TOWARDS AN AUTOMATED NOSOCOMIAL INFECTION CASE REPORTING

Framework to Build a Computer-aided Detection of Nosocomial Infection

Jimison Iavindrasana, Gilles Cohen, Adrien Depeursinge
Medical Informatics Department, University and Hospitals of Geneva, rue Micheli-du-Crest, 24, Geneva, Switzerland

Henning Müeller
Medical Informatics Department, University and Hospitals of Geneva, rue Micheli-du-Crest, 24, Geneva, Switzerland
University of Applied Sciences Western Switzerland, Sierre, Switzerland

Rodolphe Meyer, Antoine Geissbuhler
Medical Informatics Department, University and Hospitals of Geneva, rue Micheli-du-Crest, 24, Geneva, Switzerland

Keywords: Nosocomial infection, Machine learning, Feature selection, Fisher's linear discriminant.

Abstract: The prevalence survey is a valid and realistic surveillance strategy for nosocomial infection surveillance but it is resource and labor-consuming. Querying the hospital data warehouse with a set of relevant features and applying a classification algorithm on the results can reduce the amount of cases to be evaluated by the infection control practitioners. The objective of this work is to provide a framework to build a nosocomial infection model with a set of pre-selected features with Fisher's linear discriminant algorithm. Application of the methodology to two datasets provides promising results. It permits to predict respectively an average of 41.5% and 43.54% positive cases including respectively 65.37% and 82.56% true positive cases. The proposed framework can be applied to other classification algorithms, which are planned as future work.

1 INTRODUCTION

1.1 Context

Nosocomial infections (NI) are infections acquired in a hospital. In Switzerland, 70000 hospitalized patients per year are infected and 2000 deaths per year are caused by NI. A hospital aware of the quality of the patient care should have an infection prevention, control and surveillance program. The surveillance is the process of detecting these infections. Prevalence surveys are recognized as valid and realistic approaches of nosocomial infection (NI) surveillance strategies (French et al, 1983). Prevalence of NI is presented as prevalence of infected patients, defined as the number of infected patients divided by the total number of patients hospitalized at the time of study, and prevalence of infections, defined as the number of NIs divided by the total number of patients hospitalized at the time of study (Sax et al, 2002). However, a prevalence survey is resource and labor-consuming, as it requires assembling a wide range of data gathered from multiple sources. The medical record of all patients hospitalized for more than 48 hours at the time of the survey are reviewed by infection control practitioners. During this first process they extract information related to each patient and store them in a database specially developed for this purpose. The prevalence database contains 83 attributes ranging from administrative information, demographic characteristics, admission diagnoses, comorbidities and severity of illness scores, type of admission, and exposure to various risks of infection, clinical and paraclinical information, and data related to infection when present. This database is analyzed to sort out the prevalence official report.

The hospital data warehouse contains all the data in the operational system except the data of the day. Querying the hospital data warehouse and apply data mining techniques in order to report "potential cases" to be reviewed by the infection control

Iavindrasana J., Cohen G., Depeursinge A., Müeller H., Meyer R. and Geissbuhler A. (2009).
TOWARDS AN AUTOMATED NOSOCOMIAL INFECTION CASE REPORTING - Framework to Build a Computer-aided Detection of Nosocomial Infection .
In *Proceedings of the International Conference on Health Informatics*, pages 317-322
DOI: 10.5220/0001553103170322

practitioners will reduce their workload and will allow them to focus on the content of the patient record and evaluate the presence of NI.

Some of the 83 attributes of the prevalence database are acquired for administrative purposes only. The majority of these attributes are the synthesis of information from the patient records emanating from laboratory, radiology, nursing, and clinical databases. A more realistic approach to report potential cases is to find a subset of N most relevant features (N<83) and query the hospital databases on the basis of these N features. The results obtained are then classified and ranked with a classification algorithm, and only predicted positive cases i.e. lists of infected patients are reviewed by the infection control practitioners. The classification is based on a model build with the N features.

We present in this paper a framework to build a computer-aided detection of NI based on a set of N pre-selected features using Fisher's linear discriminant algorithm. For this purpose, we analyze a previous prevalence database to optimize the classification process. The retrieval of the data from the hospital data warehouse is not presented in this paper. The Fisher's linear discriminant (FLD) was chosen for its "simplicity" as it only has one parameter to optimize. One challenging characteristic of NI prevalence data is the imbalance between the positive and negative cases (respectively 11% and 89%) (Cohen et al, 2006). This important characteristic is taken into account in the proposed methodology. An application is developed to automate all the process.

1.2 NI Prevalence Data

The prevalence data we analyzed in this work is data collected at the hospitals during the 2006 survey. The dataset contains 5 data categories: 1) demographic information, 2) admission diagnosis (classified according to McCabe5 (McCabe and Jackson, 1962) and the Charlson index (Charlson et al, 1987) classifications); 3) patient information at the study date (ward type and name, status of Methicillin-Resistant Staphylococcus Aureus portage, etc); 4) information at the study date and the 6 days before (clinical data, central venous catheter carriage, workload, infection status, etc) and 5) those related to the infections i.e. for infected patients (infection type, clinical data, etc.).

In this study, we are interested in the 4 first categories of data as they are related to patient infection, which comprises 45 attributes. Most of these data are categorical except for date information (year of birth, admission date and study date) and the workload value. The dataset contains 1573 cases. The year of birth was converted into age and discretized into 3 categories (0-60; 60-75; >75) as in (Sax et al, 2002), and a new variable "hospitalization duration" was created. A Mann-Whitney-Wilcoxon statistical test on the workload value provides a significant difference between infected and non-infected patients. As it is the unique attribute having missing values (91 cases including 2 positive cases), all cases having no workload value were removed. The latter and the hospitalization duration were discretized afterwards using the minimum description length principle (Kononenko, 1995). Patients admitted for less than 48 hours at the time of the study and not transferred from another hospital were also removed. The final dataset contains 1384 cases containing 166 positive cases (11.99%). Let us denote this dataset S.

The ratio of positive cases in the dataset S is very low compared to the negative ones. The class imbalance is an important issue in machine learning since the class of interest is represented with a small number of examples (Japkowicz and Stephen, 2002). In the presence of imbalanced datasets, classification algorithms tend to classify the larger class accurately while generating more errors in the minority class. If a positive class has a ratio of 10%, a classification accuracy of 90% may be meaningless if the classification is not sensitive at all.

The class imbalance problem induces specific approaches to train classifiers and evaluate their performance. Two approaches were proposed to deal with the class imbalance problem in (Cohen et al, 2006, Estabrooks, 2004). The first one is to modify the classification algorithm or at least use an algorithm able to deal with imbalanced data. The second resamples the data to reduce the imbalance effect. The latter has the advantage of being independent of any classification algorithm.

1.3 Fisher's Linear Discriminant

The basic idea behind linear discriminant algorithms is to find a linear function providing the best separation of instances from 2 classes. Fisher's linear discriminant is looking for a hyperplane directed by w, which (i) maximizes the distance between the mean of the classes when projected on the line directed by w and (ii) minimizes the variance around these means (Fisher, 1936). An illustration of this algorithm is highlighted on the figure below (Figure 1).

Formally, Fisher's linear discriminant aims at maximizing the function:

$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (1)$$

where S_B is the scatter matrix between classes and S_W the scatter matrix within classes. This equation permits to formulate Fisher's linear discriminant as an algorithm aiming at minimizing the variance within the classes and maximizing the variance between classes. An unknown case will be classified into the nearest class centroid when projected onto a hyperplane directed by w.

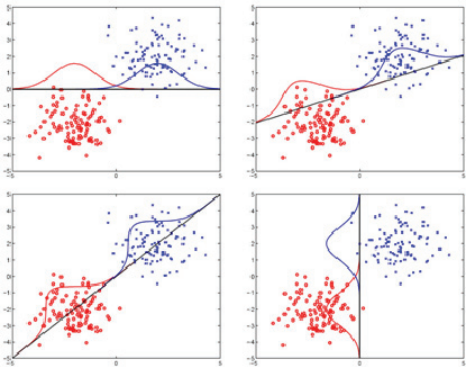

Figure 1: Illustration of Fisher's linear discriminant. The algorithm is looking for the direction providing the best separation of the classes when projected upon. In this figure, the third image (bottom left) provides the best separation of the datasets.

In a classification task, an object is member of exactly one class and an error occurs if the object is classified into the wrong one. The objective is then to minimize the misclassification rate. With Fisher's linear discriminant algorithm, the scatter matrix within classes SW is evaluated on the training datasets. To minimize the misclassification rate on unseen test sets (generalization error), a regularization factor r ($0 \leq r \leq 1$) is introduced into the computation of S_W (Hastie et al, 2001). The regularization factor r has to be optimized to minimize the misclassification error.

2 MATERIAL AND METHODS

The attributes from the dataset S were ranked according to the information gain. Afterwards, a Chi-square statistic test was applied to filter the discriminative features to be retained for an evaluation with the classification algorithm. Let us denote S1 the new dataset created with this feature selection. The dataset S1 may contain attributes which are not always documented or at least not documented in a machine readable format in the clinical database. We will remove them from the dataset S1 to obtain a second dataset S2.

We have taken the two datasets S1 and S2 described above to evaluate the discriminative power of the selected features. For classification purposes, we use the open-source toolbox MATLABArsenal. This MATLAB package contains many classification algorithms and in particular the regularized Fisher linear discriminant algorithm as described above. The MATLAB software is invoked from the java application developed for the process automation. This application also uses the WEKA api for other routine tasks such as the training/testing set splitting.

The evaluation of the predictive power of the selected features is inspired by the experimental setup described in (Rätsch et al, 2001). One hundred (100) partitions of training and testing sets were generated with the data source S1 and S2 having respectively a ratio of 60% and 40%. The original data distribution is kept in both partitions. A grid search algorithm is then applied to the first five under-sampled training sets using a 5-folds cross-validation to find the best parameters of the classification algorithm. In the five under-sampled training sets, the classes are equally distributed (50% positive cases and 50% negative cases).
The regularization factor r takes 41 values from 2-20 to 220 during this process. The best parameter of each training set was the one providing the highest recall (i.e. the parameter permitting to predict highest rate of true positive cases). The best value selected for the classification algorithm is the median of the 5 best parameters. The 100 training sets (having the original class distribution) are then used to train Fisher's linear discriminant models with this best parameter. This process allows us to build 100 models and to validate each of them on the corresponding testing set. The general performance of the classifier is computed as the mean of the 100 classification performances on the test sets. The performance of the classification algorithm with the 2 datasets (S1 and S2) is also compared with respect to the Mann-Whitney-Wilcoxon statistical test.

Table 1: List and rank of features obtained with information gain followed by a Chi-square filtering. The first column provides the rank of each attributes.

Rank	Selected attributes
1	Antibiotic therapy
2	Fever
3	Mechanical ventilation
4	Urinary tract
5	Workload value > 91.5
6	Workload value <=45.5
7	Stay at the intensive care unit during hospitalization
8	Central vein catheter
9	Hospitalization duration up to 7.5 days
10	Intensive care unit ward
11	Obstetrical ward
12	Surgery
13	McCabe score fatal < 6 months
14	No MRSA colonization
15	Actual MRSA colonization
16	McCabe score non fatal
17	Workload value between 45.5 and 91.5
18	Diabetes with organ affected
19	Transfer from another hospital as admission
20	Congestive cardiomiopathy

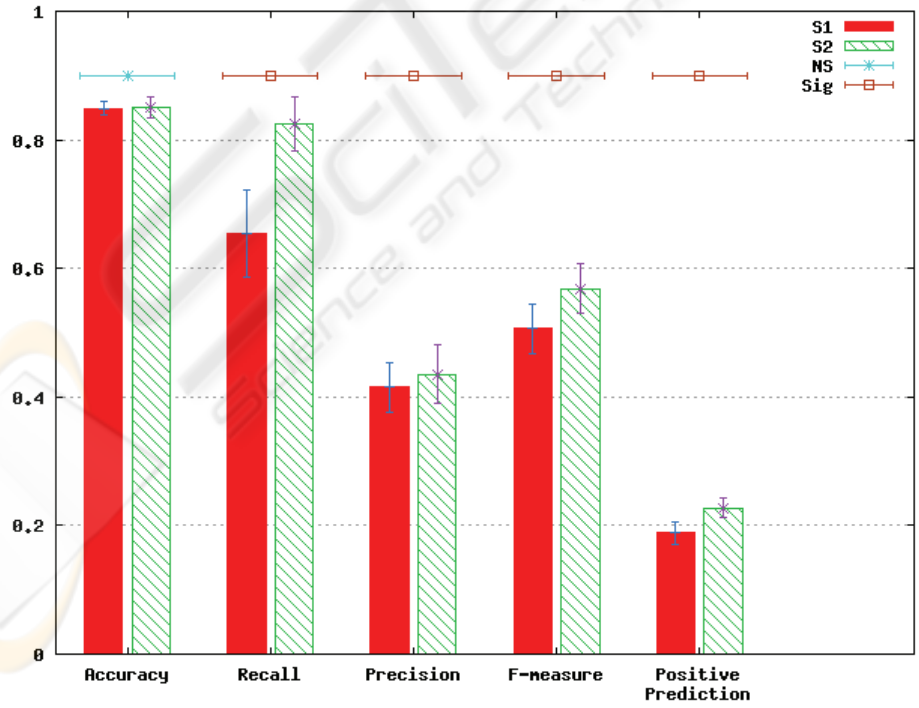

Figure 2: The mean of each performance measure on the datasets *S1* and *S2*.

3 RESULTS

3.1 Feature Selection

Twenty (20) attributes were retained from the feature selection process as summarized in the table 1. Let S1 denote the name of this subset. Two (2) clinical attributes are not always documented or at least not documented in a machine readable format in the clinical database: fever and workload value. These attributes were removed to create a second data source denoted S2 even if the information gain ranked these attributes at the fourth and sixth position.

3.2 Model Selection

The grid search algorithm applied on the two datasets S1 and S2 returned respectively r =0.5 and 1 as the best parameter. The figure 2 summarizes the performance metrics (accuracy, precision, recall, f-measure and the ratio of positive predictions) obtained with the 2 datasets in terms of their mean.

3.3 Classification Performances

Dataset S1 and S2 permit to obtain respectively a mean recall (±standard deviation) of 65.37% (±6.76) and 82.56 (±4.22), a precision (±SD) of 41.50% (±3.9) and 43.54% (±4.59), a f-measure (±SD) of 50.58(±3.83) and 56.87(±4.29) over the 100 training/testing split realizations. The mean accuracy (±SD) for S1 and S2 are 84.83%(±1.04) and 85.04%(±1.65) and the positive prediction ratios are respectively 18.82% (±1.72) and 22.73% (±1.55).

According to the results above, querying the hospital data warehouse with the features present in the dataset S1 and S2 and classify the results with the FLD algorithm, we can expect retrieving an average of (±SD) 65.37% (±6.76) and 82.56 (±4.22) of the infected patients. The mean numbers of potential cases (±SD) to be submitted to the ICP are respectively 18.82% (±1.72) and 22.73% (±1.55) of the hospitalized patients.

A Mann-Whitney-Wilcoxon statistic test provided a p value < 0.001 for accuracy, precision, f-measures and the positive prediction ratio. According to this test, there is a statistically significant difference between the accuracy, precision, f-measure and the ratio of positive prediction. The removal of the temperature and the workload features improved significantly the performance of the FLD.

4 DISCUSSION AND CONCLUSIONS

In this paper we present a framework to build a NI model based on a small number of clinical features permitting to report NI cases to be reviewed by infection control practitioners. Fisher's linear discriminant was chosen as detection algorithm. The removal of the attributes characterizing fever and workload value is not affecting the sensitivity of the classifier. This may be explained by the strong correlation between these attributes with some important features such as the antibiotherapy for the fever and surgery, stay at the intensive care unit during the hospitalization, a presence of artificial ventilation, urinary tract, and central venous catheter for the workvalue. The automation of the process needs integration of data from laboratory, radiology, nursing, and clinical databases.

Limits of this Work. The evaluation of the discriminative power of the selected features was carried out using Fisher's linear discriminant algorithm. A comparison with other classification algorithms such as Support Vector Machines (SVM) and the Kernel Fisher's linear discriminant could be a good option to improve the classification performance. The framework could also be extended with an evaluation of the best classifiers. The grid search algorithm for optimal parameters has high computational cost especially for classification algorithms with more than one parameter to optimize such as SVMs of the Kernel Fisher discriminant. A gradient descent method can be used to find the best parameter and can improve the generalization performance as described in (Chapelle et al, 2002).

Future Work. The framework introduced in this paper permits to evaluate the discriminative power of a subset of important features from the NI database. The feature selection method we have chosen in this work is based on the information gain combined with a Chi-square statistic test. More experiments with other feature selection techniques are required (Guyon, 2003). The discriminative power of the selected features will be evaluated with more than one classification algorithm. The result of these evaluations i.e. the minimal attributes required to predict most of the positive NI cases will be retained to build queries for the hospital databases in order to automatically report potential cases for the prevalence surveys. This automated nosocomial

infection reporting will permit to conduct more prevalence surveys with less cost.

ACKNOWLEDGEMENTS

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

REFERENCES

Chapelle, O, Vapnik, V, Bousquet, O, Mukherjee, S, 2002. Choosing multiple parameters for support vector machines. Mach. Learning.

Charlson, ME, Pompei, P, Ales, KL, MacKenzie, CR, 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis.

Cohen, G, Hilario, M, Sax, H, Hugonnet, S, Geissbuhler, A, 2006. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine.

Estabrooks, A, 2004. A multiple resampling method for learning from imbalanced datasets. Comput Intell.

Fisher, RA, 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics.

French, GG, Cheng, AF, Wong, SL, Donnan, S, 1983. Repeated prevalence surveys for monitoring effectiveness of hospital infection control. Lancet.

Guyon, I, Elisseeff, A, 2003. An Introduction to variable and feature selection. Mach. Learning Res J.

Hastie, T, Tibshirani, R, Friedman, J, 2001. The elements of statistical learning: data mining, inference, and prediction. Springer.

Japkowicz, N, Stephen, S, 2002. The class imbalance problem: a systematic study. Intell Data Anal J.

Kononenko, I, 1995. On biases in estimating multi-valued attributes. Eds.: Morgan Kaufmann. In Proceedings of the 14th International Joint Conference on Artificial Intelligence.

McCabe, WR, Jackson, GG, 1962. Gram-negative bacteremia, I: etiology and ecology. Arch Intern Med.

Rätsch, G, Onoda, T, Müller, KR, 2001. Soft margin for AdaBoost. Mach.Learning.

Sax, H, Pittet, D, 2002. Swiss-NOSO Network. Interhospital Differences in nosocomial infection rates: importance of case-mix adjustment. Arch Intern Med.