A COMPARATIVE STUDY OF ASSOCIATION RULE MINING TECHNIQUES AND PREDICTIVE MINING APPROACHES FOR ASSOCIATION CLASSIFICATION

Kavita Mittal  
Department of Computer Science and Engineering  
Jagannath University  
Bahadurgarh, Haryana, India.

Dr. Gaurav Aggarwal  
Department of Computer Science and Engineering  
Jagannath University  
Bahadurgarh, Haryana, India.

Dr. Prerna Mahajan  
Department of Computer Science  
Institute of Information Technology and Management  
GGSIPU, Delhi, India.

Abstract: Association Rule Mining (ARM) and classification are integrated together to build competitive classifier models called Associative Classifiers and this approach is known as Association Classification (AC). AC leads to the formation of accurate classifier consisting of significant rules capable of predicting the class of the data. This paper presents the evolution of ARM to AC highlighting the development and improvements in ARM techniques followed by AC techniques. The goal of this paper is to survey and understand different ARM and AC techniques and comparing their performance. In the literature a variety of AC algorithms have been proposed such as CBA, CMAR, MCAR, CPAR etc each adopting some or the other approach for rule learning in the initial stages. This paper also presents the importance of the rule pruning methodology with the brief survey of different methods discussed in the literature. This paper also enlightens the learning approaches adopted by different AC techniques in different domains.

Keywords: Association Rule Mining (ARM), Association Classification (AC), Rule Learning, Rule Pruning, Prediction, Class Assignment.

1. INTRODUCTION

ARM is an emerging research area in data mining that aims at extracting valuable information from huge volume of data and applying it in decision making. Today organizations are drowning in data but starving for knowledge. With the advent of internet, e-data is growing at extremely higher pace making it difficult for the competitors to position them in World Wide Trading [1]. Data mining technology helps in extracting patterns to understand the data, produce knowledge and use it for future predictions. Data mining can be classified into two categories namely descriptive mining and Prescriptive mining [1]. Descriptive mining refers to summarization and characterization of data in the repository. Prescriptive mining generate inferences on existing data to predict the future trends based on the past data. ARM is one of the descriptive mining technique used to generate associations among the items in the transactional or relational dataset.[2]. Use of ARM for building classification models(classifier) result in unique approach called Association Classification (AC). AC was introduced in 1997 for developing relationships between attribute values [3]. Association Classification is the integration of Association rule mining and classification. Classification process maps a group of attributes to a class that can be used to assign the classes of new data objects based on their attribute values. The classifiers produced by AC techniques are considered to be more accurate than traditional classification approaches like decision trees. The AC approaches are found to be successful in real world application from different domains such as academics, medical diagnosis, web filtering etc. A number of AC techniques are used in literature such as ADT, CAEP, CMAR, MMAC, MCAR, L3, CBA etc each use different method of extracting and pruning rules [4]. Building a classifier based on AC involves extraction of classification association rules from training dataset followed by selecting a subset to build the classifier. Subset of rules is obtained by evaluating complete set of class association rules and only considering the rules that cover defined training data records. After building the classifier its prediction strength is tested on the test dataset for predicting the class labels. Thus, AC attempts to explore the relationships between attribute values to assign classes aiming to obtain essential knowledge not taken care of by traditional classification methods thereby improving the classification accuracy. While building classifiers, producing complete set of rules need CPU time and many dataset scans during the training phase. The problem with AC is the generation of huge number of rule making it difficult to understand and manage. Thus to improve CPU usage and minimize the dataset scans, there is need to build classifiers with minimum number of interesting rules. This can be achieved using rule pruning approach. Rule pruning approach when applied to AC produces high quality and scalable classifier making the classifier more manageable. The AC techniques use different pruning methods most commonly used are pessimistic error, database coverage, lazy pruning discussed in literature to minimize the size of resulting classifiers. Many more methods exist each having different characteristics and application depending on the application domain. Now a days due to growing text documents solving the problem of text categorization is becoming a necessity. This work will
include the development of different pruning and prediction methods to be implemented in the association classification and then application of association classifiers for text categorization of both structured and structured data.

This paper is organized in 5 sections with section 1 discussing the introduction, section 2 describing the importance of study, section 3 presenting the methodology for conducting the research study, section 4 briefly discussing the literature review consisting of sub sections with subsection A describing the ARM techniques and their comparison, subsection B introducing Association classification techniques, subsection C discussing rule learning approaches, subsection D describing different rule pruning techniques used in the literature, subsection D briefly introducing Text Categorization using Association classification and lastly section 4 concluding the paper with future work discussions.

3. IMPORTANCE AND RELEVANCE OF THE STUDY

Research Gap

- While building classifiers, producing complete set of rules need CPU time and many dataset scans during the training phase.
- The problem with AC is the generation of huge number of rules making it difficult to understand and manage.
- Thus to improve CPU usage and minimize the dataset scans, there is need to build classifiers with minimum number of interesting rules.
- This can be achieved using rule pruning approach. Rule pruning approach when applied to AC produces high quality and scalable classifier making the classifier more manageable.
- Whether applying high confidence rule for making prediction relatively enhance the classification accuracy.

Aim of the Study

The aim of this study is to achieve the various goals such as to produce an extensive literature review on common association rule mining approaches with specific elaboration on rule pruning threshold criteria and class assignment tasks. These two phases are discussed in detail because of their importance in solving the problems in AC approach generating large number of rules, CPU and memory usage and lastly the over fitting in classification [5]. The study also focuses on the impact of minimizing the number of rules on the effectiveness and efficiency of the classifier through rule pruning technique. This study analyzes the effect of employing high confidence rule strategy for assignment of class label on test dataset based on classification accuracy. The study will proceed by developing an AC model by employing rule pruning and class assignment methods. Then the model will be exploited on particular benchmark datasets and algorithm and will be compared with other classification models.

4. METHODOLOGY

To achieve the research objectives the study will adopt the following methodology:

Figure 1: Association Classification Model

This proposed model will be adapted to work on text based training dataset. Then required thresholds (minimum support and minimum confidence) will be defined. AC system begins with processing of training data by discovering frequent item sets followed by generation of Association Classification rules. The rules are then filtered to obtain significant rules to build the classifier. Finally the classifier will be applied to the testing dataset already in preprocessed format.

5. RELATED WORK

In spite of existence of many text classifiers in literature based on different classification approach, automated text categorization is a vital area of research the needs improvement in terms of accuracy. The implementation of ARM for building ACM is still very less in Text Categorization. This section will produce detailed review on classic ARM elaborating rule pruning based association classification.
A. Association Rule Mining (ARM)
Association Rule mining being an important research branch of data mining aims to find interesting and frequent patterns, discover correlations among set of data in data repositories. The concept of ARM was introduced by Agrawal [3]. The purpose of ARM is to discover hidden relationships among different data item sets in the database. Assuming a given transaction database the ARM problem is to generate association rules considering two predefined thresholds.

Mathematical Model of ARM
Let I=\{I_1, I_2, I_3, \ldots, I_n\} be the set of n different item sets. S is the set of all the transactions in the database, T is a transaction such that T is a subset of I, where each transaction is a collection of item sets and has a unique identifier. The association rule is the instantiation of the form A→B, where A, B ∈ I are the collection of items called item sets and A, B are disjoint. A is called antecedent and B is called consequent. The rule is termed as A implies B, i.e., if |A|=k, then A is called k order set and can be expressed as A[1], A[2], A[k]. The association rule can be evaluated based on two parameters support(s) and confidence(c). The database being huge, the user is concerned about frequently used item sets that can be generated using predefined thresholds of support and confidence called minimum support (minsup) and minimum confidence (minconf).

Support can be measured as statistical significance of the association rule in the database that demonstrates the degree of representation of the rule. The greater the support, the more important is the rule [2]. In other words it is the fraction of transactions that contain both A and B.

\[ S(A \rightarrow B) = \frac{|TA| \cap |TB|}{|D|} \]

Where, \(|TA|\) is number of transactions containing item sets A, \(|TB|\) is number of transactions containing item sets A and \(|D|\) is total number of transactions in Database D. Confidence can be measured as accuracy of association rule. It measures how often items in B appear in transactions that contain A.

\[ C(A \rightarrow B) = \frac{|TA \cap TB|}{|TA|} \]

Where, \(|TA \cap TB|\) is number transactions containing both A and B, \(|TA|\) is number transactions containing item sets A. The confidence c of rule A→B is defined as c% transactions in the database D containing A also contains B and support s is s% transaction in D that contains A∪B. The goal of ARM is to discover rules having support >= minsup threshold and confidence>= minconf threshold such association rules are more strong and effective [6].

Process of ARM
ARM is a two step process consisting of extraction of all frequent item sets followed by extraction of strong association rules from the obtained frequent item sets.

| S.No | Algorithm | Description | Merit |
|------|-----------|-------------|------|
| 1    | AIS [3]   | It attempts to improve the quality of database to process queries and generate association rules by reducing the no of database scans | In reduces database scans during mining |
| 2    | Apriori   | It involves two step process to find large item sets first, then scans the database to check the support count of corresponding item sets. | Highly efficient in generating frequent item sets. |
| 3    | Apriori TiD | It is a hybrid algorithm that uses Apriori initially and later on uses Apriori TiD. It involves the | Capable of producing frequent item sets from large |
4 FP-Tree: It scans the database only twice to generate frequent item sets without any iteration process. The first scan constructs FP tree and the next scan generate frequent item sets from FP tree using a procedure called FP growth. Can produce frequent item sets involving only two database scans.

5 Continuous Association Rule Mining: It computes large item sets online. It involves at most two scans of sequential transactions to generate all large item sets. During first scan the algorithm constructs a lattice of all large item sets and then continuously removes all trivially small item sets having less parameter values than thresholds in the second scan. It requires less database scans in sequential transactional database.

6 Rapid ARM: It uses tree structure to represent original database without generating candidate item sets. It enables second scan of database by generating 1-itemsets and 2-itemsets quickly using Support Oriented Tie Item Set (SOTieIT) structure. It is efficient due to tree structure and minimizes the database scans.

7 Extended ARM: This algorithm is based on interval and ratio variables rather than nominal and ordinal variables. The rules involves quantitative variables composed with related metrics and statistics. It supports mining of rules from numeric and statistical data.

8 Generalized Disjunctive Association Rule (d-rules) Mining: It allows disjunction of various conjuncts to extract contextual interrelationships among data items. It attempts to minimize the harmful impacts as well as maximize possible benefits in the mining process.

9 Genetic Algorithm: It involves mining of positive and negative association rules in database using genetic and fitness operators and functions without taking minsup and minconf into account. It is proved to be efficient mining process not dependent upon support and confidence.

Table 2: Improved ARM Algorithm

| S.No | Algorithm | Description | Merit |
|------|-----------|-------------|-------|
| 1    | ARM with FP Tree using Positive and Negative Integration.[8] | This algorithm involves two stages: a) Rule Generation-calculates set of all positive and negative association rules followed by pruning of contradicting rules and selecting a subset of high quality. b) classification-extracts a subset of rules found in first stage and predict the class label of data object by analyzing the subset rules. It uses hybrid Approach to deal with large size dataset. | It provides more accurate and efficient classification detection of frequent item sets among large databases. |
| 2    | Weighted ARM Algorithm [2] | The weighted ARM algorithm is similar to Apriori in framework but different in functionality. The weighted support of the algorithm may be greater than 1 which contradicts the actual support that should be less than 1. Also Apriori algorithm needs to scan the database frequently. | Involves less database scans repeatedly while producing frequent item sets, thereby improving the efficiency of data mining. |
| 3    | Improved version of Apriori Algorithm [9] | This algorithm is based on four characteristics- 1) Data preparation and chooses the desired data. 2) Produce item sets that decide the rule constraints for knowledge. 3) Mine k-frequent item sets using new database 4) Produce the association rule that set up the knowledge. | It provides superior results for using knowledge base. |
| 4    | Selective Association Rule Generation [10] | This algorithm is based on defining a set of “interesting item sets” and the selectively generate rules of only these item sets. | When applied to a dataset the number of rules found was |
Building a classifier from the discovered rules and finally consists of three steps - firstly to discover rule items, then consists of a training dataset \( T \) having \( m \) distinct attributes \( A_1, A_2, \ldots, A_m \) and a list of class attributes denoted by \( C \).

The associative classification problem involves generation of rules using Apriori followed by in all cases and rule pruning approaches. The most commonly used AC algorithms are [5]:

1. **Classification based on Association rules (CBA)** - It involves generation of rules using Apriori followed by building a classifier.
2. **Classification based on Multiple Association Rules (CMAR)** - It performs classification based on weighted Chi-Square analysis applying multiple strong association rules.
3. **Classification based on Predictive Association Rules (CMAR)** - It implements greedy approach for generating rules from training dataset.
4. **Multiclass Classification based on Association Rules (MCAR)** - It involves two step process. Firstly it generates frequent rules item also called candidate rule.
items that involve more attributes based on minconf and
mindsup, followed by application of the rules on the
training dataset for building a classifier.
Other AC algorithms include ADT, CAEP, NMAC, L3 that
exists in literature.

Applications of Associative Classification

[17] introduced CCSA(Cascading of clustering based on
Schwarz Criteria and association)algorithm to perform
clustering followed by classification based on association
using Apriori association for generation of classification
rules. The algorithm was analyzed on online datasets with
Weka resulting in improved classification accuracy with
reduced number of rules.

[18] attempts to analyze the performance of
CPAR(Classification based on Predictive Association Rules)
, PRM(Predictive Rule Mining), FOIL(First order Inductive
Learner) methods on Tuberculosis dataset where CPAR and
PRM performed well in terms of accuracy, no of rules and
time consumption as compared to FOIL.

[19] tried to use multiple relational Bayesian Classification
with MCAR algorithm depending on Genetic Algorithm(GA) employed for optimization of classification
rate using association rules. The results implies that MCAR
with GA showed higher classification accuracy than
traditional MCAR.

Rule Discovery approaches in Association Classification

The first step in AC to discover item rules in two sub steps
initially to discover frequent items and then generating
rules to form a set of Classification Association Rules
(CARs) for building the classifier. Following approaches are
discovered for rule generation:

Apriori Search

In Apriori algorithm the discovery of frequent itemsets can
be achieved in few iterations involving database scan in
each iteration. The decision of frequent item is based on the
support of candidate rule item. Apriori algorithm is used by
CBA algorithm to generate frequent rule items. This
approach requires repetitive database seems consuming
more computation time at each level. Many techniques are
discussed in the literature that extends the Apriori algorithm
to generate classification rules such as Apriori TiD.

Vertical Mining Approach

This concept uses simple interaction among item IDs to find
the rules there by reducing no of database scans. Few
algorithms such as Eclat, MCAR uses this approach to
reduce the computational time needed to discover CARs.

FOIL. Decision Tree Approach

This strategy produces rules for each class in the training
dataset. The training data is divided into two subsets – one
containing positive cases and other containing negative
cases associated with a class ‘C’. The algorithm CPAR
adopts greedy AC algorithm for generating rules seeking for
the best rule condition.

Frequent Pattern (FP) Growth Approach

FP Growth algorithm makes Apriori more efficient in terms
of CPU time and memory usage. It constructs FP-tree
representing the training dataset. It perform rule items
generation in two steps. Initially it forms FP tree, then
extracting frequent rule items directly from FP tree. The
CMAR is the first AC algorithm to adapt FP growth
learning approach. The memory requirements of FP
becomes large for large datasets making it less preferable.

Association Rule Mining based on Weighted Class

In this approach weighted Association rules are produced
using weighted support and weighted confidence. The rule
items overcoming the thresholds, i.e. weighted support and
weighted confidence append to the frequent weighted rule
items set.

C. Rule Pruning methods in Association Classification

The AC techniques generally suffers from a problem of
deriving large set of rules due to highly correlated datasets
and considering all attributes for generating rules. This may
lead to insignificant and redundant rules. Such rules need to
be eliminated for effective and accurate classification.
Different pruning methods are introduced in literature based
on different concepts such as decision trees, pessimistic
estimation error, Chi-Square testing statistics etc that can be
applied either during rule generation or building the
classifier. Rule Pruning methods add constraints on rule
discovery to reduce the size of classifiers. The various
pruning techniques introduced by AC algorithms include:
Chi-Square Testing, Redundant Rule pruning, Database
Coverage, Pessimistic error estimation, Lazy Pruning,
Conflicting rules, Laplace Accuracy [4]. It was analyzed
that algorithms using lazy pruning produce large number of
rules making them unmanageable for accurate classification.
In comparision Database coverage and pessimistic error
estimation based strategies build moderate sized classifiers.
Other effective pruning methods introduced by [5] include:
High Precedence Classify correctly Pruning Method( HCP),
High Precedence Pruning Method( HP), Full Match then
High Pruning Method( FHP), Full Match Classify correctly
then High Classify correctly Pruning Method( FCHCP). The HCP pruning technique produced best results
in terms of predictive accuracy and number of rules
discovered.

[20] proposed a way of selecting the association rules based
on interestingness measures such as support, confidence,
correlation and soon. The study also presents the distribution
of the rule clusters with pattern Xi → Y over different interestingness measures.

[21] surveyed different rule pruning methods for removing
redundant rules while rule generation.

[22] proposed a method for pruning statistically insignificant
association rules in the presence of high confidence rules for
web usage. This pruning method is based on testing the
confidence of rules with low local z-scores.
The classification accuracy of the classifier. The study of pruning methods and its importance with an aim to increase learning approaches moving towards the discussion of rule classification and its techniques adapting different rule fast and efficient classifiers and adapt the same for pruning method in association rule mining for building the produced. The study attempts to use high confidence based terms of classification accuracy and number of rules produced. The experimental results proved ARC-AC performed well being global classifier in terms of effectiveness bit need to be improved by addition of partial and relative support thresholds.

[26] proposed text categorization using Association rules and Naïve Bayes classifier using word relations to derive future set from pre classified text documents and then using Naïve Bayes classified on derived feature for final categorization.

6. CONCLUSION

AC is preferred classification approach originated by integration of association rule mining and classification. AC is capable of building more accurate classifiers than the traditional classifiers such as decision trees KNN, SVM etc. This paper presents different ARM techniques with the improved versions followed by application of ARM. The main focus of paper is to understand Association Classification and its techniques adapting different rule learning approaches moving towards the discussion of rule pruning methods and its importance with an aim to increase the classification accuracy of the classifier. The study further will be experimental comparing the performance of different AC algorithms and rule pruning techniques in terms of classification accuracy and number of rules produced. The study attempts to use high confidence based pruning method in association rule mining for building the fast and efficient classifiers and adapt the same for automatic text document categorization.

REFERENCES

[1]. T. Karthikeyan1 and N. Ravikumar. A Survey on Association Rule Mining. International Journal of Advanced Research in Computer and Communication Engineering . ISSN: 2278-1021. Vol. 3, Issue 1, January 2014.

[2]. Lei Chen. The Research of Data Mining Algorithm Based on Association Rules. The 2nd International Conference on Computer Application and System Modeling (2012). ISBN 978–0–19–954186–7.

[3]. R.Agrawal, T.Imielinski, A.Swami. Mining Association Rules between sets of items in large databases. Proceedings of International Conference of Management of Data ACM-SIGMOD Washington D.C. Pages 207-216, May 1993.

[4]. Fadi Thabtah, Hussein Abdel-jaber, Mofleh Aldiabat. Rule Pruning Methods in Associative Classification Text Mining. Journal of Intelligent Computing Volume 1 Number 1 March 2010.

[5]. Jogendra prasad Bagdi. Rule Pruning with Correctly Classify in Associative Classification. JSRD - International Journal for Scientific Research & Development Vol. 2, Issue 10, 2014 | ISSN (online): 2321-0613.

[6]. Sotiris Kotsiantis, Dimitris Kanellopoulos. Association Rules Mining: A Recent Overview. GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[7]. Thabet Slimani, Amor Lazzef. Efficient Analysis of Pattern and Association Rule Mining Approaches.https://pdfs.semanticscholar.org/d91c/2deec3f4b191ca5f0f6fad3950b0f0907.pdf

[8]. Rashmi Shikhariya, Nitin Shukla. An improved association rule mining with fp tree using positive and negative integration. Journal of Global Research in Computer Science. Volume 3, No. 10, October 2012.

[9]. Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya. A Survey: On Association Rule Mining International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 1, January-February 2013, pp.2065-2069.

[10]. Rachna Somkunwar. A study on Various Data Mining Approaches of Association Rules. Rachna et al., International Journal of Advanced Research in Computer Science and Software Engineering 2 (9), September-December 2012, pp. 141-144.

[11]. Suchita Borkar, K.Rajeswari. Attributes Selection for Predicting Students’ Academic Performance using Education Data Mining and Artificial Neural Network. International Journal of Computer Applications (0975 – 8887) Volume 86 – No 10, January 2014.

[12]. Suchita Borkar, K.Rajeswari. Attributes Selection for Predicting Students’ Academic Performance using Education Data Mining and Artificial Neural Network. International Journal of Computer Applications (0975 – 8887) Volume 86 – No 10, January 2014.

[13]. Oladipupo O.O. & Oyelade O.J. Knowledge Discovery from Students’ Result Repository:Association Rule Mining Approach. International Journal of Computer Science & Security (IJCSS), Volume (4) Issue (2) Apr/2010.

[14]. Suchita Borkar, K. Rajeswari. Predicting Students’ Academic Performance Using Education Data Mining. International Journal of Computer Science and Mobile Computing. IJCSMC, Vol. 2, Issue. 7, July 2013, pg.273 – 279.

[15]. D. Magdalene Delighta Angeline. . Association Rule Generation for Student Performance Analysis using Apriori Algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013.

[16]. D. Magdalene Delighta Angeline. . Association Rule Generation for Student Performance Analysis using Apriori Algorithm. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 1, No. 1, March-April 2013.

[17]. Jeetesh Kumar Jain, Nirupama Tiwari, Manoj Ramaiya. A Survey: On Association Rule Mining International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 3, Issue 1, January-February 2013, pp.2065-2069.

[18]. Asha. T, S. Natarajan,K.N.B.Murthy. A Study of Associative Classifiers with Different Rule Evaluation Measures for Tuberculosis Prediction. IJCA Special Issue on “Artificial Intelligence Techniques Novel Approaches & Practical Applications”,AICT, 2011.
[19]. Brijkishor Kumar Gupta, Gajendra Singh Chandel. An approach to calculate minimum support confidence using MCAR with GA. International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE) Volume 5, Issue 5, May 2014.

[20]. S. Kannan1 and R. Bhaskaran. Association Rule Pruning based on Interestingness Measures with Clustering. IJCSI International Journal of Computer Science Issues, Vol. 6, No. 1, 2009.

[21]. Ashwini Batbarai1, Devishee Naidu, Survey for Rule Pruning in Association Rule Mining for Removing Redundancy. International Journal of Innovative Research in Science, Engineering and Technology. Vol. 3, Issue 4, April 2014.

[22]. Maja Dimitrijevica, Zita Bosnjak. Pruning statistically insignificant association rules in the presence of high-confidence rules in web usage data. 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014.

[23]. Karthik P, Saurabh M, U Chandrashekhar. Classification of Text Documents Using Association Rule Mining with Critical relative Support based Pruning. International Conference on Advances in Computing, Communications and Informatics (ICACCI) Sep 21-24, 2016, Jaipur, India.

[24]. Maria-Luiza Antonie, Osmar R. Zaiane. Text Document Categorization by Term Association. 2002 IEEE International Conference on Data Mining (ICDM’02), 2002.

[25]. Osmar R. Zaiane, Maria-Luiza Antonie, Classifying Text Document by associating Terms with Categories. Thirteenth Australasian Database Conference (ADC 2002), vol5, 2002.

[26]. S M Kamruzzaman and Chowdhury Mofizur Rahman. Text Categorization using Association Rule and Naïve Bayes Classifier. Asian Journal of Information Technology, Vol. 3, No. 9, pp 657-665, Sep. 2004. https://pdfs.semanticscholar.org/2e38/afc51537c203c62dd7bc645d4c10649992c.pdf