Discovery of Scalable Association Rules from Large Set of Multidimensional Quantitative Datasets

Tamanna Siddiqui, M Afshar Aalam, and Sapna Jain
Jamia Hamdard/Department of Computer Science, Haryana, India
Email: ja_zu_siddiqui@hotmail.com, mailtoafshar@rediffmail.com, hellosap@sify.com

***Abstract*— In proposed approach, we introduce the problem of mining association rules in large relational tables containing both quantitative and categorical attributes. We have proposed an algorithm for Discovery of Scalable Association Rules from large set of multidimensional quantitative datasets using k-means clustering method based on the range of the attributes in the rules and Equi-depth partitioning using scale k-means for obtaining better association rules with high support and confidence. The discretization process is used to create intervals of values for every one of the attributes in order to generate the association rules. The result of the proposed algorithm discover association rules with high confidence and support in representing relevant patterns between project attributes using the scalable k-means .The experimental studies of proposed algorithm have been done and obtain results are quite encouraging.**

***Index Terms*— Data Mining, Association rules, k-means clustering, CBA tool, Discretization, Partitioning.**

I. INTRODUCTION

Data mining, also known as knowledge discovery in databases, has been recognized as a new area for database research. The problem of discovering association rules was later introduced as a data mining approach to find out the frequent itemset from the given set of data. Given a set of transactions, where each transaction is a set of items, an association rule is an expression of the form X + Y, where X and Y are sets of items. The problem is to find all association rules that satisfy user-specified minimum support and minimum confidence constraints.

Relational tables in most business and scientific domains have richer attribute types. Attributes can be quantitative or categorical. Boolean attributes can be considered a special case of categorical attributes [2]. This research work defines the problem of mining association rules over quantitative attribute in large relational tables and techniques for discovering such rules. This is referred as the Quantitative Association Rules problem [5], [3].

The problem of mining association rules in categorical data presented in customer transactions was introduced by Agrawal R, T Lmielinski and A Swami [10]. This research work provided basic idea to several investigation efforts resulting in descriptions of how to extend the original concepts and how to increase the performance of the related algorithms [15]. The original problem of mining categorical was extended in several directions such as adding or replacing the confidence and support by other measures, or filtering the rules during or after generation, or including quantitative attributes. The use of the categorical attributes simplifies the procedure of mining the rules [19]. In the last years the application areas involving other types of attributes have increased significantly [21].

Scalability means that as a system gets larger, its performance improves correspondingly. For data mining, scalability means that by taking advantage of parallel database management systems and additional CPUs, you can solve a wide range of problems without needing to change your underlying data mining environment.

Clustering involves partitioning a given data set into several groups based on some similarity or dissimilarity measurements. Cluster analysis has been widely used in information retrieval, text and web mining, pattern recognition, image segmentation and software reverse engineering. Scalability of a clustering algorithm relies heavily on dimensionality and choice of distance function. Many clustering algorithms are good at handling low-dimensional data, involving only two or three dimensions. It is challenging to cluster data in high dimensional space and more than often speed and accuracy of an algorithm can be scaled. Some algorithms use pre processing of data to normalize many dimensions into manageable set. Each iteration of a clustering algorithm invariably has to calculate distance between points and centers. It is important to select a simple and yet elective distance function to make it efficient. Most of the times, quality of clusters depreciates as we try to improve speed of clustering algorithm [11] .

K-means is the most intuitive and popular clustering algorithm. However, the classical K-means suffers from several flaws. First, the algorithm is very sensitive to the initialization method and can be easily trapped at a local minimum regarding to the measurement (the sum of squared errors) used in the model. On the other hand, it has been proved that finding a global minimal sum of the squared errors is NP-hard even when k = 2[17]. In the proposed approach scalable k-means uses partitional

doi:10.4304/jait.3.1.69-76

clustering method to generate better association rules with high confidence and support.

The rest of the paper is organized as follows. We introduce description of some works in the literature concerning the improvement of association rule algorithms in Section 2. Section 3 is dedicated to the proposed algorithm description. Section 4 gives the illustration explaining the proposed approach. The experimental study and conclusion are presented in sections 5 and 6 respectively.

II. RELATED WORK

The concept of association between items was first introduced by Agrawal R, T Lmielinski and A Swami [10]. Since they proposed the popular Apriori algorithm [18] , the improvement of the algorithms for mining association rules have been the target of numerous studies [12] ,[23]. Many other authors have studied better ways for obtaining association rules from transactional databases [14]. Most of the efforts have been oriented to simplify the rule set and improve the algorithm performance.

Extracting all association rules from a database requires counting all possible combination of attributes [4]. Support and confidence factors can be used for obtaining interesting rules which have values for these factors greater than a threshold value. In most of the methods the confidence is determined once the relevant support for the rules is computed. Nevertheless, when the number of attributes is large, computational time increases exponentially. For a database of m records of n attributes, assuming binary encoding of attributes in a record, the enumeration of subset of attributes requires m x 2^n computational steps. For small values of n, traditional algorithms are simple and efficient, but for large values of n the computational analysis is unfeasible. When continuous attributes are involved in the rules, the discretization process is critical in order to reduce the value of n and to obtain high confident rules at the same time [25].

The main goal of association rule mining is to discover relationships among set of items in a transactional database. Association rule has been extensively studied since it was first introduced [8], [16]. A typical application of association rule mining is the market basket analysis. An association rule is an implication of the form A->B, where A and B are frequent itemsets in a transaction database and A∩B=C. The rule A-->B can be interpreted as if itemset A occurs in a transaction, then itemset B will also likely occur in the same transaction [22]. By such information, market personnel can place itemset A and B within close proximity which may encourage the sale of these items together and develop discount strategies based on such association or correlation found in the data [6] , [7] .

Therefore, association rule mining has been received a lot of attention. The sequential mining patterns [19], as well as mining quantitative association rules in large relational tables in [20] were the main area focused for research. The traditional algorithms discover valid rules by exploiting support and confidence requirements, and use a minimum support threshold to prune its combinatorial search space [13]. Two major problems may arise when applying such strategies [8]. If the minimum support is set too low, this may increase the workload significantly such as the generation of candidate sets, construction of tree nodes, comparison and test. It will also increase the number of rules considerably, which makes the traditional algorithms suffering from extremely poor performance problem. In addition, many patterns involving items with substantially different support level are produced, which usually have a weak correlation and are not really interesting to users [16]. If the minimum support threshold is set too high, many interesting patterns involving items with low supports are missed. Such patterns are useful for identifying associations among rare but expensive items such as diamond necklace, ring and earrings, as well as the identification of identical or similar web documents.

In this paper, we define the problem of mining multidimensional quantitative association rules over quantitative and categorical attributes in large relational tables and present techniques for discovering such rules.

III. PROPOSED WORK

Steps of the Proposed Algorithm

Step 1: Input Phase

The distribution of attribute values in the clusters was used for making the discretization according to the following procedure:

1. The number of intervals for each attribute is the same of the number of clusters where m is the mean value of the attribute in the in the clusters.

2. When two adjacent intervals overlap, the cut point (superior boundary of the first and inferior boundary of the next) is placed in the middle point of the overlapping region. These intervals are merged into a unique interval if one of them includes the mean value of the other or is very close to it.

3. When two adjacent intervals are separated, the cut point is placed in the middle point of the separation region. This procedure was applied for creating intervals of values for every one of the attributes in order to generate the association rules.

Step 2: Candidate Generation

Given Lk -1, the set of all frequent k – l-itemsets, the candidate generation procedure must return a superset of the set of all frequent k-itemsets. The k-means clustering helps in finding the appropriate and definite cluster with partitioning. This procedure has three parts:

a) Join Phase. Lk - 1 is joined with itself, the join condition being that the lexicographically ordered first k – 2 items are the same, and that the attributes of the last two items are different.

b) Subset Prune Phase. All itemsets from the join result which have some (k – 1)-subset that is not in Lk-1 are deleted.

c) Interest Prune Phase. If the user specifies an interest level and wants only itemsets whose support and confidence is greater than expected, the interest measure is used to prune the candidates further.

Step 3: Counting Support of Candidates.

In the process of counting support of candidates when we make a pass, we read one record at a time and increment the support count of candidates supported by the record. Thus, given a set of candidate itemsets C and a record t, we need to find all itemsets in C that are supported by t. We partition candidates into groups such that candidates in each group have the same attributes and the same values for their categorical attributes.

Step 4: Generating Rules.

We use the frequent itemsets to generate association rules. The general idea is that if, say, ABCD and AB are frequent itemsets, then we can determine if the rule AB→ CD holds by computing the ratio conf = support (ABCD)/support (AB) . If conf >= supconf, then the rule will have minimum support because ABCD is frequent. The clusters are created with a weight for the output. This is a supervised way of producing the most suitable clusters for the prediction of the output variables, which appear in the consequent part of the rules generation. We use the proposed algorithm to generate scalable association rules.

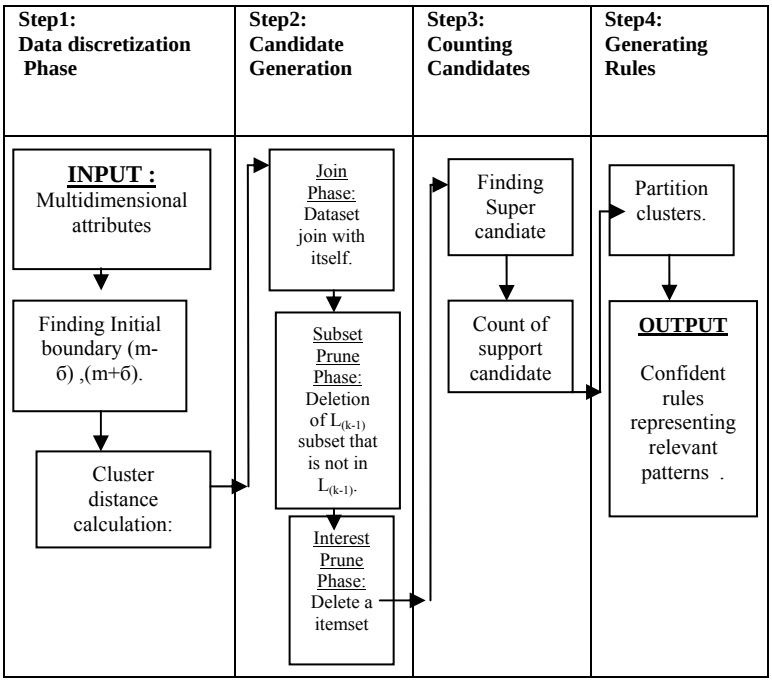

Figure 1: Block Diagram of the proposed algorithm

IV. AN ILLUSTRATIVE EXAMPLE

We use a Student data which contain the details such as marks, stream and stipend to generate the scalable association rules using the proposed approach.

Step1: INPUT phase

1. Determine the number of partitions for each quantitative attribute.

2. For categorical attributes, map the values of the attribute to a set of consecutive integers. For quantitative attributes that are not partitioned into intervals, the values are mapped to consecutive integers such that the order of the values is preserved. If a quantitative attribute is partitioned into intervals, the intervals are mapped to consecutive integers, such that the order of the intervals is preserved. From this point, the algorithm only sees values (or ranges over values) for quantitative attributes. Figure 2(a) gives the description of the student dataset which is used as an illustrative example.

Transaction Id	Marks 80…90	Marks 90...100	Stream: Medical	Stream: Nonmedical	Stipend: 1000	Stipend: 800
T001	0	1	1	0	1	0
T002	1	0	0	1	0	1
T003	0	1	1	0	1	0
T004	1	0	0	1	0	1
T005	0	1	1	0	1	0

Figure 2(a): Student dataset

Step2: Candidate Generation

a) Join Phase:

For example, let L2 consist of the following itemsets:

{(Stream: Medical) (Marks: 80...84)}
{(Stream: Medical) (Marks: 80…89)}
{(Stream: Medical) (Stipend: 800...900)}
{(Marks: 80...89) (Stipend: 800…900)}

After the join step, the result will consist of the following itemsets:

{(Stream: Medical) (Age: 80...84) (Stipend: 800...900)}

{(Stream: Medical) (Age: 80...89) (Stipend: 800...900)}

b)Subset Prune Phase : Continuing the earlier example, the prune step will delete the itemset :{ (Stream: Medical) (Age: 80..84) (Stipend: 800..900)) since its subset { (Age: 20..24) (Stipend: 800..900) } is not in L2. In step 1, we decided to partition marks into 4 intervals, as shown in Figure 2(b).

Interval
80..84
85..89
90..94
95..99

Transactionid	Marks	Stream	Stipend
T001	90..94	Medical	1000
T002	85..89	nonmedical	800
T003	90..94	Medical	1000
T004	85..89	Nonmedical	800
T005	95..99	Medical	1000

Figure 2(b): Partition marks

c) Interest Prune Phase:

K-Means clustering helps in clustering of items with the same target category are identified, and predictions for new data items are made by assuming that they are of the same type as the nearest cluster center. The k-means use clustering can help in scale out dataset and overcome the following mapping problem.

Mapping problem: We can now split the problem into two parts:

We first find which "super-candidates" are supported by the categorical attributes in the record. We re-use a hash-tree data structure described to reduce the number of super-candidates that need to be checked for a given record. Once we know that the categorical attributes of a "super-candidate" are supported by a given record, we find minconf for the candidate.

If a "super-candidate" has n quantitative attributes, the quantitative attributes are fixed for a given "super-candidate". Hence, the set of values for the quantitative attributes correspond to a set of n dimensional rectangles (each rectangle corresponding to a candidate in the super-candidate). The values of the corresponding quantitative attributes in a database record correspond to a n-dimensional point. Thus the problem reduces to finding which n-dimensional rectangles contain a given n-dimensional point, for a set of n-dimensional points. The classic solution to this problem is to put the rectangles in R*-tree [17].

If the number of dimensions is small, and the range of values in each dimension is also small, there is a faster solution. Namely, we use n-dimensional array, where the number of array cells in the j-th dimension equals the number of partitions for the attribute corresponding to the j-th dimension. We use this array to get support counts for all possible combinations of values of the quantitative attributes in the super-candidate. The amount of work done per record is only O i.e. number of dimension. Since, we simply index into each dimension and increment the support count for a single cell. At the end of the pass over the database, we iterate over all the cells covered by each of the rectangles and sum up the support counts. Using a multi-dimensional array is cheaper than using an R*-tree, in terms of CPU time. However, as the number of attributes (dimensions) in super-candidate increases, the multi- dimensional array approach will need a huge amount of memory. Thus there is a tradeoff between less memory for the R*-tree versus less CPU time for the multi-dimensional array. We use a heuristic based on the ratio of the expected memory use of the R*-tree to that of the multi-dimensional array to decide which data structure to use.

Conceptually, the table now looks as shown in Figure 2(c). After mapping the intervals to consecutive integers, we map the attributes marks in Figure 2(d) and stream in Figure 2(e).

Transactionid	Marks	Stream	Stipend
T001	3	1	1000
T002	2	2	800
T003	3	1	1000
T004	2	2	800
T005	4	1	1000

Figure2(c): After mapping attributes frequent itemsets

Value	Integer
Medical	1
Nonmedical	2

Figure2 (d): Mapping marks

Interval	Integer
80..84	1
85..89	2
90..94	3
95..99	4

Figure 2(e): Mapping stream

Step 3: Counting Support of Candidates

We use the partial completeness method to count the support of candidates generated by scalable algorithm. We first define partial completeness over itemsets rather than rules, since we can guarantee that a close itemset will be found whereas we cannot guarantee that a close rule will be found. We then show that we can guarantee that a close rule will be found if the minimum confidence level 'R'. 'R' is the minimum confidence support level required to generate efficient scalable association rules. We replace each such group with a single "super-candidate". Each "super-candidate" has two parts: (i) the common categorical attribute values, and (ii) a data structure representing the set of values of the quantitative attributes.

For example, consider the candidates:

{(Stream: Medical) (marks: 80...84), (Stipend: 800...900)}

{(Stream: Medical} (marks: 80. .89), (Stipend: 800...900)}

{(Stream: Medical) (marks: 84...89), (Stipend: 800...900)}

Step 4: Generating Rule Phase

Assuming minimum support of 40% and minimum confidence of 50%, Figure 2(f) shows some of the frequent itemsets, and Figure 2(g) some of the rules. We have replaced mapping numbers with the values in the original table in these two figures. Notice that the item (Age: 20…29) corresponds to a combination of the intervals 20...24 and 25...29. We consider that the if Minimum Support is 40% and Minimum Confidence is 50% then 3 records are used to generate the rules.

Rule	Support	Confidence
(Marks:80..89) and (Stream: nonmedical)→(stipend:800)	40%	100%
(Marks:90..99) →(Stipend)	60%	66.6%

Figure 2(f): Confidence and support generation

Itemset	Support
{(Marks:80..89)}	2
{(Marks:90..99)}	3
{(Stream: Medical)}	3
{(Stream: Nonmedical)}	2
{(Stipend:1000)}	3
{(Marks:80..89)},(Stream: Nonmedical)}	2

Figure 2(g): Scalable association rules.

We first define partial completeness over itemsets rather than rules, since we can guarantee that a close itemset will be found whereas we cannot guarantee that a close rule will be found. We then show that we can guarantee that a close rule will be found if the minimum confidence level for R is less than that for R by a certain computable)amount.

The first two conditions ensure that 'P' only contains frequent itemsets and that we can generate rules from 'P'. The first part of the third condition says that for any itemset in C there is generalization of the itemset with at most k times the support in P. The second part says that the property that the generalization has at most k times the support also holds for corresponding subsets of attributes in the itemset and its generalization. Notice that if k= 1then 'P' becomes identical to C.

The itemsets 2, 3, 5 and 7 would from a 1.5-complete set, since for any itemset X, either 2, 3, 5 or 7 is a generalization whose support is at most 1.5 times the support of X. For instance, itemset 2 is a generalization of itemset 1, and the support of itemset 2 is 1.2 times the support of itemset 1. Itemsets 3, 5 and 7 do not form a 1.5-complete set because for itemset 1, the only generalization among 3, 5 and 7 is itemset 3, and the support of 3 is more than 1.5 times the support of 1.

The mapping of Itemsets is done as shown in the Figure 2(d). There are two problems with this simple approach when applied to quantitative attributes:

a) "MinSup": If the number of intervals for a quantitative attribute (or values, if the attribute is not partitioned) is large, the support for any single interval can be low. Hence, without using larger intervals, some rules involving this attribute may not be found because they lack minimum support.

b) "MinConf": There is some information lost whenever we partition values into intervals. Some rules may have minimum confidence only when an item in the antecedent consists of a single value (or a small interval). This information loss increases as the interval sizes become larger. For example, in Figure 2(f), the rule "(stipend: O)→(stream: Nonmedical)" has 100% confidence. But if we had partitioned the attribute stipend into intervals such that O and 1 stipend end up in the same partition, then the closest rule is "(stipend: O.. 1) → (stream :nonmedical)", which only has 66.6% confidence. Figure 2(d) shows this mapping for the non-key attributes of the Student table given in Figure2 (a). Marks are partitioned into two intervals: 80...90 and 90...100.The categorical attribute, stream, has two boolean attributes ("stream: Medical" AND "stream: Nonmedical "). Since the number of values for stipend is limited to two values, Numstipend is not partitioned into intervals; each value is mapped to a boolean field. Record T001, which had (Marks: 94) now has "Marks: 90...100" equal to "l", "Marks: 80...90" equal to "O", etc.

V. EXPERIMENTAL STUDIES:

CBA originally stands for Classification Based on Associations. The experimental study is carried on dynamic simulation environment CBA which is a data mining tool developed at School of Computing, National University of Singapore. Its main algorithm was presented as a plenary paper "Integrating Classification and Association Rule Mining" in the 4th International Conference on Knowledge Discovery and Data Mining, August 23-27, 1998, New York City, USA. . However, it turns out that it is more powerful than simply producing an accurate classifier for prediction. It can also be used for mining various forms of association rules, and for text categorization or classification. This environment manages data from real projects developed in local companies and simulates different scenarios. It works with more than 20 input parameters and more than 10 output variables and generates 131061 rules. The number of records generated for this work is 400 and the variables used are sepal length, sepal width from the Iris dataset. The support (minsupp) of the rules generated is shown in figure 3.

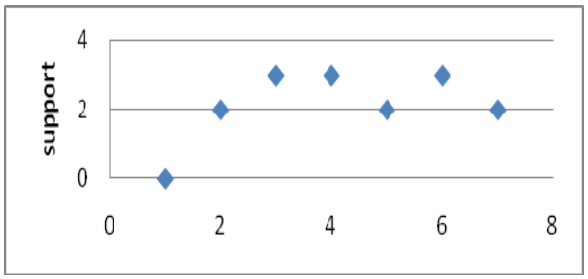

Figure 3: Minimum support of Iris Dataset .

The aim of the work is to obtain an associative model that allows studying the influence of the input variables related to the project management policy on the output variables related to the software product and the software process.

The clusters were created with a weight for the output variables three times greater than for input attributes. This is a supervised way of producing the most suitable clusters for the prediction of the output variables, which appear in the consequent part of the rules. Figure 4 gives the cluster partitioning of the Iris dataset.

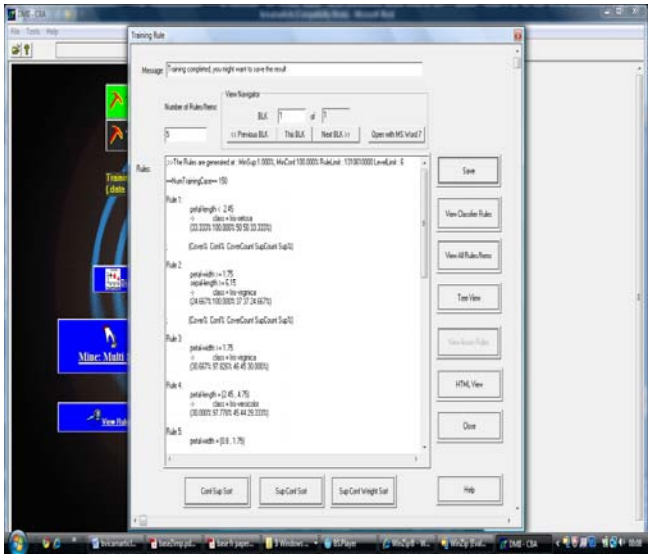

Figure 4: Cluster partition.

The Rules are generated at: MinSup: 1.000%, MinConf: 100.000% RuleLimit: 1310610000: LevelLimit: 6. We compare the number of cluster formed while generating the rules in figure 5. After comparison we can conclude that our proposed method generate more scaled and efficient association rules.

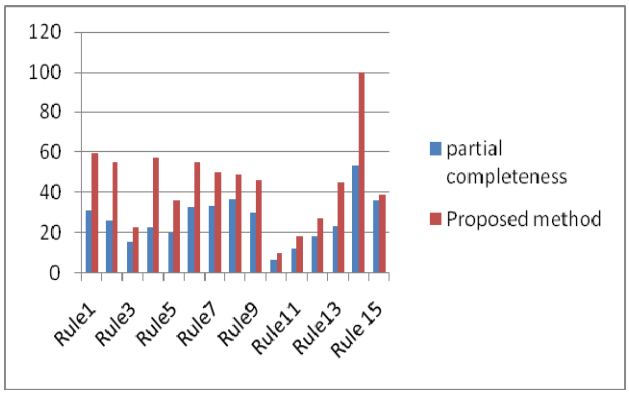

Figure 5: Cluster allocation comparison

Under the exposed conditions, 15 rules were generated.Their confidence and support factors are described above. Figure 6 shows the rule generated with our proposed method. Figure 7 shows the support for the rules generated with the proposed method.

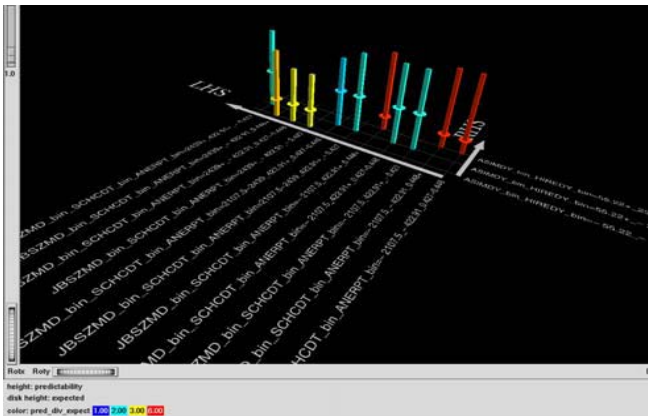

Figure 6: Rules generated.

Number	Itemset	Support
1	{Marks(80..85)}	5%
2	{Marks(80..90)}	6%
3	{Marks(80..99)}	8%
4	{Stipend(800..900)}	5%
5	{Stipend(800..1000)}	6%
6	{(Marks:80..85),(stipend:800..900)}	4%
7	{(marks:80..90),(stipend:800..1000)}	5%

Figure 7: Support for the rules generated.

VI. CONCLUSION

We introduced the problem of mining association rules in large relational tables containing both quantitative and categorical attributes. We dealt with quantitative attributes by fine-partitioning the values of the attribute and then combining adjacent partitions as necessary. We introduced a measure of partial completeness which quantifies the information lost due to partitioning. This measure is used to decide whether or

not to partition a quantitative attribute, and the number of partitions. The success of the algorithm is mainly due to the supervised multivariate procedure used for discretizing the continuous attributes in order to generate the rules. The result is an association model constituted by a manageable number of high confident rules representing relevant patterns between project attributes. This allows estimating the influence of the combination of some variables related to management policies on the software quality, the project duration and the development effort simultaneously.

We gave an algorithm for mining quantitative association rules. Our study showed that the algorithm scales linearly with the number of records. In addition, the proposed method avoids three of the main drawbacks presented by the rule mining algorithms: production of a high number of rules, discovery of uninteresting patterns and low performance.

FUTURE SCOPE

The prosposed work can be further expanded as another clustering method can be scaled up in different manner for mining more confident Association Rules form multidimensional quantitative dataset. The high scalable rules can be generated by using the other clustering methods such as principal component analysis.

REFERENCES

[1] María n. Moreno, Saddys Segrera, Vivian f. Lopez, M José polo, "A Method for Mining Quantitative Association Rules", Proceedings of the 6th WSEAS International Conference on Simulation, Modeling and Optimization, Lisbon, Portugal, September 22-24, 2006.

[2] Ramakrishnan Srikant, Rakesh Agrawal,"Quantitative Association Rules in Large Relational Tables", IBM Alma den Research Center Mining , ISSN 2309-45 vol 4 ,pg 34-67, 2006.

[3] Preetham Kumar, Ananthanarayana V S," Discovery of Multi Dimensional Quantitative Closed Association Rules by Attributes Range Method", Proceedings of the International MultiConference of Engineers and Computer Scientists 2008, Vol I IMECS 2008, pg 19-21 March, 2008, Hong Kong.

[4] Yiping Ke James Cheng Wilfred Ng , "An Information-Theoretic Approach to Quantitative Association Rule Mining " , Department of Computer Science and Engineering The Hong Kong University of Science and Technology Clear Water Bay, Kowloon, Hong Kong, ChinaIn Knowledge Discovery and Data Mining, pages 73-83, 1999.

[5] S.Prakash, R.M.S.Parvathi,"An Enhanced Scaling Apriori for Association Rule Mining Efficiency", European Journal of Scientific Research ISSN 1450-216, Vol.39 No.2, pg.257-264,2010.

[6] Agrawal R, Imielinski T, Swami," A. Database Mining: A performance Perspective", IEEE Trans. Knowledge and Data Engineering, vol. 5, 6, pg 914-925, 1993.

[7] Agrawal R., Imielinski, T. Swami, "A. Mining associations between sets of items in large databases", Proc. of ACM SIGMOD Int. Conference on Management of Data, Washington D.C., pg 207-216, 1993.

[8] R.Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", In Proceedings of the Association for Computing Machinery, Special Interest Group on Management of Data (ACM-SIGMOD), pg 207-216, May 1993.

[9] Coenen F., G. Goulbourne and P. Leng, "Tree Structures for Mining Association Rules", Data Mining and Knowledge Discovery, pg 25-51, 2004.

[10] Agrawal R, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases", Proc. of the ACM SIGMOD Washington, D.C, pg 207-216, May 1993.

[11] Grabmeier, J. and Rudolph, A., "Techniques of Cluster Algorithms in Data Mining", Data Mining and Knowledge Discovery, 6, pg 303-360, 2002.

[12] Agrawal R., Imielinski T., Swami, "A. Database Mining: A performance Perspective ", IEEE Trans. Knowledge and Data Engineering, vol. 5, pg. 914-925, 1993.

[13] Huang, Y.F., Wu, C.M, "Mining Generalized Association Rules Using Pruning Techniques", Proceedings of the IEEE International Conference on Data Mining (ICDM'02), Japan, pg 227-234, 2002.

[14] Imielinski T., A. Virmani and A. Abdulghani , " Application Programming Interface and Query Language for Database Mining", Proceedings ACM International Conference Knowledge Discovery & Data Mining, ACM Press, pg 256-261, 1996.

[15] Han J., Y. Cai, and N Cercone,"Data Driven Discovery of Quantitative Rules in Relational Databases", IEEE Trans Knowledge and Data Eng, Vol 5, pg 29-40, 1993.

[16] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In Proceedings of the 20th International Conference on Very Large Databases (VLDB), IIEEE, pages 290-297, 2002.

[17] Yu Wei , "Approximation to K-means-type Clustering", In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pg 306-315, 2003.

[18] Rakesh Agrawal, Ramakrishnan Srikant, "Algorithms for mining association rules in large databases", Proceedings of the 20th VLDB Conference Santiago, Chile, Vol -2, pg 141-182,1995.

[19] R. Agrawal and R. Srikant, "Mining sequential patterns" Proc. of 20th International Conference on Very Large Databases, Santiago de Chile, pg 487-489, 1994.

[20] Amir Netz, Surajit Chaudhuri, Jeff Bernhardt, Usama Fayyad," Integration of Data Mining and Relational Databases ", Proceedings of the 26th International Conference on Very Large Databases, Cairo, Egypt, 2000.

[21] R. Srikant and R. Agarwal, "Mining quantitative association rules in large relational tables ", In Proceedings of the Association for Computing Machinery, Special Interest Group on Management of Data, pg 1-12,1996.

[22] Agarwal R. and V. Prasad, "A Tree Projection Algorithm for Generation of Frequent Itemsets,"Parallel and Distributed Computing, 2000.

[23] Harish Verma, Eatesh Kandpal, Bipul Pandey, Joydip Dhar, "A Novel Document Clustering Algorithm Using Squared Distance Optimization Through Genetic Algorithms ", International Journal on Computer Science and Engineering Vol. 02, No. 05, pg 1875-1879, 2010.

[24] Heikki Mannila, Harmu Toivonen, and A. Inkeri Verkamo , "Efficient algorithms for discovering association rules", AAAI Workshop on Knowledge Discovery in Databases, pg 181- 192, Seattle, Washington, July 1994.

[25] Jong Soo, Park, Ming Syan Chen, and Philip S.Yu , "An effective hash based algorithm for mining association rules", Proceedings of the ACM-SIGMOD Conference on Management of Data, San Jose, California, May 1995.
[26] Agrawal R., Imielinski T., Swami A., "Mining associations between sets of items in large databases", Proceedings of ACM SIGMOD International Conference on Management of Data, Washington D.C., pg 207-216 , 1993.

Dr.Tamanna Siddiqui is a senior faculty in Computer Science Department,Jamia Hamdard, New Delhi.She has 14 years of teaching experience .She has published 21 reasearch papers in International/ national journal and conference proceedings. She has authored many books/chapters of MCA cource of Indira Gandhi Open University, New Delhi. She has delivered special lectures as a resource person at various academic institutions and conferences.Her research area includes data mining database, software engineering, soft computing and artificial intelligence.Many research scholars are doing research under her guidance in the same area.

Dr. M Afshar Alam is professor in Department of Computer Science, Jamia Hamdard,New Delhi.He has teaching experience of more than 17 years.He has authored 8 books and guided PhD research works.He has more than 30 publications in international/national journal and conference proceedings. He has delivered special lectures as a resource person at various academic institutions and conferences He is a member of expert committees of UGC,AICTE and other national and international bodies.His research areas include software re-engineering,data mining, bioinformatics and fuzzy databases.

Sapna Jain is a Phd Fellow in the Jamia Hamdard University who has obtained her MCA (Masters of Computer Application) degree from Maharishi Dayanand University, ndia. Her area of research is Scalability of data mining algorithms.

Call for Papers and Special Issues

Aims and Scope

JAIT is intended to reflect new directions of research and report latest advances. It is a platform for rapid dissemination of high quality research / application / work-in-progress articles on IT solutions for managing challenges and problems within the highlighted scope. JAIT encourages a multidisciplinary approach towards solving problems by harnessing the power of IT in the following areas:

- **Healthcare and Biomedicine** - advances in healthcare and biomedicine e.g. for fighting impending dangerous diseases - using IT to model transmission patterns and effective management of patients' records; expert systems to help diagnosis, etc.
- **Environmental Management** - climate change management, environmental impacts of events such as rapid urbanization and mass migration, air and water pollution (e.g. flow patterns of water or airborne pollutants), deforestation (e.g. processing and management of satellite imagery), depletion of natural resources, exploration of resources (e.g. using geographic information system analysis).
- **Popularization of Ubiquitous Computing** - foraging for computing / communication resources on the move (e.g. vehicular technology), smart / 'aware' environments, security and privacy in these contexts; human-centric computing; possible legal and social implications.
- **Commercial, Industrial and Governmental Applications** - how to use knowledge discovery to help improve productivity, resource management, day-to-day operations, decision support, deployment of human expertise, etc. Best practices in e-commerce, e-commerce, e-government, IT in construction/large project management, IT in agriculture (to improve crop yields and supply chain management), IT in business administration and enterprise computing, etc. with potential for cross-fertilization.
- **Social and Demographic Changes** - provide IT solutions that can help policy makers plan and manage issues such as rapid urbanization, mass internal migration (from rural to urban environments), graying populations, etc.
- **IT in Education and Entertainment** - complete end-to-end IT solutions for students of different abilities to learn better; best practices in e-learning; personalized tutoring systems. IT solutions for storage, indexing, retrieval and distribution of multimedia data for the film and music industry; virtual / augmented reality for entertainment purposes; restoration and management of old film/music archives.
- **Law and Order** - using IT to coordinate different law enforcement agencies' efforts so as to give them an edge over criminals and terrorists; effective and secure sharing of intelligence across national and international agencies; using IT to combat corrupt practices and commercial crimes such as frauds, rogue/unauthorized trading activities and accounting irregularities; traffic flow management and crowd control.

The main focus of the journal is on technical aspects (e.g. data mining, parallel computing, artificial intelligence, image processing (e.g. satellite imagery), video sequence analysis (e.g. surveillance video), predictive models, etc.), although a small element of social implications/issues could be allowed to put the technical aspects into perspective. In particular, we encourage a multidisciplinary / convergent approach based on the following broadly based branches of computer science for the application areas highlighted above:

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the "Call for Papers" to be included on the Journal's Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. "Special Issue: Selected Best Papers of XYZ Conference".
- Sending us a formal "Letter of Intent" for the Special Issue.
- Creating a "Call for Papers" for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at http://www.academypublisher.com/jait/.