigmoid with $a, b \in \mathbb{R}$. While Gaussian and Polynomial kernels satisfy Mercer's theorem, this is not always the case with the Sigmoid kernel, as discussed further in section 3.3.

$$\begin{aligned} \text{Gaussian RBF} \quad & k(x,x') = \exp\left(\frac{-||x-x'||}{\sigma^2}\right) \\ & k(x,x') = \exp\left(-(x-x')^T\Sigma^{-1}(x-x')\right) \\ \text{Polynomial} \quad & k(x,x') = \left(\langle x,x'\rangle + a\right)^d \\ \text{Sigmoid} \quad & k(x,x') = \tanh(a\langle x,x'\rangle + b) \end{aligned}$$

is amenable to a kernel formulation owing to a simple identity linking distances and inner products:

$$||x-x'|| = \sqrt{x^Tx - 2x^Tx' + x'^Tx'}, \tag{3.7}$$

which is the basis for what literature refers to as the kernel trick for distances [133]. But before we move on to the derivation of the kernel extension of the DDA method, let us briefly state two more results of substantial importance that will be necessary to be relied upon later: the definition of a Reproducing Kernel Hilbert Space and the Representer Theorem.

Definition (Reproducing Kernel Hilbert Space). *Let $\mathcal{X}$ be a non-empty set (the index set) and denote by $\mathcal{H}$ a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with inner product $\langle\cdot,\cdot\rangle$ (and the norm $||f|| = \sqrt{\langle f,f\rangle}$) if there exists a function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ with the following properties.*

1. *k has the reproducing property*

$$\langle f, k(x,\cdot)\rangle = f(x) \ \textit{for all}\ f \in \mathcal{H},\ x \in \mathcal{X};$$

   in particular, $\langle k(x,\cdot), k(x',\cdot)\rangle = k(x,x')$ for all $x, x' \in \mathcal{X}$

2. *k spans $\mathcal{H}$, i.e. $\mathcal{H} = \overline{span\{k(x,\cdot)|x \in \mathcal{X}\}}$ where $\overline{X}$ is the completion of the set X .*

In other words, feature space $\mathcal{F}$ described above may be *completed* to be an RKHS by adding the limit points of all series that are convergent in the norm induced by the inner product[2] $||f|| = \sqrt{\langle\cdot,\cdot\rangle}$ [135], which provides many benefits such as as existence of projections, etc., but most importantly has the advantage that the solutions of optimization in an RKHS under certain conditions may be found as a linear combination of a finite number of basis functions, corresponding to the cardinality N of the training data set, regardless of the dimensionality of the space where the optimization is carried out [31, 76, 112]:

[2]Please note a slight distinction in notation describing dot (inner, scalar) products: $\langle\cdot,\cdot\rangle$ and a^Ta. The former is going to be used in the generic context, e.g. in functional spaces, whereas the latter will be called upon to emphasize the vectorial treatment of the data it is applied to.

Theorem (Representer Theorem). *Denote by $\Omega : [0, \infty) \to \mathbb{R}$ a strictly monotonic increasing function, by $\mathcal{X}, \mathcal{Y}$ sets, and by $l : (\mathcal{X} \times \mathbb{R}^2)^N \to \mathbb{R} \bigcup \{\infty\}$ an arbitrary loss function. Then each minimizer $f \in \mathcal{H}$ of the regularized risk*

$$l((x_1, y_1, f(x_1)), \ \ldots \ , (x_N, y_N, f(x_N))) + \Omega(||f||_{\mathcal{H}})$$

admits a representation of the form

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x),$$

where $\alpha_i \in \mathbb{R}$ for all $1 \leq i \leq N$.

3.2 Kernel reformulation of DDA

Let us suppose that there is a space $\mathcal{F}$ where samples of training data can be mapped via $\Phi : \mathbb{R}^m \to \mathcal{F}$, such that there exists a kernel function $k(x, y) = (\Phi(x))^T \Phi(y)$, where $x, y \in \mathbb{R}^m$ and $k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. We will also assume that the discriminative transformation is sought in $\mathcal{F}$ as a projection matrix ω of size $[\mathscr{N}_{\mathcal{F}} \times d]$, where $\mathscr{N}_{\mathcal{F}}$ is the dimensionality of $\mathcal{F}$, and d is the dimension of the derived discriminative projection subspace, such that the columns of ω lie in the span of all training samples mapped in $\mathcal{F}$, by virtue of the Representer Theorem:

$$\omega = \left[\sum_i^N \alpha_i^{(1)} \Phi(z_i) \ \sum_i^N \alpha_i^{(2)} \Phi(z_i) \ \ \ldots \ \sum_i^N \alpha_i^{(d)} \Phi(z_i) \right], \tag{3.8}$$

where z_i is one of the $N_X + N_Y$ samples from the training data compound matrix Z, as defined in (R.3). The distances between images of samples x and y projected from $\mathcal{F}$ by solution ω are thus expressed as:

$$\begin{aligned} \mathscr{D}_{xy}^2(\omega) &= (\Phi(x) - \Phi(y))^T \omega\omega^T (\Phi(x) - \Phi(y)) \\ &= \mathbf{tr}\left(\omega^T (\Phi(x) - \Phi(y))(\Phi(x) - \Phi(y))^T \omega\right) \\ &= \sum_j^d \left(\sum_i^N \alpha_i^{(j)} (k(z_i, x) - k(z_i, y)) \right)^2 . \end{aligned} \tag{3.9}$$

In matrix notation (R.8) can be simplified as:

$$\mathscr{D}_{xy}^2(\omega) \equiv \mathscr{D}_{xy}^2(P) = \mathbf{tr}\left(P^T H_{xy} P\right) \tag{3.10}$$

where $P \in \mathbb{R}^{N \times d}$ is the sought nonlinear transformation represented as a matrix collecting all of the $\alpha_i^{(j)}$ coefficients, $H_{xy} = (K_x - K_y)(K_x - K_y)^T$, and $K_s = [k(z_1, s), k(z_2, s), \ldots, k(z_N, s)]^T$ denotes a vector of kernel evaluations for sample s over all of the training data.

In view of (R.9), the logarithm of the DDA optimization criterion (2.3) can now be expressed in terms of distances projected from a richer, possibly infinite-dimensional feature

space $\mathcal{F}$:

$$\begin{aligned}\log J(P) &= \frac{2}{N_X(N_X-1)}\sum_{i=1}^{N_X}\sum_{j=i+1}^{N_X}\log\Psi\left(\mathscr{D}_{ij}^W(P)\right)\\ &\quad -\frac{1}{N_XN_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\log\mathscr{D}_{ij}^B(P)\end{aligned} \tag{3.11}$$

The treatment of the obtained criterion differs only slightly compared to the linear case. Similarly to the way it is done in the DDA method, as described in equations (2.9)-(R.2) in Chapter 2, we express convex parts of the criterion by their respective approximations majorized by quadratics [63], while the concave parts are linearized. The former simple algebraic manipulation relies on the Cauchy-Schwarz inequality, while the latter is a direct consequence of the concavity of the log-function, whose combined application leads to the following approximation:

$$\begin{aligned}\mu_{\log J}(P,\bar{P}) &= \frac{1}{N_X(N_X-1)}\mathbf{tr}\left(P^T\mathbb{K}_XB(\bar{P})\mathbb{K}_X^TP\right)\\ &\quad +\frac{1}{2N_XN_Y}\mathbf{tr}\left(P^T\mathbb{K}_{XY}C\mathbb{K}_{XY}^TP\right)\\ &\quad +\frac{2}{N_XN_Y}\mathbf{tr}\left(P^T\mathbb{K}_{XY}G(\bar{P})\mathbb{K}_{XY}^T\bar{P}\right)\\ &\quad + const,\end{aligned} \tag{3.12}$$

where $\bar{P}$ is the current solution, $\mathbb{K}_X$, $\mathbb{K}_{XY}$ are Gram matrices of kernel inner products evaluated over X and all data, respectively, and B, C, G are positive semi-definite design matrices independent of P. Elements b_{ij} of B are defined as:

$$b_{ij}=\begin{cases}-\dfrac{\bar{w}_{ij}}{\Psi\left(\mathscr{D}_{ij}^W(\bar{P})\right)} & \text{if} \quad i\neq j;\\ -\displaystyle\sum_{k=1,k\neq i}^{N_X} b_{ik} & \text{if} \quad i=j;\end{cases} \tag{3.13}$$

where $\bar{w}_{ij}$ is a weight of the Huber function majorizer, that in this case is equal to 1 if $\Psi(\mathscr{D}_{ij}^W(\bar{P}))$ is less than the robustness threshold c, or $c/\Psi(\mathscr{D}_{ij}^W(\bar{P}))$ otherwise. For matrices C and G, their non-zero elements m_{ij} are defined as:

$$m_{ij}=\begin{cases}r_{ij} & \text{for } i\in[1;N_X]\\ & \text{and } j\in[N_X+1;N],\\ r_{ij} & \text{for } i\in[N_X+1;N]\\ & \text{and } j\in[1;N_X],\\ -\displaystyle\sum_{k=1,k\neq i}^{N_X+N_Y} m_{ik} & \text{for } i=j,\end{cases} \tag{3.14}$$

where r_{ij} is equal to -1 and $\frac{-1}{\mathscr{D}_{ij}^B(\bar{P})}$ for C and G, respectively. More derivation details are provided in Appendix B. Finally, taking into account theoretical considerations mentioned

in secion 2.2.3 confirmed by experimental results in section 2.4.3, we define a regularized formulation

$$\begin{aligned}\mu^{reg}_{\log J}(P,\bar{P}) &= \frac{1}{N_X(N_X-1)}\mathbf{tr}\left(P^T\mathbb{K}_X B(\bar{P})\mathbb{K}_X^T P\right)\\ &+\frac{1}{2N_XN_Y}\mathbf{tr}\left(P^T\mathbb{K}_{XY}C\mathbb{K}_{XY}^T P\right)\\ &+\frac{2}{N_XN_Y}\mathbf{tr}\left(P^T\mathbb{K}_{XY}G(\bar{P})\mathbb{K}_{XY}^T\bar{P}\right)\\ &+\lambda\left(\mathbf{tr}(P^T\mathbb{K}_{XY}P)-\Delta\right),\end{aligned} \tag{3.15}$$

where a Lagrange multiplier λ introduces an L_2 norm regularizer expressible as a trace (Representer Theorem).

The approximations used to derive $\mu_{\log J}(P,\bar{P})$ are chosen so as to ensure that the resulting expression's value is never less than the objective to be minimized, and thus provides an upper bound of the criterion (R.10). By optimizing (R.11) iteratively, every subsequent iteration achieves a goal function value that is better or at least as good as the one from the previous iteration, which leads to covergence under the practically reasonable objective boundedness assumption. In Chapter 2, section 2.4.1 this iterative process has been shown to attain more robust as well as better quality local minima, compared to the standard optimization techniques, such as gradient descent, etc.

More formally, such an iterative scheme that constitutes the core of the KDDA, the kernelized extension of the distance-based discriminant analysis method, can be written as the following algorithm:

Algorithm KDDA.

1. Assign an initial starting point $\bar{P}=\bar{P}_0\in\mathbb{R}^{N\times d}$, set convergence tolerance ϵ;
2. Find a successor point P_s : $P_s = \arg\min_P \mu_{\log J}(P,\bar{P})$ subject to a regularization constraint;
3. If $\log J(\bar{P}) - \log J(P_s) < \epsilon$, then stop;
4. Set $\bar{P}=P_s$, go to 2.

3.3 Indefinite kernels via hyperkernels

In contrast to the vast majority of kernel-based techniques for discriminant analysis and classification whose numerical stability, convergence and theoretical performance guarantees depend crucially on the positive semi-definiteness (PSD) of the underlying kernel function,

the KDDA method is free from such a restriction. Indeed, the computationally convenient convexity of the described above approximation (R.11) is due to the PSD property of matrices B and C only, which is true by construction (see Appendices A and B), and hence is not affected even when the so-called *indefinite kernels* [58, 113] are applied. These kernels do not satisfy Mercer's theorem in the strict sense and hence may produce indefinite Gram matrices, presenting some difficulties to the traditional computational methods [58]. Nevertheless, an impressive suite of indefinite kernel methods have been proposed and proven effective in practice by successfully applying jittered [36], tangent distance [60], Kullback-Leibler divergence [108], dynamic time warping [5], distance substitution [59] indefinite kernel functions. In addition to these empirical results, there exist some important theoretical contributions and facts on indefinite kernels as well, such as the recent studies on Reproducing Kernel Krein Spaces (RKKS) [113], the indefiniteness of the sigmoid kernel $k(x, x') = \tanh(ax^T x' + b)$ of neural networks for certain paramter range [95, 152], or convenient convex SVM problem formulations obtained with a broad class of conditionally positive definite kernels [133], the geometric margin interpretation attainable for indefinite kernels producing co-oriented projected and feature space separating hyperplane normal vectors [58], as well as many other results and efforts that motivate further examination of indefinite kernels in the KDDA framework, especially given the fact that KDDA by design is built to tolerate indefinite kernels. In the discussion that follows we consider the application of the hyperkernel method [114] within the KDDA framework with an important modification - the removal of the kernel PSD constraint.

3.3.1 Overview of hyperkernel method

The approach of hyperkernels [114] automatically adjusts kernel parameters in a data-dependent fashion and uses the kernel trick on the space of kernels in order to be able to control the complexity of the learned kernel function via a regularized quality functional Q_{reg}. By analogy with the definition of the regularized risk functional R_{reg} commonly used in the support vector machines [33, 153]:

$$R_{reg} = R_{emp} + \lambda ||f||^2_{\mathcal{H}} \tag{3.16}$$

the regularized quality functional Q_{reg} is a sum of a quality functional Q_{emp} and a regularization term:

$$Q_{reg} = Q_{emp} + \lambda_Q ||k||^2_{\underline{\mathcal{H}}} \tag{3.17}$$

where the former term tells how well matched kernel k is to the given data set, while the latter is the norm of the kernel in Hyper-RKHS $\underline{\mathcal{H}}$ for some positive regularization constant λ_Q. The insight of the hyperkernel approach that specifies $\underline{\mathcal{H}}$ and finds an appropriate kernel in an infinite space of possible solutions much in the same way a suitable hypothesis is found in the RKHS induced by a fixed kernel in the regularized risk minimization problem, is based on an appealing and elegant idea. Namely, the method defines a compound set $\underline{\mathcal{X}} = \mathcal{X} \times \mathcal{X}$ treating kernel k as a function $k : \underline{\mathcal{X}} \to \mathbb{R}$, which allows to extend the definition of an RKHS

for the case of a hyperkernel $\underline{k} : \underline{\mathcal{X}} \times \underline{\mathcal{X}} \to \mathbb{R}$, thus arriving at the concept of Hyper-RKHS, $\underline{\mathcal{H}}$:

Definition (Hyper Reproducing Kernel Hilbert Space). *Let $\mathcal{X}$ be a non-empty set, and denote by $\underline{\mathcal{H}} := \mathcal{H} \times \mathcal{H}$ the compound index set. The Hilbert space $\underline{\mathcal{H}}$ of functions $k : \underline{\mathcal{X}} \to \mathbb{R}$, endowed with inner product $\langle \cdot, \cdot \rangle$ (and the norm $||k|| = \sqrt{\langle k, k \rangle}$) is called a Hyper Reproducing Kernel Hilbert Space if there exists a hyperkernel $\underline{k} : \underline{\mathcal{X}} \times \underline{\mathcal{X}} \to \mathbb{R}$ with the following properties.*

1. *$\underline{k}$ has the reproducing property*

$$\langle k, \underline{k}(\underline{x}, \cdot) \rangle = k(\underline{x}) \ for \ all \ k \in \underline{\mathcal{H}};$$

   in particular, $\langle \underline{k}(\underline{x}, \cdot), \underline{k}(\underline{x}', \cdot) \rangle = \underline{k}(\underline{x}, \underline{x}')$.

2. *$\underline{k}$ spans $\underline{\mathcal{H}}$, i.e. $\underline{\mathcal{H}} = \overline{span\{\underline{k}(\underline{x}, \cdot) | \underline{x} \in \underline{\mathcal{X}}\}}$.*

More importantly, it is shown that the Representer Theorem holds for Hyper-RKHS:

Theorem (Representer Theorem for Hyper-RKHS). *Let $\underline{\mathcal{H}}$ be a Hyper-RKHS, $\mathcal{X}$ a set, Q_{emp} an arbitrary empirical quality functional. Then each minimizer $k \in \underline{\mathcal{H}}$ of the regularized quality functional*

$$Q_{reg} = Q_{emp} + \lambda_Q ||k||^2_{\underline{\mathcal{H}}}$$

admits a representation of the form

$$k(x, x') = \sum_{i,j}^{N} \beta_{ij} \underline{k}((x_i, x_j), (x, x')) \ for \ all \ x, x' \in \mathbb{R} \tag{3.18}$$

where $\beta_{ij} \in \mathbb{R}$ for all $1 \leq i, j \leq N$.

In other words, even though the optimization of Q_{reg} may be carried over a whole space of kernels, it is still possible to find an optimal solution of (3.17) by choosing among a finite number. Further, by applying the method of power series construction it is possible to derive a number of hyperkernels that satisfy all of the conditions stated in the definition of Hyper-RKHS, for instance:

$$\underline{k}(\underline{x}, \underline{x}') = (1 - \lambda_h) \sum_{i=0}^{\infty} \left(\lambda_h k(\underline{x}) k(\underline{x}') \right)^i = \frac{1 - \lambda_h}{1 - \lambda_h k(\underline{x}) k(\underline{x}')}, \tag{3.19}$$

for $0 < \lambda_h < 1$, defines harmonic hyperkernel, which in the case of Gaussian kernel becomes:

$$\underline{k}((x, x'), (x'', x''')) = \frac{1 - \lambda_h}{1 - \lambda_h \exp\left(-\sigma^2 (||x - x'||^2 + ||x'' - x'''||^2)\right)}, \tag{3.20}$$

where σ specifies a default kernel width, which by no means needs to be a close approximation to the one appropriate for the problem in question, and λ_h guides the hyperkernel's preference over various kernel widths, e.g., small λ_h emphasizes wide kernels almost exclusively, while a value close to 1 treats all widths equally. Likewise, a number of other hyperkernels are designed, such as mixed polynomial-Gaussian, translation-invariant, ARD [96, 97], etc.

3.3.2 Indefinite KDDA

Note that for $\beta_{i,j} \in \mathbb{R}$, (3.18) is not necessarily positive semi-definite [99], which is why the original hyperkernel method imposes an additional constraint and ends up solving a semidefinite optimization problem when Q_{emp} is replaced with a standard formulation of regularized risk functional (3.16). However, in the case of KDDA, we are not restricted by this PSD requirement and by virtue of the Representer Theorem for Hyper-RKHS can replace Q_{emp} with (3.15). Furthermore, the co-orientation condition [58] is automatically fulfilled by the regularization term of the KDDA formulation. Thus, the regularized quality functional minimization problem in the KDDA case becomes:

$$Q_{reg}^{KDDA} = \mu_{\log J}(P, \bar{P}, \beta, \bar{\beta}) + \lambda \left(\mathbf{tr}(P^T K(\beta) P) - \Delta\right) + \lambda_Q \beta^T \underline{K} \beta \tag{3.21}$$

where the approximation of the criterion sought to be minimized $\mu_{\log J}(P, \bar{P}, \beta, \bar{\beta})$ now depends on hyperkernel expansion coefficients $\beta_{i,j}$ collected in vector β in addition to P, $\underline{K}$ is a hyperkernel Gram matrix, $K(\beta)$ is a $N \times N$ kernel matrix obtained by reshaping an N^2-element vector $\underline{K}\beta$, and λ and Δ are regularization parameters. Finally, a practical solution scheme is obtained by breaking down (3.21) into a two-stage alternating optimization problem with a projection stage, that solves (3.21) for P while fixing current β, and a hyperkernel stage, that solves (3.21) for β while fixing current P. In summary, the iterative procedure of the KDDA method with indefinite kernels can be stated as follows:

Algorithm Indefinite KDDA.

1. Assign an initial starting point $\bar{P} = \bar{P}_0 \in \mathbb{R}^{N \times d}$, $\bar{\beta} = \bar{\beta}_0 \in \mathbb{R}^{N^2}$, set tolerance ϵ
2. Fix β and solve projection stage:
$$P = \arg\min_P \mu_{\log J}(P, \bar{P}) + \lambda \left(\mathbf{tr}(P^T K(\beta) P) - \Delta\right)$$
3. Fix P and solve hyperkernel stage:
$$\beta = \arg\min_\beta \mu_{\log J}(\beta, \bar{\beta}) + \lambda \left(\mathbf{tr}(P^T K(\beta) P) - \Delta\right) + \lambda_Q \beta^T \underline{K} \beta$$
4. If $\log J(\bar{P}, \bar{\beta}) - \log J(P, \beta) < \epsilon$, then stop
5. Set $\bar{P} = P$, $\bar{\beta} = \beta$ and go to 2

Notably, step 2 of the above algorithm involves the same optimization formulation as the one detailed in the previous section 3.2, provided that the new Gram matrices have been recomputed and fixed, such that $\mathbb{K}_{XY} \equiv K(\beta)$. The problem from step 3 essentially reduces to a large-scale convex quadratic minimization problem with a single linear constraint, instead of the original hyperkernel method's SDP problem solving which, in general, takes longer than

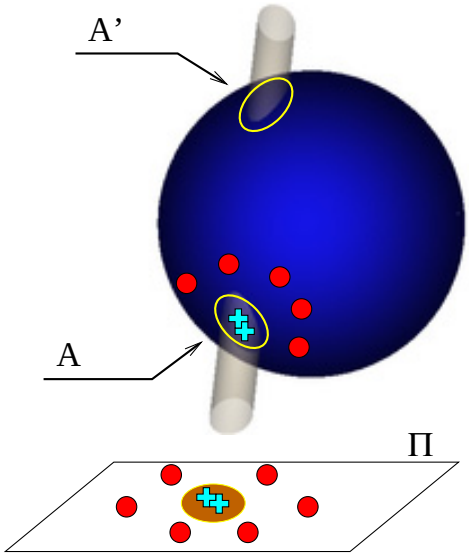


Figure 3.1: Geometric intuition: a sketch of the false positive projection of region A'.

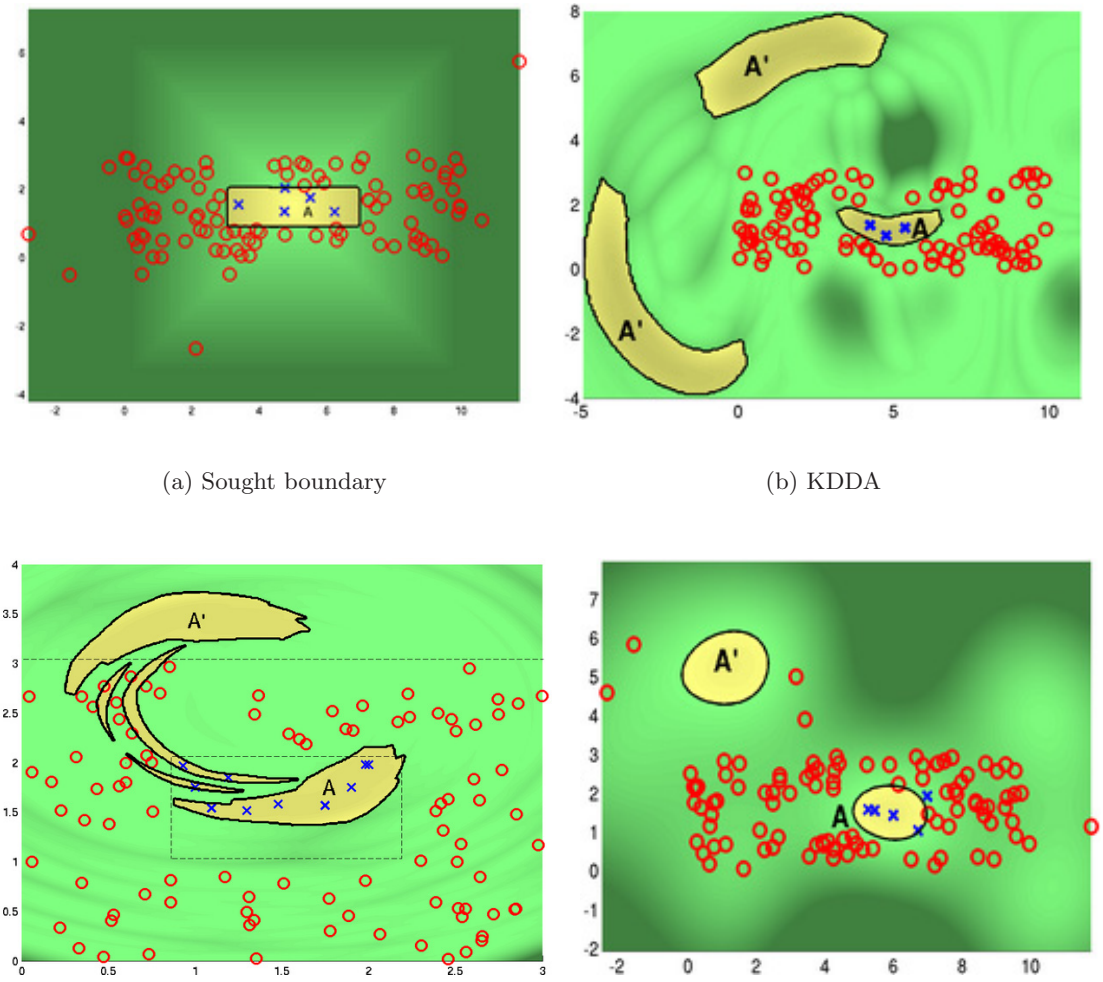
solving a quadratic program [114]. Similarly to the other variants of the algorithm discussed before, the iterative procedure for indefinite KDDA converges because of the boundedness of the objective function and stage-wise improvement at each iteration.

3.4 False positive projection effect

While the previous section concentrated on the KDDA framework extension stemming from a rather theoretical property of its formulation that tolerates the use of indefinite kernels, this section will focus on some other aspect of the KDDA method revealed through empirical evaluation. Namely, we now turn to the discussion of the *false positive projection* effect, a condition that arises mainly in unbalanced data sets when a projective nonlinear classifier utilizing a Gaussian kernel learns a decision function that erroneously associates regions of the input space with a certain, singled out target class. The explicit distinction of such target class, usually referred to as "positive", is due to natural asymmetry of the problem in question, induced by class imbalance typical for 1-against-all classification scenarios, substantial difference in prior probabilities, misclassification costs, and so on [81].

A sketch of this adverse effect is shown in Figure 3.1. Here, an asymmetric projective nonlinear classifier, such as KDDA or BiasMap [168, 169, 167], learns a discriminative projection Π that ensures maximum compactness of the positive class observations, depicted as crosses, relative to the scatter of the negative class observations, depicted as circles. One may notice, however, that the obtained decision region in the projection plane Π corresponds to two distinct parts of the spherical mapping manifold in feature space $\mathscr{F}$: A and A'. In this setup, all test data mapped into A' are classified as positive, assuming such mapping is possible [134], even though the region contains none of the training samples of the positive

class to support such a decision, and likely corresponds to the input space areas where the negative class is far more probable. Thus, a false positive projection of A' takes place. Some examples of the occurrence of this adverse condition with KDDA, BiasMap and KFD [101] classifiers are demonstrated on simple 2D data sets in Figure 3.2. The data samples belong-



(a) Sought boundary

(b) KDDA

(c) BiasMap

(d) Kernel Fisher Discriminant

Figure 3.2: An illustration of an ideal class separation boundary (3.2(a)) and examples of the false positive projection (FPP) effect occurrence in various nonlinear projective methods, (3.2(b)-3.2(d)). Input space regions subject to FPP are denoted as A'. All methods use Gaussian kernels with $\sigma = 2$.

ing to the positive and negative classes are shown as crosses and circles, respectively, while yellow-colored areas highlight the regions of input space classified as positive.

Alternatively, the false positive projection occurrence in the KDDA approach may be thought of as caused by a considerable multiplicity of solutions x^* of $\Phi(x^*)^T\omega = u$, for $u \in U$, where U is a region of projection of positive examples in Π. While this conjecture certainly

merits a separate investigation into preventive modifications such as multiplicity-reducing signed distance inequalities, its universal applicability is yet to be established. Therefore, in the following discussion we will focus only on the method-independent post-processing strategies, i.e. the techniques that do not alter the method in question, but are applied once the learning process has been completed.

3.4.1 Line tracing elimination strategy

In order to summarize the description provided in the previous section and be able to formulate a simple post-processing strategy for elimination of the false positive projection effect, we make an observation analogous to that used in the cluster assignment rule of the support vector clustering method, SVC [12], section 2.2: a data sample is subject to false positive projection if it is classified as positive, but lies across the decision boundary with respect to all of the positive class training samples. This prompts a straightforward strategy based on sampling or "tracing" classification decisions along the simplest possible linear paths between a test sample and positive class training data, leading to the following algorithm for detecting and rejecting the predictions on the test samples erroneously classified as positive.

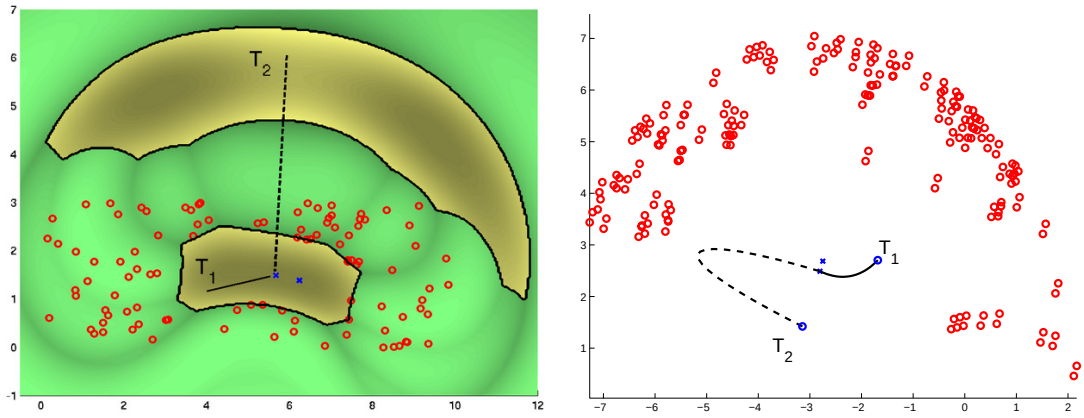
Algorithm FPP elimination by line tracing.

1. Obtain a candidate test sample t classified as positive;
2. Select sets $\mathcal{L}_i = \{\lambda t + (1-\lambda)x_i : \lambda \in \{0, h, 2h, \ldots, 1\}\}, \forall x_i \in X,\ i = 1 \ldots N_X$, and discretization step $0 < h < 1$;
3. If each of $\mathcal{L}_i$ has a sample classified as negative, declare false positive projection and reject positive classification decision on t.

An illustration of the above algorithm applied to the KDDA method is shown in Figure 3.3. Here, two sample straight lines are traced in the input space from candiadate test points T_1 and T_2. While a positive classification decision is retained on T_1, it is rejected on T_2, since on every straight line connecting it to the positive samples of the training data there exist points classified as negative. The latter fact is detected by verifying the classification decisions in the learned nonlinear projection, on points sampled from sets $\mathcal{L}_i$ using a simple uniform sampling technique as adoped in the SVC method, and switching to a Newton-Raphson root-finding routine when necessary.

3.4.2 Filter classifier elimination strategy

The simplicity of the above described elimination strategy comes at a price of having to impose crude linear constraints on the obtained decision boundary, which may negate the benefits of learning a complex nonlinear classifier. For that reason, we also consider a filter classifier



(a) 2D Input space

(b) Nonlinear projection in 2D

Figure 3.3: Applying line-tracing strategy for the false positive projection (FPP) effect elimination: positive classification decision is accepted for T_1, but rejected for T_2

elimination strategy, implemented by introducing a high-recall add-on classifier that limits the input space domain admissible for positive classification. That is, a given test sample must be predicted as positive by both classifier in question and the filter. Practically, the filter is implemented as a multiple-hyperplane classifier [82].

3.5 Experimental results

3.5.1 False positive projection elimination

For the purpose of a preliminary investigation, we conducted a series of experiments on the synthetic nested cuboid data sets, an 2D example of which was earlier shown in Figure 3.2(a). The positive and negative class observations were sampled inside and outside of randomly generated cuboids with the imbalance ratio of 100, and submitted to classification by KDDA, K-BiasMap and KFD using a Gaussian kernel with $\sigma = 2$. Due to substantial class imbalance, and in order to assess the degree of performance degradation due to false positive projection, the classification performance is separately calculated over the positive and negative class instances. The true positive rate a^+, or sensitivity, is the fraction of the positive class samples predicted correctly. Similarly, the true negative rate a^-, or specificity, is the fraction of the negative class samples predicted correctly. The overall performance is thus assessed by evaluating geometric mean accuracy

$$\mathbf{GM} = \sqrt{a^+ \times a^-} \tag{3.22}$$

that takes into account prediction accuracy on both classes [84], and specificity

$$\mathbf{SP} = a^- \tag{3.23}$$

designed to quantitatively measure the effect of false positives on classification performance. The results achieved by the three methods alone (denoted *None*, meaning no FPP elimination strategy is used) as well as their performance enhanced by the FPP elimination techniques described in sections 3.4.1 and 3.4.2 (denoted *Tracing* and *Filter*, respectively) are listed in Table 3.2. The reported figures demonstrate a statistically significant improvement in

Table 3.2: GM accuracy and specificity (in %%) for nested cuboid synthetic data set

Method	None		Tracing		Filter	
	GM	**SP**	**GM**	**SP**	**GM**	**SP**
KDDA	70.0 (±2.4)	96.6 (±1.1)	70.9 (±2.7)	99.6 (±0.1)	75.4 (±2.7)	99.6 (±0.2)
BiasMap	55.3 (±3.0)	99.1 (±0.5)	54.7 (±3.2)	99.5 (±0.2)	55.3 (±3.3)	99.6 (±0.1)
KFD	65.1 (±3.5)	73.3 (±5.9)	76.5 (±2.6)	99.6 (±0.2)	75.3 (±3.2)	99.7 (±0.1)

specificity for KDDA and KFD methods leading to an overall geometric mean accuracy increase, while at the same time pointing out the overly conservative nature of the BiasMap method where the changes are not significant.

For our content-based multimedia retrieval experiments we used ETHZ80 collection [89], whose digital images belonging to several semantic categories were represented by 286-dimensional feature vectors containing 166 global color histogram and 120 Gabor filter texture descriptors extracted by the *Viper* system [141], as was described previously in section 2.4.3. Kernel parameters were determined by cross-validation so as to maximize performance of KFD, and fixed afterwards. The obtained results for each method in terms of averaged geometric mean accuracy and specificity in the "one-against-all" classification scenario are given in Table 3.3. The reported figures generally confirm the hypothesis that false positive

Table 3.3: GM accuracy and specificity (in %%) for ETHZ80 image collection

	KFD			BiasMap			KDDA		
	none	tracing	filter	none	tracing	filter	none	tracing	filter
GM	82.7	82.7	82.7	58.0	59.2	71.4	76.8	77.2	82.2
SP	94.5	94.6	94.7	50.0	54.0	74.4	79.8	80.7	83.6

projection elimination strategies increase specificity leading to a better GM accuracy. These findings also demonstrate that even simple post-processing methods, such as line tracing, may sometimes be sufficient to enhance classification performance, while further benefits may be extracted from more sophisticated techniques, such as the filter method of an add-on high

recall classifier.

3.5.2 Evaluation of Indefinite KDDA

As a basis for comparison with the proposed method of *Indefinite KDDA*, section 3.3.2, we used related discriminant analysis techniques, already mentioned in the previous sections: Kernel Fisher Discriminant (KFD), Kernel Biased Discriminant Analysis (BiasMap), and KDDA with a fixed kernel function. Kernel parameters for these approaches were determined by cross-validation, and fixed throughout. The parameters for the Indefinite KDDA technique were set to $\Delta = 1$ by using a validation data set, while hyperkernel parameters were specified as $\lambda_h = 0.6$ to provide an adequate coverage of various kernel widths by the Gaussian harmonic hyperkernel (3.19) and $\lambda_Q = 1$ according to the recommendations from the authors of the hyperkernel approach [114]. The obtained results for each method in terms

Table 3.4: Object categorization results for the ETHZ80 image database in terms of geometric mean accuracy (in %%).

Object class	KFD	BiasMap	KDDA	Indefinite KDDA
Apple	90.35	61.56	86.02	83.21
Car	76.62	72.27	66.39	82.86
Cow	59.02	53.40	56.51	69.25
Cup	94.69	56.37	87.06	93.49
Dog	76.09	40.09	70.86	78.31
Horse	81.25	39.06	67.00	76.95
Pear	86.76	68.73	86.91	86.39
Tomato	96.66	72.73	93.45	94.05
Average	82.68	58.03	76.78	83.06

of geometric mean accuracy evaluated on the ETHZ80 digital image collection are given in Table 3.4. Here, we see that the indefinite kernel extension of the KDDA technique enhances the baseline KDDA method fine-tuned by cross-validation with a resulting increase of accuracy from 76.78% to 83.06%. In addition to that, one may observe that the proposed approach outperforms, albeit sometimes by a small margin, all other alternative discriminant analysis techniques considered. It also should be noted that in all 8 semantic category classes, the spectra of the Gram matrices at convergence contained both negative and positive eigenvalues, thus confirming the hypothesis on the usefulness of indefinite kernels.

3.6 Summary

In order to overcome the limiting assumption of linearity of the sought disciminative transformation, a kernel-based extension of the above discriminant analysis method, KDDA, is formulated whereby the optimization criterion is expressed in terms of distances projected from a feature space induced by a given kernel function.

Additionally, an application of *indefinite kernels* rendered as unrestricted linear combinations of hyperkernels is considered in the KDDA framework. The proposed formulation entails a solution of a series of quadratic minimization problems, whose computationally advantageous property of being convex is guaranteed regardless of the definiteness of the selected kernel function. This advantageous aspect of KDDA permits it to be used with kernels derived from non-metric distance measures that may better capture the perceptual similarity defining relations among higher level semantic concepts. Finally, an adverse condition referred to as the *false positive projection effect* is studied and some of its elimination strategies are assessed.

Chapter 4

Hierarchical semantic ensembles of classifiers (HSE)

In this chapter we examine automatic semantic augmentation methods where the target classes are described through general natural language as keywords, terms and semantic concepts. This problem setup is usually referred to as the semantic categorization, keyword prediction, autoannotation or automatic linguistic indexing tasks. The diversity in the problem terminology reflects the variety of contributions from numerous research domains that have been proposed to date, as we have briefly discussed in Chapter 1. For instance, an appealing idea of treating the multimedia feature data as another language to translate semantic keywords to and from is developed with the aid of generative probabilistic models by Barnard *et al.* [6, 7]. A family of methods [106, 121, 165], related to the cross-language extension of the latent semantic indexing (LSI) technique [37, 86], permit the retrieval of multimedia semantics via low-level feature queries.

Yet, the majority of the other approaches consider the multimedia autoannotation problem in the multiple-category classification framework, where unseen documents must be assigned to one or more predefined semantic categories. In [52], for instance, the authors focus on improving several popular ensemble schemes, such as OPC (one per class), PWC (pair-wise coupling) and ECOC (error-correcting output codes). The methods developed in [23, 91, 93] decompose a multiple-category classification task into a collection of binary classification problems and propose ways of recombining effectively the individual predictions from classifiers as diverse as SVM, BPM, 2D-MHMM. The semantic categories for these and many other classification-based techniques are generally assumed to be *independent, non-overlapping and sufficient to cover all of the problem domain.*

The approach presented here is also formulated as a classification-based method, but differs from the above work in the important respect that the relationships among the semantic categories derived from the individual keywords of the annotation corpora are explicitly modeled in Bayesian terms, leading to a more consistent autoannotation performance. Furthermore, the proposed method broadens the range of the derived annotation allowing to predict more general notions or semantically-related keyword groups in addition to individual key-

words present in the training data vocabulary. Another benefit of the proposed formulation is that it gives an answer to such an important question as how many keywords the system should predict and whether it is reasonable to predict anything at all.

4.1 Problem formulation

We employ a hierarchical ensemble of binary classifiers in order to perform semantic annotation of unseen images. Given a training set of annotated images $\{I_t, K_t\}_{t=1}^{n}$, where I_t and K_t represent the feature vector of a given image and its associated set of noun keywords, respectively, the concept hierarchy $H = \{C_i\}_{i=1}^{N}$ is a directed graph whose edges are defined by the "hypernym-hyponym" relationships[1] connecting the vertices, or nodes, represented by all of the unique nouns comprising the annotation vocabulary $V = \bigcup_{t=1}^{n} K_t$ and their hyponyms derived from WordNet, a semantic lexicon of English language [102]. For instance, a hypernym WordNet query for a single keyword $k = \{tree\}, k \in V$, returns an ordered set of corresponding hypernyms, as depicted in Figure 4.1(a), that establishes a path graph from the most generic semantic notion sought, an *entity*, to the actual keyword in question via a chain of related hyponym concepts, Figure 4.1(b). These paths are subsequently aggregated over the whole vocabulary V, while making certain that the duplicates links connecting the same vertices, as well as the vertices with only one child are removed. In the resulting hierarchy H, every concept C_i occupies a separate node, and is associated with a binary classifier Θ_i designed to distinguish the set of leaf concepts subsumed (directly or indirectly) by C_i, denoted as $\mathbf{L}(C_i)$, from all of the others. An example of a hierarchy derived for a simple vocabulary V:{*beach, flower, grass, mountain, rock, sky, tree*} is shown in Figure Q.1.

In order to perform the autoannotation of an unseen image represented by a low-level feature vector I_U, each concept C_i is assessed as a potential candidate. Thus, the set of possible annotations is no longer restricted to be V, as is the case for the majority of other similar techniques. The relevance of C_i is seen as a trade-off between, on one hand, how well the input data I_U fits the description of C_i from the classification accuracy point of view, and, on the other hand, how specific or non-ambiguous the candidate set of keywords $\mathbf{L}(C_i)$ is. In our method, the first of these two quantites is represented by the posterior probability of a concept given the data, $P(C_i|I_U)$, while the second one is estimated as the posterior probability of a concept given the assumption that a particular keyword k from the set of all hyponyms of C_i is chosen correctly, denoted as $P(C_i|k)$.

For a given concept C_i, the estimate of $P(C_i|I_U)$ is determined according to the following theorem, which is a reformulation of a previously established result described in [85]:

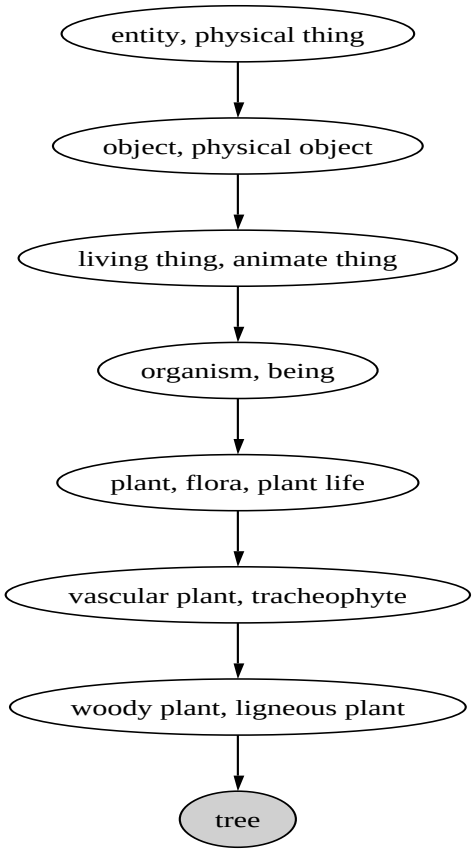
Theorem, Kumar *et al.*, 2002 [85]. *The posterior probability $P(C_i|I_U)$ for any input I_U is the product of the posterior probabilities of all the internal classifiers along a unique path*

[1]Note that A is a hyponym of B, if every A is a (kind of) B. Inversely, A is a hypernym of B, if every B is a (kind of) A.

tree

=> woody plant, ligneous plant
=> vascular plant, tracheophyte
=> plant, flora, plant life
=> organism, being
=> living thing, animate thing
=> object, physical object
=> entity, physical thing

(a) WordNet query output for keyword *tree*



(b) A corresponding path graph

Figure 4.1: Illustration of "hypernym-hyponym" relationship extraction from WordNet

from the root node to C_i, i.e.

$$P(C_i|I_U) = \prod_{l=0}^{\mathscr{D}(C_i)-1} P(C_i^{(l+1)}|I_U, C_i^{(l)}), \tag{4.1}$$

where $\mathscr{D}(C_i)$ is the depth of C_i (the depth of the root concept C_1 is 0), $C_i^{(l)}$ is the concept at depth l on the path from the root node to C_i, such that $C_i^{(\mathscr{D}(C_i))} \equiv C_i$ and $C_i^{(0)} \equiv C_1$.

In order to ensure that (4.1) is applicable in the case of classifiers with non-probabilistic outputs, such as SVM [33], a sigmoid function is fit to the raw classifier output values f_i, as described in detail in Section 4.3. As for $P(C_i|k)$, the Bayes theorem allows to express this quantity in terms of statistics of the training data as shown in (4.2):

$$P(C_i|k) = \frac{P(k|C_i)P(C_i)}{\sum_{C_j \in H} P(k|C_j)P(C_j)}, \tag{4.2}$$

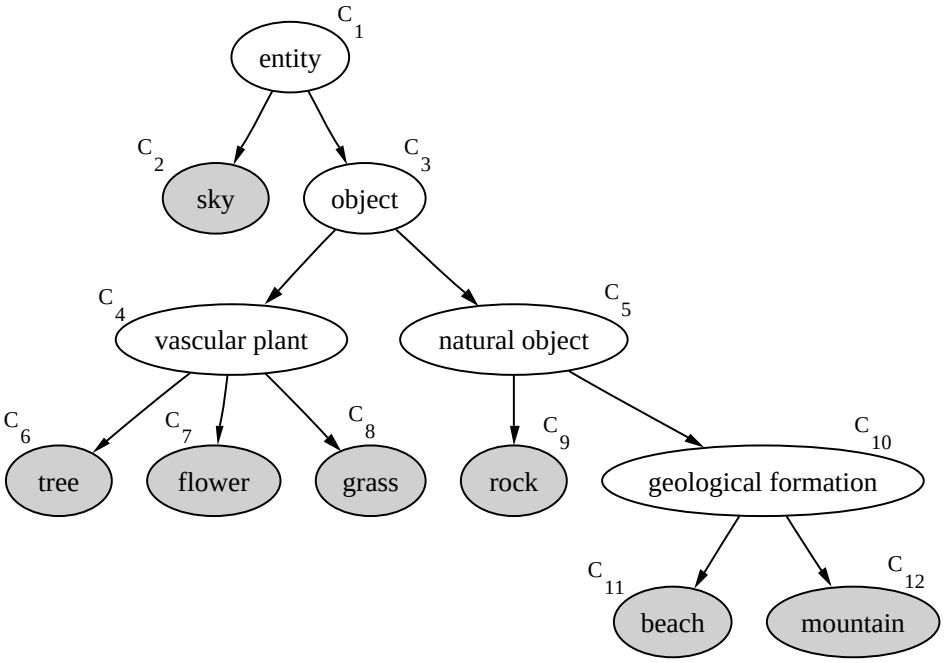


Figure 4.2: Classifier hierarchy example. Shaded nodes denote $C_i \in V$

where $P(C_i)$, a prior probability of concept C_i, is estimated from the training data as:

$$P(C_i) = \frac{\sum_{C \in \mathbf{L}(C_i)} freq^{(T)}(C)}{\sum_{C \in V} freq^{(T)}(C)}, \tag{4.3}$$

and $P(k|C_i)$, the worst-case estimate of the probability of choosing a correct annotation keyword k as an instance of C_i given the degree of generality of concept C_i, is deduced from the hyponym set cardinality information derived from WordNet:

$$P(k|C_i) = \frac{\min_{C \in \mathbf{L}(C_i)} freq^{(W)}(C)}{freq^{(W)}(C_i)}. \tag{4.4}$$

In (4.3) and (4.4), the frequency of a given concept in the training data and the cardinality of the WordNet hyponym set are denoted as $freq^{(T)}$ and $freq^{(W)}$, respectively.

Finally, assuming that the likelihood of the input data I_U given C_i is not dependent on the correctness of a particular choice of k from the hyponym set of C_i, we obtain the following result regarding concept relevance:

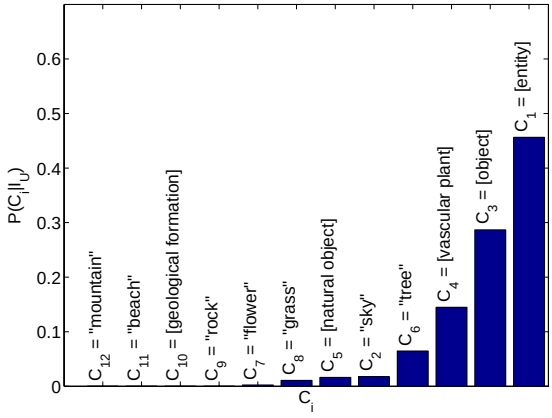
$$\rho \equiv P(C_i|I_U, k) \propto P(C_i|I_U)P(C_i|k), \tag{4.5}$$

which essentially represents a means of comparison of different hypothesis concepts $\{C_i\}$ that takes into account both the goodness of fit of the data I_U to a given concept description and the concept's inherent degree of uncertainty or specificity. The next section illustrates these notions.

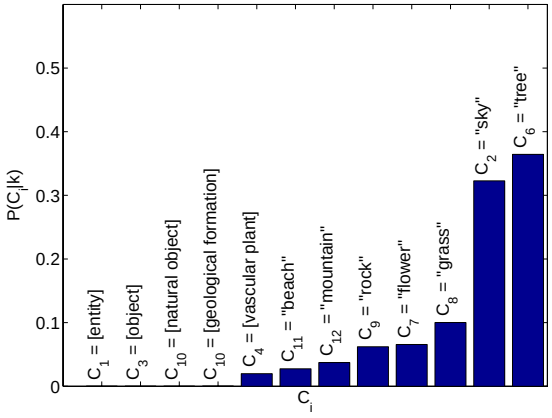
4.2 Illustrative example

Let us come back to the simplified 12-concept classifier hierarchy given in Figure Q.1. To be able to observe the effect of each of the two factors contributing to the final estimate of the concept relevance, we plot separately the computed values of $P(C_i|k)$, Figure 4.3(c), and $P(C_i|I_U)$, Figure 4.3(b)[2] , for a sample test image query depicted in Figure 4.3(a). As the

(a) Query I_U



(b) Goodness of fit $P(C_i|I_U)$



(c) Specificity $P(C_i|k)$

Figure 4.3: Individual contributions of factors $P(C_i|I_U)$ and $P(C_i|k)$

diagrams show, there is a natural tendency among the values of $P(C_i|I_U)$ to favor simpler, more general concepts, such as *object*, due to the smaller number of terms to be evaluated in product (4.1). Quite the opposite trend is noticeable among the estimates of $P(C_i|k)$ that

[2]One may note that $P(C_1|I_U) \neq 1$ as shown in the figure, contrary to what (4.1) may imply. This is explained by our use of the global prior of the root concept *entity* computed from overall WordNet statistics.

tend to promote very specific, unambiguous concepts, such as *tree*, taking into account their prior probabilities as well. This very trade-off of "Goodness of fit vs. Specificity" is captured by the concept relevance, ρ, leading to the results listed in Table 4.1 that demonstrate a

Table 4.1: Candidate concepts ranked by relevance

Rank	$-\log_2 \rho(C_i)$	Concept C_i
1	5.41	C_6 = tree
2	7.46	C_2 = sky
3	8.44	C_4 = vascular plant
4	9.84	C_8 = grass
5	12.64	C_7 = flower
6	**17.26**	$\mathbf{C_1}$ **= entity**
7	17.87	C_3 = object
8	19.42	C_5 = natural object
9	21.00	C_9 = rock
10	44.32	C_{10} = geological formation
11	55.97	C_{12} = mountain
12	56.35	C_{11} = beach

reasonable degree of coherence between the top ranking concepts C_i and the true keywords of the query $K_U = \{$*flowers, path, grass, trees*$\}$.

Another important property of the proposed method that the figures from Table 4.1 help highlight is its ability to determine exactly how many of the top-ranked concepts should be predicted. Many existing approaches [6, 7, 106] resolve this issue by specifying a tunable "refuse-to-predict" parameter that regulates the propensity of image regions to emit concepts or, as some other techniques, by simply considering a fixed number of top-ranked entries. In our case, the relevance of the root node, $\rho_1 = \rho(C_1)$, provides a natural threshold that determines the number of candidate annotation concepts to be selected. An intuitive interpretation of the negative logarithm of this quantity comes from the minimum message length (MML) principle of information theory [157], which interprets $-\log_2 \rho_1$ as the null-model hypothesis test that corresponds to transmitting all the data, since the root concept subsumes all of the other concepts, as is. According to the MML principle, any hypothesis that cannot better the null-model is not acceptable. In our example, this assertion makes us discard all of the candidate concepts ranked 6 or worse (see Table 4.1).

4.3 Probabilistic outputs for baseline classifiers

Having discussed the general formulation of the proposed method, we now turn our attention to the basic building blocks of the ensemble, namely, the individual binary classifiers.

The main criteria for selecting baseline classifiers Θ_i for each candidate concept C_i were superiority in performance and suitability for the task. Thus, we chose two main types of classifiers: support vector machines [33, 153] due to their exceptional performance record and a great degree of flexibility with various types of kernels, and the transformational approach of distance-based discriminant analysis, described in Chapers 2 and 3, as well as [79, 80], that demonstrated very competitive results on the problems specific to the target application domain.

The first of the two techniques, support vector machines (SVM) introduced in greater detail in Appendix C, produces an uncalibrated output defined as:

$$f(\mathbf{x}) = h(\mathbf{x}) + b, \tag{4.6}$$

where

$$h(\mathbf{x}) = \sum_i y_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{4.7}$$

lies in a Reproducing Kernel Hilbert Space (RKHS) $\mathscr{F}$ induced by a kernel k [135, 156]. Training an SVM minimizes regularized risk [33, 153], an estimate of the training misclassification rate plus a penalty term corresponding to the norm of h in the RHKS

$$C \sum_i (1 - y_i f_i)_+ + \frac{1}{2} ||h||_{\mathscr{F}}, \tag{4.8}$$

where $f_i = f(\mathbf{x}_i)$, which also corresponds to the minimization of a bound on the test misclassification rate [153]. The classification decision is made based on the sign of the raw output $f(\mathbf{x})$.

The other method, distance-based discriminant analysis (DDA), finds a data transformation $T \in \mathbb{R}^{m \times n}$ as a solution to the problem of minimization of criterion (R.1). The main advantages of the approach are its non-parametric nature, asymmetric class treatment appropriate for scenarios with a large degree of imbalance among the classes, and the ability to select the dimensionality of the target space automatically. When using DDA, the classification decision is made based on the class label of the nearest neighbor in the T-transformed space. An important fact that becomes evident even from the above succint overview of the SVM and DDA methods is that neither of the two techniques produces a probabilistic output, as required by (4.1). It is therefore necessary to fit posterior probabilities to the raw outputs of the classifiers, which is done by implementing the approach from [117] briefly introduced below.

Considering a posterior probability of a given concept C out of its hierarchical context to simplify the notation, we may obtain the following expression via the Bayes theorem:

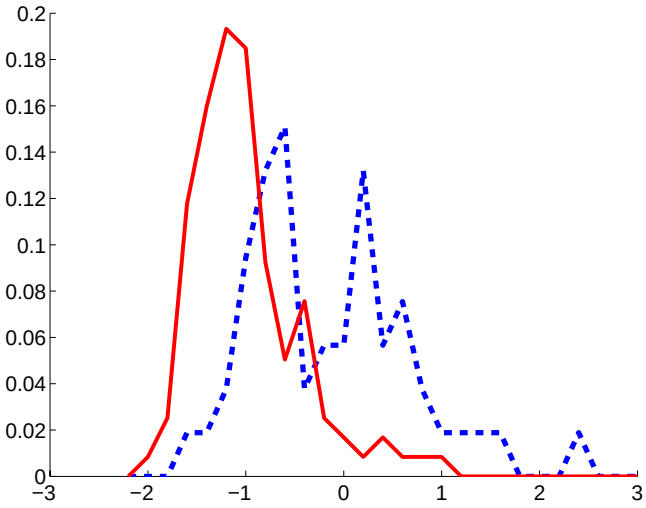
$$P(C|f) = \frac{p(f|C)P(C)}{p(f|C)P(C) + p(f|\overline{C})P(\overline{C})}, \tag{4.9}$$

where $f = f(I_U)$ is the raw output of the classifier given query I_U, which for SVM is defined by (4.6) whereas for DDA it is a signed nearest neighbor distance, $P(C)$ and $P(\overline{C})$ are

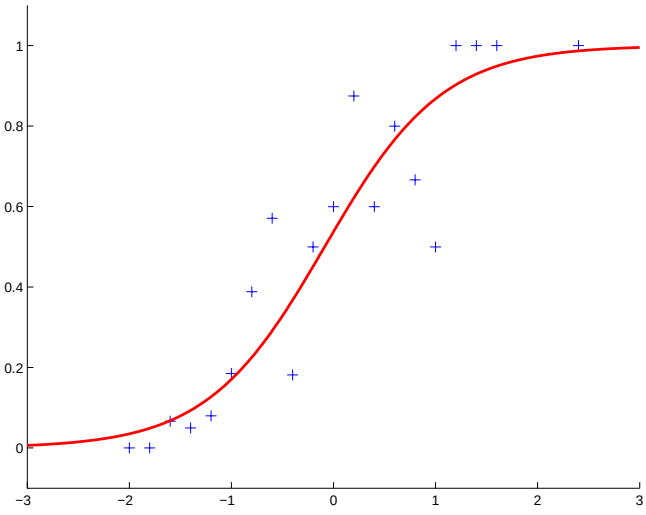
prior probabilities of C and its complement, $p(f|C)$ and $p(f|\overline{C})$ are the corresponding class-conditional densities. Assuming exponential behavior of the class-conditional densities, (4.9) simplifies to a parametrized sigmoid function:

$$P(C|f) = \frac{1}{1 + \exp(Af + B)}. \tag{4.10}$$

The parameters A and B of (4.10) are fit using maximum likelihood estimation from a



(a) Class-conditional densities of raw Gaussian SVM outputs for concept C (dashed) and its complement $\overline{C}$ (solid)



(b) Sigmoid function fit

Figure 4.4: Posterior probability fit for concept "trees"

training data set (f_i, t_i), where t_i is a concept membership indicator such that $t_i = 1$ for C,

and zero otherwise. The problem of finding A and B thus becomes that of minimizing the negative log-likelihood of the training data:

$$\min_{A,B} -\sum_i t_i \log(p_i) + (1 - t_i) \log(1 - p_i), \tag{4.11}$$

where

$$p_i = \frac{1}{1 + \exp(Af_i + B)}. \tag{4.12}$$

In order to solve (4.11) in a robust fashion, the model-trust algorithm [51] suggested by the author is used.

Figure 4.4 shows an example of applying this posterior probability fitting technique to the output values of a Gaussian kernel SVM classifier for the leaf concept *trees*. Figure 4.4(a) plots the histograms of class-conditional densities derived from tenfold cross-validation, while Figure 4.4(b) demonstrates the fit of the sigmoid function (4.10) to the posterior probabilities computed from the class-conditional densities via Bayes' rule. It should be noted that other alternative posterior fitting methods are applicable and may even be more preferrable, given the most recent (at the time of this writing) results on probabilistic interpretation of SVM [56], which opens promising venues for future research.

4.4 Experimental Results

In our experiments we have used data from two separate image collections for training and testing in an attempt to ensure collection-independent learning. The training data was derived from the Washington University annotated digital image collection [94] with about 600 images, while the testing data constituted a 254 image subset, New Zealand and Ireland sections, from Corel image database. The visual information for each training image was represented by 286-dimensional feature vector as described in Chaper 2, section 2.4.2. Annotation keywords appearing only once were eliminated from the target vocabulary V, from which a hierarchical ensemble of semantically related concepts was constructed. The resulting hierarchy is shown in Figure 4.5, where each concept is specified together with hyponym cardinality needed for calculating (4.4).

The preliminary evaluation was to be judged from the point of view of the traditional information retrieval measures of precision and recall [4, 130, 150] expressed in terms of cardinalities of three sets of abstract documents (i.e., annotated digital images in our case):

- R - set of documents relevant to a given query Q,
- A - answer set of documents produced in response to Q by the system being tested,
- Ra - the intersection of sets R and A,

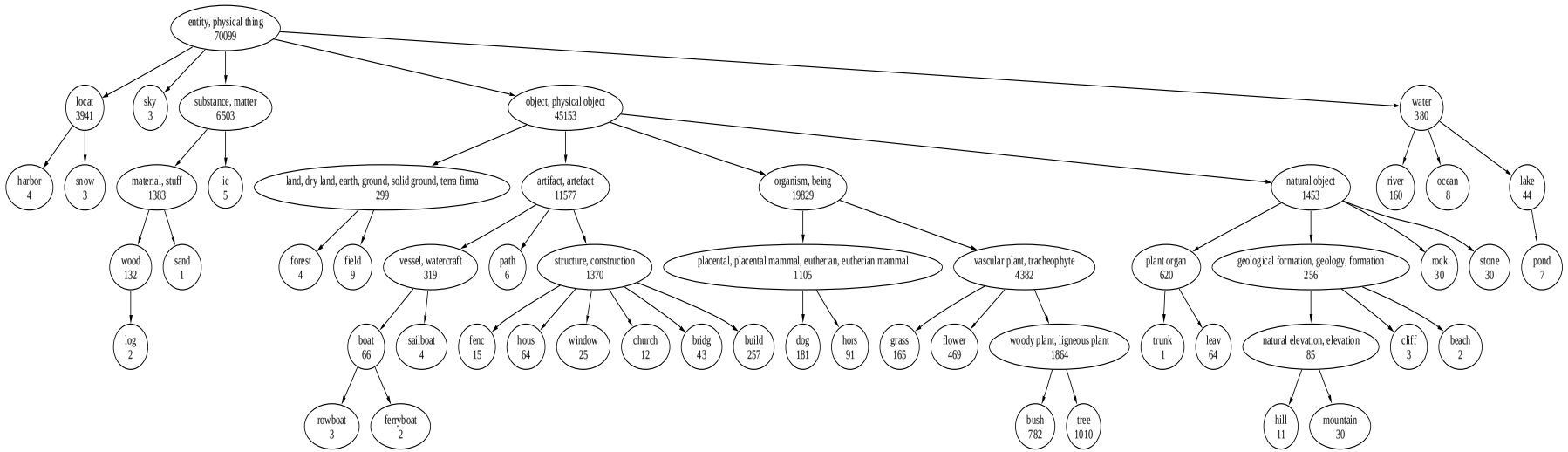


Figure 4.5: The derived semantic hierarchy of classifiers for Washington University collection [94]. Each concept is specified alongside its hyponym cardinality incremented by one to avoid zero frequency problems. For instance, concept *pond* is shown to be associated with cardinality seven because, according to the WordNet lexicon, it has six hyponyms: *fishpond, horsepond, mere, millpond, swimming hole, water hole*.

and defined as

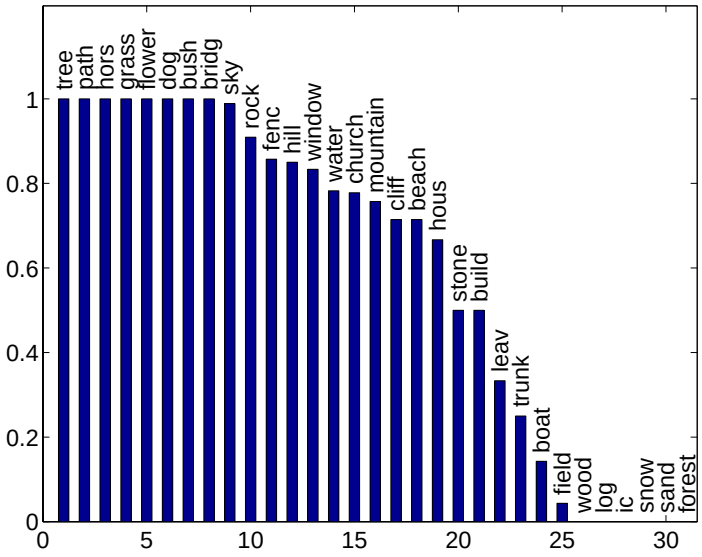
$$Recall = \frac{|Ra|}{|R|}, \tag{4.13}$$

$$Precision = \frac{|Ra|}{|A|}, \tag{4.14}$$

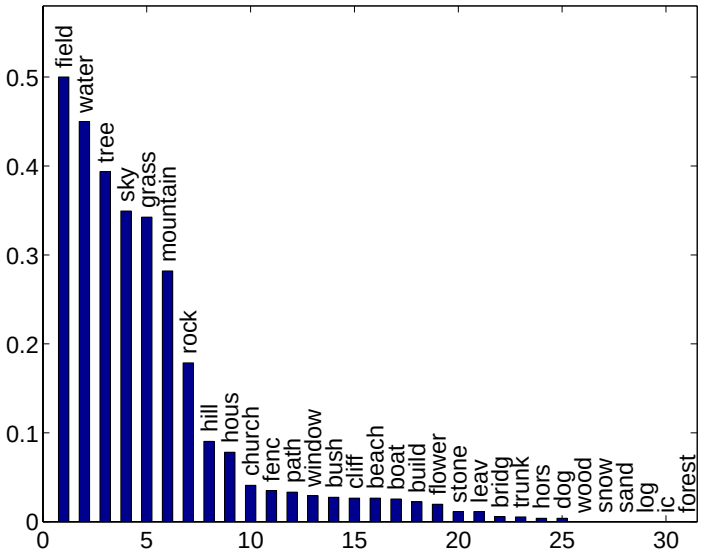
providing an estimates of the fraction of the relevant documents that have been retrieved, and the fraction of the retrieved documents that are relevant. In order to be able to judge the performance of the presented method in terms of the precision and recall indicators, we have adopted the following strategy. Whenever a non-leaf concept, $C_i \notin V$, is predicted, it is evaluated as a union of its underlying keywords, $\mathbf{L}(C_i)$, thus bridging the vocabulary gap between the derived concepts, e.g. *[vessel, watercraft]*, and the actual training data, e.g. *boat, sailboat, ferryboat, rowboat*, at the expense of precision. Using the DDA baseline classifiers [79, 80] for each concept $C_i \in H$, the following precision and recall results on the test set vocabulary were obtained (see Figure 4.6). As seen from the figure, the high recall results boosted by keyword group retrieval (see Figure 4.6(a)) do not necessarily correspond to high frequency common concepts emphasizing the importance of the concept co-occurrence factors, while the significantly lower precision values for complex concepts, such as *church, fence, boat* (see Figure 4.6(b)), indicate that these words are much more often retrieved as a group of semantically-related keywords, rather than individually. The latter observation of lower expected precision is a natural consequence of the above straightforward modification adopted to bridge the gap between the vocabulary of the training data and the one extended through WordNet. Some words shown in Figure 4.6 may appear to have been truncated, e.g. *leav*, *fenc*, *hous*, etc., due to stemming [68, 118] of the vocabulary used throughout all of the experiments.

An illustration of the automatically derived annotation is provided in Figure 4.7, showing examples occurrences of out-of-vocabulary words being replaced by a visually similar common concepts $C_i \in V$ (top-right image, *castle* $\rightarrow$ *rock*), members of the vocabulary being predicted as semantically relevant, but more common (and therefore, more likely) concepts C_i (top-left, *buildings* $\rightarrow$ *construction*), as well as other typical predictions.

In addition to the above experiments, we have compared the presented method to several popular classifier ensemble techniques, such as OPC, or 1-against-all strategy, and Max Wins algorithms [47] that combined SVM baseline classifiers. As shown in Table 4.2, the proposed hierarchical semantic ensemble (HSE) approach achieved better results despite the fact that only a fixed number of 5 top-ranked singleton concepts was allowed to be predicted, which was done in order to ensure equal conditions for all of the methods, most of which have no means of determining exactly the number of concepts in the derived annotation. The first row of Table 4.2 represents the reference point performance attained by sampling concepts according to their empirical distribution in the training data annotation, i.e. picking word *tree* first, since it is most likely to occur, then *sky*, and so on, whereas the last row shows an improvement in performance of the presented HSE method when one considers sibling



(a) Recall



(b) Precision

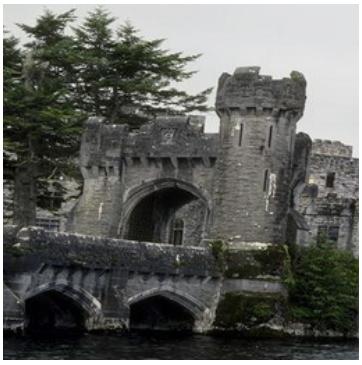
Figure 4.6: Performance indicators on test data vocabulary

concepts[3] the same, e.g. *sailboat* and *boat*.

We also examined the performance of various types of binary SVM techniques as baseline classifiers in the proposed HSE framework, as illustrated in Table 4.3. The results of these studies have confirmed earlier findings [147] stating that state-of-the-art individual classifiers

[3]Concept A is considered a sibling of concept B if $A^{(\mathscr{D}(A)-1)} = B^{(\mathscr{D}(B)-1)}$.

True annotation:
sky, street, buildings, town
Autoannotation:
sky, construction, natural object, artefact



True annotation:
sky, castle, water, tree
Autoannotation:
sky, rock, tree

True annotation:
cows, road, trees, grass
Autoannotation:
bush, tree, grass, vascular plant,
woody plant, organism

True annotation:
sky, water, mountain, trees,
Autoannotation:
sky, water, geological formation,
natural object, artefact

Figure 4.7: Autoannotation of test images

do not necessarily always lead to a better performance in ensembles, while the inadequate results for the Max Wins technique, the only scheme to be using raw classifier outputs, emphasize the importance of the role of fitted posterior probabilities in classification ensembles. However, one needs to be cautious not to invest all faith and admiration into the pointwise probability estimates of non-probabilistic classifiers in view of results from Grandvalet *et al.* [56] and Zhang [164] regarding consistency and asymptotic properties of such estimates.

4.5 Summary

We have presented a hierarchical ensemble learning method applied in the context of multimedia semantic augmentation. In contrast to the standard multiple-category classification setting that assumes independent, non-overlapping and exhaustive set of categories, the proposed approach models explicitly the hierarchical relationships among target classes us-

Table 4.2: Classifier ensemble performance restricted to top 5 keywords

Ensemble	**Baseline classifier**	**% Recall**	**% Precision**
Empirical	none	16.13	5.04
Max Wins	SVM, polyn.	8.14	3.83
Max Wins	SVM, gauss.	10.61	4.47
OPC	SVM, polyn.	20.31	7.85
OPC	SVM, gauss.	21.27	10.19
HSE	DDA	21.22	10.20
HSE+S	DDA	28.42	26.88

Table 4.3: HSE performance with respect to the choice of baseline classifier

Baseline classifier	**% Recall**	**% Precision**
SVM, linear	18.12	5.28
SVM, polynomial	18.34	5.67
SVM, gaussian	18.62	6.05
DDA	21.22	10.20

ing WordNet, in a way, bringing together a statistical classification and linguistic modeling paradigms.

A target class relevance to a query is estimated as a trade-off between the goodness of fit to a given category description and its inherent uncertainty. The latter aspect, formulated in Bayesian terms, brings an additional benefit of allowing to determine exactly the number of categories to be predicted. The promising results of the empirical evaluation confirm the viability of the proposed approach, validated in comparison to several techniques of ensemble learning, as well as with different type of baseline classifiers.

In perspective, we plan to explore further the problem of establishing correspondence between individual annotation keywords and low-level feature descriptors, and improve the proposed approach my taking advantage of the meaningful structure of the resulting hierarchical classification ensemble in order to incorporate relevance feedback from the user, thus extending the approach to the domain of interactive semantic augmentation methods.

Chapter 5

Theoretical issues

In this chapter we examine the motivational analogy between the DDA method and such techniques as support vector machine, SVM [33, 135, 153], and analytic center machines, ACM [146]. We begin considering the similarities among these approaches with a geometric interpretation of the separability constraints that gives rise to the notion of *version space*, where the solution obtained by an SVM corresponds to a Tchebycheff center and the one found by an ACM is an analytic center. In this context, the DDA method formulation is shown to provide an approximation of, or a bound for, the criterion sought to be minimized by an ACM, while having the additional benefits of allowing the uniform treatment of both separable and non-separable data set scenarios, and extending the ACM-like formulation to the case of several separating hyperplanes instead of one. In order to study more rigorously the implications of the latter setting inherent in DDA, we consider an optimal separating hyperplane classifier with an explicit extension to the multiple-hyperplane case, demonstrating the possibilities of extracting further benefits from such formulation, and deriving generalization perfromance guarantees in terms of the associated fat-shattering dimension bound.

5.1 Parallels between SVM, ACM and DDA

5.1.1 Geometric interpretation of version space

Let us consider the traditional formulation of the hard margin support vector machine (SVM) classifier[1] , that finds a hyperplane that separates two classes and maximizes the distance to the nearest data sample from either class. For a training data set comprising a set of N samples $x_i \in \mathbb{R}^m$, $i = 1 \dots N$ and their respective class labels $y_i \in \{-1, 1\}$, the sought hyperplane is determined by solving a convex optimization probem:

$$\max_{\omega} \quad \mathcal{C} \tag{5.1}$$

$$\text{subject to:} \quad y_i \langle x_i, \omega \rangle \geq \mathcal{C}, \text{ for } i = 1 \dots N, \tag{5.2}$$

$$||\omega|| = 1, \tag{5.3}$$

[1] See Appendix C for details on the support vector machine classification method.

where $\mathcal{C}$ is the separation margin, ω is a vector representation of the separating hyperplane in the generic notation without an intercept term. Then, the set of all ω which separate the training data correctly and thus represent all feasible solutions is described as

$$\mathcal{V} = \{\omega | y_i \langle x_i, \omega \rangle \geq 0; i = 1 \dots N, ||\omega|| = 1\}. \tag{5.4}$$

This set $\mathcal{V}$ is referred to as the *version space* [104, 105, 115], and has an insightful geometric interpretation. Indeed, each data sample contributes a separability constraint (5.2) that can be seen as a linear inequality

$$\langle y_i x_i, \omega \rangle \geq 0, \ for\ i = 1 \dots N, \tag{5.5}$$

which corresponds to a half-space determined by a hyperplane with normal vector $y_i x_i$ in the weight space of ω. The intersection of all these half-spaces, some of which, of course, may be redundant, with a sphere that corresponds to the length constraint (5.3) defines $\mathcal{V}$, as illustrated in Figure 5.1. It can be shown that the SVM solution is the Tchebycheff center of

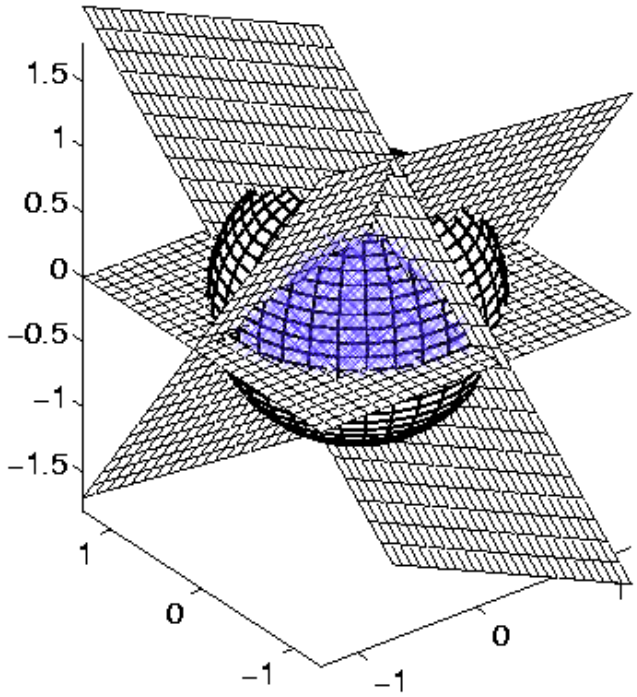


Figure 5.1: Illustration of the version space for $\omega \in \mathbb{R}^3$. Each of the three data samples introduces a half-space associated with a hyperplane defined by its normal vector $y_i x_i$, limiting the set of admissible ω. When combined, these half-spaces form a cone of feasible ω, which after intersection with the sphere corresponding to the length constraint (5.3) renders the version space $\mathcal{V}$, shown as a shaded area

the version space, which geometrically coincides with the center of the largest sphere inscribed

in $\mathcal{V}$ [136]. However, when the version space is elongated or asymmetric, the SVMs are not very effective [146]. This situation is encountered due to the fact that Tchebycheff center may no longer be in proximity of the Bayes point, a theoretically optimal solution which is known to be approximated by the center of mass of the version space [64, 66, 160]. As a result, the SVM solution in this case might provide a poor estimate of the Bayes-optimal decision boundary. While there exist some alternative solutions seeking to approximate the Bayes point directly [65, 128], the underlying problem of computing a center of mass of a polyhedron in a high-dimensional space still presents a formidable challenge. On the other hand, an easily computable center has been proposed by Sonnevend [138, 139] and referred to as *analytic center* of a convex polytope. The said center smoothly depends on data, is invariant with respect to affine transformations and can be computed by minimizing a strictly convex function over the convex polytope. Furthermore, some studies have shown that the Bayes point in the version space may be approximated reasonably well by computing the minimum of a special class of potential functions, that can be naturally adopted in analytic center methods [19]. In the discussion that follows, we examine one such method called an *Analytic Center Machine* (ACM [146]) and establish the links between its formulation and that of the DDA method described in Chapters 2 and 3.

5.1.2 Comparing ACM and DDA formulation

The concept of an analytic center comes from the domain of interior point optimization methods [20]. Let us define the slack $s_i \in \mathbb{R}$ to measure how well solution ω matches the i-th separability constraint (5.5)

$$s_i = y_i \langle x_i, \omega \rangle \geq 0, \text{ for } i = 1 \dots N, \tag{5.6}$$

that becomes negative whenever its corresponding constraint is violated, or remains positive otherwise. A set of solutions ω, compact or otherwise, for which all of the slacks s_i are positive is called a feasible region. An *interior point* is then, as the name implies, any ω located strictly inside of the feasible region. Furthermore, for each constraint we define a *potential function* [111], that goes to infinity as we approach the boundary of the feasible region:

$$\phi_i(s_i) = -\log(s_i), \text{ for } i = 1 \dots N. \tag{5.7}$$

Finally, the *logarithmic barrier* is obtained by combining the potential functions corresponding to all of the separability constraints

$$\mathbf{\Phi}(s) = -\sum_{i=1}^{N} \log(s_i), \text{ for } s = (s_1, \dots, s_N). \tag{5.8}$$

The point at which (5.8) attains its minimum is defined as the *analytic center* of the feasible region. Minimizing the above log barrier subject to a length constraint that ensures compactness of the feasible set constitutes the essence of the ACM method. Needless to say, the

ACM technique assumes that the feasible region is not empty, i.e. the data is separable, in contrast to the SVM approach that may handle non-separable data set scenarios, cf. (4.8).

Now let us recall the formulation of $\log J(T)$ (R.1), the optimization criterion of the DDA. Its second half is expressed as a geometric mean of between-class distances, and in logarithmic form is defined as

$$-\beta S_B(T) = -\frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \log d_{ij}^B(T). \tag{5.9}$$

Also, assume for the moment that T is no longer a transformation matrix, but a column vector, as per notation (5.1)-(5.3)

$$T \equiv \omega \in \mathbb{R}^m. \tag{5.10}$$

Then, the squared between-class distances can be expressed as

$$\left(d_{ij}^B\right)^2 = \langle z_i, \omega\rangle^2 + 2\langle z_i, \omega\rangle\langle -z_j, \omega\rangle + \langle -z_j, \omega\rangle^2, \tag{5.11}$$

where, according to the DDA notation (R.3), z_i and z_j are training data samples belonging to opposite classes. By taking the log of (5.11) and applying the arithmetic-geometric mean inequality

$$\sqrt[n]{a_1 a_2 \dots a_n} \leq \frac{1}{n}(a_1 + a_2 + \dots + a_n), \tag{5.12}$$

for $a_1, a_2, \dots a_n \geq 0$, we obtain

$$-\frac{1}{2}\left(\log\langle z_i, \omega\rangle + \log\langle -z_j, \omega\rangle + \frac{\log 2}{3}\right) \geq -\log d_{ij}^B \tag{5.13}$$

Finally, by summing (5.13) through the respective indices of each class, $i = 1 \dots N_X$, $j = 1 \dots N_Y$, and subsequently dividing by $N_X N_Y$, as required by the DDA formulation (5.9), we arrive at

$$\frac{1}{2}\left(-\frac{1}{N_X}\sum_{i=1}^{N_X}\log\langle z_i, \omega\rangle - \frac{1}{N_Y}\sum_{j=1}^{N_Y}\log\langle -z_j, \omega\rangle - \frac{\log 2}{3}\right) \geq \tag{5.14}$$

$$-\frac{1}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\log d_{ij}^B \equiv -\beta S_B(T), \tag{5.15}$$

where the first two terms inside the brackets of (5.14) can be seen as log barriers of the ACM method, averaged separately over the respective number of data samples in each class[2]. The result (5.14)-(5.15) may give rise to a number of motivating observations. First, the between-class portion of the DDA optimization criterion may be construed as a lower bound for, or approximation of, the weighted log barrier of ACM. Analogous results obtained by DDA and ACM in the ideal case when the training data set is perfectly separable and nearly

[2]Of course, a slight difference in notation needs to be taken into account, according to which the class labels encoded by variables $y_i \in \{-1, 1\}$ in ACM and SVM formulations are stated explicitly in DDA as different signs of the samples belonging to opposite classes, z_i and $-z_j$.

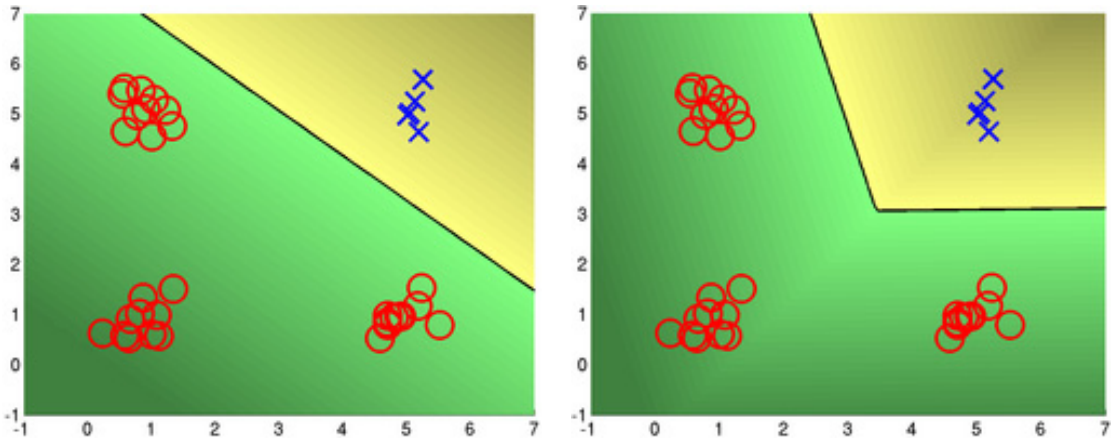
equidistantly projected, $\langle z_i, \omega \rangle \cong \langle -z_j, \omega \rangle$, can be attributed to the fact the objective functions of the two methods become similar[3] under these conditions. Second, the DDA method may be interpreted as a close relative of ACM, because of the adoption of negative logarithm potential functions, but with an important distinction: both separable and non-separable data sets are treated uniformly, since the within-class and between-class distances are non-negative, and thus are always in the admissible domain of the logarithm, save for the case of indefinite kernels, see Chapter 3, section 3.3. The latter distinction makes the DDA appear similar to the SVM, since neither of the two requires class separability, whereas ACM faces a problem with an empty feasible region when data samples are not separable. Finally, the unit column size of the of the sought transformation T (the dimensionality of the target space), as considered in (5.10), is atypical for the DDA. Indeed, the best empirical performance of the developed method has been observed when the target space dimensionality is greater than one. In the above context of SVM and ACM, this corresponds to having several separating hyperplanes used in a classifier, instead of one. This fact distinguishes the developed technique from the other methods considered above, and warrants a separate investigation of the multiple-hyperplane extension within the framework of optimal separating hyperplane classification, which is undertaken in the section that follows.

However, before considering this setting, a comment for the sake of completeness is in order. Namely, the above analysis may have established interesting parallels between the SVM, ACM and DDA methods, but it only considered the between-class distance portion of the DDA formulation, $S_B(T)$, so what about the other, within-class distance portion, $S_W(T)$? The answer to this question might appear to defy common sense logic, because this extra restriction expressed by $S_W(T)$ introduced in the problem formulation actually makes the optimization problem easier. Indeed, the empirical evidence demonstrates that diminishing the relative contribution of $S_W(T)$ to the optimization criterion results in a slower convergence, and eventually leads to severe local minima-related problems when its weight becomes virtually zero. Hence, in addition to ensuring the fulfillment of the compactness assumption among one class of data samples, $S_W(T)$ also plays an important role in the numerical stability of the DDA.

5.2 Multiple-hyperplane classification setting

As already mentioned above, this section is devoted to examining the possible benefits that can be extracted from the multiple-hyperplane extension of the optimal separating hyperplance classification. Apart from the previously discussed DDA motivation, the intuition behind the idea of introducing one or more extra hyperplanes in a classifier is exemplified in Figure 5.2, where it is shown how an additional hyperplane may improve the class separation margin, and thus have the potential to reduce the classification error rate. In the discussion that follows, we formulate the multiple-hyperplane (MH) classification problem, derive gen-

[3]For $n = 3$ and equal a_i, the two sides of inequality (5.12) are only different by a small factor of $\frac{4}{3\sqrt[3]{2}} \approx 1.058$



(a) SVM, $\gamma = 1.14$ (b) MH, $\gamma = 1.55$

Figure 5.2: SVM vs. Multiple-hyperplane (MH) method on a toy problem in 2D: an additional hyperplane leads to a better separation margin γ (both methods use linear kernels).

eralization properties of the corresponding classifier, and presents some experimental results.

5.2.1 Multiple-hyperplane classification problem formulation

The standard 2-class optimal separating hyperplane problem setting (5.1)-(5.3) can be extended trivially in order to accommodate more than one hyperplane[4]:

$$\max_{\omega_1,\ldots,\omega_{N_H}} \quad \mathcal{C} \tag{5.16}$$

$$\text{subject to:} \quad y_i \min_{j=1\ldots N_H} \langle x_i, \omega_j \rangle \geq \mathcal{C}, \text{ for } i = 1 \ldots N, \tag{5.17}$$

$$||\omega_1|| = ||\omega_2|| = \ldots = ||\omega_{N_H}|| = 1, \tag{5.18}$$

where N_H is the number of hyperplanes, each of which is defined by ω_j, for $j = 1 \ldots N_H$. Additionally, we require that the sum of distances to compound border be less or equal to the sum of signed distances to the average hyperplane $\bar{\omega}$:

$$\sum_i y_i \min_{j=1\ldots N_H} \langle x_i, \omega_j \rangle \leqslant \sum_i y_i \langle x_i, \bar{\omega} \rangle, \tag{5.19}$$

where $\bar{\omega} = \frac{1}{N_H} \sum_{j=1\ldots N_H} \omega_j$. This condition ensures some degree of flatness of the compound border avoiding overfitting. There is guaranteed to be at least one set of hyperplanes that meets this requirement. The other, more significant role of this average signed distance constraint, however, will be clarified in the following section. Finally, the decision function

[4]Note that, similarly to the DDA ideas, this formulation preserves asymmetry with respect to the observations belonging to the positive class, $y_i = +1$.

for the multiple-hyperplane classifier is specified as

$$h(x) = sgn\left(\min_{j=1\ldots N_H} \langle x, \omega \rangle\right). \tag{5.20}$$

A disadvantage of the proposed formulation is that the above optimization problem may be quite difficult due to the use of non-differentiable min-function, which necessitates the use of auxiliary numerical strategies for attaining differentiability via smoothing the loss function[5] and avoiding unacceptable local minima via annealed penalty terms. Its advantage, on the other hand, is that (5.16-5.19) are expressed in terms of dot products, and thus are easily extended to nonlinear cases via the kernel trick.

5.2.2 Generalization performance assessment

The following result establishes the generalization properties of the proposed technique.

Proposition, Error bound for MH classifier. *Consider thresholding a class* $\mathbf{F}$ *of functions* $\min_{j=1\ldots N_H} \langle x_i, \omega_j \rangle$ *with unit weight vectors on an inner product space* $\mathcal{X}$ *and fix* $\gamma \in \Re^+$. *For any probability distribution* $\mathcal{D}$ *on* $\mathcal{X} \times \{-1, 1\}$ *with support in a ball of radius* R *around the origin, with probability* $1 - \delta$ *over* l *random examples* S, *any hypothesis* $f \in \mathbf{F}$ *that has margin* $m_S(f) \geq \gamma$ *on* S *has error no more than*

$$\varepsilon(l, \mathbf{F}, \delta, \gamma) = \frac{2}{l}\left(\frac{64R^2}{\gamma^2}\log\frac{el\gamma}{4R}\log\frac{128lR^2}{\gamma^2} + \log\frac{4}{\delta}\right), \tag{5.21}$$

provided $l > 2/\varepsilon$ *and* $64R^2/\gamma^2 < l$.

Note that the above error bound is the same as presented in [33] for a single hyperplane case. In order to clarify the intuition behind this result, we need to emphasize that formulation (5.16-5.19) inherits a lot from that of an optimal separating hyperplane classifier, and thus shares many common qualities. Indeed, one may observe that the proof of a standard result on fat-shattering dimension, $fat_{\mathbf{F}}$, of an optimal hyperplane classifier [135, 10, 151] becomes applicable in the multiple-hyperplane setting, once the average signed distance constraint (5.19) has been taken into account. Consider the following proposition.

Proposition, Fat-shattering dimension bound for MH classifier. *Suppose that* $\mathcal{X}$ *is the ball of radius* R *in* $\mathbb{R}^n$, $\mathcal{X} = \{x \in \mathbb{R}^n : ||x|| \leq R\}$, *and consider the set*

$$\mathbf{F} = \{x \mapsto \min_{j=1\ldots N_H} \langle x, \omega_j \rangle : ||\omega_j|| \leq \frac{1}{\gamma}, x \in \mathcal{X}\}. \tag{5.22}$$

Then $fat_{\mathbf{F}}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2$.

[5] For instance, one possible alternative that worked well in practice was to replace $\min(a, b)$ with its differentiable approximation $\frac{1}{2}(a + b - \sqrt{(a-b)^2 + \delta})$, whose smoothness is controlled through $\delta > 0$.

Proof. The proof proceeds in a manner similar to that of [135]. Assume that $x_1, \ldots, x_r$ are γ-shattered by the multiple-hyperplane classifier. Consequently, for all class labels $y_1, \ldots, y_r \in \{-1, 1\}$, there exists a set $\omega_1, \ldots, \omega_{N_H}$ with $||\omega_j|| \leq \frac{1}{\gamma}$ for $j = 1 \ldots N_H$, such that

$$y_i \min_{j=1\ldots N_H} \langle x_i, \omega_j \rangle \geq 1, \text{ for } i = 1 \ldots r. \tag{5.23}$$

Summing (5.23) over $i = 1 \ldots r$ yields

$$\sum_{i=1}^{r} y_i \min_{j=1\ldots N_H} \langle x_i, \omega_j \rangle \geq r. \tag{5.24}$$

According to the average signed distance constraint:

$$\sum_i y_i \min_{j=1\ldots N_H} \langle x_i, \omega_j \rangle \leqslant \sum_i y_i \langle x_i, \bar{\omega} \rangle, \tag{5.25}$$

where $\bar{\omega} = \frac{1}{N_H} \sum_{j=1\ldots N_H} \omega_j$. Combining (5.24) and (5.25), we obtain:

$$\frac{1}{N_H} \sum_{i=1}^{r} y_i \left(\sum_{j=1}^{N_H} \langle x_i, \omega_j \rangle \right) \geq r, \tag{5.26}$$

which can be written in the form of an inner product as

$$\frac{1}{N_H} \left\langle \Omega, \sum_{i=1}^{r} y_i S^T x_i \right\rangle \geq r, \tag{5.27}$$

where $\Omega = \begin{bmatrix} \omega_1^T & \omega_2^T & \ldots & \omega_{N_H}^T \end{bmatrix}^T$, a column vector containing parameters of all of the hyperplanes, and $S = \mathbf{1}^T \otimes I$ for a vector $\mathbf{1}$ of all ones with length N_H. By the Cauchy-Schwartz inequality, we have

$$\frac{1}{N_H} \left\langle \Omega, \sum_{i=1}^{r} y_i S^T x_i \right\rangle \leq \frac{1}{N_H} ||\Omega|| \left\| \sum_{i=1}^{r} y_i S^T x_i \right\| \leq \frac{1}{\gamma \sqrt{N_H}} \left\| \sum_{i=1}^{r} y_i S^T x_i \right\|, \tag{5.28}$$

Combining (5.27) and (5.28) obtains

$$r\gamma \sqrt{N_H} \leq \left\| \sum_{i=1}^{r} y_i S^T x_i \right\|. \tag{5.29}$$

To bound the right-hand side of (5.29), consider the case when labels y_i are Rademacher variables, i.e., IID with $P(y_i = 1) = P(y_i = -1) = \frac{1}{2}$. From the fact that the expectation $E\{y_i y_k\} = 0$ when $i \neq k$, and that $y_i^2 = 1$, we obtain

$$\begin{aligned} E \left\| \sum_{i=1}^{r} y_i S^T x_i \right\|^2 &= \sum_{i=1}^{r} E \left\langle y_i S^T x, \sum_{k=1}^{r} y_k S^T x_k \right\rangle & (5.30) \\ &= \sum_{i=1}^{r} \left(\sum_{i \neq k} E \langle y_i S^T x_i, y_k S^T x_k \rangle + E \left\langle y_i S^T x_i, y_i S^T x_i \right\rangle \right) & (5.31) \\ &= \sum_{i=1}^{r} E ||y_i S^T x_i||^2. & (5.32) \end{aligned}$$

Since $||y_i S^T x_i|| = ||S^T x_i|| = \sqrt{N_H}||x_i|| \leq \sqrt{N_H}R$, we get

$$E\left\|\sum_{i=1}^{r} y_i S^T x_i\right\|^2 \leq N_H r R^2. \tag{5.33}$$

If the bound (5.33) is true for the expectation when Rademacher variables are used, then there must exist at least one set of labels for which it also holds true, that is

$$\left\|\sum_{i=1}^{r} y_i S^T x_i\right\|^2 \leq N_H r R^2. \tag{5.34}$$

Combining (5.29) and (5.34), we derive the sought bound

$$fat_{\mathbf{F}}(\gamma) \leq r \leq \left(\frac{R}{\gamma}\right)^2. \tag{5.35}$$

□

Bound (5.21) naturally follows, once (5.35) is introduced into the theoretical result that establishes the link between the fat-shattering dimension and generalization error [10, 151], which is listed below.

Corollary 4.14, [33]. *Consider thresholding a real-valued function space $\mathcal{F}$ with range $[-R, R]$ and fix $\gamma \in \mathbb{R}^+$. For any probability distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, 1\}$, with probability $1 - \delta$ over l random examples $\mathcal{S}$, any hypothesis $f \in \mathcal{F}$ that has margin $m_{\mathcal{S}} \geq \gamma$ on $\mathcal{S}$ has error no more than*

$$err_{\mathcal{D}}(f) \leq \varepsilon(l, \mathbf{F}, \delta, \gamma) = \frac{2}{l}\left(d \log \frac{16elR}{d\gamma} \log \frac{128lR^2}{\gamma^2} + \log \frac{4}{\delta}\right), \tag{5.36}$$

provided $l > 2/\varepsilon$, $d < l$, where $d = fat_{\mathbf{F}}(\gamma/8)$.

5.2.3 Empirical evaluation

In an experimental setup similar to those of Chapters 2 and 3, we evaluated the proposed technique in semantic multimedia retrieval experiments on the data from the ETHZ80 digital image collection [89]. The visual information for each image was represented by 286-dimensional feature vector containing 166 global color histogram and 120 Gabor filter texture descriptors extracted by the *Viper* system [141], as was described previously in section 2.4.3.

For each class, we compared the classification accuracy of the 2-class SVM [33, 153] with a Gaussian kernel tuned by cross-validation to that of the MH classifier using the same kernel parameters, but letting the number of hyperplanes vary. The outcome of these experiments demonstrated that in most cases the performance of the SVM classifier is improved by introducing extra separating hyperplanes, while the ratio of the class separation margins achieved by the two methods indicated where such improvement was most likely. The summary of results is shown in Table 5.1.

Table 5.1: Classification accuracy (in %%) per class for ETHZ80 image collection. For the MH classifier, the number in brackets beside the attained accuracy percentage indicates how many hyperplanes, N_H, were used. N_H was selected in the range from one (equivalent to SVM) to ten so as to maximize performance.

Object class	**Accuracy, MH**	**Accuracy, SVM**	**Margin ratio, (MH/SVM)**
(1) apple	97.12 (6)	96.16	1.37
(2) car	88.44 (2)	88.06	1.10
(3) cow	89.75 (5)	84.59	1.11
(4) cup	95.41 (5)	95.94	1.02
(5) dog	92.37 (6)	83.59	1.77
(6) horse	88.44 (4)	88.09	1.76
(7) pear	95.19 (5)	92.16	1.22
(8) tomato	98.38 (2)	97.66	1.01

5.3 Summary

This chapter makes an effort to establish theoretical connections and analogies with some existing machine learning methods. Through this anaylsis, an important detail of the DDA formulation reliance on several separating hyperplanes is highlighted and examined separately. The latter inquiry is performed in the context of margin-based classification. The performance of the proposed technique has been assessed theoretically by establishing a bound on generalization error, and practically by evaluating its performance in a semantic image retrieval task, providing encouraging results. Further research is warranted in order to gain a better insight into the method's theoretical properties via Rademacher complexity bounds [9, 78], and to investigate its performance in related multimedia processing applications.

Chapter 6

Conclusion

6.1 Remarks and summary

In this thesis, we have adopted a rather general view of the domain of multimedia processing applications. This view has allowed us to subdivide a great variety of methods into the two groups of interactive and automatic semantic augmentation techniques, and holistically approach the latter category from the machine learning perspective. Motivated by a number of requirements specific to the target application, we have developed a method of distance-based discriminant analysis, DDA, whose performance has proven competitive in comparison with many state-of-the-art solutions, both general and specifically designed within the automatic semantic augmentation context. Subsequently, the method has been, on one hand, extended to a kernel formulation, and, on the other hand, used as a basic building block in a more sophisticated learning machine termed as the hierarchical semantic ensemble of classifiers, HSE. From the practical point of view of semantic augmentation, the HSE method has demonstrated the best performance results, while from the machine learning perspective, the HSE technique has attested its viability as an alternative method for multiple-category classification with explicit modeling of relationships among the target class categories. Finally, an express effort has been made to establish some theoretical connections and analogies with state-of-the-art machine learning approaches, providing some new results on multpiple-hyperplane classification, MH, and opening the venues for future investigation. The below summary reiterates some important details of each of the mentioned contributions of this work.

- **Distance-based discriminant analysis, DDA**. The presented discriminant analysis method focuses on finding a transformation of the original data that enhances its degree of conformance to the compactness hypothesis and its inverse, which has been shown to lead to better performance. The classification accuracy has been shown to improve not only with the classifier of choice, NN, but also with more advanced non-linear methods, such as SVM. The latter result underlines the important alternative use of the derived transformation in the capacity of a discriminative metric, which makes it possible to combine DDA with other methods. The presented DDA formulation extends naturally

from binary to multiple class discriminant analysis problems, and allows the method to serve as a discriminating dimensionality reduction technique. In the latter case, DDA possesses the means to determine in a data-dependent fashion how many dimensions are sufficient to distinguish among a given set of classes.

- **Kernel distance-based discriminant analysis, KDDA**. In order to overcome the limiting assumption of linearity of the sought disciminative transformation, a kernel-based extension of the above discriminant analysis method, KDDA, is formulated whereby the optimization criterion is expressed in terms of distances projected from a feature space induced by a chosen kernel function. In addition to that, two particular aspects of KDDA method are examined. The first one, that opens up a possibility of using indefinite kernels, stems from a theoretical property of KDDA problem formulation convexity that holds irrespective of the definiteness of the kernel in question. The importance of being able to handle uniformly both positive semi-definite, i.e. conventional, and indefinite kernels from the point of view of semantic augmentation arises from the inherent link between the latter type of kernels and the corresponding non-metric distance measures that can better capture perceptual similarity defining relations among higher level semantic concepts. The second aspect, observed through an empirical evaluation of KDDA as well as several other projective non-linear discriminant analysis methods, lead to a separete study of elimination strategies of an adeverse condition referred to as the false positive projection effect.

- **Hierarchical Semantic Ensemble, HSE**. In contrast to the standard multiple-category classification setting that assumes independent, non-overlapping and exhaustive set of categories, the proposed HSE approach models explicitly the hierarchical relationships among target classes using the WordNet semantic lexicon, bringing together a statistical classification and linguistic modeling paradigms. A target class relevance to a query is estimated in a Bayesian framework as a trade-off between the goodness of fit to a given category description and its inherent uncertainty. An additional advantage of the HSE method is due to its ability to determine exactly the number of semantic categories to be predicted for a given test data, including the possibility of predicting nothing at all.

- **Multiple-Hyperplane Classification, MH**. An effort is made to establish theoretical connections and analogies between the presented DDA method and some state-of-the-art machine learning methods, such as support vector machines and analytic center machines. Through this anaylsis, an important detail of the DDA formulation reliance on several separating hyperplanes is highlighted and examined separately. The latter inquiry is performed in the context of margin-based classification and results in a derivation of a bound on generalization error of the introduced MH classifier.

6.2 Future perspectives

Many of the achieved results of this work have revealed promising venues for future research. In binary discriminant analysis, we have seen the benefits of asymmetry and transformational nature of examined formulation, both of which should be explored further. The apparent advantages of the iterative majorization technique highlighted through practical experience with optimizing the criterion of the developed method have shown potential to provide tremendous help in future development of optimization-based approaches. At the time of this writing, the latter conjecture has already been confirmed by a number of recent publications documenting the increasing popularity of the technique. In the area of kernel-based methods, we have shown a way to uniformly treat both indefinite kernels and traditional positive semi-definite kernels. The former type of kernels are linked to the corresponding non-metric distance measures that can better capture the perceptual similarity, hence the importance of the contribution for the future applications of machine learning algorithms for processing perceptual information. From the point of view of multiple-category classification, we have demonstrated the validity of the proposed mechanism to account for relationships among different traget class categories, which can be applied and easily extended to other domains. Also, we have already started to explore further the HSE context in order to improve the proposed approach by taking advantage of the meaningful structure of the resulting hierarchical classification ensemble in order to incorporate relevance feedback from the user, thus extending the approach to the domain of interactive semantic augmentation methods. The established theoretical connections between the proposed technique of disriminant analysis and some state-of-the-art machine learning methods have underlined unique properties of the proposed formulation and prompted the commencement of the investgation of the multiple-hyperplane setting in the large margin classification context.

The overall positive experience and encouraging performance results described throughout this thesis provide a good reason to believe that further efforts to advance machine learning technologies for semantic augmentation are both justifiable and viable.

Appendices

Appendix A

Matrix derivations for DDA

This section focuses on the intuition behind the definitions of design matrices R and G specified in (2.14) and (2.23). The derivations listed here are mostly based on those developed for the SMACOF multi-dimensional scaling algorithm [18].

Let us consider matrix R that is used in calculation of the majorizing expression of $S_W(T)$ represented by a weighted sum of within-distances. In the derivations that follow, we will assume all weights to be equal to unity, and show afterwards how this assumption can be easily corrected for. We, thus, begin by rewriting a squared within-distance in the vector form:

$$\left(d_{ij}^W(T)\right)^2 = \sum_{a=1}^{m}(x'_{ia} - x'_{ja})^2 = (\mathbf{x}'_i - \mathbf{x}'_j)(\mathbf{x}'_i - \mathbf{x}'_j)^T, \tag{A.1}$$

where $\mathbf{x}'_i$ and $\mathbf{x}'_j$ denote rows i and j from matrix $X' = XT$, representing the corresponding observations transformed by T. Noticing that $\mathbf{x}'_i - \mathbf{x}'_j = (e_i - e_j)^T X'$, (A.1) becomes:

$$\begin{aligned}\left(d_{ij}^W(T)\right)^2 &= (e_i - e_j)^T X' X'^T (e_i - e_j) \\ &= \mathbf{tr}\left(X'^T (e_i - e_j)(e_i - e_j)^T X'\right) \\ &= \mathbf{tr}\left(X'^T A_{ij} X'\right),\end{aligned} \tag{A.2}$$

where A_{ij} is a square symmetric matrix whose elements are all zeros, except for those four indexed by the combinations of i and j that are either 1 (diagonal) or -1 (off-diagonal). For instance, A_{13} for $i = 1$, $j = 3$ and $N_X = 3$ will have the following form:

$$A_{13} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}. \tag{A.3}$$

Taking into account (A.2), the sum of the squared within-distances can be expressed as:

$$\sum_{i<j}^{N_X}\left(d_{ij}^W(T)\right)^2 = \sum_{i<j}^{N_X}\mathbf{tr}\left(X'^T A_{ij} X'\right) = \mathbf{tr}\left(X'^T V X'\right) = \mathbf{tr}\left(T^T X^T V X T\right), \tag{A.4}$$

where $V = \sum_{i<j}^{N_X} A_{ij}$, for which there exists an easy computational shortcut. Namely, V is obtained by placing -1 in all off-diagonal entries of the matrix, while the diagonal elements

are calculated as negated sums of their corresponding off-diagonal values in rows or columns. That is:

$$v_{ij} = \begin{cases} -1, & \text{if} \quad i \neq j; \\ -\sum_{k=1,k\neq i}^{N_X} v_{ik} = N_X - 1, & \text{if} \quad i = j; \end{cases} \tag{A.5}$$

For instance, coming back to our previous $N_X = 3$ example, this technique produces:

$$\begin{aligned} V &= \sum_{i<j}^{N_X=3} A_{ij} \\ &= \left(\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix} \right) \\ &= \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}. \end{aligned} \tag{A.6}$$

It is not difficult to see that the same result applies to the case of non-unitary weights associated with each distance, the only difference being that instead of -1 placed into the off-diagonal elements of V, one should use the negated values of the corresponding weights. And this is exactly how the matrix formulation of $\mu_{S_W}(T, \bar{T})$, (2.15), and design matrix R, (2.14), are obtained:

$$\mu_{S_W}(T, \bar{T}) = \sum_{i<j}^{N_X} \frac{\bar{w}_{ij} \cdot \left(d_{ij}^W(T) \right)^2}{2\Psi \left(d_{ij}^W(\bar{T}) \right)} + K_1 \tag{A.7}$$

$$= \sum_{i<j}^{N_X} \frac{\bar{w}_{ij}}{\Psi \left(d_{ij}^W(\bar{T}) \right)} \left[\frac{1}{2} \mathbf{tr} \left(T^T X^T A_{ij} X T \right) + K_1' \right] \tag{A.8}$$

$$= \frac{1}{2} \mathbf{tr} \left(T^T X^T \sum_{i<j}^{N_X} \frac{\bar{w}_{ij}}{\Psi \left(d_{ij}^W(\bar{T}) \right)} A_{ij} X T \right) + K_1 \tag{A.9}$$

$$= \frac{1}{2} \mathbf{tr} \left(T^T X^T R X T \right) + K_1 \tag{A.10}$$

In order to derive the formulation of matrix G, as specified for the majorizer of $-S_B(T)$ based on Taylor series expansion (2.23), we rewrite (2.22) using the same techniques as we did in (A.2) arriving at:

$$-d_{ij}^B(T) \leq -\frac{\mathbf{tr} \left(T^T Z^T C_{ij} Z \bar{T} \right)}{d_{ij}^B(\bar{T})}, \tag{A.11}$$

where $C_{ij} = (e_i - e_{N_X+j})(e_i - e_{N_X+j})^T$ is a between-class analog of matrix A_{ij}. From (A.6), it is apparent that the same type of a computational shortcut used above to obtain V may be exploited here too. Indeed, matrix $F = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} C_{ij}$ can be quickly constructed by placing -1 in the off-diagonal elements that correspond to index locations of the between-distances,

and subsequently summing with negation to obtain the diagonal entries. An illustration of the technique for $N_X = 2, N_Y = 3$ is shown below:

$$\begin{aligned} F &= \sum_{i=1}^{N_X=2} \sum_{j=1}^{N_Y=3} C_{ij} \\ &= \begin{bmatrix} 3 & 0 & -1 & -1 & -1 \\ 0 & 3 & -1 & -1 & -1 \\ -1 & -1 & 2 & 0 & 0 \\ -1 & -1 & 0 & 2 & 0 \\ -1 & -1 & 0 & 0 & 2 \end{bmatrix}. \end{aligned} \tag{A.12}$$

This is the case of unitary weights. Again, the extension to the non-unitary weight formulation is trivial, and will involve pre-multiplying the off-diagonal entries by the appropriate quantities, which in the case of G are the reciprocals of the squares of the corresponding distances, as shown in (2.23).

In the case of a piece-wise linear approximation of the $-\log(x)$, the majorizing function of $-S_B(T)$ no longer has the same matrix G in both linear and quadratic parts of the expression. It becomes more complex due to separate derivation of a majorizer for each segment, and a summation has to be carried over all of the segments of the piece-wise linear approximation. Namely, the linear, G_L, and quadratic, G_Q, design matrices become:

$$G_L = \sum_{t=1}^{N_s} G_L^{(t)}, \tag{A.13}$$

$$G_Q = \sum_{t=1}^{N_s} G_Q^{(t)}, \tag{A.14}$$

$$G_L^{(t)} = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} C_{ij} \left(\frac{(r_t + l_t)x_t}{2|d_{ij}^B(\bar{T}) - x_t|} + \frac{l_t - r_t}{2} \right) \frac{1}{d_{ij}^B(\bar{T})}, \tag{A.15}$$

$$G_Q^{(t)} = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} C_{ij} \left(\frac{(r_t + l_t)}{4|d_{ij}^B(\bar{T}) - x_t|} \right), \tag{A.16}$$

where N_s is number of segments of the piece-wise linear approximation, and triples (x_t, l_t, r_t) are parameters of a g-family function constituting each such segment. The majorizer of $-S_B(T)$ is thus expressed as

$$\mu_{-S_B}(T, \bar{T}) = \mathbf{tr}\left(T^T Z^T G_Q Z T\right) - \mathbf{tr}\left(T^T Z^T G_L Z \bar{T}\right) + K_3, \tag{A.17}$$

for some constant K_3 independent of T. It should be noted that however more complex, matrices G_Q, G_L, G and R share an important common property with matrix V: they all are positive semi-definite.

Appendix B

Matrix derivations for KDDA

This section exemplifies the derivation of the majorizing expressions for the kernel formulation of the distance-based discriminant analysis (KDDA). We note first the matrix form of expressions for *within-* and *between-*class distances in $\mathcal{F}$:

$$\begin{aligned} \mathscr{D}_{ij}^{W}(\omega) &\equiv \mathscr{D}_{ij}^{W}(P) \\ &= \sqrt{\mathbf{tr}\left(P^T \mathbb{K}_X A_{ij} \mathbb{K}_X^T P\right)} \; for \; i = 1 \ldots N_X, \; j = i+1 \ldots N_X, \qquad \text{(B.1)} \\ \mathscr{D}_{ij}^{B}(\omega) &\equiv \mathscr{D}_{ij}^{W}(P) \\ &= \sqrt{\mathbf{tr}\left(P^T \mathbb{K}_{XY} C_{ij} \mathbb{K}_{XY}^T P\right)} \; for \; i = 1 \ldots N_X, \; j = 1 \ldots N_Y, \qquad \text{(B.2)} \end{aligned}$$

where $\mathbb{K}_X$ and $\mathbb{K}_{XY}$ are Gram matrices of inner products evaluated via kernel function over X and all data, respectively:

$$\mathbb{K}_X = \begin{bmatrix} k(z_1, x_1) & \cdots & k(z_1, x_{N_X}) \\ k(z_2, x_1) & \cdots & k(z_2, x_{N_X}) \\ \vdots & \ddots & \vdots \\ k(z_N, x_1) & \cdots & k(z_N, x_{N_X}) \end{bmatrix}, \qquad \text{(B.3)}$$

and

$$\mathbb{K}_{XY} = \begin{bmatrix} k(z_1, z_1) & \cdots & k(z_1, z_N) \\ k(z_2, z_1) & \cdots & k(z_2, z_N) \\ \vdots & \ddots & \vdots \\ k(z_N, z_1) & \cdots & k(z_N, z_N) \end{bmatrix}. \qquad \text{(B.4)}$$

Then, similarly to (A.7)-(A.10), we can derive the majorizing function of the kernel version[1] of $S_W(P)$ and design matrix B:

$$\mu_{S_W}(P,\bar{P}) = \sum_{i<j}^{N_X} \frac{\bar{w}_{ij}\cdot\left(\mathscr{D}_{ij}^W(P)\right)^2}{2\Psi\left(\mathscr{D}_{ij}^W(\bar{P})\right)} + K_1 \tag{B.5}$$

$$= \sum_{i<j}^{N_X} \frac{\bar{w}_{ij}}{2\Psi\left(\mathscr{D}_{ij}^W(\bar{P})\right)}\left[\mathbf{tr}\left(P^T\mathbb{K}_X^T A_{ij}\mathbb{K}_X P\right) + K_1'\right] \tag{B.6}$$

$$= \frac{1}{2}\mathbf{tr}\left(P^T\mathbb{K}_X^T\sum_{i<j}^{N_X}\frac{\bar{w}_{ij}}{\Psi\left(\mathscr{D}_{ij}^W(\bar{P})\right)}A_{ij}\mathbb{K}_X P\right) + K_1 \tag{B.7}$$

$$= \frac{1}{2}\mathbf{tr}\left(P^T\mathbb{K}_X^T B\mathbb{K}_X P\right) + K_1 \tag{B.8}$$

Identity (B.8) relies on the same computational shortcut we have seen in derivation of (A.6) extended to the case of non-unitary weights. Then, analogously to (A.11), we obtain a majorizing inequality for kernelized between-class distances:

$$-\mathscr{D}_{ij}^B(P) \leq -\frac{\mathbf{tr}\left(P^T\mathbb{K}_{XY}^T C_{ij}\mathbb{K}_{XY}\bar{P}\right)}{\mathscr{D}_{ij}^B(\bar{P})}. \tag{B.9}$$

Then, the final expression of the criterion majorizing function $\mu_{\log J}(P,\bar{P})$, (R.11), is derived with design matrices C and G defined as $\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y} C_{ij}$ and $\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\frac{1}{\mathscr{D}_{ij}^B(\bar{P})}C_{ij}$, and constructed efficiently through the same shortcut without requiring the summation to be carried out.

[1]Note the convention in notation: linear version of optimization criterion and related expressions depend on T, whereas their kernelized versions are expressed in terms of P.

Appendix C

Support vector machine formulation

In this section we provide a detailed account of the 2-class optimal separating hyperplane problem formulation, that constitutes the essence of the support vector machine, SVM, classification method [152]. The unique solution of the said problem is a hyperplane that separates the two classes by maximizing the margin, or the minimum distance from the hyperplane to either class, which provably leads to better classification performance on test data [33, 135, 152]. Consider the following optimization problem

$$\begin{aligned} \max_{\omega} \quad & \mathcal{C} && \text{(C.1)} \\ \text{subject to:} \quad & y_i\langle x_i, \omega\rangle \geq \mathcal{C}, \text{ for } i = 1 \dots N, && \text{(C.2)} \\ & ||\omega|| = 1, && \text{(C.3)} \end{aligned}$$

for a training data set comprising a set of N samples $x_i \in \mathbb{R}^m$, $i = 1 \dots N$ and their respective class labels $y_i \in \{-1, 1\}$, and class separation margin $\mathcal{C}$. Constraints (C.2) ensure that all data samples are at least a signed distance $\mathcal{C}$ from the decision boundary defined by ω, whereas (C.3) helps avoid equivalent solutions that are a positive multiple of each other. The formulation is simplified when (C.3) is introduced into (C.2) by dividing all inequalities by the norm of ω and changing variables so as to make $\mathcal{C} = 1/||\beta||$, which leads to an equivalent problem:

$$\begin{aligned} \min_{\omega} \quad & \frac{1}{2}||\omega||^2 && \text{(C.4)} \\ \text{subject to:} \quad & y_i\langle x_i, \omega\rangle \geq 1, \text{ for } i = 1 \dots N. && \text{(C.5)} \end{aligned}$$

This is an optimization problem with a convex objective function and linear inequality constraints, whose Langrange function is

$$\mathcal{L}_P = \frac{1}{2}||\omega||^2 - \sum_{i=1}^{N} \alpha_i \left(y_i\langle x_i, \omega\rangle - 1\right). \quad \text{(C.6)}$$

Owing to the convexity of formulation (C.4-C.5), the sought hyperplane is obtained by solving the Wolfe dual of (C.6)

$$\mathcal{L}_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle, \text{ for } \alpha_i > 0, \tag{C.7}$$

obtained by setting to zero the derivative of (C.6) with respect to ω and substituting the result back into (C.6). The strong duality also implies that the solution must satisfy complementary slackness of the Karush-Kuhn-Tucker conditions

$$\alpha_i \left(y_i \langle x_i, \omega \rangle - 1 \right) = 0, \text{ for } i = 1 \ldots N, \tag{C.8}$$

from which the sparseness of the SVM solution follows. Indeed, in the dual representation $\omega = \sum_i \alpha_i y_i x_i$ depends only on a limited number of data samples x_i lying precisely on the boundary of the separating margin, i.e. $y_i \langle x_i, \omega \rangle = 1$, and whose corresponding $\alpha_i > 0$. The rest of the data set has $y_i \langle x_i, \omega \rangle > 1$ and $\alpha_i = 0$, thus having no influence on the obtained solution. Data samples with non-zero α_i are called *support vectors*, and hence the name of the method - support vector machine. The classification of test data is performed by computing the SVM decision function

$$h(x) = sgn \langle x, \omega \rangle. \tag{C.9}$$

In practical scenarios, however, it may not be possible to separate the two classes perfectly due to such issues as noise in data samples, inadequacy of representation, difficulty of the problem, etc. In order to account for the overlap among samples, a more general SVM formulation maximizes the separation margin, but allows some observations to be on the wrong side of the margin. This is done by modifying the separability constraints (C.2):

$$y_i \langle x_i, \omega \rangle \geq \mathcal{C}(1 - \xi_i), \tag{C.10}$$

for $i = 1 \ldots N$, $\xi_i \geq 0$ and $\sum_i \xi_i \leq$ constant. The value of slack variable ξ_i denotes the proportional amount by which the corresponding sample is off to the wrong side of its margin. Since a misclassification occurs whenever $\xi_i > 1$, the bound on $\sum_i \xi_i$ is a bound on the total number of permitted misclassifications on the training data set. Through a change in variables similar to that of (C.4)-(C.5), we derive

$$\begin{aligned} \min_{\omega} \quad & \frac{1}{2} ||\omega||^2 && \text{(C.11)} \\ \text{subject to:} \quad & y_i \langle x_i, \omega \rangle \geq 1 - \xi_i, && \text{(C.12)} \\ & \xi_i \geq 0, && \text{(C.13)} \\ & \sum_i \xi_i \leq \text{ constant}, && \text{(C.14)} \end{aligned}$$

for $i = 1 \dots N$. Setting $\mathcal{K}$ to be the slack variable bound, an equivalent formulation of (C.11)-(C.14) is obtained

$$\min_{\omega} \quad \frac{1}{2}||\omega||^2 + \mathcal{K}\sum_{i=1}^{N}\xi_i \tag{C.15}$$

$$\text{subject to:} \quad y_i\langle x_i, \omega\rangle \geq 1 - \xi_i, \tag{C.16}$$

$$\xi_i \geq 0, \tag{C.17}$$

whose Langrange function is

$$\mathcal{L}_P = \frac{1}{2}||\omega||^2 + \mathcal{K}\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left(y_i\langle x_i, \omega\rangle - (1-\xi_i)\right) - \sum_{i=1}^{N}\mu_i\xi_i, \tag{C.18}$$

for $i = 1 \dots N$, and $\alpha_i, \mu_i, \xi_i \geq 0$. Then, solution ω for the non-separable case is uniquely characterized by the Wolfe dual obtained analogously to (C.7), together with the Karush-Kuhn-Tucker conditions

$$\alpha_i\left[y_i\langle x_i, \omega\rangle - (1-\xi_i)\right] = 0, \tag{C.19}$$

$$\mu_i\xi_i = 0, \tag{C.20}$$

$$y_i\langle x_i, \omega\rangle - (1-\xi_i) \geq 0. \tag{C.21}$$

Furthermore, (C.19) and (C.21) imply that at the solution, the slack variables ξ_i are given by

$$\xi_i = \left[1 - y_i\langle x_i, \omega\rangle\right]_+, \tag{C.22}$$

where subscript "+" indicates positive part. Substituting (C.22) into (C.15) leads to a regularized risk formulation of the SVM[1], well studied in the statistical learning theory [153]

$$\min_{\omega}\sum_{i=1}^{N}\left[1 - y_i f(x_i)\right]_+ + \lambda||\omega||^2, \tag{C.23}$$

where

$$f(x) = \langle x_i, \omega\rangle. \tag{C.24}$$

Finally, the dual representation of the problem solution implies that inner product formulation of (C.24) is easily extended to non-linear cases via kernel trick, as discussed in Chapter 3.

[1] cf. (3.16) and (4.8)

Appendix D

Publications

The work described in this thesis has led to the publication of the following materials.

- KOSINOV, S. Visual object recognition using distance-based discriminant analysis. Tech. Rep. 03.07, Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, Rue Général Dufour, 24, CH-1211, Geneva, Switzerland, 2003.
- KOSINOV, S., AND MARCHAND-MAILLET, S. Overview of approaches to semantic augmentation of multimedia databases for efficient access and content retrieval. In Proceedings of the 1st International Workshop on Adaptive Multimedia Retrieval (AMR 2003)/ (Hamburg, 2003).
- KOSINOV, S., AND MARCHAND-MAILLET, S. Evaluation of distance-based discriminant analysis and its kernelized extension in visual object recognition. In Proceedings of the 7th International on signal/image processing and pattern recognition (UkrObraz 2004)/ (Kijiv, Ukraine, 2004).
- KOSINOV, S., AND MARCHAND-MAILLET, S. Hierarchical ensemble learning for multimedia categorization and autoannotation. In Proceedings of the 2004 IEEE Signal Processing Society Workshop (MLSP 2004)/ (Sao Luis, Brazil, 2004), pp. 645-654.
- KOSINOV, S., AND MARCHAND-MAILLET, S. Multimedia autoannotation via hierarchical semantic ensembles. In Proceedings of the Int. Workshop on Learning for Adaptable Visual Systems (LAVS 2004)/ (Cambridge, UK, 2004).
- KOSINOV, S., MARCHAND-MAILLET, S., AND PUN, T. Iterative majorization approach to the distance-based discriminant analysis. In Proceedings of the 28th Annual Conference of the GfKl 2004/ (Dortmund, Germany, March 9-11 2004).
- KOSINOV, S., MARCHAND-MAILLET, S., AND PUN, T. Visual object categorization using distance-based discriminant analysis. In Proceedings of the 4th International Workshop on Multimedia Data and Document Engineering/ (Washington, DC, July 2004).

- KOSINOV, S., MARCHAND-MAILLET, S., AND PUN, T. Countering the false positive projection effect in nonlinear asymmetric classification. In The IEEE Symposium on Signal Processing and Information Technology (ISSPIT'05)/ (Athens, Greece, December, 18-21 2005).

- KOSINOV, S., TITOV, I., AND MARCHAND-MAILLET, S. Large margin multiple hyperplane classification for content-based multimedia retrieval. In Machine Learning Techniques for Processing Multimedia Content, ICML Workshop/ (Bonn, Germany, August, 11 2005).

- KOSINOV, S., AND MARCHAND-MAILLET, S. Visual object categorization with indefinite kernels in discriminant analysis framework. In Proceedings of SPIE Photonics West, Electronic Imaging 2006, Multimedia Content Analysis, Management, and Retrieval 2006 (EI122)/ (San Jose, USA, January, 15-19 2006).

Bibliography

[1] Ali, A. S., and Abraham, A. An empirical comparison of kernel selection for support vector machines. In *Soft Computing Systems - Design, Management and Applications* (2002), A. Abraham, J. Ruiz-del-Solar, and M. Köppen, Eds., Frontiers in Artificial Intelligence and Applications Vol. 87, IOS Press Amsterdam, Berlin, Oxford, Tokyo, Washington D.C., pp. 321–330.

[2] Andrews, S., Tsochantaridis, I., and Hofmann, T. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15*, S. T. S. Becker and K. Obermayer, Eds. MIT Press, Cambridge, MA, 2003, pp. 561–568.

[3] Arkadev, A., and Braverman, E. *Computers and Pattern Recognition*. Thompson, Washington, D.C., 1966.

[4] Baeza-Yates, R., and Ribeiro-Neto, B. *Modern Information Retrieval*. Addison Wesley, 1999.

[5] Bahlmann, C., Haasdonk, B., and Burkhardt, H. On-line handwriting recognition with support vector machines - a kernel approach. In *Eighth International Workshop on Frontiers in Handwriting Recognition* (Ontario, Canada, August 2002).

[6] Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., and Jordan, M. Matching words and pictures. *Journal of Machine Learning Research 3* (2003), 1107–1135.

[7] Barnard, K., Duygulu, P., and Forsyth, D. Recognition as translating images into text. *Internet Imaging IX, Electronic Imaging 2003 (Invited paper)* (2003).

[8] Bartlett, P. For valid generalization, the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems 9* (1997), pp. 134–140.

[9] Bartlett, P. L., and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory and Fifth European Conference on Computational Learning Theory* (2001), pp. 224–240.

[10] Bartlett, P. L., and Shawe-Taylor, J. Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods – Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. MIT Press, 1999, pp. 43–54.

[11] Basili, R., Cammisa, M., and Moschitti, A. Effective use of WordNet semantics via kernel-based learning. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)* (Ann Arbor, Michigan, June 2005), Association for Computational Linguistics, pp. 1–8.

[12] Ben-Hur, A., Horn, D., Siegelmann, H. T., and Vapnik, V. N. Support vector clustering. *Journal of Machine Learning Research 2* (2001), 125–137.

[13] Bertero, M., and Boccacci, P. *Introduction to Inverse Problems in Imaging.* Institute of Physics Publishing, January 1998.

[14] Bi, J., Chen, Y., and Wang, J. Z. A sparse support vector machine approach to region-based image categorization. In *IEEE Computer Vision and Pattern Recognition CVPR* (2005), pp. 1121–1128.

[15] Bishop, C. M. *Neural networks for pattern recognition.* Oxford University Press, Oxford, UK, 1996.

[16] Blake, C., and Merz, C. UCI repository of machine learning databases, 1998.

[17] Blei, D. M., and Jordan, M. I. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), ACM Press, pp. 127–134.

[18] Borg, I., and Groenen, P. J. F. *Modern Multidimensional Scaling.* New York, Springer, 1997.

[19] Bouten, M., Schietse, J., and den Broeck, C. V. Gradient descent learning in perceptrons: A review of its possibilities. *Physical Review E52*, 2 (1995), 1958–1967.

[20] Boyd, S., and Vandenberghe, L. *Convex Optimization.* Cambridge University Press, 2004. Available at `http://www.stanford.edu/~boyd/cvxbook.html`.

[21] Bressan, M., and Vitrià, J. Nonparametric discriminant analysis and nearest neighbor classification. *Pattern Recognition Letters 24*, 15 (2003), 2743–2749.

[22] Burges, C. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2*, 2 (1998), 121–167.

[23] Chang, E. Y., Goh, K., Sychay, G., and Wu, G. Cbsa: content-based soft annotation for multimodal image retrieval using Bayes point machines. *IEEE Trans. Circuits Syst. Video Techn. 13*, 1 (2003), 26–38.

[24] Chapelle, O., Haffner, P., and Vapnik, V. N. Support vector machines for histogram-based image classification. *IEEE Transaction Neural Networks 10*, 5 (1999), 1055–1064.

[25] Chen, Y., Zhou, X., and Huang, T. One-class svm for learning in image retrieval. In *IEEE International Conf. on Image Processing (ICIP'01)* (Thessaloniki, Greece, 2001).

[26] Chow, G. C. Tests of equality between sets of coefficients in two linear regressions. *Econometrica 28*, 3 (1960).

[27] Collobert, R., and Bengio, S. SVMTorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research 1* (2001), 143–160.

[28] Collobert, R., Bengio, S., and Bengio, Y. A parallel mixture of SVMs for very large scale problems. *Neural Computation 14*, 05 (2002), 1105–1114.

[29] Commandeur, J., Groenen, P. J. F., and Meulman, J. A distance-based variety of non-linear multivariate data analysis, including weights for objects and variables. *Psychometrika 64*, 2 (June 1999), 169–186.

[30] Courant, R., and Hilbert, D. *Methoden der mathematischen Physik. I.* Springer-Verlag, Berlin, 1968. Dritte Auflage, Heidelberger Taschenbücher, Band 30.

[31] Cox, D., and O'Sullivan, F. Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics 18* (1990), 1676–1695.

[32] Cox, I., Miller, M., Minka, T., Papathornas, T., and Yianilos, P. PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing 9*, 1 (2000), 20–37.

[33] Cristianini, N., and Shawe-Taylor, J. *An introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[34] Culotta, A., and Sorensen, J. Dependency tree kernels for relation extraction. In *42nd Annual Meeting of the Association for Computational Linguistics* (Barcelona, Spain, 2004).

[35] Cumby, C., and Roth, D. On kernel methods for relational learning. In *International Conference on Machine Learning (ICML)* (2003).

[36] DeCoste, D., and Schölkopf, B. Training invariant support vector machines. *Machine Learning 46*, 1-3 (2002), 161–190.

[37] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. Indexing by latent semantic analisys. *Journal of the American Society of Information Science 41*, 6 (1990), 391–407.

[38] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society 39*, 1 (1977), 1–38.

[39] Deng, Y., Manjunath, B. S., and Shin, H. Color image segmentation. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)* (Fort Collins, CO, 1999), vol. 2, pp. 446–451.

[40] Dietterich, T. G. Ensemble methods in machine learning. In *First International Workshop on Multiple Classifier Systems*, J. Kittler and F. Roli, Eds. Springer-Verlag, 2000, pp. 1–15.

[41] Dinkelbach, W. On nonlinear fractional programming. *Management Science A*, 13 (1967), 492–498.

[42] Duda, R. O., and Hart, P. E. *Pattern Classification and Scene Analysis*. John Wiley, June 1973.

[43] Fisher, R. A. The use of multiple measures in taxonomic problems. *Ann. Eugenics 7* (1936), 179–188.

[44] Fix, E., and Hodges, J. Discriminatory analysis: Nonparametric discrimination: Consistency properties. Tech. Rep. 4, USAF School of Aviation Medicine, February 1951.

[45] Fletcher, R. *Practical methods of optimization*. Chichester, England: Wiley, 1987.

[46] Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. Selective sampling using the query by committee algorithm. *Machine Learning 28* (1997), 133–168.

[47] Friedman, J. Another approach to polychotomous classification. Tech. rep., Stanford University, 1996.

[48] Fukunaga, K. *Introduction to statistical pattern recognition*, 2nd ed. Academic Press, New York, 1990.

[49] Fukunaga, K., and Mantock, J. Nonparametric discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 5*, 6 (1983), 671–678.

[50] Gentle, J. *Numerical linear algebra for applications in statistics*. Springer-Verlag, Berlin, 1998.

[51] Gill, P. E., Murray, W., and Wright, M. H. *Practical Optimization*. Academic Press, London and New York, 1981.

[52] Goh, K.-S., Chang, E., and Cheng, K.-T. Svm binary classifier ensembles for image classification. In *Proceedings of the tenth international conference on Information and knowledge management* (2001), ACM Press, pp. 395–402.

[53] Gosselin, P. H., and Cord, M. RETIN AL: an active learning strategy for image category retrieval. In *IEEE International Conference on Image Processing (ICIP'04)* (2004), pp. 2219–2222.

[54] Gosselin, P. H., and Cord, M. Semantic kernel learning for interactive image retrieval. In *IEEE International Conference on Image Processing (ICIP'05)* (Genova, Italy, 2005).

[55] Gosselin, P. H., and Cord, M. Semantic learning methods: Application to image retrieval. In *Conférence francophone sur l'apprentissage automatique* (Nice, France, 2005), pp. 109–110.

[56] Grandvalet, Y., Mariéthoz, J., and Bengio, S. A probabilistic interpretation of SVMs with an application to unbalanced classification. IDIAP-RR 26, IDIAP, 2005.

[57] Grauman, K., and Darrell, T. The pyramid match kernel: discriminative classification with sets of image features. In *International Conference on Computer Vision ICCV* (2005).

[58] Haasdonk, B. Feature space interpretation of SVMs with indefinite kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 4 (April 2005), 482–492.

[59] Haasdonk, B., and Bahlmann, C. Learning with distance substitution kernels. In *26th Pattern Recognition Symposium of the German Association for Pattern Recognition (DAGM 2004)* (Tübingen, Germany, 2004), Springer Verlag.

[60] Haasdonk, B., and Keysers, D. Tangent distance kernels for support vector machines. In *Proceedings of the 16th ICPR* (2002), pp. 864–868.

[61] Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. Canonical correlation analysis an overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.

[62] Hastie, T., and Tibshirani, R. Discriminant adaptive nearest neighbor classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*, 6 (1996), 607–616.

[63] Heiser, W. Convergent computation by iterative majorization: Theory and applications in multidimensional data analysis. *Recent advances in descriptive multivariate analysis* (1995), 157–189.

[64] Herbrich, R., and Graepel, T. Large scale bayes point machines. In *Advances in Neural Information Processing Systems 13* (2001), T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., MIT Press, pp. 528–534.

[65] Herbrich, R., Graepel, T., and Campbell, C. Robust bayes point machines. In *European Symposium on Artificial Neural Networks* (2000), pp. 49–54.

[66] Herbrich, R., Graepel, T., and Campbell, C. Bayes point machines. *Journal of Machine Learning Research 1* (August 2001), 245–279.

[67] Huber, P. Robust estimation of a location parameter. *Annals of Mathematical Statistics 35* (1964), 73–101.

[68] Hull, D. A. Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science 47*, 1 (1996), 70–84.

[69] Jeon, J., Lavrenko, V., and Manmatha, R. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (2003), ACM Press, pp. 119–126.

[70] Jing, F., Li, M., Zhang, H.-J., and Zhang, B. Learning region weighting from relevance feedback in image retrieval. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Orlando, Florida, 2002).

[71] Jing, F., Li, M., Zhang, H.-J., and Zhang, B. Support vector machines for region-based image retrieval. In *IEEE International Conference on Multimedia & Expo* (Baltimore,MD, 2003).

[72] Jing, F., Li, M., Zhang, L., Zhang, H.-J., and Zhang, B. Learning in region-based image retrieval. In *International Conference on Image and Video Retrieval* (2003).

[73] Joachims, T. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1999.

[74] Joachims, T. SVM-light support vector machine. Webpage, March 2002. Available at `http://www.cs.cornell.edu/People/tj/svm_light/`, Version: 5.00, Date: 03.07.2002.

[75] Kiers, H. Majorization as a tool for optimizing a class of matrix functions. *Psychometrika 55* (1990), 417–428.

[76] Kimeldorf, G. S., and Wahba, G. Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl. 33*, 1 (1971), 82–95.

[77] Kohrs, A., and Merialdo, B. Clustering for collaborative filtering applications. In *Proceedings of Computational Intelligence for Modelling, Control & Automation* (1999), IOS Press.

[78] Koltchinskii, V., and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Ann. Statist. 30*, 1 (2002), 1–50.

[79] Kosinov, S., Marchand-Maillet, S., and Pun, T. Iterative majorization approach to the distance-based discriminant analysis. In *Proceedings of the 28th Annual Conference of the GfKl 2004* (Dortmund, Germany, March 9–11 2004).

[80] Kosinov, S., Marchand-Maillet, S., and Pun, T. Visual object categorization using distance-based discriminant analysis. In *Proceedings of the 4th International Workshop on Multimedia Data and Document Engineering* (Washington, DC, July 2004).

[81] Kosinov, S., Marchand-Maillet, S., and Pun, T. Countering the false positive projection effect in nonlinear asymmetric classification. In *The IEEE Symposium on Signal Processing and Information Technology (ISSPIT'05)* (Athens, Greece, December, 18-21 2005).

[82] Kosinov, S., Titov, I., and Marchand-Maillet, S. Large margin multiple hyperplane classification for content-based multimedia retrieval. In *Machine Learning Techniques for Processing Multimedia Content, ICML Workshop* (Bonn, Germany, August, 11 2005).

[83] Krogh, A., and Hertz, J. A. A simple weight decay can improve generalization. In *Advances in Neural Information Processing Systems* (1992), J. E. Moody, S. J. Hanson, and R. P. Lippmann, Eds., vol. 4, Morgan Kaufmann Publishers, Inc., pp. 950–957.

[84] Kubat, M., and Matwin, S. Addressing the curse of imbalanced training sets: one-sided selection. In *Proc. 14th International Conference on Machine Learning* (1997), pp. 179–186.

[85] Kumar, S., Ghosh, J., and Crawford, M. M. Hierarchical fusion of multiple classifiers for hyperspectral data analysis. *Pattern Analysis and Applications 5* (2002), 210–220.

[86] Landauer, T., and Littman, M. Fully automatic cross-language document retrieval using latent semantic indexing. In *Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research* (Waterloo, Ontario, 1990), UW Centre for the New OED and Text Research, pp. 31–38.

[87] Lawrence, S., and Giles, C. Overfitting and neural networks: Conjugate gradient and backpropagation. In *Proceedings of the IEEE International Conference on Neural Networks* (2000), IEEE Press, pp. 114–119.

[88] Leeuw, J. D. Fitting distances by least squares. Tech. Rep. 130, Interdiviional Program in Statistics, UCLA, Los Angeles, CA, 1993.

[89] Leibe, B., and Schiele, B. Analyzing appearance and contour based methods for object categorization. In *International Conference on Computer Vision and Pattern Recognition (CVPR'03)* (Madison, Wisconsin, June 2003), pp. 409–415.

[90] Lewis, D. D., and Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval* (Dublin, IE, 1994), W. B. Croft and C. J. van Rijsbergen, Eds., Springer Verlag, pp. 3–12.

[91] Li, B., and Goh, K. Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of the eleventh ACM international conference on Multimedia* (2003), ACM Press, pp. 195–206.

[92] Li, J., Gray, R. M., and Olshen, R. A. Joint image compression and classification with vector quantization and a two dimensional hidden markov model. In *Data Compression Conference* (1999), pp. 23–32.

[93] Li, J., and Wang, J. Z. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Anal. Mach. Intell. 25*, 9 (2003), 1075–1088.

[94] Li, Y., and Shapiro, L. G. Object recognition for content-based image retrieval. In *Lecture Notes in Computer Science*. Springer-Verlag, 2004.

[95] Lin, H.-T., and Lin, C.-J. A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Tech. rep., Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2003. Available at `http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf`.

[96] MacKay, D. J. C. Bayesian non-linear modelling for the prediction competition. In *ASHRAE Transactions, V.100, Pt.2* (Atlanta GA, 1994), American Society of Heating, Refrigeration, and Air-conditioning Engineers, pp. 1053–1062.

[97] MacKay, D. J. C. Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems 6* (1995), 469–505.

[98] Marchand-Maillet, S., and Bruno, E. Exploiting user interaction for semantic content-based image retrieval. In *Content-based image and video retrieval (Dagstuhl seminar)* (2004), Lecture Notes in Computer Science, Springer Verlag.

[99] MARY, X. *Sous-espaces hilbertiens, sous-dualités et applications*. PhD thesis, INSTITUT NATIONAL DES SCIENCES APPLIQUEES DE ROUEN - INSA ROUEN, December 2003. ASI-PSI.

[100] Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society London (A) 209* (1909), 415–446.

[101] Mika, S., Rätsch, G., Weston, J., Schölkopf, B., and Müller, K. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing IX* (1999), Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds., IEEE, pp. 41–48.

[102] Miller, G. A. Wordnet: a lexical database for English. *Commun. ACM 38*, 11 (1995), 39–41.

[103] Mitchell, T. *Machine Learning*. McGraw Hill, 1997.

[104] Mitchell, T. M. Version spaces: An approach to concept learning. Tech. Rep. HPP-79-2, Stanford University, Palo Alto, CA, 1978.

[105] Mitchell, T. M. *Version spaces: An approach to concept learning*. PhD thesis, Stanford University, 1979.

[106] Monay, F., and Gatica-Perez, D. On image auto-annotation with latent space models. In *Proc. ACM Int. Conf. on Multimedia (ACM MM)* (November 2003).

[107] Moré, J. J., and Sorensen, D. C. Computing a trust region step. *SIAM Journal on Scientific and Statistical Computing 4*, 3 (1983), 553–572.

[108] Moreno, P. J., Ho, P. P., and Vasconcelos, N. A kullback-leibler divergence based kernel for svm classification in multimedia applications. In *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. MIT Press, Cambridge, MA, 2004.

[109] Mori, Y., Takahashi, H., and Oka, R. Image-to-word transformation based on dividing and vector quantizing images with words. *In First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)* (1999).

[110] Müller, H., Pun, T., and Squire, D. M. Learning from user behavior in image retrieval: Application of the market basket analysis. *International Journal of Computer Vision 56*, 1–3 (2004), 65–66.

[111] Nesterov, Y., and Nemirovskii, A. *Interior Point Polynomial Methods in Convex Programming: Theory and Applications*. Society for Industrial and Applied Mathematics, Philadelphia, 1994.

[112] Ong, C., Smola, A., and Williamson, R. Learning the kernel with hyperkernels. *Journal of Machine Learning Research 6* (07 2005), 1043–1071.

[113] Ong, C. S., Mary, X., Canu, S., and Smola, A. J. Learning with non-positive kernels. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning* (2004), ACM Press.

[114] Ong, C. S., Smola, A. J., and Williamson, R. C. Hyperkernels. In *Neural Information Processing Systems* (2002), vol. 15, MIT Press.

[115] Opper, M., and Haussler, D. Generalization performance of Bayes optimal classification algorithm for learning a perceptron. *Physical Review Letters 66*, 20 (May 1991), 2677–2680.

[116] Paredes, R., and Vidal, E. A class-dependent weighted dissimilarity measure for nearest neighbor classification problems. *Pattern Recognition Letters 21*, 12 (2000), 1027–1036.

[117] Platt, J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Schölkopf, and D. Schurmans, Eds. MIT Press, 1999.

[118] Porter, M. F. An algorithm for suffix stripping. *Program 14*, 3 (1980), 130–137.

[119] Pozdnoukhov, A., and Bengio, S. Tangent vector kernels for invariant image classification with SVMs. In *Proceedings of the 17th International Conference on Pattern Recognition, ICPR'04* (August 2004), vol. 3, pp. 486–489.

[120] Pozdnoukhov, A., and Bengio., S. Improving kernel classifiers for object categorization problems. In *International Conference on Machine Learning, ICML, Workshop on Learning with Partially Classified Training Data* (2005).

[121] Praks, P., Dvorsky, J., and Snasel, V. Latent semantic indexing for image retrieval systems. In *Proceedings of the SIAM Conference on Applied Linear Algebra (LA03)* (Williamsburg, USA, 2003), The College of William and Mary.

[122] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., and Riedl, J. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work* (Chapel Hill, North Carolina, 1994), ACM, pp. 175–186.

[123] Ribeiro, B., and Carvalho, P. Mercer's kernel based learning for fault detection. In *Soft Computing Systems - Design, Management and Applications* (2002), A. Abraham, J. Ruiz-del-Solar, and M. Köppen, Eds., Frontiers in Artificial Intelligence and Applications Vol. 87, IOS Press Amsterdam, Berlin, Oxford, Tokyo, Washington D.C., pp. 341–350.

[124] Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[125] Rocchio, J. Relevance feedback information retrieval. In *The Smart retrieval system experiments in automatic document processing*, G. Salton, Ed. Prentic-Hall, 1971, pp. 313–323.

[126] Rojas, M., Santos, S., and Sorensen, D. A new matrix-free algorithm for the large-scale trust-region subproblem. *SIAM Journal on Optimization 11*, 3 (2000), 611–646.

[127] Rui, Y., Huang, T., Ortega, M., and Mehrotra, S. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Video Technology 8*, 5 (1998), 644–655.

[128] Rujár, P. Playing billiards in version space. *Neural Computation 9*, 1 (1997), 99–122.

[129] S.A. Solla, T. L., and Müller, K.-R., Eds. *Learning from user feedback in image retrieval systems* (2000), MIT Press.

[130] Salton, G., and McGill, M. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[131] Salton, G., Wang, A., and Yang, C. A vector space model for information retrieval. *Journal of the American Society for Information Science 18* (1975), 613–620.

[132] Santini, S., and Jain, R. Beyond query by example. In *ACM Multimedia* (1998), pp. 345–350.

[133] Schölkopf, B. The kernel trick for distances. In *Advances in Neural Information Processing Systems 13* (2001), T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., MIT Press, pp. 301–307.

[134] Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.-R., Rätsch, G., and Smola, A. J. Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks 10*, 5 (1999), 1000–1017.

[135] Schölkopf, B., and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2002.

[136] Shawe-Taylor, J., and Williamson, R. C. A PAC analysis of a bayesian estimator. In *Tenth Annual Conference on Computational Learning Theory COLT'97* (Nashville, Tennessee, July 1997), pp. 2–9.

[137] Smith, J. R., and Chang, S.-F. Tools and techniques for color image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIF)* (1996), pp. 426–437.

[138] Sonnevend, G. An 'analytic center' for polyhedrons and new classes of global algorithms for linear (smooth, convex) programming. In *System Modelling and Optimization : Proceedings of the 12th IFIP–Conference held in Budapest, Hungary, September*

1985, A. Prekopa, J. Szelezsan, and B. Strazicky, Eds., vol. 84 of *Lecture Notes in Control and Information Sciences*. Springer Verlag, Berlin, Germany, 1986, pp. 866–876.

[139] Sonnevend, G. New algorithms in convex programming based on a notion of 'centre' (for systems of analytic inequalities) and on rational extrapolation. In *Trends in Mathematical Optimization : Proceedings of the 4th French–German Conference on Optimization in Irsee, Germany, April 1986*, K. H. Hoffmann, J. B. Hiriat-Urruty, C. Lemarechal, and J. Zowe, Eds., vol. 84 of *International Series of Numerical Mathematics*. Birkhäuser Verlag, Basel, Switzerland, 1988, pp. 311–327.

[140] Squire, D. M., Müller, W., Müller, H., and Pun, T. Content-based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99) 21*, 13-14 (2000), 1193–1198. B.K. Ersboll, P. Johansen, Eds.

[141] Squire, D. M., Müller, W., Müller, H., and Raki, J. Content-based query of image databases, inspirations from text retrieval: inverted files, frequency-based weights and relevance feedback. In *The 11th Scandinavian Conference on Image Analysis* (Kangerlussuaq, Greenland, june 1999), pp. 143–149.

[142] Su, Z., Zhang, H., and Ma, S. Using Bayesian classifier in relevant feedback of image retrieval. In *12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'00)* (Vancouver, British Columbia, Canada, 2000), p. 258.

[143] Theodoridis, S., and Koutroumbas, K. *Pattern Recognition*. Academic Press, London, 1999.

[144] Tong, S., and Chang, E. Y. Support vector machine active learning for image retrieval. In *ACM Multimedia 2001* (Ottawa, Ontario, Canada, 2001), pp. 107–118.

[145] Torkkola, K., and Campbell, W. Mutual information in learning feature transformations. In *Proc. 17th International Conf. on Machine Learning* (June 2000), pp. 1015–1022.

[146] Trafalis, T. B., and Malyscheff, A. M. An analytic center machine. *Machine Learning 46*, 1-3 (January 2002), 203–223.

[147] Tresp, V. A bayesian committee machine. *Neural Computation 12*, 11 (2000), 2719–2741.

[148] Ulusoy, I., and Bishop, C. M. Generative versus discriminative methods for object recognition. In *IEEE Computer Vision and Pattern Recognition, CVPR* (2005), pp. 258–265.

[149] van Deun, K., and Groenen, P. J. F. Majorization algorithms for inspecting circles, ellipses, squares, rectangles, and rhombi. Tech. rep., Econometric Institute Report EI 2003-35, 2003.

[150] van Rijsbergen, C. *Information retrieval*. Butterworth, London, 1979.

[151] Vapnik, V. N. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

[152] Vapnik, V. N. *The nature of statistical learning theory*. Springer, New York, 1995.

[153] Vapnik, V. N. *Statistical Learning Theory*. Wiley, New-York, 1998.

[154] Vinokourov, A., Hardoon, D., and Shawe-Taylor, J. Learning the semantics of multimedia content with application to web image retrieval and classification. In *Fourth International Symposium on Independent Component Analysis and Blind Source Separation* (Nara, Japan, 2003).

[155] Vinokourov, A., Shawe-Taylor, J., and Cristianini, N. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge, MA, 2003, pp. 1473–1480.

[156] Wahba, G. Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV. In *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, 1998, pp. 69–88.

[157] Wallace, C., and Dowe, D. Minimum message length and Kolmogorov complexity. *Computer Journal 42*, 4 (1999), 270–283.

[158] Wan, V., and Renals, S. Speaker verification using sequence discriminant support vector machines. *IEEE Trans. on Speech and Audio Processing 13* (2005), 203–210.

[159] Watanabe, H., Yamaguchi, T., and Katagiri, S. Discriminative metric design for robust pattern recognition. *IEEE Trans. Signal Processing 45*, 11 (1997), 2655–2661.

[160] Watkin, T. Optimal learning with a neural network. *Europhysics Letters 21* (1993), 871–877.

[161] Webb, A. Multidimensional scaling by iterative majorization using radial basis functions. *Pattern Recognition 28*, 5 (1995), 753–759.

[162] Weiss, S., and Kulikowski, C. *Computer Systems That Learn*. Morgan Kaufmann, 1991.

[163] Wu, G., Chang, E. Y., and Zhang, Z. Learning with non-metric proximity matrices. In *ACM International Conference on Multimedia (MM)* (Singapore, November 2005).

[164] Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics 32* (2004), 56–134.

[165] Zhao, R., and Grosky, W. From features to semantics: Some preliminary results. In *IEEE International Conference on Multimedia and Expo (II)* (2000), pp. 679–682.

[166] Zhao, R., and Grosky, W. Narrowing the semantic gap - Improved text-based document web document retrieval using visual features. *IEEE Trans. on Multimedia 4*, 2 (2002), 189–200.

[167] Zhou, X., Garg, A., and Huang, T. A discussion of nonlinear variants of biased discriminants for interactive image retrieval. In *Proceedings of CIVR'04* (Dublin, Ireland, 2004), pp. 353–364.

[168] Zhou, X., and Huang, T. Comparing discriminating transformations and SVM for learning during multimedia retrieval. In *Proceedings of the 9th ACM international Conference on Multimedia* (Ottawa, Canada, 2001), pp. 137–146.

[169] Zhou, X., and Huang, T. Small sample learning during multimedia retrieval using BiasMap. In *IEEE Computer Vision and Pattern Recognition (CVPR'01)* (Hawaii, 2001).

Résumé

R.1 Préface

Cette thèse porte sur le développement théorique et l'application pratique de méthodes d'apprentissage automatique à l'analyse et la récupération de multimédia par le contenu. Afin d'être clair et pour éviter toute confusion, une attention particulière doit être portée à la manière dont cette thèse interprète et utilise le terme *multimédia*. Ce terme possède plusieurs significations et interprétations, ses définitions évoluant et se transformant en parallèle avec le progrès technologique.

D'après sa signification littérale, dérivant du latin *multus* = "plusieurs, multiple..." et *médium* = "un système de communication, information ou distraction", le multimédia représente l'utilisation d'ordinateurs pour présenter des textes, graphiques, vidéos, animations et sons de manière intégrée. Or, l'utilisation du terme multimédia est tout à fait différente du point de vue d'un système d'apprentissage automatique, ou de tout autre système d'information multimédia en général. Un système d'information traite la même abstraction binaire d'une information digitalisée et parfois codée, quelle que soit son origine sensorielle. Comme certains auteurs le discutent, le multimédia dans ce contexte se réfère à n'importe quelle information visuelle, auditive ou textuelle, unique ou combinée. L'information multimédia digitale est immédiatement visible, audible, lisible et dans la plupart des cas compréhensible par l'utilisateur, mais pas pour le système. Cette divergence importante entre la représentation digitale et la représentation est connue sous le nom de *fossé sémantique*, la cible principale des techniques développées dans cette thèse.

R.2 Problématique

Dans cette thèse, nous identifions les approches qui s'intéressent au fossé sémantique discuté plus haut, comme étant celles d'augmentation sémantique puisque la plupart d'entre elles portent spécifiquement sur le rapprochement de représentations visuelles de documents multimédia de bas niveau et la signification de ces même documents pour en améliorer l'efficacité d'accès et de récupération. Le choix d'un terme de référence aussi général que l'augmentation sémantique nous permet d'inclure et d'analyser entièrement un grand nombre de techniques qui ont le même but: atténuer le phenomène de fossé sémantique, résolu dans chaque méthode individuelle à travers un vaste éventail de paradigmes et formulations qui se rapportent à

l'indexation, l'apprentissage, la classification, la catégorisation, la prédiction, etc. Ce choix de terminologie est lié à la manière dont le terme multimédia est per cu tout au long de cette thèse. Multimédia est compris comme étant n'importe quelle combinaison de représentation d'une information humainement compréhensible, dont les caractéristiques automatiquement calculables ne contiennent pas d'expression directe d'une signification pouvant être recherchée par l'utilisateur.

Un bon contexte pour décrire la contribution de cette thèse est défini en considérant deux groupes de méthodes pour l'augmentation sémantique: interactives - les approches adaptatives qui sont guidées par un retour de pertinence de l'utilisateur, et automatiques - celles qui tentent de dériver des corrélations utiles entre des caractéristiques représentatives du multimédia et ses aspects sémantiques en appliquant des techniques qui n'impliquent pas d'utilisateur. Les méthodes du premier groupe considèrent l'utilisateur comme étant la source ultime d'information sémantique. Elles corrigent leur solution de manière itérative en demandant et en incorporant un feedback de l'utilisateur et contiennent des contributions importantes telles que l'algorithme de Rocchio, le filtrage collaboratif, l'estimation de pertinence Bayesienne, l'apprentissage actif entre autres. D'un autre côté, les méthodes du second groupe agissent de manière complètement autonome et ne sollicitent jamais d'aide de l'utilisateur, comptant uniquement sur des techniques comme l'apprentissage automatique, la classification, l'analyse discriminante, l'indexation par sémantique sous-jacente, l'analyse de corrélation entre les langues, pour induire les significations recherchées à partir des données d'entraînement à leur disposition. Les contributions présentées dans cette thèse appartiennent au groupe des méthodes d'augmentation sémantique automatique et sont considérées du point de vue de l'apprentissage automatique. De ce point de vue, les techniques proposées sont con cues pour être capables d'améliorer leur performances sur la base des expériences et résultats précédents de fa con autonome, dans une tentative d'éliminer le besoin, ou alléger le poids placé sur l'intuition humaine dans l'analyse d'un problème posé. Malgré le fait que le besoin de connaissance et d'intuition humaine ne pourra probablement jamais être complètement éliminé à cause de l'importance de décisions intelligentes dans la représentation et la caractérisation des données, l'approche de l'apprentissage automatique présente un avantage clair et incontestable. En effet, en résolvant un problème plus général et probablement plus difficile en apprentissage automatique, la même technologie peut s'appliquer à une grande variété de problèmes particuliers sans avoir à reconstruire la solution à partir de rien à chaque fois. Une fois résolue, la même méthode générale est applicable dans les scénarios d'augmentation sémantique pour les images digitales, les signaux auditifs, les films vidéo, et autres combinaisons.

Cette thèse adopte la perspective de l'apprentissage automatique dans l'approche des problèmes d'augmentation sémantique dans l'accès et la récupération efficace à l'intériuer de collections multimédia et établit plusieurs contributions dans les domaines détaillés dans les chapitres suivants, à savoir l'analyse discriminante, l'apprentissage par machine à noyaux et la contexte adapté de classification hiérarchique.

R.3 Analyse discriminante

Le choix du domaine de l'analyse discriminante comme point de départ de notre étude reflète une tendance et une préférence délibérées pour une méthode d'apprentissage automatique discriminante par rapport à une approche alternative générative. Cette décision est motivée par un certain nombre de raisons. D'abord, le modèle discriminant a beaucoup plus de flexibilité dans les parties de l'espace d'entrée où les probabilités postérieures diffèrent de manière significative de 0 ou de 1, alors que les approches génératives modèlisent des détails de distribution de données de l'espace d'entrée ce qui peut être non pertinent pour déterminer des probabilités postérieures. En second lieu, les modèles discriminants sont en général très rapidea à produire des prévisions pour des données-tests, alors que les modèles génératifs exigent souvent une solution itérative. Troisièmement, toutes choses étant égales, on peut prévoit que les méthodes discriminants aient une meilleure performance prédictive puisqu'ells sont definies pour prévoir le label de classe plutôt que distribution commune des données de l'espace d'entrée et des labels. Ainsi, tenant compte les propriétés spécifiques de l'application destinée à l'augmentation sémantique, nous développons une technique d'analyse discriminante transformationnelle basée sur la distance, DDA. Un effort dédié est engagé à faire à la formulation de la DDA pour soutenir un caractère *non paramétrique* avec prétentions minimales sur la distribution de données de l'espace d'entrée, *asymétrique* à apparier le scénario de déploiement le plus populaire "1-against-all", et basé sur la *transformation* du domaine d'entré afin de tenir compte d'extensions, post-traitement et l'utilisation de la transformation dérivée pour fournir une métrique discriminante. Cette métrique explique les différences dans les échelles de différents caractéristiques, retire les corrélations globales et redondances parmi des caractéristiques dans une certaine mesure, et s'adapte au fait que quelques caractéristiques peuvent être beaucoup plus instructifs au sujet des labels de classe que d'autres. Afin de satisfaire ces dernieres conditions, la capacité de la DDA à extraire les caractéristiques distinctifs et réduire la dimensionnalité de données de l'espace d'entrée, tout en déterminant le nombre de dimensions suffisantes automatiquement, est d'importance cruciale. Du point de vue de l'augmentation sémantique, la DDA fournit une machine binaire d'apprentissage utilisable pour distinguer entre un certain concept sémantique et son complément, par exemple déterminer si l'image montre un object d'intérêt ou pas. Plus formellement, la formulation de la technique d'analyse discriminante basée sur la distance (DDA) est récapitulée comme suit.

Dans le cadre de l'analyse discriminante, on recherche à distinguer entre deux ou plusieurs classes ou catégories sémantiques prédéfinies de documents multimédia. En considérant au départ le cas simple de deux classes sémantiques, nous recherchons une transformation des caractéristiques qui place les exemples d'une des classes proches les uns des autres, tout en tenant les exemples de l'autre classe suffisamment eloignés dans l'espace de caractéristiques considéré. En d'autres termes, la transformation linéaire recherchée $T \in \mathbb{R}^{m \times k}$ doit transformer les valeurs de façon à respecter l'hypothèse de compacité dans le but d'améliorer la

performance de la méthode de plus proche voisin (*Neearest Neighbour*, NN). Formellement, une telle transformation est caractérisée comme étant la solution d'un problème de minimisation globale du critère suivant:

$$\log J(T) = \frac{2}{N_X(N_X-1)} \sum_{i<j}^{N_X} \log \Psi\left(d_{ij}^W(T)\right) - \frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} \log d_{ij}^B(T) \qquad \text{(R.1)}$$

où N_X, N_Y sont les nombres d'observations dans les ensembles de données X et Y correspondant aux caractéristiques des deux classes; $d_{ij}^W(T)$ dénote une distance entre les points i et j *à l'intérieur* de l'ensemble de données X transformé par T, et, de façon analogue, $d_{ij}^B(T)$ indique distances *entre* le point i de l'ensemble de données transformées X et le point j de l'ensemble des données transformées Y, $\Psi\left(d_{ij}^W(t)\right)$ est une fonction d'évaluation robuste de Huber.

Le problème de minimisation (R.1) est résolu par la technique de majorisation itérative, qui remplace l'optimisation globale d'une fonction quelconque par une suite itérative de minimisations plus simples de fonctions auxiliaires. Etant donné les propriétés de ces fonctions auxiliaires, généralement appellées *fonctions de majorization*, la procédure itérative produit une suite non-croissante de valeurs de fonction convergeant vers un point stationnaire qui est une fonction minimum locale sous certaines contraintes. Pour le critère choisi d'optimisation, (R.1), nous dérivons une fonction de majorization approximative (à une constante indépendante de T près) comme exprimé en (R.2):

$$\mu_{\log J}(T,\bar{T}) = \frac{\alpha}{2}\mathbf{tr}\left(T^T X^T R X T\right) + \frac{\beta}{2}\mathbf{tr}(T^T Z^T G Z T) - 2\beta\mathbf{tr}(T^T Z^T G Z \bar{T}), \qquad \text{(R.2)}$$

où $\bar{T}$ est le point support de la transformation recherchée, c'est-à-dire, sa valeur à l'itération actuelle, R et G sont les matrices pour les calculs des distances, et Z est la matrice qui contient les deux ensembles de données X et Y (rassemblés par lignes):

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}. \qquad \text{(R.3)}$$

À chaque itération, la minimisation de (R.2) est résolue sous la contrainte de régularisation parametrée par Δ, produisant un résultat qui devient un point support de l'itération suivante, et ainsi de suite, jusqu'à ce que la convergence soit atteinte. Ce processus de dérivation de la transformation discriminante ainsi que la classification de NN constitue l'essence de la méthode d'analyse discriminante basée sur la distance (DDA) que nous proposons et qui est récapitulée par l'algorithme ci-dessous :

Algorithme DDA.

1. assigner un premier point support $\bar{T} = \bar{T}_0 \in \mathbb{R}^{m\times k}$;
2. trouver le point successeur :

$$T_s = \arg\min_T \mu_{\log J}(T,\bar{T}) + \lambda\left(\mathbf{tr}(T^T T) - \Delta\right);$$

3. si $\log J(\bar{T}) - \log J(T_s) < \epsilon$, s'arrêter;

4. assigner $\bar{T} = T_s$, aller en 2.

Notons que le choix de la taille (en colonnes) de T à une valeur arbitraire $k \ll m$ transforme notre méthode en une technique de réduction de dimension, laquelle peut être utilisée dans différentes applications telles que la sélection de caractéristiques, la visualisation de données, *etc.* Par ailleurs, la valeur de k, c'est-à-dire la dimension jusqu'à laquelle les données peuvent être réduites sans perte de pouvoir discriminant (selon (R.1)), est déterminée avec précision par le nombre de valeurs singulières différentes de zéro de T. En effet, les distances entre les observations transformées peuvent être considérées comme distances entre les observations originales dans une métrique différente TT^T. Ceci peut être exprimé en utilisant la décomposition en valeur singulières $TT^T = USV^TVSU^T = U_kS_k^2U_k^T$. L'expression obtenue indique que l'effet de la transformation est capturé par les k premiers vecteurs singuliers gauches multipliés par leur valeurs singulières différentes de zéro. Ce nombre k donne une réponse à la question de combien de dimensions sont nécessaires pour représenter l'espace transformé.

La discussion ci-dessus sous-entend une configuration de deux classes. La généralisation de la formulation présentée à une configuration d'analyse discriminante pour un nombre de classes $K \geq 2$ ne pose pas de problèmes particuliers:

$$\log J_K(T) = \sum_{i=1}^{K-1} \left(\alpha^{(i)} S_W(T)^{(i)} - \beta^{(i)} S_B(T)^{(i)} \right). \tag{R.4}$$

A noter que (R.4) devient (R.1) pour $K = 2$.

En plus de la transformation explicitement recherchée, notre DDA peut également fournir une métrique discriminante qui tient compte des différences d'échelles pour des caractéristiques différentes, supprimer dans une certaine mesure les corrélations et les redondances globales parmi des caractéristiques et s'adapte au fait que certaines caractéristiques peuvent être beaucoup plus informatives en termes de labels de classes que d'autres. Cette observation est aisément illustrée par l'exemple de la machine à vecteur de support (ou support à vaste marges – SVM) avec noyau Gaussien:

$$k_\Sigma(x_i, x_j) = e^{-(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}, \tag{R.5}$$

pour une certaine matrice Σ de covariance et des observations x_i, x_j représentée comme des vecteurs colonne. Un choix typique de σ est la matrice d'identité multipliée par un certain facteur constant. Cependant, quand la DDA est appliquée pour prétraiter les données d'entraînement avant la procédure SVM, le classifieur SVM tire pleinemnt profit des caractéristiques les plus disriminantes extraites par la DDA. Les produits des noyaux pouvant être vus comme évalués dans une nouvelle métrique discriminante TT^T:

$$k_\Sigma(x_i, x_j) = e^{-(x_i - x_j)^T TT^T (x_i - x_j)}. \tag{R.6}$$

Cela conduit au fait que le SVM est potentiellement capable de trouver une solution plus simple, impliquant peu de vecteurs de support et améliorant les propriétés de généralisation, ce qui mène naturellement à une amélioration de performance de classification, comme confirmé par les résultats expérimentaux. Une évaluation empirique plus poussée a confirmé la performance de classification de la méthode proposée de DDA et ses extensions sur un certain nombre d'ensembles de données de référence (ensembles UCI) et sur les tâches de recherche sémantique d'images par le contenu. Les résultats encourageants ont démontré que l'approche proposée de DDA surpasse plusieurs méthodes populaires, et améliore le qualité de classification quand combinée avec d'autres techniques. Ceci fait de la DDA un excellent candidat pour l'application visée d'augmentation sémantique automatique.

R.4 Méthodes basées sur noyaux

Un grand nombre d'algorithmes linéaires d'apprentissage automatique ont été mis en valeur à travers l'utilisation de noyaux afin de pouvoir traiter des problèmes plus complexes exigeant des fonctions de décision non-linéaires. De même, la méthode DDA peut être étendue en une technique *d'analyse discriminante non linéaire* , surmontant l'hypothése de linéarité de la transformation discriminante recherchée et menant naturellement au développement de la KDDA, une extension basée sur les noyaux de la DDA. Ceci est fait en reformulant le problème en termes de distances projetées de l'espace de caractéristiques plus riche, potentiellement de dimension infinie, induit par une fonction de noyau choisie comme détaillé ci-dessous.

Soit un espace $\mathcal{F}$ duquel des échantillons des données d'entraînement peuvent être extraites par l'intermédiaire de $\Phi : \mathbb{R}^m \to \mathcal{F}$, tels qu'il existe une fonction noyau $k(x, y) = (\Phi(x))^T \Phi(y)$, où $x, y \in \mathbb{R}^m$ et $k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. Nous supposerons également que la transformation distinctive est recherchée dans $\mathcal{F}$ comme une matrice ω de projection de la taille $[\mathscr{N}_{\mathcal{F}} \times d]$, où $\mathscr{N}_{\mathcal{F}}$ est la dimension de $\mathcal{F}$, et d la dimension du sous-espace de projection discriminante dérivé, tel que les colonnes de ω se situent dans l'espace de toutes les échantillons de formation tracé dans $\mathcal{F}$. Alors, en vertu du Théorème de Representation:

$$\omega = \left[\sum_i^N \alpha_i^{(1)} \Phi(z_i) \;\; \sum_i^N \alpha_i^{(2)} \Phi(z_i) \;\; \ldots \;\; \sum_i^N \alpha_i^{(d)} \Phi(z_i) \right], \tag{R.7}$$

où z_i est un des $N_X + N_Y$ échantillons provenant des données d'entraînement de la matrice composée Z (cf (R.3)). Les distances entre les images des échantillons x et y projetées dans $\mathcal{F}$ par la solution ω sont ainsi données par:

$$\begin{aligned} \mathscr{D}_{xy}^2(\omega) &= (\Phi(x) - \Phi(y))^T \omega\omega^T (\Phi(x) - \Phi(y)) \\ &= \mathbf{tr}\left(\omega^T (\Phi(x) - \Phi(y))(\Phi(x) - \Phi(y))^T \omega\right) \\ &= \sum_j^d \left(\sum_i^N \alpha_i^{(j)} (k(z_i, x) - k(z_i, y)) \right)^2 . \end{aligned} \tag{R.8}$$

On peut simplifier (R.8) en utilisant une notation matricielle:

$$\mathscr{D}^2_{xy}(\omega) \equiv \mathscr{D}^2_{xy}(P) = \mathbf{tr}\left(P^T H_{xy} P\right) \tag{R.9}$$

où $P \in \mathbb{R}^{N\times d}$ est la transformation non-linéaire recherchée et représentée comme une matrice rassemblant tout les $\alpha_i^{(j)}$, $H_{xy} = (K_x - K_y)(K_x - K_y)^T$. $K_s = [k(z_1, s), k(z_2, s), \ldots, k(z_n, s)]^t$ dénote un vecteur d'évaluations du noyau pour l'échantillon s et tout le reste de l'ensemble des données d'entraînement.

Du fait de (R.9), le logarithme du critère de l'optimisation de DDA (R.1) peut maintenant être exprimé en termes de distances projetées dans un espace plus riche et potentiellement de dimension infinie $\mathcal{F}$:

$$\begin{aligned}\log J(P) &= \frac{2}{N_X(N_X-1)}\sum_{i=1}^{N_X}\sum_{j=i+1}^{N_X}\log\Psi\left(\mathscr{D}^W_{ij}(P)\right)\\ &\quad -\frac{1}{N_X N_Y}\sum_{i=1}^{N_X}\sum_{j=1}^{N_Y}\log\mathscr{D}^B_{ij}(P)\end{aligned} \tag{R.10}$$

Le traitement du critère obtenu diffère légèrement du cas linéaire. De même que pour la DDA, l'optimisation de (R.10) est obtenue par majorisation itérative. La fonction de majorisation est, dans ce cas:

$$\begin{aligned}\mu_{\log J}(P,\bar{P}) &= \frac{1}{N_X(N_X-1)}\mathbf{tr}\left(P^T\mathbb{K}_X B(\bar{P})\mathbb{K}_X^T P\right)\\ &\quad +\frac{1}{2N_X N_Y}\mathbf{tr}\left(P^T\mathbb{K}_{XY} C\mathbb{K}_{XY}^T P\right)\\ &\quad +\frac{2}{N_X N_Y}\mathbf{tr}\left(P^T\mathbb{K}_{XY} G(\bar{P})\mathbb{K}_{XY}^T \bar{P}\right)\\ &\quad + const,\end{aligned} \tag{R.11}$$

où $\bar{P}$ est la solution courante, $\mathbb{K}_X$, $\mathbb{K}_{XY}$ sont des matrices des produits scalaires des noyaux évalués entre X et toutes les données, respectivement, et B, C et G sont matrices positives semi-définies indépendantes de P.

Un aspect important distinguant la KDDA d'autres méthodes basées sur les noyaux est la convexité de la formulation qui persiste indépendamment du fait que le noyau fondamental est défini positif. Cette propriété est évaluée empiriquement par l'incorporation des *noyaux indéfinis* dans la KDDA. Dans plusieurs domaines d'application, ces noyaux sont connus comme correspondant à des mesures de distance qui modélisent la similitude perceptuelle.

On examine ensuite une condition défavorable dénommée "effet de projection faussement positif" et on évalue des stratégies d'élimination.

Pour l'augmentation sémantique, l'approche KDDA étend la formulation précédente au cas non linéaire, et rend également la machine de classification binaire DDA utilisable si on l'utilise avec des mesures non-métriques de dissimilarité par l'intermédiaire des noyaux indéfinis.

R.5 Ensembles hiérarchique de classifieurs

La plupart des problèmes pratiques d'augmentation sémantique ne peuvent être traités en totalité par une machine binaire simple. L'extension la plus populaire dans des cas multi-classes, multi-catégories ou multi-concepts sémantiques est de dériver autant de machines binaires que nécessaire, une pour chaque label de classe et, potentiellement établir un arbitrage parmi leurs prédictions. Dans notre étude, nous postulons qu'une telle construction et ses variations ne sont pas avantageuses dans le contexte d'augmentation sémantique. La raison fondementale est l'hypothèse d'un ensemble de concepts sémantiques cibles étant *indépendant, non-redondant, et exhaustif*. Tout en fournissant les moyens d'étendre les techniques existantes à des cas multi-catégories, ceci peut mener à des résultats contradictoires, par exemple des prédictions de combinaisons de concepts peu probables telles que le "sous-marin" et de "sable de désert" ou à estimer une erreur de fausse classification entre "fleuve" et "lac" aussi importante qu'entre "fleuve" comme "avion". Afin de réduire ces limitations, nous proposons une approche pour modéliser explicitement les rapports sémantiques hiérarchiques entre les classes visées, automatiquement dérivés et étendus grâce à un lexique sémantique. En pratique, cette méthode génère un ensemble sémantique hiérarchique (HSE) de différents classifieurs, chacun exprimé comme une machines binaire DDA. Tous travaillent ensemble en influençant des décisions de autres par des liens exprimés par la structure des relations inter-concepts

Etant donné un ensemble de documents d'etraînements annotés $\{I_t, K_t\}_{t=1}^{n}$, où I_t représente le vecteur caractéristiques d'un document donné et K_t son ensemble de mots-clés associé, la structure de HSE est déterminée par la hiérarchie de concepts $H = \{C_i\}_{i=1}^{N}$, formant un graphe orienté. Les arêtes de H sont définies par les rapports de type "hypernym-hyponym" qui lient ses noeuds C_i, représentés par tous les noms uniques du vocabulaire d'annotation $V = \bigcup_{t=1}^{n} K_t$ et leurs hyponyms (extraits de WordNet). Dans H, chaque concept C_i occupe un noeud séparé, et est associé à un classifieur binaire Θ_i conçu pour distinguer l'ensemble de concepts feuilles inclu (directement ou indirectement) par C_i (noté $\mathbf{L}(C_i)$) de tous les autres. Un exemple d'une hiérarchie dérivée pour un vocabulaire simple V:{*beach, flower, grass, mountain, rock, sky, tree*} est donné en Figure Q.1. Afin d'arriver à l'augmentation sémantique d'une image représentée par un vecteur de caractéristiques I_U, chaque concept C_i est évalué en tant que candidat potentiel. Ainsi, le choix des annotations possibles n'est plus limité par V, ce qui est le cas pour la majorité d'autres techniques semblables. La pertinence de C_i est vue comme différence entre, d'une part, la correspondance entre la représentation des données I_U et la description de la catégorie C_i et d'autre part, le niveau de détail ou de non-ambiguïté apporté par l'ensemble de mots-clés candidats $\mathbf{L}(C_i)$. Dans méthode HSE que nous proposons, la première variable est représentée par la probabilité postérieure d'un concept par rapport aux données, $P(C_i|I_U)$, tandis que la seconde variable est estimée comme $P(C_i|k)$, la probabilité postérieure d'un concept donné sachant qu'un mot-clé k de l'ensemble de tous les hyponyms de C_i est choisi. Formulé en termes bayésiens, la méthode HSE apporte

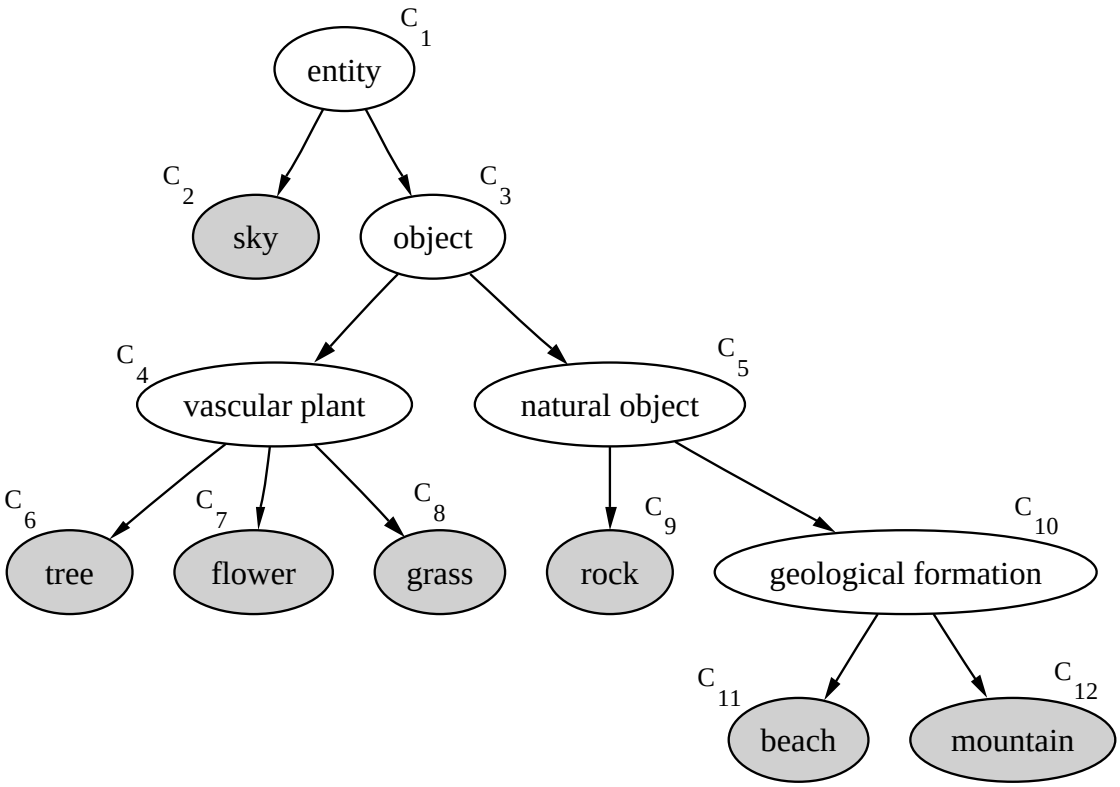


Figure R.1: Exemple de hiérarchie. Les noeuds grisés appartiennent à $C_i \in V$

la possibilité supplémentaire de pouvoir déterminer automatiquement le nombre de concepts C_i à prédire. Les résultats encourageant de l'évaluation empirique menée sur plusieurs ensembles d'apprentissage et plusieurs classifieurs confirment la viabilité de l'approche proposée.

R.6 Aspects théoriques

En plus de l'évaluation empirique de toutes les méthodes proposées, un effort particulier a été investi pour établir des relations théoriques entre nos propositions et les mthodes de classification modernes. Ainsi, des évaluations sont menées pour trouver des parallèles entre les techniques DDA, SVM et ACM (*Analytical Center Machine*). Nous commencons par la traduction géométrique des contraintes de séparabilité qui engendre la notion de l'espace de version (*version space*), dont la solution obtenue par un SVM correspond à un centre de Tchebycheff et celle définie par l'ACM est un centre analytique. Dans ce contexte, la formulation de la DDA est démontrée comme étant une approximation du critère minimisé par ACM. Des résultats analogues sont obtenus par la DDA et la technique ACM dans le cas idéal où les ensembles de données sont parfaitement séparables et leurs projections presque equidistantes. Ceci peut être attribué au fait les critères des deux méthodes deviennent semblables dans ces conditions. En outre, la méthode DDA peut être vue comme un parent proche d'ACM, en raison de l'utilisation des fonctions potentielles du "-log". On note toutefois

une distinction importante: les ensembles de données séparables et non-séparables sont traités de la même façon, car les distances sont toujours non négatives et donc toujours dans le domaine admissible du logarithme (excepté pour le cas des noyaux indéfinis, comme cités précédemment). Cette dernière dernière distinction rapproche la DDA des SVM, puisque ni l'un ni l'autre n'exige la séparabilité de données, alors que ACM ne permet pas une région adminissible vide (par exemple, si les échantillons de données ne sont pas séparable). En conclusion, la taille 1 en colonnes de la transformation recherchée nécessaire pour effectuer l'analyse ci-dessus, est atypique pour le DDA. En effet, notre meilleure performance empirique est observée quand la dimension de l'espace de cible est supérieure à 1. Pour SVM et ACM, ceci correspond à avoir plusieurs hyperplans de séparation utilisés dans un classifieur, au lieu d'un seul. Afin d'étudier plus rigoureusement des implications de la dernière configuration inhérente à la DDA, nous considérons une extension explicite au cas d'hyperplans multiples (MH), démontrant ainsi les possibilités d'extension d'une telle formulation, garantissant la généralisation de performance en termes de la notion de "fat-shattering".

R.7 Perspectives futures

Plusieurs résultats contenus dans ce travail ouvrent des voies intéressantes pour de prochaines recherches. Dans l'analyse binaire discriminante, nous avons vu le bénéfice de l'asymétrie et de l'orientation de la formulation vers une transformation de l'espace. Ceci mérite une exploration plus poussée. Les avantages apparents de la technique de majorisation itérative sont mis en valeur empiriquement. Nous préconisons l'usge de cette technique là une une optimisation est mise en jeu. Notre proposition est confortée par le nombre croissant de publications impliquant cette stratégie numérique.

Dans le domaine des méthodes basées sur les noyaux, nous avons proposé une manière uniforme de traiter les noyaux indéfinis comme les noyaux semi définis positifs traditionnels. Ce premier type de noyaux est lié à des mesures correspondant à des distances non métriques pouvant potentiellement mieux capturer la similarité perceptuelle. Ceci renforce l'importance de cette contribution à l'application d'algorithmes d'apprentissage automatique dans le traitement de l'information perceptuelle.

Pour les classifications multi-catégories, nous avons démontré la validité du mécanisme proposé pour expliquer les relations entre différentes catégories-cibles, qui peut être appliqué et facilement étendu à d'autres domaines. Nous avons d'ailleurs déjà commencé à explorer plus avant le contexte HSE afin d'améliorer l'approche proposée en prenant avantage de la structure donnée par la classification hiérarchique résultante. Ceci afin d'incorporer un retour (*feedback*) de l'utilisateur, étendant ainsi l'approche au domaine des méthodes de l'augmentation sémantiques interactives.

Les connections théoriques établies entre nos propositions sur l'analyse discriminante et les méthodes d'apprentissage modernes soulignent les propriétés uniques de la formulation proposée. Ceci nous a conduit à une étude des configurations à hyperplans multiples dans le

contexte des marges larges.

L'expérimentation ecourageante et les résultats positifs décrits tout au long de cette thèse nous donnent à penser que la recherche de progrès sur les technologies d'apprentissage automatique pour l'augmentation sémantique sont à la fois viables et justifiables.