10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.1, June 4-6, 2007, Cancún, Mexico

FEATURE SELECTION AND CLASSIFICATION OF METABOLOMIC DATA USING SUPPORT VECTOR MACHINES

S.Mahadevan * **S.L.Shah** * **C.M.Slupsky** **
T.J.Marrie ** **E.Saude** ** **D.J.Adamko** **

* *Department of Chemical and Materials Engineering, University of Alberta, Edmonton, Alberta, Canada*
** *Department of Medicine, University of Alberta, Edmonton, Alberta, Canada*

Abstract: Over the past few years there has been an explosion of biological data available for exploratory analysis. The main task of data analysis is to extract meaningful information in a way that facilitates the understanding of the complex biological processes. In order to do this, algorithms and techniques have to be developed that can be trained to learn rules and form patterns from the available data sets and then apply these rules to analyse new data. In computing science terminology this is known as machine learning. In this paper, the applicability of one such machine learning technique, namely 'support vector machines' to analyze and classify metabolomic data is explored. The paper also explores some of the feature selection algorithms which help determine important biomarkers or metabolites in data sets.

Keywords: Classification, Learning, Metabolomics, Training, Support Vector Machines, Feature Selection, Diagnosis, Data Mining

1. INTRODUCTION

Metabolomics can be defined as the field of science that deals with the measurement of metabolites in the body of an organism, in order to study the physiological processes and reactions of the body to various stimuli such as infection, disease or drug use. Metabolomics has also been applied to cell culture and micro-organisms. Such studies are of great use in early diagnosis of diseases and pre-clinical screening of candidate drugs in the pharmaceutical industry. In order to carry out such studies, analytical processes such as NMR spectroscopy and mass spectrometry are combined with statistical tools, such as multivariate analysis and machine learning tools, such as neural networks, hidden markov models and support vector machines (SVM) (Nicholson *et al.*, 1999).There have been several papers which discuss the application of multivariate techniques such as Principal component analysis (Holmes and Antti, 2002), Clustering analysis and PLS-DA (partial least squares-Discriminant analysis) (Keun *et al.*, 2003) to metabolomics. However, not much work has been done in applying state of the art machine learning tools such as artificial neural networks (Yang *et al.*, 2002) and support vector machines in analyzing metabolomics data. For quite some time now, SVMs have been used in the field of bioinformatics, especially in classifying gene expression microarray data (Guyon *et al.*, 2002), identifying protein homologies (Jaakkola *et al.*, 1999) and predicting reactions (Mu *et al.*, 2006).

Support Vector classifiers fall under the category of supervised techniques. This work focuses on

feature selection and construction of SVM classifiers which classify the given metabolomic profiles into different diseases. The data investigated in this work are metabolomic expression profiles.

Another important task in the field of bioinformatics is the issue of feature selection. This is essential in identifying the subset or combination of original variables which are responsible for producing the observed classification. Once these features are identified, further analysis is carried out to identify the reasons as to why they are important. This can be done by going back to the actual metabolic pathways that these variables (metabolites) are involved in. This helps in identifying the actual mechanism behind a particular disease and the role a particular metabolite may play in the drug design process. The pathways can also reveal discovery of new drugs which can counter that particular disease. Another advantage of feature selection is dimension reduction. The following section will briefly introduce the principles of support vector machines.

2. SUPPORT VECTOR MACHINES

The foundation for Support Vector machines was laid in 1982 by Vladimir Vapnik (Vapnik, 1982) and formally proposed by Boser (Boser *et al.*, 1992). SVMs are quite robust when it comes to handling noisy data and they are also not susceptible to the presence of outliers. The basic principle of a binary support vector classifier is as follows: given a data set comprising data from two different categories, it constructs an optimal linear classifier in the form of a hyperplane which has the maximum margin.

In the case of data sets which are not linearly separable, the data is mapped into a higher dimensional feature space, where a linear classifier is constructed. For performing this, the following kernel functions are commonly used:

- Linear kernel: $x_i^t x_j$
- Polynomial kernel: $(\gamma x_i^t x_j + constant)^d, \gamma > 0$, d is the degree of the polynomial
- Radial basis function kernel (rbf): $e^{-\gamma||x_i - x_j||^2}$
- Sigmoidal kernel: $tanh(\gamma x_i^t x_j + constant)$

The appropriate parameters for different kernels are typically chosen by performing a cross-validated grid search. For a rbf kernel the values of γ and regularization parameter c are varied and for a polynomial kernel γ, c and d are varied, while for the linear kernel only the value of c needs to be varied. The usage of sigmoidal kernel is generally avoided as it is not positive semi-definite for all values of γ. The software that was used for constructing the support vector classifiers is LIBSVM (Chang and Lin, 2001). For a more detailed account on Support vector machines see (Burges, 1998).

3. FEATURE SELECTION

Feature Selection algorithms in machine Learning can be broadly classified under two categories: i)Filter approach and ii)wrapper approach. The former is independent of the actual classifer algorithm. The selection is mainly done on the basis of a ranking system. Univariate correlation scores such as Fisher Scores are used to rank the variables that are involved. In the wrapper method, feature selection is done in conjunction with the training phase. Usually a subset of the variables are chosen and the performance of the classifier is evaluated on this subset. The subset of variables which gives the best classifier performance is chosen for final analyis. For a detailed review on Feature Selection methods see (Kohavi and John, 1997) and (Chen and Lin, 2006). In this work the performance of the filter approach (Fisher scores) and the wrapper approach (recursive feature elimination(SVM-RFE)) are compared to the case where no feature selection is done.

SVM Recursive Feature elimination is a wrapper approach which uses the norm of the weights w to rank the variables. A more elaborate version of this algorithm can be found in (Guyon *et al.*, 2002).

4. DATA

In this work two types of metabolomic data were analaysed. One is the NMR spectroscopic data and the other is the actual concentration data.

4.1 NMR spectral data

NMR spectroscopic data of human urine samples (both diseased and normal people) were obtained using a Varian 600MHz spectrometer. Details of the NMR sample preparation and spectral acquisition are beyond the scope of this paper and will be described elsewhere. The final spectral data consists of the intensity value at 65536 frequency points. Figure 1 shows a typical NMR spectrum of urine. In figure 1, the X-axis is the relative frequency (relative to internal standard) having 65536 points and the y-axis represents the corresponding intensity values. In this case 105 data samples consisting of 52 normal samples and 53 *Streptococcus pneumoniae* samples are used. The raw data can be represented as a 108x65536 matrix with rows representing different human

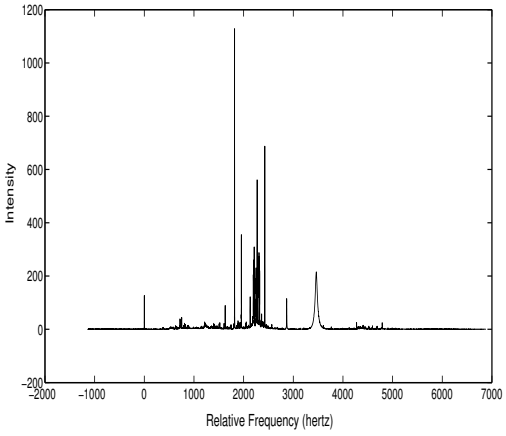

Fig. 1. Typical NMR spectrum of a urine sample

samples and the columns representing intensity values at each of the 65536 frequency points.

Some preprocessing steps were carried out before the data was further analysed. The steps were as follows:

- The raw spectral data needed to be aligned properly. In other words, the intensity values along a particular column do not correspond to a single frequency value. So the spectrum is first aligned according to the reference DSS peak (Dimethyl Silapentane sulfonate salt: this is the internal standard).
- From figure 1, it can be seen that there are tails at both the ends of the spectrum which are absolutely flat. This is the featureless part of the spectrum and can be trimmed off for each of the samples.
- Urine is essentially an aqueous solution comprising of more than 90% water. So the water peak in the spectrum is usually much taller than most of the peaks, thus dominating the spectrum. In order to remove this artefact, the region of the spectrum containing the water peak is removed.
- For dimensionality reduction, standard binning is done. Here the peaks are integrated every 18 frequency points. So effectively the dimension of the system is reduced by a factor of 18. The binning interval has to be carefully chosen. If it is too small, there is a risk of amplifying the noise in the system and also in some cases the bins will cut across a single peak. If the binning interval is too large, there may be a loss of information that is stored in the relatively narrower peaks.
- The data matrix was mean-centred and scaled to unit variance.

After pre-processing, the data matrix, say X, is of size 105x2188. There is also a class vector $Y_i \in \{-1, 1\}$,i=1,...,105. Class -1 representing normals and class 1 representing *S. pneumoniae* samples. The support vector algorithm is applied on this dataset.

4.2 Concentration data

Concentration data was obtained by quantification of 82 known urinary components (metabolites) in the NMR spectrum. This is essentially fitting the NMR peaks with known database of Lorentzian signatures for different metabolites. The concentration data consists of 118 samples comprising of 59 normals and 59 *S. pneumoniae*, each sample consists of the concentration value (measured in μ molar) of 82 metabolites. Some of the metabolites measured are: Lactose, glucose, glutamate, citrate, carnitine etc.

5. RESULTS AND DISCUSSION

5.1 Error rate calculation

The data matrix, be it concentration or spectral, is split randomly into 50% training data and 50% test data. The algorithm is trained on the training data to build a classifier model. Then this model is tested on the test data and the misclassification rate is computed. 500 iterations are repeated and the mean error rate is computed.

5.2 Concentration data

First the concentration data is taken, split into training and test data and the mean error rate was calculated. Traditional methods like Linear Discriminant analysis(LDA) and K-nearest neighbour(KNN) were also tried. Table 1 shows the summary of results.

Table 1. Concentration Data.

Method	kernel type	γ	c	d	Error Rate (average %)
SVM	linear		1000		8
SVM	polynomial	2	1000	1	7
SVM	rbf	1.22e-2	256		7.5
LDA					24.5
KNN					21

From table 1 it is observed that SVM outperforms the traditional methods in terms of the classification error rate by a significant margin. In the SVM method the polynomial kernel performs marginally better than other kernels. One disadvantage of SVM is the selection of a suitable kernel. It is difficult to say analytically which kernel is best for a given data set. The parameters for the kernel, as mentioned earlier were obtained after a grid search (Chang and Lin, 2001). Figure 2 shows

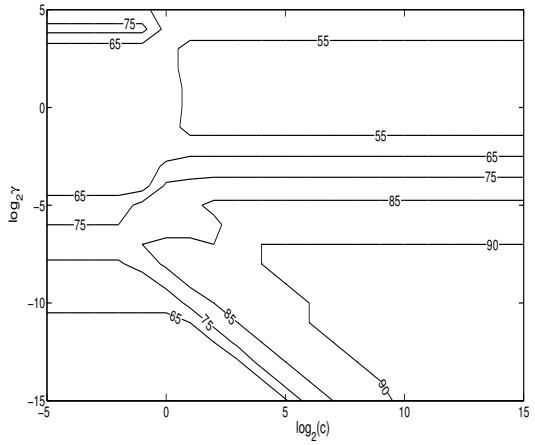

Fig. 2. Contour plot for parameter grid search

a sample contour plot for a rbf kernel generated using MATLAB. The accuracy rate is a constant along any particular line and is embedded within each line. It has to be noted that the axes are in logarithmic scale.

5.2.1. Feature Selection In order to extract the important features, the Fisher Score method and the SVM-RFE method are implemented. The results are tabulated in table 2.

Table 2. Concentration Data.

Feature Selection Method	No.of features selected	Error Rate (average %)
Fisher Score	10	4.8
SVM-RFE	17	2.8

From table 2 it can be seen that, as expected, SVM-RFE performs better than the Fisher method in terms of the classification error rate. This is because the former is a wrapper approach, where feature selection is done in conjunction with the SVM algorithm. The Fisher method is based on a uni-variate measure, hence it does not account for the mutual information contained in the features. However, both the methods perform better when compared to the case where no feature selection is done. Hence it could be concluded that feature selection reduces the dimension of the system as well as increases the separation between the two classes. In order to find the optimum number of features, the number of features is varied sequentially and the corresponding mean error rate is calculated. The one for which the error rate is the lowest is chosen. A sample plot for the SVM-RFE method is shown in figure 3. It is clear from figure 3 that the error rate attains a minimum when 17 features are selected.

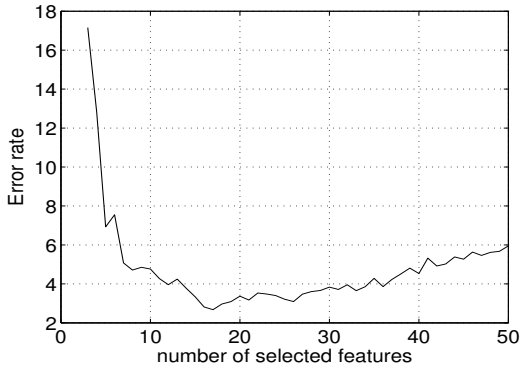

Fig. 3. Number of selected features vs error rate

5.3 Transformations

In this section the significance of making appropriate transformations are studied. Instead of the traditional unit variance scaling, other transformations such as Pareto scaling (data scaled by $\sqrt{std}$), the log and power transformations were tried on the concentration data. Classifiers were then constructed on this transformed data. Of these, log transformation performed (in terms of clasification error rate) better than the UV Scaling, while the other transformations did not fare well. The Support vector classifiers were built on the log transformed data and it was found that a polynomial kernel with $\gamma = 0.000122$ and c=5000 gives a classification accuracy rate of 98.2%. This high classification rate can be substantiated by plotting a simple 2-d scores plot (figure 4).

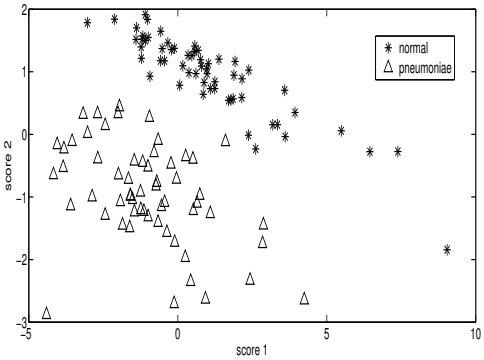

Fig. 4. PCA scores plot of log transformed data

It is evident from figure 4 that the log transformed data is almost linearly separable. So appropriate transformations can make the data more easily interpretable. This transformation was not applied on spectral data, because it contains zero elements in the matrix.

5.4 spectral data

The spectral data was split into training and test data and the mean classification error was calculated. The results are tabulated in table 3

Table 3. spectral Data.

Method	kernel type	γ	c	d	Error Rate (average %)
SVM	linear		10000		9.5
SVM	polynomial	0.25	32000	1	9
SVM	rbf	0.25	32000		7.8
LDA					18.9
KNN					14.2

From table 3 we can see that the rbf kernel gives the best classification rate. This error rate is slightly higher than the one obtained from the concentration data. In fact the spectral data encodes more information about different metabolites (in terms of peak intensities), so in this sense it is expected to perform better than the concentration data. However, it should be noted that the spectral data contains a lot of extraneous information which do not contribute to the classification thus reducing the signal to noise ratio.

5.4.1. Feature Selection Feature Selection was performed by both Fisher and SVM-RFE methods. Results are tabulated in table 4

Table 4. Spectral Data.

Feature Selection Method	No.of features selected	Error Rate (average %)
Fisher Score	10	16
SVM-RFE	40	3.9

Here again it is shown that SVM-RFE is better when compared to the Fisher Scores method. In order to explain the poor performance of the Fisher score method, these scores were plotted against the rank of the scores in figure 5. From

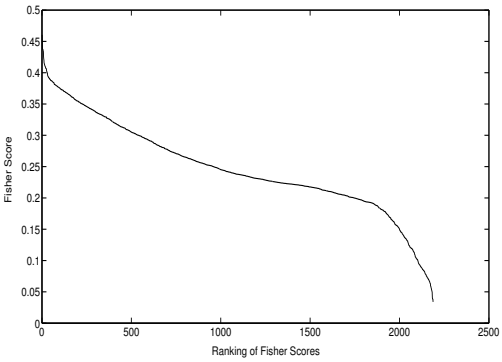

Fig. 5. Curve of F-score against features

figure 5, it can be observed that the Fisher score value for most of the features are quite significant. If a threshold is fixed, for example only the features for which the ratio of its Fisher score to the maximum score is less than 0.9, is retained. In this case almost all the features will be retained. Thus the Fisher score method does not serve the purpose of feature selection in this case.

5.5 Comparison of Selected features

The selected features in the spectral data are essentially integrated bins of the raw spectrum. These bins can be identified and the corresponding peaks in the original spectrum can be highlighted. A zoomed version of one such highlighted spectrum is plotted in figure 6. These highlighted

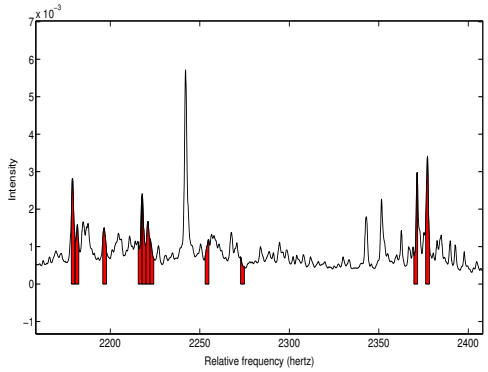

Fig. 6. Highlighted peaks of NMR Spectrum

peaks can be further analysed as to which particular metabolite they represent with the help of existing metabolite profiling databases.

For the concentration data, the first 20 important features selected by both the methods have been listed in Table 5 starting from the most important one.

From table 5, we can see that out of 20 important features, 12 features are common in both the methods. These feature numbers actually represent specific metabolites. For example feature number 29 represents citrate. The biological significance of these metabolites is beyond the scope of this work and will be discussed elsewhere.

Further work has been done to extend this binary classification to multi-class classification. Due to lack of space, this section has been omitted. However, readers can refer to the full version of this paper at www.ualberta.ca/˜sm17.

6. CONCLUSION

In this paper it has been shown that Support Vector Classifiers are highly efficient in classifying metabolomic data. Feature selection was also

Table 5. 20 important features

Feature Selection																				
F-Score	51	31	52	29	22	56	39	12	64	79	67	54	43	60	50	21	72	28	55	75
SVM-RFE	51	31	52	29	22	56	39	12	64	79	67	54	11	35	13	16	24	74	2	38

performed to highlight the important metabolites that contribute to the classification. The SVM-RFE algorithm was found to perform better than the Fisher scores filter approach. It can be concluded that these diseases are not caused by a single biomarker, but by a combination of some of the metabolites. SVM is also robust to the presence of outliers, as it is the support vectors which determine the classifier. Since the data sets have very few samples when compared to the number of features, the classifier models easily tend to overfit the system. SVM overcomes this problem by adopting a cross validation-based generalization approach. This can also be verified by observing the number of support vectors in the classifier model which is usually much less than the number of samples. This algorithm has also been successfully extended to multi-class classification of metabolomic data.

7. ACKNOWLEDGEMENT

The authors would like to thank David Chang, a graduate student for his valuable suggestions. The authors would also like to thank Dr.Brian Sykes, Dr.Brian Rowe and Dr. Allan Becker for providing the Asthma spectral data, Allison McGeer for providing the *S. pneumoniae* samples and Kathryn Rankin for analyzing the same. The authors are grateful for the software support from Chenomx Inc. Funding agencies NSERC, AFHMR, CIHR and AllerGen are acknowledged. The National High Field NMR Centre (NANUC) and the Magnetic Resonance Diagnostic Centre (MRDC) are acknowledged for their assistance and use of the facilities.

REFERENCES

Boser, B.E., I.M. Guyon and V.N. Vapnik (1992). A training algorithm for optimal margin classifiers. *Proceedings of the* 5^{th} *annnaul ACM Workshop on Computational Learning Theory* pp. 144–152.

Burges, Christopher J.C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Chang, C.C. and C.J. Lin (2001). Libsvm: A library for support vector machines.

Chen, Y.W. and C.J. Lin (2006). Combining svms with various feature selection strategies. In: *Feature Extraction, Foundations and Applications* (I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh, Eds.). Chap. 12. Physica-Verlag, Springer.

Guyon, I., S. Barnhill J. Weston and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* **46**, 389–422.

Holmes, E. and H. Antti (2002). Chemometric contributions to the evolution of metabonomics: mathematical solutions to characterising and interpreting complex biological nmr spectra. *Analyst* **127**, 1549–1557.

Jaakkola, T., M. Diekhans and D. Haussler (1999). Using the fisher kernel method to detect remote protein homologies. *Proceedings of the* 7^{th} *International Conference on Intelligent Systems for Molecular Biology*.

Keun, H.C., T.M.D. Ebbels, H. Antti, M.E. Bollard, O. Beckonert and E. Holmes (2003). Improved analysis of multivariate data by variable stability scaling: application to nmr-based metabolic profiling. *Anal Chim Acta* **490**, 265–276.

Kohavi, R. and G. John (1997). Wrappers for feature subset selection. *Artificial Intelligence* **97**(2), 273–324.

Mu, F., P.J. Unkefer and C.J. Unkefer (2006). Prediction of oxidoreductase-catalyzed reactions based on atomic properties of metabolites. *Bioinformatics* **22**(24), 3082–3088.

Nicholson, J.K., J.C. Lindon and E. Holmes (1999). 'metabonomics':understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica* **29**(11), 1181–1189.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data*. Springer Verlag. New York.

Yang, J., G.W. Xu, H.W. Kong, W.F. Zheng, T. Pang and Q. Yang (2002). Artificial neural network classification based on high-performance liquid chromatography of urinary and serum nucleosides for the clinical diagnosis of cancer. *Journal of chromatography. B, Biomedical sciences and applications* **780**, 27–33.