LETTER

A Filter Method for Feature Selection for SELDI-TOF Mass Spectrum

Trung-Nghia VU[†a)], ***Nonmember and*** **Syng-Yup OHN**[†b)], ***Member***

SUMMARY We propose a new filter method for feature selection for SELDI-TOF mass spectrum datasets. In the method, a new relevance index was defined to represent the goodness of a feature by considering the distribution of samples based on the counts. The relevance index can be used to obtain the feature sets for classification. Our method can be applied to mass spectrum datasets with extremely high dimensions and process the clinical datasets with practical sizes in acceptable calculation time since it is based on simple counting of samples. The new method was applied to the three public mass spectrum datasets and showed better or comparable results than conventional filter methods.

key words: *feature selection, filter, high dimensions, mass spectrum*

1. Introduction

SELDI-TOF MS (surface-enhanced laser desorption/ionization time-of-flight mass spectrum) is frequently used to analyze and find the differences in the proteomic patterns between the serums from patients with disease and healthy people [1], [2]. A proteomic spectrum from the SELDI method consists of tens to hundreds of thousands m/z values, each of which represents the proteomic amplitude for proteins with corresponding m/z (mass-to-charge) ratio. Since it also involves a high degree of noise, feature selection techniques with high speed and good accuracy are necessary to reduce the dimension of spectrum data to find the differences of proteomic spectrum patterns. Furthermore, it is generally impractical to apply classification algorithms to such a high dimensional spectrum data with more than 10,000 features because of calculation time and memory problems.

There have been many research efforts to apply the various types of feature selection techniques to find the differences of spectrum patterns. In [3], Petricoin et al. combined genetic algorithms and self-organising clustering methods to discover the patterns in SELDI-TOF proteomic spectra that can distinguish between serum samples from healthy women and those from women with ovarian cancer. Authors in [4] applied PCA (Principle Component Analysis) for decreasing dimensions and LDA (Linear Discriminant Analysis) with nearest centroid classifier for classification to deal with the same data in [3]. Researchers in [5], [6] used the area under the curve (AUC) criterion for feature selection and decision tree combined with boosting methods for classification in prostate cancer study. In [7], the author also used nearest centroid classifier coupled with feature selection algorithms applying on five popular cancer datasets. Feature selection methods using filter approach, which have advantages of speed, are also used to deal with mass spectrum data. In [8], information gain, GINI, and F-test are used for feature ranking with 4-class classification problem for proteome datasets. They reported GINI and entropy seemed generally give better results than F-test. In [9], researchers used T-test to rank features and chose 15 and 25 top-ranked features for classification. In [7], author applied several conventional filter methods such as T-test, P-test and KS-test for feature ranking coupled with nearest centroid classifier in a three-fold cross-validation procedure. In [10], [11], RELIEF was used in the first step to reduce the high dimension of feature space into a lower dimension.

In this paper, we propose a new filter method for feature selection for SELDI-TOF MS datasets. Firstly, the relevance of each feature for a sample is evaluated by how well the feature values of samples separates the class the sample belongs to and the other classes. Then, the relevancies for all the samples are summed up to obtain the relevance of a feature. Since the computation of relevance only consists of sorting and simple counting, the method is fast and has low computational complexity. We applied the new technique to public datasets from NCI/CCR and FDA/CBER Clinical Proteomics Program Databank: Ovarian 4-3-02, Ovarian 8-7-02 and Prostate cancer datasets [12]. The results from the experiments show performances comparable to and sometimes better than those of conventional methods such as T-Test, signal-to-noise ratio and Pearson correlation coefficient.

2. Method

In this section, we define relevance index of a feature representing how well a feature separates a set of samples into predefined classes.

2.1 Relevance Index

Suppose we are given a dataset $D = (X, Y)$ with n samples each with d features and a class label, where X is the set of feature vectors consisting of n numeric vectors in d dimensions and Y is the set of n class labels each having a value from $\{1..k\}$ representing the class a sample belongs

Manuscript received June 27, 2008.
Manuscript revised October 10, 2008.

[†]The authors are with the Department of Computer Engineering, Korea Aerospace University, South Korea.
a) E-mail: nghiavtr@kau.ac.kr
b) E-mail: syohn@kau.ac.kr

DOI: 10.1587/transinf.E92.D.346

to. Furthermore, we suppose $x_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{id})$ is a feature vector of a sample belonging to class y_i where x_{ij} represents jth feature value of x_i.

We define the relevance index of jth feature for class C by sample x_i denoted as $RV_j(x_i \mid C)$ in Formula (1).

$$RV_j(x_i \mid C) = \begin{cases} \frac{|numGreater(x_{ij}) - numLess(x_{ij})|}{n^C} & \text{if } y_i \neq C \\ 0 & \text{if } y_i = C \end{cases} \quad (1)$$

where n^C is the number of samples in the class C, and $numGreater(x_{ij})/numLess(x_{ij})$ denotes the number of samples in class C with jth feature value greater/less than x_{ij}.

A relevance index in Formula (1) has a value in the range of $[0, 1]$. The value of 1 represents that sample x_i has the jth feature value either greater or less than those of all the samples in class C. The value of 0 represents that the sample x_i is ranked the middle by jth feature value among the samples in class C. It can be regarded that as the value is closer to 1, the better the sample x_i separating from class C. Thus, the relevance index represents how well the feature helps a sample separate from class C.

Relevance index of jth feature with a given class C by all samples denoted as $RV_j(X \mid C)$ is defined in Formula (2):

$$RV_j(X \mid C) = \frac{\sum_{x_i \in X} RV_j(x_i \mid C)}{n - n^C} \quad (2)$$

The numerator in Formula (2) sums up the relevance indices by all the samples, and the denominator in Formula (2) normalizes the summation in the numerator. Since the relevance index of samples in class C are always zero, the number of samples in class C does not count in the normalization factor.

Sorting all the features by their relevance index values as in Formula (2) in a decreasing order, we rank every feature by its importance how much a feature affects the classification of samples into two groups — a group of class C samples and a group of the samples belonging to one of other classes.

In case of a two-class classification problem, the list of features sorted by their index values can be used to select the optimal feature sets since the relevance index of one class is analogous to that of the other class.

$$RV_j(X \mid D) = \frac{\sum_{C=1..k} RV_j(X \mid C)}{k} \quad (3)$$

where k is the number of classes.

The relevance index in Formula (3) is the average of the relevance values of the feature for each class in Formula (2). The index value represents how well a feature separate the samples into their classes overall. As the index value is closer to 1, the samples from a class tend to form a cluster not overlapping with the samples from other classes in terms of a feature value.

2.2 Analysis of Computational Complexity

It is not difficult to see that the computational complexity of Formula (1) are $O(n)$. We need to calculate a index for every feature, and the total computational cost to get relevance indices of all the features is $O(dn^2)$. However, the computation of the relevance value of jth feature for all the samples in $O(n \log n)$ instead of $O(n^2)$ can be completed as follows. The jth feature values are sorted in decreasing order first, and the sorted list is scanned from top to bottom to count the number of samples with the values greater (or less) than x_{ij}. Thus it takes $O(dn \log n)$ overall to complete the calculation of the indices of d features.

3. Experiments

We tested the new method with three public data sets: Ovarian 4-3-02, Ovarian 8-7-02 and Prostate of NCI/CCR in FDA/CBER Clinical Proteomics Program Databank [12].

Ovarian 4-3-02: This ovarian dataset is produced by using WCX2 protein array. It consists of 216 samples, with 100 healthy, 16 benign and 100 cancer patterns. Each pattern consists of 15,154 features (m/z values).

Ovarian 8-7-02: The spectrum was collected from WCX2 chip. The dataset consists of 253 samples including 91 control and 162 ovarian cancer patterns. Each pattern consists of 15,156 features (m/z values).

Prostate: This dataset was collected using H4 protein chip and consists of four classes representing a healthy control group, patients with benign conditions and elevated PSA value, and two stages of prostate cancer. In this experiment, we combined patterns of the first two groups into a healthy class (253 samples) and the rest of patterns (69 samples) formed the cancer class. Each sample consists of 15,154 features (m/z values).

We applied three relevance indices from the formulas (2) and (3). All the features from a data set was ranked by their relevance indices, and we selected fixed numbers (5, 10, 15, 20 and 30) of top ranked features to choose the optimal feature sets. A radial SVM was used to measure the performances of the optimal feature sets. Ten rounds of 10-fold cross validation were performed on each experiment: the training set consisting of 90% of samples was used to obtain the optimal feature set from the relevance index and build a classification model, and the testing set consisting of remaining 10% of samples were used to measure the classification accuracies from the model.

To compare the performance of our method, we used conventional filter methods including T-test [14], linear correlation coefficient of Pearson [14], and signal-to-noise ratio (P-test) [7]. The test results are shown in Table 1. In the table, OS1_1/ OS1_2 represent the result of the binary classification case based on Formula (2) in which class C is cancer/normal class. OS2 represents the result from the case using Formula (3). Our method shows the better or comparable performances than the other methods.

In case of computation time, our methods took 64 seconds for Ovarian 8-7-02 dataset with 15,156 features and 253 samples while T-test took 21 seconds. We used R environment and MS Windows XP running on Intel Core 2

Table 1 Balanced accuracies (BACC, the average of sensitivity and specificity) of methods for overall datasets. Value in parentheses presents standard deviation values. Bold figures show the highest results.

	Ovarian 4-3-02	**Ovarian 8-7-02**	**Prostate**
T-Test	0.889(0.094)	0.987(0.029)	0.864(0.120)
Pearson -Test	0.889(0.094)	0.988(0.027)	0.899(0.101)
P-Test	0.886(0.094)	0.988(0.027)	0.859(0.125)
OS1_1	**0.906(0.089)**	0.989(0.027)	0.858(0.107)
OS1_2	0.898(0.090)	0.990(0.024)	**0.903(0.087)**
OS2	0.899(0.091)	**0.991(0.021)**	0.886(0.096)

Duo CPU with 2.0 GHz clock speed and 2 GB RAM. The dataset is used frequently in the researches on feature selection and classification using proteomic spectrum data [3], [7]. Although ours takes longer computation time than the conventional methods, it can process the dataset with the reasonable number of samples within acceptable calculation time, which is large enough to extract the set of relevant features for classification and other tasks from.

4. Conclusions

We proposed a new filter method for feature selection and its applications for SELDI-TOF mass spectrum datasets. Firstly, the relevance of each feature for a sample is evaluated by how well the feature values of samples separates the class the sample belongs to and the other classes. Then, the relevancies for all the samples are summed up to obtain the relevance of a feature. We applied the new technique to public datasets from NCI/CCR and FDA/CBER Clinical Proteomics Program Databank. The results from the experiments show performances comparable to and sometimes better than those of conventional methods and acceptable calculation time on the clinical datasets with practical sizes. Thus, our method is effective to extract the set of relevant features from SELDI-TOF mass spectrum data for classification of normal and cancer samples.

Acknowledgments

This research was supported by the Internet Information Retrieval Research Center (IRC) in Korea Aerospace University. IRC is a Regional Research Center of Gyeonggi Province, designated by ITEP and Ministry of Knowledge Economy.

References

[1] J.D. Wulfkuhle, L.A. Liotta, and E.F. Petricoin, “Proteomic applications for the early detection of cancer,” Nat. Rev. Cancer, vol.3, pp.267–275, 2003.

[2] J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang, and D.W. Chan, “Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer,” Clin. Chem., vol.48, pp.1296–1304, 2002.

[3] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, and L.A. Liotta, “Use of proteomic patterns in serum to identify ovarian cancer,” Lancet, vol.359, pp.572–577, 2002.

[4] R.H. Lilien, H. Farid, and B.R. Donald, “Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum,” J. Comput. Biol., vol.10, pp.925–946, 2003.

[5] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng, and G.L. Wright, “Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men,” Cancer Res., vol.62, pp.3609–3614, 2002.

[6] Y. Qu, B.L. Adam, Y. Yasui, M.D. Ward, L.H. Cazares, P.F. Schellhammer, Z. Feng, O.J. Semmes, and G.L. Wright, “Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients,” Clin. Chem., vol.48, pp.1835–1843, 2002.

[7] I. Levner, “Feature selection and nearest centroid classification for protein mass spectrometry,” BMC Bioinformatics, 6:68, 2005.

[8] L.E. Peterson, R.C. Hoogeveen, H.J. Pownall, and J.D. Morrisett, “Classification analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles for prostate cancer,” IJCNN’06-International Joint Conf. on Neural Networks, pp.3828–3835, July 2006.

[9] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, “Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data,” Bioinformatics, vol.19, pp.1636–1643, 2003.

[10] C. Plant, M. Osl, B. Tilg, and C. Baumgartner, “Feature selection on high throughput SELDI-TOF mass-spectrometry data for identifying biomarker candidates in ovarian and prostate cancer,” Proc. Sixth IEEE International Conf. on Data Mining – Workshops, pp.174–179, 2006.

[11] E. Marchiori, C.R. Jimenez, M. West-Nielsen, and N.H.H. Heegaard, “Robust SVM-based biomarker selection with noisy mass spectrometric proteomic data,” EvoWorkshops, pp.79–90, 2006.

[12] http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

[13] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, 2000.

[14] W. Duch, “Filter methods,” in Feature Extraction: Foundations and Applications, ed. I. Guyon, S. Gunn, M. Nikravesh and L. Zadeh, pp.89–118, Studies in Fuzziness and Soft Computing, Physica-Verlag, Springer, 2006.