FSDM 2020
November 13-16, 2020
Online Conference

FSDM3269

Classification of Unbalanced Data Based on RSM and Binomial Distribution

LI Rong ZHOU Wei-bai
Guangzhou college of Commerce

Introduction

In the case of extremely unbalanced data, the results of the traditional classification algorithm are very unbalanced, and most samples are often divided into the categories of majority samples, so the accuracy of judgment of the minority classes will be reduced. In this paper, we propose a classification algorithm for unbalanced data based on RSM and binomial undersampling. We use RSM's random part features rather than all each classifier to make each training classifier reduce the dimensions, and dimension reduction makes relatively minority class samples indirectly lift. Using the above characteristics of the RSM to reduce dimension can solve the problem that unbalanced data classification in the minority class samples is too little, and it can also find the important attribute of variables to make the model have the ability of explanation. Experiments show that our algorithm has high classification accuracy and model interpretation ability when classifying unbalanced data.

Methods

We combine sampling method of binomial distribution. First, our algorithm determines the number of samples needed for the majority sample after undersampling by using binomial distribution, which makes the proportions of the minority class and the majority class similar in new data set. Then use the new data set to train RSM classifier, and get a final classification result by majority vote.

➢1: Binomial distribution sampling

We first find the cumulative distribution function of the binomial distribution

$$F(y)=\sum_{k=0}^{[y]}\binom{n}{k}p^{k}(1-p)^{n-k}$$

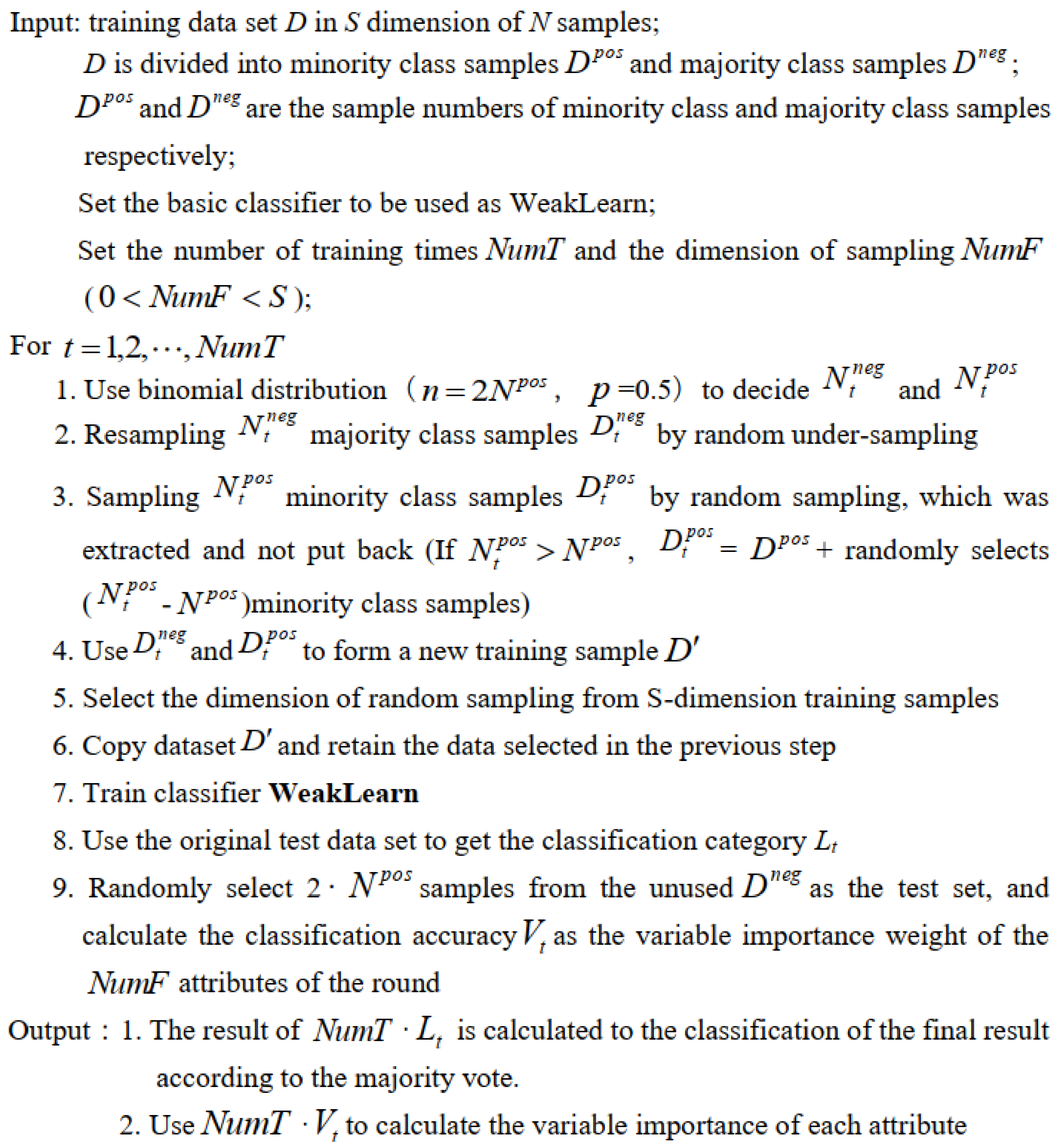
After a two-item sampling method, a new sample with a small difference in sample size between the minority and majority classes will be obtained, and this new sample will be classified using the RSM classifier and a final classification will be obtained.

➢2: RSM algorithm

We take advantage of the fact that RSM uses a few variables at a time to build a classification model to calculate the variable importance of each attribute of the dataset

$$w_i=\frac{\sum_{t=1}^{NumT}V_t}{n_i}$$

specific algorithm



```
Input: training data set $D$ in $S$ dimension of $N$ samples;
    $D$ is divided into minority class samples $D^{pos}$ and majority class samples $D^{neg}$;
    $D^{pos}$ and $D^{neg}$ are the sample numbers of minority class and majority class samples respectively;
    Set the basic classifier to be used as WeakLearn;
    Set the number of training times $NumT$ and the dimension of sampling $NumF$ ($0 < NumF < S$);
For $t=1,2,\cdots,NumT$
    1. Use binomial distribution（$n=2N^{pos}$, $p=0.5$）to decide $N_t^{neg}$ and $N_t^{pos}$
    2. Resampling $N_t^{neg}$ majority class samples $D_t^{neg}$ by random under-sampling
    3. Sampling $N_t^{pos}$ minority class samples $D_t^{pos}$ by random sampling, which was extracted and not put back (If $N_t^{pos} > N^{pos}$, $D_t^{pos} = D^{pos}$ + randomly selects ($N_t^{pos}$ - $N^{pos}$)minority class samples)
    4. Use $D_t^{neg}$ and $D_t^{pos}$ to form a new training sample $D'$
    5. Select the dimension of random sampling from S-dimension training samples
    6. Copy dataset $D'$ and retain the data selected in the previous step
    7. Train classifier **WeakLearn**
    8. Use the original test data set to get the classification category $L_t$
    9. Randomly select $2\cdot N^{pos}$ samples from the unused $D^{neg}$ as the test set, and calculate the classification accuracy $V_t$ as the variable importance weight of the $NumF$ attributes of the round
Output：1. The result of $NumT\cdot L_t$ is calculated to the classification of the final result according to the majority vote.
        2. Use $NumT\cdot V_t$ to calculate the variable importance of each attribute
```

Conclusions

There are unbalanced data problems in many different fields. This target receives much attentions in recent years related to relative issues. The past research mainly combined with a multiple classifier and many people focus on Bagging and Boosting classifier. Few people notice RSM classifier which has not only the advantages of multiple classifier, but also the ability of reducing classification dimension, indirectly promote the minority class sample effect. We use RSM combined with binomial distribution sampling to propose a binomial random subspace classification algorithm for unbalanced data. This method integrates the concepts of RSM and resampling. Experiments show that our algorithm has a higher, more balanced and more stable classification accuracy when classifying unbalanced data, and it also has the ability of model interpretation.