Neurocomputing 105 (2013) 3–11

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Feature selection for high-dimensional imbalanced data

Liuzhi Yin [a], Yong Ge [b,*], Keli Xiao [b], Xuehua Wang [c], Xiaojun Quan [d]

[a] School of Management, University of Science and Technology of China, Hefei, China
[b] Rutgers Business School, Rutgers University, Newark, NJ, USA
[c] School of Management, Dalian University of Technology, Dalian, China
[d] Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong

ARTICLE INFO

Available online 27 October 2012

Keywords:
Feature selection
Imbalanced data
Hellinger distance
AUC
F-measure

ABSTRACT

Given its importance, the problem of classification in imbalanced data has attracted great attention in recent years. However, few efforts have been made to develop feature selection techniques for the classification of imbalanced data. This paper thus fills this critical void by introducing two approaches for the feature selection of high-dimensional imbalanced data. To this end, after introducing three traditional methods, we study and illustrate the challenges of feature selection in imbalanced data with Bayesian learning. Indeed, we reveal that the samples in the larger classes have a dominant influence on these feature selection methods. However, the samples in rare classes are essential for the learning performances of rare classes. Based on these observations, we provide a new feature selection approach based on class decomposition. Specifically, we partition the large classes into relatively smaller pseudo-subclasses and generate the pseudo-class labels accordingly. Feature selection is then performed on the new decomposed data for computing the goodness measurement of features. In addition, we also introduce a Hellinger distance-based method for feature selection. Hellinger distance is a measure of distribution divergence, which is strongly skew insensitive as the class prior information is not involved for computing the distance. Finally, we theoretically show the effectiveness of these two approaches with Bayesian learning on synthetic data. We also test and compare the performances of the proposed feature-selection methods on some real-world data sets. The experimental results show that both decomposition-based and Hellinger distance-based methods can outperform existing feature-selection methods with a clear margin on imbalanced data.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

An emerging critical challenge in data mining and machine learning research is to address so-called "imbalanced classes" in real-world data. The class imbalance problem refers to the issues that occur when a data set is dominated by a major class or classes which have significantly more instances than the other rare/small classes in the data. Typically, people have more interests in learning rare classes in the data. For example, in customer churn analysis, the churn class tends to be a rare class because usually there are a small number of customers who will switch from one service provider to another. However, it is important for companies to predict and prevent the customer turnover. Many other similar examples can be observed in real-world applications, such as risk management, web categorization, medical diagnosis/monitoring, and network intrusion analysis.

The class imbalance problem becomes even severe when the dimensionality of data is high. In this situation, feature selection usually becomes essential to the learning algorithm because high-dimensional data tends to decrease the efficiency of most learning algorithms. This is also widely known as the curse of dimensionality. While feature selection has been extensively studied, its importance and unprecedented problem for class imbalanced data is recently realized and there is increasing attention from the machine learning and data mining communities.

In this paper, we investigate feature selection issues in the imbalance situation. Specifically, after briefly introducing three traditional feature selection methods, we illustrate the inefficiencies of traditional methods for imbalanced data. Along this line, we provide a study with Bayesian learning to explore the challenges caused by imbalanced data for traditional feature selection methods. Indeed, for the imbalanced data, the traditional methods for computing the goodness of feature will be mainly influenced by the instances from the majority classes. This biased influence will lead to the selection of features which are not suitable for predicting rare class(es).

Instead, we provide a feature selection method based on class decomposition. This method first divides the large classes into relatively smaller pseudo-subclasses with relatively uniform sizes. Feature selection is then performed on the new decomposed data for computing the goodness measurement of features. In addition, we also introduce a Hellinger distance-based method

* Corresponding author.
E-mail address: yongge@pegasus.rutgers.edu (Y. Ge).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.04.039

for feature selection. Hellinger distance is a measure of distribution divergence, which is more tolerant to skewed class distributions, since the class prior information is not involved for computing the distance. Finally, we theoretically show the effectiveness of these two approaches with Bayesian learning on synthetic data. We also test and compare the performances of the proposed feature-selection methods on real-world data sets. The results show that both decomposition-based and Hellinger distance-based method outperform existing feature-selection methods with a clear margin on imbalanced data in terms of various performance metrics.

Related work. In the following, we briefly introduce the related work in two relevant areas: classification of imbalanced data and feature selection.

To deal with the challenges imposed by imbalanced class distributions, many learning algorithms have been proposed. For instance, the sampling based method is one of the simplest yet effective ones. The over-sampling scheme replicates the instances from the small classes to match the size of large classes [6]. In contrast, under-sampling cuts down the large class sizes to achieve a similar balanced effect [15]. Drummond and Holte [2] provided detailed comparisons on these two resampling schemes. Liu et al. [13] borrowed bagging and boosting ideas and built multiple classifier with majority class example in order to overcome the main deficiency of under-sampling, which is that many majority class examples are ignored. Another popular method is the cost-sensitive learning scheme which takes the cost matrix into consideration during model building and generates a model that has the lowest cost. Also, Margineantu et al. [8] examined various methods for incorporating cost information into the C4.5 learning algorithm. In addition, Joshi et al. [17] proposed PNrule, a two-phase rule induction algorithm, to handle for mining rare classes. Y. Tang et al. [24] incorporated different rebalance heuristics, including cost-sensitive learning, over-sampling and under-sampling in SVM modeling and introduced four SVM variations to tackle the imbalance learning problem. A survey about this topic can be found in [11].

In the literature, there are a large number of feature selection methods [31–34,27,26,10,28,18] which can be generally grouped into two categories: classifier-dependent and classifier-independent methods. Classifier-independent methods usually use some statistics to measure the prediction power of each variable, such as the relevance between variables and the class label. Correlation criteria are one typical method to measure the linear correlation between variable and label [10]. Fisher criterion score is also one linear correlation measure which has been used in feature selection [14]. But nonlinear correlation between variable and label cannot be measured with these two methods, while often this kind of nonlinear correlation is essential for building a prediction model. Mutual information is also a criterion commonly used in feature selection [25,7], which is able to capture the nonlinear correlation. But one weakness of mutual information is that high computation usually will be caused because of the estimation of distribution, especially for continuous variable. Also information gain is employed by measuring the number of bits of information obtained for category prediction by knowing the presence or absence of a variable. The same weakness happens for information gain as mutual information. In addition, χ^2 measures the lack of independence between variable and label and can be compared to the chi-square distribution with one degree of freedom to judge extremeness. But similar to mutual information and information gain, the computation of chi-square score has a quadratic complexity. In addition, some other classifier-independent methods, e.g. *t*-test method, are employed for feature selection. Furthermore, Chow et al. [4] proposed a new data distribution factor to select appropriate clusters, which combines the compactness and separation together with a newly introduced concept of singleton item. For classifier-dependent methods, typically they measure the prediction power of each variable by training and testing a learning model [35,29,28]. The goodness of variable is measured by the performance of classification. The performance of classification can be validated by various criteria that involve classification error, false positive rate, false negative rate, F-measure and ROC (receiver operating characteristic). Compared with classifier-independent methods, this kind of method obviously will cause more computation which actually depends on special classifier. For example, for SVM with RBF kernel, the training processing is very time-consuming. All these prior works on the above were developed for general feature selection problem and may not be suitable for imbalanced data.

Feature selection on imbalanced data is under explored in the literature, while there are some previous studies in the literature. For instance, Mladenic et al. [16] discussed the feature selection issues for unbalanced class distribution. However, this work is limited to the Naive Bayesian classifier. They have the conclusion that Odds Ratio is the best measure for Bayesian learning on imbalanced document data. Also, Zheng et al. [35] proposed a feature selection method for imbalanced text documents by adjusting the combination of positive and negative features in the data. Their method sticks to the traditional goodness-measures of features. Finally, Chen et al. [29] proposed to use the ROC curve instead of accuracy for evaluating the classification performances. They showed that the feature-selection performances measured by the ROC curve are more robust for imbalanced data. All these prior works are only exploratory in nature, and the problem of developing a general and systematic approach for feature selection on imbalanced data remains pretty much open. This paper presents such an attempt.

2. Traditional feature selection methods

In this section, we mainly introduce three traditional feature selection methods: Correlation-based, Fisher and Mutual Information approaches, which will be used in this study.

Correlation coefficient. The correlation coefficient is a statistical test that measures the linear relationship between two variables. Correlation coefficients can range from -1 to 1. The absolute value of the coefficient gives the strength of the relationship. Absolute values which are closer to 1 indicate a stronger relationships. The sign of the coefficient gives the direction of the relationship: a positive sign means that the two variables increase or decrease in a consistent way, while a negative sign shows that one variable decreases as the other increases, and vice versa.

In learning problem, the correlation coefficient is used to evaluate how accurately a feature predicts the class or target independent of the context of other features. All features are then ranked based on the correlation scores. If the covariance $cov(X_i,Y)$ between a feature X_i and the target Y and the variances of the feature and target, $var(X_i)$ and $var(Y)$ are available, the correlation can be calculated as

$$\rho = \frac{cov(X_i,Y)}{\sqrt{var(X_i)\cdot var(Y)}} \tag{1}$$

Eq. (1) can only be used when the true values for the covariance and variances are known. If these statistics are unknown, an estimate of the correlation can be obtained using Pearson's correlation coefficient over a sample of the population (x_i,y). The formula uses the mean of each feature and the target to calculate the coefficient

$$\rho = \frac{\sum_{k=1}^{m}(x_i^k-\overline{x_i})(y_k-\overline{y})}{\sqrt{\sum_{k=1}^{m}(x_i^k-\overline{x_i})^2\cdot\sum_{k=1}^{m}(y_k-\overline{y})^2}} \tag{2}$$

where m is the number of data points. Then we use the coefficient of determination, or ρ^2, to enforce a ranking of the features according to the goodness of linear fit between the target and features. We should note that only linear relationships between a feature and the target can be found by this correlation. Thus, the feature and the target may be perfectly related in a non-linear manner, but the correlation could be equal to 0. This restriction can be overcome by using simple non-linear preprocessing techniques on the feature before computing the correlation coefficients in order to obtain a goodness of non-linear relationship fit between features and the target [10]. Besides, another issue with using correlation coefficients is how to rank features based on coefficient. If features are according to their values, some strong correlated features with target, which have negative coefficient, may not be chosen. On the other hand, if features are chosen based on their absolute values, we may not select a ratio of positive to negative features that give the best results [29]. How to decide the ratio often takes a lot of empirical testing, though it can result in extremely strong result.

Mutual information. Mutual information is a criterion using information theory, which is commonly used in statistical language modeling of word associations and related application [25,7]. The mutual information between each variable and the target is

$$I(i) = \int_{x_i}\int_{y} p(x_i,y)\log\frac{p(x_i,y)}{p(x_i)p(y)}\,dx\,dy \tag{3}$$

where $p(x_i)$ and $p(y)$ are the probability densities of x_i and y, and $p(x_i,y)$ is the joint density. The criterion $I(i)$ is the measure of dependency between the density of variable x_i and the density of the target y.

The difficulty is that the densities $p(x_i)$, $p(y)$ and $p(x_i,y)$ are often unknown and are hard to estimate from data. It will become easier for the case of discrete or nominal variables because the integral becomes a sum:

$$I(i) = \sum_{x_i}\sum_{y} P(X=x_i,Y=y)\log\frac{P(X=x_i,Y=y)}{P(X=x_i)P(Y=y)} \tag{4}$$

The probability can be estimated from frequency counts. But the estimation will become harder with a larger number of classes and variable values. While the estimation may be unreliable if the number of observations is not sufficient.

For the continuous case, which is the hardest, one can consider discretizing the variables or approximating their densities with a non-parametric method such as Parzen windows.

Fisher criterion. Fisher's criterion of a single feature is defined as the following. Given a feature X_i in a two-class data set, we denote instances in class 1 as X^1, and those in class 2 as X^2. Assume that $\overline{X}_i^k$ is the average of the feature in class k. The Fisher score (F-score for short) of the feature is

$$F(X_i) = \frac{(\overline{X_i^1}-\overline{X_i^2})^2}{(S_i^1)^2+(S_i^2)^2} \tag{5}$$

where

$$(S_i^k)^2 = \sum_{x\in X_i^k}(x-\overline{X_i^k})^2 \tag{6}$$

The numerator indicates the discrimination between two classes and the denominator indicates the scatter within each class. The larger the F-score is, the more likely this feature is discriminative. Therefore, this score can be a criterion of feature selection, which is simple and generally quite effective [14].

For data with more than two classes, F-score can be extended as the following. Given a feature X_i with m classes, denote the set of instances in class k as X_i^k, and $|X_i^k| = l_k$, $k=1,\ldots,m$. Assume that $\overline{X}_i^k$ and $\overline{X}_i$ are the average of the feature in X_i^k and X_i, respectively. The Fisher score of this feature is defined as

$$F(X_i) = \frac{S_B(i)}{S_W(i)} \tag{7}$$

where

$$S_B(i) = \sum_{k=1}^{m} l_k(\overline{X}_i^k-\overline{X}_i)^2 \tag{8}$$

$$S_W(i) = \sum_{k=1}^{m}\sum_{x\in X_i^k}(x-\overline{X}_i^k)^2 \tag{9}$$

Note that Eq. (7) is equivalent to Eq. (5) when $k=2$. A negative aspect of using Fisher criterion to obtain feature ranking is that it also does not reveal mutual information among features.

3. Challenge: feature selection in imbalanced data

In this section, we illustrate the challenges of feature selection imposed by imbalanced data. To make the discussion more specific and more tangible, we illustrate the challenges using a specific feature selection method, i.e., correlation-based feature selection. Also, we employ Bayesian learning in the discussion.

For classification, feature selection is used to improve the classification performance in terms of accuracy and computation [29]. Here, we focus on the improvement of classification accuracy. In the discussion, we only show the classification boundary and condition probability for one dimension, but the similar result can be generalized to the multiple/entire attributes as well. As we know, for the binary classification problem in one dimension, the classification error is proportional to the shadow area in Fig. 1 [23], where $p(x/y_1)$ and $p(x/y_2)$ are the class condition probabilities. After feature selection, the shadow area usually tends to be smaller than before because the selected features tend to be more discriminative between classes [22], as shown in Fig. 2, where x corresponds to different attributes before and after feature selection. This kind of bias of feature selection has been shown in [22]. However, this positive effect of feature selection often does not happen when data is imbalanced for traditional feature selection methods, such as correlation approaches or Fisher-based approaches. To better illustrate, we generate some synthetic data to further explain this.

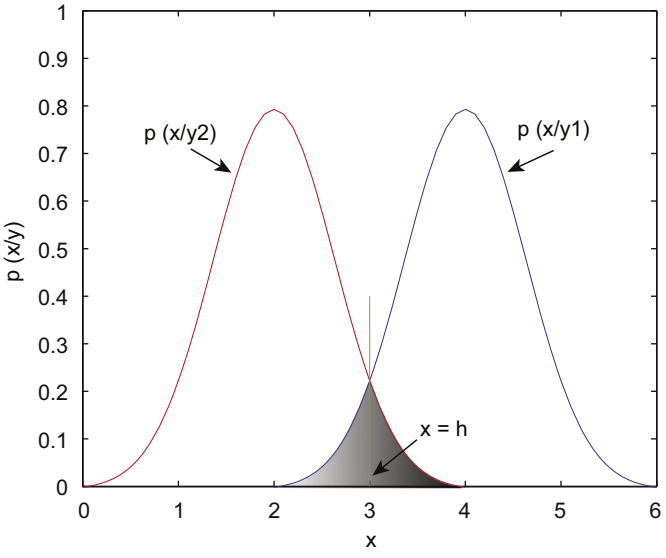

Fig. 1. One dimension class-conditional probability.

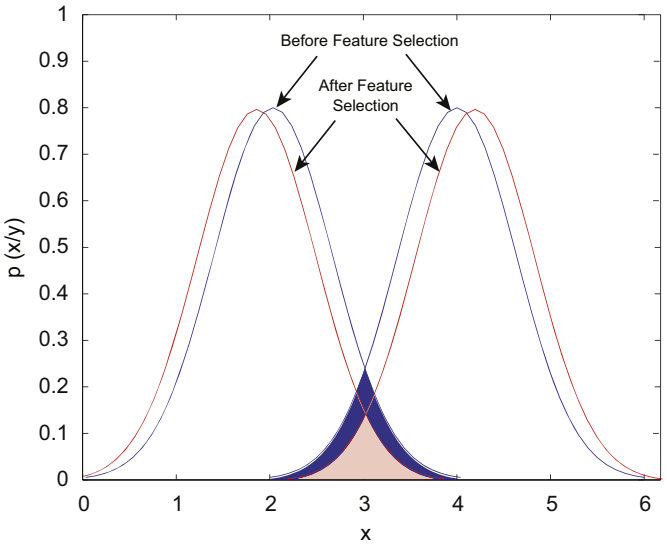

Fig. 2. Decision boundary before and after feature selection.

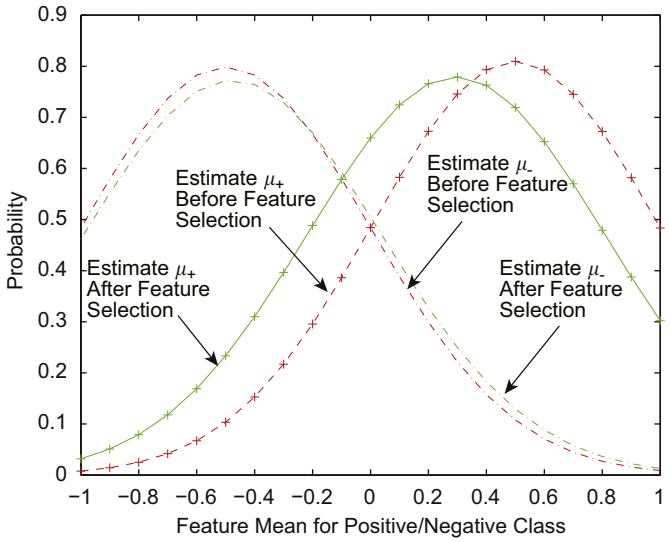

Fig. 3. Decision boundary for imbalanced data.

Table 1
Synthetic data set-class conditional means and standard deviation.

Class	+	−
Mean	$\mu_+ = 0.1$	$\mu_- = -0.1$
Standard deviation	$\sigma_+ = 1$	$\sigma_- = 1$

3.1. Synthetic data generation

We generate the continuous synthetic data with binary classes. The class conditional distribution of the attributes is Normal and has values as shown in Table 1. All features are independent and have identical properties. We created continuous data sets with 100 attributes and 1000 instances. To keep the data set imbalanced, the number of instances in two classes is 900 and 100, respectively.

3.2. Classification error illustration

Since the underlying distribution of class-condition was assumed to follow the Normal distribution, The distribution of positive or negative class-conditional mean will be $N(\mu_+, \sigma_+/\sqrt{n})$ or $N(\mu_-, \sigma_-/\sqrt{n})$. We estimate the μ_+ and μ_- with maximum likelihood estimation. For feature selection, we choose a correlation-based method to select top 10 attributes. The simulation is repeated 100 times, the average results, i.e. the average estimated μ_+ and μ_-, are depicted in Fig. 3.

As can be seen, after feature selection, the means of negative and positive class-condition probabilities shift close to each other, which indicates the class-condition Gaussian distributions will move close to each other and consequently causes that the classification error becomes larger than before feature selection. This clearly illustrates that feature selection will be really challenging when the samples are imbalanced.

4. Decomposition-based feature selection

In this section, we introduce a decomposition-based feature selection approach, which can deal with the challenge imposed by imbalanced data.

Three feature-selection criteria introduced in Section 2 lead to three measures for goodness of features. For simplicity, we represent them as *Corr*, *MI* and *Fisher*, respectively. A higher score indicates the better feature for all these three measurements. With the above definitions of three methods, we can find that when the samples are really imbalanced, the calculation of scores will be mainly influenced by the majority class. For example, given a two-class problem, class 1 is the majority class and class 2 is the small class. Suppose that the attribute distribution in each class follows the Normal distribution with different means, for the calculation of *Fisher* in Eq. (5), S_i^1 tends to be larger than S_i^2 because the number of instances in class 1 is much larger than that of class 2 for imbalanced samples. This negative imbalanced influence is mainly caused by the really skewed ratio of the numbers of samples in different classes. Furthermore, this negative imbalanced influence would be worse if noise exists in reality. In addition, for *Corr* and *MI*, the calculations are negatively influenced by the skewed ratio as well.

To overcome this challenge, in this paper, we propose a decomposition-based framework for feature selection. Specifically, as shown in Fig. 4, there are three phases in this framework. In Phase I, we employ K-means clustering on class i ($i = 1,2,\ldots,C$) according to the user preset cluster number $K(i)$ to decompose the majority class into relatively balanced pseudo-subclasses. Then we change the instance labels of class i with the subclass labels provided by the K-means clustering, thus forming a multi-class data set with $\sum_{i=1}^{C} K(i)$ subclasses. We get the pseudo-labels for the samples with the pseudo-subclasses. Note that for some minor class, decomposition is unnecessary, which means that $K(i) = 1$. In Phase II, we measure the goodness of each feature with the pseudo-labels and the traditional measurement of the goodness of each feature. Next, we rank the features according to the goodness based on the calculated scores. Then, we select the top k good features and release the pseudo-labels to original labels. Finally, in Phase III, we can do classification or other jobs with the selected features and the released labels.

Note that there are some points needed to be further addressed. In Phase I, the cluster number $K(i)$ is larger than 1 for the relative large class i, equal to 1 for relative small class. In practice, we first estimate the ratio of the size of the majority class to the size of the minor class and assign $K(i)$ accordingly to adjust the data set to a relatively balanced situation.

While we employ k-means clustering algorithm to decompose the majority class, the choices of clustering algorithms in our framework is not limited to k-means. Any other clustering algorithm which can produce clustering with relatively balanced

(Decomposition-based Framework for Feature Selection)

Input: Tr: a training sample set.
La: the label for Tr.
M: the number of selected features.
K: a vector specifies the number of local clusters for each class.

Output: Tr': the training sample after feature selection.

Procedure:

Phase I: local clustering
1. for class i=1 to C //C represents #classes
2. clusterLabel(i)=Clustering(TR(i), K(i));
3. La(i)*=changeLabel(La(i), clusterLabel(i));
4. end for

Phase II: score calculation
5. for feature j=1 to N // N represents # of feature
6. Score(j)=scoreMeasure(feature(j), La*);
7. end for
8. ranking feature according to score,
9. feature with higher score coming first
10. Tr'=Tr(sample set with top-M-scores feature)

Phase III: validating
11. learningModel=build(Tr',La);
12. Te'=SelectFeature(Te); // select the corresponding
13. features for testing sample
14. Predict(learningModel, Te');
(for validating only)

Fig. 4. The decomposition-based framework for feature selection.

size, such as EM (Expectation–Maximization) clustering algorithms, can also be used in this framework.

In addition, although we mainly study our decomposition-based method with three traditional feature selection methods introduced in Section 2 in this paper, our decomposition-based framework can be applied to general imbalanced data together with more feature selection methods, such as entropy-based method and classifier-dependent methods.

Finally, the proposed decomposition-based framework is computationally efficient. First, if the K-means is used within the large class, the time required in the clustering phase is modest-basically linear in the number of data points [23]. Also, since the number of relatively large class in a data set is often very small. So the time for clustering is at most $O(n)$, where n is the number of total samples. Thus we can keep the computational cost of the proposed frame at the same level as the original classifier.

5. Feature selection based on the Hellinger distance

In this section, we introduce an alternative feature selection method based on the Hellinger distance for imbalanced data. The Hellinger distance is a measure of distributional divergence [5,12]. Let P and Q denote two probability measures that are continuous distributions with respect to a third probability measure λ. The definition of the Hellinger distance can be given as

$$d_H(P,Q)=\sqrt{\int_{\Omega}(\sqrt{P}-\sqrt{Q})^2\,d\lambda} \tag{10}$$

This definition does not depend on λ. It can also be defined for a countable space Φ:

$$d_H(P,Q)=\sqrt{\sum_{\phi\in\Phi}(\sqrt{P(\phi)}-\sqrt{Q(\phi)})^2} \tag{11}$$

The range of the Hellinger distance is in $[0,\sqrt{2}]$. And it is symmetric and non-negative, implying $d_H(P,Q)=d_H(Q,P)$. The Hellinger distance allows us to capture the notion of "affinity" between the probability measures on a finite event space. If $P=Q$, then $distance=0$ (maximal affinity) and if P and Q are completely disjoint then $distance=\sqrt{2}$ (zero affinity). Therefore, the better feature we want to select is the one that carries the minimal affinity between the class. The minimal affinity means that this feature is most discriminative between classes. Thus, the Hellinger distance can be used to measure the prediction power of features to classify the samples. The higher the distance is, the better the corresponding feature is.

Table 2
Some notations.

Corr	Traditional correlation-based method
Fisher	Traditional Fisher-based method
MI	Traditional mutual information method
D-Corr	Decomposition-based framework with correlation
D-Fisher	Decomposition-based framework with Fisher
D-MI	Decomposition-based framework with Mutual information
Hell	Hellinger distance-based method.

For the efficiency of computation, we discretize all continuous features into p partitions or bins. Assuming a two-class problem, let X_+ be class + and X_- be class −. Then we are interested in calculating the distance in the normalized frequencies aggregated over all the partitions of the two class distributions X_+ and X_-. The Hellinger distance between X_+ and X_- is

$$d_H(X_+,X_-)=\sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{|X_{+j}|}{X_+}}-\sqrt{\frac{|X_{-j}|}{X_-}}\right)^2} \tag{12}$$

This distance formulation is strongly skew insensitive as the prior does not influence the distance calculation [5]. It essentially captures the discrimination between the distributions of feature values given the two different classes. There is no factor of class prior information.

6. Experimental results

In this section, we present the experimental results to validate the performances of the decomposition-based and Hellinger distance-based methods for feature selection on imbalanced real-world data.

6.1. The experimental setup

Experimental tools. In the experiments, we used three types of classifiers: Support Vector Machines (SVMs), decision trees, and Bayesian learning. Their corresponding implementations are LIBSVM [3], C4.5 and Bayesian learning [30]. Also, we applied K-means, a widely used clustering method which tends to produce clusters with relatively uniform sizes, as the clustering method in our decomposition-based framework. During the K-means clustering, when the data sets are relatively low-dimension, squared Euclidean distance was used as the distance measure; and when the data sets have relatively high dimension, the cosine similarity was used instead. The reason is that Euclidean distance in a high-dimensional space is not very meaningful. Finally, some notations are given in Table 2. For example, D-Corr means we combine decomposition-based framework with correlation measure to rank the features. The default classifier used for classification is SVMs. If other classifiers are used instead, e.g., C4.5, we will explicitly denote it by "Corr(C4.5)".

Experimental data sets. For our experiments, we used a number of benchmark data sets that were obtained from different application domains. They almost have a large number of features, and significant imbalanced between classes. Some characteristics of these data sets are shown in Table 3. These data sets have been

Table 3
Some characteristics of experimental data sets.

Data set	# Objects	#Features	#Classes	Class ratio
CNS [20]	90	7129	2	2:1
LYMPH [21]	77	7129	2	58:19
OVARY [9]	66	6000	2	50:16
NIPS [19]	200	13,649	2	160:40

Table 4
Comparison cross data sets, methods and metrics.

	MI	D-MI	Fisher	D-Fisher	Corr	D-Corr
Comparisons on the CNS data set						
F-measure for rare class	0.66	0.70	0.54	0.60	0.69	0.77
F-measure for major class	0.87	0.85	0.85	0.87	0.89	0.88
AUC of ROC	0.59	0.63	0.44	0.50	0.53	0.54
Comparisons on the LYMPH data set						
F-measure for rare class	0.65	0.70	0.54	0.59	0.1667	0.32
F-measure for major class	0.86	0.90	0.85	0.87	0.8462	0.8682
AUC of ROC	0.46	0.54	0.50	0.60	0.412	0.5935
Comparisons on the OVARY data set						
F-measure for rare class	0.66	0.70	0.54	0.60	0.5833	0.6087
F-measure for major class	0.87	0.85	0.85	0.87	0.9074	0.9174
AUC of ROC	0.59	0.63	0.44	0.50	0.5112	0.6062
Comparisons on the NIPS data set						
F-measure for rare class	0.58	0.65	0.67	0.71	NaN	NaN
F-measure for major class	0.90	0.91	0.91	0.92	0.8725	0.8827
AUC of ROC	0.50	0.53	0.5	0.50	0.4511	0.5963

Note: NaN means that the precision for rare class is 0.

used to validate the performance of the feature selection method in [29]. However, as we discussed before, this feature-selection method [29] is classifier-dependent and is beyond the research scope of this paper.

Another real-world data set we used is the KDDCUP'09 data set. This data set is about customer relationship management. The target is to estimate the churn, appetency and up-selling probability of customers. But in this experiment, we only consider churn problem. There are 50,000 samples and 15,000 features for the original data set. Considering the efficiency and effectiveness we random sample 10,000 samples and remove the 260 categorical attributes for our experiment. Thus, for our experimental setting, there are totally 10,000 samples and 14,740 features The class ratio of churn problem is still around 12 : 1 after the random sampling. More detail about the data sets can be found in [1].

Evaluation metrics. The F-measure and Micro-AUC of ROC [34] are used to validate the performance. And for the classification, we use fivefold cross-validation to evaluate the results.

6.2. The performance of the decomposition-based framework

Since we have multiple data sets, feature selection methods and evaluation metrics, the comparisons can be obtained in many combinations. Thus, we summarize the results in Table 4 except for the KDDCUP09 data, where we select top 40 features with different feature selection methods. The F-measure and AUC of ROC comparisons are shown in Table 4. As can be seen, our decomposition-based framework can improve the classification performance in terms of both F-measure and AUC of ROC with a clear margin.

Table 5
A performance comparison with C4.5.

	Fisher (C4.5)	D-Fisher (C4.5)	MI (C4.5)	D-MI (C4.5)
Comparisons on LYMPH data set				
F-measure for major class	0.88	0.86	0.814	0.832
F-measure for minor class	0.64	0.60	0.389	0.53
AUC of ROC	0.75	0.76	0.625	0.737
Comparisons on OVARY data set				
F-measure for major class	0.863	0.854	0.776	0.768
F-measure for minor class	0.533	0.483	0.353	0.303
AUC of ROC	0.709	0.741	0.57	0.587

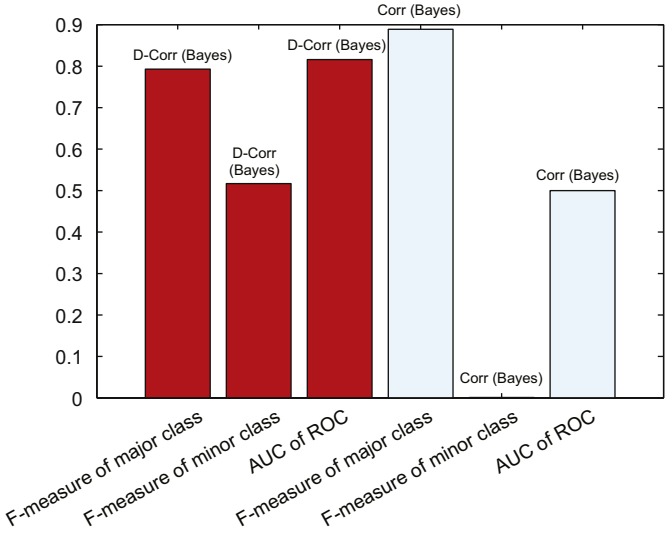

Fig. 5. Performance comparisons with Bayesian learning on the NIPS data set.

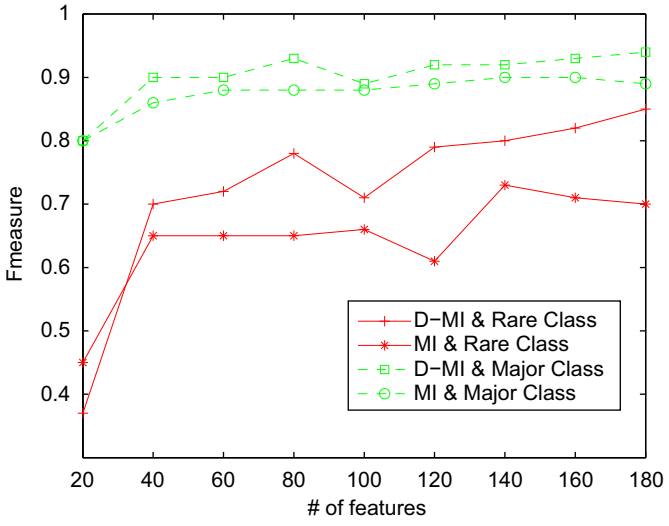

Fig. 6. The performance with respect to # of features on LYMPH.

The results for different classifiers also illustrate the effectiveness of our decomposition-based framework. For example, on the CNS data set, we show the comparison of D-MI and MI in Table 5. As we can see, with C4.5 classifier, we can obtain better classification performance with our decomposition-based framework than

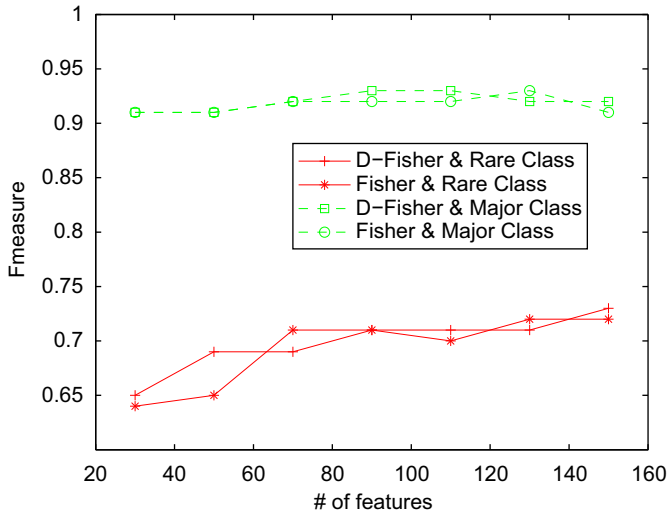

Fig. 7. The performance with respect to # of features on NIPS.

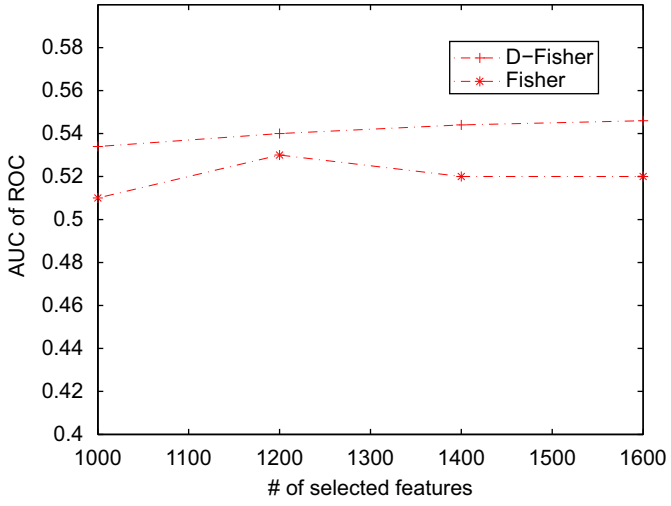

Fig. 8. A comparison of D-Fisher and Fisher on KDDCUP09.

traditional methods. In addition, for Bayesian learning, the similar improvement is obtained. For example, we show the performance on the NIPS data in Fig. 5. We can observe that traditional correlation-based approach results in even 0 F-measure for rare (small) class with Bayesian learning.

Furthermore, we show the performance versus the number of selected features using MI and D-MI on the LYMPH data set in Fig. 6. As can be seen, D-MI significantly outperforms MI for most different numbers of features when using SVMs. In addition, we can observe that the performance becomes better with the increasing of the number of features. This change is very sharp for initial increasing number of features, and it becomes slow when the number of features is over certain value. Also there are some accident drop when the number of features increases. This is probably caused by two reasons: the variance of classifier on different sample sets and the redundant features introduced. Similar results can be found for other traditional methods. For example, for Fisher and D-Fisher, the results on the NIPS data set, similar trend is shown in Fig. 7. For all cases, selecting a small set of features with decomposition-based framework outperforms traditional feature-selection methods.

For the KDDCUP09 data set, we only use SVMs as the classifier. The number of selected features is over 1000. Fig. 8 shows the comparisons of AUC between Fisher and D-Fisher. As can be seen,

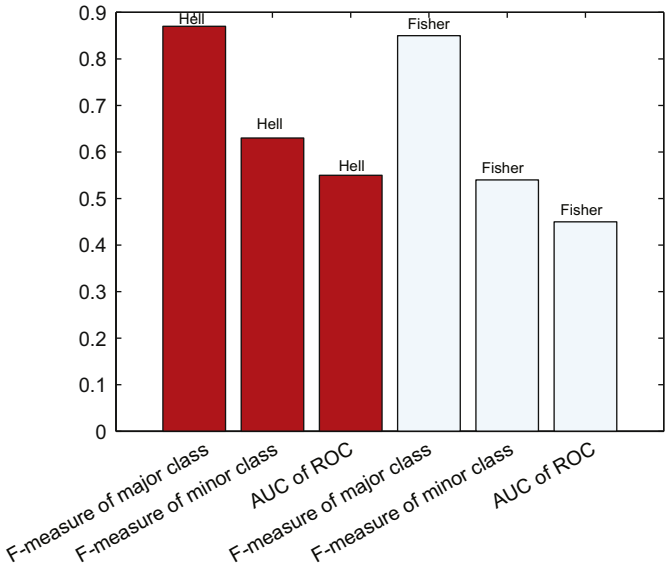

Fig. 9. A comparison of Hell and Fisher on CNS.

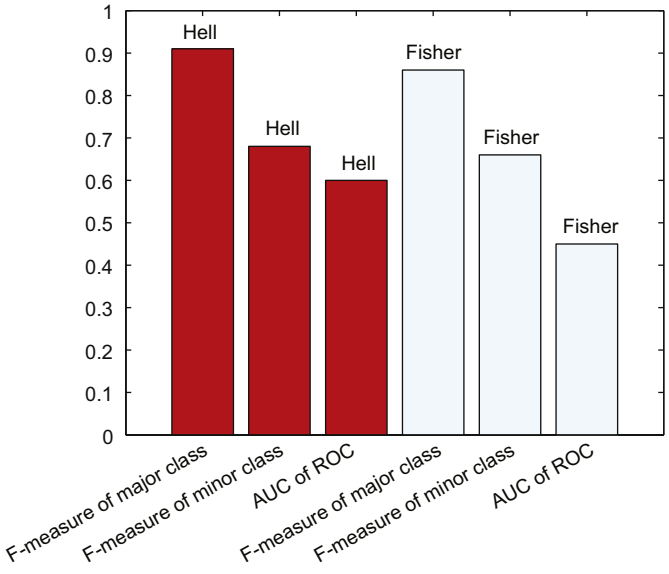

Fig. 10. A comparison of Hell and Fisher on LYMPH.

when we select 1000, 1200, 1400 or 1600 features, our decomposition-based framework consistently outperform Fisher in terms of AUC.

6.3. *The performance of feature selection based on the Hellinger distance*

In this subsection, we show the performance of the Hellinger distance for feature selection. The comparisons of F-measure and AUC between Fisher and Hell on CNS, LYMPH and NIPS are shown in Figs. 9–11, where the number of selected features is 40. As can be seen, Hell outperforms Fisher on CNS and LYMPH data sets with a significant margin, but Hell is less competitive on the NIPS data set. Hell can also lead to better performances than other traditional methods. For example, compared with MI, the improvement of Hell is shown in Fig. 12. In addition, we show the performance of the Hellinger distance versus the number of selected features on the CNS data set in Fig. 13. As can be seen, the

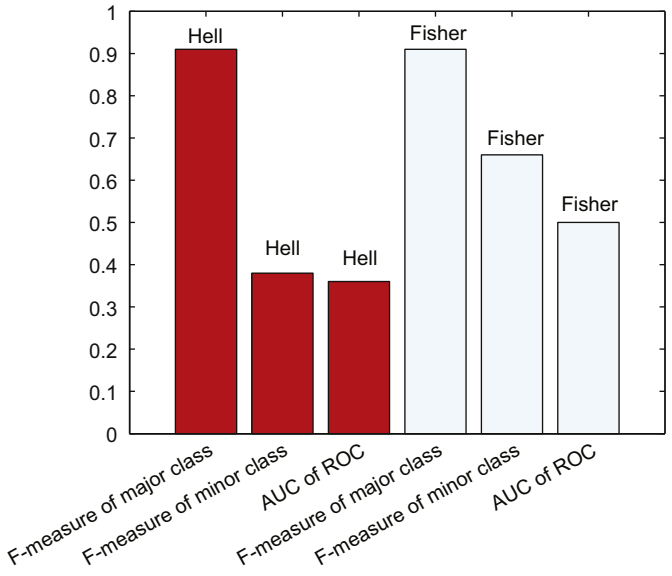

Fig. 11. A comparison of Hell and Fisher on NIPS.

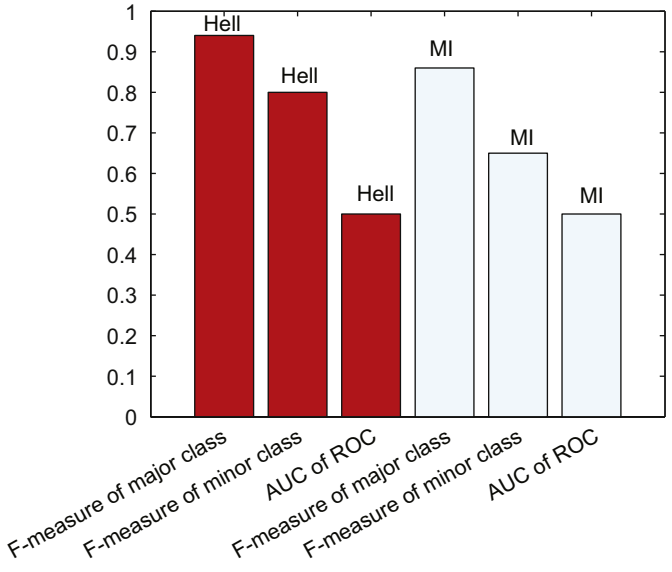

Fig. 12. A comparison of Hell and MI on LYMPH.

performance in terms of F-measure or AUC overall increases with the number of selected features, while there is some fluctuation.

7. Concluding remarks

In this paper, we studied the feature selection issues in imbalanced data, where a critical challenge for feature selection is the dominant influence of samples from large classes on the traditional feature selection methods. To deal with this challenge, we proposed two different approaches. The first one is to first decompose large classes into pseudo-subclasses with relatively balanced sizes and then measure the goodness of features with the decomposed data. The decomposition-based framework reduces the biased influence of imbalanced class distributions on feature selection methods. Alternatively, we also developed a Hellinger distance-based method. Hellinger distance is insensitive to the class distributions, since the computation of this distance does not involve the class information. Therefore, the Hellinger

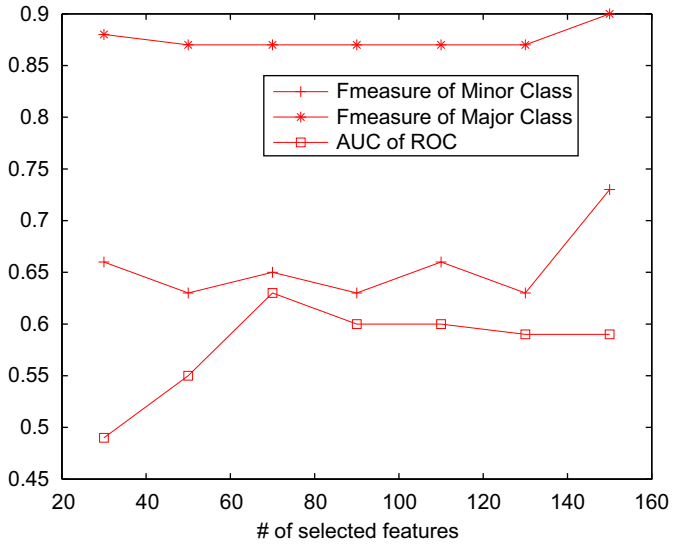

Fig. 13. The performance of Hell with respect to the number of selected features on CNS.

distance-based feature selection method can also handle the challenge imposed by imbalanced class distributions.

We validated the performances of the proposed two approaches using various real-world data. Specifically, we compared the proposed methods with three traditional feature rank methods, correlation, Fisher and Mutual information. As demonstrated in the experimental results, both decomposition-based and Hellinger distance-based methods outperform traditional feature-selection methods with a clear margin in terms of F-measure, AUC and ROC.

References

[1] 〈http://www.kddcup-orange.com/〉.
[2] C. Drummond, R. Holte, C4.5, class imbalanced, and cost sensitivity: why under-sampling beats over-sampling, in: The 20th International Conference on Machine Learning Workshop on Learning from Imbalanced Data Sets, 2003.
[3] C.-C. Chang, C.-J. Lin, LIBSVM: A Library for Support Vector Machines, 2001.
[4] T.W.S. Chow, P. Wang, E.W.M. Ma, A new feature selection scheme using a data distribution factor for unsupervised nominal data, IEEE Trans. Syst. Man Cybern. B: Cybern. PAMI-38 (2) (2008) 499–509.
[5] D.A. Cieslak, N.V. Chawla, Learning decision trees for unbalanced data, in: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008, pp. 241–256.
[6] C. Ling, C. Li, Data mining for direction marketing: problem and solutions, in: Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1998.
[7] I. Dhillon, S. Mallela, R. Kumar, A divisive information-theoretic feature clustering algorithm for text classification, J. Mach. Learn. Res. 3 (March) (2003) 1265–1287.
[8] D. Margineantu, T. Dietterich, Learning Decision Trees for Loss Minimization in Multi-Class Problems, Technical Report, Oregon State University, 99-30-03.
[9] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (2002) 572–577.
[10] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. (Special Issue on Variable and Feature Selection) 3 (March) (2003) 1157–1182.
[11] G. Weiss, Mining with rarity: a unifying framework, ACM SIGKDD Explor. Newslett. (Special Issue on Learning from Imbalanced Datasets) 6 (1) (2004) 7–19.
[12] T. Kailath, The divergence and Bhattacharyya distance measures in signal selection, IEEE Trans. Commun. Technol. 15 (1) (1967) 52–60.
[13] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. B: Cybern. PAMI-39 (2) (2009) 539–550.
[14] W. Malina, On an extended Fisher criterion for feature selection, IEEE Trans. Pattern Anal. Mach. Intell. PAMI-3 (5) (1981) 611–614.

[15] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one-sided selection, in: Proceedings of the 14th Annual International Conference on Machine Learning, 1997.
[16] D. Mladenic, M. Grobelnik, Feature selection for unbalanced class distribution and naive Bayes, in: Proceedings of the 16th International Conference on Machine Learning, 1999.
[17] Mahesh V. Joshi, R. Agarwal, Vipin Kumar, Predicting rare classes: can boosting make any week learner strong? in: Proceeding of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.
[18] Y. Pang, Y. Yuan, X. Li, Effective feature extraction in high-dimensional space, IEEE Transaction on Syst. Man Cybern. B: Cybern. PAMI-38 (6) (2008) 1652–1656.
[19] S. Roweis, 〈http://www.cs.toronto.edu/roweis〉.
[20] P. Scott, P. Tamayo, et al., Prediction of central nervous system embryonal tumour outcome based on gene expression, Nature 415 (January) (2002) 436–442.
[21] M.E.A. Shipp, Diffuse large b-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning, Nature Med. 8 (2002) 68–74.
[22] S.K. Singhi, H. Liu, Feature subset selection bias for classification learning, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 849–856.
[23] P.-N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Addison-Wesley, 2005.
[24] Y. Tang, A.S.K. Yan-Qing Zhang, Nitesh V. Chawla, SVMs modeling for highly imbalanced classification, IEEE Transaction on Syst. Man Cybern. B: Cybern. 39 (1) (2009) 281–288.
[25] K. Torkkola, Feature extraction by non-parametric mutual information maximization, J. Mach. Learn. Res. 3 (March) (2003) 1415–1438.
[26] M. Wang, X.-S. Hua, J. Tang, R. Hong, Beyond distance measurement: constructing neighborhood similarity for video annotation, IEEE Trans. Multimedia.
[27] M. Wang, X. Sheng Hua, R. Hong, J. Tang, G.-J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circuits Syst. Video Technol. 19 (5) (2009).
[28] X. Wen Chen, J.C. Jeong, Minimum reference set based feature selection for small sample classifications, in: Proceedings of the 24th International Conference on Machine Learning, 2007.
[29] X. Wen Chen, M. Wasikowski, Fast: a ROC-based feature selection metric for small samples and imbalanced data classification problems, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2008.
[30] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., Morgan Kaufmann, San Francisco, 2005.
[31] Z. Xu, R. Jin, J. Ye, M.R. Lyu, I. King, Non-monotonic feature selection, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009.
[32] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: Proceedings of the 14th International Conference on Machine Learning, 1997.
[33] Y. Yang, H.T. Shen, Z. Ma, Z. Huang, X. Zhou, L21-norm regularized discriminative feature selection for unsupervised learning, in: Proceedings of International Joint Conferences on Artificial Intelligence, 2011.
[34] Y. Yang, F. Wu, F. Nie, H.T. Shen, Y. Zhuang, A.G. Hauptmann, Web personal image annotation by mining label correlation with relaxed visual graph embedding, IEEE Trans. Image Process. 21 (3) (2012) 1339–1351.
[35] Z. Zheng, X. Wu, R. Srihari, Feature selection for text categorization on imbalanced data, ACM SIGKDD Explor. Newslett. (Special Issue on Learning from Imbalanced Datasets) 6 (1) (2004) 80–89.

Liuzhi Yin is currently doing a Ph.D. degree in Department of Statistics and Finance of University of Science and Technology of China (USTC). He received B.E. degree from USTC in 2005. His general area of research is Random Matrix Theory and its application in finance and biology. He has published several technical papers in peer reviewed journals and conference proceedings of China.

Yong Ge received the B.E. degree in Information Engineering from Xi'an Jiao Tong University, Xi'an, China, in 2005 and the M.S. degree in Signal and Information Processing from the University of Science and Technology of China, Hefei, China, in 2008. He is currently a Ph.D. candidate in the Department of Management Science and Information System at Rutgers, the State University of New Jersey, NJ, USA. His research interests include data mining, spatio-temporal data analysis, and recommender systems.

Keli Xiao is a Ph.D. candidate in the Finance and Economics Department at Rutgers Business School, Rutgers, the State University of New Jersey, USA. Before that, he have received the master degree of quantitative finance at Rutgers in 2009 and the master degree of computer science at Queens College-CUNY, USA, in 2008. His research interests include data mining, recommender systems, financial fraud detection and the asset bubbles in housing markets.

Xuehua Wang received the B.S. degree in applied mathematics from Dalian University of Technology, Dalian, China, in 1990 and the M.S. degree in applied mathematics from Dalian University of Technology, Dalian, China, in 1993, and the Doctor degree in system engineering from Dalian University of Technology, Dalian, China. She is currently a Professor in the department of management and economics at Dalian University of Technology, Dalian, China. Her research interests include data mining, complex system analysis.

Xiaojun Quan is currently a Ph.D. student in the Department of Computer Science, City University of Hong Kong. He received the B.E. degree in computer science from the Chang'an University in 2005 and the M.E. degree in computer science from University of Science and Technology of China in 2008. His research interests include data mining, information retrieval and question answering.