# **Data Categorization using various Classification Methods**

**Keyur V. Boghani<sup>1</sup> Disha D. Sanghani<sup>2</sup>** <sup>1</sup>M.E. Student <sup>2</sup>Assistant Professor

<sup>1,2</sup>Department of Information Technology

<sup>1,2</sup>Shantilal Shah Engineering College, Gujarat, India

Abstract— The motive of this paper is to compare different classification methods on various datasets. Three different datasets having diverse data category are used for categorization. These sample data are provided by University of California, Irvine (UCI). Five classifiers i.e. NaiveBayes. NaiveBayesMultinomialText, KStar, IterativeClassifierOptimizer. DecisionTable and RandomTree are compared on three datasets i.e. Labour, Soybean and Weather. Comparison is carried out based on classification accuracy of each classification method for every dataset. Practical simulation is performed using Weka tool for Data Mining.

Keywords- Classification, NaiveBayes, Soybean, Data Mining, KStar, DecisionTable, Weka, RandomTree, Categorization

## I. INTRODUCTION

The technique of extracting required information from vast amount of data is called "data mining". Various technological solutions are now available to meet the challenges of data aggregation, data management, information integration and accessibility. Data mining is computing process of identifying patterns from large datasets involving methods at the combination of machine learning, statistics and database systems. It is an important process where classification methods are applied to extract useful patterns. It is the interdisciplinary subfield of computer science and information technology. Overall goal of the data mining process is to gain information from a dataset and convert it into an understandable format for further use. Along with the raw analysis step, it also includes database and data management process, data preprocessing, model and inference analysis, interestingness metrics, complexity calculations, post-processing of found structures, visualization as well as online updating. Hence, Data mining may be considered as an analysis step of the "Knowledge Discovery in Databases" process which is popularly known as KDD.

## II. METHODOLOGY FOR CLASSIFICATION

Weka tool is used for classification of datasets. Three distinct data sets are classified using five classification algorithms i.e. NaiveBayes, NaiveBayesMultinomialText, KStar, IterativeClassifierOptimizer, DecisionTable and RandomTree. The full form of WEKA is Waikato Environment for Knowledge Learning. Weka is a computer program which was created by a student from University of Waikato in New Zealand for the motive of recognizing information from raw data accumulated from agricultural sector. Standard data mining tasks like data preprocessing, clustering, classification, regression, association, feature selection etc. are supported by Weka. It is an open source application that is freely available for all.

In Weka, datasets should be formatted in ARFF format. The Weka Explorer would use these data sets automatically. Classify tab in Weka Explorer is used for data classification. Several different classifiers like bayes, functions, rules, trees etc. are by default provided in Weka tool.

#### A. Steps to Perform Classification on Data Set in Weka Tool:

- 1) Step 1: Open input dataset from local directory.
- Step 2: Apply classification algorithm on entire dataset. 2)
- 3) Step 3: Note down result of classification accuracy provided as Correctly Classified Instances in percentage from Classifier output window.
- 4) Step 4: Continue performing from Step 2 for different classification methods for selected dataset.
- 5) Step 5: Open another dataset file available in ARFF format and repeat from Step 2 again.
- Step 6: Compare and analyze various classification 6) algorithms based on classification accuracy provided for different datasets and discover the efficient classification algorithm for particular data set.

Practical simulation is carried out on a system with configuration of Intel Pentium Processor P6100, 3 GB DDR3 Memory and 500 GB HDD. Experimental analysis is performed three times and average classification accuracy is calculated and considered for comparison.

## **III.** LITERATURE REVIEW

#### A. Nadir Omer Fadl Elssied, Othman Ibrahim & Waheeb Abu-Ulbeh

Objective of this paper is to improve classification accuracy and classification time for spam e-mail classification. Computational time of SVM classifiers is decreased by reducing the number of support vectors. Authors have proposed the K-means SVM (KSVM) algorithm based on hybrid of SVM and K-means clustering. The number of clusters is also a significant input parameter here.

## 1) Experimental Results:

KSVM significantly outperforms SVM and many other spam detection methods in terms of classification accuracy (effectiveness) and time consuming (efficiency).

#### B. Mr. C. Balakumar & Dr. D. Ganeshkumar

Six decision tree algorithms which are basically used as classifiers namely J48 or C4.5, Rndtree, BFtree, REPtree, LMT and simple CART are compared. Test results are shown in WEKA tool. The goal of this research work is to create a decision tree model and train the model so that it can predict the value of a target variable based on several input variables.

## 1) Experimental results:

Among all the decision tree classifiers compared in this paper, the execution time, accuracy and low false positive

rate has been satisfactory only in Rndtree classifier. The accuracy of RndTree is 99%.

## C. Deepak Kanojia & Mahak Motwani

In this research paper, authors have compared kNN classifier and Naïve Basian classification method for a large amount of database in character datatype.

1) Method:

Comparison between Naïve Basian and kNN classifier is carried out based on subsets of features. Class label contains four categories of character data type classified as life and medical transcripts, arts and humanities transcripts, social science transcripts & physical science transcripts. kNN classifier outperformed Naïve Basian classifier when number of samples were small. But as the size of dataset increases, Naive Basian classification method provides satisfactory classification results compared to kNN classifier.

Results clearly show that overall performance of Naïve Basian classifier is better than that of kNN classification method in terms of classification accuracy.

## D. Savita Pundalik Teli & Santoshkumar Biradar

Here, an algorithm for email classification based on Naïve Bayesian theorem is proposed by the authors. Moreover, the mails are classified on the bases of email body.

## 1) Proposed Method:

Authors have introduced three stages for spam e-mail separations. First stage initiates with creation of classification rules. Classifier training is carried out in the second phase. And the last stage performs the decisive email classification.

# E. R. Kishore Kumar, G. Poonkuzhali & P. Sudhakar

In this research paper, spam dataset is analyzed by the authors using TANAGRA data mining tool to explore the efficient classifier for email spam classification.

## 1) Method:

Feature construction and feature selection is carried out initially to extract the relevant features. Then, various classification algorithms are applied over e-mail dataset and cross validation is performed for each of these classifiers. The best classifier is discovered based on the error rate, precision and recall.

The Rnd tree classification algorithm applied on relevant features after fisher filtering has produced more than 99% accuracy in spam detection.

# IV. EXPERIMENTAL RESULTS

Comparison of six different classification algorithms for distinct datasets is carried out in Weka tool based on classification accuracy. Classification accuracy is calculated as number of instances correctly classified among hundred test samples. Observations of classification accuracy for each classifier are shown in following table. We can clearly observe that NaiveBayes has performed well with Soybean dataset, KStar classifier outperformed other algorithms for Labour dataset and Weather dataset is most efficiently classified using RandomTree classification method.

0				
Classification Method	Labo	Soybe	Weath	Avg
	ur	an	er	

NaiveBayes	89.47	92.97	64.28	82.2 4
NaiveBayesMultinomial Text	64.91	13.47	64.28	47.5 5
Kstar	89.47	87.99	35.71	71.0 6
IterativeClassifierOptim izer	87.72	93.85	57.14	79.5 7
DecisionTable	75.44	84.33	57.14	72.3 0
Average	78.95	84.04	78.57	80.5 2



## V. CONCLUSION

We have compared performances of different classification methods on various datasets. Three benchmark datasets are used for comparison in our analysis. It was observed that the performance of a classification technique varies on distinct datasets. Several factors like type of dataset, types of attributes, system configuration, number of tuples and attributes etc. affect the performance of the classifier method. IterativeClassifierOptimizer outperformed other classification methods for Soybean dataset. Overall, NaivesBayes classifier has provided satisfactory classification results with all datasets.

In future, this research work can be carried out further by focusing on enhancement of classification accuracy by choosing appropriate classification method for particular types of datasets. A fusion of various classification algorithms may also be performed for additional improvement of classification.

## References

- Nadir Omer Fadl Elssied, Othman Ibrahim and Waheeb Abu-Ulbeh: An Improved of Spam E-Mail Classification Mechanism using K-Means Clustering. In: Journal of Theoretical and Applied Information Technology, VOL. 60, and NO. 3, February 2014.
- [2] Mr. C. Balakumar and Dr. D. Ganeshkumar: A Data Mining Approach on Various Classifiers in Email Spam Filtering. In: International Journal for Research in Applied Science & Engineering Technology, VOL. 3, NO. 1, May 2015.
- [3] Deepak Kanojia & Mahak Motwani, "Comparison of Naive Basian and K-NN Classifier" International Journal of Computer Applications, Volume 65, No. 23, March 2013
- [4] Savita Pundalik Teli and Santoshkumar Biradar: Effective Email Classification for Spam and Non-Spam. In: International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, No. 6, June 2014.
- [5] R. Kishore Kumar, G. Poonkuzhali and P. Sudhakar: Comparative Study on Email Spam Classifier using Data Mining Techniques. In: International MultiConference of Engineers & Computer Scientists, Vol. 1, and March 2012.