Classification of Mass Spectrometry Proteomics Data: Manifold and Supervised Distance Metric Learning

Qingzhong Liu1,2, Andrew H. Sung1,2, Bernardete M. Ribeiro3*, Mengyu Qiao1

1Department of Computer Science
2Institute for Complex Additive Systems Analysis
New Mexico Tech, Socorro, NM 87801, USA
{liu, sung, mqiao}@cs.nmt.edu
3Department of Informatics Engineering
University of Coimbra, Coimbra, Portugal
bribeiro@dei.uc.pt corresponding author

ABSTRACT
Mass spectrometry has become a widely used measurement in proteomics research. High dimensionality of features and small dataset are two major challenges in mass spectrum data analysis. To address the high dimensionality issue, we study feature extraction and feature selection. A well-known approach is to detect peak values and apply support vector machine recursive feature elimination (SVMRFE) to choose feature sets for classification. In this paper, we successfully apply a distance metric learning to classify proteomics mass spectrometry data. Experimental results show that distance metric learning can successfully be applied to the classification of proteomics data and the results are comparable to the best results by applying SVM to the feature sets chosen with the use of SVMRFE. Our results also show the promising potential of manifold learning in feature reduction. Additionally, the results also indicate that peak detection may not be the optimal choice in preprocessing proteomics data.

Keywords
Proteomics, mass spectrum, feature selection, classification, distance metric learning

1. INTRODUCTION
Mass spectrometry (MS) establishes an effective framework for biomedical diagnosis and protein identification [1].

The mass spectrum data mining usually contains four steps, preprocessing, feature extraction, feature selection and classification. Williams et al [2] proposed a robust algorithm for computing the baseline correction of MALDI-MS spectra. On the other hand, the electronic noise is generated from the electronic instrument, and it is usually randomly distributed in the spectra. Chen et al [3] designed a wavelet-based denoising, which applies wavelet transformation and removes the certain amount of lower value wavelet coefficients. The denoised data are normalized to provide a systematic representation of the spectra. The next crucial step is to extract features from the spectra, and then form the initial complete feature set. The simplest way is to extract every data point as a discriminative feature, and generate a huge feature set including more than 15000 features [4, 5]. On the other hand, a more elaborated algorithm for peak detection and alignment performs a more aggressive feature extraction [6, 7, 8]. “The curse of dimensionality” in mass spectrometry data requires a powerful feature selection algorithm to choose the discriminative feature subset. A popular method to deal with the feature selection is to apply support vector machine recursive feature elimination (SVMRFE) and select a small subset of peaks as input variables for the classification [12, 13, 15].

In this paper, we apply a distance metric learning to the classification of proteomics mass spectrometry data and the results are equivalent to the best results by applying SVM to the feature sets chosen with the use of SVMRFE. We also compare the testing results on the reduced feature sets with the use of a manifold learning. The remainder is organized as follows. Preprocessing is discussed in section 2; a distance metric learning for Large Margin Nearest Neighbor (LMNN) classification, proposed by Weinberger etc. is described in section 3; experimental results are presented in section 4; conclusions follow.

2. PREPROCESSING
The preprocessing includes spectra resampling, wavelet denoising, baseline correction, normalization, peak detection and alignment.

2.1 Spectra resampling and wavelet denoising
The mass spectrum data is in a discrete format and the intervals are not equal in the whole spectrum. For high-resolution data, the high-frequency noise and redundant data points harm the quality of the dataset. So, we have to set the common low-frequent mass value to every sample spectrum in order to give a unified representation. By using spline interpolation, we resample the data and confine the interval to a unified size. Before resampling, the sample spectrum has a little variation from the true spectrum. The data is resampled to a standard discrete data which could be analyzed in frequency domain. The electrical noise is
generated during the mass spectrum acquisition by the instrument and it is almost random distributed noise. The next step is using discrete wavelet transform to eliminate the electrical noise. Then, we apply polynomial filter of second order to smooth the signal and improve data quality.

2.2 Baseline correction and normalization

The baseline effect is introduced by the chemical contamination and changes the true protein distribution. To minimize the chemical noise, the baseline should be subtracted from the spectrum. In order to obtain the baseline, the local minima should be computed by assigning an appropriate window size. Then, we use spline interpolation to fit the baseline. In order to compare sample spectra, we also need to normalize the data.

2.3 Peak detection and alignment

The final feature acquisition step of MS preprocessing is to spot the peak position and its magnitude. In our mass spectrum experiment, the peak detection method proposed by Coombes et al. [10] is performed on mean spectrum rather than individual spectra, and we used the ad hoc method based on signal to noise ratio to select the large peaks [6]. Performing peak detection on mean spectrum reduces the false peak generated by random noise and identifies some undetectable peaks in individual spectrum.

3. DISTANCE METRIC LEARNING

Many machine learning algorithms, for instance, K-Nearest Neighbor (KNN) classifier, rely on the distance metric for the input data patterns for (supervised or unsupervised) learning. With the given class labels for training samples, supervised distance metric learning can be divided into global distance metric learning [19] and local distance metric learning [20-24].

For unsupervised distance metric learning or manifold learning, the main idea is to learn an underlying low-dimensional manifold where geometric relationship between most of the observed data is preserved.

Inspired by the work on neighborhood component analysis [22] and metric learning by energy-based models [25], Weinberger et al. proposed a distance metric learning for Large Margin Nearest Neighbor classification (LMNN). Specifically, the Mahalanobis distance is optimized with the goal that k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin [26].

4. MATERIALS AND EXPERIMENTS

The following two mass spectrometry datasets have been tested in our experiment.

1. High resolution time-of-flight (TOF) mass spectrometry (MS) proteomics data set from surface-enhanced laser/desorption ionization (SELDI) ProteinChip arrays on 121 ovarian cancer cases and 95 controls.

http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp

2. The breast cancer QC SELDI spectra data set was studied by Pusztai et al. [27]. Here we utilized data of 57 controls and 51 cases. The data set may be downloaded at http://bioinformatics.mdanderson.org/Supplements/Datasets

We process the dataset according to the methods described in section 2 for peak detection. And then apply LMNN to the detected peak spectra data. Since SVMRFE is a classical feature selection based on the weights of the support vectors [12] and is widely used in the classification of proteomics data with the use of SVM, we also compare the results by using LMNN with Euclidean distance, Mahalanobis distance, and energy-based classification [26] and SVM combining with SVMRFE on the detected peak data. In each experiment, 80% samples are randomly chosen for training and the remaining 20% samples are tested. We repeated the experiments 10 - 100 times and compared the average testing results.

We also apply the manifold learning to the features chosen by peak-detection and the features filtered by rank-sum test without peak-detection, and obtain the reduced features that are mapped from high-dimension to low-dimension, then we apply a support vector machine and KNN to the reduced feature sets and compare the testing results.

4.1 Peak detection with SVMRFE

Figure 1 lists the average testing accuracy by applying SVM to SVMRFE on ovarian and breast cancer data sets, respectively, where data were preprocessed with the use of peak-detection algorithms before the use of SVMRFE.
Table 1 below lists the average testing accuracy by applying LMNN classifiers to the peak data sets.

Table 1 Average testing accuracy with LMNN classifiers

<table>
<thead>
<tr>
<th>Data set</th>
<th>Classifier</th>
<th>Testing accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian cancer</td>
<td>LMNN_energy</td>
<td>99.3%</td>
</tr>
<tr>
<td></td>
<td>LMNN_Eucliden</td>
<td>84.6%</td>
</tr>
<tr>
<td></td>
<td>LMNN_Mahalanobis</td>
<td>99.0%</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>LMNN_energy</td>
<td>81.8%</td>
</tr>
<tr>
<td></td>
<td>LMNN_Eucliden</td>
<td>84.6%</td>
</tr>
<tr>
<td></td>
<td>LMNN_Mahalanobis</td>
<td>81.7%</td>
</tr>
</tbody>
</table>

Table 2 lists the best averaging testing accuracy by applying SVM to feature sets (by SVMRFE) from the peak data sets.

Table 2. Best average testing accuracy of applying SVM to feature sets ranked by SVMRFE

<table>
<thead>
<tr>
<th>Data set</th>
<th>Testing accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian cancer</td>
<td>98.0%</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>89.5%</td>
</tr>
</tbody>
</table>

Comparing the results listed above, we can see that, although the testing results in table 1 are not as good as the best result by applying SVM to SVMRFE feature sets for breast cancer data set, the results by applying LMNN classifiers based on energy classification and Mahalanobis distance are better than the best result obtained by applying SVM to SVMRFE feature sets. It indicates that LMNN classifiers with energy classification and Mahalanobis metric are highly competitive for the classification of proteomics data. We also test the classification results by applying LMNN to the feature sets chosen by SVMRFE, and compare the best testing result and the least number of the features corresponding to the best result in each experiment against the results with the use of SVM, shown in table 3. Experimental results indicate comparable results by applying LMNN and support vector machines.

Table 3. The highest testing accuracy and the least number of the features corresponding to the highest testing results

<table>
<thead>
<tr>
<th>Data set</th>
<th>Classifier</th>
<th>Testing accuracy</th>
<th>Number of corresponding features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ovarian cancer</td>
<td>LMNN_energy</td>
<td>99.5%</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>LMNN_Eucliden</td>
<td>98.4%</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>LMNN_Mahalanobis</td>
<td>99.5%</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>99.3%</td>
<td>11</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>LMNN_energy</td>
<td>92.6%</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>LMNN_Eucliden</td>
<td>90%</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>LMNN_Mahalanobis</td>
<td>92.5%</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>94%</td>
<td>68</td>
</tr>
</tbody>
</table>

4.2 Results of applying manifold learning for feature reduction with/without peak detection

We also preprocess the data by using peak-detection algorithms (method one) and filtered the data sets by using rank sum test without peak-detection (method two); then we apply manifold learning, a locally linear embedding (LLE) proposed in [29] to reduce the features, finally we employ a SVM and a KNN to classify the data sets. Figure 2 gives the comparison of the testing accuracy of these two methods, with and without peak-detection on the ovarian data. Figure 3 plots the testing results on breast cancer data.

![Fig. 2. The testing results of the LLE reduce feature sets on peak-detection and rank sum test, ovarian cancer data.](image)

![Fig. 3. LLE reduce feature sets on peak-detection and rank sum test, breast cancer data.](image)
These results seem to imply that peak detection cannot include all useful features, i.e. some useful features may not be located at the peak locations. Further study is needed. Also, compare with the results of SVMRFE, the results indicate that unsupervised metric or manifold learning holds good promise in dealing with dimension reduction.

5. CONCLUSIONS

In this paper, we compared a supervised distance metric learning, large margin nearest neighbor classifier and SVM for classification of mass spectrometry proteomics data. Experiments give good results of applying distance metric learning to proteomics data, comparable to the results by applying SVM. Preliminary results also show the potential of manifold learning in feature reduction, though the present results are not as comparable to those of SVM. Our experimental results also indicate that peak detection may not be the optimal choice in preprocessing proteomics data.

6. ACKNOWLEDGEMENTS

Support for this research received from ICASA, a division of New Mexico Tech, is gratefully acknowledged.

7. REFERENCES