isrjournals

International Journal of Advanced Research in Computer Science Engineering and Information Technology

Volume: 4, Issue: 3,Special Issue: 3 ,Apr,2016 ,ISSN_NO: 2321-3337

Performance Enhancement of imbalanced data using Meta-cost algorithm

M.kiruthiga[1], P.Sangeetha[2]

[1]Department of CSE, P. A. College of Engineering and Technology, Coimbatore, India.
[2]Assisstant Professor, Department of CSE, P. A. College of Engineering and Technology, Coimbatore, India.

ABSTRACT— *Class imbalance is one of the major issues in classification. It degrades the performance of data mining. It mostly occurs by the non-experts labeling the object. Online outsourcing systems, such as Amazon's Mechanical Turk, allow users to label the same objects with lack of quality. It frequently increases the cost of misclassification which arise due to imbalance.Thus, a meta-cost algorithm is projected to handle the problem of imbalanced noisy labeling and to reduce the misclassification cost. The main objective is to generate the training dataset and integrate labels of examples. This method is used to resolve the issue of minority sample and also able to deal with imbalanced multiple noisy labeling. The algorithm is applied to the imbalanced dataset collected from UCI repository and the obtained result shows that the meta-cost algorithm performs better than other methods.*

Index Terms –repeated labeling, majority voting, positive and negative labels.

1. INTRODUCTION

The online crowd sourcing systems such as Rent-A-Coder and Amazon Mechanical Turk is to acquire required services, generate ideas from a large group of people. It allows number of non-expert labelers to label the object inexpensively. Thus online crowd sourcing are gainful while comparing to traditional expert labeling methods. The cheap labels are noisy due to missing of the expertise, preference and enthusiasm. It causes imbalanced class distribution with lack of labeling quality.

Considering repeated labeling is determining multiple labels for all data points [12]. Preceding research describes repeated labeling strategies can improve the labeling quality by integrating the repeated labels using Majority Voting (MV) integration strategy. For example, considering a multiple noisy label set {+, -, +, -, +} and applying the MV, as a result final label "+" is assigned to this example since "+" obtains the highest voting.

A preceding scenario strategy of using Majority Voting (MV) for multiple noisy labels, it finalizes the class label based on the highest number of voting predicted. It assumes that all data points are uniformly distributed by integrating the labels. But the real is mislabeling are not distributed uniformly. In binary classification, labelers provide high probability for the one and significantly less probability for other [11]. When the labels are imbalanced, the count of negative labels obtained is far more than that of positive labels. When MV is applied the negative examples outnumber positive ones and the training set hold no positive examples. Sheng introduced an agnostic algorithm Positive Label frequency Threshold (PLAT) to use skewed noisy labels to stimulate an integrated label for each example [17]. It mostly handles the

issues of imbalanced noisy labeling datasets. The organization of the paper is as follows. In section 2, the related works are reviewed. In Section 3, the estimation of accuracy is analyzed. Section 4 describes the working of an agnostic algorithm. In section 5, we compare the performance of our algorithm with other method. Section 6 provides the conclusion and future work.

2. RELATED WORK

An imbalanced datasets is learned based on a combination of the SMOTE algorithm and the boosting procedure to improve the overall F-values and to get better prediction performance on the minority class [2]. He et al. evaluated the learning performance over the imbalanced learning scenario by providing a review on the state-of-the-art technologies, and the current assessment metrics [6]. Donmez described Interval Estimate (IE) Threshold to predict the experts with the highest estimated accuracy for labels [4]. Kumar defines the supervised learning methods where unsupervised counter-parts are outperformed frequently since the learner are provided with more information can permit to learn a desired pattern effectively [8].

Smyth et al. described the remote sensing applications for training the pattern recognition algorithms to detect objects of concern by considering ground-truth data as basis [14]. [7] Kajino et al. projected a convex optimization formulation for learning from crowd's .To estimate without the true labels the personal models are build for each individual crowd workers. Lo et al. described the Cost-Sensitive learning problems [10].

3. ACCURACY ESTIMATION

The true positives proportion and true negatives proportion with the total number of cases is described as accuracy and it is examined. The minority class is used as positive class and majority class as negative class; the accuracy is calculated using following equation (1),

$$\text{Accuracy} = \frac{\text{True positive+True negative}}{\text{Total number of true cases}} \quad (1)$$

The true positive (TP) is the number of correctly labeled items that belong to the positive class. The true negative (TN) is the number of correctly labeled items that belong to the negative class. The false positive (FP) is the number of items incorrectly labeled as belonging to the positive class. The false negative (FN) is the number of items incorrectly labeled as belonging to the negative class. Based on the number of instances in the test data, the correct classifiers prediction is found. The provided value and the measured values are accurately the same when 100% accuracy is obtained.

3.1 Imbalanced labeling impact on MV and PLAT

A data set containing a proportion tp of true positive examples and tn of true negative examples is considered, the class distribution is balanced if tp ≤ 0.5. A variable V is distinct to control the mislabeling percentage on the positive data points. It reflects the imbalanced labeling level, the higher level of imbalance. The labeling quality can be integrated on positive examples Pp, and Pn on negative examples if the labeling quality is same for all labelers, then Pp = (tp+Vp-V)/d and Pn = (p+V-Vp-tp)/(tn) are calculated. When applying the majority voting, we

can use Bernoulli model to calculate the integrated quality q of multiple noisy labels by using. Then α which is the ratio of the labeled number of positive examples (Pos) and negative examples (Neg) are evaluated as follows,

$$\alpha=Pos/Neg= [tpq_p+(1-d)(1-qn)] / [(tn)qn+tp(1-qp)] \quad (2)$$

The accuracy of learning model will eventually decreases when α is reduced and the number of positive examples in the final training set will also declines. It gives raise to imbalanced noisy labeling and also results in low quality labeling. Thus MV is easy to understand but for imbalanced multiple noisy labeling, the *MV* does not work [16] at all. [17] The PLAT Algorithm improves the performance of imbalanced dataset but still it doesn't reduce the cost for misclassification of attributes and class. It just spits the postive and negative classes based on estimated threshold level. The true label reovery ability of PLAT algorithm is higher than MV but while considering the total cost, the meta-cost algorithm provides the highest performance.

4. COST-SENSITIVE LEARNING (CSL) METHODS

The CSL is to build a model with minimum misclassification costs based on equation (3),

$$TotalCost = C(-, +) * FN + C(+, -) * FP \quad (3)$$

The cost-sensitive learning methods are categoried into cost-sensitive classifier and meta-cost algorithm, to lower the misclassification cost. The MetaCost is a method for creating cost-sensitive classifiers by Domingos [3]. It is a wrapper algorithm that defines any classifier can be used. The algorithm introduces a bias based on a cost matrix C (i,j) in the training data.

In tic-tac-toe dataset [1], considering a specific sample $s_i =< x_i, y_i >$ and it associates a multiple noisy label set that enclose $L_{Pos}^{(i)}$ positive labels and $L_{neg}^{(i)}$ negative labels. If the new label is different from the true label, the misclassification cost is assigned. For an example x, the probability of each class j as P(j/x) have to be found eveni f we know the potential cost of misclassifying the example as class i for each possible $\sum_j^1$ C(i/j). We can also relabel the example to its least costly prediction. The new label react the Bayes optimal prediction which seeks to minimize the conditional risk R (i/x),

$$R(i/x) = \sum_j^1 P(j/x)C(i/j) \quad (4)$$

The conditional risk R(i/x) is the expected cost of predicting that an example x belongs to class i. In the MetaCost algorithm, the obtained optimal predictions are relabeled to the examples in the training set with their given estimated probabilities and misclassification costs. The predictions produced by the base classifier should then be sensitive to the cost of misclassifying examples. To obtain the efficient result, the meta-cost algorithm is introduced to process the noisy dataset more effectively.

Meta-cost Algorithm

- S is the training set.
- L is the classification learning algorithm.

• C_M is a cost matrix.
• m is the number of examples in each resample.
• n is the number of resamples to generate.

STEP 1: For i in range 1 to m

a) Create S_i as a resample of S with m examples.

b) Create model Mi by applying L to S_i.

STEP 2: For each example x in S do

(a) For each class j do

i. Create $P(j|x) = (1/\sum^1_j) \sum^1_j P(j|x, Mi)$

(b) Change the class of x to the class k that minimize $\sum^1_j P(j|x)C_M(k, j)$

STEP 3: Create final model M by applying L to S.

If L does not produce class probabilites, MetaCost sets P (j|x, M_i) = 1 for the class L produces P (j|x, M_i) = 0 for all other classes. When calculating P (j|x), we can also choose to not include those M_i where x belong to the corresponding resample S_i. The advantage of this is that the model M produced will have a lower statistical bias. The algorithm is to create m resamples S_i from set S. In each resample there will be a different bias in the distribution of classes, thus creating different models M_i by applying L. Based on it, you create a bias in the training set L by relabeling each x to the class that gives the lowest predicted total cost. P (j|x) can be seen as a prediction of the confusion matrix for the model. In step 2 b) the algorithm is essentially relabeling the class of x to the class that is in the final model M gives the lowest total cost and the accuracy is calculated. Thus, the algorithm solves the imbalance problem and improves the label quality [12] [13].

5. EXPERIMENTS

The performance of meta-cost algorithm is estimated on conducting experiment on tic-tac-toe dataset is shown in Fig. 5.1. The dataset includes 10 attributes with 958 examples. It is categorized into 626 positive label and 332 negative labels. The meta-cost algorithm is applied and the accuracy is calculated and is shown in Fig 5.5. The meta-cost algorithm produces the highest accuracy value and reduces the total expected cost than other method is shown in Table 2.

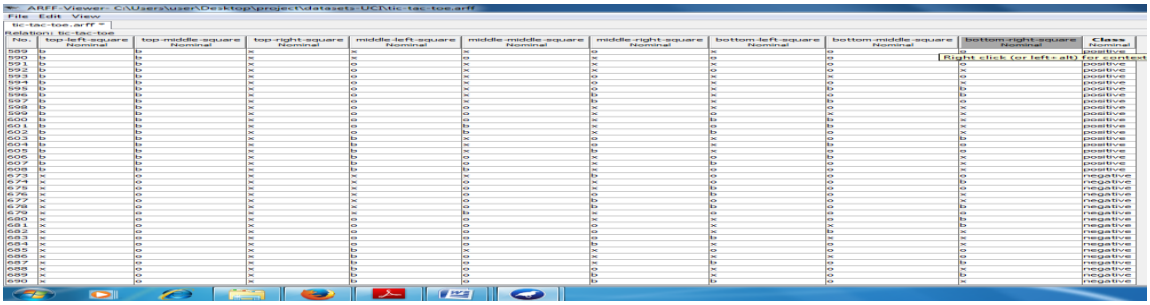

Figure 5.1 Tic-tac-toe dataset

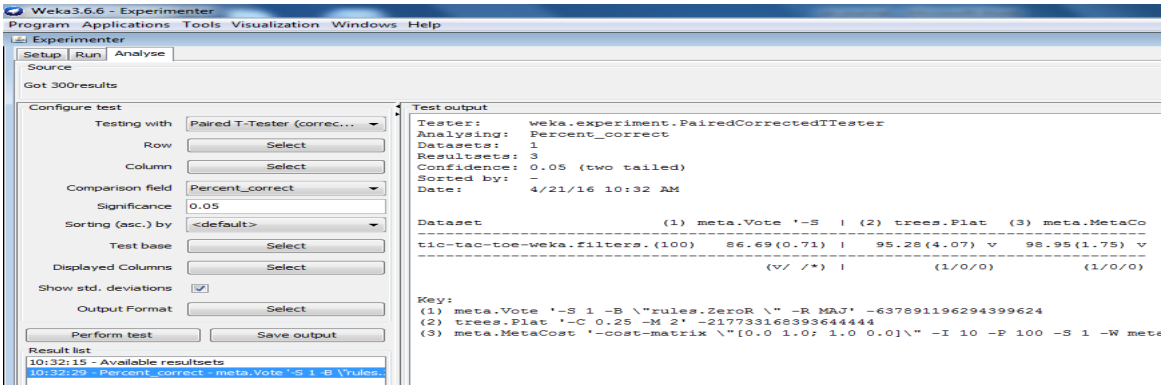

Figure. 5.5 Accuracy estimation for MV, PLAT and Meta-Cost Algorithm

TABLE 2
Performance comparison of algorithms on Tic-tac-toe dataset

Algorithms	Training data		Testing data	
	Accuracy	Error rate	Accuracy	Error rate
MV	57	42.26	86.6	13.31
PLAT	91	8.86	95.8	4.17
CSC	95	4.86	97	2.87
Meta-Cost	98.43	1.56	98.78	1.21

TABLE 3
Misclassification cost of imbalanced dataset

METHODS	MISCLASSIFICATION COST
MV	626
PLAT	128
CSC	61
Meta-cost	16

6. CONCLUSION

In this paper the meta-cost algorithm performs well on the imbalanced labeling dataset and it does not require any knowledge of labelers labeling quality. The meta-cost algorithm is suitable for both single and multiple labeling. The experimental result shows that the meta-cost algorithm performs well in reducing the total misclassification cost of imbalanced datasets.

REFERENCE

#1. C. L. Black and C. J. Merz. UCI repository of machine learning database [Online]. Available: http://archive.ics.uci.edu/ml/, 1998.

#2. N. V. Chawla, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

#3. Domingos P,' Meta-cost: A general method for making classifiers cost-sensitive,' in KDD, pp 155-164.

#4. P. Donmez, J. G. Carbonell, and J. Schneider, "Efficiently learning the accuracy of labeling sources for selective sampling," in Proc.15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2009, pp. 259–268.

#5. H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

#6. H. Kajino, Y. Tsuboi, and H. Kashima, "A convex formulation for learning from crowds," in Proc. 26th AAAI Conf. Artif. Intell., 2012, pp. 73–79.

#7. A. Kumar and M. Lease, "Modeling annotator accuracies for supervised learning," in Proc. 4th ACM WSDM Workshop Crowd sourcing Search Data Mining, 2011, pp. 19–22.

#8. X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory under sampling for class imbalance learning," in Proc. IEEE 6th Int. Conf. Data Mining, 2006, pp. 965–969.

#9. H. Y. Lo, J. C. Wang, H. M., Wang, and S. D., Lin, "Cost-sensitive multi-label learning for audio tag annotation and retrieval," IEEE Trans. Multimedia, vol. 13, no. 3, pp. 518–529,2011.

#10. C. Parker, "On measuring the performance of binary classifiers," Knowl. Inform. Syst., vol. 35, no. 1, pp. 131–152, 2013.

#11. V. S. Sheng, "Simple multiple noisy label utilization strategies," in Proc. IEEE 11th Int. Conf. Data Mining, 2011, pp. 635–644.

#12. V. S. Sheng, F. Provost, and P. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, nosiy labeler," in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2008, pp. 614–662.

#13. P. Smyth, M. C. Burl, U. M. Fayyad, P. Perona, and P. Baldi, "Inferring ground truth from subjective labeling of venus images," Adv. Neural Inform. Syst., vol. 8, pp. 1085–1092, 1995.

#14. R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast— But is it good?" in Proc. Conf. Empirical Methods Natural Lang. Process., 2008, pp. 254–263.

#15. P. Welinder and P. Perona, "Online crowdsourcing: Rating annotators and obtaining cost-effective labels," in Proc. Workshop Adv.Comput. Vis. Humans Loop, 2010, pp. 25–32.

#16. J. Zhang, X. Wu, and Victor S. Sheng,"Imbalanced Multiple Noisy Labeling", vol 27, feb 2015.