POSITIVE LABEL FREQUENCY THRESHOLD ALGORITHM FOR IMBALANCED CLASS DISTRIBUTION

M. Kiruthiga[1]
[1](Department of CSE, P. A. College of Engineering and Technology, Pollachi,Coimbatore,India, kiruthigamanickam31@gmail.com)

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. Thus, an agnostic algorithm Positive LAbel frequency Threshold (PLAT) is projected to handle the problem of imbalanced noisy labeling. 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 PLAT performs better than other methods.*

***Keywords*—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 [11]. 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 and completes the quality of labels are higher. 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 [10]. For example, mostly labeling on minority examples is error-prone and it is not unusual. In this scenario, the algorithm handles minority as the positive class. While 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 outnumbers positive ones and the training set hold no positive examples.

We introduced an agnostic algorithm PLAT to use skewed noisy labels to stimulate an integrated label for each example. 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 [5]. Donmez described Interval Estimate (IE) Threshold to predict the experts with the highest estimated accuracy for labels [3]. 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 [7]. 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 [13]. [6] 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. Strapparava et al. presents the Affective Text task to focus on the labeling of emotions and valence classification in news headlines, and is intended as an exploration of the connection between emotions and lexical semantics [15].

Lo et al. described the Cost-Sensitive learning problems [9]. It is based on audio tag annotation task and the cost sensitive classification issues are solved by considering the tag count as costs. Our work is different and the examples

are given the higher priority. Two classes are treated equally in our work.

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),

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

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 psitive (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.The accuracy is the evaluation of classifier on a set of test data. 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.

A. Imbalanced labeling impact on mv

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,

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

For example, the class distribution is balanced, if the value of d=0.5 and 0.5 < p < 1 with increase in number of labels and decrease in α value by applying MV. Thus 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. If the distribution of class is imbalance, then the outcome will be worse. Thus MV is easy to understand but for imbalanced multiple noisy labeling, the MV does not work [17] at all. Certain sampling techniques [4], [8] may also be used but the limitation over that method is the important information also gets eliminated.

3. PLAT ALGORITHM

The threshold algorithm is to check and create an effective label for multiple noisy label dataset. In mushroom 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. Using it the frequency of positive and negative labels are determined using equation (3),

$$Freq_p = \frac{Lpos}{Lpos+Lneg} = 1\text{-}Freq_n \quad (3)$$

To obtain the efficient result, we introduced the technique Positive Label Frequency threshold (PLAT) algorithm to process the noisy dataset more effectively. The sample set is considered as input that contains the examples with multiple noisy label set. Finally, the positive and negative are listed.

Algorithm

1. For each i∈Sample_set do
2. Calculate F_i and insert it into frequency_table
3. Initialize final labels of samples to be negative
4. Sort (frequency_table) in ascending order of F
5. N0:=size of (sample set)
6. $N_{L1} := N_{R1} := 0$
7. P=EstimateThresholdPosition(frequency_table, N0, N_{L1}, N_{R1})
8. $P0_{max} = (N_{L1} - N_{R1}) * N_{R1}/(N_{L1} + N_{R1}) + N_{R1}$
9. L=size of(frequency_table)-1
10. $Pos = 0$
11. While L>P do
12. Category (F_L)=pos
13. $Pos = Pos + sizeof\ (items(F_L))$
14. F_m=($F_0 + F_t$)*θ
15. L=P
16. While
17. $F_L > F_m$&$Pos + sizeof\ (items(F_L)) < P0_{max}$ do
18. Category (F_L)=pos
19. $Pos = Pos + sizeof\ (items(F_L))$
20. For i=0 to size of(frequency_table)-1 do
21. Insert items(F_i) into list_p or list_n according to category (F_L) value
22. Return list_p and list_n.

Initially we have to split the given frequency table into multiple range intervals. We can directly classify the samples whose values are greater than the specified threshold T value as positive samples. For interval with the $Freq_p$ values less than and equal to the threshold T, then the middle value of $Freq_p$ of the interval Fm is computed. We consider that the data points whose $Freq_p$ values are greater than fm and close to threshold T have high probability to be positive. The remaining data points have high probability to be negative. The algorithm shows that the category is found by the positive and negative cases proportion. Finally the algorithm return the positive and negative lists and the accuracy is calculated. Thus, the algorithm solves the imbalance problem and improves the label quality [12].

A. Estimate threshold position algorithm

The sorted frequency_table, N0 as input, the position P in sorted frequency_table whose value is treated as threshold T is evaluated.

1. Add position 0 to max_set
2. For i=1 to size of (frequency_table)-2 do
3. a_0=size of(items(F_i))-sizeof(items($F_i - 1$))

4. b_0=sizeof(items($F_i + 1$))-sizeof(items(F_i))
5. if $a_0 \geq 0 \& b0 \leq 0 diff\ (F_i, F_{min\_set(last)},)$
6. then add i into max_set
7. if $a_0 \leq 0 \& b0 \geq 0 diff\ (F_{min\ \ set(last)}, F_{i_0})$
8. then add i into minima_set
9. $P_0 = argmax_j\{sizeof\ (items\ (F_{j_0}))|F_{j_0} < 0.5, j \in \max\ \_set\}$
10. $P_1 = argmax_k\{sizeof\ (items\ (f_{j_0}))|f_{j_0} < 0.5, k \in \max\ \_set\}$
11. If P_0 & P_1 are not found then valley = $argmin_l\{sizeof\ (items\ (F_l\ ))|F_{p0} < F_l\ < F_{p_1}, l \in min\_set$
12. If valley found then P=valley else P=P_0
13. N_{L1}=$\sum_{i=0}^{p} sizeof\ (items(Fi))$
14. While $N_{L1} < N0/2$ do
15. P=P+1;
16. $N_{L1} = N_{L1} + sizeof\ (items(Fp))$
17. N_{R1}=N0-N_{L1}
18. Return P, N_{L1} and N_{R1}

This algorithm describes that the probability of positive and negative sample computation and return the position for each example and then it is used in the PLAT algorithm to list the positive and negative cases.

4. EXPERIMENTS

The performance of PLAT algorithm is estimated on conducting experiment on mushroom dataset listed in Table 1. The mushroom dataset includes hypothetical samples corresponding to 23 species of gilled mushrooms in the agaricus and lepiota family. Each species is identified as edible, poisonous.

TABLE 1
Dataset Used in Experiment

Dataset	Mushroom
Attributes	23
Examples	8124
Positive label	3916
Negative label	4208

The mushroom dataset is shown in Fig. 5.1.The PLAT algorithm is based on the distribution of positive and negative labels and the accuracy is calculated for it based on the labeling. Each non-numeric attribute is converted to numeric values and the missing attribute is assigned to zero. Then, the positive and negative labels are assigned to the each tuples.

Using majority voting, the accuracy is evaluated and is shown in Fig. 5.2. The position is estimated to each element as shown in Fig 5.3. Then the accuracy is evaluated for PLAT algorithm and is shown in Fig 5.6.

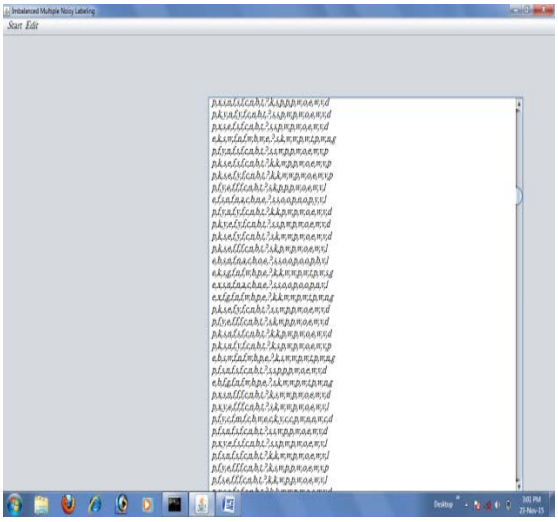

Fig. 5.1 Mushroom dataset

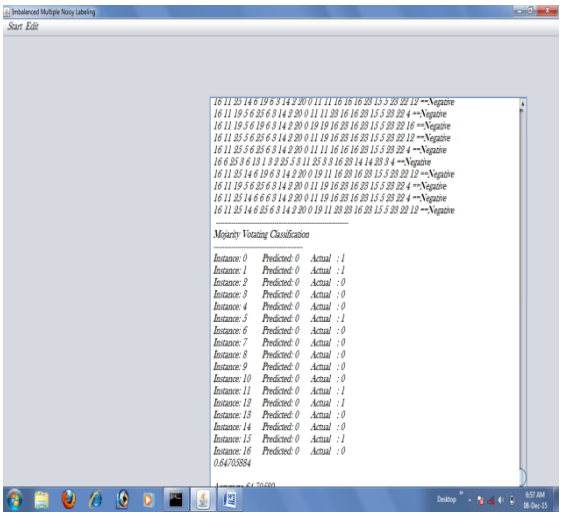

Fig. 5.2 Accuracy of Majority Voting

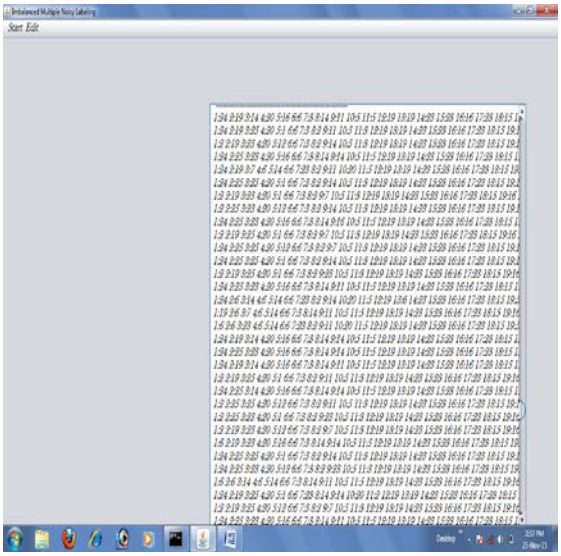

Fig. 5.3 Position Estimation using Estimate Threshold Position Algorithm

Finally, the PLAT algorithm is compared with majority voting method. Under imbalanced class distribution, the performances of both methods are evaluated.

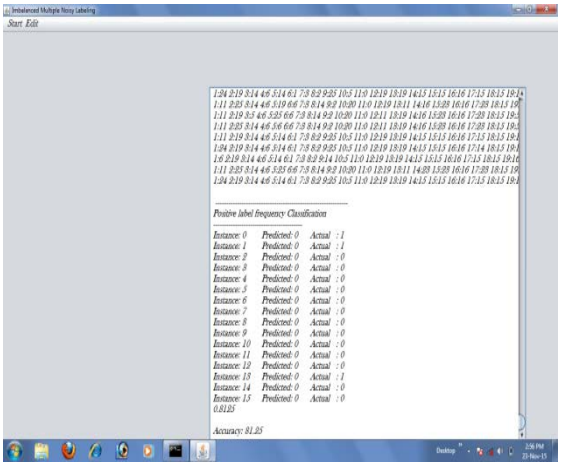

Fig. 5.4 Accuracy estimation for PLAT Algorithm

The PLAT algorithm is agnostic and it produces the highest accuracy value when comparing with the Majority Voting is shown in Table 2.

TABLE 2

Performance comparison on mushroom dataset

METHOD	ACCURACY
MV Method	64.5
PLAT Algorithm	81.5

The MV and PLAT algorithm are compared based on the accuracy result and shown in Fig 5.5.

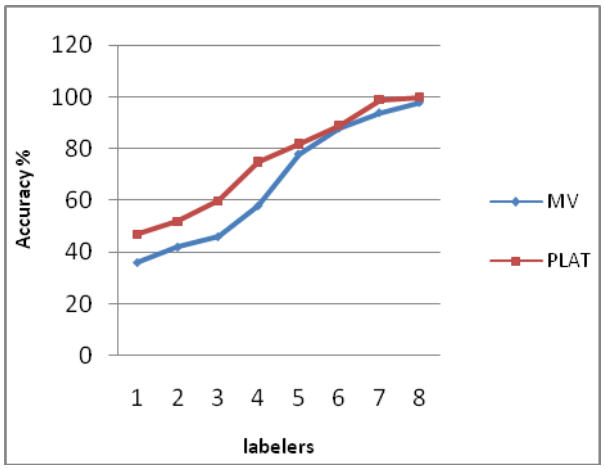

Fig. 5.5 Results of compared methods on mushroom datasets

5. CONCLUSION

In this paper the PLAT algorithm performs well on the imbalanced labeling dataset and it does not require any knowledge of labelers labeling quality and it can be used for both balanced and imbalanced labeling. The experimental result shows that it performs well and in future the cost-sensitive learning can be studied to reduce the misclassification cost.

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] 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.

[4] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," Comput. Intell., vol. 20, no. 1, pp. 18–36, 2004.

[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, Jun. 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. Process. 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] C. Strapparava and R. Mihalcea, "SemEval-2007 Task 14: Affective text," in Proc. 4th Int. Workshop Semantic Eval., 2007, pp. 70–74.

[16] 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

[17] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in Proc. Adv. Neural Info. Process. Syst. 22, 2009, pp. 2035–2043.

[18] J. Zhang, X. Wu, and Victor S. Sheng,"Imbalanced Multiple Noisy Labeling", vol 27, feb 2015.