MEDINFO 2007
HISA Ltd Australia
© 2007 The authors. All rights reserved.

Applying Features Selection to Optimise the Cost-Benefit of Diagnostic Classifiers

Frank Lin[1], Vitali Sintchenko[1,2], Fanrong Kong[2], Gwendolyn Gilbert[2], Enrico Coiera[1]

[1]*Centre for Health Informatics, The University of New South Wales, Sydney, Australia*
[2]*Centre for Infectious Diseases and Microbiology, Institute of Clinical Pathology and Medical Research, Sydney West Area Health Service, Sydney, Australia*

Abstract

BACKGROUND: Classification by supervised machine learning (SML) is increasingly used in clinical diagnosis. However, the benefits of feature selection (FS) prior to the construction of diagnostic classifier remain to be elucidated.

METHODS: Group B streptococcus (GBS) data were trained by SML to differentiate the invasive strains from the normal flora. Five feature ranking algorithms were used to rank the relative merits of 19 GBS features prior to classification. Areas under the ROC curve (AUC) using 10-fold cross-validation of 7 classifiers were obtained. The benefit of FS was measured by the minimum number of features (N_p) required to achieve the desired level of AUC (AUC_p).

RESULTS: For the same number of features, all classifiers achieved better AUCs under feature selection. At $AUC_{0.9}$ (90% relative to the best AUC), the median N_p achieved by FS was 4, compared to 13 using random selection and 18 under the worst scenario. The reduction of AUC from 0.724 to 0.702 was not statistically significant (p=0.21).

CONCLUSION: Feature selection algorithms have the potential to improve the cost-benefit of diagnostic classifiers through the elimination of the non-discriminatory features without compromising the overall classification performance.

Keywords:

machine learning; feature selection; diagnosis

Introduction

Individualised medicine based on genomic and proteomic data has a promising future in clinical medicine. To enable individualised medicine on a practical level, factors such as accuracy, cost, and clinical utility of the tests need to be carefully evaluated. With the current advances in biotechnology, large amount of biological data can be generated rapidly and accurately. Supervised machine learning (SML) classifiers are increasingly used in analysing complex correlations in reaching a clinical diagnosis or in predicting outcomes of various diseases [1]. The issue of cost, however, remains a potentially limitation, ranging from bench-side material to expert interpretations [2]. Excluding non-discriminatory features can lead to cost reduction and potentially faster turnover time. In addition, the inclusion of non-discriminatory features not only can increase the training burden (curse of dimensionality), it could also affect the overall classifier performance.

In this paper, the benefits of feature selection (FS) algorithms were studied by extracting a subset of discriminatory features from the original unselected dataset prior to the training of SML diagnostic classifier. Feature selection has been known to facilitate dimensionality reduction to optimise measurement, storage, and efficiency of supervised classifiers [3]. To determine the optimal subset of features, we presented a graphical method for determining the optimal cut-off.

The dataset studied in this paper was the serotypic and genotypic data of a Gram positive bacterium, group B streptococcus (GBS). GBS is an important genitourinary pathogen that frequently finds its carriage in pregnant women. It has a potential to cause serious infection in neonates including sepsis, pneumonia, and meningitis. The trained classifier aims to distinguish the invasive isolates from the normal colonising flora based on various bacterial features. The classification aims to assist clinicians in identifying the high risk patients, such that by instituting prompt antimicrobial prophylaxis, the risk of potential morbidity and mortality can be reduced.

Methods

The group B streptococcus (GBS) dataset was used to construct classifiers for predicting the virulence of bacteria based on its serotype and genotype. Eighteen genotypes were determined by multiplex PCR and reverse line blot, including molecular serotype (MS), protein genetic profile (PGP), mobile genetic elements (*GBSi1, ISSag1, ISSag2, ISSag4, IS1381, IS861,* and *IS1548*), and antibiotic-resistance genes (*tetM, int, tetO, ermB, ermTR, aph, aad, mre,* and *mef)* [4][5][6][7]. Together with the capsular serotype (CS), 19 bacterial features were selected from the database. A total of 778 GBS isolates were used for classifier training, including 590 invasive and 188 colonising isolates that were collected from clinical cultures and routine antenatal swabs.

Seven SML classifiers were used in this study, including naïve Bayes (NB), Id3 and J48 decision trees, support vector machine (SVM) trained by sequential minimal optimisation (SMO) algorithm, logistic regression (LR), and decision table (DT). The bacterial features were ranked according to their relative discriminatory power by variable ranking algorithms based on correlation criteria (symmetrical uncertainty, and Chi-squared attribute selection) or information-theoretic criteria (Information gain and gain ratio, and ReliefF). Classifiers were trained and evaluated by adding each bacterial feature incrementally in sequence as determined by each FS algorithm. The performance of classifiers was evaluated by area under the ROC curve (AUC) using 10-fold cross validation. The classification and evaluation were performed using Waikato Environment for Knowledge Analysis (WEKA) [8].

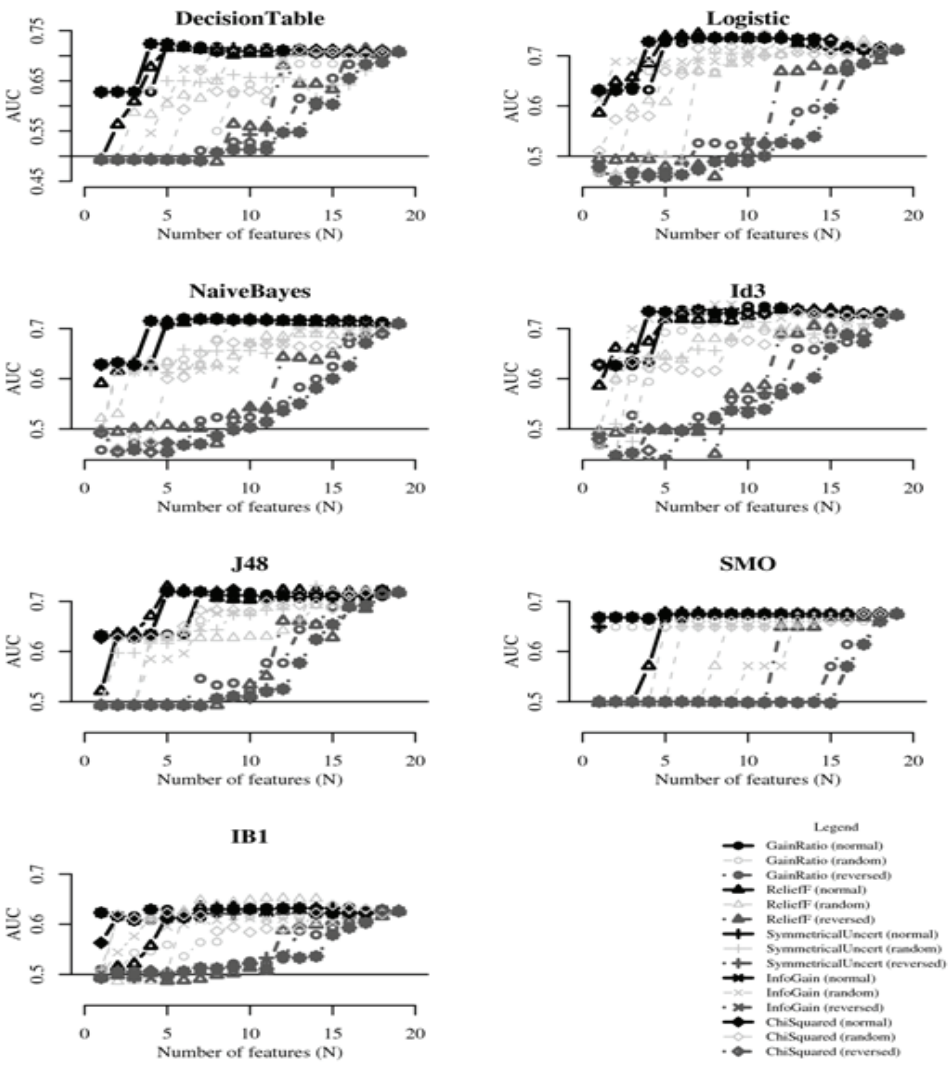

Figure 1 - AUC versus number of features (N) for 7 classifiers. (logistic: logistic regression; SMO: SVM trained by SMO algorithm; IB1: nearest-neighbour classifier)

To study the benefit conferred by the FS algorithms (normal), two extra feature selection schemes were performed. The theoretically "worst" ranking order was generated by reversing the normal rank produced by FS for each rank (reversed). In addition, the ranks were randomly shuffled to act as control, representing the

scenario where N features were selected by chance (random).

A cut-off value p (between 0.0 and 1.0) was defined to quantify the relative performance measured by AUC such that

$$AUC_p = 0.5 + p \times (AUC_{max} - 0.5)$$

In addition, N_p was defined as the minimum number of the first few features required to achieve AUC_p. The values of N_p were used to plot a cumulative gain chart for each classification scheme.

Results

The best AUC (AUC_{max}) achieved by FS using all 19 features under 10-fold classification were approximately 0.65-0.75 among the 7 classifiers (Table 1). At $N=19$, all AUC converged to the same point in all three studied ranking schemes (normal, random, and reversed). In general, the AUCs showed an increasing trend with more features involved in the classifier training. The AUCs for each classifier are shown in Figure 1.

The median N_p of the three ranking schemes across the classifiers were obtained and plotted in the cumulative gain chart as shown in Figure 2, which the ranking achieved AUC_p with less number of genes than the random and reversed set for all points. The median N_p at $p=0.90$ for the 7 classifiers were 4, 13, and 18 for the normal, random, and reversed ranks respectively. The median reduction in AUC (from 0.724 to 0.702) was not considered statistically significant (p=0.21, two-tailed unpaired t-test). The top-4 features selected by FS algorithms are shown in Table 2. Features CS, PGP, *GBSil* and *ISSag2* were identified as top-4 by >50% of the selected algorithms.

Classifier	N_p			AUC	
	normal	random	reversed	AUC_{max}	$AUC_{0.9}$
DT	4	13	18	0.724	0.702
NB	4	13	18	0.721	0.699
J48	5	13	18	0.731	0.708
IB1	4	9	18	0.652	0.637
LR	4	11	18	0.745	0.721
Id3	4	10	18	0.748	0.723
SMO	2	13	18	0.677	0.659
Overall	4	13	18	0.724	0.702

Table 1 - The median of the minimum number of features (Np) to achieve relative performance p=0.90 for all 7 classifiers. (DT: decision table; NB: Naive Bayes classifier; J48: J48 decision tree; Id3: Id3 decision tree; IB1: nearest neighbour classifier; LR: logistic regression; SMO: SVM by SMO algorithm)

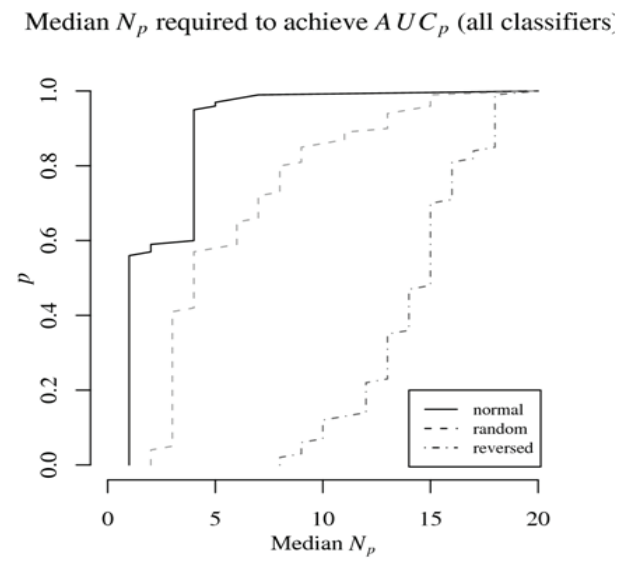

Figure 2 - Median N_p for all values of p

Rank	Feature Selection Algorithm				
#	ChiSq	SymUnc	GainRatio	InfoGain	ReliefF
1	**CS**	**CS**	**CS**	**CS**	MS
2	**PGP**	**PGP**	***GBSi1***	**PGP**	**PGP**
3	***GBSi1***	***GBSi1***	*mre*	***GBSi1***	*IS1381*
4	***ISSag2***	***ISSag2***	***ISSag2***	***ISSag2***	*intf*

Table 2 - Top-4 features as selected by feature selection algorithms. Bold font indicates the feature is identified by more than 50% of FS. (Feature selection algorithms: ChiSq: Chi-squared; SymUnc: Symmetrical Uncertainty)

Discussion

This paper evaluated the utility of feature selection algorithms in selecting a subset of features that best correlates bacterial informe random and reversed-FS ranking method for all number of features, suggesting the ability of FS algorithms in selecting the most discriminatory set of features for classification.

We used a graphical method in this paper for determining the optimal number of features to participate in the classifier training. For the desired level of AUC_p, the minimum number of features N_p can be determined by plotting a cumulative gain chart of the relative best AUC (p) that enables the visual representation of features-performance trade-off. Subsequently the corresponding N_p can be used

to select a discriminative set of features from the rank produced by FS algorithms given a desired level of classifier performance. This technique may be useful in data with larger dimensions to determine the optimal number of features, as the lengthy list generated by FS may not be immediately obvious for processing.

The benefit of FS in diagnostic classifiers can be demonstrated in Table 1 with $p=0.90$. By reducing the average number of features from 19 to 4, there was a potential reduction of over 75% on the cost of material given a <10% of relative AUC trade-off. The decrease in AUC did not represent a statistically significant decrease in performance. For our GBS classification task, the cost of the "unhelpful" PCR primers and hybridisation probes can be eliminated from the final product. According to Wilson's criteria, the cost-benefit is an important consideration in a diagnostic screening test [9]. Compared with the classifier trained with full set of features, the FS-optimised classifier potentially increased its acceptability as a screening tool, in which the cost of material was reduced through the elimination of the need non-discriminatory features where the benefits is preserved as measured by AUC.

Conclusion

Feature selection algorithms have the ability to optimise the cost-benefit of diagnostic classifiers by eliminating non-discriminatory features without compromising the overall classifier performance. This paper demonstrated that feature selection should be an integral part in the construction of machine learning-based diagnostic classifiers.

Acknowledgements

We would like to thank Ms Heather Gidding for maintaining the GBS database. This project is funded by Australian National Health and Medical Research Council.

References

[1] Sajda P. Machine Learning for Detection and Diagnosis of Disease. Ann Rev Biomed Eng 2006; 8:537-65

[2] McPherson E. Genetic diagnosis and testing in clinical practice. Clin Med Res. 2006 Jun;4(2):123-9.

[3] Guyon I. An introduction to variable and feature selection. J of Machine Learning Research 2003(3):1157-82

[4] Kong F, Gowan S, Martin D, James G, Gilbert GL. Serotype identification of group B streptococci by PCR and sequencing. J Clin Microbiol. 2002 Jan;40(1):216-26.

[5] Kong F, Gowan S, Martin D, James G, Gilbert GL. Molecular profiles of group B streptococcal surface protein antigen genes: relationship to molecular serotypes. J Clin Microbiol. 2002 Feb; 40(2):620-6.

[6] Kong F, Martin D, James G, Gilbert GL. Towards a genotyping system for Streptococcus agalactiae (group B streptococcus): use of mobile genetic elements in Australasian invasive isolates. J Med Microbiol. 2003 Apr; 52(Pt 4):337-44.

[7] Zeng X, Kong F, Wang H, Darbar A, Gilbert GL. Simultaneous detection of nine antibiotic resistance-related genes in Streptococcus agalactiae using multiplex PCR and reverse line blot hybridization assay. Antimicrob Agents Chemother. 2006 Jan; 50(1):204-9.

[8] Witten IH and Frank E. Data Mining: Practical machine learning tools and tecnniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[9] Wilson JMG, Jungner G. Principles and Practice of Screening for Disease. WHO Chronicle 1968; 22(11):473

Address for correspondence

Dr Frank Lin
frank.lin@student.unsw.edu.au
Centre for Health Informatics
The University of New South Wales
45 Beach Street, Coogee
NSW 2034, Australia