Active Learning algorithm for Threshold of Decision Probability on Imbalanced Text Classification based on Protein-Protein Interaction Documents

Guixian Xu[1,2], Zhendong Niu[1], Xu Gao[3], Yujuan Cao[1], Yumin Zhao[1]

1. School of Computer Science, Beijing Institute of Technology, Beijing, China
2. College of Information Engineering, Minzu University, Beijing, China
3. North China Grid Company Limited, Beijing, China
{xuguixian2000, zniu, quishuichangtian, oblivion}@bit.edu.cn; gao.xu@ncgc.com.cn

Abstract—**The study of host pathogen protein-protein interactions (PPIs) is essential to understand the disease-causing mechanisms of human pathogens. A large number of scientific findings about PPIs are generated in the biomedical literatures. Building a document classification system can accelerate the process of mining and curation of PPI knowledge. With more and more imbalanced dataset appearing, how to handle the imbalanced classification problem is becoming a hot topic in machine learning field. In this paper, we propose an Active Learning algorithm for Threshold of Decision Probability (ALTDP) to solve problem of misclassifying the minority class based on imbalanced host pathogen PPIs data set. The results demonstrate the proposed approach is significant to improve the accuracy of classification on imbalanced data set.**

Keywords: imbalanced text classification; machine learning; protein-protein interaction

Ⅰ. INTRODUCTION

With more and more infectious diseases emerging/reemerging, there have been major initiatives for large-scale genomic and proteomic projects to study the basic biology and disease-causing mechanisms of human pathogens [1, 2]. Therefore, a large number of important scientific discoveries regarding these pathogens and their host responses are generated and often buried under the increasing volume of biomedical literature. Building a document classification system can accelerate the process of mining and curation of PPI knowledge which can facilitate the development of preventative and therapeutic strategies against human pathogens.

Recently, the imbalanced classification problem is becoming a hot topic in data mining field. Usually the strategies of solving imbalanced problem have three types: re-sampling, cost sensitive learning, and adjusting algorithms to bias the rare class [3]. In this paper, we present an an Active Learning algorithm for Threshold of Decision Probability (ALTDP) to solve problem of misclassifying the minority class based on imbalanced host-pathogen protein-protein interactions corpus consisting of 1,360 MEDLINE abstracts. The experimental results show the proposed approach is efficient to improve the accuracy of classification on imbalanced data set.

In the following, we first introduce the research background and related work. The experimental method is discribed next. We then present the results and discussion, and conclude our work.

Ⅱ. BACKGROUND AND RELATED WORK

A. Related work on imbalanced dataset classification

There are usually three types of strategy to solve the imbalanced problem: re-sampling, cost sensitive learning, and adjusting algorithms to bias the rare class [3]. Data sampling techniques include two approaches: one is to overcome the problem of class imbalance by either removing some data from the majority class, the other is to add some artificially created or duplicated data to the minority class. Some researchers investigated the effect of sampling methods for imbalanced data distributions [4,5]. An ensemble classifier was constructed in which each classifier was trained on a subset of the majority class and on the whole minority class in [6].

There are some comparison studies of classification methods on imbalanced data categorization. The comparisons included the comparison of fuzzy classifier and C4.5 decision tree [7], the comparison of Vector quantization SVM (VQSVM) and traditional SVM [8], the comparison of Lazy Bagging classification algorithm and C4.5 [9]. It was shown that each former algorithm which was proposed for imbalanced classification outperformed the latter in the experiments [7-9]. [10] investigated the effect of dataset size and class distribution on imbalanced classification performance with 11 classifiers of WEKA tool. The results show the data sampling is effective at alleviating the problem of rare events. Some studies focused on feature selection

978-0-7695-3958-4/10 $26.00 © 2010 IEEE
DOI 10.1109/DSDE.2010.28

IEEE computer society

Authorized licensed use limited to: University of Louisiana at Lafayette. Downloaded on April 20,2023 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

techniques in text classification with imbalanced data [11,12]. Active learning has also been adopted to tackle the imbalanced text classification problems as in [13, 14].

B. Machine learning algorithm

Logistic Regression (LR) is utilized for prediction of binary-event probabilities by fitting a generalized linear model to given data points in statistics. LR has gained popularity for high-dimensional classification such as text processing where linear boundaries are usually adequate to separate the classes [15].

Logistic Regression function [16] is defined as followes:

$$f(t) = \frac{1}{1+e^{-t}}$$

t is the input and f(t) is the output. f(t) confined to values between 0 and 1 represents the probability of a particular outcome. Here

$$t = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k$$

b_0 is called the intercept and $b_1, b_2, \ldots, b_k$ are called the regression coefficients of $x_1, x_2, \ldots, x_k$ respectively.

Usually, in classification field, the decision probability of threshold of LR is assigned 0.5 when the class distribution is balanced. If the class distribution is imbalanced and the decision probability still is 0.5, the classifier will misclassify some test examples of minority/positive class to the majority/negative class. The majority class test examples are hardly misclassified to the minority class. This means that 0.5 is high as the threshold of decision probability when LR is used to the imbalanced classification. It will lead to the serious positive examples error cost

In this paper, we present an active learning algorithm for threshold of positive decision probability based on LR to solve the imbalanced text classification on the host-pathogen protein-protein interactions documents.

III. METHOD

A. Data collection selection

Most pathogen protein-protein interaction (PPI) information annotated in knowledge bases is for viral proteins or PPI within bacteria. The annotated Host Pathogen protein-protein interactions corpus coming from [17] is used to the experiment. This corpus called PH-data was generated from two different sources. One is from UniProtKB database [18]. The other is from PubMed. Two domain experts reviewed and manually annotated this set, and categorized the abstracts as positive or negative. PH-Data consists of 1,225 negative abstracts and 135 positive ones.

B. Data preprocessing

The text is normalized by changing nouns in plural forms into singular forms, verbs in past tense into present tense based on the SPECIALIST lexicon, a component in the Unified Medical Language System (UMLS) [19]. We also replace punctuation marks with spaces and changed uppercase letters to lowercase ones. For example, the sentence "This entry process requires the Shigella Ipa proteins that are secreted by a type III secretion apparatus and that act in concert to fine tune cell responses." is normalized to "this entry process require the shigella ipa protein that be secrete by a type iii secretion apparatus and that act in concert to fine tune cell response".

C. Active learning for Threshold of Decision Probability

For the classification evaluation, F measue is usually chosen as the main evaluation criteria and it is calculated as F=2*P*R/(P+R), where P is the precision (i.e., TP/(TP+FP)) and R is the recall (i.e., TP/(TP+FN)). TP is the number of true positives; FP is the number of false positives; and FN is the number of false negatives. When the decision probability of LR on PH-Data is assigned 0.5, the classifier misclassifies some test examples of the positive class to the negative class and the negative class test examples are hardly misclassified to the positive class. Therefore precision is almost equal to 1 and recall is obviously lower than precision. This will not get an ideal F value.

We present the Active Learning algorithm for Threshold of Decision Probability (ALTDP) based on the changing trend of F performance. With ALTDP, we can get the optimal threshold in which the classifier will get an optimal F performance.

We assume Data is the two-class imbalanced training text collection, we randomly select 80% of Data as the training data set (TR) of ALTDP, the rest of Data is as the testing data set (TE) of ALTDP and regarded as the human expert of active learning for the judgment of performance of F.

ALTDP is described as follows.

ALTDP algorithm:

Input: TR,TE,LR,Init, ζ

//TR is the training data set

//TE is the testing data set

//LR is the classification algorithm

Authorized licensed use limited to: University of Louisiana at Lafayette. Downloaded on April 20,2023 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

//Init is the initial threshold
//ζ is the changing degree of threshold
Output:
Dth //Dth is the decision threshold
Begin
(1) Build classifier C with LR on TR data set;
(2) Temp_th=Init;
(3) Use C to classify TE and get the sorted positive probability rank list of each article in TE;
(4) Compute the current F score based on Init value;
(5) Temp_f=current F;
(6) Temp_th=Init-ζ;
(7) Compute the current F score based on Temp_th value;
(8) If (Temp_f<current F)
(9) {
(10) Temp_th=Temp_th-ζ;
(11) Temp_f=current F;
(12) Goto (7);
(13) }
(14) Else
(15) {
(16) Dth=Temp_th;
(17) Exit;
(18) }

ALTDP will produce a lower threshold of decision probability than 0.5. With this threshold, the higher F value of classification will be achieved. This means that decreasing the threshold will make the classifier reduce the misclassification error of minority class examples.

D. Machine learning details

We use LR-TRIRLS [15] to implement logistic regression. LR-TRIRLS implements a fast optimization method to fit a logistic model to given data points, and it can accommodate high-dimensional features. LR-TRIRLS accepts either dense numeric feature vectors or sparse binary feature vectors. The latter is adopted in our experiments. LR-TRIRLS is a parameter-free tool. Any model parameter is not needed to assign.

E. Experiment design

We use 10 runs of 5 fold cross validation on PH-Data in the experiments. There are 50 data sets generated. Firstly, we compute the optimal threshold of decision probability (Dth) with ALTDP.(ζ=0.01). For removing stopwords (RS) and not removing stopwords (NS), two average thresholds (Dth1 and Dth2) will be learned from all four folds of 50 data sets. The stopwords list comes from Pubmed stopwords [20].

Secondly, we apply Dth value to the experiments. The experiments are designed to i) compare removing stopwords (RS) with not removing stopwords (NS), and ii) compare default threshold 0.5 and ALTDP threshold. For each run, the same 5 fold partitions are used for the following four settings: (RS, 0.5), (RS, Dth1), (NS, 0.5), and (NS, Dth2).

For the experiment evaluation, we use F value as the main evaluation criteria. We present the average F scores on multiple runs.

Ⅳ. RESULT AND DISCUSSION

Figure 1 shows the changing trend of F value when thresholds varies from 0.5 to 0.1 at RS and NS settings with LR. We can observe that when the class distribution is imbalanced, F value increases significantly with the decreasing of threshold. When threshold decreases to some degree (the optimal Dth) , F value will reach to the max value. With the continuing decrease of threshold, the F value will decrease. This expresses the feasibility of ALTDP algorithm. At RS setting, Dth1 computed with ALTDP is equal to 0.17. At NS setting, Dth2 computed with ALTDP is also equal to 0.17.

When we get the optimal decision probability threshold, we can use it the second step. Table 1 shows the F values at four settings: (RS, 0.5), (RS, Dth1), (NS, 0.5), and (NS, Dth2).

From table 1, it can be seen that at RS setting, the performance of (RS, Dth1) outperforms (RS, 0.5). At NS setting, the performance of (NS, Dth2) outperforms (NS, 0.5). This expresses that the threshold got with ALTDP is significant for improving the classification accuracy and has the great practical value. If the stopwords are removed, it is not very important for improving the classification performance of LR.

Considering the implementation efficiency of classifier, we remove the stopwords from the PH-data to build the classifier with LR. We will use 0.17 as the decision threshold for imbalanced PH-Data classification.

Ⅴ. CONCLUSION

In this paper, we have reported the ALTDP algorithm to overcome the shortcoming of imbalanced classifier which tends to support the majority class. On the imbalanced host pathogen protein-protein interaction data set, the experimental results show that the proposed method can improve the classification performance of imbalanced

Authorized licensed use limited to: University of Louisiana at Lafayette. Downloaded on April 20,2023 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

classification. An automated classification system can be constructed to detect the host pathogen protein-protein interaction relevance of MEDLINE abstracts.

VI. ACKNOWLEDGEMENTS

This work is supported by the fund provided by National Natural Science Foundation of China 60705022.

Figure 1. The relation of F value and threshold.

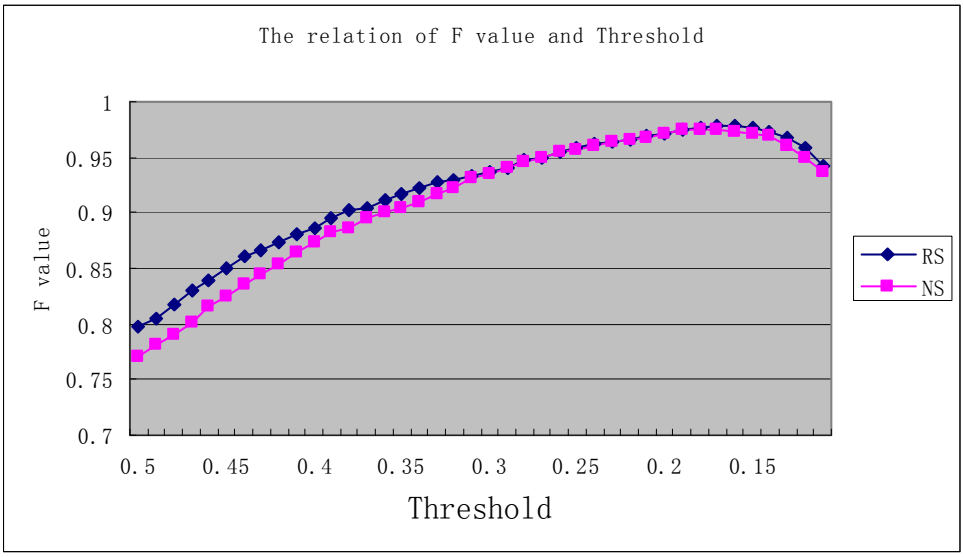

Table 1. The F values of four settings: (RS, 0.5), (RS, Dth1), (NS, 0.5), and (NS, Dth2).

Experiment setting	(RS, 0.5)	(RS,Dth1)	(NS, 0.5)	(NS,Dth2)
F value	0.877	0.987	0.862	0.987

REFERENCES

[1] CG Zhang, BA Chromy and SL McCutchen-Maloney. Host–pathogen interactions: a proteomic view. *Expert Review of Proteomics*, 2(2):187-202, 2005.

[2] K Nomura, S DebRoy, YH Lee, N Pumplin, J Jones and SY He. A Bacterial Virulence Protein Suppresses Host Innate Immunity to Cause Plant Disease. *Science*, 313 (5784): 220-223, 2006.

[3] CG Weng and Josiah Poon. A data complexity analysis on imbalanced datasets and an alternative imbalance recovering strategy. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. 2006.

[4] W Ng and M Dash. An Evaluation of Progressive Sampling for Imbalanced Data Sets. In *Sixth IEEE International Conference on Data Mining - Workshops*. 2006.

[5] SJ Yen, YS Lee, CH Lin and JC Ying. Investigating the Effect of Sampling Methods for Imbalanced Data Distributions. In *2006 IEEE International Conference on Systems, Man, and Cybernetics*. 2006.

[6] M Molinara, M.T. Ricamato, F. Tortorella.Facing Imbalanced Classes through Aggregation of Classifiers. In *14th International Conference on Image Analysis and Processing*. 2007.

[7] S Visa, A Ralescu. The Effect of Imbalanced Data Class Distribution on Fuzzy Classifiers – Experimental Study. In *2005 IEEE International Conference on Fuzzy Systems*, pages 749-754, 2005.

[8] T Yu, T Jan, S Simoff and J Debeham. A Hierarchical VQSVM for Imbalanced Data Sets. In *Proceedings of international Joint Conference on Neural Networks*. 2007.

[9] X Zhu. Lazy Bagging for Classifying Imbalanced Data.In *Seventh IEEE International Conference on Data Mining*. Pages 763-768, 2007.

[10] C Seiffert,TM Khoshgoftaar, JV Hulse, A Napolitano. Mining Data with Rare Events: A Case Study. In *19th IEEE International Conference on Tools with Artificial Intelligence*. pages 132-139,2007.

[11] G Forman. An Extensive Emporical Study of Feature Selectiong Metrics for Text Classification. *Journal of Machine Learning Research*, 3: 1289-1305, 2003.

[12] X Chen, M Wasikowski. FAST: A ROC-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems. *KDD'08*, pages: 124-132. 2008.

[13] S Ertekin, J Huang, CL Giles. Active Learning for Class Imbalance Problem. *SIGIR* 2007.

[14] S Ertekin1, J Huang, L´eon Bottou, CL Giles. Learning on the Border:Active Learning in Imbalanced Data Classification. *CIKM'07*.

[15] P Komarek and A Moore. "Making logistic regression a core data mining tool: A practical investigation of accuracy, speed, and simplicity." Institute, Carnegie Mellon University: 685-688,2005.

[16] http://en.wikipedia.org/wiki/Logistic_regression

[17] G Xu, L Yin, M Torii, Z Niu, C Wu, Z Hu and H Liu. Document Classification for Mining Host Pathogen

Authorized licensed use limited to: University of Louisiana at Lafayette. Downloaded on April 20,2023 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.

Protein-Protein Interactions. *2008 IEEE International Conference on Bioinformatics and Biomedicine*, pages 461-466，2008.

[18] CH Wu, R Apweiler, A Bairoch, DA Natale, WC Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, R Mazumder, C O'Donovan, N Redaschi and Baris Suzek The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* 2006.

[19] O Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32: D267-D270,2004.

[20] http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp&rendertype=table&id= pubmedhelp. T43

Authorized licensed use limited to: University of Louisiana at Lafayette. Downloaded on April 20,2023 at 10:23:51 UTC from IEEE Xplore. Restrictions apply.