Volume 2, No. 1, Jan-Feb 2011

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Efficient Data Mining Technique Using Associate Rule

Deepchnad Ahirwal
Department of Information Technology
Samrat Ashok Technological Institute
Vidisha, M.P. (India).
deepchand7818.ahirwal@gmail.com

Abstract- In recent years there has been a massive proliferation of data that is accessible electronically, ranging in form from highly structured databases, such as article collections and company records, to web sites and pages that vary greatly in content. Since its introduction, association rules mining technique became one of the most frequently used data-mining techniques. Association rules have exhibited an excellent ability to identify interesting association relationships among a set of binary variables describing huge amount of transactions. Although the technique can be relatively easily generalized to other variable types, the generalization can result in a computationally expensive algorithm generating a prohibitive number of redundant rules of little significance. This danger especially applies to quantitative (and ordinal) variables. These issues are tackled in this thesis and an alternative approach to the quantitative association rule mining is presented and verified. In the proposed approach, quantitative or ordinal variables are not immediately transformed into a set of binary variables. Instead, simple arithmetic operations are applied in order to construct the compound attributes and the algorithm further searches for areas of increased association which are finally decomposed into conjunctions of atomic attributes.

Keywords- Data Mining, Association Rule and Efficient Search Engine.

I. INTRODUCTION

At present, more and more databases containing large quantities of data are available. These industrial, medical, financial and other databases make an invaluable resource of useful knowledge. The task of extraction of useful knowledge from databases is challenged by the techniques called data-mining techniques. One of the widely used data-mining techniques is association rules mining.

Association rules mining identifies associations (patterns or relations) among database attributes and their values. It is a pattern-discovery technique which does not serve to solve classification problems (it does not classify samples into some target classes) nor prediction problems (it does not predict the development of the attribute values). Association rules mining generally searches for any associations among any attributes present in the database.

Association rule (AR) is commonly understood [1] as an implication X → Y in a transaction database D = $\{t_1 \ldots t_m\}$. Each transaction t_i □ D contains a subset of items I = $\{i_1 \ldots\ldots i_n\}$. X and Y are disjoint itemsets, it holds X; Y □ I and X ∩ Y = Φ. The left hand side of this implication is called antecedent, the right hand side is referred to as consequent. The transaction database D can also be viewed as a boolean dataset where the boolean values of attributes in records express occurrence of items in transactions.

Association rules mining algorithms are used to identify association rules. The first association rules mining algorithm APRIORI was described by Agrawal, Imielinski, and Swami in 1993. The algorithm based on antimonotonicity property of frequent itemsets became the standard and benchmark for its descendants. The original dataset used in [1] contained market basket data, ie. naturally boolean (binary, two-valued) attributes. Either an item (attribute) has been purchased by a customer and is in his/her market basket, indicated by a value of 1 (true), or it has not, indicated by a value of 0 (false). Rules generated by the APRIORI algorithm showed direct associations between the purchased items such as examples with toothbrushes and toothpastes. Association rule mining problem poses the question of efficiency. The number of potential rules X → Y defined by X □ Ix □ $\{ix_1;\ldots\ldots ; ix_n\}$, Y □ I_y □ $\{iy_1;\ldots. ; iy_m\}$, where Ix and Iy are disjoint, is equal to $2^{(m+n)}$. When general datasets are considered, the AR mining problem is known to be NP-complete. In restricted cases, for example in sparse boolean datasets (where it holds all ti □ D; |ti| <= O(log|I|)) lower complexity bounds have been proved to hold. Finding rules in quantitative data further strengthens importance of efficiency. An increase in the number of values that can be associated with any given variable increases the number of rules exponentially, thus causing execution time to increase significantly. There are different methods in the literature for handling quantitative variables and categorical values. Quantitative association rule mining algorithms commonly attack this problem by splitting the numeric values into discrete intervals, and converting a single variable with D intervals into D binary variables and solving the classical association rules mining problem then. These methods and algorithms for quantitative association rule mining presented so far in literature are researched in this thesis. Then, an innovative quantitative association rule mining algorithm is proposed.

The major goals of the thesis can be summarized in following points:

A. To research and summarize current work and literature related to the quantitative association rules (QAR) mining topic. On the basis of results and findings from the research, the main goal of this thesis is to propose and describe an innovative QAR mining algorithm. The proposed Quantitative Association Rules mining algorithm challenges some of the problems connected with QAR mining.

B. The main idea of the Quantitative Association Rules mining algorithm is in a trade-off between time complexity and completeness of the algorithm. The fundamental distinction from other approaches (algorithms) is that the Quantitative Association Rules mining algorithm is not based on the complete search of the state space, and therefore it cannot be guaranteed all valid rules are found. Algorithm uses new method for construction of compound condition and explores possibility of usage of genetic algorithms for searching of areas of strong associations.

The goal of the proposed approach is to gain two main advantages during the QAR mining:

C. Reduction of the time complexity as introduced before, the time complexity of QAR algorithms is one of the biggest problem connected with QAR mining, the QUARG algorithm significantly reduces the time complexity, so that the QAR mining can be used even on larger data domains with reasonable time costs.

D. Reduction of the number of redundant rules during the generation of quantitative association rules many redundant rules are generated, nevertheless this topic is not in the focus of attention of former or even current research, this thesis goal is to propose a definition of redundancy of quantitative association rules and to propose the Quantitative Association Rules mining algorithm which prohibits generation of redundant rules.

E. External or background knowledge, i.e. knowledge which is not directly present in the database from which the rules are generated, became more and more used during the process of rules mining (and knowledge discovery in databases in general). The association rules mining problem is formalized in the pure attribute-valued format. Employment of background knowledge enables to exploit resources outlying the transaction database. The next goal of this thesis is to explore possibilities of incorporation of background knowledge mechanisms into the QAR mining process and implementation into the Quantitative Association Rules mining algorithm.

F. Next goal of the thesis is validate and verify the functionality of Quantitative Association Rules mining algorithm on real-life datasets and to solve some of the practical task from the area of association rules mining.

II. BACKGROUND

The association rules were introduced and after the first association rules mining algorithms were described, the problem of mining association rules from databases with quantitative attributes emerged. QAR mining requires adaptation of original APRIORI approach. In this section the overview of QAR mining theory and overview of approaches, which were introduced so far are presented. First the focus will be on QAR mining in general, then in second part the focus will move toward the QAR mining for the genomic datasets. In the end of this section a brief summary of related work is provided. An essential preprocessing technique used by the majority of presented algorithms is the discretization. It is therefore suitable to begin this section with a short introduction to the discretization topic and introduce basic discretization techniques [1] and [3].

For each algorithm it is necessary to define character of input data, or data the algorithms can deal with. The data mining literature contains a variety of terms describing different types of data or attributes. In this work the following division is used:

[a] Qualitative attributes their values are only divided in categories, but not numerical Measures-

[b] nominal attributes, which values are exhaustively divided into mutually exclusive categories with no rankings that can be applied to these categories (names, colors).

[c] ordinal the categories into which the values are classified can be ordered (evaluation of an action = {very good, good, bad, very bad}).

[d] Quantitative attributes attributes that are measured on a numerical scale and to which arithmetic operations can be applied.

[e] discrete have a measurement of scale composed of distinct numbers with gap in between (number of cars).

[f] continuous can ideally take any value (height, distance).

Association rules mining (AR mining) algorithms work with and are optimized for nominal attributes. They can be quite easily generalized to work with ordinal attributes. Generalization of AR algorithms for quantitative attributes is possible, but not less straight solution. QAR mining algorithms were introduced to work with quantitative attributes. Most of the QAR algorithms rely upon preprocessing of the quantitative attributes.

The preprocessing follows this main objective: to reduce a potentially infinite number of values of the quantitative attributes. For preprocessing the discretization and consecutive mapping are used [2].

A. Quantitative association rules Overview-

Recently, different aspects of the quantitative association rules mining problem have been studied. Several approaches have been presented in the literature, authors mostly agree that standard AR mining algorithms used for quantitative variables or attributes are ineffective and often provide us with redundant or illogical rules.

Quantitative association rules mining problem was first introduced in [59]. An example of quantitative association rule can be "people between 50-60 year of age have at least 2 cars": <Age : 50:60> → <Num_Cars: 2…..>. The authors deal with quantitative attributes by fine-partitioning; they combine adjacent partitions as necessary. The process of generating quantitative rules consists of these steps:

[a] setting the number of intervals for quantitative attributes and following discretization (partitioning),

[b] mapping quantitative attributes into the boolean attributes,

[c] combining adjacent intervals, counting support, gaining all frequent itemsets,

[d] using all frequent itemsets for rules generation,

[e] interesting rules selection.

The authors map the quantitative association rules problem to the classic association rules problem - instead of having just one field in the table for each attribute, they have as many fields as the number of attribute values. The most suitable discretization form is discussed to avoid the problems with execution time and too many generated rules. There is a trade -off between faster execution time with fewer intervals and reducing information loss with more

intervals. The information loss can be reduced by increasing the number of intervals, at the cost of increasing the execution time and potentially generating many uninteresting rules.

The partial completeness measure is presented to handle the amount of the information loss by partitioning (discretization). The intuition behind the partial completeness is as follows: let R be the set of rules obtained by considering all the ranges over the raw values of quantitative attributes. Let R' be the set of rules obtained by considering all the ranges over the partitions of quantitative attributes. The way to measure the information loss when we go from R to R' is to see for each rule in R how 'far' the 'closest' rule in R' is. The further away the closest rule, the greater the loss. By defining close rule to be generalizations, and using the ratio of support of the rules as a measure of how far apart the rules are, the measure of partial completeness is derived.

Quantitative data are dichotomized using simple thresholds as the basis for boolean classification. The dichotomized data are used in association rule algorithms to find interesting rules and patterns in this data. Once significant associations are found, the dimensionality on the selected interesting variables can be increased. Boolean Analyzer is used as an association rule algorithm, to look at rules. Finding significant rules is strongly dependent on how thresholds for booleanization are defined. Authors are discussing several statistical measures to identify the thresholds and dependency of itemsets. The probabilistic interestingness measure is used to identify interesting rules.

The conclusions are following:

[i] Expert, mean and median boolean thresholds can find known rules.
[ii] Mean and median thresholds may yield more accurate results than thresholds from an expert.
[iii] There are problems with mode as an automated method.
[iv] Rules found using all three methods correlate very well with regard to their PIM values and hence their ordering [5] and [6].

In recent, introduced a new definition of quantitative association rules based on statistical inference theory. This definition reflects the intuition that the goal of association rules is to find extraordinary and therefore interesting phenomena in databases. An association rule indicates an association between a subset of the population, described by the left-hand side of the rule, and an extraordinary behavior of this subset, described by the right-hand side of rule, thus, the general structure of an association rule is: populationsubset → extraordinarybehavior. The best way to describe the behavior of some population is by describing its probability distribution (or at least, by providing important information about this distribution). The concept of sub-rules is introduced, which can be applied to any type of association rule.

Impact rules detect useful interactions between combinations of categorical and numeric variables. Impact rules are characterized as follows: a training set is a finite set of records, where each record is an element to which we apply Boolean predicates called conditions, and which is associated with a numeric value called the target. An impact rule consists of a conjunction of conditions, called the antecedent, and one or more statistics, called the consequent, describing the impact on the target of selecting the training set records that satisfy the antecedent. Impact rule analysis may seek a finite number of impact rules that individually optimize some function of quality, usually one of the rule statistics. The author presents an alternative strategy for impact rule discovery, utilizing the OPUS search algorithm, that has lower computational requirements than a frequent itemset approach and which avoids in many cases the requirement that arbitrary constraints (namely, minimum cover) imposed on the considered rules. Using OPUS for impact rule discovery, search may be constrained to find the top n impact rules on some measure, and pruning may remove branches that cannot lead to an impact rule that satisfies that constraint. For many measures, this constraint is sufficient to provide efficient search. Where the measure does not facilitate effective pruning, additional constraints, such as minimum cover, can be employed.

Optimized association rules are permitted to contain uninstantiated attributes and the problem is to determine instantiations such that either the support or the confidence of the rule is maximized. To generalize optimized association rules problem in three ways:

[a] association rules are allowed to contain disjunctions over uninstantiated attributes,
[b] association rules are allowed to contain an arbitrary number of uninstantiated attributes,
[c] uninstantiated attributes can be either categorical or numeric [7] and [9].

A new kind of rule, ordinal association rule, which extracts implications between conjunctions of any type of attributes. These rules are based on an objective measure, intensity of inclination, which evaluates the 'smallness' of the number of transactions which contradict the rule. This measure prunes out the transformation step of data (i.e. the discretization step of numeric attributes and the step of complete disjunctive coding) thereby avoiding obtaining a prohibitive number of rules which have little significance and have many redundancies. Ordinal association rules extracted by this measure reveal the overall behavior of transactions in the database. The author focuses on the technique for mining of specific rules based on extracted ordinal association rules in order to, on the one hand remove the transformation step of initial attributes, and on the other hand obtain a variable discretization of numeric attributes i.e. dependent on association of attributes. This technique is split into two steps, each step going towards a more significant degree of specialization. The first step allows us to obtain specific ordinal association rules from ordinal association rules and, the second to extract association rules from specific ordinal association rules. Moreover, this technique is particularly suitable for sparse data since it does not seek frequent itemsets and then does not use support.

In previous provide an algorithm for quantitative association rules mining. This algorithm is more time consuming than standard algorithms. Authors also do not address the problem of attributes combination in detail. Data mining procedure KL-Miner mines for patterns based on evaluation of two dimensional contingency tables. Data mining procedure KL-Miner mines for patterns of the form R C/Cond. Here R and C are categorical attributes, the attribute R has categories r_1, …… r_K, the attribute C has categories c_1, ……. C_L. Further, Cond is a Boolean attribute. The procedure deals with data matrices. The attributes R and C correspond to columns of the analyzed data matrix.

Boolean attribute Cond is derived from the other columns of the data matrix. The intuitive meaning of the pattern R C/ Cond is that the attributes R and C are in relation given by the symbol when the condition given by the Boolean attribute Cond is satisfied. The symbol is called KL-quantifier. It corresponds to a condition imposed on the contingency table of R and C. The pattern R C / Cond is verified on the contingency table of R and C in data matrix M / Cond. Here M is the analyzed data matrix and M / Cond is a data matrix consisting of all rows of M satisfying Cond. There are various KL-quantifiers, some of them correspond to simple conditions concerning frequencies, and some others are of statistical nature. Association rules mining in preference ordered data, i.e. in databases with quantitative attributes, is introduced. The authors call attributes in these databases criteria and are stressing the importance of semantic correlation between criteria. For example, in decision about credit granting, where two criteria are considered, "month salary" and "evaluation of bank risk concerning payment of a credit", these criteria are semantically correlated in the following sense: an improvement of month salary should not deteriorate the evaluation of bank risk. The paper introduces consideration of criteria in association rules. Some of basic concepts have been inspired by dominance-based decision rules of multiple criteria classification problems. In multiple criteria classification problems criteria are semantically correlated with preference ordered decision classes. The algorithm proposed by authors has higher time consumption than original APRIORI algorithm. The author used fuzzy concepts for generation of quantitative rules. Two types of rules are presented: deviation rules and tendency rules.

The former type of rule is basically a fuzzy counterpart to the approach in [7]. The latter type of rule is able to represent gradual dependencies between attributes. This becomes possible through the use of fuzzy partitions for the attributes' domains.

Genetic algorithms are used for QAR generation to overcome the problem of discretization. Using this approach authors avoid the discretization phase - most suitable partitions are found by genetic algorithms. An evolutionary algorithm is used to find the most suitable amplitude of the intervals that confirm a k-itemset, so that they have a high support value without the intervals being too wide. This approach optimizes support of the rules, however the confidence is not addressed, and therefore rules with high confidence can be missed [8] and [10] and [11].

B. Quantitative association rules for genomic datasets-

Currently, large quantities of gene expression data are generated and made available in publicly accessible databases. Data mining and automated knowledge extraction from this kind of data belong to the major contemporary scientific challenges. Genomic datasets (gene expression data) contain lots of quantitative attributes, hence the quantitative association rules mining algorithms are suitable for data mining tasks on this kind of data.

Though, regarding association rules, genomic datasets data represent a difficult mining context. First, the data is high-dimensional which asks for an algorithm scalable in the number of variables. Second, as the expression values are typically quantitative variables, computational demands increases and may result in output with a prohibitive number of redundant rules. Third, the data is often noisy which may also cause a large number of rules of little significance. For data-mining tasks in this kind of datasets, clustering is one of the most often used methods the most similar genes are found so that the similarity among genes in one group (cluster) is maximized and similarity among particular groups (clusters) is minimized. Although very good results are gained by this method there are three main drawbacks:

[a] One gene has to be clustered in one and only one group, although it functions in numerous physiological pathways.

[b] No relationship can be inferred between the different members of a group. That is, a gene and its target genes will be co-clustered, but the type of relationship cannot be rendered explicit by the algorithm.

[c] Most clustering algorithms make comparisons between the gene expression patterns in all the conditions examined. They therefore miss a gene grouping that only arises in a subset of cells or conditions.

The need for adaptation of original APRIORI based algorithms for mining rules with quantitative attributes was identified shortly after the APRIORI algorithm was introduced. Respecting the orderliness of quantitative values was proved as necessary. Discretization is one of the first issues mentioned in almost all works dealing with quantitative attributes and quantitative rules mining. Discretization of the quantitative values is understood as a tool for reduction of theoretically infinite time complexity of QAR mining algorithms, as the number of quantitative values can theoretically be infinite. Discretization for quantitative association rules mining purposes is firstly described and the authors used and recommended simple and in many cases unsuitable equip depth discretization followed by a decomposition of one quantitative attribute into several binary attributes. This leads to a significant increase in time complexity. In some further works more convenient distance-based approach to discretization is used.

Some authors just reduced the quantitative values to the boolean values. Exactly contrary approaches completely omit the discretization step with the argument that discretization causes the information loss. But the rules generated with such method are different from the classical rules and their understandability is questioned. New measures of quality for quantitative association rules are proposed. The authors point out that quantitative rules require new semantical understanding and current measures of quality are therefore unsuitable. Statistical values, as mean value, are used. None of the measures which were specially proposed for quantitative rules become widely accepted and used. In most works and experiments presented in literature, the measures based on classical measures as support, confidence and lift are used. Time consumption is another important and discussed issue. As the number of possible attributes values grows the time complexity of QAR mining increases exponentially. The time consumption of quantitative association rules mining. It is even more significant issue when taking into account the genomic datasets. These datasets contain thousands of attributes, therefore the question of time costs be-comes crucial. To my best knowledge there is no algorithm described in the literature so far exhibiting significant reduction of time consumption when compared to the complete search approach and simultaneously posing no limitation on the form and character of generated rules.

Discussion of the possible trade-off between the completeness and the time costs of QAR mining technique is missing in current literature [1], [11] and [12].

III. PROPOSED TECHNIQUE

On the basis of the state-of-the-art research the innovative algorithm for quantitative association rules mining is proposed. Generally the algorithm consists of these basic parts (steps):

[a] ***Data preprocessing-*** quantitative values of attributes are discretized, their values are mapped to the consecutive row of integers beginning with 1, time complexity is thus reduced and also values of attributes are normalized.

[b] ***Identification of areas of strong associations -*** the algorithm identifies areas, where strong associations between antecedent and consequent are expected, searched space is reduced in this step, association rules are searched only in these areas of strong associations, the rest of state space is not searched by the algorithm.

[c] ***Decomposition of areas of strong associations -*** each area is decomposed into a set of antecedent or consequent atomic conditions (triplets), these conditions are combined into a set of candidate rules.

[d] ***Candidate rules verification -*** every rule from the set of candidate rules is verified, the rules which satisfy defined quality and interestingness measures thresholds are marked as valid rules.

A. Preprocessing of Atomic Attributes-

The input of the quantitative association rules algorithm is a database DB containing N attributes with values from a quantitative domain D_q. Particular attributes a_1, a_2, ..., a_N present in the database are called raw atomic attributes. Values of raw atomic attributes need to be preprocessed before start of the rules mining. During the preprocessing phase the values of each raw atomic attribute are discretized. Discretized values of each attribute are further mapped on consecutive row of integers, beginning with 1. Number one represents the lowest values and D the highest values, where D is the degree of discretization for particular attribute, ie. the number of discretization bins into which the quantitative values are discretized. As a result of the preprocessing phase we receive the set of preprocessed attributes p_a_1, p_a_2... p_a_N. Each preprocessed atomic attribute takes values $v \in \{1, Dn\}$; where Dn is the degree of discretization of n-th attribute ($n \in \{1, N\}$).

B. Atomic and Compound Attributes-

An association rule X → Y consists of an antecedent (left-hand side of a rule) and consequent (right hand side of a rule). There should be no principal limitation in the number of attributes on antecedent or on consequent side of the rule, i.e. antecedent (consequent) of a rule can consist of one attribute or a combination of several attributes. We have to deal with the situations:

[a] Antecedent (consequent) attribute is an atomic attribute - the value of antecedent (consequent) attribute is equal to the value of the atomic attribute.

[b] Antecedent (consequent) attribute is a compound attribute - we have to combine values (vectors) of two or more atomic attributes and represent them by one vector.

C. Time Complexity-

Assume we have a total number of N attributes and M records in a database. Then N is also the maximum number of atomic attributes and time complexity for atomic attributes creation is O (N X M). The total number of subsets (combinations) of atomic attributes Natrib_comb is given by

$$N_{atrib\_comb} = \sum_{k=1}^{K_{max}} \binom{N}{k}$$

A graphical overview of time complexity of a complete search approach is provided in Figure below.

Time complexity of the complete search approach is determined by the total number of verifications of candidate rules and number of records in a database. The total number of verifications of candidate rules is the product of the number of possible antecedent-consequent combinations N_{ant_cons} and the number of verifications between one antecedent-consequent combination $N_{verif\_iant\_cons}$.

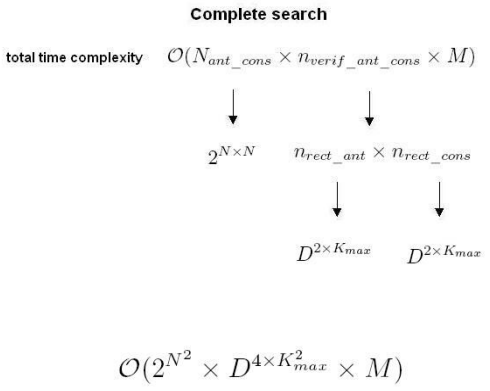

Figure: 1 Time complexity of complete search approach

D. Redundancy of Quantitative Association Rules-

The notion of redundancy is used to verify whether incompleteness of the quantitative association rules (QUARG) algorithm is rather a positive or negative feature. In other words, checks whether the QUARG algorithm reduces redundancy or rather looses non-recurring rules. Redundancy of rules generated by the QUARG algorithm and redundancy of rules generated by the complete search approach is also compared. Experiments performed on several datasets are presented to test incompleteness of the QUARG algorithm.

The definition of redundancy for quantitative association rules showed that redundancy of set of rules generated by the QUARG approach is significantly lower than redundancy of set of rules generated by the complete search approach. For datasets, from which a high numbers of rules are generated (thousands and tens of thousands rules), over 90% of rules generated by complete search are redundant rules. At the same time, the QUARG algorithm generates 20-30% of redundant rules.

The input of the algorithm is a quantitative database with attributes a_1, a,……a_N. These attributes are called atomic attributes. The QUARG algorithm consists of the following basic steps:

[a] Pre-processing of attributes: values of all atomic attributes are discretized and mapped to consecutive row of integers beginning with 1, 1 represents the lowest value of an atomic attribute. For this phase of algorithm the inspiration arises and the idea of mapping the discretized values of quantitative attributes on the consecutive row of integers. For the discretization step where this important fact is stated: distance among particular records in 2-D space is important during quantitative rules mining. Authors introduced Distance based association rules and they identified clusters of records. The QUARG algorithm uses k-means discretization. The output of this step is a set of preprocessed atomic attributes $a_1, a_2 \ldots a_N$.

[b] Construction of compound attributes: both antecedent and consequent of a rule consist of a compound attribute. Compound attribute is a combination of one or more preprocessed atomic attributes. For next algorithm steps, it is necessary to represent the values of compound attributes by a single number, despite they are composed of several atomic attributes. Generally it is the projection from N-D space to 1-D space described by Formula 1, where N is the number of atomic attributes in a compound attributes.

[c] Areas of interest identification and decomposition: Areas of increased associations between antecedent and consequent are identified in difference matrix. Genetic algorithms are used to identify the areas of interest. The areas of interest are then decomposed and the candidate rules are extracted. Main inspiration for this algorithm step is taken, where the idea of contingency tables of differences is presented. The areas of interest border the state space from which the candidate rules are extracted, so the reduction of time complexity of the QUARG algorithm takes place in this step. Output of this step is a list of candidate rules.

[d] Verification of candidate rules: Candidate rules gained by the decomposition of areas of interest are verified in the last step of the algorithm. Minimum confidence, minimum support and minimum lift thresholds are set up for elimination of weak rules.

IV. RESULTS

Using the QUantitative Association Rule mininG (QUARG) approach the number of verifications is approximately 10% of complete search approach. For the limit of two atomic attributes in one compound attribute the results are displayed in Figure . Time consumption of complete search approach is determined practically explicitly by the number of verifications, but QUARG approach has some other minor contributors to the time consumption beside the candidate rules verification step. The candidate rules verification takes in 85% of the time consumption, aside from this step, the biggest contributor is identification of the areas of interest by genetic algorithms (15% of the time consumption), the ratio presented. Time consumption of the QUARG algorithm is significantly lower even if we take into account the other steps of the algorithm.

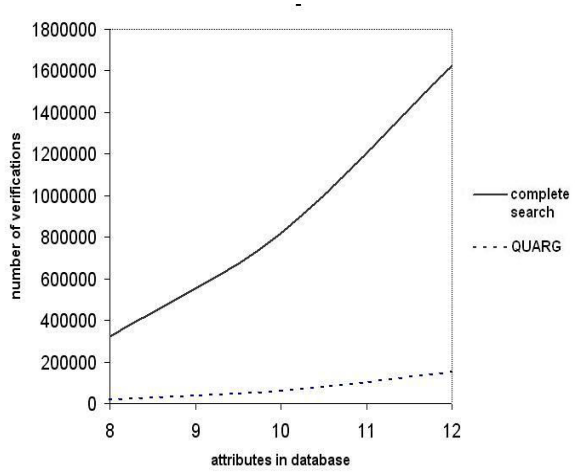

Figure: 2 Number of verifications for maximum number of 2 atomic attributes on antecedent and consequent side of rule.

The dependencies of the time complexity on three parameters: number of attributes in a database, maximum number of atomic attributes in a compound attribute and the degree of discretization.

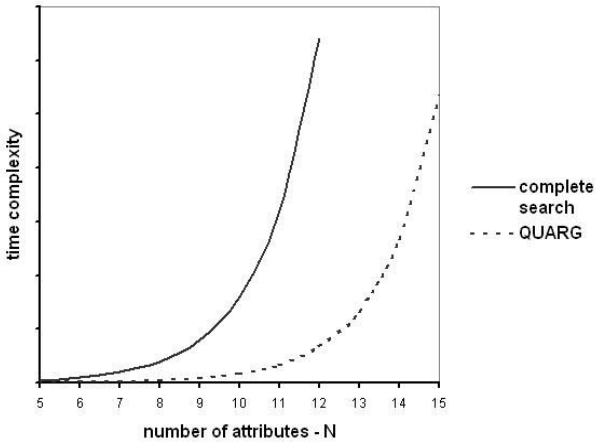

Figure: Time complexity of Association Rule mining, with fixed parameters Kmax = 2, D = 5 and M = 1000.

V. CONCLUSION

In this thesis, The innovative algorithm for quantitative association rules mining (QUARG) was proposed and described. The thesis completely described the proposed algorithm preprocessing, atomic and compound attributes construction, areas of interest identification and their following decomposition into candidate rules. Its main innovations are: in the way it constructs compound attributes (compound conditions), usage of genetic algorithms for areas of strong associations identification, decomposition of antecedents and consequents into rules.

The goal of proposal of a new QUARG algorithm was to challenge two major problems connected with QAR mining:

A. time complexity high time complexity of QAR mining algorithms protect their full application on larger datasets, the main idea of the proposed QUARG algorithm lies in significant reduction of the time complexity, QUARG algorithm has significantly lower

time cost when comparing to standard algorithms, the searched space (and the time complexity) is reduced during the phase of focusing on areas of interest, these areas of interest bound the searched space and candidate rules are generated only from inside of these areas, the rest of state space in not searched, for identification of areas of interest the genetic algorithm technique is used.

B. redundant rules lots of redundant rules are generated during the process of QAR mining, nevertheless this topic is not in the center of attention of current research works, in this thesis the definition of redundancy of quantitative association rules is provided, the notion of redundancy is used to demonstrate that incompleteness of the proposed QUARG algorithm more likely causes decrease in redundancy rather than omission of valuable rules.

VI. REFERENCES

[1] R. Forsati, M. R. Meybodi and A. Ghari Neiat,"Web Page Personalization based on Weighted Association Rules", IEEE 2009 International Conference on Electronic Computer Technology.

[2] Khalid Iqbal and Dr. Sohail Asghar, "Generating Hierarchical Association Rules with the Use of Bayesian Network", IEEE 2009 Third International Conference on Network and System Security.

[3] M. Karaolis, Student Member, IEEE, l.A. Moutiris, FESC, L. Papaconstantinou," AKAMAS: Mining Association Rules Using a New Algorithm for the Assessment of the Risk of Coronary Heart Events", Proceedings of the 9th International Conference on Info~ation Technology andApplications in Biomedicine, ITAB 2009, Larnaca, Cyprus, 5-7 November 2009.

[4] Prof Thivakaran.T.K, Rajesh.N, Yamuna.P, Prem Kumar.G, "PROBABLE SEQUENCE DETERMINATION USING INCREMENTAL ASSOCIATION RULE MINING AND TRANSACTION CLUSTERING", IEEE 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies.

[5] Zhongmei Zhou, "Mining Strongly Associated Rules", IEEE 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery.

[6] M. Atzmueller, F. Puppe, and H-P. Buscher. Exploiting background knowledge for knowledge-intensive subgroup discovery. In International Joint Conference on Artificial Intelligence, pages 647_652, Edinburgh, Scotland, 2005.

[7] Y. Aumann and Y. Lindell. A statistical theory for quantitative association rules. Journal of Intelligent Information Systems, 20:255_283, 2003.

[8] J. Baumeister, M. Atzmüller, and F. Puppe. Inductive learning for case-based diagnosis with multiple faults. In ECCBR '02: Proceedings of the 6th European Conference on Advances in Case-Based Reasoning, pages 28_42, London, UK, 2002. SpringerVerlag.

[9] R.J. Bayardo and R. Agrawal. Mining the most interesting rules. In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 145_154, New York, NY, USA, 1999. ACM.

[10] C. Becquet, S. Blachon, B. Jeudy, J-F Boulicaut, and O. Gandril. Strong-associationrule mining for large-scale gene-expression data analysis: a case study on human SAGE data. Genome Biology, 3:531_537, 2002.

[11] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J.M. Carazo, and A. Pascual-Montano. Integrated analysis of gene expression by association rules discovery). BMC Bioinformatics, page 7:54, 2006.

[12] J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In Proceedings of the Twelfth International Conference on Machine Learning, pages 194_202, Tahoe City, CA, 1995.