*eHealth Beyond the Horizon – Get IT There*
*S.K. Andersen et al. (Eds.)*
*IOS Press, 2008*



# Novelty Detection using One-class Parzen Density Estimator. An Application to Surveillance of Nosocomial Infections

Gilles COHEN[a,1], Hugo SAX[b], Antoine GEISSBUHLER[a]
[a]*Medical Informatics Service, University Hospital of Geneva, 1211 Geneva , Switzerland*
[b]*Department of Internal Service, University Hospital of Geneva, 1211 Geneva , Switzerland*

**Abstract.** Nosocomial infections (NIs) - those acquired in health care settings - represent one of the major causes of increased mortality in hospitalized patients. As they are a real problem for both patients and health authorities, the development of an effective surveillance system to monitor and detect them is of paramount importance. This paper presents a retrospective analysis of a prevalence survey of NIs done in the Geneva University Hospital. The objective is to identify patients with one or more NIs based on clinical and other data collected during the survey. In this classification task, the main difficulty lies in the significant imbalance between positive and negative cases. To overcome this problem, we investigate one-class Parzen density estimator which can be trained to differentiate two classes taking examples from a single class. The results obtained 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 88.6%. These results suggest that one-class Parzen density estimator can provide an effective and efficient way of overcoming data imbalance in classification problems.

**Keywords.** Pattern Recognition, One-Class Classification, Parzen density estimator, Data Imbalance, Nosocomial Infections, Infection control.

## Introduction

A major and constant concern in Health Care Institutions is Infection control, particularly of nosocomial[2] origin, which directly engage the hospital responsibility.

[1] Corresponding Author: Gilles COHEN, Service d'Informatique Médicale, Hôpitaux Universitaires de Genève, Rue Micheli-du-Crest, 24, Genève, Switzerland, E-mail: gilles.cohen@sim.hcuge.ch
[2] A nosocomial infection is a disease that presents itself in hospitalized patients in whom the infection was not present at the time of the admission.

This leads to focus increasingly on surveillance to detect and monitor infections, nosocomial or not. Data are collected to assess the magnitude of the problem, detect outbreaks, identify risk factors, target control measures on high-risk patients or wards, and evaluate prevention programs. Finally, the surveillance aims to decrease infection risk and consequently improve patients' safety.

Two methods are generally used for surveillance: (1) trans-sectional assessment (i.e. prevalence studies), which gives estimates on a large population at relatively low cost; or (2) prospective, ongoing surveillance (incidence studies). The latter is the method of choice, however, this method is arduous, unfeasible at a hospital level, and currently recommended only for high-risk, i.e., critically ill patients. Prevalence surveys, which constitute an alternative and more realistic approach, are being recognized as a valid surveillance strategy and are increasingly used. 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 of the surveys), and the limited capacity to identify risk factors. However, the data they provide are sufficient to measure the magnitude of the problem, evaluate a prevention program, and help to 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].

## 1. Data collection and preparation

The University Hospital of Geneva (HUG) has been performing yearly prevalence studies since 1994 [2]. Their methodology is as follows: the investigators visit every ward 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, X-ray and microbiology reports are reviewed, and additional information is eventually obtained by interviewing nurses or physicians in charge. Each nosocomial infection is recorded according to modified Centres for Disease Control criteria. The survey includes only infections still active at any point during the six days preceding the visit. Collected variables include demographic characteristics, admission date, admission diagnosis, comorbidities, McCabe score, type of admission, provenance, hospitalization ward, functional status, previous surgery, previous intensive care unit 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. What is particularly time-consuming is the careful examination of each available information for all patients, in order to detect those who might be infected. This pilot study was aimed at applying machine learning techniques to data collected in the 2002 prevalence survey in order to detect nosocomial infections using the factors described above.

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.

## 2. The imbalanced data problem

The major difficulty inherent in the data (as in many medical diagnostic applications) is the highly skewed class distribution. Out of 683 patients, only 75 (11% of the total) were infected and 608 were not. The problem of imbalanced datasets is particularly crucial in applications where the goal is to maximize recognition of the minority class[3]. The issue of class imbalance has been actively investigated and remains largely open; it is handled in a number of ways [3], including: oversampling the minority class, building cost-sensitive classifiers [4] that assign higher cost to misclassifications of the minority class, stratified sampling on the training instances to balance the class distribution [5] and rule-based methods that attempt to learn high confidence rules for the minority class [6]. In this paper we investigate another way of biasing the inductive process to boost sensitivity (i.e., capacity to recognize positives) based on one-class Parzen density estimator. Experiments conducted to assess this approach are described in Section 5 and results are discussed in Section 6.

## 3. Classification algorithm

### *3.1. One-class classification*

In typical classification problems a discriminative hypothesis h is found based on training cases from all c (≥2) classes, so that it can classify each new unseen case into one of c classes with the lower possible generalisation error.

While many pattern recognition problems fall into this category, some other problems are best formulated differently as *one-class* or *novelty detection* problems. The one-class classification task can be expressed as the ability 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 a previously unseen test case has been produced by the underlying distribution of the training set of normal cases. While it seems to be similar to conventional classification problems (i.e. two-class), one-class classification differs in the way the classifier is trained. It is trained only by cases from the majority class, and never sees those from the minority class. It must estimate the boundary that separates those two classes based only on data which lies on one side of it. The one-class approach is particularly attractive in situations where cases from one class are expensive or difficult to obtain for model construction. In a one-class setting, novel or abnormal cases can be detected by constructing a real-valued density estimation function. The most straigthforward method is to estimate the density of the training data and to set a threshold on this

[3] For convenience we identify positive cases with the minority and negative cases the majority class.

density. In this study such an approach is used : a one-class Parzen density estimator. We briefly describe it in the following section.

### *3.2. One-class Parzen Density Estimator*

The Parzen Window [7,8,9] is a simple kernel based estimator for estimating density function. The main idea of this non-parametric method is to place a Gaussian kernel for each pattern within the pattern vector determining the position of the center of the kernel. The Gaussian kernels have a smoothing parameter that control the smoothness of the kernels to give appropriate classification performance. The smoothing parameter is normally global to the entire set of Gaussian kernels. Let $p(x)$ be the density function to be approximated. Consider a training set X of n i.i.d. examples drawn according to $p(x)$, the Parzen window estimate of $p(x)$ based on the X is $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \kappa(x - x_i)$ where $\kappa(.)$ is a kernel function with localized support and its exact form depends on n. The most commonly used is the Gaussian kernel : $\hat{p}(x) = \frac{1}{n(2\pi)^{d/2}\sigma^d}\sum_{i=1}^{n} exp\left\{-\frac{\|x - x_i\|^2}{2\sigma^2}\right\}$ here $x_i$ is an example in the training set, $\sigma$ is the smoothing function or bandwith and d is the dimensionality of the feature space. The density estimator $\hat{p}(x)$ obtained from the training set give us a quantitative measure of the degree of novelty for each new example. This is used to reject examples where the estimate $\hat{p}(x) < \rho$ for some threshold $\rho$, effectively generating a new class of "novel" data. Thus any point where the likelihood $\hat{p}(x)$ is below some threshold is considered to be novel. This approach was used by Bishop [10] who gives the following justification : Denote $C_1$ as normality and $C_2$ as novelty. The corresponding prior probabilities are $P(C_1)$ and $P(C_2)$ and the probability density function are $p(x|C_1)$ and $p(x|C_2)$. According to Bayes, $x \in C_1$ if $P(C_1|x) = p(x|C_1)\,P(C_1) > P(C_2|x) = p(x|C_2)\,P(C_2)$. Deciding whether a new example x is abnormal or novel depends on the comparaison between $p(x|C_1)$ and $p(x|C_2)P(C_2)/\,P(C_1)$, where the latter is equivalent to a threshold ρ on $\hat{p}(x)$.

## 4. Experimental Setup

### *4.1. Evaluation strategy*

The experimental goal was to assess the ability of one-class Parzen density estimator to cope with imbalanced datasets. To train one-class Parzen density estimator we experimented with different values for the parameter ρ for a fixed σ obtained by maximizing the "pseudo-likelihood" computed using cross validation (see Eq. 1).

$$\sigma = \underset{\sigma}{\text{argmax}}\left\{\frac{1}{n}\sum_{i=1}^{n}\log(f(x_i))\right\} \text{ where } f(x_j) = \frac{1}{(n-1)\sigma}\sum_{i=1, i\neq j}^{n}\kappa\left(\frac{x_j - x_i}{\sigma}\right) \quad (1)$$

For both approaches generalization error was estimated using 5-fold cross-validation. The complete dataset was 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 in this approach, the training sets consisted only of non infected patients whereas the test sets contained both infected and non infected patients according to the original class distribution. 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.

### *4.2. Performance Metrics*

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)/N, where N=TP+TN+FP+FP is the total number of cases.

## 5. Results

**Table 1.** Performance of one-class Parzen density estimator for different parameter settings using an RBF Gaussian kernel. $\rho_{perc}$ is the percentile corresponding to the threshold $\rho$ It means that the lowest $\rho_{perc}$ of the training data is rejected by applying the corresponding threshold.

| Parameters | | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| $\sigma$ | $\rho_{perc}$ | % | % | % |
| 0.14 | 5 | 65.7 | 12.9 | 98.3 |
| | 10 | 69 | 23.6 | 97.1 |
| | 20 | 72 | 39.6 | 92 |
| | 40 | 72,9 | 68.3 | 75.8 |
| | 50 | 70.6 | 88.6 | 59.5 |

**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% |

Tables 1 summarize performance results for the one-class Parzen density estimator. They show the best results obtained by training classifiers using different parameter configurations on non infected cases only.

Clearly, for the one-class Parzen density estimator, highest sensitivity is reached when $\rho$ corresponds to a rate of training data rejection of 50%; the price to pay for such

a result is that many non infected cases are equally labelled abnormal, thus yielding low specificity. One-class approach lead to significant improvements in sensitivity over classical symmetrical SVMs. Moreover a major limitation of one-class Parzen density estimator is its relatively high computational demand during the testing phase.
In a previous study on the same nosocomial dataset [11], 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 in these previous experiments.

## 6. Conclusion

We analyzed the results of a prevalence study of nosocomial infections in order to detect infected patients. The major hurdle, typical in medical diagnosis, is the problem of rare positives. To address this problem we investigated the applicability of a one-class algorithm proposed by [10]. Experimental results reported in this paper are encouraging. From the point of view of sensitivity, one-class Parzen density estimator attain the highest level (88.6 %) 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. Overall we feel that one-class Parzen density estimator are a promising approach to the detection of nosocomial infections and can become a reliable component of an infection control system.

## References

[1] G. G. French, A. F. Cheng, S. L. Wong, and S. Donnan. Repeated prevalence surveys for monitoring effectiveness of hospital infection control. *Lancet*, 2:1021–23, 1983.
[2] S. Harbarth, C. Ruef, P. Francioli, A. Widmer, D. Pittet, and S.-N. Network. Nosocomial infections in Swiss university hospitals: a multi-centre survey and review of the published experience. *Schweiz Med Wochenschr*, 129:1521–28, 1999.
[3] N. Japkowicz. The class imbalance problem: A systematic study. *Intelligent Data Analysis Journal*, 6(5), 2002.
[4] P. Domingos. A general method for making classifiers costsensitive. In *Proc. 5th International Conference on Knowledge Discovery and Data Mining*, pages 155–164, 1999.
[5] M. Kubat and S. Matwin. Addressing the curse of imbalanced data sets: One-sided sampling. In *Procs˙of the Fourteenth International Conference on Machine Learning*, pages 179–186, 1997.
[6] K. Ali, S. Manganaris, and R. Srikant. Partial classification using association rules. In *Proc. 3rd International Conference on Knowledge Discovery in Databases and Data Mining*, 1997.
[7] B. Silverman. Density Estimation. Chapman and Hall, 1986.
[8] E. Parzen. On estimation of a probability density function and mode. Annals of Mathematical Statistics, 33:1065–1076, 1962.
[9] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2000.
[10] C. M. Bishop. Novelty detection and neural network validation. *Vision, Image and Signal Processing, IEE Proceedings-*, 141(4):217–222, 1994.
[11] G. Cohen, M. Hilario, H. Sax, and S. Hugonnet. Asymmetrical margin approach to surveillance of nosocomial infections using support vector classification. In *Intelligent Data Analysis in Medicine and Pharmacology*, 2003.