31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Model based clustering for tandem mass spectrum quality assessment

Jiarui Ding, Jinhong Shi and Fang-Xiang Wu*, *Member, IEEE*

***Abstract*— Several computational methods have been proposed to assess the quality of tandem mass spectra. These methods range from supervised to unsupervised algorithms, discriminative to generative models. Unsupervised learning algorithms for tandem mass spectra are not probabilistic model based and they don't provide probabilities for spectra quality assessment. In this study, the distribution of high quality spectra and poor quality spectra are modeled by a mixture of Gaussian distributions. The Expectation Maximization (EM) algorithm is used to estimate the parameters of the Gaussian mixture model. A spectrum is assigned to the high quality or poor quality cluster according to its posterior probability. Experiments are conducted on two datasets: ISB and TOV. The results show about 57.64% and 66.38% of poor quality spectra can be removed without losing more than 10% of high quality spectra for the two spectral datasets, respectively. This indicates clustering as an exploratory data analysis tool is valuable for the quality assessment of tandem mass spectra without using a pre-labeled training dataset.**

I. INTRODUCTION

Automatic quality assessment of tandem mass spectra is an important module in the peptide identification pipeline. Spectral quality assessment can be used to discover false negatives which are identifiable spectra but misidentified [1] and eliminate false positives which are unidentifiable spectra but also misidentified by some peptide identification algorithms [2]. In addition, it can help to find post-translational peptides [3]. Finally, the common use of quality assessment algorithm is as a pre-filer to filter out the unidentifiable spectra before peptide identifications [4]. Because mass spectrometry is a high-throughout technology, computational algorithms for automatic quality assessment are needed to speedup the analysis of tandem mass spectra.

In the past, several supervised machine learning algorithms have been proposed to assess the quality of tandem mass spectra [3]. For supervised machine learning, a labeled training dataset is needed to train a classifier, and the trained classifier is used to classify spectra as high quality or poor quality. Ideally, the spectra of the training set should be identified by several peptide identification algorithms and manually validated, i.e., the set should be correctly labeled without or with very few falsely labeled spectra. However, such spectral data sets are hard to obtain in most cases. Worse still, tandem mass spectrometers may produce different spectra even for the same peptide under different experimental conditions. Therefore, the training and testing spectra may not come from the same probability distribution and the trained classifier may fail to discriminate poor quality spectra from high quality ones. The performance of classifiers can be improved by training a specific classifier for each experiment.

Besides, an alternative choice for quality assessment of tandem mass spectra is clustering algorithms which do not need a labeled training set. Recently, several clustering algorithms have been used for quality assessment of tandem mass spectra [5], [6]. However, these algorithms do not use a probabilistic model of the spectral feature data and thus do not provide posterior probabilities for the assignment of spectra to clusters. The probabilistic models give us a meaningful way to cluster data and we can easily measure the fitness of data to models. The posterior probabilities from a probabilistic model are very useful, for example, we can use them to make predictions [7].

In this paper, we use the probabilistic model based clustering algorithm to perform quality assessment of tandem mass spectra. In addition to the ease of analyzing spectra by assuming a probabilistic model, the model on one dataset can be used to initialize the EM algorithm to fit other datasets, and thus the algorithm provides an automatic method to assess the qualities of tandem mass spectra. The remainder of this paper is organized as follows. Section II introduces the model based clustering algorithm. In Section III, two datasets, the ISB and the TOV datasets, are used to investigate the performance of the algorithm. Section IV concludes this study and gives some directions for further improvement.

II. METHODS

According to the different definitions of clustering, the existing clustering methods can be classified as combinatorial, mode seeking algorithms and model based algorithms. Combinatorial algorithms do not assume a probability distribution on the data, and samples are assigned to clusters by optimizing an objective function [8]. The mode seeking methods take a nonparametric approach to find the modes of the probability density function of data, and a sample is assigned to its nearest mode [9]. The mode seeking methods may be good choices if the structure of data is very complex and can not be modeled by a simple parametric probability distribution. The probabilistic model based methods assume

*Corresponding author.

Manuscript received April 7, 2009. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

J. Ding is with the Department of Mechanical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 5A9, Canada. jiarui.ding@usask.ca

J. Shi is with the Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 5A9, Canada. jinhong.shi@usask.ca

F-X. Wu is with the Department of Mechanical Engineering and associated with Division of Biomedical Engineering, University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 5A9, Canada. fangxiang.wu@usask.ca

samples are independently and identically distributed from a predefined probability density function such as a mixture of Gaussian distributions. After inferring the unknown parameters, clustering is achieved by assigning samples to different Gaussian components [10]. When the assumed probabilistic distribution is correct, the model based algorithm can achieve good clustering results.

To use machine learning algorithms for automatic quality assessment of tandem mass spectra, each spectrum should be represented by a fixed length feature vector. In this study, we use the top 10 features selected by the $SVM\text{-}RFE$ algorithm as described in [11].

A. Model based clustering

After exploratory data analysis and from previous research [4], the distribution of high quality spectra and poor quality spectra can be modeled by a mixture of Gaussian distributions:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \tag{1}$$

where K is the number of mixture components and here $K = 2$; one component corresponds to high quality spectra while the other component corresponds to poor quality spectra. π_k is the mixture coefficient. $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$ is a Gaussian density function with its mean of μ_k and covariance matrix of Σ_k, and $\mathbf{x}$ is a feature vector.

For Gaussian mixture models, it is difficult to use the maximum likelihood method to estimate the parameters because there exists a summation over k that occurs in the log-likelihood function. However, we can introduce a latent variable $\mathbf{z}$ which is the label of $\mathbf{x}$. Here $\mathbf{z}$ is a K-dimensional latent variable. The value of the k-th element of $\mathbf{z}$ satisfies $\mathbf{z}_k \in \{0, 1\}$ and $\sum_{k=1}^{K} z_k = 1$. The distribution of $\mathbf{z}$ is specified by the mixture coefficients

$$p(z_k = 1) = \pi_k \tag{2}$$

The joint distribution of $\mathbf{x}$ and $\mathbf{z}$ is

$$p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k} \pi_k^{z_k} \tag{3}$$

The posterior probability of $\mathbf{z}$ given $\mathbf{x}$ is

$$p(\mathbf{z}|\mathbf{x}) = \frac{\prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k} \pi_k^{z_k}}{\sum_{\mathbf{z}_k} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k} \pi_k^{z_k}} \tag{4}$$

Note that only one k makes $z_k = 1$. Thus

$$p(z_k = 1|\mathbf{x}) = \frac{\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)\pi_k}{\sum_{k=1}^{K} \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)\pi_k} \tag{5}$$

Suppose that we are given data $\mathbf{X}$ which is an $N \times D$ matrix. The n-th row $\mathbf{x}_n^T$ is a feature vector which represents the quality of the n-th spectrum. The corresponding latent variable matrix is $\mathbf{Z}$, which is an $N \times K$ indicator matrix and the value of z_{nk} satisfies $z_{nk} \in \{0, 1\}$ and $\sum_{k=1}^{K} z_{nk} = 1$.

Given $\mathbf{X}$ and $\mathbf{Z}$, the likelihood function of μ, Σ, π becomes

$$p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)^{z_{nk}} \pi_k^{z_{nk}} \tag{6}$$

The log-likelihood function becomes

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (\ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) + \ln \pi_k) \tag{7}$$

Now suppose that we already know $\mu_k^i, \Sigma_k^i, \pi_k^i$, then the posterior distribution for z_{nk} is (E-step)

$$p(z_{nk} = 1|\mathbf{x}_n, \mu_k^i, \Sigma_k^i, \pi_k^i) = \frac{\pi_k^i \mathcal{N}(\mathbf{x}_n|\mu_k^i, \Sigma_k^i)}{\sum_{k=1}^{K} \pi_k^i \mathcal{N}(\mathbf{x}_n|\mu_k^i, \Sigma_k^i)} \tag{8}$$

Now compute the score function

$$\begin{aligned} Q &= \sum_{z_{nk}} \ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) p(z_{nk} = 1|\mathbf{x}_n, \mu_k^i, \Sigma_k^i, \pi_k^i) \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1|\mathbf{x}_n, \mu_k^i, \Sigma_k^i, \pi_k^i)(\ln \mathcal{N} \\ &\quad * (\mathbf{x}_n|\mu_k, \Sigma_k) + \ln \pi_k) \end{aligned} \tag{9}$$

Maximizing Q under the constraint of $\sum_{k=1}^{K} \pi_k = 1$ by the use of Lagrange multiplier, we get (M-step)

$$\begin{aligned} N_k &= \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{x}_n, \mu_k^i, \Sigma_k^i, \pi_k^i) \\ \pi_k^{i+1} &= \frac{N_k}{N} \end{aligned} \tag{10}$$

$$\mu_k^{i+1} = \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{x}_n, \mu_k^i, \Sigma_k^i, \pi_k^i) \mathbf{x}_n \tag{11}$$

$$\begin{aligned} \Sigma_k^{i+1} &= \frac{1}{N_k} \sum_{n=1}^{N} p(z_{nk} = 1|\mathbf{x}_n, \mu_k^i, \Sigma_k^i, \pi_k^i) \\ &\quad * (\mathbf{x}_n - \mu_k^{i+1})(\mathbf{x}_n - \mu_k^{i+1})^T \end{aligned} \tag{12}$$

Given initial values for π, μ and Σ, the EM algorithm alternates between the E-step and the M-step, and finally finds a local maximum of the incomplete likelihood function (integrate out $\mathbf{Z}$ in Equation (7)).

III. Results and discussion

A. Datasets and performance evaluation

In this study, the ISB and the TOV datasets are used to investigate the performance of the model based clustering algorithm. Below is a brief description of the two datasets.

(1) ISB dataset consists of $37,044$ tandem mass spectra from 18 control mixture proteins [12], and these spectra were searched using Sequest against a human protein database appended with sequences of the 18 proteins. 2772 spectra were determined to be correctly identified after manual validations. These data were also analyzed by InsPecT, and annotated another 820 possibly modified peptides [13]. These 3592 spectra were labeled as "high" quality in this study, and other spectra were labeled as "poor" quality.

TABLE I

THE CLUSTERING RESULTS OF THE EM ALGORITHM.

Experiments	ISB		TOV	
	AUC	TNR	AUC	TNR
1	0.7647	57.64%	0.8214	66.33%
2	0.7647	57.64%	0.8214	66.33%
3	0.7647	57.64%	0.8214	66.33%
4	0.7647	57.64%	0.8214	66.36%
5	0.7647	57.64%	0.8214	66.33%
6	0.7647	57.64%	0.8214	66.36%
7	0.7647	57.64%	0.8214	66.36%
8	0.7647	57.64%	**0.8592**	**58.32%**
9	0.7647	57.64%	0.8214	66.33%
10	0.7647	57.64%	0.8214	66.36%

(2) TOV dataset consists of $22,576$ tandem mass spectra, and these spectra were searched against a subset of the Uniref100 database (release 1.2, http://www.uniprot.org) containing $44,278$ human protein sequences using Sequest. 2197 spectra were determined to be correctly identified after validated by PeptideProphet [14] (PeptideProphet scores are equal or greater than 0.9). All these 2197 spectra were labeled as "high" quality in this study, and all the other spectra were labeled as "poor" quality.

To evaluate the performance of the EM algorithm, we reported true positive rates (TPR, the fraction of positives correctly classified as positives), true negative rates (TNR, the ratio of negatives correctly classified as negatives) and false positive rates (FPR, the fraction of negatives misclassified as positives). We also reported receiver operating characteristic (ROC) curves, which are a plot of TPR as a function of FPR. The area under the curve (AUC) was used for comparing classification results. $AUC = 1$ means perfect classification and 0.5 indicates random guess.

B. The clustering results of the EM algorithm

The EM algorithm has been run 10 times on ISB and TOV datasets. The clustering results are shown in Table I. The TNRs are calculated when TPRs are fixed at 90%. The proposed clustering algorithm can remove about 66.36% of poor quality spectra while losing only 10% of interpretable spectra for TOV dataset. While for the spectra of the ISB dataset, about 57.64% of poor quality spectra can be safely removed without losing more than 10% of high quality spectra.

Table II shows the clustering results for the threshold of zero, i.e., a spectrum is assigned to the cluster with the larger posterior probability. Even using this simple threshold, about 53.47% $(= 17853/(17853+15599))$ of poor quality spectra can be removed while losing only 6.26% of high quality spectra for the spectra of ISB dataset. For the spectra of TOV dataset, about 53.73% $(= 10949/(9430+10949))$ of poor quality spectra can be removed while losing only 3.41% $(= 75/(2122+75))$ of high quality spectra. In other words, more than 53% of poor quality spectra can be removed by using the zero threshold while very few of high quality spectra are lost.

TABLE II

THE DISTRIBUTION OF SPECTRA IN DIFFERENT CLUSTERS WITH THRESHOLD OF ZERO. FOR ISB DATASET, THE NUMBERS ARE THE AVERAGE OF THE 10 RUNS. FOR TOV DATASET, THE NUMBERS ARE THE AVERAGE OF 9 RUNS (EXCLUDING THE 8-*th* RUN)

Dataset	Predicted High Quality	Predicted Poor Quality
ISB		
High Quality	3367	225
Poor Quality	15599	17853
TOV		
High Quality	2122	75
Poor Quality	9430	10949

TABLE III

THE CLUSTERING CENTERS OF THE EM ALGORITHM. THE POTENTIAL SALIENT FEATURES ARE HIGHLIGHTED. THE MAJORITY OF SPECTRA IN CLUSTER 1 ARE HIGH QUALITY SPECTRA WHILE THOSE IN CLUSTER 2 ARE POOR QUALITY SPECTRA.

#	ISB		TOV	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
$\mathbf{B}_5$	**0.51**	**-0.54**	**0.60**	**-0.63**
F_7	-0.20	0.21	0.07	-0.07
$\mathbf{W}_1$	**0.54**	**-0.57**	**0.65**	**-0.69**
F_4	0.00	0.00	0.46	-0.48
B_3	0.04	-0.04	-0.31	0.33
$\mathbf{W}_4$	**0.49**	**-0.51**	**0.54**	**-0.57**
$\mathbf{W}_7$	**0.56**	**-0.59**	**0.67**	**-0.70**
$\hat{\mathbf{W}}_4$	**-0.84**	**0.89**	**-0.77**	**0.81**
F_5	-0.11	0.12	0.37	-0.39
$\mathbf{W}_{10}$	**0.53**	**-0.55**	**0.63**	**-0.66**

C. The salient features for EM algorithm

Irrelevant features generally have little power for clustering methods to discriminate poor quality spectra from high quality ones. For this reason, we want to find the discriminative features for the EM clustering algorithm and we call them salient features. Salient features can be found from the cluster centers of EM algorithm. Since each feature is normalized to have mean of 0 and variance of 1, the features with large absolute values between two cluster centers could be salient features for cluster analysis.

Table III lists the cluster centers from the 10 runs of the EM algorithm. For ISB dataset, the numbers are the average of the 10 runs. For TOV dataset, the numbers are the average of 9 runs (excluding the 8-*th* run). From the clustering centers of each dataset, some features have nearly the same values in both clusters while values for other potential salient features vary a lot. These potential salient features are highlighted in Table III.

Figure 1 plots the absolute values of feature differences between two cluster centers in descending order. For ISB dataset, from Figure 1 (a), the four features with small absolute values of feature difference may be discarded because their values are far smaller compared to other six features. For TOV dataset, the cluster center differences do not show a distinct partition line compared to those of ISB dataset but they show a similar trend of decrease. The EM algorithm has been applied to the dimension-reduced feature sets in

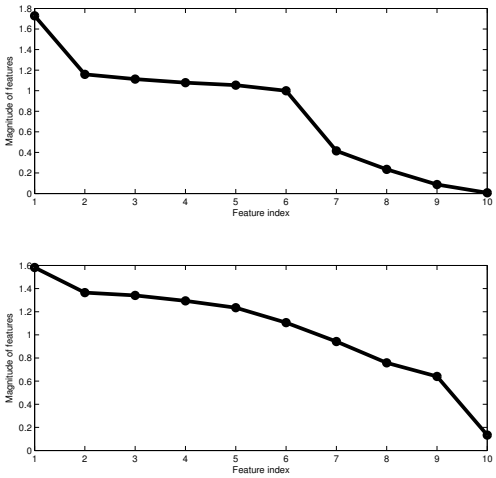

Fig. 1. Plot of the absolute values of clustering center difference in descending order for ISB dataset (a) and TOV dataset (b).

TABLE IV

THE CLUSTERING RESULTS WHEN USING SIX SALIENT FEATURES.

Experiments	ISB		TOV	
	AUC	TNR	AUC	TNR
1	0.7674	58.16%	0.8290	66.99%
2	0.7675	58.16%	0.8290	66.99%
3	0.7674	58.16%	0.8289	66.87%
4	0.7674	58.16%	0.8214	66.36%
5	0.7674	58.16%	0.8290	66.99%
6	0.7674	58.16%	0.8290	66.99%
7	0.7674	58.16%	0.8290	66.99%
8	0.7674	58.16%	0.8290	66.99%
9	0.7674	58.16%	0.8290	66.99%
10	0.7674	58.16%	0.8290	66.99%

which only the six features with large absolute values of cluster center difference are retained. From Table IV, it can be seen that the clustering results are better than those using the whole 10 features.

D. Determine the quality of spectra in each cluster

From the cluster centers, we can easily determine the spectra in which cluster are of high quality or poor quality. From the definition of B_5, W_1, W_4, W_7 and W_{10} [11], the high quality spectra should have larger value for these features than poor quality spectra do. In cluster 1, the values of these features are larger than those in cluster 2. $\hat{W}_4$ is the ratio of the number of peaks which have a relative intensity greater than 1% of the total intensity to the total number of peaks in a spectrum. For this feature, it is difficult to image whether the high quality spectra should have larger values or not. For this reason, we compute the mean for both high quality and poor quality spectra of this feature in ISB dataset, and the values are -0.77 and 0.08, respectively. Clearly, the high quality spectra have smaller values for this feature. For both ISB and TOV datasets, the value of $\hat{W}_4$ in cluster 1 is smaller than that in cluster 2.

IV. CONCLUSIONS AND FURTHER IMPROVEMENT

This study uses a mixture of Gaussian distributions to model the distribution of spectral feature data. Experimental results show the mixture of Gaussian distribution is a reasonable model of the spectral feature data.

From Table III, the cluster centers of the two datasets are similar although they are not exactly the same. Therefore, although we may not use the obtained probabilistic model on one dataset to model the distribution of spectral features in other datasets, the model can be used to initialize the EM algorithm to fit other datasets. We will further analyze the use of model based clustering algorithms for automatic quality assessment of tandem mass spectra in subsequent studies.

V. ACKNOWLEDGMENTS

We would like to thank Dr. Andrew Keller from Institute for Systems Biology for generously providing the ISB dataset and Dr. Guy G.Poirier from Laval University for providing the TOV dataset used in this paper. Jiarui Ding thanks the University of Saskatchewan for funding him through a graduate scholarship award.

REFERENCES

[1] A. Nesvizhskii, F. Roos, J. Grossmann, M. Vogelzang, J. Eddes, W. Gruissem, S. Baginsky, and R. Aebersold, "Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data," *Molecular & Cellular Proteomics*, vol. 5, no. 4, pp. 652–670, 2006.

[2] K. Flikka, L. Martens, J. Vandekerckhove, K. Gevaert, and I. Eidhammer, "Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering," *Proteomics*, vol. 6, no. 7, pp. 2086–2094, 2006.

[3] J. Wong, M. Sullivan, H. Cartwright, and G. Cagney, "msmsEval: tandem mass spectral quality assignment for high-throughput proteomics," *BMC Bioinformatics*, vol. 8, no. 1, 2007.

[4] F. Wu, P. Gagne, A. Droit, and G. Poirier, "Quality assessment of peptide tandem mass spectra," *BMC Bioinformatics*, vol. 9, no. suppl:6, p. S13, 2008.

[5] F. Wu, J. Ding, and G. Poirier, "An approach to assess peptide mass spectral quality without prior information," *International Journal of Functional Informatics and Personalised Medicine*, vol. 5, no. 2, pp. 140–155, 2008.

[6] J. Ding, J. Shi, and F. Wu, "Quality assessment of tandem mass spectra by using a weighted k-means," *Clinical Proteomics*, 2009, accepted.

[7] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[8] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. University of California Press, 1967, pp. 281–297.

[9] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[11] J. Ding, J. Shi, A. Zou, and F. Wu, "Feature selection for tandem mass spectrum quality assessment," in *IEEE International Conference on Bioinformatics and Biomedicine*, 2008, pp. 310–313.

[12] A. Keller, S. Purvine, A. Nesvizhskii, S. Stolyar, D. Goodlett, and E. Kolker, "Experimental protein mixture for validating tandem mass spectral analysis." *OMICS*, vol. 6, no. 2, pp. 207–12, 2002.

[13] S. Tanner, H. Shu, A. Frank, L. Wang, E. Zandi, M. Mumby, P. Pevzner, and V. Bafna, "Inspect: fast and accurate identification of post-translationally modified peptides from tandem mass spectra," *Anal. Chem.*, vol. 77, no. 14, pp. 4626–4639, 2005.

[14] A. Keller, A. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Anal. Chem.*, vol. 74, no. 20, pp. 5383–5392, 2002.