A New Approach For Imbalanced Data Classification Based On Maximize F1-measure Learning

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*— The class imbalance problem occurs when instances in one class are more than that in another. It has been reported to severely hinder classification performance of many traditional classification algorithms and many researchers have paid a great deal of attention to this field. Different kinds of methods have been proposed to solve the problem these years, such as resampling methods, integrated learning method. Their main idea is to rebalance the original dataset by changing the weight or proportion of instances in different classes, so that it can be applied to traditional classification algorithms. However, these conventional class imbalance handling methods might suffer from the loss of potentially useful information, unexpected mistakes or increasing the likelihood of overfitting because they may alter the original data distribution.

In this study, we propose a new method for imbalanced data sets which is different from previously proposed solutions to the class imbalance problem. We first put forward the idea that treat the performance measures as training target, then designed the loss function and build a model based on artificial neural network to solve the problem. The experimental results on 8 imbalanced data sets show that our proposed method is usually superior to the conventional imbalance data handling methods. Moreover, its performance is also good when the class imbalance ratio is low, i.e. classes are more imbalanced.

Keywords- Imbalanced data classification, F1-measure, Minimize lossing learning

# Introduction

In recent years, imbalanced classification problems have attracted a lot of researcher’s interests due to its challenges in various real-world applications. Different from standard classification problems, an imbalanced task involves a data set that has an imbalanced class distribution, i.e., the number of instances in one class(majority class) is outnumbered by the number of instances in another class(minority class)[1]. There exist imbalanced class distributions in many real world domains, such as detecting oil spills from satellite images [2], identifying fraudulent credit card transactions[3], medical diagnosis[4], anomaly detection[5], finance risk management[6], software defect prediction[7], and bioinformatics[8].

In abovementioned imbalanced classification problems, we always care more about the minority class because they contain more vital information. However, a problem usually occurs because traditional classification algorithms tend to be biased towards to the majority class [9,10].

This template, created in MS Word 2000 and saved as “Word 97-2000 & 6.0/95 – RTF” for the PC, provides authors with most of the formatting specifications needed for preparing electronic versions of their papers. All standard paper components have been specified for three reasons: (1) ease of use when formatting individual papers, (2) automatic compliance to electronic requirements that facilitate the concurrent or later production of electronic products, and (3) conformity of style throughout a conference proceedings. Margins, column widths, line spacing, and type styles are built-in; examples of the type styles are provided throughout this document and are identified in italic type, within parentheses, following the example. Some components, such as multi-leveled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

# Related Work

## 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 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.

## 分类评价标准

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

In imbalanced classification problems, a specific metric is needed to evaluate the performance of the classifier.

Confusion matrix for binary classification

|  |  |  |
| --- | --- | --- |
|  | Positive prediction | Negative prediction |
| Positive class | True positive (TP) | False negative (FN) |
| Negative class | False positive (FP) | True negative (TN) |

The performance of classifier is usually measured by accuracy () in conventional methods, but in the imbalanced classification problem, the misclassification of the minority class costs more than the majority one, so the researchers used a variety of criterion to measure the classification results.

The receiver operating characteristic (ROC) graphic is commonly used as an evaluation criterion. The ROC graphic depicts the trade-off between TPrate(,y-axis) and FPrate(,y-axis), the area under the ROC curve (AUC) is a useful metric for classifying performance because it gives the probability that a randomly selected pair of samples (one positive and one negative) would have their predicted probabilities correctly ordered.[ Zou Q, Xie S, Lin Z, et al. Finding the Best Classification Threshold in Imbalanced Classification ☆[J]. Big Data Research, 2016, 5:2-8.]

The F1-value is a trade-off between precision (P) and recall (R) which is described as follows:

,

The number of true positive samples has the greatest influence on the F1-value, so it is used in imbalanced classification problems in which the positive samples are more important, so in this paper, we use F1-value to evaluate the performance of the classifier.

# Proposed Method

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

## Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the Units

* Use either SI (MKS) or CGS as primary units. (SI units ade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often e in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop.

    

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)

An excellent style manual for science writers is [7].

# Experiment and Analysis

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save ar.

## Authors and Affiliations

The template is designed so that author affiliations are not s.

### For author/s of only one affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

### For author/s of more than two affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the “Columns” icon from the MS Word Standard toolbar and then select “1 Column” from the selection palette.

#### Highlight author and affiliation lines of affiliation 1 and copy this selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign number of columns: Place your cursor to the mns.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

## Figures and Tables

### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

a. Sample of a Table footnote. *(Table footnote)*

1. Example of a figure caption. *(figure caption)*

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

# Conclusion

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

##### 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. Lizhi Peng, Hongli Zhang, Bo Yang, Yuehui Chen, “A new approach for imbalanced data classification based on data gravitation,” Information Sciences 288 (2014) 347–373
2. M. Kubat, R.C. Holte, S. Matwin, “Machine learning for the detection of oil spills in satellite radar images”, Mach. Learn. 30 (1998) 195–215.
3. T. Fawcett, F. Provost, “Adaptive fraud detection”, Data Min. Knowl. Discov. 1(1997) 291–316
4. M.A. Mazurowski, P.A. Habas, J.M. Zurada, et al, ‘Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance”, Neural Networks: Off. J. Int. Neural Network Soc. 21 (2008) 427.
5. W. Khreich, E. Granger, A. Miri, R. Sabourin, “Adaptive roc-based ensembles of hmms applied to anomaly detection”, Pattern Recognit. 45 (2012) 208–230.
6. I. Brown, C. Mues, “An experimental comparison of classification algorithms for imbalanced credit scoring data sets”, Expert Syst. Appl. 39 (2012) 3446–3453.
7. L. Pelayo, S. Dick, “Applying novel resampling strategies to software defect prediction”, in: Annual Meeting of the North American on Fuzzy Information Processing Society, pp. 69–72.
8. H. Yu, J. Ni, J. Zhao, ACOSampling: an ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data,

Neurocomputing 101 (2013) 309–318.

1. S.García,F.Herrera,Evolutionary under sampling for classification with imbalanced datasets: proposals and taxonomy,Evol. Comput. 17 (2009) 275–306.
2. Y.Tang, Y.Q. Zhang, N.V. Chawla, S. Krasser, SVMs modeling for highly imbalanced classification, IEEE Trans. Syst. Man Cybern. PartB 39 (1) (2009) 281–288.
3. .K. Elissa, “Title of paper if known,” unpublished.
4. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
5. 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].
6. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.