DOI: 10.23977/jeis.2021.060202
ISSN 2371-9524

Comparison of binary classifier and outlier detection in different equilibrium

Fujie Sun*, Yinjie Tang, Zhaohao Wu

South China Agricultural University, Guangzhou, Guangdong, 510642

*Corresponding author

Keywords: outlier detection, binary classifier, unbalanced data set

Abstract: In the fields of financial risk control and mechanical production, the data sets with abnormal problems are always extremely unbalanced, because the most of abnormal problems occur hardly. Using this unbalanced data set to train the binary classifier, the result is often not ideal. Although there are Ensemble Learning and Grid Search methods to improve the F1 and accuracy of the classifier, in order to simplify the model, it is better to regard this financial risk control problem as an outlier detection problem than a binary classification problem. This paper uses the public data set on Kaggle, and compares the performance of the commonly used binary classification algorithms including Bayesian, Decision Tree, Random Forest, Logistic Regression Classifier, K-Nearest Neighbor (KNN), AdaBoost, One-Class SVM, Isolation Forest and Local Outlier Factor on balanced and unbalanced data sets respectively. According to the experimental results, this paper find that Bayesian is more suitable when the data set is small. Random Forest is more suitable for balanced data. For medium and large data sets with extremely unbalanced data, the effect of using One-Class SVM is better and more stable, and the effect of stable model is more important than that of unstable one.

1. Introduction

1.1 Problem introduction

In this paper, the experience of trial-and-error method is transformed into real experiments for model verification. Especially in the field of financial risk control, in extremely unbalanced data sets, outlier detection algorithm is often widely used instead of binary classification algorithm, because binary classifier can not classify as many companies with financial fraud as possible, and that will bring disaster to the stock and market. This paper uses simple binary classification algorithm and unsupervised outlier algorithm to compare balanced and unbalanced data sets respectively, and uses AUC, F1 scores, confusion matrix and normalized diagonal scores (see 1.2 for details) as model evaluation indicators.

1.2 Normalized diagonal scores

In the classification problem of extremely unbalanced data sets, F1 is generally high and can not correctly evaluate the pros and cons of classifiers. That leads to high accuracy, but can not find as many negative classes as possible.Therefore, according to the characteristics of confusion matrix, this paper creatively defines a method to evaluate the advantages and disadvantages of classifiers on unbalanced data sets, which is named normalized diagonal scores.

Since the confusion matrix of the ideal perfect classifier should be as shown in the figure below:

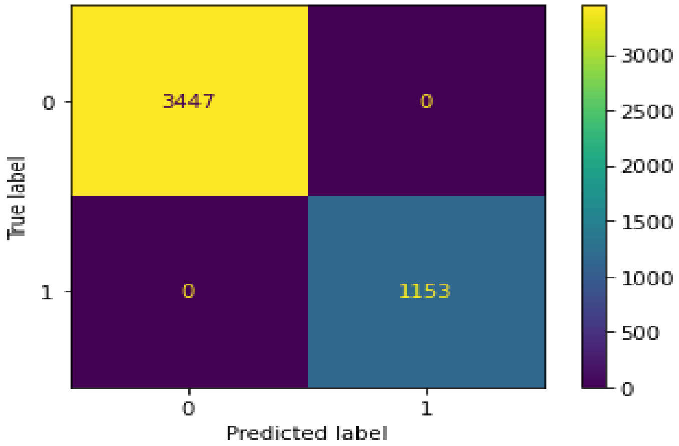

Figure 1: Confusion matrix

That is, all the numbers in the matrix may be concentrated in the main diagonal.

Let the confusion matrix K be:

$$K = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

$X = |K| = A \times D - B \times C$; The larger the X is, the better the classification performance of the model is.

When $X = sup(X)$, if and only if B = C = 0 and A and D are not equal to 0, it is the most ideal model.

However, considering the impact of data of different scales, the above indicators need to be further improved.

Let the confusion matrix Y = $\begin{bmatrix} g*A & g*B \\ g*C & g*D \end{bmatrix}, \forall g \in N$, The new indicators are:

$$f(X) = \frac{AD - BC}{(A+B+C+D)}$$

The results are as follows:

$$f(gX) = f(Y) = \frac{g^2AD - g^2BC}{g(A+B+C+D)} = \frac{g(AD-BC)}{(A+B+C+D)} = gf(X)$$

That is, the new evaluation index has the property of linear multiplication, and is not easy to be affected by the data scale.

For general cases, the key is to find g (which can be regarded as the maximum common factor of the four elements of the confusion matrix).

According to the properties of the greatest common factor, the essence of finding the greatest common factor of four elements is to find the intersection of the greatest common factor of the combination of two elements.The intersection is transitive, so it only need to find three combinations, such as: (A,B),(A,C),(A,D).

Therefore, the normalized diagonal scores have the following form:

$$\overline{X} = \frac{1}{g}\frac{AD - BC}{(A+B+C+D)}$$

Where G is the maximum common factor of (A, B, C, D).

2. Experiment and result

2.1 Comparison of binary classification and outlier detection in balanced data

This paper downloads the classified data set on whether users default on loans on Kaggle. The ratio of positive samples to negative samples is 5:1 and the dimension is 23, so it can be regarded as a balanced data set. In the experiment, data cleaning and feature engineering are carried out, such as missing value filling, variable conversion and desensitization, data standardization and so on.

Then, Grid Search and various classifiers and outlier detection algorithms are applied to train the model, and the final training results are shown in the following table.

Table 1: Comparison of models under balance conditions

	Bayes	Decision Tree	Random forest	AdaBoost	KNN	Isolation Forest	One-class SVM
AUC	0.7277	0.6066	0.7591	0.7638	0.7152	0.7465	0.7039
F1-scores	0.6672	0.7205	0.7990	0.7934	0.7805	0.7969	0.7375
Normalized diagonal-scores	251.537	180.653	255.101	243.320	196.687	54.643	104.410

2.2 Comparison of binary classification and outlier detection in unbalanced data

Similarly, this paper finds another unbalanced data set of the same task (loan default) on Kaggle, in which the ratio of positive class to negative class is 100:1 and the dimension is 23, so it is a unbalanced data set.

In the experiment, data cleaning and feature engineering are also carried out on this data set, and Grid Search and various model algorithms are applied on this data set. The final model training results are as follows:

Table 2: Comparison of models under unbalanced conditions

	Bayes	Decision Tree	Random forest	AdaBoost	KNN	Isolation Forest	One-class SVM
AUC	0.6855	0.5078	0.67201	0.7191	0.5658	0.7232	0.7530
F1-scores	0.9277	0.9787	0.98420	0.9837	0.9842	0.7926	0.8388
Normalized diagonal-scores	5.705	0.546	0	-0.042	0	2.617	12.497

3. Comparison between binary classifier and outlier detection algorithm

According to tables above and figures that there is little difference between the model effects of binary classification algorithm and outlier detection algorithm in the case of label balance.Random Forest and Bayesian are more suitable for balanced data than the others.

While in the case of label imbalance, the model effect of outlier detection algorithm is significantly better than the binary classification algorithm. The One-Class SVM algorithm is better in both the AUC socres and diagonal scores than the others, which can verify that the model effect is excellent.

4. The interpretation of unbalanced data set is suitable for outlier detection

For the binary algorithms, there are different ideas in essence.In the traditional binary classification algorithm, the aim is to minimize the number of errors in the classification process.They assume that the error costs of false positive and false negative are equal,so they are not suitable for class unbalanced data.

The feasibility explanation of outlier detection algorithm can be roughly divided into two categories: linear model and model based on similarity measurement. In outlier detection, for unbalanced data, model only focus on learning the law of positive samples, so as to find out the data that does not conform to the law of positive samples, that is, negative samples.

Linear model: Taking the classic One-Class SVM as the analysis object, the purpose is to find the maximum interval of feature points in the positive and negative samples. We can calculate the weighted Euclidean distance from each sample to the hyperspace composed of K eigenvectors (the smaller the eigenvalue is, the greater the weight is). We can also directly analyze the covariance matrix, and take the Mahalanobis distance of the sample (the distance from the sample to the distribution center when considering the relationship between features) as the anomaly degree of the sample.

Similarity based models: the distribution of abnormal points and normal points is different, so the similarity is low.For example, the simplest KNN can be used for anomaly detection. An example of a sample and its k-th nearest neighbor can be regarded as an outlier. Obviously, the k-nearest neighbor distance of the outlier is larger.Obviously, there are few data points and low density in the space where abnormal points are located.Similarly, Isolation Forest calculates the number of hyperplanes required to "isolate" a sample by dividing hyperplanes. In a low density space, isolating a sample requires less division times.

Therefore, outlier algorithm is more to find the maximum difference between positive and negative samples.

References

[1] Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." Data Mining, 2008. ICDM&apos; 08. Eighth IEEE International Conference on. IEEE, 2008.

[2] Ma J, Perkins S. Time-series Novelty Detection Using One-class Support Vector Machines. Procedding of International Joint Conference on Neural Networks, 2003.