An application of one-class support vector machine to nosocomial infection detection

Cohen, Gilles; Hilario, Mélanie; Sax, Hugo; Hugonnet, Stéphane; Pellegrini, Christian; Geissbuhler, Antoine

MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

An Application of One-class Support Vector Machines to Nosocomial Infection Detection

Gilles Cohen[a], Mélanie Hilario[b], Hugo Sax[c], Stéphane Hugonnet[c], Christian Pellegrini[b], Antoine Geissbuhler[a]

[a] *Medical Informatics Service, University Hospital of Geneva, 1211 Geneva,Switzerland*
[b] *Artificial Intelligence Laboratory, University of Geneva, 1211 Geneva, Switzerland*
[c] *Department of Internal Medicine, University Hospital of Geneva, 1211 Geneva, Switzerland*

Abstract

Nosocomial infections (NIs)---those acquired in health care settings---are among the major causes of increased mortality among hospitalized patients. They are a significant burden for patients and health authorities alike; it is thus important to monitor and detect them through an effective surveillance system. This paper describes a retrospective analysis of a prevalence survey of NIs done in the Geneva University Hospital. Our goal is to identify patients with one or more NIs on the basis of clinical and other data collected during the survey. In this two-class classification task, the main difficulty lies in the significant imbalance between positive or infected (11%) and negative (89%) cases. To cope with class imbalance, we investigate one-class SVMs which can be trained to distinguish two classes on the basis of examples from a single class (in this case, only "normal" or non infected patients). The infected ones are then identified as "abnormal" cases or outliers that deviate significantly from the normal profile. Experimental results are encouraging: whereas standard 2-class SVMs scored a baseline sensitivity of 50.6% on this problem, the one-class approach increased sensitivity to as much as 92.6%. These results are comparable to those obtained by the authors in a previous study on asymmetrical soft margin SVMs; they suggest that one-class SVMs can provide an effective and efficient way of overcoming data imbalance in classification problems.

Keywords:

Nosocomial infections, Prevalence, Infection control, Surveillance, One-class learning, Support Vector Machines.

Introduction

Infection control is a major and constant concern of health care institutions; nosocomial[1] infections, in particular, tend to attract particular attention insofar as they directly engage the responsibility of hospital authorities. Thus the increasing emphasis on surveillance to monitor and detect infections, nosocomial or not. It provides data to assess the magnitude of the problem, detect outbreaks, identify risk factors for infection, target control measures on high-risk patients or wards, or evaluate prevention programs. Ultimately, the goal of surveillance is to decrease infection risk and consequently improve patients' safety.

Two methods are generally applied to perform surveillance: (1) trans-sectional assessment (i.e. prevalence studies), because of their capability to give estimates on a large population at relatively low cost; or (2) prospective, ongoing surveillance (incidence studies). The gold standard is the latter, which consists in reviewing on a daily basis all available information on all hospitalized patients in order to detect all nosocomial infections. However, this method is labor-intensive, infeasible at a hospital level, and currently recommended only for high-risk, i.e., critically ill patients. As an alternative and more realistic approach, prevalence surveys are being recognized as a valid surveillance strategy and are becoming increasingly performed. Their major limitations are their retrospective nature, the dependency on readily available data, a prevalence bias, the inability to detect outbreak (depending on the frequency the surveys are performed), and the limited capacity to identify risk factors. However, they provide sufficiently good data to measure the magnitude of the problem, evaluate a prevention program, and help allocate resources. They give a snapshot of clinically active NIs during a given index day and provide information about the frequency and characteristics of these infections. The efficacy of infection control policies can be easily measured by repeated prevalence surveys [1].

Materials and Methods

Setting and data collection

The University Hospital of Geneva (HUG) has been performing yearly prevalence studies since 1994 [2]. The methodology of prevalence surveys is as follows. The investigators visit all wards of the HUG over a period of approximately three weeks. All patients hospitalized for 48 hours or more at the time of the study are included. Medical records, kardex, X-ray and microbiology reports are reviewed, and additional information eventually obtained by interviews with nurses or physicians in charge of the patient. All nosocomial infections are recorded according to modified Centres for Disease Control (CDC) criteria. Only infections still active at any point during the six days preceding the visit are included. Collected variables include demographic characteristics, admission date, admission diagnosis, comorbidities, McCabe score, type of admission, provenance, hospitaliza-

[1] A nosocomial infection (from the Greek word *nosokomeion* for hospital) is one that develops during a patient's hospitalization whereas it was not present or incubating at the time of the admission.

tion ward, functional status, previous surgery, previous intensive care unit (ICU) stay, exposure to antibiotics, antacid and immunosuppressive drugs and invasive devices, laboratory values, temperature, date and site of infection, fulfilled criteria for infection.

Although less time-consuming than prospective surveillance, a prevalence survey nevertheless requires considerable resources, i.e., approximately 800 hours for data collection and 100 hours for entering data in an electronic data base. Due to this important effort, we can afford to perform such studies only once a year. What is particularly time-consuming is the careful examination of all available information for all patients, in order to detect those who might be infected. The aim of this pilot study is to apply data mining techniques to data collected in the 2002 prevalence study in order to detect nosocomial infections on the basis of the factors described above.

Data preprocessing

The dataset consisted of 688 patient records and 83 variables. With the help of hospital experts on nosocomial infections, we filtered out spurious records as well as irrelevant and redundant variables, reducing the data to 683 cases and 49 variables. In addition, several variables had missing values, due mainly to erroneous or missing measurements. These values were assumed to be missing at random, as domain experts did not detect any clear correlation between the fact that they were missing and the data (whether values of the incomplete variables themselves or of others). We replaced these missing values with the class-conditional mean for continuous variables and the class-conditional mode for nominal ones.

The class skew problem

Our nosocomial dataset shares the class imbalance problem observed in many real world applications, especially in the medical domain. Out of 683 patients, only 75 (11% of the total) were infected and 608 were not. Data imbalance is particularly detrimental in classification problems where the heavily under-represented class is precisely the class of interest. There have been several proposals for coping with imbalanced datasets [3], including: oversampling the minority class, under-sampling or downsizing the majority class [4], or a combination of both [5, 6]; building cost-sensitive classifiers [7] that penalize more heavily misclassification of the minority class; and rule-based methods that attempt to learn high confidence rules for the minority class [8]. In this paper we investigate another alternative, known as recognition-based learning or novelty detection, which consists in simply ignoring one of the two classes and learning from a single class [9, 10]. This approach is quite atypical: to minimize generalization error in a classification problem involving $c \geq 2$ classes, the standard approach is to build a discriminating hypothesis h based on training cases from all c classes..

One-class classification can be viewed as an attempt to distinguish between new cases similar to members of the training set and all other cases that can occur. In a probabilistic sense, one-class classification is equivalent to deciding whether an unknown test case is produced by the underlying distribution that corresponds to the training set of normal cases. While it appears similar to conventional binary classification problems, one-class classification differs in the way a classifier is trained. It is trained only on cases from one class, and never sees those from the second class. It must therefore estimate the boundary that separates those two classes based only on data which lie on one side.

Learning with one-class Support Vector Machines

Support vector machines [11, 12] (SVMs) are learning machines based on the *Structural Risk Minimization principle* (SRM) from statistical learning theory. They were originally introduced for solving two-class pattern recognition problems. An adaptation of the SVM methodology in order to handle classification problems using data from only one class has been proposed by [13, 14]. This adapted method, termed one-class SVM, identifies "abnormal" cases amongst the known cases and assumes them to belong to the complement of the "normal cases". Schölkopf et al. formulate the one-class SVM approach as follows:

Consider a training set $\{x_i\}$ i=1,...,n $x_i \in \Re^d$ and suppose them distributed according to some unknown underlying probability distribution P. We want to know if a test example x is distributed according to P or not. This can be done by determining a region R of the input space X such that the probability that a test point drawn from P lies outside of R is bounded by some a priori specified value $\nu \in (0,1)$. This problem is solved by estimating a decision function *f* which is positive on R and negative elsewhere.

$$f(x) > 0 \quad \text{if } x \in R \text{ and } f(x) < 0 \text{ if } x \notin R \qquad (1.1)$$

A non linear function $\Phi : X \rightarrow \Im$ maps vector x from the input vector space X endowed with an inner product to a Hilbert space $\Im$ termed feature space. In this new space, the training vectors follow an underlying distribution P', and the problem is to determine a region R' of $\Im$ that captures most of this probability mass distribution. In other words the region R' corresponds to the part of the feature space where most of the data vectors lie. To separate as many as possible of the mapped vectors from the origin in feature space $\Im$ we construct a hyperplane H(w,ρ) in feature space defined by

$$H(w, \rho) = \langle w, \Phi(x) \rangle - \rho \qquad (1.2)$$

where w is the weight vector and ρ the offset, as illustrated in Fig. 1.

The maximum margin from the origin is found by solving the following quadratic optimization problem.

$$\text{Minimize} \cdot \frac{1}{2}\left(\langle w, w \rangle + \frac{1}{\nu n} \cdot \sum_{i=1}^{n} \zeta_i - \rho\right) \qquad (1.3)$$

$$\text{subject to } (\langle w, \Phi(w) \rangle \geq \rho - \zeta_i) \qquad \zeta_i \geq 0$$

where ξ_i are so-called slack variables that penalize the objective function but allow some of the points to be on the wrong side of the hyperplane, i.e. located between the origin and H(w,ρ) as depicted in Fig.1. $\nu \in (0,1)$ is a parameter that controls the trade-off between maximizing the distance from the origin and containing most of the data in the region created by the hyperplane. It is proved in [14] that ν is an upper bound on the fraction of outliers i.e. training errors, and also a lower bound on the fraction of support vectors.

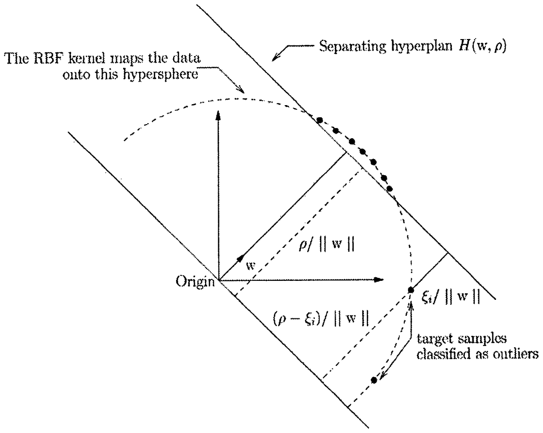
Separating hyperplan $H(\mathrm{w}, \rho)$
The RBF kernel maps the data onto this hypersphere
$\rho/ \|\mathrm{w}\|$
w
Origin
$\xi_i/ \|\mathrm{w}\|$
$(\rho - \xi_i)/ \|\mathrm{w}\|$
target samples classified as outliers

Figure 1 - Schematic 2D overview of a one-class SVM classifier. In the feature space, the vectors are located on a hypersphere. The hyperplane H*(w,ρ) separates the training vectors from the rest of the surface of the hypersphere*

Let $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ be n non negative Lagrange multipliers associated with the constraints, the solution to the problem is equivalent to the solution of the Wolfe dual [15] problem.

$$Minimize \ \frac{1}{2}\alpha_i\alpha_j\langle\Phi(x_i), \Phi(x_j)\rangle$$

suject to

$$0 \le \alpha_i \le \frac{1}{\nu n}, \qquad \sum_{i=1}^{n} \alpha_i = 1 \qquad (1.4)$$

the solution for w is

$$w = \sum_{i=1}^{n} \Phi(x_i) \qquad \text{where} \quad 0 \le \alpha_i \le \frac{1}{\nu n} \qquad (1.5)$$

and the corresponding decision function is :

$$f(x_j) = \operatorname{sgn}(\alpha_i\langle\Phi(x_i), \Phi(x_j)\rangle - \rho) \qquad (1.6)$$

All training data vectors x_i for which $f(x_i) \le 0$ are called *support vectors* (SVs); these are the only vectors for which $\alpha_i \neq 0$. SVs are divided in two sets : the *margin* SVs, for which $f(x_i) = 0$, and the *non-margin* SVs, for which $f(x_i) < 0$.

Notice that in (1.4) only inner products between data are considered; for certain particular maps Φ, there is no need to actually compute $\Phi(x_i)$ and $\Phi(x_j)$; the inner product can be derived directly from x_i and x_j by means of the so-called "kernel trick". A kernel K is a symmetric function that fulfills Mercer's [11, 15] conditions. The main property of functions satisfying these conditions is that they implicitly define a mapping from X to a Hilbert space $\mathfrak{I}$ such that

$$K(x_i, x_j) = \langle\Phi(x_i), \Phi(x_j)\rangle \qquad (1.7)$$

and thus can be used in algorithms using inner products. Accordingly, the hyperplane (1.2) in feature space $\mathfrak{I}$ becomes a non linear function in the input space X.

$$f(x_j) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i K\langle x_i, x_j\rangle - \rho\right) \qquad (1.8)$$

There are many admissible choices for the kernel function $K(x_i, x_j)$. The most widely used in one-class SVMs is the Gaussian Radial Basis Function RBF kernel:

$$K(x_i, x_j) = e^{(-|x_i, x_j|^2)/(2\sigma^2)} \qquad (1.9)$$

where σ is a parameter that controls the "width" of the kernel function around x_i. Since $\langle\Phi(x_i),\Phi(x_i)\rangle = K(x_i, x_i) = e^0 = 1$ with an RBF kernel, the training data in $\mathfrak{I}$ lie on a region on the surface of a hypersphere centered at the origin of $\mathfrak{I}$ with radius 1 as depicted in Fig. 1.

Finally we obtain the decision function of Eq. (1.8) with

$$\rho = \sum_{i=1}^{n} \alpha_i K\langle x_i, x_j\rangle \quad \text{for any } \alpha_i \text{ satisfies } \ 0 < \alpha_i < \frac{1}{\nu n}$$

which defines the contour of the region R in input space by cutting the hypersurface defined by the weighted addition of SVM kernels at a given altitude ρ.

Experimentation and Results

The experimental goal was to assess the ability of one-class SVMs to cope with imbalanced datasets. To train one-class SVM classifiers we used an RBF kernel (Eq. (1.9)) and experimented with different values for the ν and σ parameters. Generalization error was estimated using 5-fold cross-validation. The extreme imbalance between the classes precluded the use of 10-fold cross-validation, which would have resulted in an overly small number of infected test cases per fold. The complete dataset was thus randomly partitioned into five subsets. On each iteration, one subset (comprising 20% of the data samples) was held out as a test set and the remaining four (80% of the data) were concatenated into a training set. Note that while the test set should reflect the original class distribution for error estimation to be plausible, one-class learning dictates restriction of the training set to a single class (in this case non infected patients. Error rates estimated on the test sets were then averaged over the five iterations. Overall performance was quantified using the metrics discussed in the following section.

Performance Measures

A widely used performance metric in classification is accuracy, i.e. the fraction of correctly classified data points in the test set. When the prior probabilities of the classes are very different, such metrics might be misleading. For instance, on a dataset with a 95%-5% class distribution, it is straightforward to attain 95% accuracy by simply assigning each new case to the majority class. Despite the impressive accuracy, such a solution is inacceptable in medical diagnosis, as the classifier would have failed to recognize a single diseased case (assuming healthy cases are the majority). Performance metrics that dissociate errors specific to each class are needed.

To discuss alternative performance criteria we adopt the standard definitions used in binary classification. TP and TN stand for the number of true positives and true negatives respectively, i.e., positive/negative cases recognized as such by the classifier. FP and FN represent respectively the number of misclassified positive and negative cases. In two-class problems, the accuracy rate on the positives, called sensitivity, is defined as : TP/(TP+FN), whereas the accuracy rate on the negative class, also known as specificity, is : TN/(TN+FP). Classification accuracy is simply (TP + TN)/ TP+TN+FP+FN.

In medical diagnosis [16], biometrics and recently machine learning [17], a more flexible way of assessing a classification method is the receiver operating characteristic (ROC) curve. A ROC curve plots sensitivity versus 1-specificity for different thresholds of the classifier output. Based on the ROC curve, one can decide how many false positives (respectively false negatives) one is willing to tolerate and tune the classifier threshold to best suit a certain application. A random assignment of classes to data would result in a ROC curve in the form of a diagonal line from (0,0) to (1,1).

Table 1: Performance of one-class SVMs for different parameter settings using an RBF Gaussian kernel.

Parameters		Accuracy	Sensitivity	Specificity
ν	σ	%	%	%
0.05	**10^{-4}**	**74.56**	**92.60**	**43.73**
	0.1	75.49	80.60	65.60
	0.15	72.51	70.39	74.40
	0.17	71.69	66.94	77.87
0.2	10^{-4}	75.69	79.28	68.27
	0.06	74.97	77.14	69.87
	0.07	74.97	76.82	70.40
	0.1	74.36	74.67	72.27

Findings

Table 1 summarizes performance results for one-class SVMs. It shows the best results obtained by training classifiers using different parameter configurations on non infected cases only.

Clearly highest sensitivity is attained when both ν and σ are small. As explained above (see "One-class Support Vector Machines"), ν is an upper bound on the fraction of outliers that can be ignored. Recall that in our application problem the outliers are the abnormal (infected) cases. With smaller values of ν, more abnormal cases are taken into account, which explains the higher sensitivity at the cost of decrease in specificity. Furthermore, when σ is small, the system puts a Gaussian of narrow width around each data point and hence most of the infected test cases are correctly recognized as abnormal. As the value of sigma increases, the region of influence of each Gaussian becomes larger and the normal cases tend to dominate the abnormal cases, thus increasing specificity at the cost of sensitivity. It is crucial to tune the ν and σ parameters in determining the balance between normality and abnormality as there is no explicit penalty for false positive in one-class classification, contrary to the two class formulation [18]. Since the goal of this study is to identify infected cases, the solution retained is that which achieves maximal sensitivity.

In a previous study on the same dataset [18], we investigated a support vector algorithm in which asymmetrical margins are tuned to improve recognition of rare positive cases. Table 2 shows the best performance measures obtained previously.

Table 2: Best performance of SVMs with symmetrical and asymmetrical margin.

SVM Classifier	Accuracy	Sensitivity	Specificity
Sym. Margin	89.6%	50.6%	94.4%
Asym. Margin	**74.4%**	**92%**	**72.2%**

A comparison of Tables 1 and 2 shows that both one-class and asymmetrical margin SVMs lead to significant improvements in sensitivity over classical symmetrical SVMs. While the maximal sensitivity attained by one-class SVMs (92.6%) is slightly higher (92%) than that of the asymmetrical margin approach, the latter achieves significantly higher specificity.

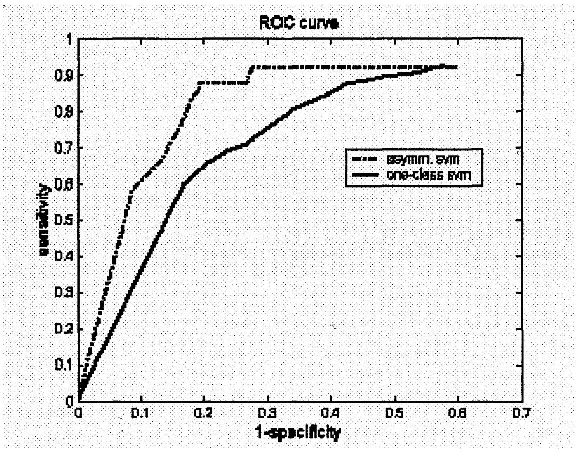

Figure 2 - ROC curve for one-class SVMs classifiers varying σ. Also plotted is the ROC curve for asymmetrical SVMs

In Fig. 2 a ROC curve shows the trade-off between sensitivity and false positive rate (1-specificity) . Ideally, we want high sensitivity (to detect most of the infected patients) and a low false positive rate (to avoid mistakenly classifying non infected patients as infected). To allow for a direct comparison between asymmetrical-margin and one-class SVMs, the ROC curve of asymmetrical SVMs is also plotted. It is clear that while both approaches attain the same maximal level of sensitivity, asymmetrical SVMs do so at a much lower cost in specificity.

Conclusion and future work

We analyzed the results of a prevalence study of nosocomial infections in order to detect patients with infections. The major hurdle, typical in medical diagnosis, is the problem of rare positives. To address this problem we investigated the applicability of an algorithm proposed by [14] to estimate the support of a distribution. Experimental results reported in this paper are encouraging. From the point of view of sensitivity, one-class SVMs attain the highest level (92.6%) observed by the authors throughout a series of studies on the problem. However, the price paid in terms of loss in specificity is quite exhorbitant, and domain experts must decide if the high recognition rate is worth the cost of treating false positive cases. From this point of view, asymmetrical-margin SVMs might prove preferable in that they maintain a more reasonable sensitivity-specificity trade-off.

In the near future, we intend to prospectively validate the classification model obtained by performing in parallel a standard prevalence survey. We also plan to improve overall accuracy of one-class SVMs by enhancing the resolution in the support region boundaries via *conformal transformation*, an approach described in the context of the two-class SVMs by [19]. Overall we feel that one-class SVMs are a promising approach to the detection of nosocomial infections and can become a reliable component of an infection control system.

Acknowledgments

The authors are grateful for the dataset provided by the infection control team at the University of Geneva Hospitals.

References

[1] French G, Cheng AF, Wong SL, Donnan S. Repeated prevalence surveys for monitoring effectiveness of hospital infection control. *Lancet* 1983;2:1021-23.

[2] Harbarth S, Ruef C, Francioli P, Widmer A, Pittet D, Network S-N. Nosocomial infections in swiss university hospitals: a multicentre survey and review of the published experience. *Schweiz Med Wochenschr* 1999;129:1521-28.

[3] Japkowicz N. Learning from Imbalanced Data Sets: A comparison of various Strategies. In: AAAI Workshop on Learning from Imbalanced Data Sets; 2000; Menlo Park, CA: AAAI Press; 2000.

[4] Kubat M, Matwin S. Addressing the Curse of Imbalanced Training Sets: One sided selection. In: Kaufmann M, editor. Proceedings of 14th International Conference in Machine Learning; 1997; San Francisco, CA; 1997. p. 179-186.

[5] Chawla N, Bowyer K, Hall L, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence 0and Research* 2002;16:321-357.

[6] Cohen G, Hilario M, Sax H, Hugonnet S. Data Imbalance in surveillance of nosocomial infections. In: International Symposium on Medical Data Analysis; 2003; Berlin; 2003.

[7] Domingos P. MetaCost: A General Method for Making Classifiers Cost-Sensitive. *Knowledge Discovery in Data Mining* 1999:155-164.

[8] Ali K, Manganaris S, Srikant R. Partial Classification using Association Rules. In: *Knowledge Discovery in Databases and Data Mining*; 1997. p. 115-118.

[9] Manevitz LM, Youssef M. One-class SVMs for document classification. *Journal of Machine Learning Research* 2001;2.

[10] Kowalczyk A, Raskutti B. One Class SVM for yeast regulation prediction. *SIGKDD Explorations* 2002;4(2).

[11] Vapnik V. *Statistical Learning Theory*: Wiley; 1998.

[12] Cortes C, Vapnik V. Support vector networks. *Machine Learning* 1995;20(3):273-297.

[13] Schölkopf B, Williamson RC, Smola AJ, Shawe-Taylor J, Platt J. Support vector method for novelty detection. In: Adv. in Neural Information Processing Systems 12; 2000: MIT Press; 2000. p. 582-588.

[14] Scholkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson RC. Estimating the support of a high-dimensional distribution. *Neural Computation* 1999;13:1443-1471.

[15] Christianini N, Taylor JS. *An introduction to Support Vector Machines*; 2000.

[16] Centor RM. Signal detectability. *Medical Decision Making* 1991;11:102-6.

[17] Provost F, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. In: *Proc. Fifteenth International Conference on Machine Learning (ICML98)*: Morgan Kaufmann, San Francisco CA; 1998. p. 445-453.

[18] Cohen G, Hilario M, Sax H, Hugonnet S. Asymmetrical Margin Approach to Surveillance of Nosocomial Infections Using Support Vector Classification. In: Intelligent Data Analysis in Medicine and Pharmacology; 2003; Cyprus; 2003.

[19] Amari S, Wu A. Conformal transformation of kernel functions: A data-dependent way to improve support vector machine classifiers. *Neural processing Letter* February 2002;15(1):59-67.

Address for correspondence

Gilles Cohen,
Medical Informatics Service,
University Hospital of Geneva,
Rue Micheli-du-Crest 24, 1211 Geneva ,Switzerland
Tel : ++41 22 372 7550, Fax : ++41 22 372 8680
Gilles.Cohen@sim.hcuge.ch