Two stage Decision Tree Learning from Multi-class Imbalanced Tweets for Knowledge Discovery

Salina Adinarayana
Department of Information Technology, Shri Vishnu Engineering College for Women Bhimavaram, INDIA
s.suhasini2k9@gmail.com

E.Ilavarasan
Department of Computer Science Engineering, Pondicherry Engineering College Pondicherry, INDIA

Abstract— Data Mining is an efficient technique for knowledge discovery from existing databases. The existing algorithms performance degrades when applied to the multi class imbalance dataset. The imbalance nature of twitter data set also hinders the process of efficient knowledge discovery. In this paper, we proposed an efficient learning approach for knowledge discovery from multi class imbalance datasets specifically designed for opinion mining. The proposed Under Sampled Imbalance Decision tree Learning (USIDL) approach uses decomposition of multi class into number of binary class samples followed by a unique technique for under sampling the instances from majority subset of each binary sample. The experimental results suggest that the proposed technique performs better than the existing C4.5 algorithm on six evaluation metrics.

***Keywords**- Knowledge Discovery, Decomposition, multi class binary sample,Imbalance Decision tree Learning, Under Sampling, USIDL*

*****

I. INTRODUCTION

Mining opinion targets and opinion words from online reviews are important tasks for fine-grained opinion mining, the key component of which involves detecting opinion relations among words. The challenging task of opinion mining is about the data, which is available for sentiment analysis is of with different formats or styles. In order to compare users with different opinions on a specific topic and analyze large volume of opinion data, we use a real world twitter dataset. The tweets are manually selected from the social blogging sites. The dataset consists of reviewer's id, reviewer's name, tweets they have made and time and year. Then we have to transform the data into a format that will be more easily and effectively processed using data mining techniques to conduct the analysis for knowledge discovery.

In imbalanced datasets, the class of interest is generally a small fraction of the total instances, but misclassification of such instances is often expensive.

Significant growth of research on the class imbalance problem for binary class datasets is going on, multi-class datasets have received considerably less attention. This is partially due to the fact that the multi-class imbalance problem is often much harder than its related binary class problem, as the relative frequency and cost of each of the classes can vary widely from dataset to dataset.

The multi-class classification problem is an extension of the traditional binary class problem where a dataset consists k classes instead of two. While imbalance is said to exist in the binary class imbalance problem when one class severely outnumbers the other class, extended to multiple classes the effects of imbalance are even more problematic. That is, given k classes, there are multiple ways for class imbalance to manifest itself in the dataset. One typical way is there is one "super majority" class which contains most of the instances in the dataset.

Decomposition techniques are become a powerful tool in the data mining community convert multi-class problems into binary class problems. While decomposing the problem into binary class, consider size of the generated decomposition as important consideration.

The dataset in which one class is underrepresented when compared to other class i.e. the ratio of instances in one class predominantly more than the other class then such a class is known as class imbalance data.

The vast number of publications in the field of opinion mining indicates the recent importance given by the researchers for opinion mining. There are numerous contributions in the field of opinion mining and decision trees. We have considered some of the main recent proposals as follows:

Rifkin et al., [1] have discussed various techniques on how to decompose Multiclass classification problem into several binary class problems.

Christoph Dorn et al., [2] have built a framework for best team discovery using the techniques of genetic algorithm and simulated annealing. The unique self adjusting mechanism improves and reaches to the best combination of the connectivity for team formation. T Michele Filannino et al., [3] have proposed an approach using different feature types for general domain temporal identification and normalization. The rules based conditional post-processing techniques are used for unique implementation.

Desheng Dash Wu et al., [4] have proposed a efficient framework for analyzing stock market using support vector machines ad auto regressive conditions. AlexanderPakWu et al., [5] have proposed an approach for micro-blogging, for the task of sentiment analysis using linguistic analysis on twitter corpus. AameraZ.H.Khanet al., [6] have proposed a sentiment analysis method using a lexicon based entity-level approach.

Saif Hassan et al., [7] have proposed an approach which uses both the semantic feature set and sentiment-topic feature set for efficient sentiment analysis. The features in the twitter corps are replaced with the modified features using the unique techniques proposed. Apoorv Agarwal [8] have proposed a method for efficient twitter data analysis for appropriate feature selection using POS-specific prior polarity. Hao Wang et al., [9] have proposed a specific approach on twitter corpus for sentiment analysis of public towards presidential candidates in the 2012 U.S. elections.

Efthymios Kouloumpiset al., [10] have proposed an approach for sentiment analysis using prominent linguistic features on twitter messages. The information for sentiment analysis is analyzed using both the lexical and feature resources. AndranikTumasjanet al., [11] have conducted a content analysis on 100,000 messages on political discussion from twitter corps using LIWC text analysis software. Daniel M. Romeroet al., [12] have proposed a methodology which finds the weights of the links on twitter with different users on different topics of discussion for their crucial roles. Alec Goet al., [13] have investigated different approaches using Naive Bayes, Maximum Entropy, and SVM classifiers for classifying sentiment of messages on micro-blogging services like Twitter. These are some of the recent contribution which used twitter for opinion mining.

López, Victoria, et al.,[17] have reviewed the topic of classification with imbalanced datasets, and focused on approaches to deal classification of imbalanced datasets and effect of data intrinsic characteristics in learning from imbalanced datasets. Hsu, Chih-Wei, and Chih-Jen Lin [18] have discussed decomposition implementations for altogether methods of multi class problem and compared their performance with three methods based on binary classifications.

II. Proposed Technique

This section presents the detailed architecture of the proposed technique Under Sampled Imbalance Decision tree Learning (USIDL) which consists of two stage process namely decomposition followed by under sampling the binary samples obtained in stage2. The detailed working principles of the USIDL technique are explained below:

Stage 1:
The class imbalance twitter corpus consists of the multi class opinions. In the existing opinions, it consists of positive, negative and neutral opinions. Most of the researchers concentrated on binary class imbalanced datasets making positive opinion set as majority class and negative opinions as minority class and leaning neutral opinions aside. Here we are concentrating on multi class imbalanced dataset. A decomposition technique is employed to convert the multi class opinions into ‘n’ number of binary class opinion samples$bc_1, bc_2, bc_3, \ldots, bc_n$. The decomposition is done using One-Versus-All Decomposition (OVA) which is discussed as follows:

i. In this decomposition, given c classes, c classifiers are built such that each one considers one of the classes to be the positive class while the remaining are combined into a negative class.
ii. When a new instance is seen, each classifier returns a probability estimate for the instance.
iii. An overall probability estimate is then obtained by combining each of the individual probability estimates into a vector of length c, and normalizing.

Stage 2:
i. Each bc_i is again decomposed in into majority and minority class samples for further process. The majority subset of each bci is considered for keen observation for knowledge discovery, since the main idea behind this proposal is to perform under sampling for all ‘n’ binary class samples.

ii. In the process of under sampling the instances from each bc_i’s majority subset is to be reduced by identifying and removing the noisy, outlier and weak range of instances using feature pruning technique discussed in section V. This results generation of strongest majority set for that binary sample.

iii. Using this strongest majority set generated in step ii of stage 2 and the minority set of the binary sample formed earlier, perform comparison analysis of this USIDL with bench mark algorithm using evaluation metrics discussed in section IV.

Repeat this two stage learning technique for all the n binary samples formed in stage 1 to discover the desired knowledge from the selected imbalanced dataset.

III. Feature Pruning to eliminate unnecessary Features

It works on the principle for each sentence in the database, if it contains any frequent feature, extract the nearby adjective. If such an adjective is found, it is considered as an opinion word. A nearby adjective refers to the adjacent adjective that modifies the noun/noun phrase is a frequent feature. If no such feature is found then it will be considered as irrelevant and will be removed. The noisy and outlier instances can be easily identified by analyzing the intrinsic properties of the instances. The range of weak instances can be identified by first identifying the weak features in the majority subset. The correlation based feature selection [14] technique selects the important features by following the inter correlation between feature - feature and the inter correlation between feature and class.

The features which have very less correlation are identified for elimination. The ranges of instances which belong to these weak features are identified for elimination from every bci ‘s majority subset. The percentage of instances removed from every majority class can be determined by using the correlation based feature subset filter which uses the concept of feature to feature correlation and feature to class correlation .

The base algorithm C4.5 is applied to the new twitter dataset formed new-bc$_k$ with the combination of enhanced majority subset and minority subset for evaluation of metrics with the proposed technique.

IV. Experimental Design and Evaluation Criteria’s

The important function to consider when building a decision tree is known as the splitting criterion. The experimental design used in some of the existing publications is 70-30 % split of the data for training and testing. One of the limitations of this approach is regarding proper divisibility of the data for validation. The 70 percent of the data used for training may or may not include the instances of the both classes. In random scenario, if training data contains mostly only one class instances then the build model is weak and the results generated are also weak.

To overcome the short comings of the existing publications on splitting criterion, we employ technique similar to the decomposition described in Section IV. That is, given the set of classes C, we consider each unique pair of subsets: C1 ⊂ C, C2 = C\C1 and consider all classes in C1 as the positive class, and all classes in C2 as the negative class1.Further use the 10 fold Cross Validation (CV) for ‘n’ runs. The empirical experiments conducted in large number of papers conform that the best ‘n’ value as 10. In 10 fold CV, the dataset is split into 10 independent subsets. In first run, 9 subsets are used for training and the reaming 1 subset is used for testing. We have first implemented this evaluation criterion for one random binary sample bck out of ‘n’samples formed, and then repeated the evaluation for all the other n-1 samples.

We have used the open source tool Weka[15] to implement our proposed technique and evaluated the performance on all the binary samples bcn for n=1,n using accuracy, AUC, precision, F-measure, TP Rate, TN Rate, FP Rate and FN Rate measures. The formulas of all the measures are given in the following equations.

The Area under Curve (AUC) [12] measure is computed by,

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \qquad (1)$$

The Precision [11] measure is computed by,

$$\Pr = \frac{TP}{(TP)+(FP)} \qquad (2)$$

.

The F-score [11] Value is computed by,

$$F - Score = \frac{2 \times \Pr \times \mathrm{Re}}{\Pr e + \mathrm{Re}} \qquad (3)$$

Where Pr is Precision calculated using eq-2 and Re is Recall which is calculated [11] using the equation,

$$\mathrm{Re} = \frac{TP}{(TP)+(FN)} \qquad (4)$$

The False Positive Rate [11] measure is calculated by,

$$FalsePositiveRate = \frac{FP}{(FP)+(TN)} \qquad (5)$$

The False Negative Rate [11] measure is computed by,

$$FalseNegativeRate = \frac{FN}{(TP)+(FN)} \qquad (6)$$

V. Dataset used in This Decision tree Learning

The dataset used in this work are tweets with multi class opinions, after decomposition of the multi class tweets into ‘n’ number of binary class datasets bcn for n=1,n, a random sample of binary class bck considered for analysis which consists of 361 opinions, in which 304 are positive opinions and 57 are negative opinions as shown in Table 1. The imbalance ration(IR) of the bck-sample is found to be 5.34.

TABLE I. The UCI datasets of BC$_K$ sample and its properties

S.No	Dataset	Instances	bck-Majority	bck-Minority	IR
1	Twitter	361	304	57	5.34

VI. Results and Discussion

The comparative study of our proposed learning technique is done with C4.5 [15] benchmark algorithm. The results of the technique on the twitter bc$_k$-sample using weka tool are presented in the table 2.

The accuracy of the proposed technique is increased from 81.36 to 92.52. The AUC value of the proposed approach reduced from 0.508 to 0.496 due to some of the unique properties of the dataset which reduced due to the under sampling approach. The precision value of the proposed approach is increased from 0.840 to .929 which helps in improving the prediction of the specific class.

The F-score value of the proposed approach improved from 0.896 to 0.961. The recall value of the proposed approach is improved from 0.962 to 0.996. The FP Rate of the proposed technique is increased from 0.976 to 1.0. The FN rate of the proposed technique is decreased from 0.038 to 0.004.

TABLE II. Test Results of C4.5 versus USIDL on evaluation metrics

SNo.	Measure	C4.5	USIDL
1	Accuracy	81.36±3.02●	92.52±1.79
2	AUC	0.508±0.082	0.496±0.016
3	Precision	0.840±0.015●	0.929±0.013
4	F-Score	0.896±0.019●	0.961±0.010
5	Recall	0.962±0.036●	0.996±0.015
6	FP Rate	0.976±0.069●	1.000±0.000
7	FNRate	0.038±0.036●	0.004±0.015

● Marked value represents the win of USIDL over C4.5

The evaluation metrics discussed in table 2 for the compared C4.5 algorithms versus the proposed technique on the bc$_k$ sample dataset are show in the Figure 1.

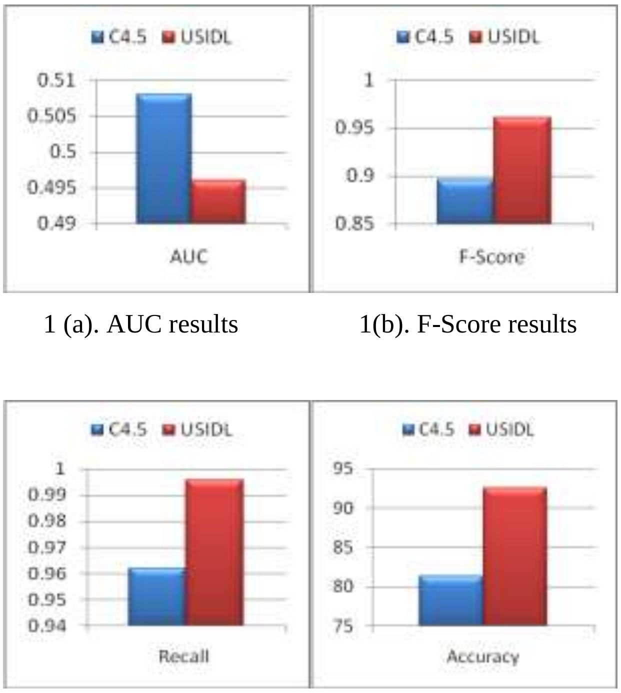
The results shows that there is a good improvement in f-score, recall, accuracy, precision, and FP rate. This improvement is due to the unique technique used by the proposed technique for under sampling the unnecessary and noisy instances from the majority subset of all decomposed binary samples for the better sentiment analysis.

From Table 2 and Figure 1 we can see that our proposed technique had given a proper solution for the investigated question. The slight fall in the result of AUC, is due to the removal and shifting of instances from majority subset to minority subset. The theoretical concept presented, support the results of the empirical findings in the proposed experiments. We have repeated the USIDL technique on the remaining n-1 decomposed binary samples and it outperforms when we have done the comparative analysis with C4.5 results on the above discussed evaluation metrics.

VII. CONCLUSIONS

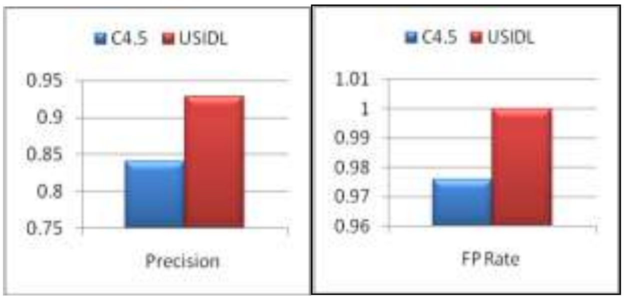
In this work, we have proposed a novel decision tree learning technique for opinion mining from multi class imbalance twitter corpus. The proposed technique uses decomposition of multi class sample into multiple binary samples and then applied under sampling on the instances from every binary sample's majority subset. The experimental results suggest that the proposed learning technique performs well on six evaluation measures out of seven when compared with bench mark decision tree learning algorithm C4.5.

In future work, an efficient technique using more replication of instances in multi class imbalance tweets of ecommerce transactions data in minority subset using unique techniques can be done to mine opinions for product recommendation.



1 (a). AUC results 1(b). F-Score results

1(c). Recall results 1(d) .Accuracy results



1(e) AUC results 1(f). F-Scor results

Figure 1. Test results of C4.5 and USIDL on bc$_k$ binary sample dataset.

REFERENCES

[1]. Rifkin, Ryan,"Multiclass classification." Lecture Notes, Spring08. MIT, USA (2008).

[2]. Christoph Dorn, Florian Skopik, Daniel Schall, SchahramDustdar," Interaction mining and skill-dependent recommendations formulti-objective team composition", Data & Knowledge Engineering 70 (2011) 866–891.

[3]. Michele Filannino, GoranNenadic," Temporal expression extraction with extensive feature type selection and a posteriori label adjustment", Data &Knowledge Engineering 100 (2015) 19–33

[4]. Desheng Dash Wu, LijuanZheng, and David L. Olson," A Decision Support Approach for Online StockForum Sentiment Analysis", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. 44, NO. 8, AUGUST 2014.

[5]. Alexander Pak, Patrick Paroubek " Twitter as a Corpus for Sentiment Analysis and Opinion Mining", Conference Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010,Valletta, Malta, ISBN 2-9517408-6-7.

[6]. AameraZ.H.Khan, Dr. Mohammad Atique, Dr. V. M. Thakare, "Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis" National Conference on "Advanced Technologies in Computing and Networking"-ATCON-2015Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477.

[7]. Saif, Hassan; He, Yulan and Alani, Harith (2012). Alleviating data sparsity for Twitter sentiment analysis. In: 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at the 21st International Conference on theWorld Wide Web (WWW'12), 16 April 2012, Lyon, France, CEUR Workshop Proceedings (CEUR-WS.org), pp. 2–9.

[8]. ApoorvAgarwal, BoyiXie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau," Sentiment Analysis of Twitter Data", Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 30–38, Portland, Oregon, 23 June 2011.cO2011 Association for Computational Linguistics.

[9]. Hao Wang, Dogan Can, Abe Kazemzadeh," A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle", Proceedings of the 50th Annual

Meeting of the Association for Computational Linguistics, pages 115–120, Jeju, Republic of Korea, 8-14 July 2012. c 2012 Association for Computational Linguistics.

[10]. EfthymiosKouloumpis, TheresaWilson, Johanna Moore,” Twitter Sentiment Analysis: The Good the Bad and the OMG!”, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media,

[11]. AndranikTumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe,” Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[12]. Daniel M. Romero, Jon Kleinberg,” The Directed Closure Process in Hybrid Social-Information Networks,with an Analysis of Link Formation on Twitter”,Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[13]. Go, A.; Bhayani, R. & Huang, L. (2009), 'Twitter Sentiment Classification using Distant Supervision', Processing, 1--6.

[14]. Mark A. Hall,” Correlation-based Feature Selection for Machine Learning”,PhD Thesis, The University of Waikato, aprill 1999.

[15]. Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

[16]. J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA:Morgan Kaufmann, 1993.

[17]. Victoria López ,Alberto Fernández, Salvador García ,Vasile Palade , Francisco Herrera. "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics." Information Sciences 250 (2013): 113-141.

[18]. Hsu, Chih-Wei, and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines." IEEE transactions on Neural Networks 13.2 (2002): 415-425.