A Quantile K-means Bayes Approach to Classification for Imbalanced Data

Yanzhu Hu*, Xinghao Zhao and Song Wang

School of Automation, Beijing University of Posts and Telecommunication, Beijing, China

*Corresponding author e-mail: wongsangwongsang@163.com, yzhu@263.net, zhxh@bupt.edu.cn

Abstract. This paper focuses on the classification of imbalance data. In Machine Learning, a data set is imbalanced when the class proportions are highly skewed. A natural way of handling imbalanced data is to attempt to equalise the class frequencies and train the classifier of choice on balanced data. A new approach called Quantile K-means Bayes was proposed to solve the problem. The first focus is on a modified q-classifier. The second focus is on combine the k-means and Bayes algorithm using the data density. The proposed approach is evaluated by 101 benchmark data sets from KEEL collection. A comparison of the proposed approach and other conventional approaches is presented in terms of the G-mean. It can be seen that the proposed approach is able to acquire good performance among the other conventional approaches do. Therefore, this novel approach is an added value for the classification problem for imbalance data.

1. Introduction

Class imbalance arises when the number of examples belonging to one class is much greater than the number of examples belonging to another [1]. Imbalanced data sets are common in fields such as bioinformatics, medicine, security, finance, software development, and satellite imaging, DNA sequences, gene recognition, engineering [2]. The class of interest is usually the minority class, standard classifiers are driven by accuracy and operate under the assumption that the data sample is a faithful representation of the population of interest, which often results in the minority class being ignored [3].

The existing methods for classification of imbalanced data can be categorised as follows[1]: The algorithm-level category, The data-level category, The cost-sensitive methods and the classifier ensembles Examples include decision tree algorithms insensitive to the class sizes, like Hellinger Distance Decision Tree (HDDT) [4], and a SVM classifier with different penalty constants for different classes Oversampling [5], SMOTE [6], These techniques can be jointly applied to increase the size of the minority class while simultaneously decreasing the majority. The best known methods are the cost-sensitive versions of AdaBoost: AdaCost [7]. In general, according to [8], algorithm level and cost-sensitive approaches are more data-dependent, whereas data level and ensemble learning methods are more versatile. However, the above algorithms ignore the density information in feature space. Moreover, the calculation is complex, which leads to low classification accuracy.
Therefore, the new processing way was proposed that can be used to deal with the imbalance data, for two-class imbalanced classification task, based on modified k-means. Before that it makes use of the q-classifier and for the sake of utilizing the data density for class labels. In the end Bayes algorithms be used to achieve accurate classification.

2. Related Definition
Here list some definitions of imbalanced dataset and notation above all, which will be used in the following chapters. The learning data is expressed as \( D_i = \{(X_i, Y_i) | i \in 1, 2, \ldots, N\} \), where \( X_i \) represents \( m \)-dimensional feature and \( Y_i \in \{0, 1\} \) represents the output response and that is a binary ordinal response. In general, researchers use imbalance ratio (IR) to quantify the degree of imbalance. It is assuming that the majority class labels are 0 and the minority class labels are 1, and assuming that \((X, Y)\) denote an independent generic data point with distribution \( P \) as well as. The goal of the method for imbalance data is to make an accurate prediction for the minority data, which can classify it to label 0 or label 1 precisely. In other words, build an accurate classifier for \( Y \) given \( X = x \) when the learning data is imbalanced.

First of all, the definition of imbalance ratio (IR) is the proportion of the cardinality of the majority and minority samples \([1]\), which is \( \frac{N_0}{N_1} \), when the dataset satisfy \( IR \gg 1 \) the imbalance occurs.

3. Method
As mentioned above, the classification error plays an important role in the algorithms to for the classifier in many methods. In lots of literatures adopt measures at the expense of misclassifying minority class samples to classify majority correctly, which exits problems obviously. Thus a better approach is to achieve good performance under both type of classification error.

3.1. The q*-classifier
The goal of classifying imbalance data is to achieve an optimal result of the output. Thus, denoting q*-classifier as \( \delta_{q^*} \) in the first place, which assigns an instance \( x \) to the minority class if its data density for minority class labels, \( f_{X|Y}(x|1) \), is larger than the data density for majority class labels, \( f_{X|Y}(x|0) \):

\[
\delta_{q^*}(x) = 1_{\{f_{X|Y}(x|1) \geq f_{X|Y}(x|0)\}}
\]

It is easy to see that the classifier is based on the conditional density of the features, \( f_{X|Y} \), what’s the meaning of this step is to removes the effect of the prevalence of the minority class labels.

By this way can it be able to handle imbalance, which is convenient to describe the classifier based on density theoretically in terms of the conditional density of the data. For the sake of reducing the difficulty to realize it, it can be rewritten the \( \delta_{q^*}(x) \) into the follow form:

\[
\delta_{q^*}(x) = 1_{\{\Delta_D(x) = 1\}}, \quad \Delta_D(x) = \frac{f_{X|Y}(x|1)}{f_{X|Y}(x|1)} \frac{p(x)(1-\pi)}{p(x)}
\]

Where, \( \pi = P\{Y=1\} \) represent the marginal probability and \( p(x) = P\{Y=1|X=x\} \) is conditional class probability function.

3.2. Naive Bayes algorithm
In order to overcome the difficulty mentioned above, the propose way made full use of the Bayes method to calculate the probability function. Firstly, cluster the majority \( X \) into k sub-class by K-means as \( X = [c_1; c_2; \ldots; c_k] \), however, the size of one sub-class may be much larger than the other sub-class so that the clustering result obtained by original K-means algorithm may introduce new imbalance. It is common practice to introduce a threshold \( T \) to ensure the number of data in sub-class are not much
different. It could be called Threshold K-means algorithm. Up to now, the Naive Bayes algorithm can be used. As the follow equation show:

\[ P(c | x) = \frac{P(c)P(x | c)}{P(x)} \]

Where \( P(c) \) the priori probability of class is, \( P(x | c) \) is the conditional probability of the sample \( x \) belonging to \( c \). \( P(x) \) is a constant factor after normalization. It’s necessary to note that all of this is based on the assumption that features are independent. After that the formula above can be computed as following show:

\[ P(c | x) = \frac{P(c)P(x | c)}{P(x)} = \frac{P(c)}{P(x)} \prod_{i=1}^{m} P(x_i | c) \]

Where \( x_i \) prefers to the value of the \( i^{th} \) variable of \( x \), \( P(x_i | c) \) is the conditional probability of the \( i^{th} \) variable. It is assumption that \( x_i \) obeys Gaussian distribution \( P(a_i | c) \sim N(\mu_{c,i}, \sigma_{c,i}^2) \), where \( \mu_{c,i}, \sigma_{c,i}^2 \) refer to the mean and variance of the \( i^{th} \) variable. \( P(x_i | c) \) is able to calculate as follow [9]:

\[ P(x_i | c) = \frac{1}{\sqrt{2\pi\sigma_{c,i}}} \exp \left( -\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2} \right) \]

Thus, the classification result can be achieved as follows:

\[ h(x) = \arg \max_{c \in \mathcal{F}} P(c) \prod_{i=1}^{m} P(x_i | c) \]

4. Experiments

4.1. Performance Metrics for Imbalanced Data

As for the Imbalance problems, our goal is to find a classifier that achieves both high TNR and TPR values. The class of interest is chosen to be the “positive” class (\( \omega^+ \)) and the other class is the “negative” class (\( \omega^- \)) [1]. The confusion matrix for a given classifier which assigns the respective labels + and − is:

| True labels | + | - |
|-------------|---|---|
| +           | A | B |
| -           | C | D |

Table 1. The matrix of a given classifier

In order to evaluate the performance for Imbalance data, the G -mean was used to estimate it, which is the geometric mean of TNR and TPR, as the follow equation show:

\[ GM = \sqrt{TNR \times TPR} = \sqrt{\frac{AD}{(A+D)(C+D)}} \]

The advantages is to replace misclassification rate in imbalanced data. Because of that the G -mean is close to 1 only when both the true negative and true positive rates are close to 1 and the difference between the two is small. Our goal is to find a classifier that achieves both high TNR and TPR values in imbalance problems.
4.2. Data sets
The data sets from the KEEL data collection was used to test methods [10], in which contains 101 binary imbalanced datasets. The data sets of the repository are not all completely independent of each other, however, it’s the most widely used and authoritative date sets. Many characteristics of the data sets as the follow Table 2 shows:

Table 2. Three Scheme comparing Characteristics of the data sets from the KEEL collection.

| Name         | #Attributes (R/I/N) | #Examples | IR  | Name         | #Attributes (R/I/N) | #Examples | IR  |
|--------------|--------------------|-----------|-----|--------------|--------------------|-----------|-----|
| abalone-17_vs_7-8-9-10 | 8 (7/0/1)         | 2338      | 39.3| dermatology-6 | 34 (0/34/0)     | 358       | 16.9 |
| abalone19    | 8 (7/0/1)         | 4174      | 129 | ecoli-0 vs 1 | 7 (7/0/0)        | 220       | 1.86 |
| abalone-19_vs_10-11-12-13 | 8 (7/0/1)     | 1622      | 49.7| ecoli-0-1_vs_2-3-5 | 7 (7/0/0)    | 244       | 9.17 |
| abalone-20_vs_8-9-10 | 8 (7/0/1)      | 1916      | 72.7| ecoli-0-1_vs_5 | 6 (6/0/0)       | 240       | 11  |
| abalone-21_vs_8 | 8 (7/0/1)       | 581       | 40.5| ecoli-0-1-3-7_vs_2-6 | 7 (7/0/0)   | 281       | 39.1 |
| abalone-3_vs_11 | 8 (7/0/1)      | 502       | 32.5| ecoli-0-1-4-6_vs_5 | 6 (6/0/0)    | 280       | 13  |
| abalone9-18 | 8 (7/0/1)        | 731       | 16.4| ecoli-0-1-4-7_vs_2-3-5-6 | 7 (7/0/0)  | 336       | 10.6 |
| car-good    | 6 (0/0/6)        | 1728      | 24  | ecoli-0-1-4-7_vs_5-6 | 6 (6/0/0)   | 332       | 12.3 |
| car-vgood   | 6 (0/0/6)        | 1728      | 25.6| ecoli-0-2-3-4_vs_5 | 7 (7/0/0)    | 202       | 9.1  |
| cleveland-0_vs_4 | 13 (13/0/0)   | 177       | 12.6| ecoli1       | 7 (7/0/0)       | 336       | 3.36 |

4.3. Experiments Results
An experiment was designed to explore the performance of our proposed theoretical findings in the above section with respect to the imbalance date. The aim is to test the theory and verify whether the results have better G-mean results or not among some popular ways.

As for the experimental protocol, it is used 10- fold cross-validation repeated 250 times in the whole 101 benchmark data sets. Each data set is split into ten sets of equal size; nine is used for training and the other for testing, and repeat this method until all data is processed. The G-mean for each procedure is reported in Figure 2. The cross-validation folds are drawn using stratified sampling so that the imbalance ratio is mirrored in the training and testing parts.

![Figure 1. Impact of class imbalance on minority class performance](image-url)
From figure 1, it can be seen that the error rate of the minority class are more than ten times as much as that of the majority class when the IR arrive at in between 5:1 and 10:1. Even if the IR during 1:1 and 3:1, the problem is still obvious.

![Some G-mean Results](image_url)

**Figure 2.** The values of G-mean for some methods

As for the performance for the imbalance data compared to other already existed methods are shown in the Figure 2. It is observed that the proposed approach has similar presentation with respect to others in the same Date sets, which is obviously demonstrated on the line graph. It appears that in the first and the sixth unit, it is the best among the ways.

5. Conclusion

Imbalanced data sets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed by two classes: The majority (negative) class and the minority (positive) class. These types of sets suppose a new challenging problem for Data Mining, since standard classification algorithms usually consider a balanced training set and this supposes a bias towards the majority class.

A new processing technique was proposed to classify the imbalance data. Firstly, combining q*-classifier, the classifier is based on the conditional density of the features, so that is can removes the effect of the prevalence of the minority class labels. In the next step, introducing the K-means algorithm into classification problem and before that join a T to ensure the number of data in sub-class are not much different. In the end, applying the Naive Bayes algorithm to compute the density probably.

Despite the idea May simplicity someway, the methods have proved competitive when compared with other state-of-the-art approaches. In the future the proposed approach will extend the ideas in this article to multiple-class unbalanced problems.

Acknowledgments

This work was partially supported by Beijing Science and Technology Planning Project (No. Z191100001419001), Beijing Natural Science Foundation (No.4192042) and the National Key Technology R&D Program of China (No.2015BAK40B03).

References

[1] Kuncheva L I, Arnaiz-González, Álvar, Diez-Pastor, José-Francisco, et al. Instance Selection
Improves Geometric Mean Accuracy: A Study on Imbalanced Data Classification [J]. 2018.

[2] Lu C, Ke H, Zhang G, et al. An improved weighted extreme learning machine for imbalanced data classification [J]. Memetic Computing, 2019, 11 (1): 27 - 34.

[3] Chen G, Liu Y, Ge Z. K-means Bayes algorithm for imbalanced fault classification and big data application[J]. Journal of Process Control, 2019, 81: 54 - 64.

[4] D. A. Cieslak, N.V. Chawla, Learning decision trees for unbalanced data, in: Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases – Part I, ECML PKDD ’08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 241 - 256, http://dx.doi.org/10.1007/978-3-540-87479-9 - 34.

[5] G. Batista, R. Prati, M. Monard, A study of the behavior of several methods for balancing machine learning training data, ACM SIGKDD Explor. Newslett. 6 (1) (2004) 20 – 29.

[6] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321 – 357.

[7] W. Fan, S.J. Stolfo, J. Zhang, P.K. Chan, AdaCost: misclassification cost-sensitive boosting, in: Proceedings of the Sixteenth International Conference on Machine Learning, ICML ’99, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1999, pp. 97 - 105.

[8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid- based approaches, IEEE Trans. Syst. Man Cybern., Part C: Appl. Rev. 42 (4) (2012) 463 - 484, http://dx.doi.org/10.1109/TSMCC.2011.2161285.

[9] O’Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data [J]. Pattern recognition, 2019, 90: 232 - 249.

[10] Information on https://sci2s.ugr.es/keel/imbalanced.php#subA.