Guo Qiang

SMOTE ALGORITHM IN IMBALANCED DATA

This article mainly introduces the basic principle of SMOTE algorithm for processing unbalanced data and the specific implementation method of R language, combined with the characteristics of unbalanced data, uses and improves SMOTE algorithm.

INTRODUCTION

The so-called unbalanced classification problem refers to the pattern classification problem in which the number of training samples is distributed unbalanced among classes. Specifically, the number of samples of some classes is far less than that of other classes. When traditional machine learning methods are used to solve these unbalanced classification problems, the performance of classifiers often drops significantly, and the resulting classifiers are highly biased.

At present, the main algorithms for dealing with biased data can be divided into two categories [1]: undersampling and oversampling. Undersampling is to achieve the purpose of balancing data by removing the number of samples in most classifications of training data, while oversampling is mainly to form a new small number of classification samples to achieve the purpose of balancing data. Synthetic minority oversampling technique (SMOTE) is a solution based on the idea of oversampling, which balances the class distribution by adding artificially synthesized minority class samples to the data, reduces the possibility of overfitting, and improves the performance of the classifier on the test set. SMOTE provides a new direction for solving the imbalance problem, becomes an effective preprocessing technique for dealing with imbalanced data, and has been successfully applied in many different fields. SMOTE has promoted the generation of methods to solve imbalanced classification problems, and made significant contributions to new supervised learning paradigms [2]. According to the distribution of minority samples, the ADASYN algorithm adaptively changes the weights of different minority samples, and automatically judges the number of new samples that need to be synthesized for each minority sample, which is an important improvement for SMOTE to calculate the sample weight.

I. SMOTE ALGORITHM

The SMOTE algorithm was first proposed by Chawla et al. [3] in the Journal of Artificial Intelligence. This is an oversampling method. Its main idea is to achieve the purpose of balancing samples by inserting new samples in some similar minority samples. The characteristic of the SMOTE algorithm is that it does not simply copy samples in a random oversampling manner, but adds new samples that do not exist. Synthesize new samples between class samples, thus effectively alleviating the overfitting problem caused by random oversampling. Assuming that there are minority class samples, for each sample x, search its k (usually 5) minority class nearest neighbor samples; if the upsampling rate is N, randomly select N samples among its k nearest neighbor samples, denoted as $y_1, y_2, ..., y_n$, perform random linear interpolation between minority class samples $x and y_j (j = 1, 2, ..., N)$, and construct new minority class samples p_j.

$$p_j = x + rand(0,1) * (y_j - x), j = 1, 2, ..., N \quad (1)$$

In formula (1), rand(0,1) represents a random number in the interval (0,1). Merging these newly synthesized minority class sample points into the original data set can generate a new training set. Assuming that the attribute value of a minority class sample is (6, 4), and the attribute value of the nearest neighbor point jumping out from the minority class is (4, 3), a random number between 0 and 1 is randomly generated rand(0, 1) = 0.2. Then the calculation process of the new synthetic sample is:

$$\begin{aligned} p_j &= x + rand(0,1) * (y_j - x) \\ &= (6,4) + 0.2 * ((4,3) - (6,4)) \\ &= (6,4) + 0.2 * (-2,-1) \\ &= (5.6, 3.8) \end{aligned}$$

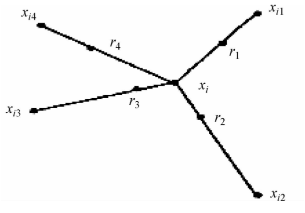

Fig. 1 Schematicdiagramof SMOTEsynthesis

Figure 1 depicts the process of artificially synthesizing new data. xi is a certain minority class instance, xi1, xi2, xi3, xi4 are four neighbors of xi respectively, r1, r2, r3, r4 are four new artificial data generated, this method generates new minority class data increases generality, rather than causing overfitting like exact replicating samples.

The SMOTE algorithm abandons the practice of random oversampling and copying samples, which can prevent the problem of easy overfitting in random oversampling and improve the performance of the classifier. However, the SMOTE algorithm also has the following two disadvantages:

- Since a new sample is generated for each minority class sample, the problem of overlapping generated samples is prone to occur.
- In the SMOTE algorithm, there is a problem of over-generalization, which is mainly attributed to the method of generating synthetic samples.

ADASYN: Adaptive Synthetic [4]. According to the learning difficulty of different minority samples, the weight distribution is used to synthesize more training data for a small number of difficult-to-learn category samples. The ADASYN algorithm adaptively changes the weights of different minority samples according to the distribution of minority samples, automatically determines the number of new samples that need to be synthesized for each minority sample, and synthesizes more new samples for samples that are difficult to learn, thereby compensating for bias. state distribution. ADASYN can not only reduce the learning bias caused by the original unbalanced data distribution, but also adaptively shift the decision boundary to focus on those samples that are more difficult to learn. ADASYN supports two classifications and also supports multi-category classification.

II. SMOTE algorithm in R language

The specific algorithm of SMOTE can be realized through the DMwR package in R software. The following is the main program. Install and load the DMwR package. Application install. packages () installs the data package, and library () loads the DMwR package. The specific steps are:

$$rm(list = ls())$$

$$install.packages(''DMwR\, denpendencies = T)$$

$$library(DMwR)$$

Read in data. Suppose the original data is data. csv, including the grouping variable class, which takes a value of 1 or 0 (negative data is 0, positive data is 1). With the R software read. csv() reads in the data, and reads in the variable names in the raw data. Set the name of the data set after reading to data. The specific steps are:

$$data = read.csv(''data.csv\, header = T)$$

SMOTE algorithm. In the SMOTE algorithm, the samples of the majority class are under-sampled, and the samples of the minority class are over-sampled, so a specific number of samples will be set. Related parameters and explanations: formula, set the independent variables and grouping variables in the data set; data specifies the data set to be processed, perc.over, and perc.under options define the times of oversampling and undersampling respectively. The specific steps are:

$$newdata = SMOTE(formula, data,$$

$$perc.over =, perc.under =)$$

Organize data. Use the table() statement to view the proportion of positive data and negative data in the new data set. The specific steps are:

$$table(''newdata\$class'')$$

III. Conclusion

For classification problems, there may be serious bias in the categorical dependent variable, that is, the proportion between categories is seriously out of balance, which leads to biased conclusions. In order to solve the problem of unbalanced data, the SMOTE algorithm, which synthesizes a small number of over Sampling technology, which is a common method for dealing with unbalanced data at present. This article introduces the basic principles of the SMOTE algorithm to deal with unbalanced data problems and how to install and run the SMOTE algorithm on the R language. That is, to analyze and simulate a minority of class samples, and add artificially simulated new samples to the data set, so that the categories in the original data are no longer seriously unbalanced.

IV. References

1. Alberto F, Maria J, Francisco H, On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data sets[J]. Expert systems with Applications, 2009,36:9805-9812
2. CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]
3. CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of artificial intelligence research, 2002, 16(1): 321-357.
4. HE Haibo, BAI Yang, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]//Proceedings of 2008 IEEE International Joint Conference on Neural Networks. Hong Kong, China, 2008: 1322-1328.

Guo Qiang, student of department of Information Technologies in Automated Systems Department, BSUIR, 254951872@qq.com.

Supervisor, German Oleg, Associate Professor, Candidate of Technical Sciences, the Belarusian State University of Informatics and Radioelectronics, ovgerman@tut.by.