Fuzzy Improved Decision Tree Approach for Outlier Detection in SMS

Priyanka Maan
Department of CSE
ITM University, Gurgaon
Haryana, India

Meghna Sharma
Department of CSE & IT
ITM University, Gurgaon
Haryana, India

ABSTRACT
Spam is one of the serious problems faced by internet community globally. Spam Detection is a critical issue in business world. In this paper an intelligent three stage model is presented to perform the spam inclusive outlier identification. The SMS textual dataset is taken as input and than its filtration is done. After that this textual information is converted to the statistical information using fuzzy and assign the weights to dataset. The decision tree algorithm is than applied on this fuzzy weighed dataset to classify the dataset. This algorithm is defined to separate the spam and non spam data values. A comparison of existing Bayesian and proposed Fuzzy based decision tree approach is done. The results shows that the recognition rate is improved using the proposed approach. The work is implemented in weka integrated java environment.

Keywords
Spam Detection, Outlier detection, Data mining, fuzzy logic, decision tree (DT), weka

1. INTRODUCTION
Data mining itself considered as an intelligent processing over the dataset to identify the relationship between data values, pattern or the identification of the trends. Outlier or Anomaly detection has become an important task in pattern recognition for a variety of applications like network intrusion detection, fraud detection, medical monitoring or manufacturing [1]. Outliers are data records which are very different from the normal data but occur only rarely. If they are completely missing within the training data the problem is also called one-class classification or novelty detection.

For solving problem of outlier detection many different approaches have been proposed: statistical methods, model and distance based approaches as well as profiling methods. Various methodologies and categories surveys are given in [2], [3] and [5].

Various data mining approaches are there which can be used to mine the cybercrime like DT, Bayesian classification, SVM, neural network [6]. Many approaches used by researchers for spam detection includes decision trees modified approaches, naive bayes classifier, neural network etc [10], [11], [12] and [13]. But here fuzzy improved decision tree approach is used for detecting spam messages in sms.

SMS spam detection is one of the most critical business problems. The work is defined to analyze the SMS messages under the attribute and value analysis to perform the fraud detection and identification.

The presented research work is defined using hybrid model with combination of decision tree and fuzzy logic. The decision is here applied as the main classifier to perform the SMS classification and fuzzy logic is applied to assign the weightage to the data values.

2. LITERATURE SURVEY
Many surveys are conducted in past for spam detection. One of them for the detection of Web Spam categorize algorithms into three categories based on the type of information they use: content-based methods, link-based methods, and methods based on non-traditional data such as user behaviour, clicks, HTTP sessions[4]. Many research focuses on study of algorithms for spam detection under content based filtering.[7], [8] and [9].

Chung I Chang[14] has defined a work on integrated sequence mining for fuzzy timer interval specification. Author defined a pattern for discovery to the system and frequent sequential pattern generation over the database. Author defined a formation for successive pattern generation. Author defined the algorithm for pattern mining under fuzzy interval specification. Author defined a work on integrated form for sequential pattern generation and specification.

Koosha Golmohammadi [1] has defined a work to detect the frauds over the security markets. The paper presents an overview of fraud detection in securities market as well as a comprehensive literature review of data mining methods that are used to address the issue. Author identifies the best practices that are based on data mining methods for detecting known fraudulent patterns and discovering new predatory strategies.

N N R Ranga Suri [16] has presented a work on data ranking so that the effective mining and outlier identification over the dataset. Author proposes a two-phase algorithm for detecting outliers in categorical data based on a novel definition of outliers. In the first phase, this algorithm explores a clustering of the given data, followed by the ranking phase for determining the set of most likely outliers. The proposed algorithm is expected to perform better as it can identify different types of outliers, employing two independent ranking schemes based on the attribute value frequencies and the inherent clustering structure in the given data.

Amir A. Sheibani [17] has defined an opinion mining approach for message spam identification. This kind of spam is called product review spam. This papers aims to represent a literature review of what have done on review spam recognition.

Ms.K.Mouthami [18] has defined a work on sentiment analysis to perform message classification of message reviews. Author defined a work on sentiment analysis applied in social media. In proposed work, a new algorithm called Sentiment Fuzzy Classification algorithm with parts of speech

tags is used to improve the classification accuracy on the benchmark dataset of Movies reviews dataset.

Farkhund Iqbal [19] defined a work to identify the criminal network based on chat log analysis. Author propose a unified framework using data mining and natural language processing techniques to analyze online messages for the purpose of crime investigation. Presented framework takes the chat log from a confiscated computer as input, extracts the social networks from the log, summarizes chat conversations into topics, identifies the information relevant to crime investigation, and visualizes the knowledge for an investigator.

3. PROPOSED APPROACH

Web Communication is performed in the form of chats, emails or some other message communication system. SMS communication is one of such kind of common communication performed through mobile phones. But as the messages are transmitted in outer environment. In real world, all kind of users are involved with their specific activities. These activities also include the illegal activities such as spamming, cyber threatening, blackmailing etc. These kinds of SMS data are required to process to identify the spam data or the outliers. In this present work, an intelligent three stage model is presented to perform the spam inclusive outlier identification.

The work has accepted the SMS textual dataset as input. In first stage of work, the textual information processing is done to reduce the information size and to perform filtration. This filtration stage includes the removal of non-keywords or stop list words from the list and to classify the positive and negative sense words. The negative sense words are the words that are common part of spams. After this stage, the filtered message set is obtained. In second stage, this textual information is converted to the statistical information using fuzzy logic. This information processing is obtained by applying the weights to different kind of keywords and by performing the keyword frequency analysis. As the statistical information set is obtained, the decision tree algorithm is applied on this fuzzy weighed dataset. The decision tree is here defined to classify the dataset. This algorithm is defined to separate the spam and non spam data values. The work is implemented in weka integrated java environment.

In this work, a java inclusive weka work is defined for data classification. The weka tools are here considered to fetch the data and for classification algorithms.

4. RESEARCH DESIGN

In this work, an effective SMS information processing approach is defined to manage the dataset and to perform the message criticality analysis.

As the model begins, the input to the system is passed in the form of textual dataset. This textual dataset is processed by a series of textual mining operations. These mining operations include the removal of stop list words, identification of positive words, negative words. The negative word dataset is generated manually by analyzing the spam message contents that often found in mail or messages.

After this stage, the filtered dataset is obtained along with associated positive and negative word set. In second stage of this model, the textual information set is converted to statistical form. The statistical information conversion is based on the keyword weight analysis. These weights are then passed under the fuzzy rules and transformed to fuzzy improved data values. These fuzzy weighted values are then passed to the decision tree algorithm to perform the message classification. The classes obtained here include normal messages, spam and ham messages.

Spam messages are here considered as the outlier to the message set that are required to ignore to process under the system. The model of this work is given here.

This work is also defined as the comparative model with an existing processing model.

According to this existing processing model the textual training and testing sets are analyzed under mining operation and generated the statistical information set for both the training and testing dataset.

Later on the Bayesian network is applied to perform the classification on this dataset. This probabilistic classification approach is defined to perform the recognition.

The flowchart below shows the proposed work for spam detection.

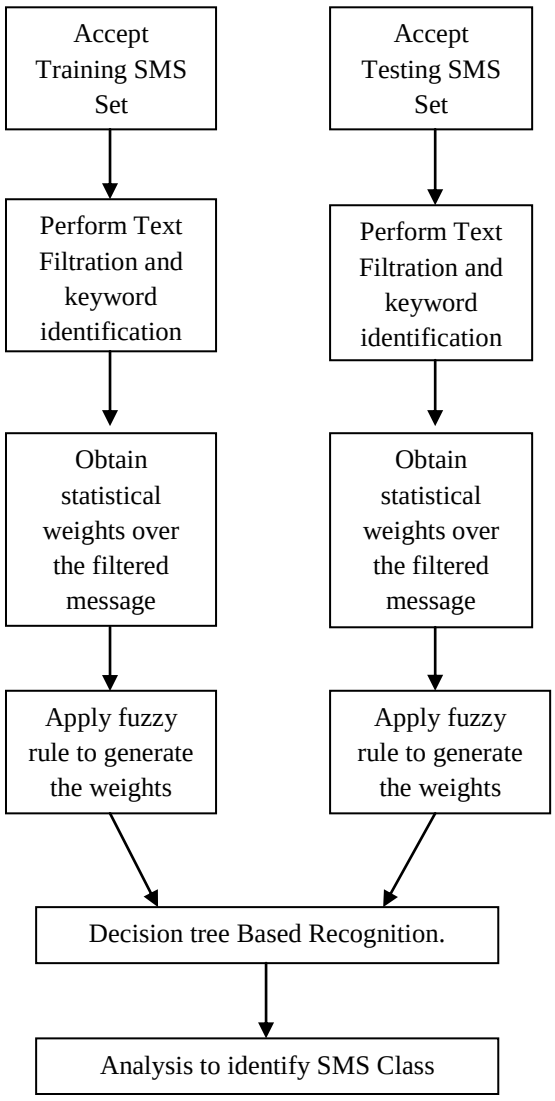

Figure 1: Flowchart of Proposed Work

5. RESULT ANALYSIS

To present the work effectiveness, the work is here tested under multiple test datasets. The analysis is here done in terms of recognition rate obtained for the effective recognition of message class. This section also covered the comparative results obtained from the Bayesian probabilistic algorithm.

The results are calculated for 2 size of test data set: 500 and 1000. Results for 500 records size test data are shown.

Test Set I (500 instances)

Table 1: Test Set Properties

Properties	Values
Train Set Size	5505
Test Set Size	500
Number of Features Collected	7
Number of Class	3

Here table 1 is showing the properties of test set been analyzed for the recognition process. The results of the classification process are given for the existing and proposed work for complete dataset are given here under.

Table 2: Analysis for Recognition on Complete Dataset

Dataset Properties	Existing	Proposed
Number of Instance	500	500
Correctly Recognized	430	475
Incorrectly Recognized	70	25
Recognition Rate	86%	95%
True Positive	93.86%	96.11%
False Negative	6.14%	3.89%

The table 2 shows that the result obtained in case of proposed work are more effectively identified as compared to the existing approach.

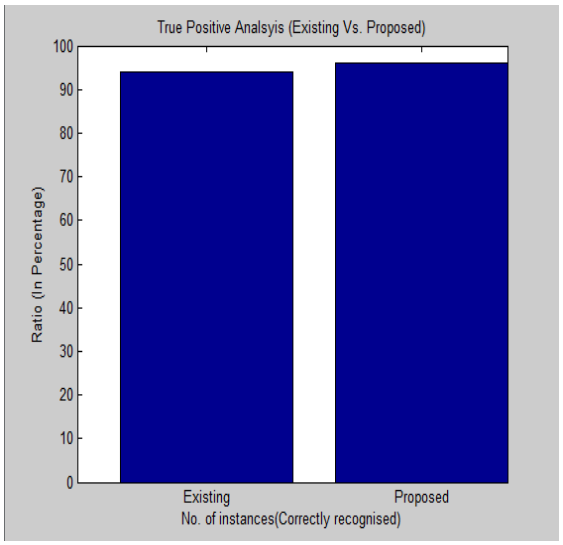

Figure 2 : True Positive Analysis

Here figure 2 is showing the true positive analysis on existing and proposed approach. Here figure shows the true positive rate is here improved.

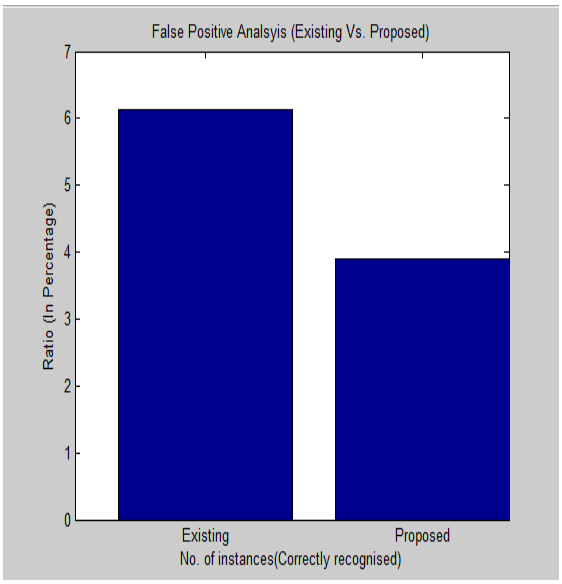

Figure 3: False Positive Analysis

Here figure 3 is showing the false negative analysis on existing and proposed approach. Here figure shows the false negative rate is here improved.

Table 3: Analysis for Recognition on Normal Messages

Dataset Properties	Existing	Proposed
Number of Instance	500	500
Normal Messages	427	427
Correctly Recognized	382	420
Incorrectly Recognized	45	7
Recognition Rate	89.46%	98.36%

Here table 3 is showing the recognition rate obtained for normal messages. The result shows that the presented work is effective to provide the effective recognition on normal messages.

Table 4: Analysis for Recognition on Spam Messages

Dataset Properties	Existing	Proposed
Number of Instance	500	500
Spam Messages	65	65
Correctly Recognized	48	53
Incorrectly Recognized	17	12
Recognition Rate	73.84%	81.54%

Here table 4 is showing the recognition rate obtained for spam messages. The result shows that the presented work is effective to provide the effective recognition on normal messages.

Table 5: Analysis for Recognition on Work Messages

Dataset Properties	Existing	Proposed
Number of Instance	500	500
Work Messages	8	8
Correctly Recognized	0	2
Incorrectly Recognized	8	6
Recognition Rate	0%	25%

Here table 5 is showing the recognition rate obtained for work messages. The result shows that the presented work is effective to provide the effective recognition on normal messages.

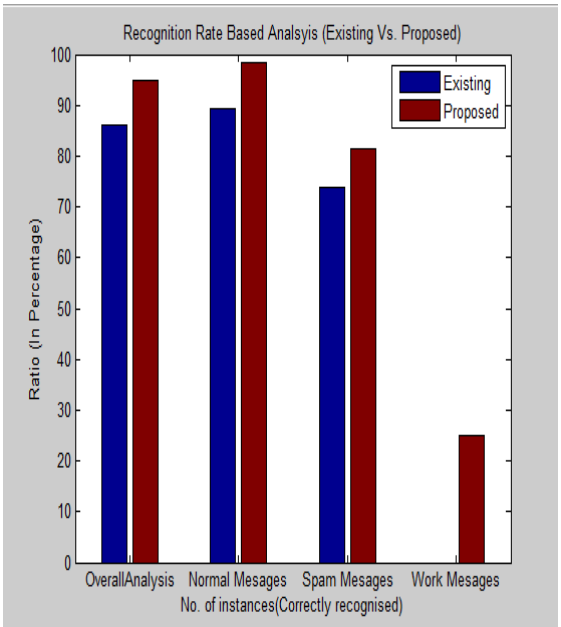

Figure 4: Recognition Rate Analysis (Existing Vs. Proposed')

Here figure 4 is showing the recognition rate based analysis obtained for the defined test set. The results show that for all kind of messages, the recognition obtained for the proposed work is better.

6. CONCLUSION

One of the complex problems for identifying the outlier over the SMS. In this work an intelligent text mining effective statistical information mining approach defined for spam message identification. The work is defined as a three stage model. In first stage of this model, the SMS message processing and filtration is performed. In second stage, the information processing is done using fuzzy based weighted approach. In final stage, the decision tree is applied for data classification. The work has divided the messages in three classes called ham messages, spam messages and normal messages. The results show that the work has improved the recognition rate.

The presented work is about to perform the detection of SMS outlier using a hybrid based on. The work can be extended in future under different dimensions.

I. The presented work is applied for spam detection; same work can be applied on some other datasets such as fraud detection etc.

II. The work is here applied using fuzzy inclusive decision tree approach in future the enhancement can be done by including some optimization approach with it.

7. REFERENCES

[1] Koosha Golmohammadi, Osmar R. Zaiane, Data Mining Applications for Fraud Detection in Securities Market,2012 European Intelligence and Security Informatics Conference 978-0-7695-4782-4/12 © 2012 IEEE.

[2] V. Hodge and J. Austin. A survey of outlier detection methodologies. Artif. Intell. Rev., 22(2), 2004.

[3] M. Markou and S. Singh. Novelty detection: A review-part 1: Statistical approaches. Signal Processing, 83(12), 2003.

[4] Nikita Spirin, Jiawei Han, Survey on Web Spam Detection: Principles and Algorithms, SIGKDD Explorations Volume 13, Issue 2.

[5] M. Markou and S. Singh. Novelty detection: A review - part 2: Neural network based approaches. Signal Processing, 83(12), 2003.

[6] Priyanka Maan, Meghna Sharma, A Study on Mining Approach under Cyber Crime Analysis , IJETCAS 15-183; 2015.

[7] Sahil Puri, Dishant Gosain, Mehak Ahuja, Ishita Kathuria, Nishtha Jatana, Comparison and Analysis of Spam Detection Algorithms, IJAIEM, Volume 2, Issue 4, April 2013.

[8] N.S. Kumar, D.P. Rana, R.G.Mehta, Detecting E-mail Spam Using Spam Word Associations, IJETAE Volume 2, Issue 4, April 2012.

[9] R.Malarvizhi, K.Saraswathi, Content-Based Spam Filtering and Detection Algorithms- An Efficient Analysis & Comparison, IJETT–Volume 4 Issue 9- Sep 2013.

[10] Zhen Yang, Xiangfei Nie, Weiran Xu, and Jun Guo, An Approach to Spam Detection by Naive Bayes Ensemble Based on Decision Induction, ISDA'06, 2006.

[11] Matthias Reif, Markus Goldstein, Armin Stahl, Anomaly Detection by Combining Decision Trees and Parametric Densities, 2008 IEEE.

[12] Kelton Costa; Patricia Ribeiro; Atair Camargo; Victor Rossi; Henrique Martins; Miguel Neves; Ricardo Fabris, Comparison of the Techniques Decision Tree and MLP for Data Mining in Spams Detection To Computer Networks, 2013 IEEE.

[13] Semih Ergin, Sahin Isik, The Investigation on the Effect of Feature Vector Dimension for Spam Email Detection with a New Framework.

[14] CHUNG-I CHANG," An Integrated Sequential Patterns Mining with Fuzzy Time-Intervals", 2012 International Conference on Systems and Informatics (ICSAI 2012) 978-1-4673-0199-2/12 ©2012 IEEE

[15] Mohammad Zaid Pasha, Nitin Umesh, A Comparative Study on Outlier Detection Techniques, International Journal of Computer Applications, Volume 66– No.24, March 2013.

[16] N N R Ranga Suri,"An Algorithm for Mining Outliers in Categorical Data through Ranking", 978-1-4673-5116-4/12@2012 IEEE

[17] Amir A. Sheibani," Opinion Mining and Opinion Spam", 6'th International Symposium on Telecommunications (IST'2012) 978-1-4673-2073-3/12©2012 IEEE

[18] Ms.K.Mouthami," Sentiment Analysis and Classification Based On Textual Reviews".

[19] Farkhund Iqbal," Mining Criminal Networks from Chat Log", 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology 978-0-7695-4880-7/12 © 2012 IEEE.