A feature selection algorithm for multilayer perceptron based on simultaneous two-sample representation

Shudong Liu
School of Information and Security engineering
Zhongnan University of Economics and Law
Wuhan, China
Corresponding: upt.mymeng@gmail.com

Ke Zhang
School of Information and Security engineering
Zhongnan University of Economics and Law
Wuhan, China

Xu Chen
School of Information and Security engineering
Zhongnan University of Economics and Law
Wuhan, China

***Abstract*— Classification is one of the hot topics of machine learning domains, its main task is to learn a classification model from training data and predict the labels of unknown samples. To date, many classification models have been proposed and are widely used in various real-world applications, e.g., naive Bayes (NB), logistic regression (LR), support vector machine (SVM) have been successfully employed in spam recognition, bank loan credit scoring and network rumor recognition, respectively. Imbalance learning is an important branch of classification task in machine learning domains. Data-level, algorithm-level and ensemble solutions are the three main methods proposed thus far to address imbalance learning. To alleviate the issues of data explosion and feature selection for a multilayer perceptron with simultaneous two-sample representation, in this paper, we propose a novel feature selection method based on the pairwise samples distance constraint, which considers the class labels of paired samples, select the features which push two similar samples closer together and pull two different samples farther apart. Finally, we conduct experiments on four high-dimensional DNA microarray datasets. The experimental results demonstrate that our proposed algorithms outperform some state-of-the-art algorithms in terms of F-measure and G-mean.**

Keywords—Multilayer perceptron, Imbalance learning, feature selection, information gain, supervised learning

I. Introduction

Over past decades, imbalance learning has attracted lots of attention from both the academia and industry. Several workshops have been held by international conferences, and many special issues have been released by major academic journals. In addition, imbalance learning has been listed as one of the top ten difficult problems in the field of data mining at ICDM'05[1]. With the development of big data and deep learning techniques, we will face much more difficulties in imbalance learning field, e.g., imbalance learning algorithm for big data processing platform, the new approach for synthetic minority sampling, and different tradeoff learning strategies for high-imbalanced datasets. Overall, the existing imbalance learning algorithms can be divided into three groups [2,3]: data-level solution [4,5], algorithm-level solution [6,7] and ensemble-based solution [8,9].

Generally, most of classifiers need to train a map $f: X \rightarrow Y$, where $X \in R^n$ is the feature space of all samples, and $x_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, \cdots x_i^{(n)})$ refers to the n-dimensio -nal feature vector, Y is the class label vector, and $|Y|$ represents the number of distinct class labels of all samples. For example, Y is a 2-dimensional vector $Y = \{y_1, y_2\}$ for a two-class classification. Dumpala et al. [10] propose an imbalance learning algorithm based on a new data representation, called multilayer perception (MLP) based on simultaneous two-sample representation (S2SMLP). The input of S2SMLP is no longer a sample x_i, but a pair of samples $\{x_i, x_j\}$, and the output is the class label $\{y_i, y_j\}$ of the two samples $\{x_i, x_j\}$, which provides a new way to learn the relationship among intra- and inter-class variables. Furthermore, a sample x_i can be coupled with each x_j from training dataset. The advantage of S2SMLP is that the label of sample x_i judges by the voting result of all output y_i in each couple $\{y_i, y_j\}$. Obviously, the volume of S2SMLP's training data is $|X|$, and the input of S2SMLP is a dimensional vector. The data representation may result in a high-dimensional big training dataset, and it will become more serious in the era of big data.

To solve these problems, in this paper, we propose a novel feature selection method based on the pairwise samples distance constraint for high-dimensional data, which considers the labels of each pair of samples, and selects the features which push two same class samples closer together and pull two different class samples farther apart. We build an optimization objective function based on the pairwise samples distance constraint, and compute a candidate value for each feature according to its contribution to the pairwise samples distance constraint.

II. Related Works

Like feature selection methods in traditional supervised learning algorithms, feature selection methods for imbalance learning can be divided into three groups: filter-based, wrapper-based and embedded-based methods. The

978-1-7281-7647-5/20/$31.00 ©2020 IEEE

DOI 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00059

filter-based method evaluates each feature and filters it which has a lower score. The advantage of this method is that it does not depend on the classifier, and it has high efficiency and strong scalability, but the selected features are not optimal for a specific classifier [11]. The latter two methods depend on the classifier, and produce well-targeted feature subsets. The wrapper-based method directly takes the accuracy of the classifier as the evaluation indicator for feature selection, and then selects the optimal feature subset for the classifier. The disadvantage is that it needs to exhaust all possible feature combinations, it requires a large amount of computation, and usually results in over-reliance on the classifier, moreover, it often overfits the training dataset. The embedded-based method integrates feature selection with the evaluation process through a global optimization algorithm. No matter which method is used, it would come down to ranking all features, and the ranking criterion usually contributes the classification accuracy [12] or loss function of feature discrimination [13].

In recent years, high-dimensional feature selection for imbalance learning has attracted many scholars' attention. For example, Liu et al. [11] propose an effective selection method, which investigates imbalanced problem by optimizing F-measures. Alshawabkeh et al. [14] propose a novel embedding feature selection algorithm that takes advantage of the training samples' mean margins of boosting to select features. Li et al. [15] propose a feature selection algorithm based on weighted mutual information, which assigns different weights to the samples based on the fuzzy C-means clustering method.

By taking features' corresponding distributions over all classes into consideration, Alibeigi et al. [16] propose a feature-ranking algorithm based on features' probability density estimation over all classes, which could weaken the dominant role of the majority class and enhance the influence of the minority class. Zhou et al. [13] propose an online streaming feature selection method based on neighborhood rough set theory. Maldonado et al. [12] propose a family of backward feature elimination algorithms, which integrated the feature elimination process with training SVM by different learning strategies. Moayedikia et al. [17] propose a wrapper feature selection algorithm based on symmetrical uncertainty and harmony search. Additionally, Maldonado et al. [18] also propose an embedded feature selection method for SVM classifiers. It uses support vector data description and cost-sensitive SVM to deal with imbalanced data. Chen et al. [19] propose a feature selection algorithm for imbalanced data based on neighborhood rough sets, which evaluates the significance of features by the uneven distribution of the classes.

III. HIGH-DIMENSIONAL FEATURE SELECTION STRATEGY FOR S2SMLP

In general, the goal of feature selection is to transform the samples from a high-dimensional space to a low-dimensional space, which pushes similar samples closer together and pulls different samples farther apart from each other. This cognitive knowledge inspires, and we select those features which can push similar samples closer together and pull different samples farther apart.

Given a pair of samples $\{x_i, x_j\}$, we define an n-dimensional vector $p_{ij} = [d_1, d_2, \cdots d_n]$ as a selection indicator. If $d_i = 1$, it means that the ith feature of two samples is selected, otherwise, it is not. Many such pairs of samples can be constructed from training data, and we achieve d_i for each pair. Finally, the features are selected in terms of the selection indicator, which satisfies the following formula.

$$\begin{aligned} &max \sum\nolimits_i d_i \\ s.t. \quad &\textstyle\sum_{ij} dis\left(p_{ij}^T x_i, p_{ij}^T x_j\right) < \sigma \end{aligned} \qquad (1)$$

Where σ is a threshold. For the sake of avoiding explosion of high-dimensional feature combination. We narrow the search space, for a pair of samples $\{x_i, x_j\}$ with the same class label, $m(m < n)$is the number of selected features, we select $\lambda(\leq n)$ features from n features, which can push similar samples closer. Then, the selection indicator satisfies the following formula.

$$\begin{aligned} &\textstyle\sum_{ij} dis\left(p_{ij}^T x_i, p_{ij}^T x_j\right) < \sigma \\ s.t. \quad &\textstyle\sum_i d_i = \lambda \end{aligned} \qquad (2)$$

The pairs of majority samples are much more than minority samples in simultaneous two-sample representati -on. It is unreasonable if the two different pairs of samples have similar weights in formula 2, moreover, the influence of the minority class for feature selection is probably overwhelmed by the minority class, therefore we need to further modify formula 2.

$$\begin{aligned} &\frac{\sum_{p_{ij}, x_i \in Ma, x_j \in Ma} dis\left(p_{ij}^T x_i, p_{ij}^T x_j\right)}{M(M-1)} + \frac{\sum_{p_{ij}, x_i \in Mi, x_j \in Mi} dis\left(p_{ij}^T x_i, p_{ij}^T x_j\right)}{N(N-1)} \\ s.t. \quad &\textstyle\sum_i d_i = \lambda \end{aligned} \qquad (3)$$

Similarly, for a pairs $\{x_i, x_k\}$ with different class labels, we define an n-dimensional vector $q_{ik} = [d_1, d_2, \cdots d_n]$as a selection indicator, where $d_i \in \{0,1\}$. The selected features labeled with $d_i = 1$ means that they can pull different samples farther apart, namely:

$$\begin{aligned} &\max \textstyle\sum_{q_{ik}} dis\left(q_{ik}^T x_i, q_{ik}^T x_j\right) \\ s.t. \quad &\textstyle\sum_i d_i = \lambda \end{aligned} \qquad (4)$$

In our following experiment, in order to preset λ by a uniform metric on all experimental datasets, we transform λ to η , where $\eta = \left(\lambda/_d\right) * 100\%$ and d is the dimension of feature space of an original dataset. Therefore, λ refers to the percentage of the dimension of feature space d, and we can preset parameter through parameter η.

There are probably multiple feature combinations simultaneously satisfying above two formulas 3 and 4, which would produce multiple results of the selected indicators p_{ij} and q_{ik}. Then, we normalize them.

$$\tilde{p}_{ij}=\frac{1}{\sum_{\alpha=1}^{C}\binom{n_\alpha}{2}}\sum_{ij,y_i=y_j}p_{ij} \qquad (5)$$

$$\tilde{q}_{ik}=\frac{1}{\sum_{\alpha,\beta,\alpha\neq\beta}^{C}n_\alpha n_\beta}\sum_{ik,y_i\neq y_j}q_{ik} \qquad (6)$$

$\tilde{p}_{ij}=[p_1,p_2,\cdots,p_n]$ and $\tilde{q}_{ik}=[q_1,q_2,\cdots q_n]$ are calculated by the above two formulas. Candidate features exist with high frequency in both $\tilde{p}_{ij}$ and $\tilde{q}_{ik}$. To find them out, we compute a candidate values for each feature be following formula.

$$s_i=\left|\frac{p_i-q_i}{q_i+p_i}\right| \qquad (7)$$

We rank all candidate features in terms of and select features. A feature selection algorithm for S2SMLP is shown in Algorithm 1.

Algorithm 1 Feature selection algorithm for S2SMLP

```
Input: training dataset $\{X,Y\}$, λ and the number of selected features $m(<n)$
Output: m-dimensional feature subset selected from n-dimensional feature space
for $\forall\{x_i,x_j\}\in X$ do
    if $y_i=y_j$ then
        Calculate $p_{ij}$ by formula 3
    else
        Calculate $q_{ik}$ by formula 4
    end if
end for
Calculate $\tilde{p}_{ij}$ and $\tilde{q}_{ik}$ by formula 5 and 6
for $\forall i$
    Calculate $s_i$ by formula 7
end for
Rank all features in terms of $s_i$
Select $Top-m$ features
```

IV. EXPERIMENTS AND RESULTS ANALYSIS

A. Evaluation Criteria

We compare the performance of our proposed algorithm through $F-measure$ and $G-mean$, which are commonly used to evaluate imbalance learning algorithms. The confusion matrix of the classification results is listed as follows (in Table 1).

Table 1. Confusion matrix of classification

	Predicted Positives	Predicted Negatives
Actual Positives	True Positives (TP)	False Negatives(FN)
Actual Negatives	False Positives(FP)	True Negatives(TN)

Positive predictive value (precision) is defined as:

$$Precision=\frac{TP}{TP+FP}$$

True positives rate (recall or sensitivity) is defined as:

$$Recall=TPR=\frac{TP}{TP+FN}$$

True negative rate (specificity) is defined as:

$$TNR=\frac{TN}{TN+FP}$$

$F-measure$ is defined as:

$$F-measure=\frac{2\times Precison\times Recall}{Precision+Recall}$$

$G-mean$ is defined as:

$$G-mean=\sqrt{TPR\times TNR}$$

B. Experimental Datasets

four high-dimensional DNA microarray datasets from KEEL dataset repository and UCI machine learning repository are used in our following experiments. Four high-dimensional DNA microarray datasets are described in Tables 2.

TABLE 2. Description of DNA microarray data

Dataset	# Samples	# Features	# Class	IR
COLON	59	2000	2	22/37
CNS	60	7129	2	21/39
OVARIAN	66	6000	2	16/50
LUNG	181	12534	2	31/150

C. Experimental Results and Analysis

In order to evaluate our proposed feature selection algorithm in this paper, we compare our proposed algorithm with four state-of-the-art feature selection algorithms on four high-dimensional DNA microarray datasets (as shown in Table 2). The four state-of-the-art feature selection algorithms are listed as follows: information gain (IG) [20], minimal redundancy, maximal relevance (mRMR)[21], fast correlation-based filter solution (FCBF) [22] and nearest neighbor feature distinguishing feature weighting algorithm (Relieff) [23](K of KNN is 5 in our experiment).

(1) Parameter λ tuning

According to the research results in [24], the Manhattan distance has better performance in higher dimension space. Therefore, the Manhattan distance is used to calculate the distance between two samples in formula 3 and 4. According to formulas 3 and 4, we need to preset λ in order to find out the two indicative vectors p_{ij} and q_{ik}. In this part of our experiment, in order to preset λ by a uniform metric on all experimental datasets, we transform λ to η, where $\eta = \left(\lambda/_d\right) * 100\%$ and d is the dimension of feature space of an original dataset. Therefore, λ refers to the percentage of the dimension of feature space d, and we can preset λ through η. We let η take different values between 5% and 100% and investigate the performance of our proposed algorithm in terms of $F-measure$ and $G-mean$. The results are shown in Table 3 and Table 4.

TABLE 3. The results of different η on $F-measure$

	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
COLON	0.714	**0.786**	**0.820**	**0.793**	0.816	0.798	0.796	0.796	0.796	0.796	0.796
CNS	0.545	**0.636**	**0.748**	**0.697**	0.688	0.697	0.683	0.697	0.697	0.697	0.697
OVARIAN	0.915	**0.931**	**0.939**	**0.931**	0.931	0.931	0.931	0.931	0.931	0.931	0.931
LUNG	0.892	**0.926**	**0.928**	**0.912**	0.912	0.912	0.912	0.912	0.912	0.912	0.912

TABLE 4. The results of different η on $G-mean$

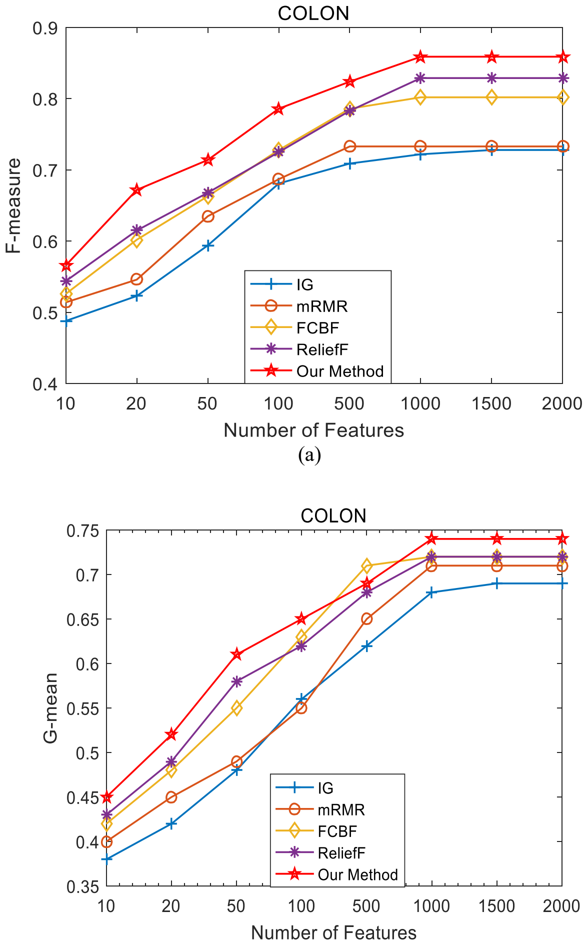
	5%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
COLON	0.612	**0.671**	**0.698**	**0.724**	0.682	0.665	0.664	0.651	0.668	0.668	0.668
CNS	0.656	**0.724**	**0.748**	**0.743**	0.698	0.693	0.669	0.693	0.693	0.693	0.693
OVARIAN	0.893	**0.946**	**0.96**	**0.945**	0.944	0.944	0.944	0.944	0.944	0.944	0.944
LUNG	0.925	**0.943**	**0.943**	**0.941**	0.941	0.941	0.941	0.941	0.941	0.941	0.941

We can see from Table 3 and Table 4 that different values of η have a significant effect on $F-measure$ and $G-mean$, and when η is about 10%-30% of the dimension of feature space d, $F-measure$ and $G-mean$ have the best performance on the four experimental datasets . Experimental setup and results

In order to reduce the computational complexity and maintain the performance of our proposed algorithm, λ is set to 20% of the dimension of feature space d of the four experimental datasets in the following experiments.

(2) Performance comparisons

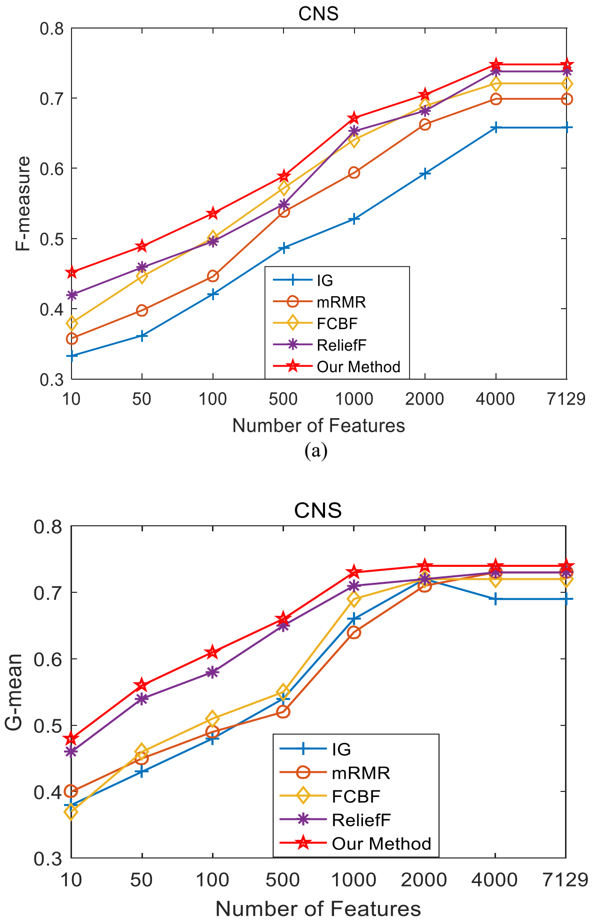
We compare our proposed feature selection algorithm with four state-of-the-art algorithms on four high-dimensional DNA microarray datasets. We tune m, which is the number of selected features from original n-dimensional feature vector of each experimental datasets, and investigate the performance of our proposed algorithm and four existing algorithms. The experimental results are shown in Figures 1-4. On the whole, our proposed algorithm outperforms the four state-of-the-art feature selection algorithms in terms of $F-measure$ and $G-mean$. Specifically, on two small datasets COLON and CNS, our proposed algorithm performs better than IG, mRMR and FCBF, but it has no obvious advantage compared to ReliefF. The possible reason for this experimental result is that the performance of those algorithms is optimal on a smaller subset. In addition, our proposed algorithm is obviously better than the other four methods on datasets the OVARIAN and LUNG, which suggests that our algorithm is more suitable for S2SMLP.



(a)

(b)

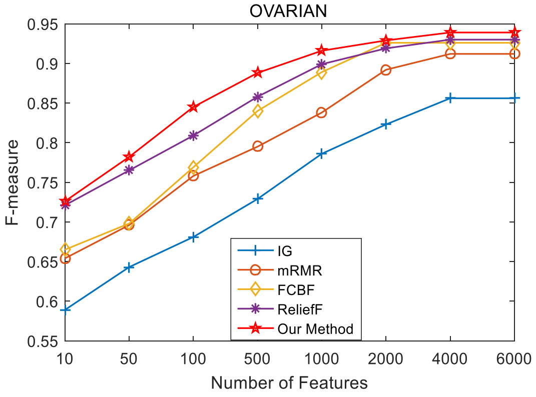
FIGURE 1. Comparison of experimental results on COLON dataset



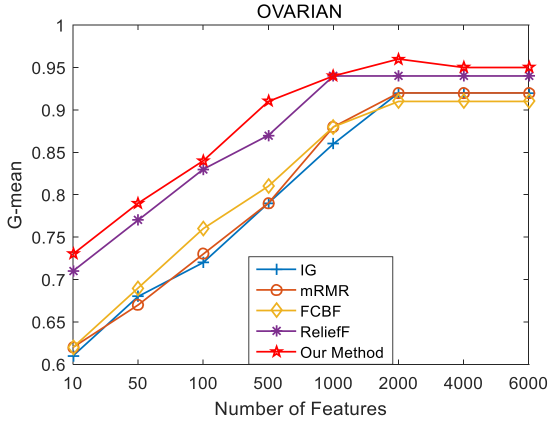
(a)

(b)

FIGURE 2.Comparison of experimental results on CNS dataset

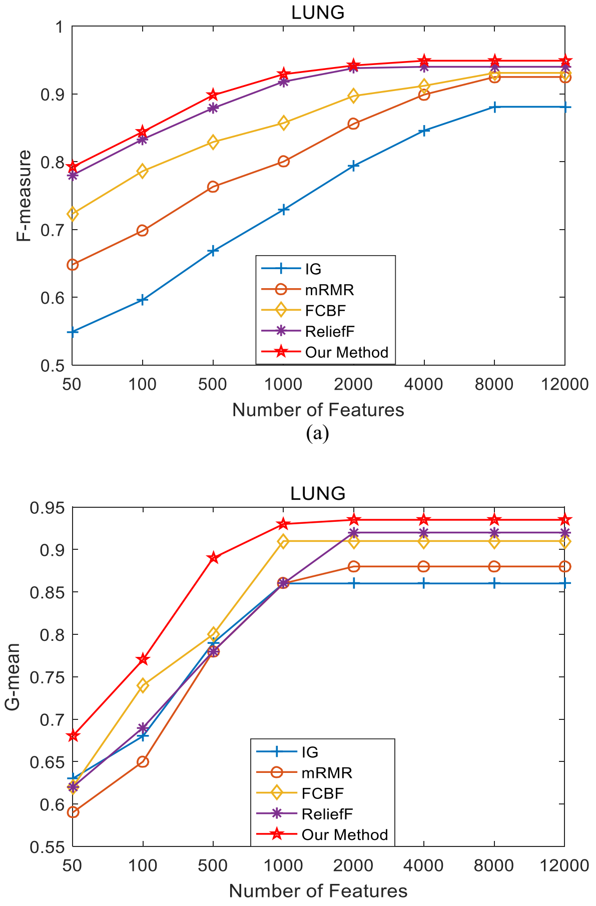


(a)



(b)

FIGURE 3. Comparison of experimental results on OVARIAN dataset



(a)

(b)

FIGURE 4. Comparison of experimental results on LUNG dataset.

V. CONCLUSION

Imbalance learning is widely used in many fields, such as churn prediction, abnormal activity recognition, software defect prediction, micro-blog sentiment analysis, multi-label learning, rating prediction, etc. In recent years, many researchers from the academics and industry have paid more and more attention to it. In order to address the issue of data explosion and feature selection for S2SMLP, in this paper, we propose a novel feature selection method based on the pairwise samples distance constraint, which considers the class label of the pairwise samples and selects these features which push two similar samples closer together and pull two different samples farther apart. Finally, we conduct experiments on four-high dimensional DNA microarray datasets. The experimental results demonstrate that our proposed algorithms outperform some state-of-the-art algorithms in terms of $F-measure$ and $G-mean$.

ACKNOWLEDGMENT

The work was supported by the National Natural Science Foundation of China (61602518, 71872180) and the Fundamental Research Funds for the Central Universities, Zhongnan University of Economics and Law(2722020JCT033).

REFERENCES

[1] Q. Yang, X. Wu, “10 challenging problems in data mining research,” International Journal of Information Technology and Decision Making, vol.5, no.4, pp.597-604,2006.

[2] R. Guermazi, I. Chaabane, M. Hammami, “AECID: asymmetric entropy for classifying imbalanced data,” Information Sciences, vol.467, pp.373- 397, 2018.

[3] F. Wu, X. Jing, S. Shin, W. Zuo, J.Yang, “Multiset feature learning for highly imbalanced data classification,” in Proc. of the thirty-first AAAI Conference on Artificial Intelligence(AAAI), pp. 15 83-1589, 2017.

[4] O. Loyola-Gonzalez, J. F. Martinez-Trinidad, J.A. Carrasco-Ochoa, “Study of the impact of resampling methods for contrast pattern based classifiers in imbalanced databases,” Neurocomputing, vol.175, pp.935-947, 2016.

[5] C.Lin, T. Hsieh, Y. Lin, “Minority Over sampling in Kernel Adaptive Subspaces for Class Imbalanced Datasets,” IEEE Transactions on Knowledge and Data Engineering, vol.30, no.5,pp. 950-962, 2018.

[6] S.Decherchi, W. Rocchia, “Import vector domain description: a kernel logistic one-class learning algorithm,” IEEE Transactions on Neural Networks and Learning Systems, vol.28, no.7, pp. 1722-1729, 2017.

[7] Z. Ferdowsi, R. Ghani, R. Settimi, “Online active learning with imbalanced Classes,” in Proc. of IEEE 13th International Conference on Data Mining (ICDM’13), pp.1043-1048, 2013.

[8] C. Cao, Z. Wang, “IMCStacking: cost-sensitive stacking learning with feature inverse mapping for imbalanced problems,” Knowledge-Based Systems, vol. 150, pp. 27-37, 2018.

[9] Y. Wang, D. Wang, N. Geng, Y. Wang, “Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection,” Applied Soft Computing, vol.77, pp.188-204, 2019.

[10] S. Dumpala, R. Chakraborty, S. Kopparapu, “A novel data representation for effective learning in class imbalanced scenarios,” in Proc. of the Twenty-seventh International Joint Conference on Artificial Intelligence (IJCAI), pp.2100-2106, 2018.

[11] M. Liu, C. Xu, Y. Luo, et all., “cost-sensitive feature selection by optimizing F-measures,” IEEE Transactions on Image Processing, vol.27, no.3, pp.1323-1335, 2018.

[12] S. Maldonado, R.Weber, F. Famili, “Feature selection for high dimensional class-imbalanced data sets using support vector machines,” Information Science, vol.286, pp.228-246, 2014.

[13] P. Zhou, X. Hu, P. Li, X. Wu., “Online feature selection for high-dimensional class-imbalanced data,” knowledge-Based Systems, vol. 136, pp.187-199, 2017.

[14] M. Alshawabkeh, J. A. Alsm, J. G. Dy, “Feature weighting and selection using hypothesis margin of boosting,” in Proc. of IEEE 12th International Conference on Data Mining, pp.41-50, 2012.

[15] K. Li, M. Yu, Y. Liu, T. Li, “Feature selection method based on weighted mutual information for imbalanced data,”. International Journal of Software Engineering and Knowledge Engineering, vol.28, no.8, pp.1177-1194, 2018.

[16] M. Alibeigi, S. Hashemi, A. Hamzeh, “DBFS: an effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets,” Data and Knowledge Engineering, vol.81-82, pp.67-103, 2012.

[17] A. Moayedikia, K.L. Ong, Y. Boo, W. Yeoh, “Feature selection for high dimensional imbalanced class data using harmony search,” Engineering Applications of Artificial Intelligence, vol.57, pp.38-49, 2017.

[18] S. Maldonado, J. Lopez, “Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification,” Applied Soft Computing,vol.67, pp. 94-105, 2018.

[19] H. Chen, T. Li, X. Fan, C. Luo C, “Feature selection for imbalanced data based on neighborhood rough sets,” Information Sciences, vol.483, pp.1-20, 2019.

[20] M. Hall, L. Smith, “Practical feature subset selection for machine learning,” Computer Sciences, vol.98, pp.181-191,1998.

[21] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Trans. Pattern Anal. Mach. Intell. 2005, 27 (8):1226-1238.

[22] L.Yu, H. Liu, “Feature selection for high-dimensional data: a fast correlation-based filter solution,” in Proc. of the 20th International Conference on Machine Learning, pp. 856-863, 2003.

[23] I. Kononenko, “Estimating attributes: analysis and extensions of relief,” in Proc.of European Conference on Machine Learning, pp.171-182, 1994.

[24] C. C. Aggarwal, A. Hinneburg, D.A. Keim, “On the surprising behavior of distance metrics in high dimensional space,” in Proc. of International Conference on Database Theory, pp. 420-434, 2001.