**RESEARCH ON CLASSIFICATION**

**OF IMBALANCED DATA SETS based on Maximize F1-Measure Learning**

# Abstract

With the rise of mobile Internet technology, more and more raw data has been collected for analysis and mining. The data in many fields are seriously unbalanced. The number of samples belonging to different categories is extremely different. Traditional machine learning methods usually use global classification accuracy as the training target and perform poorly on the unbalanced data set. Therefore, the unbalanced data classification algorithm has gradually become a subject of concern.

At present, there are two main types of imbalanced data set classification, namely data resampling and integrated learning. Their main idea is to make the original data "balanced" by changing the weight or proportion of samples in the original data set. State, which applies to the traditional classification algorithm, the results of such algorithms and data distribution is closely related to the complex process of regulation is often required in order to get a better result and generalization ability is not strong. Therefore, this paper proposes an algorithm to solve the problem of unbalanced data set classification directly with the F1 value as the training target, and has achieved good results

Introduction

With the advent of the era of large data and a variety of networked systems become larger and larger, more complex structure, a variety of surveillance, security, financial and other systems every day access to massive raw data, so the original The analysis of data and the excavation of potential information play a crucial role in the classification and decision-making process. Although the existing methods have achieved great success in the field of data mining, there are still many problems in many practical problems. Unbalanced data is one of the relatively new and arduous challenges.

The problem of unbalanced data classification is classified learning problem when the number of samples in different categories is very different. For example, if there are more samples in one class (positive class, majority class) than in other classes (negative classes, minority classes), then this classification problem is called unbalanced data set classification problem.

The problem of unbalanced sample classification has been concerned for more than ten years, and there are many related researches during this period. Generally, all the common solutions can be divided into two categories:

**(1) classical imbalance classification method** The main idea of this class of algorithms is to unbalanced data set through a series of processing or classification process to take differentiated treatment in the data set of examples, so that the original data into a relative "equilibrium ", And then solve the problem of imbalance;

**(2) The traditional classifier optimization method** For unbalanced data set classification, this kind of algorithm is far less than the proportion of classical unbalanced processing methods, and they do not have a common processing law. Their general idea is to further refine the training results of the conventional classification algorithm for unbalanced data so that the resulting model can handle the unbalanced data set.

A) Classical unbalance treatment method：According to the existing research results, the classical methods to solve the unbalanced data set can be summarized as follows:

**(1) the original training data reconstruction** by resampling technology to change the majority of the number of samples or a small number of categories, so that the original data has become relatively balanced;

**(2) Integrated learning method** The integrated learning method is used to train multiple weak classifiers. Finally, the classification results are obtained by voting or weighting. In the training process, the weights are changed or the original data sets are partitioned to solve the imbalance problem.

**(3) sensitive cost learning methods**  The sensitive cost learning method changes the weights of the original data in the evaluation criteria, usually using the artificial sensitive cost matrix to help calculate the classification loss to solve the imbalance problem.

**B) Traditional classifier optimization method**

A common feature of the traditional classifier optimization algorithms is that they are no longer balanced as unbalanced data sets as classical unbalanced processing methods, and their main idea is that by modifying The training process of the classifier or the classification process to adapt to the unbalanced data set not only by optimizing the training process of the algorithm to reduce the imbalance distribution on the training process, or the use of normal training train the model, through a series of other processes The adjustment of the model, or get the ordinary model in the classification phase and the classical classification phase of the different methods to solve the problem of imbalance.

For the existing methods of unbalanced dataset classification, most of them transform the original data set into a balanced state by using some strategy, and then use the classical machine learning method to solve the problem. The main idea of classical machine learning algorithm is independent identically distributed If the distribution of the original data set is changed, the classical machine learning method can not properly fit the original data set, resulting in unstable results or poor results. The algorithm in this paper solves the above problems by minimizing loss learning and directly evaluating the standard F1 value as the training target by unbalanced data set classification.

Minimize loss learning:

For most of the existing machine learning algorithms, we usually think that the training data set and the test data set satisfy the independent co-distribution, this distribution is also considered the true sample spatial distribution, that is, we summed up the training data and test data summarized Bias to classify unknown samples in real space. The training goal of the machine learning algorithm is to find a special hypothesis in all hypothesis spacesto make the test data set of the size reach the set minimum value of loss, and the formula 2-22 reaches the minimum value. For the traditional loss function, since it is usually considered that the points in the space are independent of each other, the loss function can be transformed into the form of 2-23, that is, the overall loss is transformed into the sum of the individual losses of all samples.

So according to Eqs. 2-22 and 2-23, and based on the hypothesis of independent co-distribution between training data and real data, or training hypothesis that sampling data is real space, we can finally convert 2-22 to 2-24, Into the form of 3-10, which is why through the training data set can be real space training or fitting.









For most of the traditional machine learning methods, their training idea is usually to construct the error function between the output of a single sample and the target output, and add the errors of all the samples as the total loss of the training set. The ideal loss function is the 0-1 loss function. If the output and the target output belong to the same category, the loss is 0, and the difference is 1, as shown in Equation 3-11.



And the average of 0-1 loss does not necessarily apply to all the problems, so the idea of minimizing loss of learning is proposed, by using a custom loss function instead of 0-1 loss training, in order to adapt to different problems, This idea was first applied to structured support vector machines.

The concrete idea is to convert a single sample to a single output, such as Eq. (3-12), into the form of Eq. (3-13). In the real training problem, we use the training set's feature set and target output set to solve Therefore, the training space in Eq. (3-13) is the target space. Assume that h becomes the global assumption, corresponding to all the sample inputs and all the classifier outputs. And the overall loss is transformed from the form of Eq. 3-10 to the loss of each sample to the form of Equations 3-14. Assumptions 3-13 are equivalent to solving the results of Eqs. 3-15. In this paper, the traditional machine learning classification algorithm is transformed into the idea of minimizing the global loss by using the method of independence to solve the loss, so as to establish the loss through the unbalanced dataset classification evaluation criteria, and to adapt to the unbalanced sample classification problem.









Unbalanced Data Set Classification Algorithm Based on Maximum F1 Value:

1、（imbalance learning and minimize lose learning）

Assume that there is a one-dimensional imbalanced data set, which contains most samples and a few samples. The probability density curve is shown in Figure 3-1, and it is assumed that the sample ratio of the majority class to the minority class is. It is obvious that the basic idea of the traditional classifier is to maximize the global accuracy rate as the final training target. For the two types of boundary parts, even if their probability density sizes are similar, because most of the classes are different from the minority class sample base, The number of class samples will be far more than the number of samples of a small number of categories, the final classification line is likely in the figure near the location of the line b, in favor of the side of a few classes.

The idea of the classical unbalanced dataset classification algorithm is to directly reduce the sample ratio between most classes and a few classes by using some method. Usually, the number of sample points of the two classes will be the same or very close, and then the traditional classification Algorithm, if the probability density curve of the original dataset is the same as that shown in Figure 3-1, since there is no such problem that the two sample bases are different, the classification limit which leads to the highest global classification accuracy should be the line a . This line is based on the abscissa of the intersection of two kinds of probability density curves as the threshold of demarcation. The minority of the left side of the boundary line and the right side of the majority class are misclassified samples, which are easy to prove by the area method.

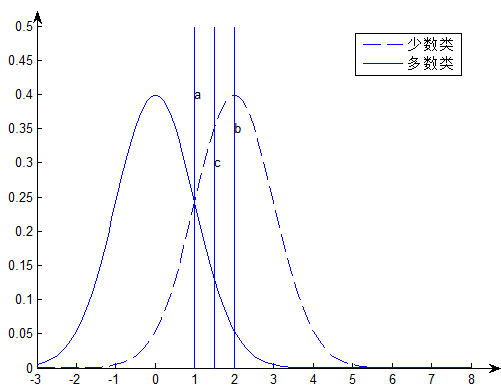


Figure 3-1 Probability density curve of the dataset

However, due to the change of the sample space, we can only think that the solution (line a) is the best classification line in the current changed sample space and can not determine whether the demarcation point is the best point on the original data set. In general, there are special evaluation criteria to evaluate the effect of the final classification, while the classical unbalanced classification method usually only increases the unbalanced classification index of the original space, but usually it can not prove that the solution must be The optimal solution in the original sample space.

Therefore, this algorithm is based on minimizing the loss of learning, using the commonly used imbalance data set evaluation index F1 value to construct a special loss function, used in neural network model to solve the problem of unbalanced data set classification.

**2、the construction of loss function**

In this algorithm, we choose F1 as the optimization target, so we can set the loss function as (1-F1). For the training of neural network, the minimum value of loss is the same as the maximum of F1, As long as the sign before the gradient can be changed, so here will minimize the loss of the concept extended to maximize the objective function, both 3-15 into 3-16 type of form.



For the neural network model in this section, we still use the sgn function of the traditional neural network shown in Eqn. 3-17 as the final classification criterion. For the output on the whole training set, we use it to represent, and the target output is still used . In order to express the final value of F1, we first need to find the recall rate (Recall) and accuracy (precision), according to observe the confusion matrix and the relationship between the confusion matrix can be seen to meet the parameters TP 3- 18, and the formula 2-3 can be converted into the form of the formula 3-19, and the formula 2-6 can become the form of the formula 3-20. So the final F1 value can be expressed as the form 3-21.











However, by observing Equation 3-21, we can find that since the sum of the sequences is 0 and 1, although the training goal can be expressed by the entire training set, the F1 value is still included because the process involves a step operation Is discrete and does not establish a direct numerical connection to our neural network output and to the connection layers between the various nodes. In the structured SVM, we use spatial traversal and double optimization to solve arbitrary objective function, and therefore need to spend a huge amount of time. In this paper, we give up this idea, but use neural Network output layer sigmod function specific to the nature of equation 3-5 will be established with the contact.

The algorithm uses neural network training process is to take the first use of the current state of the network classification, and then solve the loss and the loss has been optimized to the next better state the idea of training in the evaluation process to transform, no longer Using the current classification of the results of neural network, but using the current output of the shape of the probability of solving the expected value of 3-5, and to optimize the expectations, so that both the establishment of output and parameters and the direct link between the target, It is also possible to increase the probability that the target will acquire a higher value by optimizing the expected value, so that the meaning of training is not lost.

However, the establishment of the expectation of contact with us can not be used in the exact solution in Section 2.2.3, first of all because the exact solution takes time, which is contrary to the design of fast algorithm in this section the original intention, can be seen in Section 2.2.3 The space between the expectation and the neural network can not be directly established, so it is impossible to directly establish the relationship between the expectation and the sample. So, we can not directly establish the relation between the expectation and the neural network weight, To solve these two major problems we use the approximate relation in Equation 3-22 below.



Because of the expectation and covariance relationship, 3-23, the expectation of the numerator is the expectation of the numerator in Equation 3-22, respectively, and the expectation of the denominator is due to the subsequent back-propagation algorithm In the hope that some samples can be brought into, rather than for a part of the sample will be discarded classifier output, so continue to use the form of the square rather than directly expected form, but even if the square will eventually close to 0 or 1, so its overall value is close. For the covariance of the latter two variables in Eq. 3-23, we can find that X and Y are both in the numerator and denominator, so the opposite trend of X and Y is the root-sine covariance. We can conclude that the covariance will always be zero, so there is a relation 3-24. When the algorithm optimizes the approximation to the right side, the expectation can be expanded and converged to a global optimum because the expected value is the upper bound of the approximation and relatively close. Solution or local optimal solution, has reached training purposes.





To sum up, we turn the original neural network training target 3-12 into the form of equation 3-13, and design the objective function according to the commonly used F1 value in the evaluation standard of unbalanced sample classification. Then, the discrete function approximation As a continuous function so that the function can be perfectly associated with each training set sample output, and then with the neural network parameters, and proved that the training process of neural network can optimize the approximation function while optimizing the training set On the final F1 value, to meet the initial design.

**3, the training process**

For the neural network model, the most effective weight updating strategy is the back propagation algorithm. Since the final result of this algorithm will train the objective function to the maximum value, the updating process is as follows: . Which represents the learning rate, the size of the impact of the neural network convergence rate and convergence accuracy, and occasionally may also affect the final convergence in which a very good solution near. The set-up method is usually a set method is also an annealing method to set a relatively large initial value and with the training process continues to shrink.





Therefore, in order to find out the update amount of each weight, we need to solve the partial differential of the whole objective function F1 for each parameter in each node. Where wij is the ith weight of node j, netj is the result of the inner product of node j, which is the result of sigmod (equation 3-2), and xji is the input of node i corresponding to the i-th weight. In the process of training is always known, as long as the current state of the node weights and input can be obtained. What follows is that the first half of the result in Equation 3-26 results in the final output of the partial derivative for each node output.



In order to solve the partial differential of each node, we need to classify all the nodes into two classes, one is the output node and the other is the hidden layer node. Since the final objective function described in the previous section is directly related to the output, The output node can solve the partial differential directly through the objective function value in the current state. For the hidden layer node, it needs to solve the partial differential through the downstream node of the node, which is the main idea of the chain propagation algorithm.

For the output node, we use the chain rule, with the function of the request is decomposed into the form 3-27. Where oj is a component of h (x) in equation 3-22, so the last term in Eq. 3-27 can be solved by 3-28, and the former can be solved directly The solution procedure is shown in Equation 3-29. For the sake of simplicity, the constant coefficient in front of the previous objective function is omitted in this formula, which has no effect on the optimization process.







For hidden nodes, there is no way to directly use the objective function, so update the connection parameters according to the downstream node of each hidden node. The update scheme is shown in Figure 3-30. Because this algorithm is a binary classification problem, there is only one output node. And the method of solving the output node and the derivative of the function are introduced into Eqn. 3-30, the form of Eqs. 3-31 can be obtained where the partial differential of the output node is the first component of the node, And the number of hidden nodes.





The above is the original thought deduction process of the loss minimization learning algorithm for the unbalanced data set and the updating of the weights at the final training time.

Experimental results and analysis:

The experimental datasets in this chapter are all from the UCI machine learning dataset. For the data set selection process, we mainly select the data sets that have appeared in other unbalanced dataset classification algorithm research. The following 8 data sets The parameters are shown in Table 4-1.

表4-1 数据集参数表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Data set | Number of sample | Fractional proportion | Feature |
| 1 | YEAST | 1484 | 12.60% | 8 |
| 2 | Abalone | 4177 | 8.02% | 8 |
| 3 | Glass | 214 | 23.83% | 10 |
| 4 | Breast Canser | 699 | 34.50% | 9 |
| 5 | Vehicle Silhouettes | 946 | 23.43% | 18 |
| 6 | Haberman | 305 | 26.47% | 3 |
| 7 | Ecoli | 335 | 2.69% | 7 |
| 8 | Credit | 30000 | 22.12% | 24 |

In this section, the SMOTE algorithm, adaboost algorithm, structured support vector machine algorithm, classical neural network algorithm, sensitive cost learning algorithm and the algorithm of this paper are compared. The results are shown in the following table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | SMOTE | Adaboost | SSVM | ANN | SCL | ML-ANN |
| 1 | 0.747 | 0.717 | 0.645 | 0.667 | 0.717 | 0.846 |
| 2 | 0.406 | 0.399 | 0.486 | 0.352 | 0.331 | 0.581 |
| 3 | 0.980 | 0.922 | 0.980 | 0.837 | 0.866 | 1.000 |
| 4 | 0.984 | 0.943 | 0.961 | 0.984 | 0.956 | 0.994 |
| 5 | 0.995 | 0.997 | 0.962 | 0.945 | 0.979 | 0.995 |
| 6 | 0.459 | 0.423 | 0.632 | 0.413 | 0.660 | 0.647 |
| 7 | 0.762 | 0.696 | 0.727 | 0.941 | 0.941 | 0.941 |
| 8 | 0.534 | 0.435 | -- | 0.503 | 0.516 | 0.542 |

As can be seen from the table above, the algorithm in this paper has achieved some success in unbalanced dataset classification algorithm, and its result is usually better than the previous algorithms

Conclusion：

In this paper, we start from the unbalanced dataset classification and evaluation criteria, and instead of using the global accuracy rate of the traditional classification method as the training target, instead of training data classification F1 value as the training target, directly from the unbalanced data set classification effect is poor Fundamental to start, cut in. In this paper, we start from the loss function of the classifier, instead of using the traditional loss function. Instead, we construct the loss function associated with the F1 value directly, and take the approximation expectation of the current classifier output to solve the F1 value. The expectation is the lower bound of the expected value of F1, to confirm the feasibility of its optimization, and because it no longer has the original F1 function of the discrete, but directly with the classifier output to establish a link, it can reverse propagation algorithm Iterative method to solve the problem of unbalanced data set training.

For the algorithm in this paper, the following problems can be studied or optimized:

(1) In the strict sense, our algorithm does not establish the direct mathematic relation between the expectation of F1 value and the output of neural network model or model parameters. If we can overcome this difficulty, it may make the classification accuracy Further enhancements.

(2) In this paper, the algorithm can not be parallel computing, and can not use the traditional artificial neural network learning methods such as the batch to speed up the process of solving the algorithm run time is much higher than the classic neural network. So if we can solve this problem, we should be able to qualitative leap in the performance of algorithm time.