Association Rule Generation using Attribute Information Gain and Correlation Analysis for Classification

L Kiran Kumar Reddy[1]
Dr.S.Phani Kumar[2]

Associate Professor, Dept of CSE,
SCIENT Institute of Technology, Ibrahimpatnam, Hyderabad

Prof & Head, Department of CSE,
GITAM University, Hyderabad

Abstract

The discovery of association rules is an important data-mining task for which many algorithms have been proposed. However, the efficiency of these algorithms needs to be improved to handle real-world large datasets. Association rule mining often generates a huge number of rules, but a majority of them either are redundant or do not reflect the true correlation relationship among data objects. Some strong association rules are based on support and confidence can be misleading and this has been one of the major bottlenecks for successful application of association rule mining. In this paper we propose an efficient classification approach for generating intersecting association rules using dataset attributes information gain and its Correlation association analysis. Experiment evaluation shows high accuracy achieved in classification in compare to existing classifiers.

Keywords: *Association rule, Classification, Information Gain, Correlation Analysis, Attribute Selection.*

1. INTRODUCTION

Association rules are used to discover the relationships, and potential associations, of items or attributes among huge data. These rules can be effective in uncovering unknown relationships, providing results that can be the basis of forecast and decision. They have proven to be very useful tools for an enterprise as it strives to improve its competitiveness and profitability. According to a study by Meta Group, over 70% of Fortune's top 1000 have established data warehousing projects, to integrate their internal databases and use data mining technologies in an effort to discover meaningful information. The mined results will be consulted in the executive decision-making process [13]. The application and development of association rules is a popular area of data mining research [11].

Existing classification and rule learning algorithms in machine learning [16] mainly use heuristic/greedy search to find a subset of regularities (e.g., a decision tree or a set of rules) in data for classification[4][5]. In the past few years, extensive research was done in the database community on learning rules using exhaustive search under the name of association rule mining. The objective there is to find all rules in data that satisfy the user-specified minimum support and minimum confidence. Although the whole set of rules may not be used directly for accurate classification, effective and efficient classifiers have been built using the rules.

Ultimately, only the user can judge if a given rule is interesting, and this judgment, being subjective, may differ from one user to another. However, objective interestingness measures, based on the statistics “behind” the data, can be used as one step toward the goal of weeding out uninteresting rules from presentation to the user[19][20]. This paper presents, an effective classification approach for generating interesting association rules based on dataset attributes information gain and its Correlation analysis known as IG-ACA. It works on the two phases, as in Phase-1 attribute selection using Information gain (IG) and in Phase-2 Attributes Correlation analysis (ACA) for rule generation.

The rest of the paper is organized as follows, section-2 describes background study related to association rules and classification related works. Section-3 presents the proposed association rule generation approach and algorithms. Section-4 describes the experiment and results evaluations and section-5 presents the paper conclusion.

ISSN: 2278 – 1323

2. BACKGROUND STUDY

Building effective classification systems is one of the central tasks of data mining and machine learning tasks .Since its introduction, Association Rule Mining [1][2], has become one of the core data mining tasks, and has attracted tremendous interest among data mining researchers and practitioners. It has an elegantly simple problem statement, that is, to find the set of all subsets of items (called itemsets) that frequently occur in many database records or transactions, and to extract the rules telling us how a subset of items influences the presence of another subset. The existing techniques are, however, largely based on heuristic/greedy search. They aim to find only a subset of the regularities (e.g., a decision tree or a set of rules) that exist in data to form a classifier [6][10][11].

Classifying real world instances is a common thing anyone practices through his life. One can classify human beings based on their race or can categorize products in a supermarket based on the consumers shopping choices. In general, classification involves examining the features of new objects and trying to assign it to one of the predefined set of classes [3]. Given a collection of records in a data set, each record consists of a group of attributes; one of the attributes is the class. The goal of classification is to build a model from classified objects in order to classify previously unseen objects as accurately as possible.

There are many classification approaches[17][18] for extracting knowledge from data such as divide-and-conquer [7], separate-and-conquer [8], covering and statistical approaches [9]. The divide-and-conquer approach starts by selecting an attribute as a root node, and then it makes a branch for each possible level of that attribute. This will split the training instances into subsets, one for each possible value of the attribute. The same process will be repeated until all instances that fall in one branch have the same classification or the remaining instances cannot be split any further. The separate-and-conquer approach, on the other hand, starts by building up the rules in greedy fashion (one by one). After a rule is found, all instances covered by the rule will be deleted. The same process is repeated until the best rule found has a large error rate. Statistical approaches such as Naïve Bayes [9] use probabilistic measures, i.e. likelihood, to classify test objects. Finally, covering approach [15] selects each of the available classes in turn, and looks for a way of covering most of training objects to that class in order to come up with maximum accuracy rules.

On the other hand, a training data set often generates a huge set of rules. It is challenging to store, retrieve, prune, and sort a large number of rules efficiently for classification. Many studies [1][5] have indicated the inherent nature of a combinatorial explosive number of frequent patterns and hence association rules that could be generated when the support threshold is small (i.e., when rare cases are also be included in the consideration). To achieve high accuracy, a classifier may have to handle a large set of rules, including storing those generated by association mining methods, retrieving the related rules, and pruning and sorting a large number of rules. In this paper, we present a new approach IG-ACA based on attribute gain and correlation analysis for accurate and efficient classification to overcome this problem.

3. PROPOSED ASSOCIATION RULE GENERATION APPROACH

In general, given a training data set, the task of classification is to build a classifier from the training data set such that it can be used to predict class labels of unknown objects with high accuracy. This approach is to explore association relationships between object attributes and class labels. The idea is natural since it utilizes frequent attributes information gain to and class labels Correlation over training data set to generate the effective classification rules.

3.1. Attribute Selection Based on Information Gain

Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task or redundant. Claude Shannon [12] on information theory studied the value or "information content" of messages. Such an approach minimizes the expected number of tests attributes needed to classify a given tuple. The attribute with the highest information gain (IG) is chosen as the attribute for the association rule generation and it evaluates attributes by measuring their information gain with respect to the class.

Let *C* be set consisting of *c* data samples with *m* distinct classes. The training dataset c_i contains sample of class information. Then expected information *I*, of C needed to classify a given sample is calculated by equation-1.

$$I(C) = -\sum_{i=1}^{m} \frac{c_i}{c} log_2 \left(\frac{c_i}{c}\right) \qquad (1)$$

where, $\frac{c_i}{c}$ is the probability that an arbitrary sample to a class c_i. A log function to the base 2 is used, because the information is encoded in bits. *Information (C)* is just the average amount of information needed to identify the class label of a tuple in *C*.

Now, suppose we have to select the tuples in *C* on some attribute *A* having *v* distinct values, $\{a_1, a_2, \ldots, a_v\}$, as observed from the training data. Then the information, *I* of *A* in C needed is calculated by equation-2.

$$I_A(C) = -\sum_{j=1}^{v} \frac{|c_j|}{|c|} \times I(c_j) \qquad (2)$$

The term $\frac{|c_j|}{|c|}$ acts as the weight of the j^{th} value sets. $I_A(C)$ is the expected information required to classify a tuple from *C* based on the attribute *A*, and actual information gain *(IG)* of the attribute *A* is calculated using equation-3.

$$IG(A) = I(C) - I_A(C) \qquad (3)$$

The attribute *A* with the highest information gain is chosen as the primary attribute for the Correlation analysis.

3.2. Attribute Correlation Analysis

We propose an attribute Correlation Analysis (ACA) algorithm to correlate to other classifying attributes in relate to the classified class of the attributes in trained set data. It allows us to analyze categorical data correlation relationship between two attributes, *A* and *B*.

Algorithm-1: *Attribute Correlation Analysis*

Method: *ACA_Analysis (T, HG_A, A)*

Inputs: *T* → array of trained transaction datasets.
HG_A → highest gains attribute
A → array of attributes

Outputs: Each items of *HG_A* values correlation patterns as, *HG_A_Pattern[]*

```
HG_A_Items[] ← Unique items of Attribute HG_A
                 from T

for each item i in A[]
 A_Items[] ← Unique items of Attribute A[i] from T
   for each item j in HG_A_Items[]
     c_label[i][j]←T.ClassLabel(A[i],
                        HG_A_Items[j])
   End for
End for

for each label i in c_label
      for each label j in c_label
         attLabel ← c_label[i][j]
            if attLabel == true then
            HG_A_Pattern[i,j].add( Item.A[i])
            End if
         End for
End for
```

The obtained items pattern of highest gain attribute from the algorithm-1 will be used for the rule generation. It is highly desirable for data mining systems to generate only interesting patterns. This would be much more efficient for users and data mining systems, because neither would have to search through the patterns generated in order to identify the truly interesting ones.

Algorithm-2: *Rules Generation*

Method: *Rule_Generation (HG_A_Pattern)*

Inputs: *HG_A_Pattern* → array of generated patterns of HG_*A*.

Outputs: *Rules Patterns*

```
      HG_A_Items[] ← Unique items of Attribute
  HG_A from T

for each item k in HG_A_Items[]
  Item_Patterns[]←HG_A_Pattern.get(HG_A_Items[k])

   for each pattern p in Item_Patterns []
          Rules_pattern[k] ← Rules_pattern[k].append
                             (Item_Patterns[p])
   End for
End for
```

The generated patterns are essentially efficient discovery for classification. We evaluate our proposal with a synthetic datasets to evaluate the performance in the next section.

4. EXPERIMENT EVALUATION

An extensive experimental evaluation is performed over synthetic databases to observe the efficiency of the classification patterns generation. To perform an experiment analysis of our proposal we use Java tool. Initially we implement attribute selection using information gain and then implement attribute correlation analysis to generate the associated rules patterns.

4.1 Datasets

We create a synthetic trained database required for the employee selection based on three attributes and 14 transaction tuples with class labels and each attributes has related qualifying values as shown in Table-1.

Table-1: Attributes and its Item values

Attributes	**Items Values**
Age_ Level	Youth, Middle_aged, Senior
Qualification	Diploma, Graduate, Master
Experience	Yes, No
Conduct	Bad, Good, Excellent

Using the Table-1 attributes we created a transaction database as shown in Table-2. It presents a training set with class label tuples. The class-label "*Qualify for Selection*" has two distinct values as, "*yes*" and "*No*".

Table-2: Clas-Labled Trained Transaction Table of Employee selection

T.Id	**Age_ Level**	**Qualification**	**Experience**	**Conduct**	**Class:** ***Qualify for Selection***
1	Youth	Master	No	Bad	No
2	Youth	Master	No	Good	No
3	Middle_age	Master	No	Good	Yes
4	Senior	Graduate	No	Good	Yes
5	Senior	Diploma	Yes	Good	Yes
6	Senior	Diploma	Yes	Bad	No
7	Middle_age	Diploma	Yes	Excellent	Yes
8	Youth	Graduate	No	Good	No
9	Youth	Diploma	Yes	Good	Yes
10	Senior	Graduate	Yes	Good	Yes
11	Youth	Graduate	Yes	Excellent	Yes
12	Middle_age	Graduate	No	Excellent	Yes
13	Middle_age	Master	Yes	Good	Yes
14	Senior	Graduate	No	Excellent	No

4.2 Results

In phase-1 of the implementation we identified the highest information gain attribute among of the database in Table-2. We compute the database information, *I(D)* and obtain that *I(D) = 0.940 bits*, and also compute each database attribute information gain to find the highest gain attribute. Table-3 shows the computed information gain results of the attributes.

Table-3: Attributes Information Gain Values

Attribute	**Gain Value** ***(in Bits)***
Age_Level	0.694
Qualification	0.024
Experience	0.151
Conduct	0.048

Table-3 results shows "*Age_Level"* attributes shows high gain values in compares to other attributes. So, for the next phase of the analysis we select "*Age_Level"* as highest gain value. Using highest gain value we perform attributes Correlation Analysis and we obtain some each attribute patterns as shown below Table-4.

Table-4: ACA Pattern derived based on "*Age_Level*"

Attributes Correlation Analysis

	Master	*Graduate*	*Diploma*
Youth	N	Y	Y
Middle_age	Y	Y	Y
Senior	NA	Y/N	Y/N

	With Exp.	*No. Exp.*
Youth	Y	Y/N
Middle_age	Y	Y
Senior	Y/N	Y

	Bad	*Good*	*Excellent*
Youth	N	Y/N	Y/N
Middle_age	NA	Y	Y
Senior	N	Y	N

Pattern Derived

Youth	*{Graduate, Diploma}*
Middle_age	*{Master, Graduate, Diploma }*
Senior	*{Graduate, Diploma}*
Youth	*{With Exp, No. Exp.}*
Middle_age	*{ With Exp, No.Exp.}*
Senior	*{ With Exp, No. Exp.}*
Youth	*{Good, Excellent}*
Middle_age	*{ Good, Excellent }*
Senior	*{ Good, Excellent }*

Table-5: Rule generated based on "*Age_Level*"

Youth	*{Graduate, Diploma}{With Exp, No. Exp.}{Good, Excellent}*
Middle_Age	*{Master, Graduate, Diploma }{With Exp, No.Exp.}{ Good, Excellent }*
Senior	*{Graduate, Diploma}{With Exp, No.Exp.}{ Good, Excellent }*

Based on the pattern obtained in Table-4 we generate the required association rules for classification. The generated pattern will be effective for the run time classification. Table-5 present the final rule generated for the classification.

5. CONCLUSION

Association rule mining often generates a huge number of rules, but a majority of them either are redundant or do not reflect the true correlation relationship among data objects. In this paper we presents, an effective classification approach for generating interesting association rules based on dataset attributes information gain and its Correlation analysis known as IG-ACA. It executes on the two phases, as in Phase-1 attribute selection using Information gain (IG) and in Phase-2 Attributes Correlation analysis (ACA) for rule generation. We implement the proposed work over a synthesis dataset to evaluate the rule generation. Effective rules are generated on execution. In feature work, we evaluate the work in compare to the existing classifiers.

ACKNOWLEDGEMENT

A manuscript effort of such a great worth is only possible with the help of several people, directly or indirectly. First and foremost I have immense happiness in expressing my sincere thanks to my guide, Dr.S. Phani Kumar, Prof & Head, CSE Dept, GITAM University, Hyderabad for his continuous support and valuable suggestions whenever it was needed. I am very much thankful to all my near and dears whose presence always inspires me to do better. My happiness culminates, when I recall the co-operation extended by my friends during the completion of this task. A final and heartily thanks go to my parents without them I am nowhere.

REFERENCES

[1]. By XING Xue CHEN Yao WANG Yan-en,"Study on Mining Theories of Association Rules and Its Application" .International Conference on Innovative computing and communication Asia – Pacific Conference on Information Technology and Ocean Engineering 978-0-7695-3942-3/10, 2010.

[2]. Ila Chandrakar and A. Mari Kirthima, "A Survey On Association Rule Mining Algorithms", In international Journal Of Mathematics and Computer Research, ISSN:2329-7167,vol-1, November,2013

[3]. J. Pei, J. Han, and R. Mao, "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets," Proc. ACM SIGMOD Workshop Research Issues in Data Mining and Knowledge Discovery, pp. 21-30, 2000.

[4]. Devasri Rai, A.S.Thoke and Keshri Verma, "Enhancement of Associative Rule based FOIL and PRM Algorithms", Proc.. 2012.

[5]. W. Li ,J.Han and J.Pie CMAR: Accurate and efficient Classification based on multiple class-association rules. In ICDM01, 2011,pp.369{376.san jose, CA ,Nov.2001.

[6]. M.J. Zaki and C.J. Hsiao, "Charm: An Efficient Algorithm for Closed Itemset Mining," Proc. Second SIAM Int'l Conf. Data Mining, pp. 34-43, 2002.

[7]. P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the Right Objective Measure for Association Analysis," Information Systems, vol. 29, pp. 293-313, 2004

[8]. K. Wang, S. Zhou, and Y. He. Growing decision tree on support-less association rules. In KDD'00, Boston, MA, Aug. 2000.

[9]. D. Burdick, M. Calimlim, J. Flannick, J. Gehrke, and T. Yiu, "Mafia: A Maximal Frequent Itemset Algorithm," IEEE Trans.Knowledge and Data Eng., vol. 17, no. 11, pp. 1490-1504, Nov.2005.

[10]. J. R. Quinlan, "An empirical comparison of genetic and decision-tree classifiers". In Proc. 1988 Int. Conf. Machine Learning (ML'88), pages 135–141, San Mateo, CA, 1988.

[11]. R. Agrawal, T. Imilienski, A. Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, DC, 1993 pp. 207–216.

[12]. C. E. Shannon and W.Weaver, "The mathematical theory of communication". University of Illinois Press, Urbana, IL, 1949.

[13]. META Group, Data warehouse marketing trends/opportunities: an in-depth analysis of key market trends, META Group 1998.

[14]. By CH.Sandeep Kumar, K.Shrinivas, Peddi Kishor T.Bhaskar: "An Alternative Approach to Mine Association Rules" 978-1-4244- 8679-3/11,IEEE, 2011.

[15]. Claudia Marinica and Fabrice Guillet, "Knowledge-Based Interactive Postmining of Association Rules Using Ontologies," IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 6, June 2010.

[16]. J. Li, "On Optimal Rule Discovery," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 4, pp. 460-471, Apr. 2006.

[17]. Fayyad, U., and Irani, K. (1993). Multi-interval discretisation of continues-valued attributes for classification learning. IJCAI-93, pp. 1022-1027, 1993.

[18]. R.J. Bayardo, Jr., R. Agrawal, and D. Gunopulos, "Constraint-Based Rule Mining in Large, Dense Databases," Proc. 15th Int'l Conf. Data Eng. (ICDE '99), pp. 188-197, 1999.

[19]. Xiong,H."Mining strong affinity association patterns in data sets with skewed support distribution"Computing & processing (software & hardware) 0-7695-1978-4/09 IEEE 2003

[20]. Yihua Zhong, Yuxin Liao, "Research of Mining Effective and Weighted Association Rules Based on Dual Confidence", Fourth International Conference on Computational and Information Sciences, 2012, IEEE.