

Text Classification based on Association Rule Mining Technique

Meenakshi Mishra
Gyan Ganga Institute
Of Technology and Sciences,
Jabalpur, India.

Santosh K. Vishwakarma
Gyan Ganga Institute
Of Technology and Sciences,
Jabalpur, India.

ABSTRACT

In this paper association rule mining (ARM) and classification are explained and performed. The limitation of Classification without ARM is also discussed. The Paper also considers the use of association rule mining in classification approach in which a comparative study of Naïve Bayes Classifier and KNN is performed for this purpose. Finally, a comprehensive experimental study against FIRE data set is presented to evaluate and compare traditional and association rule based classification techniques with regards to classification performance.

Keywords

Data Mining, Text Classification, Association Rule Mining, Classification with ARM, Naïve Bayes, KNN.

1. INTRODUCTION

Data mining is the process of extracting the knowledge from large set of data. Data mining is often defined as finding hidden information in a database. Classification is possibly the most common and most accepted data mining technique. Example of classification application include image and pattern appreciation, medical diagnosis, loan agreement, detecting faults in industry application, and classifying financial market trends. Estimation and prediction may be viewed as types of classification. Text Classification is the mission of assigning predefined categories to text documents. It can provide theoretical views of document collections and has important applications in real world.

Association rule defined association between two items. Association rules are commonly used by retail stores to aid in marketing, advertising, floor placement, and inventory control. Association rule mining is to find out association rules that convince the predefined minimum support and confidence from a given database. The difficulty is usually decomposed into two sub problems. One is to discover those item sets whose occurrences go beyond a predefined entry in the database; those item sets are called frequent or large item sets with the constraints of minimal confidence and second one is to create association rules from those large item sets with the constraint of minimal confidence. To improve the performance of classification we perform text classification with association rule mining that gives more efficient result than traditional classification.

2. LITERATURE REVIEW

Al Deen et al. [2] published a paper on title “Classification based on association rule mining techniques”. In this paper classification and association rule mining algorithm are discussed and established. Particularly, the problem of association rule mining, and the investigation and comparison of popular association rules algorithm. The classic problem of classification in data mining will be also discussed. This submission also considers the use of association rule mining

in classification approach in which a recently proposed algorithm is demonstrated for this purpose. Finally a complete experimental study in opposition to 13 UCI data sets is presented to evaluate and compare traditional and association rule based classification techniques with regards to classification precision, number of derived rules, rules features and processing time.

Rahman et al. [3], Published a paper on title “Text Classification using the Concept of Association Rule of Data Mining”. In this document, a procedure of Classifying text using the Concept of Association rule of data mining is discussed. Association rule mining technique has been used to develop feature set from pre-classified text documents. Naïve Bayes classifier is then used on derived features for final classification.

Karthik et al. [4], published an approach for text classification is proposed using association rule mining (ARM) with critical relative support (CRS) based pruning. CRS is probable to reduce the size of association rule base and improve classification performance without compromising accuracy.

Shuanghui Luo [5] published a paper on title “Distributed Classification Based on Association rules (DCBA) Algorithm”. This Paper will focus on the combination of association rule mining and classification rule mining. Data mining has been used in market data analysis, catalog design, web log sequence, DNA analysis, high throughput drug design etc. This research is to combine the existing data mining algorithms and distributed techniques to develop a distributed CBA algorithm and distributed CMAR algorithm and apply them to mine very large and distributed databases.

Bing et al. [6] published a Paper on title “Integrating Classification and Association Rule Mining”. In this paper, they propose to integrate these two mining techniques. The integration is completed by focusing on mining a special subset of association rules, called class association rules (CARs). An efficient algorithm is also given for building a classifier based on the set of exposed CARs. Tentative results show that the classifier built this way is, in general, more accurate than that produced by the state-of-the-art Classification system C4.5. In addition, this combination helps to solve a number of problems that exist in the current classification systems.

Yang et al. [7] published a paper on title “A Classification algorithm based on Association Rule Mining”. In this paper, a classification algorithm is present that is based on Trie-tree that named CARPT, which remove the frequent items that cannot generate frequent rule directly by adding the count of class labels. And it compress the storage of catalog using the two dimensional array of vertical data format, reduce the number of scanning the database significantly, at the same time, it is convenient to count the support of candidate sets.

So, time and space can be saved effectively. The research results show that the algorithm is feasible and effective.

Irina Tudor [8] published a paper on designation “Association Rule Mining as a Data Mining Technique”. In this effort, an Association rule mining represents a data mining technique and its objective is to find exciting association or correlation relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many companies are becoming concerned in mining association rules from their databases to increase their profits. For example, the discovery of appealing association relationships among huge amounts of business operation records can help catalog design, cross marketing, loss leader examination, and other business decision manufacture processes. A classic example of association rule mining is market basket analysis. This procedure analyzes customer buying habits by judgment associations between the different items that customers place in their “shopping baskets”.

Prathibhamol C. P et al. [9] published a paper on designation “Anomaly Detection based Multi Label Classification using Association Rule Mining (ADMLCAR)”. In this paper an Anomaly Detection based Multi Label Classification using Association Rule Mining (ADMLCAR) is used for solving MLC problem. Traditionally, most of the multi label classification problem is solved by either of the two methods: Problem transformation, Algorithm adaptation. But the method discussed in this paper aims at a novel method different from traditional solution to multi label classification problem. For clustering, ADMLCAR uses k-means algorithm and for association rule mining purpose it uses vertical data format. To predict the test data instances, this method locates the adjacent cluster. Once the clusters are recognized it uses oversampling principal component analysis (PCA) within the nearest cluster with respect to test instances. Oversampling PCA is used to emphasize the need for confirming the fact that test instance’s label set will not only be confined to its nearest cluster label set. This is because, anyways the test instance will be identified to a nearest cluster by earnings of Euclidean distance measure but as clustering is unsupervised the nearest cluster may contain many objects entities of different label sets.

Rajan et al. [10] published a paper on designation “A Method for Classification based on Association Rules using Ontology in Web Data”. This paper shows a new method based on association rule mining and ontology for the classification of web pages. This work is pruning of association rules, generated by mining process. The main complexity arises due to the fact that there are various integers of text documents that are considered for generating the association rules using the A-priori algorithm. But these rules that were generated are not based on the semantic knowledge. In order to obtain the most accurate rules we gone for the construction of the ontology, based on the domain knowledge. With this domain knowledge we design various operators which are helpful in reducing the rules generated. Thus the various rules that we get are semantically correct with regards to the domain selected. We use the high confidence value based classifier for classifying the given text document to that particular domain. Association rules are mined from this matrix using A-priori algorithm. Based on the high confidence value, a new text document is classified into one of the predefined classes. In general, from association rule mining, a huge amount of association rules are mined. All the association rules generated may not be useful for the classification purpose. So, In order to reduce the irrelevant association

rules, it needs semantic knowledge. For this purpose, propose new domain specific ontology to overcome this weakness of association rule mining method.

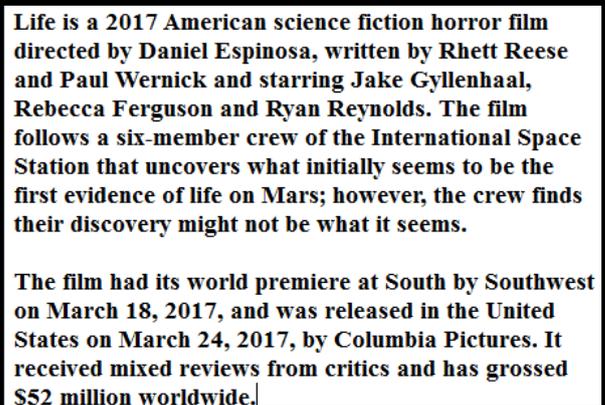
Senthil [11] presents a work on designation “Association Rule Based Classification”. In this work, in this thesis, we focused on the construction of classification models based on association rules. Although association rules have been predominantly used in favor of data exploration and explanation, the interest in using them for prediction has rapidly increased in the data mining community. In order to mine only policy so as to can be used for classification, we modified the well known association rule mining algorithm Apriori to handle user-defined input constraints. We considered constraints that require the presence/absence of particular items, or that limit the number of items, in the antecedents and otherwise the consequents of the rules. It developed a characterization of those item sets with the intention of will potentially form rules that satisfy the given constraints.

Zhonghua et al. [12] published a paper based on “A New Class Based Associative Classification Algorithm”. In this work, applying association rule into classification can improve the truth and obtain some valuable rules and information that cannot be captured by other classification approaches. On the other hand, the rule generation procedure is very time-consuming when encountering large data set. Besides, traditional classifier building is organized in several separate phases which may also degrade the effectiveness of these approaches. In this paper, a new class based associative classification be in motion toward (CACA) is proposed. The class label is taken good advantage of in the rule mining step so as to cut down the searching space. The projected algorithm furthermore coordinates the rule creation and classifier construction phases, shrinking the rule mining space when building the classifier to help speed up the rule construction. Experimental effect suggested that CACA is building better performances in accuracy and efficiency in Associative classification approaches.

3. DATASET

The Dataset used in the present system is related with movies, sports, and technology text data. It is a collection of nine files. In the dataset, there are 3 Class names as:

1. **Movies:** It contains three text files related with movies data.
2. **Sport:** It contain three text files related with sports data.
3. **Technology:** It contains three text files related with Technology data.



Life is a 2017 American science fiction horror film directed by Daniel Espinosa, written by Rhett Reese and Paul Wernick and starring Jake Gyllenhaal, Rebecca Ferguson and Ryan Reynolds. The film follows a six-member crew of the International Space Station that uncovers what initially seems to be the first evidence of life on Mars; however, the crew finds their discovery might not be what it seems.

The film had its world premiere at South by Southwest on March 18, 2017, and was released in the United States on March 24, 2017, by Columbia Pictures. It received mixed reviews from critics and has grossed \$52 million worldwide.

Figure 1: Text Data

4. METHODOLOGY

In the present system, we have enumerated the theoretical analysis of the method that is: Text Classification Based on Association Rule Mining Technique. In this work, we have used Rapid Miner version 7.5 with extension of Text mining. We believe that Rapid miner workflow approach entices systematic research and facilitates its implementation in combination with Rapid Analytics. Methodology follows -

4.1 Identification of research dataset.

4.2 Preprocessing-

Data preprocessing is a data mining technique that involves transforming rare data into an understandable format. Real-world data is often incomplete, incompatible, and or missing in certain behaviors or trends, and is likely to contain many errors. The steps of preprocessing is following-

Tokenization - In lexical analysis, tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other significant basics called tokens. The evidence of tokens becomes input for more processing such as parsing or text mining.

Case transform – This operator transforms all characters in a document to either lower case or upper case, in that order.

Stop word Removing - Stop words are words which are filtered out before or after processing of natural language data. Though stop words usually refer to the most common words in expressions, there is no single universal list of stop words used by all natural language processing tools, and indeed not all tools even use such a list. Some tools explicitly avoid removing these stop words to support phrase search.

Stemming - Stemming is the process of reducing inflected (or sometimes derived) Words to their word stem, base or root form—normally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the identical stem, even if this stem is not in itself a valid root.

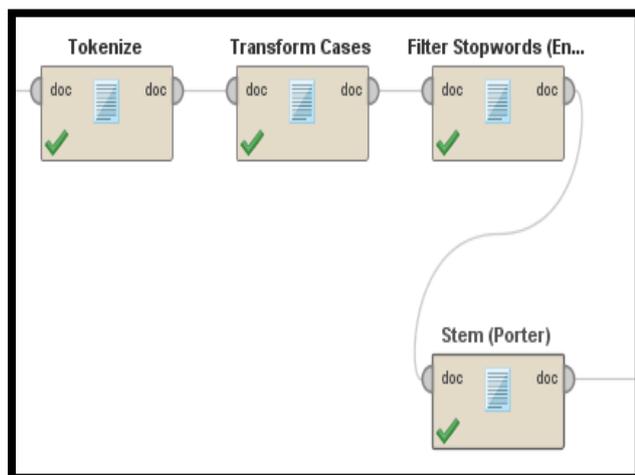


Figure 2: Preprocessing

4.3 Apply Association Rule Mining using FP-growth in Rapid Miner (Tool).

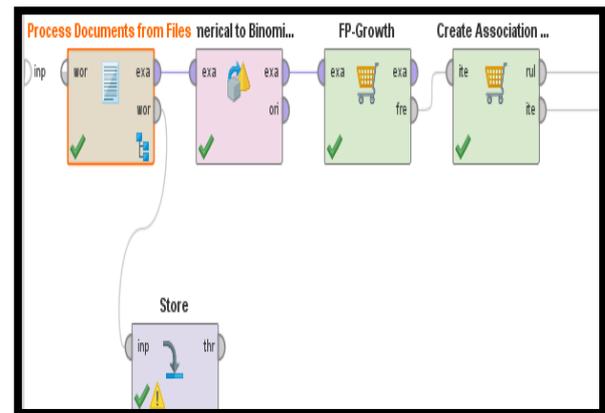


Figure 3: Association Rule Mining

For Association Rule Mining we have use five operators in rapid miner. These are –

Process Documents from files – It is largely used for text Processing. It generates word vectors from a text collection stored in multiple files.

Numerical to Binominal - This operator changes the type of the certain numeric attributes to a binominal type. It also maps all morals of these attributes to equivalent binominal values.

FP-Growth - This operator efficiently calculate all frequent item sets from the given Example Set using the FP-tree data structure. It is requisite that all attributes of the input Example Set should be binominal.

Create Association Rules - This operator generates a set of association rules from the given set of common item sets.

4.4 Intermediate Result

No.	Premises	Conclusion
1	win	game
2	s	game
3	player	game
4	plai	game
5	oppon	game

Figure 4: Association Rules

4.5 Apply Classification using naive Bayes method

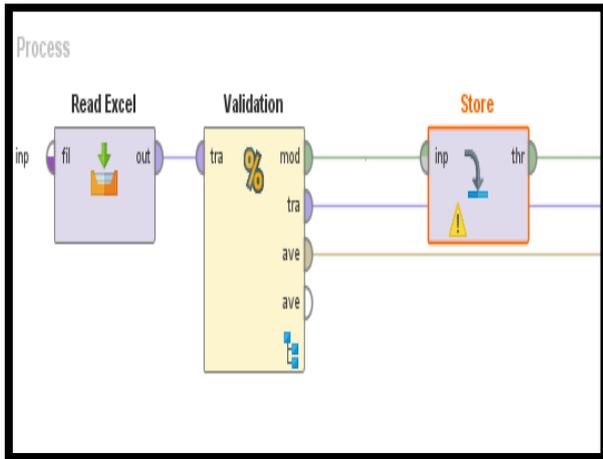


Figure 5: Training

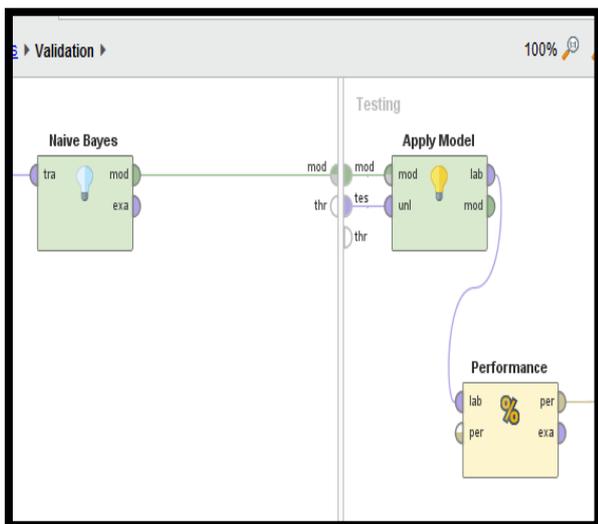


Figure 6: Validation

4.6 Bayesian Classification

Naive Bayes Classifier is a supervised learning technique used for machine learning. Naive Bayes classifiers are a family of simple probabilistic classifiers based on Bayes theorem with strong (naive) *naive* (independence) comes from the assumption that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption. Especially for small sample sizes, naive Bayes classifiers can outperform the more powerful alternatives. Being relatively robust, easy to implement, fast, and accurate, naive Bayes classifiers are used in many different fields. In Naive bayes classifier different models that assign class labels to problem sample, represented as vectors of feature values, where the class labels are drawn from some finite set. Naive Bayes is a collection of algorithms: all naive Bayes classifiers feature of a particular class is different from other feature of the same class. For example, a fruit may be considered to be an apple if it has following features is red, round, and about 10 cm in diameter.

5. RESULT ANALYSIS

Accuracy - Accuracy is an evaluation metrics on how a model perform. It defines Relative number of correctly classified examples or in other words percentage of correct predictions. The accuracy is calculated by taking the percentage of correct predictions over the total number of examples. Accuracy can be finding by:

$$Accuracy = \frac{1}{|I|} \sum_{j=1}^{|I|} \frac{|a_j \cup b_j|}{|a_j \cap b_j|}$$

Kappa - The Kappa statistic is a metric that compares an Observed Accuracy with an expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers along with themselves.

$$Kappa = \frac{(observed\ accuracy - expected\ accuracy)}{(1 - expected\ accuracy)}$$

Correlation - Returns the correlation coefficient between the label and prediction attributes.

$$Correlation(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2]^{1/2}}}$$

Weighted mean Precision - The weighted mean precision is calculated by taking the average of precision of every class.

$$Precision = \frac{1}{|I|} \sum_{j=1}^{|I|} \frac{|a_j \cap b_j|}{|b_j|}$$

5.1 Performance Table

Table 1

Classifier	Naive Bayes	FP Growth + KNN (M)	FP-Growth + Naive Bayes
Accuracy	85.56%	86.21%	93.33%
Kappa	0.854	0.743	0.925
Correlation	0.500	0.955	0.904
Weighted mean Precision	16.67%	56.73%	61.67%

From above performance table, in simple classification we get 85.56% accuracy and classification with ARM provides 93.33% accuracy.

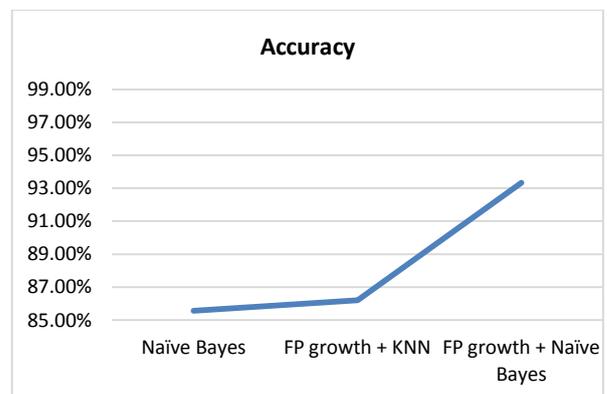


Figure 7: Accuracy

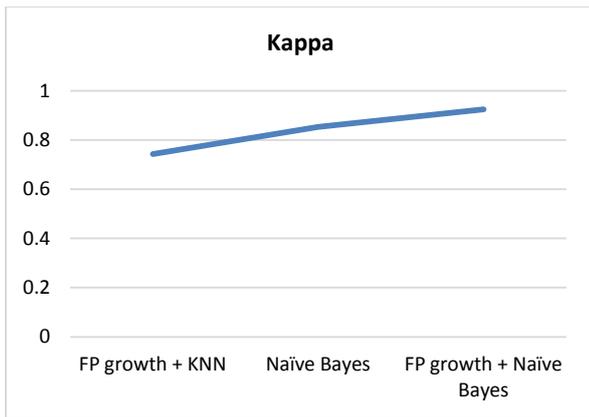


Figure 8: Kappa

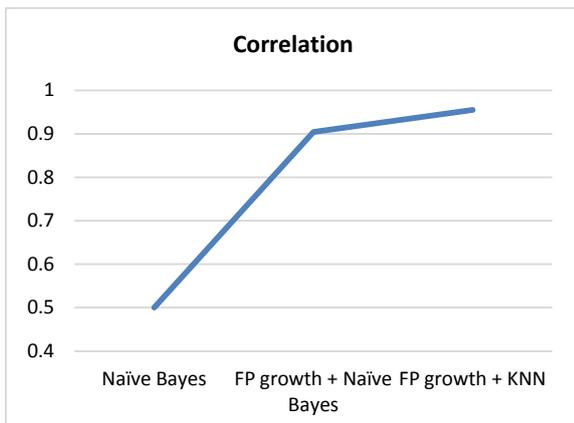


Figure 9: Correlation

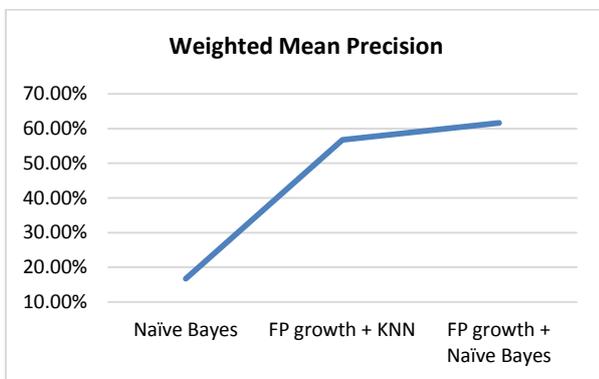


Figure 10: Weighted Mean Precision

6. CONCLUSION

In this work, Association Rule Mining can be efficiently used for text document Classification. This paper proposes a framework to integrate classification and association rule mining. The comparative study of text classification and classification with ARM is effective for improving the accuracy of prediction. In this work classification provide more accuracy than simple classification.

The Classification accuracy in this work is 93.33%. This result shows the accurateness of classification that is done.

Traditional Classification is not yet possible to determine accurate class that is predefined. Classification with ARM is more powerful than other traditional techniques. The Naïve Bayes Classifier is more accurate with ARM for the reason that ARM produce large association rule for particular words that are in text data. The Pre-processing is another benefit for producing efficient number of rules.

7. REFERENCES

- [1] Abdullah S. Ghareb, Abdul Razak Hamdan, Azuraliza Abu Bakar, 2012 "Association Rule Mining as a Data Mining Technique", 4th Conference on data mining and optimization.
- [2] "Alaa Al Deen" Mustafa Nofal and Sulieman Bani-Ahmad, 2010 "Classification Based on Association-rule mining techniques", Ubiquitous Computing and Communication Journal, Vol. 5.
- [3] Bangaru Veera Balaji, Vedula Venkateswara Rao, 2013 "Improved Classification Based Association Rule Mining", International Journal of Advance Research in Computer and Communication Engineering, Vol. 2.
- [4] Bing Liu, Wynne Hsu and Yiming Ma, 1998 "Integrating classification and Association Rule Mining", AAAI.
- [5] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, S M Kamruzzaman 2010 "Text Classification using the Concept of Association Rule Of Data Mining", International conference on information Technology.
- [6] Karthik P, Saurabh M, and U Chandrasekhar, 2016 "Classification of Text Documents Using Association Rule Mining with Critical Relative Support Based Pruning", ICACCI.
- [7] Prathibhamol C. P, Amala G.S and Malavika Kapadia, 2016 "Anomaly Detection based Multilevel Classification using Association Rule Mining (ADMLCAR)", Intl. Conference on Advances in Computing, Communication and Information (ICACCI).
- [8] R.Hubert Rajan, Julia Punitha Malar Dhas, 2012 "A method for Classification based on Association Rules using Ontology in web data", International Journal of computer applications, volume 49.
- [9] Senthil K. Palanisamy, 2006 "Association Rule Based classification".
- [10] Yang Junrui, Xu Lisha, 2012 "A Classification Algorithm Based on Association Rule Mining", International Conference on Computer Science And Service System.
- [11] Yannis Haralambous, Phillipe Lenca, 2014 "Text Classification using Association Rules Dep dependency pruning and Hyperonymization", DMNLP.
- [12] Zhohanghua Tang and quin Liao, 2007 "A new Class based Associative Classification Algorithm", IAENG International journal of Applied Mathematics (IJAM).
- [13] Shuanghui Luo, 2002 "Distributed Classification Based on Association Rules (DCBA) algorithm"