Minimal Set of Attributes Required to Report Hospital-Acquired Infection Cases

Jimison Iavindrasana, Gilles Cohen, Adrien Depeursinge, Rodolphe Meyer, Antoine Geissbuhler
Department of Medical Informatics
University and Hospitals of Geneva
Rue Micheli-du-Crest 24, CH-1211 Geneva, Switzerland
{jimison.iavindrasana, gilles.cohen,adrien.depeursinge,rodolphe.meyer,antoine.geissbuhler}@sim.hcuge.ch

Abstract

Data collection for hospital-acquired infection prevalence study is resource consuming. Data mining techniques can be applied to data extracted from the hospital data warehouse in order to report potential cases to be reviewed by infection control practitioners (ICP). The objective of this paper is to investigate the minimal set of attributes required for an automated cases reporting. Information gain and SVM recursive feature elimination combined with a chi-square filtering were used to select the most important features in the prevalence database of the 2006 survey. The temperature and workload were included within the 20 most important features. These attributes are not well documented and removed from the list of important features. The results obtained with the resulting dataset were acceptable because the ICP will have to analyze the electronic health record of only 22.73% of hospitalized patients.

1 Introduction

Hospital-acquired infections or nosocomial infections (NI) are those infections acquired in a hospital, independently of the reason of the patient admission. NI appears after 48 hours after the patient admission. These infections may be related to medical procedures such as the implantation of infected urinary tracts or simply occur during the hospitalization where the micro-organisms are transmitted from other patients, medical staff or are a consequence of the hospital environment contamination.

In Switzerland, 70 000 hospitalized patients per year are infected and 2 000 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 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]. From these formulas, the prevalence survey is resource and labor consuming: the electronic health record (EHR) of all patients admitted for more than 48 hours the day of survey should be analyzed by infection control practitioners (ICP). If necessary, additional information is obtained by interviews with nurses or physicians in charge of the patient.

The University Hospitals of Geneva has been performing yearly prevalence survey since 1994. The prevalence database contains 83 attributes ranging from administrative information, demographic characteristics, admission diagnoses, comorbidities and severity of illness scores, type of admission, exposure to various risks of infection, clinical and paraclinical information, and data related to infection when present. One of the main characteristic of the prevalence data is the nature of the attribute values: most of them are nominal. An attribute value summarizes the presence or absence of a particular risk factor or a sign and symptom of an infection like central venous catheter or an antibiotic treatment that can be found in the EHR of the patient. Only the year of birth and the workload values are numerical. Another important characteristic of the prevalence data is the imbalance between the positive and negatives cases: there are around 10% of positive cases.

IT can bring a valuable support for NI surveillance. The hospital data warehouse contains all the data in the hospital operational system except the data of the day. A complex data processing may be implemented to extract all necessary information from the data warehouse, analyze and summarize the information and populate the prevalence database. This approach is too ambitious because of the nature of data to be analyzed: some of them may be found in free text or not well documented in the EHR and require an intervention of a specialist. The most realistic approach is to query the hospital data warehouse in order to populate the N most important item of the prevalence database (N<<83 where 83 is the numer of attributes in the prevalence database) and apply data mining techniques to report "potential cases" to be reviewed by the ICP. The potential cases are those predicted as positive cases by a classification algorithm. The main advantage of this approach is the reduction of their workload, and will allow them to evaluate the presence of NI on subset of patients; they can have much more time to analyze the content of the patient record and

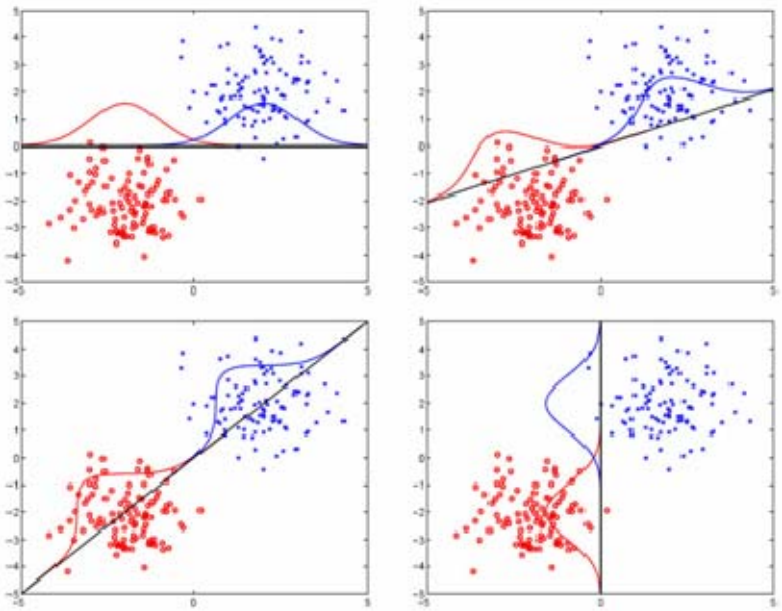

Figure 1: Illustration of Fisher's linear discriminant. The algorithm searches for the direction providing the best separation of the classes when projected upon. In this figure, the third image (bottom left) provides the best separation of the datasets.

may found new risk factors. Indeed, the most consuming time during a NI prevalence survey is the data collection, which represents approximately 800 hours per year [Cohen et al, 2004].

Various data mining techniques were applied at the University Hospitals of Geneva to support NI prevalence survey since 2002. We can quote among others the use of different form of the SVM algorithm optimization including asymmetrical margin approach [Cohen et al, 2003], one-class SVM [Cohen et al, 2004], a comparison of SVM with other classification algorithms [Cohen et al, 2006]. In this paper, we are analyzing the minimal set of features necessary to report potential cases to be reviewed by ICP. We will remove important features if they are not well documented in the EHR. The differences with the previous works lie in the objective of the experimentation, the methodology and the dataset used.

2 Background

2.1 Nosocomial infection prevalence data:

In this work, we performed a retrospective analysis of prevalence data collected at the University Hospitals of Geneva (6 hospitals and 2'200 beds) during the 2006 NI prevalence survey. The dataset contains five data categories: 1) demographic information, 2) admission diagnosis (classified according to McCabe [McCabe and Jackson, 1962] and the Charlson index classifications [Charlson et al, 1987]); 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 six 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 four first categories of data as they are related to patient infection, which comprises 45 attributes. The dataset contains 1573 cases.

To homogenize the data values, we transformed all numerical data into nominal ones. 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 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%). And finally, all attributes were binarized. Let us call this dataset S.

2.3 Class imbalance problem

The class imbalance problem is an important problem 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, 2004; 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.

2.4 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 (FLD) 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 property is highlighted on the figure below (Figure 1).

Formally, FLD aims at maximizing the function:

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

where S_B is the scatter matrix between classes and S_W the scatter matrix within classes. This equation permits to formulate FLD 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*.

In a classification task, an object is a 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 FLD algorithm, the scatter matrix within classes S_W 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.

3 Material and methods

Two feature selection algorithms were used independently. The first based on the information gain (IG) of each attribute [Quinlan, 1986] and the second based on the combination of attributes using SVM Recursive Feature Elimination (RFE) [Guyon, 2002]. These 2 algorithms return all the features ranked by order of importance. To filter the most important ones, a chi-square statistic test was performed to filter the discriminative features to be retained for an evaluation with a classification algorithm. These feature selection algorithms were applied to 100 training sets build from the original dataset S. The significant attributes retained by both feature selection algorithms over the 100 training datasets were retained to build a second dataset S1. Afterwards, we removed the important features in S1 which are not well documented in the EHR to obtain a third dataset S2. We then evaluated the performance of the FLD algorithm on these two datasets. For classification purposes, we used the open-source toolbox MATLABArsenal[1]. This MATLAB package contains many classification algorithms and in particular the regularized FLD algorithm as described above. FLD was chosen as it has only one parameter so easier to optimize.

The methodology adopted to evaluate the performance of the FLD 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% (approximately 830 cases) and 40% (approximately 554 cases). The original data distribution is kept in both partitions i.e. 11.99% of positive cases. Five balanced dataset (50% of positive cases and 50% of negative cases) were created from the first five training sets. A grid search algorithm is then applied to these down-sampled datasets using a 5-folds cross-validation to find the best parameters of the classification algorithm. The regularization factor *r* takes values from 2^{-20} to 1 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) and the highest precision. The best value selected for the classification algorithm is the median of the 5 best parameters obtained with the five down-sampled data. The 100 training sets (having the original class distribution) are then used to train the FLD models with this best parameter. This process allowed 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.

4 Results

4.1 Feature selection

Twenty (20) attributes were retained from the two feature selection algorithms; IG and SVM RFE returned the same features after the Chi-square filtering. The hospitalization duration up to 7.5 days, was retained as a discriminative attribute. Two admission diagnoses are discriminative: those classified as "non fatal" and "fatal in less than 6 months" according to the McCabe classification, transfer as admission, "congestive cardiomyopathy" and "diabet with organ affected" as comorbidities. In the third data category, the intensive care unit and obstetrical wards, absence or actual MRSA colonization are the most discriminative attributes. In the fourth data category, an antibiotic treatment, fever, a surgery, a stay at the intensive care unit during the hospitalization, a presence of artificial ventilation, urinary tract, central venous catheter and the 3 categories of workload value were significantly discriminative.

[1] http://www.informedia.cs.cmu.edu/yanrong/MATLABArsenal/MATLABArsenal.htm (last accessed October 2008)

Table1: List and rank of features obtained after applying IG and SVM RFE followed by a Chi-square filtering. The first column provides the rank of each attributes. The two algorithms provided the same features but not with the same rank.

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

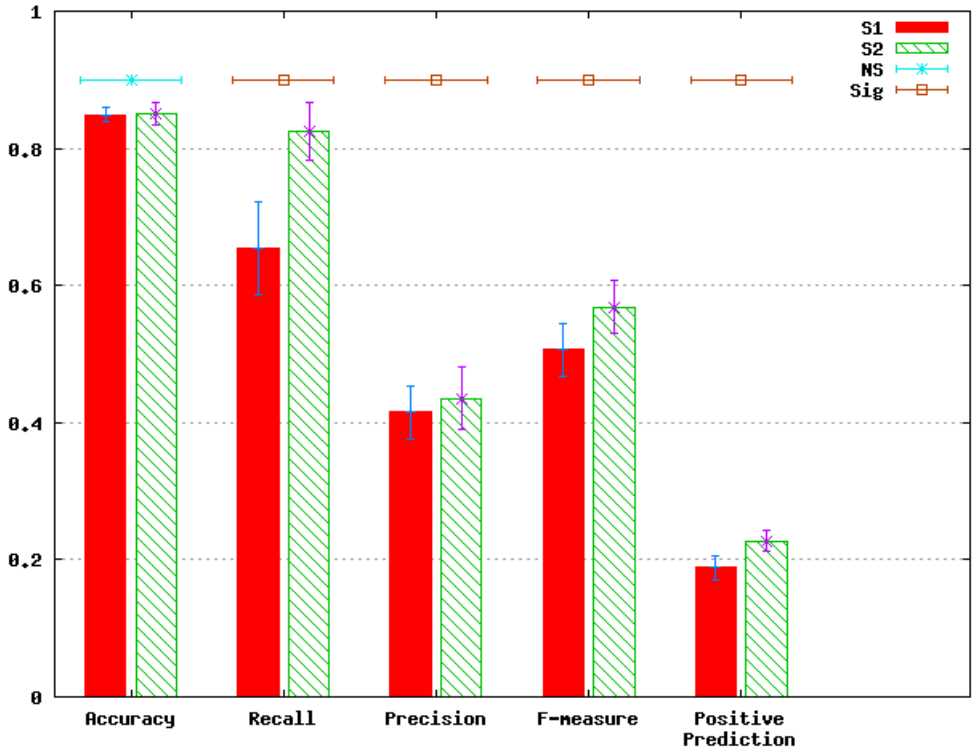

Figure 2. The mean and standard deviation of each performance metrics on the datasets S1 and S2. Sig (respectively NS) indicates (no) significant difference between the performance measure on the two datasets.

Table 1 summarizes the features returned by both IG and SVM RFE. The feature selections described above provided two clinical features, which 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 the second dataset S2.

4.2 FLD performances

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

Dataset S1 and S2 permit to obtain respectively a mean recall (±SD) 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 data 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, if we query 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 the performance of the FLD.

5 Discussion and conclusion

In this paper we investigated the minimal set of features necessary to report potential cases to be reviewed by infection practitioners. IG and SVM RFE were used to select these features and Fisher's linear discriminant was chosen as classification algorithm. The removal of the attributes characterizing fever and workload value significantly improved the performance of the classifier. This result may surprise as these attributes was retained by the IG and SVM RFE as important features to predict a NI. However, this phenomenon is not new in statistical and machine learning domain. This phenomenon is called redundancy or negative interaction [Kludas et al, 2008]. This redundancy cannot be evaluated with IG as it only evaluates quantity of information brought by each attribute to the value of the class. It was not also handled by the SVM RFE because it eliminates X features per iteration (1 in our case) and do not try all possible combinations of features.

The results we obtained with the S2 dataset are acceptable. The ICP will only review the EHR of 22.73% of hospitalized patients. The precision value indicates that 43.54% of these patients are infected and they represent 82.56% of all infected patients. The precision rate is satisfactory because the data of non-infected patients are necessary for statistical tests in order to identify the most important risk factors. The ICP will also have enough time to identify new risk factors from the infected patients' EHR and propose new preventive measures for the future.

5.1 Limits of this work

The evaluation of the discriminative power of the selected features was carried out using Fisher's linear discriminant algorithm because of its simplicity: one parameter (0<r<=1) to optimize during the grid search process. A comparison with other classification algorithms such as Support Vector Machines (SVM) and the Kernel FLD has to be carried out. The grid search algorithm for algorithm parameters optimization has high computational cost especially for classification algorithms with more than one parameter to optimize such as SVM or the Kernel FLD. A gradient descent method could converge more rapidly to the best parameter and can improve the generalization performance as described in [Chapelle et al, 2002].

5.2 Future work

The results obtained were promising and in the future, we plan to evaluate the discriminative power of the selected features 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 than the usual method for conducting prevalence studies.

Acknowledgments

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

References

[Chapelle et al, 2002] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing multiple parameters for support vector machines. Mach. Learning. 2002;46(1-3):131-59.

[Charlson et al, 1987] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40:373-83.

[Cohen et al, 2003] Cohen G, Hilario M, Hugonnet S, Sax H. Asymmetrical Margin Approach to Surveillance of Nosocomial Infections Using Support Vector Classification. IDAMAP; 2003.

[Cohen et al, 2004] Cohen G, Hilario M, Sax H, Hugonnet S, Pellegrini C, Geissbuhler A. An application of one-class support vector machine to nosocomial infection detection. In Proceedings of MedInfo:2004;11:716-20.

[Cohen et al, 2006] Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artificial Intelligence in Medicine. 2006;37(1):7-18.

[Estabrooks, 2004] Estabrooks A. A multiple resampling method for learning from imbalanced datasets. Comput Intell. 2004,20(1):18-36.

[Fisher, 1936] Fisher RA. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 1936;7:179–88.

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

[Guyon et al, 2002] Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning. 2002;46:389-422.

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

[Japkowicz and Stephen, 2002] Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal J 2002;6(5):429-49.

[Kludas et al, 2008] Kludas J., Bruno E., and Marchand-Maillet S. Can Feature Information Interaction help for Information Fusion in Multimedia Problems? In Proceedings of MMIU:2008.

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

[McCabe and Jackson, 1962] McCabe WR, Jackson GG. Gram-negative bacteremia, I: etiology and ecology. Arch Intern Med. 1962;110:847-55.

[Quinlan, 1986] Quinlan JR. Induction of decision trees. Machine Learning, 1, 81-106, 1986.

[Rätsch et al, 2001] Rätsch G, Onoda T, Müller KR. Soft margin for AdaBoost. Mach.Learning. 2001;42(3):287-320.

[Sax et al, 2002] Sax H, Pittet D, Swiss-NOSO Network. Interhospital Differences in nosocomial infection rates: importance of case-mix adjustment. Arch Intern Med. 2002;162(21):2437-42.