The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

The Imbalanced Problem in Mass-spectrometry Data Analysis

Hao-Hua Meng[1] Guo-Zheng Li[2] Rui-Sheng Wang[3]
Xing-Ming Zhao[4] Luonan Chen[4]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200072, China
[2]Department of Control Science and Engineering, Tongji University, Shanghai 201804, China
[3]School of Information, Renmin University of China, Beijing 100872, China
[4]Institute of System Biology, Shanghai University, Shanghai 200444, China

Abstract In many cases, protein mass-spectrometry data are imbalanced, i.e. the number of positive examples is much less than that of negative ones, which generally degrade the performance of classifiers used for protein recognition. Despite its importance, few works have been conducted to handle this problem. In this paper, we present a new method that utilizes the EasyEnsemble algorithm to cope with the imbalance problem in mass-spectrometry data. Furthermore, two feature selection algorithms, namely PREE (Prediction Risk based feature selection for EasyEnsemble) and PRIEE (Prediction Risk based feature selection for Individuals of EasyEnsemble), are proposed to select informative features and improve the performance of the EasyEnsemble classifier. Experimental results on three mass spectra data sets demonstrate that the proposed methods outperform two existing filter feature selection methods, which prove the effectiveness of the proposed methods.

Keywords Mass-spectrometry, Feature selection, Ensemble

1 Introduction

Protein mass-spectrometry is a potential technique for high-throughput disease classification and biomarker identification. Thereby, fast and accurate detection of diseases, such as early cancer detection, may revolutionize the field of medical diagnosis. Typically, serum samples are analyzed by a mass spectrometer, producing a high dimensional abundance histogram. Next, informative features are extracted from the high dimensional data and presented to a classifier. In turn, the classifier outputs a decision about the status of the patient with respect to a particular disease (e.g., healthy or diseased) [1, 6]. The nature of relatively high dimensionality but small sample size in mass-spectrometry data cause the known problem of 'curse of dimensionality'. Therefore, selecting a small number of discriminative features from thousands of features is essential for successful protein recognition [14].

Feature selection, a process of selecting a subset of features from the original ones, is frequently used as a preprocessing technique in data mining. It has been proved effective in reducing dimensionality, improving mining efficiency, increasing mining accuracy, and enhancing result comprehensibility [3, 5]. In the field of systems biology, the most

widely used procedures of feature selection are based on a score which is calculated for all features individually and features with the best scores are selected [13, 18, 10]. Feature selection procedures output a list of relevant features which may be experimentally analyzed by biologists. This method is often denoted as univariate feature selection (filter methods), whose advantages are its simplicity and interpretability [10]. Embedded feature selection has been proposed by Guyon et al. [4, 5], which has lower complexity than wrapper feature selection. It depends on the used classifiers, so it produces better performance for the used classifiers than filter feature selection.

Though feature selection helps to improve the performance of classifiers, imbalance of mass-spectrometry data sets reduces performance of the previously proposed methods [6]. Few works on unbalanced bio-medical data sets have been conducted, Yang et al. [13] proposed two evaluation scores for feature selection in imbalanced microarray data, while their experiments adopted prediction accuracy to valuate the methods. Since accuracy maybe fails to find the accuracy of minor positive sample, the experimental results is not confident enough. Li et al. [7] propose a novel algorithm PRIFEAB for an imbalanced drug activities data set, which used embedded feature selection with asymmetric bagging. Zhao et al. have also studied how to handle imbalanced problems in protein classification [15] and gene function prediction [16].

For the imbalance problem of data sets, many works [9, 19, 17] have been done in machine learning field. They mainly used over-sampling, under-sampling or mixture of over-sampling and under-sampling strategies, of which the EasyEnsemble classifier using under-sampling proposed by Liu et al. [9] achieved interesting results. Combining with the EasyEnsemble classifier [9], we propose embedded feature selection with an evaluation criterion prediction risk [8] for analysis of imbalanced mass-spectrometry data sets, where two algorithms PREE and PRIEE are proposed to perform feature selection for classification of imbalanced data sets. They will be compared with two filter methods GS1 and GS2 which are also designed to solve the imbalanced problem of data sets.

The remainder of this paper is arranged as follows: In Section 2, The Easyensemble algorithm is shortly introduced and then two novel algorithms PREE and PRIEE are presented in detail. In Section 3, benchmark data sets and experimental setting are described. In Section 4, comparative experiments on several imbalanced benchmark mass-spectrometry data set are described. At last, conclusions are given in Section 5.

2 Computational Methods

2.1 The EasyEnsemble Classifier

The EasyEnsemble classifier [9] is an under-sampling algorithm [11, 19], which independently samples several subsets from negative examples. For each subset, a classifier is built. All generated classifiers are then combined for the final decision by using Adaboost [2].

2.2 Embedded Feature Selection

Embedded feature selection has been proposed by Guyon et al. [4, 5]. Prediction risk is an embedded evaluation criterion, which evaluates the features by computing the change of training accuracy when the features are replaced by their mean values,

$$R_i = \mathrm{BAC} - \mathrm{BAC}(\bar{\boldsymbol{x}}^i) \tag{1}$$

where BAC means the BAC value by applying EasyEnsemble on the training set, and BAC($\bar{\boldsymbol{x}}^i$) means the BAC value on the training set with the ith feature replaced by its mean value. BAC is defined in subsection 3.3. The feature corresponding to the least R_i is removed, because its change causes the least difference and it is the least important one.

We propose embedded feature selection for the unbalanced data sets by using the prediction risk criterion. Two algorithms are proposed, one uses EasyEnsemble as a whole machine to evaluate features which is named as PREE (Prediction Risk based feature selection for EasyEnsemble). Another one uses AdaBoost weak learner in EasyEnsemble to evaluate features and select features for individuals of EasyEnsemble which is named as PRIEE (Prediction Risk based feature selection for Individuals of EasyEnsemble). PREE is summarized in Algorithm 1, and PRIEE is in Algorithm 2.

Algorithm 1 The PREE algorithm

Input: Training data set $S_r = \{(\mathbf{x}, \mathbf{y})\}$, Number of selected features P
Output: Ensemble model **N**

1: Begin
2: Train the ensemble model $\boldsymbol{N}$ on the training set S_r by using EasyEnsemble.
3: Calculate the BAC value on the training subset, and all the prediction risk values $\boldsymbol{R}$ by using Equation (1).
4: Rank $\boldsymbol{R}$ in the descending order, and select the top P features as the optimal feature subset.
5: Generate the optimal training subset $S_{r-optimal}$ from S_r according to the above optimal features.
5: Re-train the model $\boldsymbol{N}$ on the optimal training subset $S_{r-optimal}$.
6: End

3 Experimental Data Sets and Settings

3.1 Mass-spectrometry Data Sets

Three mass-spectrometry data sets are used [12], where two are for Ovarian Cancer studies and one is for Prostate Cancer studies. there are 15154 features (data points in the m/z axis) in the original data set. For efficient computation, we select every other feature and 7577 features are used. Information about these data sets are summarized in Table 1, where Size is the number of examples, the positive is used as minor class while the union of all other classes is used as major class, #min/#maj is the size of minor/major class, and Ratio is the size of major class divided by that of minor class.

3.2 Settings

Our proposed algorithms, i.e. PREE and PRIEE are compared with two existing class-imbalance feature selection methods, i.e. GS1 and GS2 [13]. GSi used

$$compact_i(p)/scatter(p), (i = 1, 2)$$

Algorithm 2 The PRIEE algorithm

Input: Training data set $S_r = \{(\mathbf{x}, \mathbf{y})\}$, Number of individuals T, Number of selected features P
Output: Ensemble model $\mathbf{N}$

1: Begin
2: **for** $k = 1 : T$ **do**
3: Generate a training subset S_{rk}^{-} from negative training set S_r^{-} by using the Bootstrap sampling technique, the size of S_{rk}^{-} is the same with that of S_r^{+}.
4: Train the individual model N_k on the training subset $S_{rk}^{-} \bigcup S_r^{+}$ by using AdaBoost with weak classifiers $h_{k,j}$ and corresponding weights $\alpha_{k,j}$, i.e.

$$N_k(x) = \mathrm{sgn}\left(\sum_{j=1}^{n_k} \alpha_{k,j} h_{k,j}(x) - \theta_k\right)$$

.
5: Calculate the BAC value on the training subset, and all the prediction risk values $\boldsymbol{R}$ using Equation (1).
6: Rank $\boldsymbol{R}$ in the descending order, and select the top P as the optimal feature subset.
7: Generate the optimal training subset $S_{rk-optimal}$ from S_{rk} according to the above optimal features.
7: Re-train the individual model N_k on the optimal training subset $S_{rk-optimal}$.
8: **end for**
9: Ensemble the obtained models $\mathbf{N}$ in the way like

$$\mathbf{N}(x) = \mathrm{sgn}\left(\sum_{k=1}^{T}\sum_{j=1}^{n_k} \alpha_{k,j} h_{k,j}(x) - \sum_{k=1}^{T} \theta_k\right)$$

10: End

Table 1: Experimental mass-spectrometry data sets

Dataset	Size	Feature	#min/#maj	Ratio
Ovarian0403	116	7577	16/100	6.25
Ovarian0807	253	7577	91/162	1.78
Prostate0703	322	7577	63/259	4.11

to evaluate the importance of genes, where $compact_i(p)$ represents intra-class variations of a data set, while $scatter(p)$ means inter-class variations.

In EasyEnsemble, 5 subsets are sampled, i.e. T is set to 5 during experiments, on each an ensemble containing 15 weak learners is trained. Thus, the final ensemble generated by EasyEnsemble also contains 15*5=75 weak learners. In all experiments, we use the

k-nearest neighbor as the weak learner with $k = 1$. To compare with EasyEnsemble, we also implement a normal ensemble, whose difference with EasyEnsemble is in Step 3 of Algorithm 1. Here normal ensemble generates a training subset S_{rk} from S_r with double size of S_r^+ instead of only generating S_{rk}^- from S_r^- like in EasyEnsemble, then N_k is trained on S_{rk}.

To compare the results of feature selection fairly, we use the 3-fold cross validation procedure. Using the top ranked genes selected by a feature selection method, together with their expression values in the training dataset, one can build an EasyEnsemble that will decide for each testing example the class it belongs to. Only the expression values for those selected genes in the testing example are used for such a decision making. This is a standard way to test the quality of those selected genes, to examine how well the resulting classifier performs. Note that testing examples are not included in the training dataset.

3.3 Learning and Performance Measurement

Detailed results of prediction accuracy

$$\text{ACC} = \frac{\text{\# correctly predicted examples}}{\text{\# whole examples}},$$

true positives ratio

$$\text{TPR} = \frac{\text{\# correctly predicted positive examples}}{\text{\# whole positive examples}}$$

and true negatives ratio

$$\text{TNR} = \frac{\text{\# correctly predicted negative examples}}{\text{\# whole negative examples}}$$

are presented, where #A means the number of A. TPR also names as sensitivity and TNR names as specificity. We also use balanced accuracy BAC = (TPR+TNR)/2. Since the class distribution of the used data set is skew, prediction accuracy (ACC) may be misleading. Therefore, besides ACC, the BAC, TPR and TNR values are averaged on the 3-fold cross validation method for the analysis of experimental results.

4 Results and Discussions

4.1 Computational Results

Table 2 lists the best results of different measures, i.e. BAC, ACC, TPR and TNR. Here EN/ALL and EA/ALL mean results are obtained by using normal ensemble or EasyEnsemble on all the features, GS1/*P*, GS2/*P*, PRIEE/*P* and PREE/*P* mean results are obtained by using EasyEnsemble on the data set with the optimal *P* features selected by using GS1, GS2, PRIEE or PREE respectively.

From Table 2, we can see that:

(1) Using the top 100 features, PRIEE and PREE obtain better results than that without feature selection in terms of BAC, ACC and TPR. PRIEE also obtains the best results among all the feature selection methods in terms of BAC, ACC and TPR.

Table 2: The best results (together with the number of selected features) obtained by using different feature selection methods

	EN/ALL	EA/ALL	GS1/*P*	GS2/*P*	PREE/*P*	PRIEE/*P*
			BAC			
Ovarian0403	0.526	0.706	0.550(8)	0.561(97)	0.730(70)	0.858(57)
Ovarian0807	0.958	0.733	0.807(44)	0.838(95)	0.931(78)	0.974(22)
Prostate0703	0.909	0.937	0.791(56)	0.800(84)	0.909(92)	0.945(74)
Average	0.798	0.792	0.716(36)	0.733(92)	0.857(80)	0.926(51)
			ACC			
Ovarian0403	0.862	0.810	0.810(1)	0.810(1)	0.914(39)	0.845(48)
Ovarian0807	0.640	0.735	0.818(44)	0.858(72)	0.937(78)	0.976(73)
Prostate0703	0.820	0.947	0.867(22)	0.873(72)	0.950(82)	0.960(74)
Average	0.774	0.831	0.832(22)	0.847(48)	0.934(66)	0.927(62)
			TPR			
Ovarian0403	0.063	0.563	0.250(8)	0.313(97)	0.500(70)	0.875(57)
Ovarian0807	0.923	0.727	0.769(44)	0.802(34)	0.912(78)	0.978(22)
Prostate0703	0.841	0.921	0.683(56)	0.683(84)	0.841(92)	0.937(43)
Average	0.609	0.737	0.567(36)	0.599(72)	0.751(80)	0.930(40)
			TNR			
Ovarian0403	0.990	0.850	0.930(1)	0.930(4)	1.000(39)	0.880(100)
Ovarian0807	0.994	0.740	0.883(92)	0.920(72)	0.953(97)	0.988(23)
Prostate0703	0.977	0.954	0.934(23)	0.931(3)	0.981(93)	0.970(74)
Average	0.987	0.848	0.916(38)	0.927(26)	0.978(76)	0.946(65)

(2) Embedded feature selection algorithms i.e. PRIEE and PREE perform much better than filters like GS1 and GS2 in terms of BAC, ACC and TPR.
(3) The ratio of minor class to major class is critical to the value of TPR by three feature selection methods, e.g. ratio of the Ovarian0403 data set is 6.25, GS1, GS2, PREE and EA, EN do not obtain satisfactory results on this data set in terms of TPR and BAC.
(4) GS2 performs better than GS1 does on all the data sets by using all the measures.

4.2 Discussions

As for the above experimental results, they are quite interesting. We give some explanations as below:

(1) PREE and PRIEE are two embedded feature selection methods, while GS1 and GS2 are two filter methods. Here we find PREE and PRIEE perform better than GS1 and GS2 in terms of BAC, which indicates embedded feature selection has more power to help to improve generalization performance than filter does when it is a classification task. We consider that embedded feature selection uses classifiers to evaluate features, so the selected top features can help improve performance of classification.
(2) PRIEE performs better than PREE in terms of BAC and TPR. This is because feature selection selects features for individuals in PRIEE, which not only improves generalization performance of individual weak learners, but also improves diversity

among weak learners. Therefore, it greatly improves performance of EasyEnsemble. While PREE only selects features for EasyEnsemble. Therefore, PRIEE obtains better performance than PREE.

(3) GS2 and GS1 are used as two filter feature selection methods, though the previous works [13] show GS2 has similar performance with GS1, it is not the case in our paper. There is a weight factor to overcome imbalance of data in GS2, not in GS1, which did not exhibit advantages in the previous work, because the previous work used prediction accuracy as the measure for performance, this can not accurately reflect the true intrinsic business. In our paper, we use BAC for imbalanced binary classification, GS2 shows its advantages are expected.

(4) From this paper, we find different measures give somewhat different results. For imbalanced problems, we consider BAC and TPR or sensitivity for classification of the minor positive examples, TNR or specificity for classification of major negatives. If we want a balanced accuracy, we prefer to BAC, a balanced result of TPR and TNR. ACC does not accurately reflect the ability of classifiers because the major class may cover it.

5 Conclusions

To address the imbalanced problem in analysis of mass-spectrometry data sets, we propose to apply EasyEnsemble and embedded feature selection. By Comparing with the existing algorithms GS1 and GS2, experimental results show that our proposed two algorithms PREE and PRIEE can greatly improve the prediction ability in terms of BAC and TPR. Since this is a protein classification problem, the positive sample is few but important, BAC and TPR are more proper than ACC to measure generalization performance of classifiers.

This work proposes embedded feature selection for imbalanced mass spectrometry data sets, with two novel algorithms based on EasyEnsemble. Extension of this paper includes testing the proposed algorithms with more imbalanced classifiers and improving efficiency of the process of feature selection.

Acknowledgments

This work was supported by the Natural Science Foundation of China under grant no. 20503015 and 60873129, the STCSM "Innovation Action Plan" Project of China under grant no. 07DZ19726, the Shanghai Rising-Star Program under grant no. 08QA14032 and Systems Biology Research Foundation of Shanghai University.

References

[1] R Aebersold and M Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.

[2] E Bauer and R Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1-2):105–139, 1999.

[3] A L Blum and P Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

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

[5] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[6] Ilya Levner. Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics*, 6:68, 2005.

[7] Guo-Zheng Li, Hao-Hua Meng, Wen-Cong Lu, Jack Y. Yang, and Mary Q. Yang. Asymmetric bagging and feature selection for activities prediction of drug molecules. *BMC Bioinformatics*, 9(Suppl 6):S7, 2008.

[8] Guo-Zheng Li, Jie Yang, Guo-Ping Liu, and Li Xue. Feature selection for multi-class problems using support vector machines. In *Lecture Notes in Artificial Intelligence 3173 (PRICAI2004)*, pages 292–300. Springer, 8 2004.

[9] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory under-sampling for class-imbalance learning. In *Proceedings of International Conference on Data Mining*, pages 965–969. IEEE Press, 2006.

[10] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[11] G. M. Weiss. Mining with rarity - problems and solutions: A unifying framework. *SIGKDD Explorations*, 6(1):7–19, 2004.

[12] Gordon Whiteley. Biomarker profiling, discovery and identification (low resolution SELDI-TOF datasets part), 2008.

[13] Kun Yang, Zhipeng Cai, Jianzhong Li, and Guohui Lin. A stable gene selection in microarray data analysis. *BMC Bioinformatics*, 7:228, 2006.

[14] Xuegong Zhang, Xin Lu, Qian Shi, Xiu-Qin Xu, Hon-Chiu E Leung, Lyndsay N Harris, James D Iglehart, Alexander Miron, Jun S Liu, and Wing H Wong. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, 7:197, 2006.

[15] Xing-Ming Zhao, Xin Li, Luonan Chen, and Kazuyuki Aihara. Protein classification with imbalanced data. *Proteins*, 70(4):1125–1132, 2007.

[16] Xing-Ming Zhao, Yong Wang, Luonan Chen, and Kazuyuki Aihara. Gene function prediction using labeled and unlabeled data. *BMC Bioinformatics*, 9:57, 2008.

[17] Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explorations*, 6(1):80–89, 2004.

[18] X Zhou and D P Tuck. MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics*, 23:1106–1114, 2006.

[19] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006.