Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31 - August 4, 2005

Entropy Based Disease Classification of Proteomic Mass Spectrometry Data of the Human Serum by a Support Vector Machine

Terje Kristensen and Gaurav Kumar
Department of Computer Engineering,
Bergen University College, Nygårdsgaten 112,
N-5020 Bergen, Norway
E-mail: tkr@hib.no

Abstract—**Disease diagnostics using proteomic patterns is a new platform that is developed to detect early-stage cancer. Proteomic pattern analysis uses the overall pattern to diagnose disease states without the need to identify the components within the pattern. The patterns are generated from Mass Spectrometry (MS) data, and an algorithm is developed to decipher the patterns within the mass spectrometry data to discriminate between serum taken from healthy and cancer-affected individuals.**

There is need for cancer biomarkers with more accurate diagnostic capability. Use of MS is such a technique. Mass spectrometry data of the human serum consist of intensities of various ions present in the sample. A typical sample can have about 15000 different ions present. A very important question then is which ions are the best classifiers.

We have used an information-theoretical concept, information gain, to measure how well a given attribute separates the training examples according to the target classification. The method measures the drop in the entropy of the system caused by selecting a particular attribute. The lower the drop, the better the attribute. Our algorithm first selects the attributes with highest information gain and then classifies the diseased and healthy data based on these attributes using Support Vector Machines (SVM). The method achieves very strong performance.

IndexTerms- **Mass spectrometry, proteomics, entropy, information gain, SVM.**

I. Introduction

Disease diagnostics using proteomic patterns [1] has recently been developed as a diagnostic tool which does not rely on the identification of the proteins detected. The ability to discriminate patterns from serum acquired from healthy individuals, from serum of cancer-affected individuals is the most important aspect of this technique. Proteomic pattern diagnostics is a type of pattern diagnostics based on the analysis of a huge amount of data to find disease patterns in the proteins expressed. Serum proteomic signatures from mass spectrometry data are used as a diagnostic classifier of proteomic signatures from high dimensional MS data.

Such an approach has given very promising results in detection of early stage cancer [10]. The blood proteome is changing constantly as a consequence of the perfusion of organ systems. Small peptide fragments are removed from the actual disease organ and are contained in the blood serum. These fragments contain low molecular weight molecules which exist below the range of detection of conventional techniques. As a result researchers have turned to mass spectrometry that exhibits optimal performance in the low mass range [6], [9].

MS is a powerful analytical tool for determining masses of bio-molecules in a complex sample mixture. Such a technique is used to identify compounds. Detection of compounds can be accomplished with very minute quantities. This means that compounds can be identified at very low concentration in chemically complex mixtures. Experimental conditions that effect the molecular composition of a sample will also affect its mass spectrum. Mass spectrometry is used to test for the presence of different kinds of molecules, and the presence of such molecules may indicate an enzymatic change, a disease state or a certain cell type condition.

Mass spectrometry data consists of a set of m/z values (m is the atomic mass and z is the charge of the ion) and the corresponding relative intensities of all molecules present with that m/z ratio. The mass spectrometry data of a chemical sample thus is an indication of presence or absence of the actual molecules. The data might therefore be used to predict the presence of a disease condition and distinguish it from a sample taken from a healthy individual or any other living organism in general with a circulatory system.

Since the mass spectrometry data consist of intensities of thousands of molecules in a sample, it cannot be analysed manually. Computational methods such as artificial neural networks (ANN) should be suitable to do such an analysis. In two earlier papers we have shown that use of ANN techniques are suitable, for instance, to classify between eukaryotic or prokaryotic cells [4], [5]. However, in this paper we want to study how SVM can be used to perform an analysis to discriminate between different types of cancers.

The database used consists of individuals suffering from either ovarian or prostatic cancer, in addition to healthy persons. A problem when using a MS technique is how to select the most suitable attributes from a database of about 15000 attributes, to train the network. In this paper we have

0-7803-9048-2/05/$20.00 ©2005 IEEE

used an information-theoretic measure based on the entropy concept to discriminate between the most important attributes [7].

II. Entropy and Information Gain

Given a collection S, containing positive and negative examples of some target concept, the entropy S relative to this boolean classification is defined by

$$Entropy(S) = -\ p+\log_2 p+ \ \ -\ p^{-}\log_2 p^{-} \qquad (1)$$

where p+ is the proportion of positive examples in S and p^- is the proportion of negative examples in S.

In (1) we have defined entropy for boolean classification. For a general case, if the target attribute can take on n different values, the entropy of S relative to a n-wise classification would be

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (2)$$

where pi is the proportion of S belonging to class i. The entropy concept in (2), characterises the impurity of an arbitrary collection of examples, can now be used to define another important concept in information theory, the information gain.

The information gain measures the expected reduction in entropy. Given the entropy as a measure of the impurity in a collection of training examples, we now define a measure of effectiveness of an attribute in classifying the training data. The information gain simply measures the expected reduction in the entropy caused by partitioning the examples according to this attribute. More precisely, the information gain, Gain(S,A) of an attribute A, relative to a collection of examples S, is defined as

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} Entropy(Sv_i) \qquad (3)$$

where $v_1, v_2, ...v_n$ is the set of all possible values of attribute A, Sv_r is the subset of S for which the attribute A has the value v_r i.e., $Sv_r = \{s \ \varepsilon\ S \mid A(s) = v_r\}$.

In the above representation, the first term is just the entropy of the original collection S, and the second term is the expected value of the entropy after S is partitioned using attribute A. The expected entropy described by the second term is simply the sum of the entropies of each subset Svr, weighted by the fraction of examples $|Sv_r| / |S|$ that belong to Sv_r. Gain(S,A) is therefore the expected reduction in entropy caused by knowing the value of attribute A. An important point to note is that the above representation cannot be used if the attribute A is continuous-valued.

A. Incorporating the Continuous-Valued Attributes

Our initial definition of information gain restricts the attributes to take on a discrete set of values. This restriction can easily be removed so that continuous valued decision attributes can be incorporated. This can be accomplished by dynamically defining new discrete values attributes that partition the continuous attribute value into a discrete set of intervals. In particular, for a attribute A that is continuous-valued, the algorithm can dynamically create a new boolean attribute Ac that is true if $A < c$ and false otherwise. The choice of the threshold c is based on the maximum information gain achieved.

Once the best classifying attributes have been selected the SVM can be used to train the data based on these selected attributes.

III. Svm Theory

SVM is a computationally efficient learning technique that is now being widely used in pattern recognition and classification problems [2]. This approach has been derived from some of the ideas of statistical learning theory regarding controlling the generalization abilities of a learning machine [11], [12].

In this approach the machine learns an optimum hyper-plane that classifies the given pattern. By use of kernel functions, the input feature space by applications of a non-linear function can be transformed into a higher dimensional space where the optimum hyper-plane can be learnt. This gives a flexibility of using one of many learning models by changing the kernel functions.

A. SVM Classifier

The basic idea of a SVM classifier is illustrated in Fig.1. This figure shows the simplest case in which the data vectors (marked by 'X' s and 'O' s) can be separated by a hyper-plane. In such a case there may exist many separating hyper-planes.

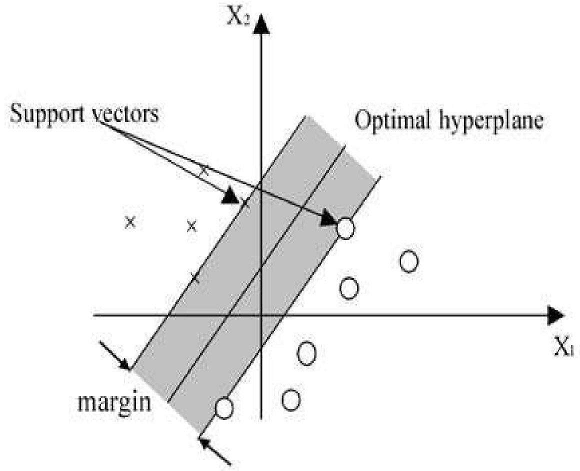

Fig. 1. Support Vector Machines classification defined by a linear hyper-plane that maximizes the separating margins between the classes.

Among them, the SVM classifier seeks the separating hyper-plane that produces the largest separation margins.

In the more general case, in which the data points are not linearly separable in the input space, a non-linear transformation is used to map the data vectors into a high-dimensional space (called feature space) prior to applying the linear maximum margin classifier. To avoid the potential pitfall of over-fitting in this higher dimensional space, a SVM uses a kernel function in which the non-linear mapping is implicitly embedded. A function qualifies as a kernel function if it satisfies the Mercer's condition [11].

With the use of a kernel function, the discriminant function in a SVM classifier has the following form :

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i \, K(\mathbf{x_i}, \mathbf{x}) + b \qquad (4)$$

where K(-,-) is the kernel function, $\mathbf{x_i}$ are the support vectors determined from the training data, y_i is the class indicator (e.g. +1 and -1 for a two class problem) associated with each $\mathbf{x_i}$, N is the number of supporting vectors determined during training, α is the Lagrange multiplier for each point in the training set and b is a scalar representing the perpendicular distance of the hyper-plane from origin..

Support vectors are elements of the training set that lie either exactly on or inside the decision boundaries of the classifier. In essence, they consist of those training examples that are most difficult to classify. The SVM classifier uses these borderline examples to define its decision boundary between the two classes.

B. The SVM Kernel Functions

The kernel function plays a central role of implicitly mapping the input vectors into a high dimensional feature space, in which better separability is achieved. The most commonly used kernel functions are the polynomial kernel given by :

$$K(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i}^T \mathbf{x_j} + 1)^p \text{ , where } p > 0 \text{ is a constant} \qquad (5)$$

or the Gaussian radial basis function (RBF) kernel given by

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\|\mathbf{x_i} - \mathbf{x_j}\|^2 / 2\sigma^2 \right) \qquad (6)$$

where $\sigma > 0$ is a constant that defines the kernel width. Both of these kernels satisfy the Mercer's condition mentioned above.

IV. SVM TRAINING

In the experiments we have used databases from the NIH and FDA Clinical Proteomics Program Data Bank [8]. The data set consisted of different m/z values and their intensities. Each of the above databases is divided into healthy and diseased sets. The files in each of these datasets are comma delimited. The above algorithm helps us to find the best attributes for classification of diseased and healthy samples.

Since the m/z values are continuous, we need to find a 'c' for each of the entropy calculations. This value is decided based on the maximum information gain achieved. Once this value is found, we can use the same method as we used to handle the entropy calculation for discrete values. A table is then prepared for all the m/z attributes versus the reduction in entropy they bring about. This table relates the usefulness of each attribute.

Once the best attributes are found we use the LIBSVM toolbox [3] to classify the spectrums into diseased and healthy categories, based on only those attributes. We see that using this method we can reduce the number of attributes needed for classification of diseased versus healthy sample. In previous experiments using SVM one selects 7-8 attributes randomly and use these for classification [10]. Our method makes the whole process more organised and also gives better results.

During the training phase the variables in the kernel function and the regularization parameter C have to be determined. The training samples were divided into m equal subsets of equal size and a methodology based on *one-versus one* was used in the classification regime. The experiments were done with various parameters settings. The model with the best generalization, e.g least error, was then selected.

V. EXPERIMENTS AND RESULTS

Once the best attributes have been estimated we used these attributes for SVM classification. We performed the training of SVM using the best 5 attributes followed by the best 4 attributes and so on. Finally, we train the SVM using only the best attribute. The results have been summarized in table 1 given underneath.

Accuracy %	C = 1	C = 100	C = 500
Best 5	99.2	100	100
Best 4	99.2	100	100
Best 3	99.2	100	100
Best 2	98.8	100	100
Best 1	96.8	97.6	98.1

Table 1. SVM performance classification of the 1-5 best attribute values.

From table 1 we see that perfect classification is achieved by using only 3-5 attributes. In [10], 7-8 attributes are selected randomly and used for classification. By use of the method presented in this paper we are able to give priority to the different attributes, based on their information gain, which gives optimal classification.

The columns in table 1 denote the performance of the SVM classifier at various values of the cost parameter C. The larger the value of C, the better the classification. The rows denote the performance in percent, dependent on the number of attributes taken for training of the SVM. We see that a larger number of attributes give a higher accuracy.

VI. Conclusion

Proteomic pattern diagnostics represents a new paradigm for disease detection. This type of analysis requires only a small amount of blood from which MS spectra are generated. The most promising aspects with such an analysis are a very high throughput because the MS spectra can be determined in very short time. In such an analysis the pattern itself, independent of the identity of the proteins, is the discriminator. This may be done before the identity of the proteins is determined.

The MS platform is promising to use for cancer diagnostics. By use of MS one can generate complex proteomic spectra from an extreme small volume of blood in short time. Combined with nano-technology this platform can generate new tools, created at the intersection between proteomics and the nano-technology.

In such a future perspective we might also introduce nano-harvesting agents into the blood serum that are able to diagnose on the fly, based on the MS data taken. Such nano-particles with their diagnostics cargo, can communicate remotely with a computer, and the status of the blood serum may be checked to reveal the signatures of these biomarkers.

In this paper we have developed a new method that predicts the best attributes to be used to classify the most appropriately disease attributes. Once these attributes have been determined, using the concept of information gain, we can train a SVM network for classification of cancer vs. non-cancer samples.

A very high classification accuracy (100%) is obtained in the experiments using the best 5 attributes for training of the SVM. This result is superior to previously applied methods such as ANN and probabilistic classification [10], as far as we know.

Acknowledgements

The authors wish to acknowledge associate professor Alvhild Alette Bjørkum, Faculty of Engineering, Biomedical Laboratory Sciences Program, Bergen University College for letting us know about this new platform of proteomic pattern analysis, based on MS data, and how it can be used to diagnose early-stage cancer. Without her introduction to this field this paper would not have been possible.

References

[1] Aebersold R., Mann,M. Mass spectrometry based on proteomics. Nature 422, 2003.

[2] Burges, C.J.C A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 2(2), 1998.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for Support Vector Machines. available online at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[4] Kristensen,T., Patel,R. Classification of Eukaryotic and Prokaryotic Cells by a Backpropagtion network. In Proceedings of I.E.E.E International Joint Conference on Neural Computing, IJCNN 2003, Portland, Oregon, USA.

[5] Kristensen, T. Prototypes of ANN Biomedical Pattern Recognition Systems. In Proceedings of IASTED International Conference on Simulation and Modelling (ASM 2002) Crete, Greece, 2002.

[6] Liotta, A.L., Ferrari,M., Petricoin,E. Clinical proteomics written in blood. Nature 2003, 425:905.

[7] Mitchell,T.M. Machine Learning. International Editions 1997. McGraww-Hill Companies, 1997.

[8] NIH and FSA clinical Proteomics Program Databank, 2002. http://www.clinicalproteomics.steem.com

[9] Petricoin,E., Liotta,A.L.. Seldi-tof-based serum proteomic pattern diagnostics for early detection of cancer. Current Opinion in Biotechnology 2004, 15: 24-30.

[10] Lilian,R.H., Farid,H., Donald,B.R. Probabilistic Disease Classification of Expression-Dependent Proteomic Data from Mass Spectrometry of Human Serum, Journal of Computational Biology Volume 10, Number 6, 2003

[11] Vapnik,V.N. Statistical Learning Theory. Wiley, New York, 1998.

[12] Vapnik,V.N. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10,Sep 1999.