MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010
© 2010 IMIA and SAHIA. All rights reserved.
doi:10.3233/978-1-60750-588-4-764

Measuring the effectiveness of hospital-acquired infection prevention

Jimison Iavindrasana[a], Gilles Cohen[a], Adrien Depeursinge[a], Henning Müller[a,b], Rodolphe Meyer[a], Hugo Sax[c], Antoine Geissbuhler[a]

[a] *Medical Informatics Department, University and Hospitals of Geneva, Switzerland*
[b] *University of Applied Sciences Western Switzerland, Sierre, Switzerland*
[c] *Infection Control Program, University and Hospitals of Geneva, Switzerland*

Abstract

This article deals with data on nosocomial infections acquired in the Geneva University Hospitals. Goal of the work is to derive a model from a hospital-acquired infection (HAI) prevalence survey of year Y and apply them to a prevalence survey of years Y+1, Y+2. This analysis permits to evaluate the effectiveness of preventive measures taken after the prevalence survey in year Y. It also analyzes the robustness of the SVM algorithm on time-variable attributes. The model build on the dataset of year Y gives better results than in a previous study. The application of the model on the Y+1 and Y+2 prevalence surveys shows simultaneously improvements and deteriorations of 5 performance measures. This highlights the effectiveness of prevention and reduces the risk of HAI after the prevalence survey of year Y. We introduce a new method to detect redundancy in a dataset with the SVM algorithm.

Keywords:

Hospital infections, Program evaluations, Machine learning.

Introduction

A hospital is a facility providing medical care to sick people. Independent of the reason for admission, a patient may acquire new infections inside a hospital due to the presence of microorganisms. These hospital-acquired infections (HAI) usually appear 48 hours after the patient admission. The infectious agents can be transmitted from other patients or by health care workers during medical procedures. In Switzerland, 70’000 hospitalized patients per year are infected and 2’000 deaths per year are caused by HAI. Many prevention and surveillance programs are carried out to prevent and/or reduce the risk of HAI. A prevention program, for example, includes hygiene measures permitting to isolate or eliminate infectious agents such as washing hand before any contact with patients, the use of gloves, use of masks, disinfection, sterilization, etc. A surveillance program aims at detecting infections. *French et al.* proved that a repeated prevalence surveys is a valid and realistic approach for infection control and surveillance [1]. The prevalence of infection is the number of infected patients divided by the total number of hospitalized patients at the time of the study [2]. The infection prevalence rate can also be used as an indicator of the quality of patient care.

However, the HAIs are not always documented in the electronic health record (EHR) of the patients and the infection control practitioners have to carry out a survey to obtain the prevalence rate. For this purpose, the EHR of all hospitalized patients admitted for more than 48 hours are analyzed. If necessary, additional information is obtained by interviews with nurses or physicians in charge of the patient. This survey is performed during one to three months on a yearly basis at the University and Hospitals of Geneva since 1994. This survey is labor intensive and it cannot be carried out all year long.

In a previous study, we extracted the most important features of a HAI database allowing the prediction of an HAI infection [3]. Fisher’s linear discriminant was used to evaluate the predictive power of these features and it provided good results. However, maximum margin classifiers such as support vector machine (SVM) are more appealing with respect to generalization performance from a theoretical viewpoint. A maximum margin classifier looks for an optimal hyperplane separating the training dataset so that the distance of training points to the optimal hyperplane is maximized. This supposes that the training data are separable. Finding the optimal hyperplane is equivalent to resolving the following optimization problem:

$$\min\left(w^T w\right) \textit{subject to } y_i\left(w^T x_i + b\right) \geq 1 \qquad (1)$$

In the relation (1), w is a vector perpendicular to the hyperplane, b is a scalar value, $\left\{x_i \in \mathbf{R}^d, y_i \in \{-1,1\}\right\}_{i=1}^N$ are the training points, *N* the number of examples and *d* the number of variables. If certain conditions hold and using the Lagrangian formulation, the previous problem is equivalent to its dual (2), which is a quadratic optimization problem and which can be solved using several techniques.

A SVM is a maximum margin classifier using only points on both sides of the margin or support vectors (points x_i for which the Lagrangian multipliers $\alpha_i > 0$) to build a model.

$$\max_{\alpha\geq 0} \sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j x_i^T x_j \quad st \sum_i \alpha_i y_i = 0 \quad (2)$$

For non-separable training datasets, penalty variables ξ_i are introduced to soften the constraints of the maximum margin formulation (1). The penalty variables are drawn as follows: $0 < \xi_i \leq 1$ if the points are on the correct side of the hyperplane and $\xi > 1$ if the point is on the wrong side. A cost variable *C* is also introduced to control the trade-off between the width of the margin and the points within the margin. The goal of the SVM classifier is then to maximize the margin while minimizing the total sum of the penalties and thus the equation (1) becomes:

$$\min\left(w^T w\right) + C\sum_i \xi_i^p \;\; st\; y_i\left(w^T x_i + b\right) \geq 1 - \xi_i, \xi_i \geq 0 \qquad (3)$$

However, the dual problem is often solved because the duality theory provides a convenient way to deal with constraints. The dual optimization problem can also be written in terms of dot products permitting the use of the kernel functions. The kernel trick allows applying the maximum margin algorithm to a transformed version of a non-separable dataset (feature space) via a mapping function ϕ. The related dual problem can be expressed as:

$$\max_{\alpha}\; 2\alpha^T e - \alpha^T\left(G\left(K\right) + \frac{1}{C} I_n\right) \quad st\, \alpha \geq 0, \alpha^T y = 0 \qquad (4)$$

In the previous relation, *e* is the n-vector of ones, $\alpha \in \mathbf{R}^N$, *G(K)* the Gram matrix and defined by $G_{ij}\left(K\right) = \left[K\right]_{ij} y_i y_j = k\left(x_i, x_j\right) y_i y_j$, I_n is a diagonal matrix of 1 and $\alpha \geq 0$ means $\alpha_i \geq 0$ for all i=1,…,n. The transformation function ϕ is integrated in the definition of the kernel matrix *K*. One kind of such kernel is the Gaussian kernel or RBF kernel expressed as $K\left(x_i, x_j\right) = \phi(x_i)^T \phi(x_j) = e^{-\left\|x_i - x_j\right\|^2 / 2\sigma^2}$. For such a kernel, the misclassification cost *C* and the kernel parameter σ need to be optimized.

Many researchers consider SVM as one of the best classification algorithm due to its theoretical foundation based on structural risk minimization implying a better generalization performance [4]. However, SVM can provide bad results used with wrong parameters. The usual way to find the parameters of SVM is to scan a range of possible values of the parameters, evaluate the classifier with a data splitting methods such as cross-validation or bootstraping and then select those providing the best performance. A better method is to evaluate the SVM with the leave-one-out procedure during grid search. This process is expensive with respect to computation time and a more efficient way to choose the SVM parameters is to take advantage of the underlying theory especially the bound of the leave-one-out error.

For the SVM with an RBF kernel and in the case of non-separable training data, *Vapnik* showed that the leave-one-out error is upper bounded by $4R^2\left\|w\right\|^2$ (the radius margin bound) [4]. *R* is the radius of the smallest sphere containing all $\phi(x_i)$ and is a solution to the following optimization problem:

$$\max_{\beta}\; 1 - \beta^T K \beta \quad st\, \beta_i \geq 0, e^T\beta = 1$$

This bound of the leave-one-out error can be used to estimate the parameter σ of the RBF kernel and the soft margin parameter *C*. The reader is referred to [5] for a survey of SVM error bound estimation. To obtain the radius margin bound for non-separable training data, we perform the following change: $\tilde{w} = \begin{bmatrix} w \\ \sqrt{C}\xi \end{bmatrix}$ and $\tilde{\phi}(x_i) = \begin{bmatrix} \phi(x_i) \\ y_i e_i \big/ \sqrt{C} \end{bmatrix}$ as the i-th training data. The kernel function becomes:

$\tilde{K}(x_i, x_j) = K(x_i, x_j) + \delta_{ij}/C$, where δ_{ij} is the Kronecker symbol. The new radius margin bound is $\tilde{R}^2\left\|\tilde{w}\right\|^2$ where $\tilde{R}^2$ is the objective value of:

$$\max_{\beta}\; 1 + \frac{1}{C} - \beta^T\left(K + \frac{1}{C}\right)\beta \quad st\, \beta \geq 0, e^T\beta = 1 \qquad (4)$$

The relation (4) can be solved using optimization techniques such as the gradient descent algorithms [6]. The use of the radius margin bound to estimate the parameters of the SVM is attractive thanks to the speed of its resolution.

Even if preventive measures are taken along the year to reduce and/or prevent the risk of HAI, many signs and symptoms and/or risk factors remain reliable to diagnose an infection (e.g. antibiotics treatment, fever, use of devices such as catheter or urinary tract, etc.). The goal of this paper is to evaluate the robustness of the SVM with respect to generalization performance i.e. its capacity to predict future unseen prevalence surveys. For this purpose, we predict the presence of an HAI on patient enrolled in the 2007 and 2008 prevalence survey from a model build on the 2006 data. This evaluation can also provide an insight of the effectiveness of the preventive measures taken between two prevalence surveys. A longer-term objective is to build an automated prevalence survey tool using information within the hospital data warehouse.

Materials and Methods

Datasets and software

As introduced in the previous section, we use a 2006 prevalence survey to build the HAI model and the 2007 and 2008 prevalence surveys for evaluation. We use three versions of these datasets as did in a previous study [3]. The first dataset, called S, contains all features from the prevalence database: demographic information; admission diagnostic according to the McCabe score and the Charlson index classification; patient information at the study date (ward type and name, status of Methicillin-Resistant Staphylococcus Aureus portage, etc); and information at the study date and the six days before (clinical data, central venous catheter carriage, workload, infection status, etc). After a first data cleaning and binarization, this dataset contains 60 features and 1384 cases including 166 positive ones (11.99%). The second dataset, called S1, contains 20 features obtained after application of 2 feature selection methods (information gain [7] and SVM RFE [8]). The third dataset S2 is obtained from S1 but without the fever and workload features as the values of these features are not systematically gathered in clinical practice. We also highlighted in our previous study the redundancy or the negative interaction of these two features with the others making the learning with the dataset S and S1 challenging. The 2007 (resp. 2008) prevalence survey contains 1528 (1467) unique cases including 153 (156) positive cases. The ratio of positive cases turns around 10% and 12% for the 3 years.

We use libsvm with L2 implementation (i.e. *p=2* in relation (3)) and having a radius margin bound resolution implementation using gradient descent algorithm [9]. The software is executed on a linux machine having quad-processor of 2.33GHz frequency and 3Gb of memory.

Model selection and evaluation

We implement the same strategy as in our previous study: 105 random training and testing splits are created for S, S1 and S2. The number of splits is taken arbitrarily. The SVM parameters are obtained on 5 random training sets. The gradient descent algorithm is applied to each of the five training sets using 4 initialization points. For each of the initialization point, 3x5 cross-validations are performed and provide 60 couples of the SVM parameters. As the datasets present imbalance ratio on positive and negative cases, we arbitrarily correct the imbalance by taking equal numbers of positive and negative cases before performing cross-validations. The radius margin optimization may converge to the final SVM solution from each initialization point but we take the results having less absolute value of covariance on the 2 parameters.

The evaluation of the model on 2006 data is done on the 100 remaining training/testing splits i.e. 100 models are created with the best parameters and are evaluated on the corresponding test set. The mean of f-measure, precision, recall or sensitivity, specificity and accuracy over the 100 test sets is used as performance metric. The 2007 and 2008 prevalence data are also evaluated with the 100 models. The prediction of a case is the mean prediction over 100 models using a majority vote.

Results

Four initialization points are considered for model selection: $init1 = (e,1)$, $init2 = (e^2, e^2)$, $init3 = (e, e^{2^{-10}})$ and $init4 = (e^5, e^5)$ where *e* denotes the exponential function. The 60 SVM parameters of the dataset S (respectively S1 and S2) are obtained in 81 (respectively 56 and 53) seconds. Table 1 provides a summary of the obtained parameters with respect to their mean, median and standard deviation. The last line of Table 1 provides the absolute value of the covariance of the SVM parameters. All initialization points converge to the same value of *C* and *Sigma* but the initialization point *init4* (respectively *init3* and *init4*) provides less covariance of the SVM parameters for the dataset S (respectively S1 and S2).

These best parameters are used to build 100 models on the 2006 prevalence dataset; the results are depicted in Figure 1 and more details can be found in table 2. The horizontal bar on the top of figure 1 indicates if the results obtained with S, S1 and S2 are not significantly different based on the Mann-Wilcoxon mean test. This is the case for precision, specificity and accuracy of S and S1 and the Sensitivity of S1 and S2.

The evaluation of these models on the 2006, 2007 and 2008 data based on the features used is summarized in table 2. The mean of the f-measure, precision, sensitivity, specificity and accuracy of the models are reported. Confusion matrices are also reported for an intuitive reading of the results (A1, A2, A3, B1, B2, B3, C1, C2, C3). The numbers in the confusion matrix is the rounded mean of true positive, true negative, false positive and true negative obtained over the 100 models.

Discussion

As we have seen in the previous section, the use of the radius margin bound is attractive for SVM model selection with respect to its computational efficiency. A cross-validation procedure can take days if the step of the variables is thin. The computational efficiency of the radius margin bound allows us to carry out more experiments with several values of the ratio between the positive and negative examples.

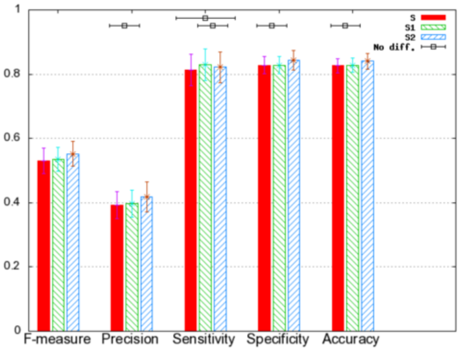

Figure 1 – Performance metrics of the models on the 2006 prevalence survey data.

For the 2006 dataset, the down-sampling methodology penalized the precision at the expense of recall. The obtained results are better than those from a previous study especially with the dataset S [3]. We have seen in the relations (1), (2), (3) and (4) that the SVM formulations are independent of the features present in the dataset. This was illustrated by the equivalence of the precision, sensitivity and accuracy between the datasets S and S1. The sensitivity of the three datasets S, S1 and S2 are equivalent and the other performance metrics of S2 are improved. From these results, we propose a new method to detect redundancy in a dataset using the SVM algorithm: if the removal of a subset of features keeps the recall unchanged while the other performance measures are improved then the subset of features has a negative interaction with the others. Many experiments are needed to confirm this proposal.

Table 1 - Best parameters C and Sigma according to the datasets and the initialization points on the dataset S, S1 and S2

		Dataset S				Dataset S1				Dataset S2			
		Init 1	**Init 2**	**Init 3**	**Init 4**	**Init 1**	**Init 2**	**Init 3**	**Init 4**	**Init 1**	**Init 2**	**Init 3**	**Init 4**
C	**Mean**	3.395	0.678	1.152	0.577	0.406	0.408	0.392	0.400	1.273	0.363	0.362	0.358
	Median	0.555	0.556	0.561	**0.549**	0.389	0.388	**0.391**	0.397	0.356	0.350	0.349	**0.348**
	Std dev	21.529	0.445	4.173	0.128	0.085	0.077	0.051	0.060	7.914	0.065	0.081	0.061
Sigma	**Mean**	0.069	0.067	0.068	0.065	0.199	0.199	0.198	0.197	0.232	0.230	0.230	0.227
	Median	0.066	0.065	0.071	**0.064**	0.189	0.194	**0.193**	0.193	0.205	0.201	0.204	**0.199**
	Std dev	0.041	0.038	0.040	0.029	0.086	0.085	0.076	0.082	0.128	0.124	0.115	0.108
Abs(Covariance (C,Sigma))		0.1906	0.005	0.035	**0.002**	0.0001	0.002	**0.000**	0.000	0.2123	0.001	0.002	**0.002**

Table 2 - Results obtained when applying the best parameters on the prevalence datasets of the year 2006, 2007 and 2008 according to the number of features in datasets S, S1, and S2

	2006			2007			2008		
Dataset S	**A1**	L+	L-	**A2**	L+	L-	**A3**	L+	L-
	P+	54	84	P+	69	35	P+	57	35
	P-	12	404	P-	84	1340	P-	99	1276
	F-MEASURE		0.5297	F-MEASURE		0.6008	F-MEASURE		0.6270
	PRECISION		39.29%	PRECISION		66.35%	PRECISION		61.96%
	SENSITIVITY		81.29%	SENSITIVITY		54.90%	SENSITIVITY		63.46%
	SPECIFICITY		82.78%	SPECIFICITY		97.45%	SPECIFICITY		97.33%
	ACCURACY		82.60%	ACCURACY		92.21%	ACCURACY		90.87%
Dataset S1	**B1**	L+	L-	**B2**	L+	L-	**B3**	L+	L-
	P+	55	84	P+	60	37	P+	56	40
	P-	11	404	P-	93	1338	P-	100	1271
	F-MEASURE		0.5377	F-MEASURE		0.6132	F-MEASURE		0.6108
	PRECISION		39.79%	PRECISION		61.86%	PRECISION		58.33%
	SENSITIVITY		82.88%	SENSITIVITY		60.78%	SENSITIVITY		64.10%
	SPECIFICITY		82.79%	SPECIFICITY		97.31%	SPECIFICITY		96.95%
	ACCURACY		82.80%	ACCURACY		91.49%	ACCURACY		90.46%
Dataset S2	**C1**	L+	L-	**C2**	L+	L-	**C3**	L+	L-
	P+	54	77	P+	65	41	P+	62	44
	P-	12	411	P-	88	1334	P-	94	1267
	F-MEASURE		0.5547	F-MEASURE		0.5936	F-MEASURE		0.5936
	PRECISION		41.87%	PRECISION		61.32%	PRECISION		58.49%
	SENSITIVITY		82.13%	SENSITIVITY		57.52%	SENSITIVITY		60.26%
	SPECIFICITY		84.27%	SPECIFICITY		97.02%	SPECIFICITY		96.64%
	ACCURACY		84.02%	ACCURACY		91.56%	ACCURACY		90.59%

With respect to our longer-term objective and with the features existing in S2 we can expect to have 41.87% true positive cases, which represent 82.13% of all infected patients when we build the model with the 2006 data. When the models were applied to datasets from 2007 and 2008, the "profile" of the performance changed: the sensitivity goes down considerably while the other performance measures increase. For the 2007 data, with the features in S2 we retrieved 61.32% true positive cases representing only 57.52% of infected patients. This drastic change of the performance measures can be explained as a consequence of the effectiveness of the prevention measures of the hospital. In other words, the prevention measures did not reduce the prevalence rate (around 10% from 2006 to 2008) but changed the signs and symptoms importance in our datasets and/or the risks of contracting the infections. A quick measure of the information gain followed by a chi-square filtering on the 3 datasets, as in [3] while we created the dataset S1, highlights a change in the order of the attributes and the appearance of new important ones. In machine learning this phenomenon is called "concept drift" [10]. Usually, the concept drift makes the model built on old data inconsistent with new data. If we want to use SVM to achieve our long-term objective, we need to take into account concept drift and implement, for example, an incremental method proposed in [11] i.e. exploiting the models build on previous prevalence survey datasets to predict actual cases.

Another important point in the analysis of the data warehouse data to support an automated prevalence system is the number of features to be extracted from the data warehouse, which was the topic of our previous study. A McNemar test (with Bonferoni's adjustment) was carried out in order to compare the performance obtained on the datasets S, S1 and S2 of years 2007 and 2008. This test indicates that the features present in S provide the best performance followed by S2. This indicates that if all the information of S can not be acquired for a new patient we have to use only those present in the dataset S2 (and not S1).

Conclusion

This study allows evaluating the robustness of SVM in a situation where the distribution of the features is changing over time. The sensitivity decreases and the other metrics increase when we apply a model build from datasets of the year Y on datasets of the year Y+1 and Y+2. This situation allows highlighting the effectiveness of the preventive measures taken after the prevalence survey in year Y. The use of the radius margin bound to select the SVM parameters was effective and allows us to carry out more experiments with respect to the manipulation of the ratio of positive and negative cases. We propose a new method to find redundant subsets of features in a dataset. We also found that the features present in S or S2 are necessary to build models. However it would be better to investigate the performance obtained using a dataset S3 including only common features obtained after the application of the information gain followed by a chi-square filtering on the prevalence datasets from the year 2006, 2007 and 2008. In the future, we plan to create models focusing on precision rather than the recall by using an asymmetrical misclassification cost. The main challenge is the adaptation of the radius margin bound for asymmetrical misclassification cost.

Acknowledgments

This work has been partially funded by the European Commission Sixth Framework Programme @neurIST project (IST-027703, see http://www.aneurist.org) and the KnowARC project (IST-032691), the Swiss National Science Foundation grant 200020–118638/1, and by the Geneva University Hospitals research grants 05-I-13 and 05–9–II.

References

[1] French GL, Wong SL, Cheng AF, Donnan S. Repeated Prevalence Surveys for Monitoring Effectiveness of Hospital Infection Control. The Lancet. 1989; 334:1021-1023.

[2] Sax H, Pittet D, and the Swiss-NOSO Network. Interhospital Differences in Nosocomial Infection Rates: Importance of Case-Mix Adjustment. Arch Intern Med. 2002; 162:2437-2442.

[3] Iavindrasana J, Cohen G, Depeursinge A, Meyer R, Geissbuhler A. Minimal Set of Attributes Required to Report Hospital-Acquired Infection Cases. In: Proc. Intelligent Data Analysis in bioMedecine And Pharmacology (IDAMAP). Washington, USA: 2008.

[4] Vapnik V. Statistical learning theory. New York, USA: John Wiley; 1998.

[5] Chapelle O, Vapnik V, Bousquet O, Mukherjee S. Choosing Multiple Parameters for Support Vector Machines. Mach Learn. 2002; 46:131-159.

[6] Bonnans FJ, Gilbert CJ, Lemaréchal C, Sagastizébal CA. Numercial optimization, theoretical and practical aspects. Springer Verlag; 2003.

[7] Quinlan JR. Induction of decision trees. Mach Learn. 1986; 1:81-106.

[8] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn. 2002; 46:389-422.

[9] Chung K, Kao W, Sun C, Wang L, Lin C. Radius Margin Bounds for Support Vector Machines with the RBF Kernel. Neural Comput. 2003; 15:2643-2681.

[10] Tsymbal A. The Problem of Concept Drift: Definitions and Related Work. Tech. report. Dublin, Ireland: 2004.

[11] Klinkenberg R, Joachims T. Detecting Concept Drift with Support Vector Machines. In: Proc. The Seventeenth International Conference on Machine Learning (ICML). San Francisco, CA, USA: 2000. pp. 487-494.

Address for correspondence

Rue Gabrielle-Perret-Gentil 4, CH-1211 Geneva 14, Switzerland
jimison.iavindrasana@sim.hcuge.ch