RESEARCH ON CLASSIFICATION

OF IMBALANCED DATA SETS BASED ON MAXIMIZE F1-MEASURE LEARNING (use style: *paper title*)

Subtitle as needed *(paper subtitle)*

Changchen Zhang1, Jingwei You2, Ziqiao Lv3etc

*(Affiliation 1*): dept. name of organization,name of organization, acronyms acceptable, City, Country

*(Affiliation 2*) : dept. name of organization,name of organization, acronyms acceptable, City, Country

*(Affiliation 3*) : dept. name of organization,name of organization, acronyms acceptable，City, Country

e-mail address if desired

*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.

Keywords-component; formatting; style; styling; insert (key words)

# Introduction *(Heading 1)*

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.

**(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 cannot 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

Identify applicable sponsor/s here. *(sponsors)*

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 H. The training goal of the machine learning algorithm is to find a special hypothesis in all hypothesis spaces to 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.

(2-22)

(2-23)

(2-24)

(3-10)

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.

(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

## 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.

Insert the figure

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 cannot 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 cannot 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.

## 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 sigmoid 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 cannot 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 cannot be directly established, so it is impossible to directly establish the relationship between the expectation and the sample. So, we cannot 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.

## the training process

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.

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 is the weight of node j, is the result of the inner product of node j, which is the result of sigmoid (equation 3-2), and is the input of node i corresponding to the 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 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 datasets. For the data set selection process, we select the data sets that are used in other imbalance classification algorithms. The experimental parameters of the 8 datasets are as follows:

Table 4-1 parameters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Dataset | 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, We compare the proposed algorithm with the SMOTE algorithm, Adaboost algorithm, structured support vector machine algorithm, classical neural network algorithm, sensitive cost learning algorithm. 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 |

The parameters of SVM are set after trying different kernel functions and whose result is the best, the parameters of ANN is default in WEKA, we use same dataset as the training set and test set, so the results indicate ability of the algorithm to fit the test space. In the results, the traditional algorithms have good performance on some datasets like No.3,4,5 , because the results have been good and there are fluctuations in the results, but the proposed method targets at the approximation of F1, so the algorithm proposed in this paper has little advantages over the traditional ones. When it comes to imbalanced datasets on which the traditional methods have poor performance like No.1,2,6 and so on, the results of the proposed method are considerable. The dataset 8 is special, we find that the features of this dataset are discrete, and there are many useless features, and the characteristic values differ greatly, but the algorithm in this paper takes discrete feature as continuous one, and there is no feature selection nor binaryzation, so the performance is not that good.

In the table , the experiment results of the improved SMOTE algorithm, AdaBoost.M1 algorithm, a structured support vector machine algorithm (SSVM), the cost sensitive learning method (Sensitive cost learning, SCL), and two kinds of algorithms in this paper are recorded. The 3 classic unbalanced classification methods in the table depend on traditional machine learning algorithms, using Naive Bayesian logistic regression classifier as base classifiers, the results in the table is the best ones, the main reason for the selection of the two kinds of classifiers is that SMOTE algorithm and the AdaBoost.M1 algorithm will increase the time complexity of the original algorithm, in order to make the process of modeling the time consuming acceptable, we choose two kinds of relatively simple structure, and they are commonly used in traditional machine learning classification with a good performance.

From the table we can see that the two algorithms in this paper can achieve a similar performance with the classical method, the overall classification results on 8 data sets is slightly better than the classic balance methods over F1 value.

The classical imbalance classification methods address the problem by changing the distribution of original samples, so it has a high upper limit of this kind of unbalanced classification method because there may happen to make the ideal sampling results or weight distribution, so the proposed algorithm is not better than classical unbalanced classification algorithm overall.

It has been proved that the algorithm has the ability to fit the training space, and it also proves that the algorithm in this paper can be used as a machine learning algorithm to solve the imbalanced problems. However, just fitting the training set may not be perfect for engineering problems, machine learning methods need to have good generalization ability and cannot be overfitting, so in this section we will compare the proposed algorithm with the front test on the generalization ability of algorithm.

We usually use the method of cross validation to measure the generalization ability of the algorithm, the dataset is divided into K parts, and the K tests are carried out, each of the K parts is taken as a test set, and other collections are used as the training set to verify the algorithm. Because we want to verify the generalization ability of the classification algorithm on imbalanced data sets, and the minority class samples may be very little, for examples there are only 9 minority samples in some training samples, so K should not be taken particularly large, in the experiment we choose K to be 3. In this section we selected four kinds of algorithms whose generalization ability is good: the traditional neural network, AdaBoost algorithm, structured support vector machine and the proposed method in this paper, the cross validation results are in the following table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Adaboost | ANN | SSVM | ML-ANN |
| 1 | 0.749 | 0.724 | 0.699 | 0.782 |
| 2 | 0.276 | 0.377 | -- | 0.545 |
| 3 | 0.879 | 0.865 | 0.865 | 0.879 |
| 4 | 0.932 | 0.945 | 0.935 | 0.946 |
| 5 | 0.805 | 0.927 | 0.916 | 0.927 |
| 6 | 0.248 | 0.328 | 0.476 | 0.575 |
| 7 | 0.667 | 0.875 | 0.727 | 0.941 |
| 8 | 0.505 | 0.434 | -- | 0.712 |

From the table we can see that both the proposed algorithm and the existing algorithm, the cross validation results are similar to the prior results on the 8 datasets in this paper, so the proposed algorithm has reliable generalization ability, the algorithm’s ability of training space fitting is true and reliable.

Form the results we can conclude from the table that the proposed algorithm has achieved some success in imbalanced classification, 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 use F1 value as evaluation criteria instead of the global accuracy rate used in the traditional classification method, we try to address the fundamental problem in which the classifier has poor performance. In this paper, we start from the loss function of the classifier, instead of using the traditional loss function, we construct the loss function associated with the F1 value directly, and take the approximation expectation of the current classifier output to compute the F1 value. We proved the assumption that the approximated expectation is the lower bound of the expectation value of F1 to confirm the feasibility of optimization for F1, and the F1 value is associated with the output of classifier and is no longer discrete. So the training can be complete by back propagation algorithm.

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

(1) In the strict sense, the algorithm does not establish a direct mathematic relation between the expectation of F1 value and the output of neural network model or with the model parameters. If we can overcome this difficulty, the classifier will have a better performance.

(2) In this paper, the algorithm cannot be parallel and cannot use the traditional artificial neural network learning methods such as the batch to speed up, so the proposed is much slower than classic neural network. So if we can solve this problem, there will be a qualitative leap of performance of the algorithm.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try “R. B. G. thanks”. Put sponsor acknowledgments in the unnum-bered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.