Postprint: KFDA-Boosting Algorithm for Imbalanced Data Classification

Authors: Wang Lai, Fan Chongjun, Yang Yunpeng, Yuan Guanghui

Date: 2018-05-02T00:00:00+00:00

Abstract

Imbalanced data distributions and nonlinear feature characteristics increase classification difficulty, particularly in recognizing minority classes within imbalanced datasets, thereby impacting overall classification performance. To address this issue, we propose the KFDA-Boosting algorithm, which combines the capability of Kernel Fisher Discriminant Analysis (KFDA) to effectively extract nonlinear feature representations with the ensemble learning concept of Boosting. To validate the effectiveness and superiority of the proposed algorithm for imbalanced data classification, we employ G-mean, precision, and recall of the minority class as evaluation metrics, conduct performance assessments on ten datasets from the UCI repository, and perform comparative experiments against six classification algorithms including Support Vector Machines. Experimental results demonstrate that, particularly for data with high imbalance ratios or nonlinear characteristics, the KFDA-Boosting algorithm can effectively identify minority classes and achieves significant classification performance with good stability compared to other classification algorithms.

Full Text

Preamble

KFDA-Boosting Algorithm Oriented to Imbalanced Data Classification

Wang Lai[1], Fan Chongjun[1]†, Yang Yunpeng[1], Yuan Guanghui[2] ,[2]

[1]Business School, University of Shanghai for Science & Technology, Shanghai 200093, China
[2] School of Information Management & Engineering, [2] Experimental Center, Shanghai University of Finance and Economics, Shanghai 200433, China

Abstract: The imbalance of data distribution and the nonlinearity of data characteristics increase the difficulty of classification, particularly in recognizing minority class samples in imbalanced datasets, thereby affecting overall classification performance. To address this problem, we propose the KFDA-Boosting algorithm, which combines the capability of Kernel Fisher Discriminant Analysis (KFDA) to effectively extract nonlinear features from samples with the ensemble learning concept of Boosting. To verify the effectiveness and superiority of the proposed algorithm for imbalanced data classification, we employ G-mean, precision, and recall of the minority class as evaluation metrics and select ten datasets from the UCI repository to test the KFDA-Boosting algorithm' s performance against six other classification algorithms, including Support Vector Machine. The results demonstrate that for imbalanced data classification, especially for data with high imbalance ratios or nonlinear characteristics, the KFDA-Boosting algorithm can effectively identify minority classes and achieves significantly better overall classification performance and stability compared to other algorithms.

Keywords: Kernel Fisher discriminant analysis; ensemble learning; imbalanced data; classification

0 Introduction

Accurately identifying minority class samples using their features to improve overall classification performance holds substantial value and significance for addressing imbalanced data classification problems. Imbalanced data classification has important applications in many domains, such as disease diagnosis, text recognition, and intrusion detection. Imbalanced data refers to datasets where one or some classes have far more samples than others. For imbalanced data classification, greater attention is paid to minority class samples, and the misclassification cost for minority classes is relatively high. Meanwhile, as data volume increases, relationships among data increasingly exhibit nonlinear characteristics, even strong nonlinearity, which further increases the difficulty of recognizing minority classes. Therefore, effectively utilizing sample features, particularly nonlinear features, to accurately identify minority class samples and thereby improve overall classification performance is of great value and significance for solving imbalanced data classification problems.

In recent years, research on imbalanced data classification has primarily focused on two levels: the data level and the algorithm level. At the data level, the main approach is to achieve balance among classes through resampling, including undersampling and oversampling. Research on resampling methods has mainly evolved around Laurikkala's neighborhood cleaning algorithm [1] and Chawla et al.' s SMOTE algorithm [2]. For instance, Zheng et al. [3] proposed a SMOTE-SVM traffic incident detection algorithm for imbalanced datasets, which uses the SMOTE algorithm to oversample incident data to reduce imbalance. Similarly,

Yi and Yang et al. [4, 5] first balanced sample sizes across classes through improved SMOTE algorithms before applying classification algorithms to handle imbalanced data. However, oversampling may introduce additional noise, while undersampling may lose some useful and important information.

At the algorithm level, the main approaches include cost-sensitive learning and ensemble learning methods. Cost-sensitive learning assigns different costs to different types of errors to achieve the goal of minimizing the total misclassification cost [6, 7]. For example, Zou et al. [8] designed a cost-sensitive decision tree algorithm for imbalanced data in customer value segmentation to achieve effective customer value identification. Shi et al. [9] proposed a classification algorithm combining cost-sensitive learning with random forests for imbalanced datasets. Ensemble learning mainly includes Boosting [10] and Bagging [11] algorithms. Research on using ensemble learning to solve imbalanced data classification problems has mostly involved improvements based on the Boosting algorithm proposed by Freund and Schapire [12, 13, 14]. For instance, Ying et al. [15] incorporated LDA into the Boosting algorithm to build weak classifiers for customer churn prediction. Although LDA-Boosting improved classification efficiency, it could not achieve ideal results for classifying imbalanced data with nonlinear features. Li et al. [16] used KNN as weak classifiers, applied BPSO for feature extraction, and then used the Adaboost-KNN algorithm for classification, but the selection of optimal feature subsets was prone to falling into local optima, thereby affecting the final classification results.

The above two levels have not fully considered the effective utilization of sample features, particularly nonlinear features. To address this problem, considering that Boosting, as an effective classification learning method, assigns higher weights to difficult-to-learn samples, causing classifiers to focus on those samples in subsequent training and thereby improving classification performance for imbalanced data to some extent, and that Kernel Fisher Discriminant Analysis can very effectively extract nonlinear features, we combine these two algorithms and propose the KFDA-Boosting algorithm. The KFDA-Boosting algorithm uses Kernel Fisher Discriminant Analysis to effectively extract nonlinear discriminant features and leverages the ensemble learning concept of Boosting to improve its classification performance. Finally, we conduct simulation experiments on ten datasets selected from the UCI repository to test the feasibility and effectiveness of the KFDA-Boosting algorithm for imbalanced data classification and compare its performance with six other classification algorithms, hoping to demonstrate the algorithm's effective identification of minority classes and improved overall classification performance.

1.1 Boosting Algorithm Concept

This paper primarily uses binary classification problems as examples to elaborate on the Boosting algorithm concept [17, 18]. Given a weak classifier and training

set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where x_i is an n-dimensional column vector representing the i-th sample, and $y_i \in Y = \{+1, -1\}$ represents the class label of the i-th sample.

First, each sample in the training set is assigned an initial weight. Then, in each iteration, a weak hypothesis $h_t : X \to \{-1, 1\}$ is generated, and sample weights are updated accordingly—that is, increasing the weights of misclassified samples and decreasing the weights of correctly classified samples. This causes the weak classifier to focus on previously misclassified samples in the next iteration. After T rounds of iteration, the weak hypotheses generated in each round are weighted according to their classification accuracy to obtain the final hypothesis.

1.2 Improved Kernel Fisher Discriminant Analysis

The basic idea of Kernel Fisher Discriminant Analysis is to nonlinearly map an input space $\mathbb{R}^n$ to another feature space F, and then use Fisher discriminant analysis in the feature space to achieve classification of the input space [19, 20].

Considering the characteristics of the Boosting algorithm, this paper improves the original Kernel Fisher Discriminant Analysis by assigning corresponding weights to each sample. The specific approach and process are as follows.

In the feature space F, each sample' s corresponding weight is $D_t(i)$, which is regarded as the weight of the i-th sample in the t-th round of learning. In the transformed space F, the mean of each class sample $\tilde{m}_{t,\phi}$ and the within-class scatter matrix $S^{\phi}_{t,W}$ are respectively:

$$\tilde{m}_{t,\phi} = \sum_{i:y_i=k} D_t(i)\phi(x_i), \quad k = \pm 1$$

$$S^{\phi}_{t,W} = \sum_{k=\pm 1} \sum_{i:y_i=k} D_t(i)(\phi(x_i) - \tilde{m}_{t,\phi})(\phi(x_i) - \tilde{m}_{t,\phi})^T$$

From the above $\tilde{m}_{t,\phi}$ and $S^{\phi}_{t,W}$, the between-class scatter matrix $S^{\phi}_{t,B}$ and within-class scatter matrix $S^{\phi}_{t,W}$ can be expressed as:

$$S^{\phi}_{t,B} = (\tilde{m}^{\phi}_{t,+1} - \tilde{m}^{\phi}_{t,-1})(\tilde{m}^{\phi}_{t,+1} - \tilde{m}^{\phi}_{t,-1})^T$$

$$S^{\phi}_{t,W} = \sum_{k=\pm 1} \sum_{i:y_i=k} D_t(i)(\phi(x_i) - \tilde{m}^{\phi}_{t,k})(\phi(x_i) - \tilde{m}^{\phi}_{t,k})^T$$

According to the Fisher discriminant criterion, the expression of the Fisher criterion function at this time is:

$$J_t^{\phi}(w) = \frac{w^T S_{t,B}^{\phi} w}{w^T S_{t,W}^{\phi} w}$$

Thus, the Fisher discriminant function in the feature space is obtained to achieve Fisher discriminant. However, if the dimensionality of feature space F is very high or even infinite, the optimal discriminant vector cannot be solved directly through the above formula. To address this problem, we introduce kernel functions. This paper uses the RBF kernel function (i.e., Gaussian radial basis kernel function) as the mapping function:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{\sigma^2}\right)$$

where x, y are corresponding sample values, and σ is a constant that determines the degree of nonlinearity.

According to the reproducing kernel theory, any solution in the high-dimensional space can be expressed as a linear combination of training samples in that space, that is:

$$w = \sum_{i=1}^{m} \alpha_i \phi(x_i) = \Phi\alpha$$

where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)^T \in \mathbb{R}^m$ is the linear coefficient of each element $\phi(x_i)$.

By projecting the mean $\tilde{m}_{t,\phi}$ of the training samples in feature space F onto w, we can obtain $\tilde{m}_{t,\phi}^T w$. Then, the numerator and denominator of the right fraction in Equation (6) can be transformed into the following forms:

$$w^T S_{t,B}^{\phi} w = \alpha^T M_t H M_t^T \alpha$$

$$w^T S_{t,W}^{\phi} w = \alpha^T H_t \alpha$$

where M_t and H_t are matrices derived from the weighted samples.

Combining Equations (6), (11), and (13), we can obtain the Fisher discriminant in the feature space. According to the properties of the generalized Rayleigh quotient, the solution can be obtained by solving the eigenvalue problem. The projection of feature space $\phi(x)$ onto w is:

$$w^T \phi(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x)$$

1.3 KFDA-Boosting Algorithm Flow

In imbalanced data, the minority class corresponds to the positive class and the majority class corresponds to the negative class. By incorporating the improved Kernel Fisher Discriminant Analysis into the Boosting algorithm framework, we obtain the KFDA-Boosting algorithm flow as follows.

For the training set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where y_i represents the label of the i-th sample, $y_i \in Y = \{+1, -1\}$; the weak learning algorithm is the improved KFDA; the number of iterations is T; and the distribution of the i-th sample in the t-th iteration is denoted as $D_t(i)$.

gives the confusion matrix for binary classification problems.

In traditional classification algorithms, classification accuracy is usually adopted as the performance evaluation metric, that is:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Typically, in imbalanced data, the number of positive class samples is small, so TP will not be large and may even be 0, while TN is large, making the final classification accuracy relatively high but ignoring the classifier' s correct recognition rate for the positive class. Therefore, accuracy cannot truly reflect classifier performance in a meaningful way.

Given the above limitations of classification accuracy, this paper uses the G-mean value as the overall classification performance evaluation metric:

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

where $TPR = \frac{TP}{TP+FN}$ and $TNR = \frac{TN}{TN+FP}$. The G-mean value, as a commonly used evaluation metric for imbalanced data classification, measures the classification performance of both positive and negative classes using TPR and TNR respectively. Larger values indicate better classification performance. If either of these two values is poor, the G-mean value will be unsatisfactory.

To further measure the classification performance on the positive class, we introduce precision and recall for the positive class, defined as follows:

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

KFDA-Boosting Classification Algorithm

Input: Training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where $x_i \in X$, $y_i \in Y = \{+1, -1\}$; number of iterations T.

Step 1: Initialize sample weights: $D_1(i) = \begin{cases} \frac{1}{2N_+} & \text{if } y_i = +1 \\ \frac{1}{2N_-} & \text{if } y_i = -1 \end{cases}$, where N_+ and N_- are the total numbers of positive and negative samples respectively.

Step 3: Train weighted KFDA for sample distribution D_t: Obtain the projection direction α_t.

Step 6: Obtain weak classifier $h_t(x) = \begin{cases} +1 & \text{if } w_t^T \phi(x) > \theta_t \\ -1 & \text{otherwise} \end{cases}$, where threshold θ_t is determined by the weighted average of the projections of class means onto α_t, with weights being the corresponding class sample counts.

Step 14: Update sample weights: for $i = 1$ to m, $D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$, where Z_t is a normalization factor.

Output: Final hypothesis $H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$.

---

3 Experiments and Results

3.1 Data Sources

We selected ten datasets from the UCI repository as test data. To treat the selected datasets as binary classification problems, we made the following provisions: if a dataset has two classes, the class with fewer samples is designated as the positive class (e.g., Sonar, Ionosphere datasets); if a dataset has multiple classes (more than 2), one class is designated as the positive class and the remaining classes are merged as the negative class. After these provisions and processing, the imbalanced datasets for binary classification are obtained and sorted by imbalance level (IL) in ascending order, as shown in .

3.2 Data Preprocessing

To prevent excessive differences among attribute values from affecting the algorithm' s iteration process, we normalized the raw data before conducting algorithm experiments. For any feature attribute in the data table, we selected the maximum value of that feature attribute and then divided all samples' values of that attribute by this maximum value to obtain the normalized values for each sample. The calculation formula is as follows:

$$x'_{ij} = \frac{x_{ij}}{\max_j(x_{ij})}, \quad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, N$$

where $\max_j(x_{ij})$ represents the maximum value of the j-th feature attribute. After this processing, the feature attribute values of each sample are converted to $[0, 1]$, which also eliminates the influence of measurement units and facilitates subsequent iterative calculations.

3.3 Comparative Experiments and Results Analysis

This experiment adopted five-fold cross-validation by randomly dividing each dataset into five equal parts. Each time, one part was used as the test set while the other four parts served as the training set. Finally, the average values of the evaluation metrics (including accuracy, G-mean, minority class precision, and recall) obtained from the five experiments were taken as the final evaluation results for the algorithm. Although accuracy has limitations for evaluating imbalanced data classification, we calculated this metric mainly for comparative purposes.

To verify the effectiveness of the KFDA-Boosting algorithm for imbalanced data classification, we conducted comparative experiments with six other algorithms: Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), Kernel Fisher Discriminant Analysis (KFDA), Cost-Sensitive Decision Tree (CS-DT), and SMOTE-SVM. SVM, KFDA, and KFDA-Boosting use the same kernel function (RBF kernel), and the maximum number of iterations is set to 200. For comparison, the parameter settings for SMOTE-SVM and CS-DT follow those in references [3] and [8] respectively.

The final results of accuracy, G-mean, positive class precision, and recall for the seven algorithms are shown in through . To more intuitively compare and analyze the classification effects of each algorithm, the test results in Tables 3-6 are plotted as corresponding line charts in [Figure 1: see original paper] through [Figure 4: see original paper].

As can be seen from Figures 1 and 2, compared with DT, SVM, ANN, and two improved classification algorithms (CS-DT, SMOTE-SVM), the G-mean value of the proposed KFDA-Boosting algorithm is the highest on five datasets in , while on other datasets, the G-mean values are not far from the corresponding maximum values, indicating that the proposed algorithm has good overall classification performance. When the imbalance ratio of the dataset gradually increases, compared with traditional DT and SVM algorithms, CS-DT and SMOTE-SVM show varying degrees of improvement in G-mean values across datasets, leading to better overall classification performance. When the imbalance ratio increases to a certain extent, the proposed algorithm still maintains a large G-mean value and is relatively superior to the corresponding values of CS-DT and SMOTE-SVM improved algorithms, such as on the Page-Blocks and Yeast test sets. Meanwhile, although the accuracy of DT, SVM, and ANN algorithms all reaches above 90%, their corresponding G-mean values are relatively small.

Furthermore, Figures 3 and 4 show that as the imbalance ratio gradually in-

creases, the average values of both positive class precision and recall of KFDA-Boosting on corresponding datasets are large, while the corresponding values of other algorithms are relatively small. This situation is particularly evident in tests on imbalanced data with nonlinear characteristics, such as the Yeast dataset, indicating that the proposed algorithm can effectively utilize the nonlinear features of samples and has strong recognition ability for minority class samples. This also confirms from another perspective that classification accuracy sometimes cannot effectively measure classifier performance for imbalanced data classification.

Additionally, it can be observed that for the same dataset, the G-mean value and positive class precision of the KFDA-Boosting algorithm are greater than the corresponding values of KFDA. Except for the Ionosphere dataset, the positive class recall of the proposed algorithm is greater than that of KFDA or comparable to the optimal value, indicating that the proposed algorithm significantly improves classification performance compared with using KFDA alone, particularly for positive class samples.

Moreover, we calculated the variance of G-mean values for various classification algorithms as 0.1173, 0.0889, 0.0263, 0.0063, 0.0069, 0.0039, and 0.0025 respectively. This demonstrates that for different datasets, the proposed KFDA-Boosting algorithm exhibits better overall classification stability compared with other classification algorithms.

---

4 Conclusion

For imbalanced data classification, where data characteristics increasingly exhibit nonlinearity, making minority class recognition more difficult, this paper proposes a classification method based on improved Kernel Fisher Discriminant Analysis and Boosting algorithm, namely the KFDA-Boosting algorithm. This algorithm can effectively utilize sample features, especially nonlinear features, to achieve nonlinear discrimination of original data and ensure optimal separability of samples. Finally, through testing on ten datasets from the UCI repository, the experiments demonstrate that for data with large imbalance ratios or nonlinear characteristics, the proposed algorithm achieves remarkable classification performance. Compared with DT, SVM, ANN, KFDA, CS-DT, and SMOTE-SVM, the KFDA-Boosting algorithm can effectively identify minority classes, exhibits good overall classification performance, and demonstrates better stability. This proves the feasibility and effectiveness of the algorithm in handling imbalanced data classification problems.

To expand the applicability of the proposed algorithm to imbalanced data classification, future research will consider multi-class problems and corresponding evaluation metrics, and further improve the classification performance of the algorithm for imbalanced data.

References

[1] Laurikkala J. Improving identification of difficult small classes by balancing class distribution [C]// Proc of the 8th Conference on AI in Medicine. 2001.

[2] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority oversampling technique [J]. Artificial Intelligence Research, 2002, 16 (3): 321-357.

[3] Zheng Wenchang, Chen Shuyan, Wang Xuanqiang. SMOTE-SVM traffic incident detection algorithm for imbalanced datasets [J]. Journal of Wuhan University of Technology, 2012, 34 (11): 58-62+123.

[4] Yi Baiheng, Zhu Jianjun, Li Jie. Imbalanced SVM classification of credit risk for small loan companies based on improved SMOTE [J]. Chinese Journal of Management Science, 2016, 24 (03): 24-30.

[5] Yang Yi, Lu Chengbo, Xu Genhai. A refined Borderline-SMOTE method for imbalanced datasets [J]. Journal of Fudan University: Natural Science Edition, 2017, 56 (05): 537-544.

[6] Jiang Shengyi, Xie Zhaoqing, Yu Wen. Research on cost-sensitive naive Bayes classification for imbalanced data [J]. Journal of Computer Research and Development, 2011, (S1): 387-390.

[7] Li Yong, Liu Zhandong, Zhang Haijun. Survey of ensemble classification algorithms for imbalanced data [J]. Application Research of Computers, 2014, 31 (5): 1287-1291.

[8] Zou Peng, Mo Jiahui, Jiang Yihua, et al. Customer value segmentation based on cost-sensitive decision trees [J]. Management Science, 2011, 24 (2): 20-29.

[9] Shi Yanwen, Wang Hongjie. Cost-sensitive random forest classifier based on new impurity measure [J]. Computer Science, 2017, 44 (S2): 98-101.

[10] Schapire R E. The strength of weak learnability [C]// Proc of the 2nd Annual Workshop on Computational Learning Theory. 1989: 197-227.

[11] Breiman L. Bagging predictors [J]. Machine Learning, 1996, 24 (2): 123-140.

[12] Li K, Fang X, Zhai J, et al. An imbalanced data classification method driven by boundary samples-boundary-boost [C]// Proc of International Conference on Information Science and Control Engineering. 2016: 194-199.

[13] Hu Xiaosheng, Wen Juping, Zhong Yong. Dynamic balanced sampling based ensemble classification method for imbalanced datasets [J]. CAAI Transactions on Intelligent Systems, 2016, 11 (02): 257-263.

[14] Qin Mengmei, Qiu Jianlin, Lu Pengcheng, et al. AdaBoost-based class imbalance learning algorithm [J]. Application Research of Computers, 2017, 34 (11): 3229-3232+3254.

[15] Ying Weiyun, Lin Nan, Xie Yaya, et al. Customer churn prediction using LDA Boosting algorithm [J]. Journal of Applied Statistics and Management,

2010 (3): 400-408.

[16] Li Yijing, Guo Haixiang, Li Yanan, et al. A Boosting-based ensemble learning algorithm for classification in imbalanced data [J]. Systems Engineering—Theory & Practice, 2016 (1): 189-199.

[17] Wang Lulin. Research on Boosting classification algorithm for imbalanced samples [D]. Harbin: Harbin Institute of Technology, 2013.

[18] Li Xiang. Application and research of Boosting classification algorithm [D]. Lanzhou: Lanzhou Jiaotong University, 2012.

[19] Chang Zhipeng, Cheng Longsheng. Multi-parameter automatic optimization algorithm for kernel Fisher discriminant analysis [J]. Systems Engineering and Electronics, 2013, (01): 212-217.

[20] Li Jianyun, Qiu Wanhua. Kernel Fisher discriminant analysis method for evaluating consumer credit risk [J]. Systems Engineering—Theory Methodology Applications, 2004 (6): 548-552+556.

Note: Figure translations are in progress. See original paper for figures.