Improved opinion mining using combination samplings for class imbalance twitter corpus sentiment analysis

P Ratna Babu[1] and Dr. Bhanu prakash Battula[2]
[1]*Sri Chundi Ranganayakulu Engineering College, AP, India*
[2]*Tirumala Engineering College, Narasaraopet, AP, India*

***Abstract*—**Data Mining is a popular knowledge discovery technique. In data mining decision trees are of the simple and powerful decision making models. One of the limitations in decision trees is the complexity and error rate. Inspired by human learning strategies, we propose a decision tree structure which mimics human learning by performing continuous improved learning while learning imbalance twitter corpus. In this paper, we propose a novel Improved Opinion Mining using Combination Sampling (IOMCS) method based on over and under sampling strategies. Extensive experiments, using Naive Bayes (NB) decision tree as base classifier for proposed approach, show that the accuracy of our method is improved than the state-of-the-art methods.

***Index Terms*—** *Knowledge Discovery, Data Mining, Classification, Decision Trees, Improved Opinion Mining using Combination Sampling (IOMCS).*

I. INTRODUCTION

In Machine Learning community, and in data mining works, classification has its own importance. Classification is an important part and the research application field in the data mining [1]. A decision tree gets its name because it is shaped like a tree and can be used to make decisions. ―Technically, a tree is a set of nodes and branches and each branch descends from a node to another node. The nodes represent the attributes considered in the decision process and the branches represent the different attribute values. To reach a decision using the tree for a given case, we take the attribute values of the case and traverse the tree from the root node down to the leaf node that contains the decision." [2]. A critical issue in artificial intelligence (AI) research is to overcome the so-called ―knowledge-acquisition bottleneck" in the construction of knowledge-based systems. Decision tree can be used to solve this problem. Decision trees can acquire knowledge from concrete examples rather than from experts [3]. In addition, for knowledge-based systems, decision trees have the advantage of being comprehensible by human experts and of being directly convertible into production rules [4].

A decision tree not only provides the solution for a given case, but also provides the reasons behind its decision. So the real benefit of decision tree technology is that it avoids the need for human expert. Because of the above advantages, there are many successes in applying decision tree learning to solve real-world problems. In this paper, we proposed an improved approach using decision tree learning on imbalance twitter corpus.

The paper is organized as follows. In Sect. 2 we present the recent advances in decision tree learning. This will directly motivate the main contribution of this work presented in Sect. 3, where we propose a new framework for Improved Opinion Mining using Combination Sampling (IOMCS). Evaluation criteria's for decision tree learning is presented in section 4. Experimental results are reported in Sect. 5. Finally, we conclude with Sect. 6 where we discuss major open issues and future work.

II. RELATED WORK

Alexander Pak et al., [5] have focused on using Twitter, the most popular microblogging platform, for the task of sentiment analysis by automatically collect a corpus for sentiment analysis and opinion mining purposes using linguistic analysis. Syed Akib Anwar Hridoy et al., [6] have discussed a methodology which allows utilization and interpretation of twitter data to determine public opinions about the iPhone 6 for male–female specific analysis. Manogna Medurua et al., [7] have analyzed the twitter posts about government issues and political reforms. The proposed framework uses Twitter as the platform to analyze the emotions of the users using Sentiment Analysis.

Amir Karami et al., [8] have proposed a computational public opinion mining approach to explore the discussion of economic issues in social media during an election. Current related studies use text mining methods independently for election analysis and election prediction; this research combines two text mining methods: sentiment analysis and topic modeling. Pooja Khanna et al., [9] have presented a very easy, cost and time effective approach that expose the opinions of much larger public (not bounded by any geographical boundaries) which otherwise would have been not possible. The study presents an exhaustive study on the efficiency of R language in opinion mining and how opinion data can be extracted from twitter database. Vaibhavi N Patodkar et al., [10] have presented a novel solution to target-oriented sentiment summarization and SA of short informal texts with a main focus on Twitter posts known as "tweets". They also developed a general solution to sentiment classification when we do not have any labels in a target domain but have some labeled data in a different domain, regarded as source domain.

Myneni Madhu Bala et al., [11] have presented sentiment analysis was done on tweets concerning the natural disasters and female specific analysis. M.V.Sangameswar et al., [12] have presented a novel utility of flume and HDFS on the analysis of twitter data is to extracted to HDFS through FLUME to the Hive is utilized for extraction and analysis of some data. Ali Hasan et al., [13] have provided a comparison of techniques of sentiment analysis in the analysis of political views by applying supervised machine-learning algorithms such as Naïve Bayes and support vector machines (SVM).

Pierre Ficamos et al., [14] have proposed so as to estimate the sentiment of a tweet using the pre-processing step required to extract features from Twitter data. This method requires to extract topics from the training dataset, and train models for each of these topics. The method allows to increase the accuracy of the sentiment estimation compared to using a single model for every topic. Naw Naw et al., [15] have aimed to perform social media sentiment analysis by applying machine learning approach of Artificial Intelligence (AI) on National Educational Rate and Crime Rate occurred in different countries. Muqtar Unnisa et al., [16] have collect information from social networking sites like Twitter and the same is used for sentiment analysis. The processed meaningful tweets are cluster into two different clusters positive and negative using unsupervised machine learning technique such as spectral clustering automated approach.

III. THE PROPOSED METHOD

In this section, the Improved Opinion Mining using Combination Sampling (IOMCS) approach is presented.

The IOMCS approach follows an approach for continuous improvement using over and under sampling strategies. It performs nearest neighbour technique for identification of noisy and outlier instances.

The different components of our new proposed framework are elaborated in the next subsections.

Step 1: Preparation of Majority and Minority Subset

The datasets is partitioned into majority and minority subsets. In hybrid sampling strategy, we concentrate on both minority and majority sub space improvement. In the first stage, we perform over sampling on minority data subset for further analysis to generate synthetic instances. In the later on stage, we perform under sampling on majority data subset for further analysis to reduce instances.

Step 2: Selection of novel subset of Instances from minority Subspace

In over sampling, minority subset can be further analyzed to find the missing or noisy instances so that we can eliminate those. For finding noisy, boarder line and missing value instances one of the ways is to go through a pre-processing process and to apply distance measure and remove the unwanted opinions.

Step 3: Generating Synthetic Instances by using novel subset in minority Subspace

The prominent instances remained in the minority subset are to be resampled i;e both replicated and hybridized instances are generated. The percentage of synthetic instances generated will range from 0 – 100 % depending upon the percentage of difference of majority and minority classes in the original dataset. The synthetic minority instances generated can have a percentage of instances which can be a replica of the pure instances and reaming percentage of instances are of the hybrid quality of synthetic instances generated by combing two or more instances from the pure minority subset.

Step 4: Performing under sampling from majority Subspace

The majority instances, which are excess in percentage than the minority subset are reduced by following the intelligent inexact technique for removal. In this technique the influential features are selected and retrieved for further utilization. The weak or less influential features are selected for removal of instances which are in the border line and range of misclassification. The process of finding such instances can be done by applying techniques of polarity finding in a semi group of instances using KNN (K Nearest Neighbor) searching algorithm. The main principle of investigation in the KNN approach is to find the percentage of opposite polarity instances in the group for identification of mostly misclassified or outlier instances.

Step 5: Forming the Strong Dataset

The oversampled minority subset and the under sampled majority subset are combined to form a strong and balance dataset, which is used for learning of a base algorithm. In this case we have used NB [7] as the base algorithm. Combining of the synthetic minority instances to the original dataset, results in the formation of an improved and almost balanced dataset. The imbalance dataset improvement can be made into balance or almost balance depending upon the pure majority subset generated. The maximum synthetic minority instances generated are limited to 100% of the pure minority set formed. Our method will be superior than other oversampling methods since our approach uses the only available pure instances in the existing minority set for generating synthetic instances. Our method will be superior than other under sampling methods since our approach performed under sampling using the instance specific technique for instance removal.

IV. EXPERIMENTAL DESIGN AND EVALUATION CRITERIA'S

In this paper, we used Twitter opinion mining dataset consisting of 1155 opinions which are divided into training

and testing part. It provides imbalance corpus data with 944 positive and 208 negative opinions for sentiment analysis.

Table 1: The Twitter datasets and their properties

S.no.	Dataset	Instances	Missing	Features	IR
1.	Twitter	1155	No	993	5.52

The twitter opinion mining dataset sample instances with features and class can be seen below,

Twitter Datasets:
@relation Twitter

@attribute Twitter numeric
@attribute body string
@attribute class {pos,neg}

@data
1229709107,'anyone feel motivated the fri afternoon prior to a holiday? wanted to get lots done... but i want jammies and judge judy... \"SIR!\" <3 her ',pos

1231217680,'I had the same issue with dominions site. Fixed it by using internet explorer ',neg
……
…….
…….
In most of the cases, the analysis of the twitter dataset was done assuming it as a balance dataset. We propose to analyze the twitter dataset as an imbalance dataset, the reason is, almost all the real world datasets are in imbalance nature. The existing algorithms are not efficient in discovering the hidden knowledge from the imbalance twitter dataset. We proposed a novel IOMUS algorithm for efficient knowledge discovery from the imbalance twitter dataset.

Preparation of the Dataset:

Take the twitter imbalance dataset and convert the string to vector by following morphological approach. After the morphological conversion of dataset, the numbers of features generated are very high. The most important features required for the further analysis should be identified. In this work the approach used to identify the important feature subset is by considering the feature to feature correlation and feature to class correlation. The dataset with important subset of features is considered for further analysis using our proposed IOMCS approach.

Pre Processed Twitter Datasets:
@relation Twitter
@attribute = numeric
@attribute About numeric
@attribute Agis numeric
@attribute Alt numeric
@attribute Although numeric
@attribute Amazing numeric
@attribute And numeric
@attribute Are numeric
@attribute As numeric
@attribute August numeric
@attribute BTW numeric
@attribute Beach numeric
@attribute Beatz numeric
@attribute Beautiful numeric
@attribute Been numeric
@attribute Behaviour numeric
@attribute Bella numeric
@attribute Best numeric
……
…….
…….
@attribute class {pos,neg}

@data
{0 1229709107,6 1,19 1,186 1,233 1,241 1,251 1,253 1,293 1,357 1,390 1,407 1,419 1,455 1,464 1,470 1,485 1,491 1,492 1,528 1,574 1,649 1,747 1,764 1,803 1,804 1}
{0 1231217680,114 1,294 1,443 1,483 1,675 1,698 1,747 1,792 1,834 1,875 1,917 1,921 1,941 1,942 1,992 neg}

{0 1229063765,233 1,269 1,402 1,483 1,521 1,605 1,656 1,745 1,747 1,764 1,877 1,889 1,896 1,897 1,905 1,944 1,950 1,985 1,987 1,992 neg}
……
…….
…….

The experimental methodology used for experimental simulation is 10 fold cross validation. In 10 fold cross validation the data source is divided into 10 equal partitions. In each run, one of the folds is used for testing and remaining folds are used for training the model. The mean of 10 runs are used for computing of evaluation metrics such as accuracy, AUC, TP rate, TN rate etc…

We performed the implementation of our new algorithms within the Weka [28] environment on windows 7 with i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM. The validation of the results is done using 10 fold cross validation, in which the dataset is split into 10 subsets and in each run nine subset are used for training and the remaining subset is used for testing. In 10 runs, the testing subset is altered and average measures for the 10 runs are generated.
We used the positive/negative polarity of the opinions from Twitter dataset. The dataset is highly imbalanced; the majority class is “positive” with 944 opinions, while the minority class is “negative” with 208 opinions. In our

experiments, 10 fold cross validation technique is used for experimental validation. We evaluate our proposed approach with decision tree evaluator C4.5 and REP.

Table 2 summarizes the results obtained using C4.5, REP and the proposed IOMCS. We evaluated seven measures: accuracy, AUC, precision, recall, f-score, FP rate and FN rate. F-score is a more informative score since it considers both precision and recall measures. The evaluation metrics used in the paper are detailed below,

Accuracy is the percentage of correctly classified instances. AUC can be computed simple as the micro average of TP rate and TN rate when only single run is available from the classification algorithm.

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

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

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \quad \text{----------} \quad (2)$$

The Precision measure is computed by,

$$\text{Precision} = \frac{TP}{(TP) + (FP)} \quad \text{----------} \quad (3)$$

The Recall measure is computed by,

$$\text{Recall} = \frac{TP}{(TP) + (FN)} \quad \text{----------} \quad (4)$$

The F-score value is computed by,

$$F - score = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{----------} \quad (5)$$

V. RESULTS AND DISCUSSION

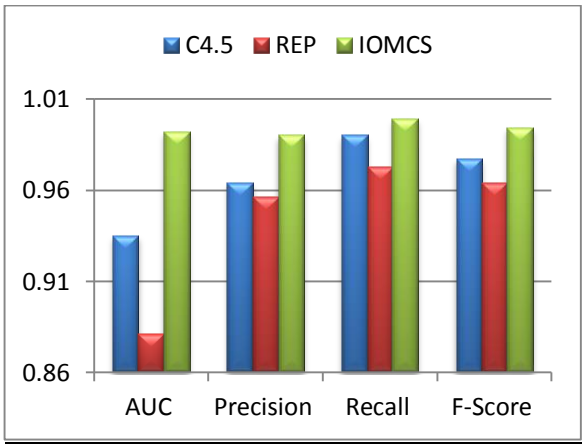
In this section, the results of the IOMCS approach are compared and discussed. The results are summarized as follows.

Table 2 shows the detailed experimental results of the mean classification accuracy, AUC, Precision, Recall, F-score, FP Rate and FN Rate of C4.5 and REP Tree on imbalance opinion mining data sets. From Table 2 we can see accuracy performance of IOMCS model that it can achieve substantial improvement over C4.5 on all the data set (7 wins) which suggests that the IOMCS model is potentially a good technique for decision trees. The IOMCS method can also gain significantly improvement over REP (7 wins) and is comparable to two state-of-the-art technique for decision trees.

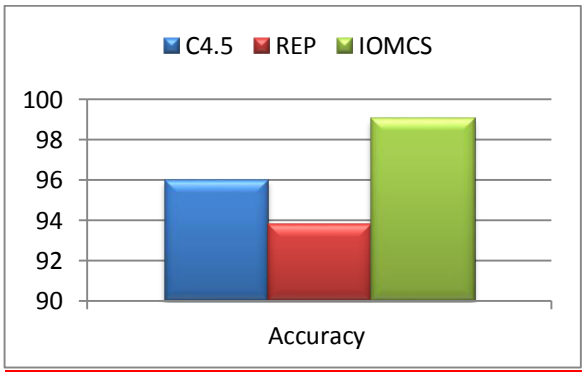
Table 2 Summary of tenfold cross validation performance on the Twitter datasets

Measure	C4.5	REP IOMCS
Accuracy	95.99±1.89●	93.80±2.63●
99.04±0.91	AUC	0.935±0.046●
0.881±0.065●	0.992±0.015	Precision
0.964±0.019●	0.956±0.021●	0.990±0.010
Recall	0.990±0.011●	0.973±0.019●
0.999±0.004	F-Score	0.977±0.011●
0.964±0.015●	0.994±0.005	FP Rate
0.209±0.111●	0.252±0.124●	0.056±0.054
FN Rate	0.010±0.011●	0.027±0.019●
0.001±0.004		

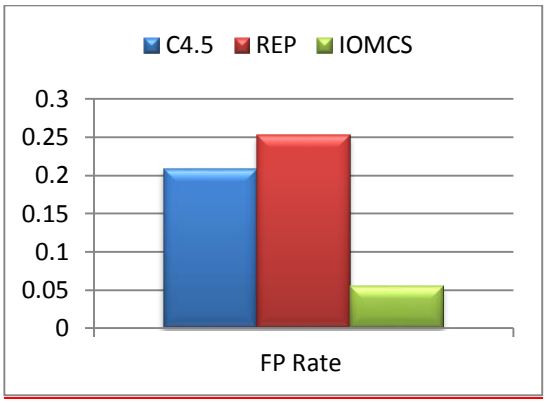
●Bold dot indicates the win of IOMCS on C4.5 algorithm;



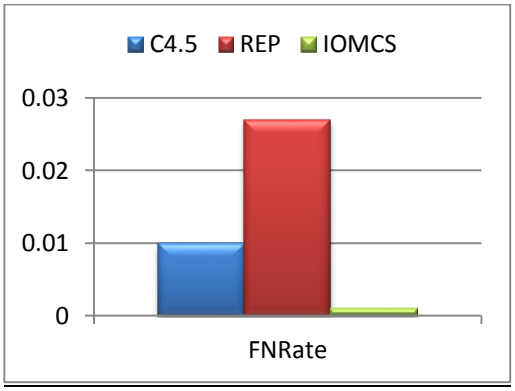
1(a)



1(b)



1(c)



1(d)

Fig. 1 (a) – (**d)** Test results on AUC, precision, recall, F-score, accuracy, FP Rate and FN Rate between C4.5, REP verses IOMCS on imbalance twitter datasets.

Figure 1(a) shows the detailed experimental results of the mean AUC, Precision, Recall ad F-measure of C4.5, REP and IOMCS on the imbalance twitter data sets. Figure 1(b) – (d) presents the results IOMCS verses C4.5, and REP in terms of accuracy, FP Rate and FN Rate which have achieved substantial improvement.

VI. CONCLUSION

In this paper, we proposed an Improved Opinion Mining using Combination Sampling (IOMCS) for decision trees. The proposed algorithm mimics human learning approach. We posited that by applying human learning in machine spaces will lead to an improved performance due to dynamic planning. To test this hypothesis we ran experiments on imbalance twitter corpus. We then compared this method with traditional benchmark algorithms. From these results it is apparent that our proposed IOMCS approach is a competitive one and improved the evaluation measures.

VII. REFERENCES

[1]. Juanli Hu, Jiabin Deng, Mingxiang Sui, A New Approach for Decision Tree Based on Principal Component Analysis, Proceedings of Conference on Computational Intelligence and Software Engineering, page no:1-4, 2009.

[2]. Shane Bergsma, Large-Scale Semi-Supervised Learning for Natural Language Processing, PhD Thesis, University of Alberta, 2010.

[3]. J. Durkin. Expert Systems: Design and Development, Prentice Hall, Englewood Clis, NJ, 1994.

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

[5]. Alexander Pak, Patrick Paroubek," Twitter as a Corpus for Sentiment Analysis and Opinion Mining",

[6]. Syed Akib Anwar Hridoy, M. Tahmid Ekram, Mohammad Samiul Islam, Faysal Ahmed and Rashedur M. Rahman,"Localized twitter opinion mining using sentiment analysis", Anwar Hridoy *et al. Decis. Anal. (2015) 2:8*, DOI 10.1186/s40165-015-0016-4

[7]. Manogna Medurua, Antara Mahimkarb, Krishna Subramanianc, Puja Y. Padiyad, Prathmesh N. Gunjgur," Opinion Mining Using Twitter Feeds for Political Analysis", *International Journal of Computer (IJC) (2017) Volume 25, No 1, pp 116-123*

[8]. Amir Karami, London S. Bennett, Xiaoyun He," Mining Public Opinion about Economic Issues: Twitter and the U.S. Presidential Election",

[9]. Pooja Khanna, Sachin Kumar, Sumita Mishra, Anant Sinha," SENTIMENT ANALYSIS: AN APPROACH TO OPINION MINING FROM TWITTER DATA USING R", International Journal of Advanced Research in Computer Science, 8 (8), Sept–Oct 2017,252-256.

[10]. Vaibhavi N Patodkar, Sheikh I.R.,"Twitter as a Corpus for Sentiment Analysis and Opinion Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 12, December 2016.

[11]. Myneni Madhu Bala, M. Srinivasa Rao, M Ramesh Babu," SENTIMENT TRENDS ON NATURAL DISASTERS USING LOCATION BASED TWITTER OPINION MINING, International Journal of Civil Engineering and Technology (IJCIET) Volume 8, Issue 8, August 2017, pp. 09–19.

[12]. M.V.Sangameswar, Dr. M.Nagabhushana Rao, N.S.Murthy," Twitter Data Analysis on Natural Disaster Management System", International Journal of Engineering Trends and Technology (IJETT) – Volume-45 Number8 -March 2017

[13]. Ali Hasan, Sana Moin, Ahmad Karim and Shahaboddin Shamshirband," Machine Learning-Based Sentiment Analysis for Twitter Accounts", Math. Comput. Appl. 2018, 23, 11; doi:10.3390/mca23010011

[14]. Pierre FICAMOS, Yan LIU," A Topic based Approach for Sentiment Analysis on Twitter Data", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 12, 2016

[15]. Naw Naw and Aye Chan Mon," Social media data analysis in sentiment level by using support vector machine", Journal of Pharmacognosy and Phytochemistry 2018; SP1: 609-613.

[16]. Muqtar Unnisa, Ayesha Ameen, Syed Raziuddin," Opinion Mining on Twitter Data using Unsupervised Learning Technique", International Journal of Computer Applications (0975 – 8887) Volume 148 – No.12, August 2016.