DRP Report: Quantitative Association Mining From Bottom Up and Heuristic Search Perspectives

Hao Wang

Committee Chair: Dejing Dou
Committee Member: Stephen Fickas, Daniel Lowd, Allen Malony, Christopher Wilson

November 20, 2013

Abstract

The traditional association mining focuses on discovering frequent patterns from the categorical data, such as the supermarket transaction data. The quantitative association mining (QAM) is a nature extension of the traditional association mining. It refers to the task of discovering association rules from quantitative data instead of from categorical data. The discrepancies between the two types of data lead to different analytical methods and mining algorithms. Several properties and interestingness measures that play important roles in the traditional association mining do not apply anymore in the quantitative situation. In this paper, we propose two quantitative association mining algorithms from the bottom up and heuristic search perspectives respectively. They take two new interestingness measures, density and correlation, which are better fits for the quantitative situation. The algorithms can find strong correlated intervals in a generally less correlated environment. Experiment results from neuroscience and health social network data validate the feasibility of our algorithms.

1 Introduction

The traditional association mining focuses on discovering the interesting patterns or rules from large categorical datasets. The categorical data is the data type with domain of limited possible values for each attribute, such as the supermarket transaction data in Table 1(a). The first association mining algorithm, Apriori algorithm, was proposed in 1993 by Agrawal and Srikant [15] and was later extended by many other researches [8] [10][11][16][19]. These algorithms dedicate to discover association rules based on user defined support and confidence thresholds. For categorical data, the association rule is an expression in the form of $X \Rightarrow Y$ [s,c], in which X and Y are sets of items. An item

(a) Categorical Data: Supermarket Transaction Data

Transaction ID \ Item	Milk	Bread	Egg	Wine	Fish	Vegetable
T1	NO	NO	NO	YES	NO	NO
T2	YES	YES	NO	NO	NO	NO
T3	YES	YES	NO	NO	YES	NO
T4	YES	YES	NO	YES	YES	NO
T5	NO	YES	YES	NO	YES	YES

(b) Quantitative Data: Temperature

Month \ City	Temperature Hawaii	Temperature L.A.	Temperature S.D.
Jan	72.9	56.8	57.4
Feb	73.0	58.6	58.6
Mar	74.4	59.6	59.6
Apr	75.8	62.0	62.0
May	77.5	64.1	64.1
Jun	79.4	66.8	66.8
Jul	80.5	71.0	71.0
Aug	81.4	72.6	72.6
Sept	81.0	72.6	71.4
Oct	79.6	71.4	67.7
Nov	77.2	67.7	62.0
Dec	71.4	62.0	57.4

Table 1: Example of Categorical and Quantitative Data

is a possible value in the domain of attribute, such as Milk = YES in Table 1(a) for the supermarket transaction data. The s and c stand for the support and confidence measures respectively. The support is a measure of the occurrence frequency of involved items in the association rule. For association rule $X \Rightarrow Y$, the $Support(X \Rightarrow Y)$ is defined as:

$$Support(X \Rightarrow Y) = P(XY) = P(X \cap Y).$$

The Confidence is a measure of quality or precision of the association rule. For association rule $X \Rightarrow Y$, it represents the probability of set Y given the condition that X happens. It is defined as

$$Confidence(X \Rightarrow Y) = P(Y|X) = P(X \cap Y)/P(X).$$

A concrete example of association rule using the data in Table 1(a) is, $Bread = YES \Rightarrow Milk = YES$ $[60\%, 75\%]$. It indicates the fact that "60% of customers who come to a supermarket buy both milk and bread, and among all the customers who buy bread, 75% of them also buy milk." In this example, the 60% and 75% stand for the support and confidence measures respectively. Following the idea of association mining, quantitative association mining (QAM) was proposed by Srikant et al. in 1996 [16]. One significant difference between QAM and traditional association mining is the output of QAM is in the form of "frequent" intervals intead of frequent itemsets. For example, one quantitative association rule for the meteorological data in Table 1(b) is

$$Temperature\ Hawaii[70\ F^o, 80\ F^o] \Rightarrow Temperature\ S.D.[60\ F^o, 70\ F^o], [42\%, 56\%],$$

in which $[70\ F^o, 80\ F^o]$ and $[60\ F^o, 70\ F^o]$ are two intervals defined on the attributes *Temperature Hawaii* and *Temperature S.D.* respectively. This rule shows that "for 42%

days of a year, the temperatures at Hawaii and S.D. fall into the ranges [70 F^o,80 F^o] and [60 F^o, 70 F^o] respectively, and while the temperatures of Hawaii belongs to [70 F^o, 80 F^o], 56% days of the temperatures at San Jose fall into [60 F^o, 70 F^o]. The 42% and 56% in the example above are the support and confidence measures respectively.

Although the results from these two types of association rule mining looks similar, the problem formulation and data mining process are nevertheless quite different. Since the domain of attribute in categorical data take only limited number of possible values, the values repeat themselves constantly in different data tuples. Traditional association mining algorithms rely on the repetition of data values in the sense that the initialization of support and confidence measures is based on a counting process on those frequently repeated items. This process, however, does not apply anymore in the quantitative situation since quantitative data seldom or never repeat themselves. For quantitative data, if we apply the traditional association mining on quantitative data directly, the counting process will mostly returns items with support counts exactly one; therefore no meaningful association rules can be discovered. Moreover, in quantitative association mining the interestingness measures support and confidence do not serve as useful and accurate measures either. As the output of QAM is in the form of "frequent" intervals, any set of data values could fall into one interval by a series of consecutive adjacency. Statistically the larger intervals an association rule have, the larger support and confidence it will result in. In this situation, the full range association rule, for example,

$$\begin{aligned} &Temperature\ Hawaii[73.0\ F^o, 81.4\ F^o] \\ &\Rightarrow Temperature\ S.D.[57.4\ F^o, 72.6\ F^o], [100\%, 100\%], \end{aligned} \tag{1}$$

will be the always and only best output. This rule has both the maximum support and confidence, however, it is trivial and does not convey any useful information.

The failure of support and confidence measures in QAM urges the needs of new interesting measures. In this paper we introduce two new interestingness measures: density and correlation that better fit with the QAM problem. The density in our paper is defined as the average number of data instances over unit length interval. This measure evaluates the occurrence frequency of data values under unit length interval. It represents the significance of a quantitative association rule in the way of how frequently the data conform with the association rule. For example, considering the two association rules below from Table 1(a),

$$Temperature\ Hawaii[70\ F^o, 80\ F^o] \Rightarrow Temperature\ S.D.[55\ F^o, 65\ F^o],$$

$$Temperature\ Hawaii[80\ F^o, 90\ F^o] \Rightarrow Temperature\ S.D.[65\ F^o, 75\ F^o],$$

they have intervals with equal length on both attributes. However, there are seven data instances that conform with the first rule while only three conform with the second one. This fact implies that data are more likely to fall into the first rule rather than the second one within unit length of interval. The first rule therefore covers data more efficiently and represents pattern with better quality. The correlation in this paper is defined by the Spearman's rank correlation coefficient [6]. The rules with strong correlated intervals

reveal useful information from a different perspective. The data in strong correlated intervals not only happen together, but also associate with unanimous trends of raising and falling. The association rules with high correlation measure are very useful in practice. For example, in the stock market, the demand of stocks and bonds generally raises as the price falls. This market principle of price and demand is the corner stone of a stable financial market. However, under certain circumstance such as a potential economic crisis, when the price falls below certain threshold and crash in the stock market is triggered, the demand would decrease together with the decline of the price for certain price range. In this example, the association rules from historical transaction data with either positive or negated correlation on certain price ranges will provide the investors with useful information on when and how to reduce risk in the financial investment.

There are various methods to sovlve the QAM problem and they generally fall into three perspectives, top down, bottom up and heuristic search respectively. In the top down perspective, domain knowledge or data statistics are used to assist the data mining process. The domain knowledge and data statistics can be used to simplify the data mining model or guide the path of the data mining process. In the bottom up perspective, the data mining problems are initialized as a set of sub problems in which the interestingness measures are trivially computable from data. The output of the QAM is then constructed as union of these sub problems by various data mining algorithms. In the heuristic search perspective, the problem is initialized as a set of results with inferior or non optimal interestingness measures, then these interestingness measures are iteratively improved by heuristic search algorithms till global optimal is reached.

In this paper, we propose two algorithms for the QAM problem from two of these three perspectives, bottom up and heuristic search, respectively. In the bottom up perspective, we propose to model the quantitative data with a hypergraph representation and use the average commute time distances to approximate the interestingness measures we used, density and correlation. In the heuristic search perspective, we propose an iterative crossover schema on genetic algorithm to reduce the heuristic search complexity and an optimization procedure after crossover to increase the accuracy of each iterative step in the genetic algorithm. We have also designed top down algorithm which deploys the domain knowledge such as Ontology to enhance QAM process, however, the experiment results for this perspective are not of satisfactory quality. This is primarily because we do not have suitable Ontologies in hand which fit with the need of the QAM problem. The research and discussion from the top down perspective is left for future work.

The rest of this paper is organized as follows: We give a brief selection of related works in section 2. We present the problem definition about the data representation and interestingness measure in section 3. We make a detailed description of our method from both the bottom up and heuristic search perspective in section 4. We report experiment results in 5. We discuss the future work of QAM in section 6 and conclude the paper in section 7.

2 Related work

Quantitative association mining is an intriguing and promising problem since after all 90% of our real life data is quantitative [17] such as the stock market price data, wireless sensor data and meteorological data. In previous research, various quantitative association mining algorithms have been proposed from different perspectives. Fukuda et al. [19][20] proposed several methods that either maximize the support with predefined confidence or maximize the confidence with predefined support by merging up adjacent instance buckets. Srikant and Agrawal [16] dealt with quantitative attributes by discretizing quantitative data into categorical data. From the theoretical perspective, Aumann and Lindell [21] in 1999 introduced a new definition of quantitative association rules based on the distribution of the quantitative data using statistical inference. Wijsen and Meersman [9] analyzed the computational complexity of mining quantitative association rules. Fuzzy set theory was introduced to the quantitative association mining problem to deal with the crisp boundary problem. Kuok et al. [13] proposed a fuzzy quantitative association rule mining algorithm which introduces the fuzzy set theory and generates fuzzy intervals instead of crisp intervals. Delgado et al. [14] introduced the fuzzy set theory into quantitative data mining by defining fuzzy sets on the domain of quantitative attributes. In these algorithms above, the algorithms that use discretization method suffer from information loss during the discretization and usually suffer from the *catch-22* problem [16] and the crisp boundary problem [21] as well. The algorithms using fuzzy methods can generate rules with fuzzy boundaries but they also require the help of domain experts to define the fuzzy concepts on the quantitative intervals, this task could be time consuming and subjective, especially when the set of attribute is large.

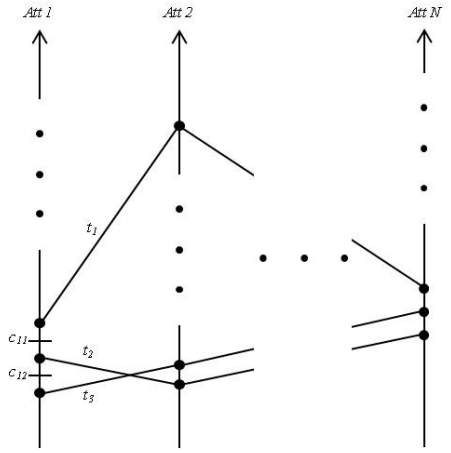

Figure 1: Quantitative Data Representation

3 Problem Definition

3.1 Data Representation

As illustrated in Figure 1, it shows an example of quantitative data with values sorted in ascending order on attributes. Let $A = \{Att_1, Att_2, ..., Att_M\}$ be the set of attributes, in which M is the number of total attributes. The domains of these attributes are either numerical or categorical. Let $I = \{Inst_1, Inst_2, ..., Inst_N\}$ be the set of data instances, in which N is the number of total data instances. The value of the ith attribute Att_i on the jth data instance $Inst_j$ is denoted as D_{ij}. The cutting points $c_{i1}, c_{i2}, ..., c_{i(N-1)}$ are the averages of each two adjacent values on corresponding attribute att_i. The interval boundaries of the QAM rules are defined on these cutting points. The set of attributes, instances and intervals for an association rule are denoted as S_{Att}, S_{Inst} and S_{Inter} respectively. A QAM association rules R is in the form $R = \cup_{Att_i \in S_{Att}} Att_i[c_{il}, c_{ih}]$, [density d, correlation c].

3.2 Interestingness Measure

As mentioned briefly in section 1, the support and confidence are not proper interestingness measures for the QAM problem. In this paper, we propose two interestingness measures, density and correlation, to replace the role of support and confidence in QAM. The goal of our quantitative association mining is therefore to discover both dense and strongly correlated intervals. The definitions of these two measures are presented in the current section.

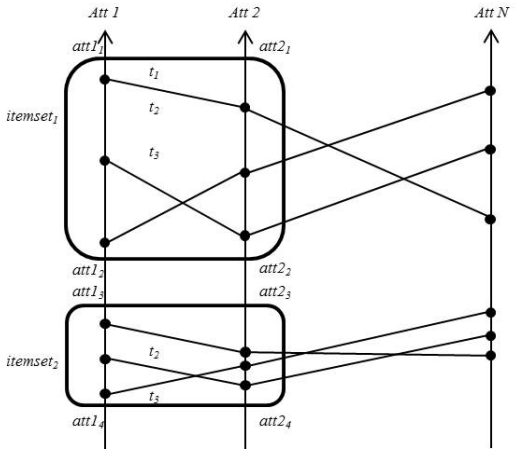

Figure 2: Density Metric

Density The density measure in our paper is defined as the average number of data

instances over unit length interval, i.e.,

$$Dens(rule) = \frac{\sum_{i \in S_{att}} n_{inst}/(c_{ih} - c_{il})}{n_{att}}, \tag{2}$$

in which n_{inst} and n_{att} stand for the number of data instances and number of attributes in the association rule respectively. The density measures the average distances between the data values in the association rule or from another perspective the amount of data values that unit interval contains. It evaluates the significance of an association rule in the way that how frequently this association rule may apply in future application. For example, itemsets $itemset_1$ and $itemset_2$ in Figure 2 are of the same support and confidence measure. The distributions of the data in the two itemsets are nevertheless different. The data values in $itemset_2$ has higher density and shorter average distances than the ones in $itemset_1$. Therefore, within the same size of interval, data will be more likely to fall into $itemset_2$ rather than $itemset_1$ in future application.

The density measure in the QAM replaces the function of support in the traditional association mining. In categorical data, the distance between each item is either 0 for the same items or 1 for different items. The support in traditional association mining algorithms is in essence the statistics of repetition of the same items. The bigger support value an association rule has, the more frequent and likely this rule is going to appear in future instance. Different from the categorical data, the values in quantitative data have very few or no repetition. The counting process for the support measure will return a support count of exactly one for each data values. Moreover, the distances between data values in the quantitative data varies depending on the density/sparsity of the data. These distances contains important information in QAM as the reasons described above, however, the support measure could not capture this information in the quantitative situation.

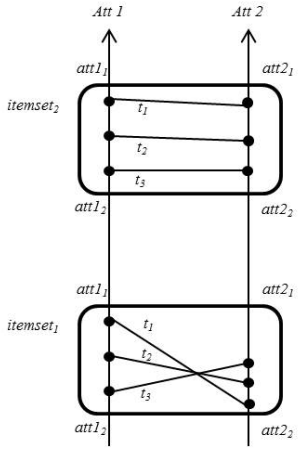

Figure 3: Correlation Metric

Correlation The correlation in this paper is defined by the Spearman's rank corre-

lation coefficient [6]:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}},$$

in which $\bar{x}$ and $\bar{y}$ stand for the average order ranks of attribute x and y respectively. The value of the Speareman's rank correlation coefficient is between [-1, 1], in which +1 and -1 indicate the maximum positive and negative correlations respectively. As shown in Figure 3, the values of attribute 2 in itemset 1 raises strictly along with the values in attribute 1, the correlation measure is therefore +1. In itemset 2, the values of attribute 2 fall when values of attribute 1 raise, the correlation is -1. Intervals with 1 or -1 are of strong correlation and the uncorrelated intervals have correlation close to 0.

The correlation is a measure used to evaluate the quality or precision of the association rules. It plays a similar role as the confidence measure in the traditional association mining. The confidence in the traditional association mining intends to capture the causality relationship of two itemsets, i.e., under the condition one itemset happens, the occurrence probability of another itemset. In the quantitative situation, however, the confidence measure is not accurate enough to measure the precision of association rule. As the output of QAM is in the form of intervals, the larger interval one rule has, statistically the more data instances it will contain. Therefore, the increasing in length of intervals in an association rule will result in an increasing confidence value under the assumption of similar data distribution. In the extreme case, for association rules with full right side intervals $Att_i[min_b, max_b]$, the confidence will always be 100%. This rule, however, contradicts with the goal of finding the causality relationship between itemsets, since after all, the rule is trivial and does not convey any useful information.

4 Method

4.1 Bottom Up Perspective

As mentioned in the previous section, the density and correlation are better interestingness measures for the QAM problem. However, to discover the dense and strongly correlated interval sets from the bottom up perspective is a complex process as the combination of correlation is not linear and recalculation of it is required each time the intervals are combined. For association rule with attribute set and interval set $\{S_{att}, S_{inter}\}$, the correlation is defined as the average correlations between each pair of two intervals in the association rule. The computation of this correlation requires a complexity of $C^2_{n_{att}} n^2_{inst} = O(N^2M^2)$ in which C^n_m stands for the amount of combination. Using the bottom up merge mining method, for example, if we initialize N intervals on each attribute and merge the adjacent intervals with the minimum correlations, a computation of correlation is required for each pair of intervals which are possible to be merged. This computation requires a complexity of $O(NC^2_N C^2_{n_{att}} N^2) = O(N^5M^4)$. As the amount of combination of the attribute sets is $O(M!)$, the complexity of bottom up mining that applied on all combination of attribute sets is then $O((M+4)!N^4)$. Therefore, to use the straight forward merge mining algorithm for discovering the dense and strong correlated intervals is a non trackable process. In this section, we present our hypergraph based method which efficiently solve the QAM problem from the bottom

up perspective. The density and correlation measure are both taken into consideration through the hypergraph representation of the data. These measures are captured by the average commute time distance between vertices in the hypergraph.

4.1.1 Hypergraph representation of quantitative data

Hypergraph is a generalization of regular graph that each edge is able to incident with more than two vertices. It is usually represented by $G = (V, E, W)$, in which V, E and W are the set of vertices, edges and weights assigned to corresponding edges respectively. The incident matrix of a hypergraph G is defined by H in which

$$H(v, e) = \begin{cases} 1 & if\ v \in e \\ 0 & if\ v \notin e \end{cases} \tag{3}$$

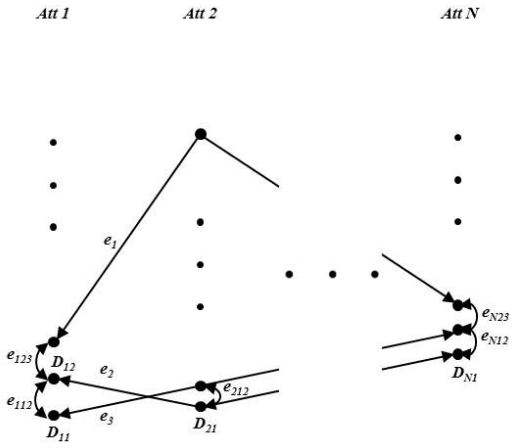

Figure 4: Hypergraph Representation of Quantitative Data

As shown in Figure 4, each data value D_{ij} corresponds to a vertex in the hypergrah. Each pair of vertices with adjacent values, D_{ij} and $D_{i(j+1)}$, are connected through a hyperedge with weight proportion to the inversion of the distance between them. The vertices in the same data instance with value D_{i1}, D_{i2},...,D_{iM} are connected by a hyperedge as well with user defined weight.

Zhou et al. [4] generalized the random walk model on hypergraph and defined the average commute time similarity S_{ct} and the Laplacian similarity S_{L_+}. The average commute time similarity $n(i, j)$ is defined by

$$n(i, j) = V_G(l^+_{ii} + l^+_{jj} - 2l^+_{ij}), \tag{4}$$

in which l^+_{ij} is the ith and jth element of matrix $\mathbf{L}^+$, $\mathbf{L}$ is the hypergraph Laplacian defined by

$$\mathbf{L} = \mathbf{D}_v - \mathbf{HWD}_e^{-1}\mathbf{H}^T \tag{5}$$

, and $\{.\}^+$ stand for Moore-Penrose pseudoinverse. D_v and D_v denote the diagonal matrix containing the degree of vertices and edges respectively and $V_G = tr(\mathbf{D_v})$ is the volume of hyper graph. The average commute time distance is defined by the inversion of normalized average commute time similarity [7]. As mentioned in [12] the commute-time distance $n(i, j)$ between two node i and j has the desirable property of decreasing when the number of paths connecting the two nodes increases. This intuitively satisfies the property of the effective resistance of the equivalent electrical network [5].

4.1.2 Density, Correlation and Average Commute Time distance

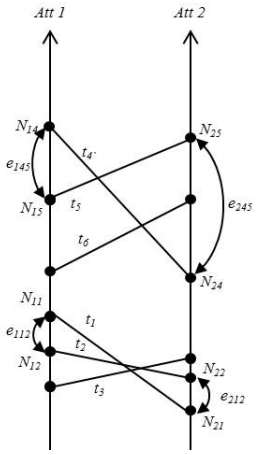

Figure 5: Commute time distance, Density and Correlation

The average commute time distance is a random walk distance. It represents the average distance one walks through when traversing from one node to another in a graph though all random paths. The relationship between the random walk average commute time distance and density and correlation is illustrated as follows. As in Figure 5, the data instances t_1, t_2 and t_3 are dense and strongly negative correlated and the data instances t_4, t_5 and t_6 are sparse and not strongly correlated. As mentioned in the previous section, the weight of the hyperedge between each two adjacent values is the inversion of their distance in the original data. For the denser data instances set t_1, t_2 and t_3, the hyperedges between them are with higher weight. The distance of a random walk through the direct hyperedge is therefore shorter than the edges between t_4, t_5 and t_6. Moreover, the non direct random walk paths on strong correlated data instances, for example, $N_{11} \rightarrow N_{21} \rightarrow N_{22} \rightarrow N_{12}$ is shorter than loosely correlated data instances, for example, $N_{11} \rightarrow N_{21} \rightarrow N_{22} \rightarrow N_{12}$ because the corresponding entities on the other correlated attributes are closer between each other. Therefore, the average commute time distance between the denser and strongly correlated vertices is relatively shorter than the sparse and not strongly correlated vertices. For the reason above, the

average commute time distance from random walk model is capable in capturing both the density and correlation measures for our quantitative association mining problem.

4.1.3 Algorithm Description

Input: $\boldsymbol{D}$, φ, s, c
Initialize: $\boldsymbol{IS} = \boldsymbol{i}_1, \boldsymbol{i}_2, ..., \boldsymbol{i}_N$, $\boldsymbol{i}_k = t_k$, $\boldsymbol{M}$, rule = $\emptyset$
till: size($\boldsymbol{M}$)=1
do: for i_m, i_m, dist(i_m, i_n)=min($\boldsymbol{D}$)
i_m = Merge($\boldsymbol{S}_m$,$\boldsymbol{S}_n$)
if $\frac{max(dist(\boldsymbol{S}_m), dist(\boldsymbol{S}_n))}{dist(\boldsymbol{S}_{mn})} \geq \varphi$
rule += get_rule($\boldsymbol{S}_m$,$\boldsymbol{S}_n$)
Update($\boldsymbol{M}$)

Table 2: Algorithm Description

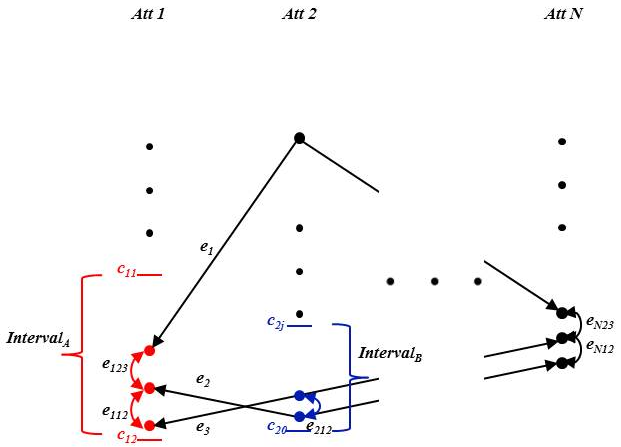

Figure 6: Algorithm in Bottom Up Perspective

As described in previous section, the average commute time distance on hypergraph model is able to capture both the density and correlation measures for the QAM problem. Based on this fact, we propose a bottom up mining algorithm for discovering association rules in quantitative data. The pseudo code of our algorithm is shown in Table 2. With the definition of average commute time similarity in section 4.1.1, we can update the distances of each adjacent data instances using the average commute

time distance. The distance between the adjacent data instances are the inversion of the average commute time similarity. After the update of hypergraph distances, a merge mining process is applied simultaneously on all the attributes to generate our association rules. For each attribute, a distance matrix M_i is maintained for each intervals. At initialization there are in total $N - 1$ intervals with each data value corresponding to one interval. Each row or column in matrix M_i corresponds to one interval in the attribute and the corresponding entry in matrix contains the distance measure between two intervals. φ is the user defined threshold for the automatic rule generator which is described in Section 4.1.4. After each merge, the distance of the new interval is updated based on the previous two intervals.

In every few iterations, for each generated interval, we scan the corresponding intervals on the rest attributes. the distance metrics, density and correlation are calculated for these pairs of intervals. if the distance metric is above the user predefine threshold such as the interval a and interval b in figure 6, then the two intervals are combined into one attribute interval set $\{S_{att}, S_{inter}\}$ and the following merge mining process continues on this set. The merge mining process of this interval halts when the interesting measures of rule generation is satisfied as described in section 4.1.4.

4.1.4 Automatic Rule Generation

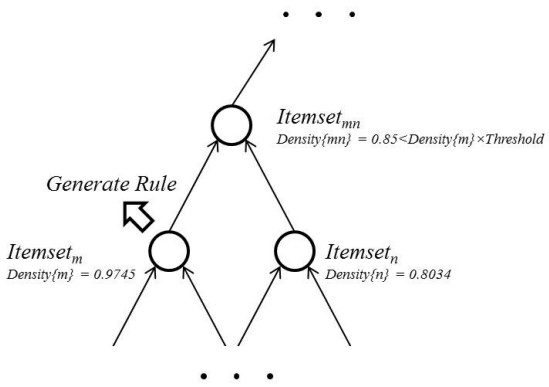

Figure 7: Automatic Rule Generation

In most of the previous association mining algorithms, the rule generation and the frequent itemset discovery are usually two independent phases. As pointed out by Srikant [16] the algorithms often generate too many rules with similar intervals, support and confidence. This problem is called the *too many rule* problem. To solve this problem, people define different interestingness measures to filter out uninteresting and similar rules. However, as algorithm may apply on various datasets with different

characteristic, it is hard to define a universal interestingness measure that satisfies all the dataset perfectly.

In this paper, we deploy an association rule generation mechanism that works along with the process of frequent itemset discovery and deploy the nature characteristic of the data. As show in Figure 7, the process of our merging mining algorithm can be represented as a binary tree, each node represents an interval set and it might merge with some other intervals in the merging process. In this process, the association rules are generated when either the density, confidence or correlation of the new interval set is lower than some user defined percentage φ of the previous set of intervals.

Using this mechanism, the generated association rules has the potential to overlap with each other. It means in our algorithm there is no "crisp boundary" problem that similar values were separated into different intervals. Moreover, it deploys the nature statistic distribution of the dataset that an association rule will always be generated whenever a good enough interval set was discovered.

4.2 Heuristic Search Perspective

In the present section, we present our metaheuristic search algorithm for the QAM problem. This algorithm contains an iterative crossover procedure on the genetic algorithm framework and optimization procedure is used after crossover to guide the search path of the genetic algorithm. In section 4.2.1 we give a short background of the genetic algorithm, then in section 4.2.2 we present the individual representation of the genetic algorithm. In section 4.2.3 the iterative crossover schema and in section 4.2.4 optimization procedures are introduced.

4.2.1 Genetic Algorithm

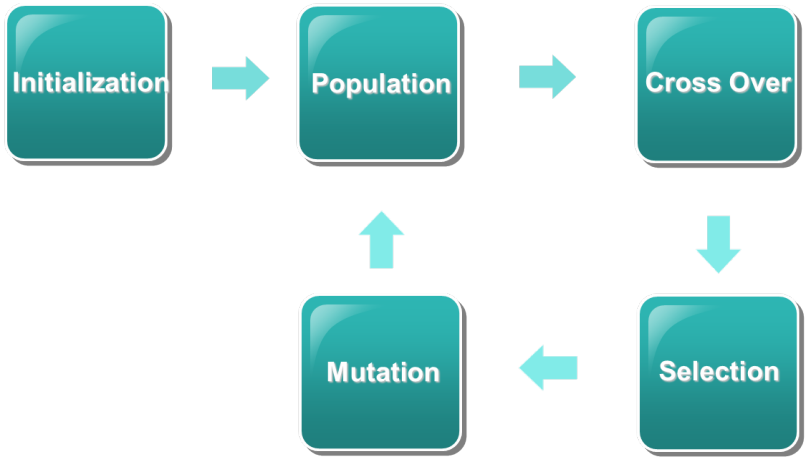

Figure 8: Genetic Algorithm

Genetic algorithm is a popular metaheuristic search algorithm that imitates the nature selection process. At the initialization of the algorithm, the population which contains a fix number of individuals is initialized. For our QAM problem, each individual contains a set of attributes and a set of data instances. These initial individuals can be generated either randomly or based on domain knowledge and statistics. The algorithm proceeds through a series of crossover, mutation, selection processes on the population as shown in Figure 8. In the crossover step a child is generated at added into population from each pair of individuals which are called the parents. The child will takes part of data or attributes from both of the two parents through a random selection procedure as shown in figure 3. The mutation step randomly alters part of the individual so that the individual has the ability to jump out of a local optimal. The selection process filter out inferior individuals based on a fitness function. The same number of individuals with top fitness functions are left for the next round iterative process.

Input: $\boldsymbol{A}$, $\boldsymbol{B}$

Execute: for each a in $\boldsymbol{A}$ and b in $\boldsymbol{B}$

```
if a == b
   add C a
else
   if ( rand(0,1) ) add C a
   if ( rand(0,1) ) add C b
```

Output: C

Table 3: Random selection procedure in crossover for both the attribute sets and instance sets

4.2.2 Individual Representation

In our work, each individual is represented as a pair of attribute and instance sets {Att, Inst}, in which Att $\subseteq \{att_1, att_2, ..., att_M\}$ and Inst $\subseteq \{inst_1, inst_2, ..., inst_N\}$. The instances in the individual has the shape constraint of a hyper rectangle. The hyper rectangle and set of intervals on attributes are in fact equivalent based on the fact that the space defined by the set of intervals is always a hyper rectangle. At the convergence of the genetic algorithm, association rules are generated from each individual within constant time. Comparing with the model in section 3, the data of the individual is modeled by the set of instances instead of the set of intervals. The reason for this representation is because the interval representation does not fit well with the genetic algorithm with reasons stated bellow.

With the interval representation, crossover of intervals cannot proceed in some situations, for example, when the intersection of the parents attribute sets are empty. Even in the situation when the parents share one or few common attributes, the crossover can only perform on these few common attributes. The power of crossover is therefore largely limited. Furthermore, crossing over intervals could readily lead to children that do not conform with the grounding of crossover. The interval of the children might

not relate with any of the two parents. For example, two non overlapping intervals on attribute i that $Inter_A = Att_i[al, ah]$ and $InterB = Att_i[bl, bh]$, in which $ah \geq al \geq bh \geq$ bl. The possible child from crossover $Inter_{AB} = Att_i[bh, al]$ does not include any component of the two parents but the unrelated interval in between. This crossover result contradicts the purpose of using the genetic algorithm, i.e., to borrow superior characteristics from both the parents.

The instance set is a better individual representation in the sense that the crossover of data instance is not constrained by the set of attributes and vice versa. The crossover is performed through a random selection procedure as shown in Table 3. With this representation, the crossover can proceed between any pair of individuals even when they have empty intersections of attribute set or instance set. Each child individual borrows part of instances or attributes from both of its parents.

At last, as the goal of our quantitative association mining is to find the dense and strongly correlated intervals, we should constraint the data instances of each individual with an hyper rectangle. All instances inside the hyper rectangle are included, while the ones outside are not. This is because the space of a set of intervals only equivalent with the set of data inside of a hyper rectangle. The hyper plane of the hyper rectangle is in fact defined on the boundary of the intervals. The scattered or random shaped instance set does not conform with the association rule representation.

4.2.3 Iterative Crossover

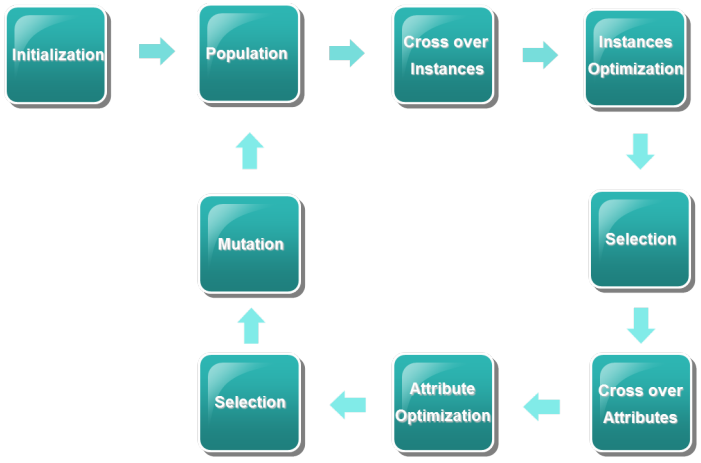

Figure 9: Genetic Algorithm with Iterative Cross Over and Optimization Procedure

With the individual representation described in the previous section, we propose an iterative crossover schema as shown in Figure 9. The ability to apply the iterative crossover is one of the advantages of instance attribute representation, {Att, Inst}. Comparing with other optimization problem, the QAM problem has a special structure: the solution space is composed by the two mutually dependent sets, attribute set

and instance set (or interval set). For any quantitative association rule, the set of data instances is only optimal under this specific attribute set, and vise versa. This structure results in a catch-22 problem, i.e., for any association rule either one of the attribute and instance sets cannot be specified before the other.

To deal with this dilemma, an iterative crossover schema was designed in our genetic algorithm. Under this schema, the crossover was performed iteratively on data instance set and attribute set by fixing the other. For example, the instance crossover for individual A, {att_a, $inst_a$} and B, {att_b, $inst_b$} willwilla generate two children C, {att_a, rand($inst_a$, $inst_b$)}, D, {att_b, rand($inst_a$, $inst_b$)}. And in the attribute crossover step, individual A, {att_a, $inst_a$} and B, {att_b, $inst_b$} will generate C, {rand(att_a, att_b), $inst_a$}, D, {rand(att_a, att_b), $inst_b$}. The rand( ) function is the random selection procedure defined as in Table 3.

4.2.4 Optimization Process

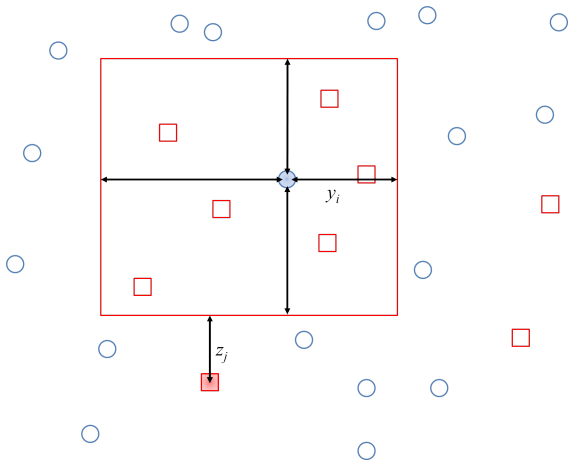

Figure 10: Optimization Procedure

Another advantage of the {Att, Inst} individual representation is that an optimization procedure can be introduced after crossover. The crossover in genetic algorithm involves in several randomized procedures. It is expected to generate children which are similar to both of their parents. However, due to the random process, it might also generates inferior individuals that are further away from the optimization goal in the solution space. Introducing optimization process after the crossover is able to guide the search path of genetic algorithm. Although due to the catch-22 problem described in the past section, direct optimization is not applicable for QAM problem, a feasible solution is applicable under our iterative crossover schema. After instance or attribute crossover step, we proceed to optimize the crossovered instance or attribute set by fixing the other one. Moreover, as described in section 4.2.2, since the instance set is required to be a set in the shape of hyper rectangle, the instances after crossover is

constrained into an axis aligned hyper rectangle by our optimization procedure. The optimization process is described in the following section.

Let us denote the set of instances in individual by I and the set of instances that are not in individual by O. They are denoted by squares and circles in Figure 10 respectively. As described in section 4.2.2 we constraint our data instances as by axis aligned hyper rectangles. The optimization goal is therefore to find a axis aligned hyper rectangle with as many squares and as less circles as possible. For each hyper rectangle selected, two types of errors are defined. Type A error indicates the error when a data instance is inside the hyper rectangle but not a selected instance and type B error indicates that when a selected data instance is outside the hyper rectangle. The two types of errors are represented in the following formulas by y and z respectively.

$$\begin{aligned} & y_{ik} = max\{w_k I_i - r_k, 0\} \\ & z_{jk} = max\{-w_k O_j + r_k, 0\} \\ & i = 1, 2...n_a, \\ & j = 1, 2...n_b, \\ & k = 1, \dots, 2n_{att} \end{aligned} \tag{6}$$

in which k is the number of hyper plane of the hyper rectangle, w_k is a vector in the form of (0, 0, ..., ± 1, 0, 0, ...) which corresponds to one of the hyperplanes of, in which n_{att} is the dimension of the hyper rectangle, i.e., the number of attributes for this individual. The hyper plane is represented by $\{w_k x = c_k, k = 1, \dots, 2n_{att}\}$. The problem is therefore can be formulated as a mathematic programming problem with error function represented as bellow:

$$\begin{aligned} & \min_{w,\gamma,y,z,r} \sum_{i=1}^{n_a} \min_{k \in K} y_{ik} + \sum_{j=1}^{m_b} \sum_{k \in K} z_{jk} \\ & s.t. \quad I_i w_k - \gamma_k \geq y_{ik} \\ & \qquad O_j w_k - \gamma_k \leq z_{ik} \end{aligned} \tag{7}$$

Unfortunately, this mathematic programming problem is neither convex nor linear. It is therefore hard to find the solution which corresponds to the optimal hyper rectangle. To solve this problem, we approximate the objective function using an linear lower bound with methods similar in [18]. Let $u_{ik} = y_{ik} x_i, i = 1, ..., n_a$ and $\sum_{i=1}^{n_a} = 1$ where $x_i \in \{0, 1\}, i = 1, ..., n_a$. Then the u_{ik} is bounded by its piecewise linear and convex envelope

$$u_{ik} \geq \max\{y_{ik} - M x_i - M, 0\}$$

where M is the upper bound on y_{ik}. By applying this under estimation, the optimization

problem is converted to a mix integer linear programming problem

$$\begin{aligned} \min_{w,\gamma,y,z,r} & \sum_{i=1}^{n_a} \sum_{k \in K} u_{ik} + \sum_{j=1}^{m_b} \sum_{k \in K} z_{jk} \\ s.t. \quad & O_i w_k - \gamma_k \geq y_{ik} \\ & O_j w_k - \gamma_k \leq z_{ik} \\ & \sum_{k \in K} x_{ik} = 1 \\ & u_{ik} - y_{ik} - M x_{ik} + M \geq 0 \end{aligned} \tag{8}$$

Then the IBM CPLEX/MILP [1] tool is used to solve this mix integer linear programming problem.

5 Experiment Result

To validate the feasibility of our algorithm, we test both our two algorithms on the data from the NEMO [2] and SMASH [3] projects. The experiment results for the NEMO data and the SMASH data are presented in section 5.1 and 5.2 respectively.

5.1 Experiment on the NEMO data

NEMO [2] is the abbreviation of the Neural ElectroMagnetic Ontologies system. It is designed for the purpose of neuro science data sharing and analysis. The data used from NEMO system is the event-related potentials (ERP) or the "brainwave" data. In the past decades, the studies on ERP data have resulted in many complex neural patterns that can be used to predict human behavior cognition, and neural function. The ERP data is measured through the electro encephalography, i.e., the measure of electrical activity on the scalp using electrodes. The values of ERP data are therefore all quantitative. Each tuple in the data represents the activities in various parts of the cortical network at some specific time. The study of density and correlation of the ERP data has the potential to reveal the connection and function pattern in the cortical network.

In Table 4 we list the top association rules by our algorithm from the NEMO data from the bottom up perspective. The rules are selected by weighted sum of correlation and density and ranked by correlation. Each attribute reflexes the activities measured in certain part of the cortical network at some specific time. For example, the attribute IN-LOCC stands for left occipital lobe which is in charge of the visual spacial processing, color discrimination, and motion perception of human being. The association rules discovered by our algorithm show high interestingness measures and reveal interesting patterns. For example, the first rule in Table 3(a), IN-LOCC [-1.90 , -0.14] $\Rightarrow$ IN-ROCC [-1.92 , -1.15] has a fairly high density value 5.77 and maximum correlation 1.00. This rule reveals a correlation between left occipital lobe and the right occipital lobe. The left occipital lobe in human brain is response for the vision and eye control functions while the right occipital lobe is in charge of reading and writing skills. The relation between IN-LOCC and IN-ROCC implies the potential connection between

(a) Top Ten Rules with Ranked by Density and Correlation

attribute2	range	attribute2	range	density	correlation	corr gain
IN-LOCC	[-1.90,-0.14]	IN-ROCC	[-1.92,-1.15]	5.77	1.00	1.00
IN-LPTEM	[-0.88,-0.33]	IN-RFRON	[-3.66,-1.38]	3.53	1.00	1.20
IN-RATEM	[0.87,2.52]	IN-RFRON	[-3.86,-1.33]	3.10	-1.00	12.08
IN-LATEM	[3.45,1.31]	IN-LFRON	[-5.48,-1.89]	3.02	-1.00	27.21
IN-RPTEM	[-0.95,6.45]	IN-LPTEM	[-0.88,6.35]	2.39	0.99	0.99
IN-LATEM	[-0.58,1.75]	IN-RFRON	[-5.14,-0.28]	2.33	-0.99	11.31
IN-RORB	[-1.05,-0.07]	IN-LPTEM	[0.22,3.34]	2.12	-1.00	12.25
IN-LPTEM	[-0.88,5.49]	IN-LPAR	[-5.88,14.67]	1.45	-0.99	1.62
IN-RPTEM	[-0.95,5.58]	IN-LPAR	[-5.88,14.67]	1.44	-0.98	1.61
IN-ROCC	[-1.92,12.01]	IN-LOCC	[-1.89,12.10]	1.02	0.99	1.00

(b) Top Ten Rules Ranked by Correlation Gain

attribute1	range	attribute2	range	density	correlation	cor gain
TI-max	[236.00,308.00]	IN-RFRON	[-39.26,-0.34]	0.58	0.90	66.89
IN-LFRON	[0.99,7.11]	IN-RATEM	[2.41,17.28]	2.02	1.00	32.95
IN-LATEM	[0.45,1.31]	IN-LFRON	[-5.48,-1.89]	3.02	-1.00	27.21
IN-LFRON	[0.99,7.11]	IN-LATEM	[2.43,17.38]	1.94	1.00	27.21
IN-RORB	[-1.05,-0.07]	IN-LPTEM	[0.22,3.34]	2.12	-1.00	12.25
IN-RORB	[-9.08,-0.49]	IN-LPTEM	[-0.88,0.20]	1.20	0.99	12.25
IN-RATEM	[0.87,2.52]	IN-RFRON	[-3.86,-1.33]	3.10	-1.00	12.08
IN-LATEM	[-0.58,1.75]	IN-RFRON	[-5.14,-0.28]	2.33	-0.99	11.31
IN-RPTEM	[0.16,2.69]	IN-RORB	[-7.88,-0.49]	1.03	-0.99	8.20
TI-max	[236.00,308.00]	IN-LORB	[-20.24,-0.34]	0.73	0.96	7.45

Table 4: Quantitative Association Mining Results From The Neuro Science Data From The Bottom Up Perspective

the left and right lobe in the cortical network. It also reveals the possible affiliation of functions between the vision and reading, writing skills.

Note that even if the correlation for some rules is high enough in Table 4, the correlation gain is relatively low. The correlation gain is defined as the ratio between the correlation of the rule and the correlation of the attributes involved. High correlation gain implies that even though the two attributes are largely uncorrelated most of time, strong correlation might still be found under some intervals. The rule with high correlation gain is valuable in the sense that it reveals the pattern that does not show up under a general picture, i.e., the hidden pattern. On the other hand, for the rules with low correlation measure, most of the contribution of the correlation comes from the attributes but not the interval itself. For example, for the rule IN-RPTEM [-0.95, 6.45] $\cup$ IN-LPTEM [-0.88, 6.35], even if it has a near full correlation 0.99, the correlation gain is 0.99 which means correlation for these two intervals is even lower than the overall correlation between these two attributes. These rules are therefore not useful since the correlation between the two attributes gives us all the information that we need already. This correlation gain measure provides us another perspective to evaluate the precision of an association rule.

In Table 5(a) and 5(b) we list the experiment results from the heuristic search perspective. Comparing with Table 4, the result from the bottom up perspective and heuristic perspective share some similarities. For example, both tables contain rules associated with IN-LOCC, IN-ROCC and IN-LATEM, IN-RATEM. Specificlly, note that the first two rules in 5(b) indicate high correlation gain from IN-RATEM and IN-

(a) Top Five Rules with Ranked by Density and Correlation

attribute1	range	attribute2	range	density	corr	corr gain
IN-LATEM	[-2.16,1.75]	IN-RATEM	[-2.22,3.36]	5.6	1.00	4.34
IN-LFRON	[-0.43,7.11]	IN-RFRON	[-0.44,6.51]	5.86	1.00	1.01
IN-LORB	[-2.14,12.65]	IN-RFRON	[-2.16,6.51]	4.81	1.00	1.06
IN-RORB	[-0.44,20.83]	IN-LFRON	[-3.06,7.11]	3.53	1.00	4.76
IN-LPAR	[-0.18,14.67]	IN-RPAR	[-0.19,14.56]	3.14	1.00	1.03

(b) Top Five Rules with Ranked by Correlation Gain

attribute1	range	attribute2	range	density	corr	corr gain
IN-RATEM	[1.90,17.28]	IN-LFRON	[-4.13,7.11]	2.67	0.99	33.00
IN-LATEM	[0.99,17.38]	IN-LFRON	[-4.13,7.11]	2.86	1.00	27.21
IN-LPTEM	[0.20,14.56]	IN-RORB	[-9.08,-0.56]	3.34	0.98	12.26
IN-RATEM	[-0.43,17.28]	IN-RFRON	[-0.44,6.51]	3.72	0.89	11.32
IN-LATEM	[-0.42,17.38]	IN-RFRON	[-0.44,6.51]	3.15	0.88	11.31

Table 5: Quantitative Association Mining On Neuro Science Data From The Heuristic Search Perspective

LATEM both to IN-LFRON[-4.13, 7.11]. These two rules together reveal a potential high correlation between IN-RATEM and IN-LATEM. This guess is in fact validated by the first rule in Table 5(a) that IN-LATEM [-2.16 , 1.75] $\Rightarrow$ IN-RATEM [-2.22 , 3.36] with 1.00 correlation.

5.2 Experiment on the SMASH data

SMASH [3] is the abbreviation of Semantic Mining of Activity, Social, and Health Project. It was designed for the purpose of learning the key factors that spread the healthy behaviors in social network. It consists of distributed personnel device and web-based platform that collect data from both social and physical activity. The data collected in this project include social connections and relations, physical activities and biometric metrics from the subjects. After preprocessing, the input data in our experiment have the following indicators for the physical activities and biomedic metrics. The physical activity indicator "Ratio No.Steps" is the change ratio of steps that the subjects walked through in two consecutive periods of time. Three biomedical metrics HDL, LDL and BMI are used for the health indicators. The HDL and LDL stand for the high density lipoprotein and low density lipoprotein respectively. The rate of HDL usually relates with decreasing rate of heart related disease and the reverse case for LDL. The BMI stands for body mass index which is a common indicator of the obesity level. The study of this data set dedicates to discover the relations between physical activities and rate of heart disease conditions.

In Table 6 we list our experiment results from the SMASH system in the bottom up perspective. The association rules in Table 6(a) and 6(b) are ranked by correlation and correlation gain respectively. The second and third rule in Table 6(b) Ratio HDL [0.79 , 0.91], Ratio No.Steps [0.78 , 3.94] and Ratio LDL [1.00 , 1.02] Ratio No.Steps [0.39 , 3.82] demonstrate two origins of the high correlation gain. Although the correlation for the second rule is only 0.56, it results in an even higher correlation gain than the third rule. For the third rule, although the correlation is maximum which

(a) Top Five Rules with Ranked by Density and Correlation

attribute1	range	attribute2	range	density	correlation	corr gain
Ratio LDL	[1.00,1.02]	Ratio No.Steps	[0.39,3.82]	245.39	1.00	22.11
Ratio BMI	[0.97,1.12]	Ratio LDL	[0.82,1.17]	118.22	0.99	3.55
Ratio LDL	[0.97,1.26]	Ratio BMI	[0.89,1.00]	70.14	0.94	45.22
Ratio HDL	[0.79,0.95]	Ratio BMI	[0.94,1.03]	83.87	0.89	9.41
Ratio BMI	[0.98,1.14]	Ratio No.Steps	[0.27,1.66]	120.77	0.82	5.13

(b) Top Five Rules Ranked by Correlation Gain

attribute2	range	attribute2	range	density	correlation	corr gain
Ratio HDL	[0.97,1.26]	Ratio BMI	[0.89,1.00]	70.14	0.94	45.22
Ratio HDL	[0.79,0.91]	Ratio No.Steps	[0.78,3.94]	53.34	0.57	28.15
Ratio LDL	[1.00,1.02]	Ratio No.Steps	[0.39,3.82]	245.39	1.00	22.11
Ratio HDL	[0.78,1.25]	Ratio No.Steps	[0.27,3.96]	255.20	0.37	18.41
Ratio LDL	[0.88,1.02]	Ratio No.Steps	[0.01,3.82]	305.55	0.72	16.12

Table 6: Quantitative Association Mining On Social Health Network Data From Bottom Up Perspective

is 1.00, part of the high correlation in it comes from more of the two attributes rather than from the interval.

Comparing these two tables, the results in Table 5(b) contain more rules related with No.Steps than the ones in Table 5(a). The reason is possibly because the correlation between biomarkers are more regular and stable in the general picture than the correlation between physical activities and biomarkers. For example, the level of HDL and LDL are both in someway related with the cholesterol level in human body. In most cases high HDL level couples with low cholesterol level and the reverse case for LDL. The situation for physical activity is nevertheless different. Research indicates that physic activities have various impacts on the obeisity status respect to the short term and long term effects. The human subjects might carry out more activities for various even contradicting reasons. For example, people might increase the amount of physical activities due to an increasing of BMI in previous period, however, some other subjects who wants to build up their body might work out more due to the decreasing of weight. Moreover, increasing amount of physical activities might lead to increase in food intaking which results in higher cholesterol level in the following period.

In table 7(a) and 7(b), we list the top five rules from our heuristic search algorithm. The results are roughly similar with the ones in table 6.

6 Discussion and Future Work

Our experiment results validate the feasibility of our algorithm. They show that our algorithms are capable in finding the set of intervals with both high density and correlation gains. Nevertheless, some improvements are still needed for our methods.

In section 4.2.4, an optimization procedure is introduced after the crossover to guide the search path of the genetic algorithm. After the crossover of instances, the instances in the child individual is optimized into an axis aligned hyper rectangle. All the data instances for the individual reside inside the hyper rectangle while the rest

(a) Top Ten Rules with Ranked by Density and Correlation

attribute1	range	attribute2	range	density	correlation	corr gain
Ratio LDL	[0.96,1.00]	Ratio HDL	[0.94,1.00]	322.75	0.88	3.22
Ratio LDL	[0.86,0.94]	Ratio HDL	[0.86,1.04]	387.50	-0.70	2.55
Ratio BMI	[1.00,1.02]	Ratio No.Steps	[0.73,3.80]	155.75	0.68	4.15
Ratio LDL	[1.07,1.15]	Ratio No.Steps	[0.83,1.17]	221.85	0.66	11.44
Ratio LDL	[0.77,1.14]	Ratio No.Steps	[1.04,1.12]	160.50	0.58	10.04

(b) Top Ten Rules Ranked by Correlation Gain

attribute2	range	attribute2	range	density	correlation	corr gain
Ratio HDL	[0.94,1.04]	Ratio NO.Steps	[0.98,1.16]	287.50	0.57	56.10
Ratio HDL	[0.90,1.25]	Ratio NO.Steps	[0.90,0.99]	172.44	0.55	53.50
Ratio LDL	[0.65,0.90]	Ratio BMI	[0.97,0.99]	151.46	0.60	25.12
Ratio LDL	[1.07,1.15]	Ratio No.Steps	[0.83,1.17]	221.42	0.66	11.44
Ratio LDL	[0.90,1.05]	Ratio No.Steps	[0.75,0.98]	147.13	0.59	10.11

Table 7: Quantitative Association Mining On Social Health Network Data From Heuristic Perspective

instances are not. This procedure tries to make the instances in the children individuals as dense as possible and make the calculation of correlation a feasible procedure as well. However, the optimization algorithm after the crossover of attributes is not established yet. Although there are various feature selection algorithms that select the optimal subset of attributes. None of them could select the feature based on the correlation of attributes. This part of selecting the subset of attributes after crossover is left for future work.

Moreover, note that the optimization in section 4.2.4 uses the IBM CPLEX/MILP tool to find the optimal hyper rectangle that maximize the selected and minimize the non selected instances in hyper rectangle. The complexity of the mix integer optimization process varies depending on the structure of data and individuals. In practice, the result of the optimization result might not be the optimum output as well. With the power of evolution and selection, the optimization process in the evolution process might not need to generate an optimal result based on the current dual set. One possible way of the agile optimization procedure could start from either of the two parents which are already an individual with good fitness function after several iterations. Another way is to run the optimization process only in a few iterations. Then the marginal improvement of the interestingness measure is judged that optimization halts if it barely improves or be discarded if deteriorated.

7 Conclusions

In this paper, we present our two algorithms in terms of the quantitative association mining problem. Although the QAM seems like a nature extension of the traditional association mining, the mining process for it is quite different. We propose to use two new interestingness measures, density and correlation, that better fit with the QAM problem since the support and confidence measures from the traditional association mining do not fit with the QAM anymore. Further more, we design two algorithms that can discover association rules with both high density and correlations measures from

the quantitative data. Experiment results validate both our algorithms from the bottom up and heuristic search perspective.

References

[1] http://www.ibm.com/software/commerce/optimization/

[2] http://aimlab.cs.uoregon.edu/NEMO/web/

[3] http://aimlab.cs.uoregon.edu/smash/

[4] Dengyong, Z., Jiayuan, H., Scholkopf, B.: Learning with hypergraphs: Clustering, classification, and embedding. pp. 1601–1608 (2006)

[5] Doyle, P., Snell, L.: Random Walks and Electric Networks. Mathematical Association of America (1984)

[6] Freedman, D., Pisani, R., Purves, R.: Statistics. W. W. Norton & Company (2007)

[7] Haishan, L., Paea, L.P., Ruoming, J., Dejing, D.: A hypergraph-based method for discovering semantically associated itemsets. In: International Conference on Data Mining. pp. 398–406 (2011)

[8] Javeed, Z.M.: Scalable algorithms for association mining. IEEE Transaction on Knowledge and Data Engineering 12(3), 372–390 (2000)

[9] Jef, W., Robert, M.: On the complexity of mining quantitative association rules. Data Mining and Knowledge Discovery 2(3), 263–281 (1998)

[10] Jiawei, H., Jian, P., Yiwen, Y., Runying, M.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery 8(1), 53–87 (2004)

[11] Jochen, H., Ulrich, G., Gholamreza, N.: Algorithms for association rule mining a general survey and comparison. SIGKDD Explorations Newsletter 2(1), 58–64 (2000)

[12] Lovasz, L.: Random walks on graphs: A survey. In: Miklos, D., Sos, V.T., Szőnyi, T. (eds.) Combinatorics, Paul Erdős is Eighty, vol. 2, pp. 353–398 (1996)

[13] Man, K.C., Ada, F., Hon, W.M.: Mining fuzzy association rules in databases. International Conference on Management of Data pp. 41–46 (1998)

[14] Miguel, D., Daniel, S., Maria, M.B., Vila, M.M.A.: Mining association rules with improved semantics in medical databases. Artificial Intelligence in Medicine 21(1-3), 241–245 (2001)

[15] Rakesh, A., Ramakrishnan, S.: Fast algorithms for mining association rules in large databases. In: International Conference on Very Large Data Bases. pp. 487–499 (1994)

[16] Ramakrishnan, S., Rakesh, A.: Mining quantitative association rules in large relational tables. International Conference on Management of Data pp. 1–12 (1996)

[17] Salvador, G., Julian, L., Jose, S., Victoria, L., Francisco, H.: A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transaction on Knowledge and Data Engineering 25(4), 734–750 (2013)

[18] Seo, R.H.: Pattern classification by concurrently determined piecewise linear and convex discriminant functions. Computing and Industry Engineering 51(1), 79–89 (2006)

[19] Takeshi, F., Yasuhido, M., Shinichi, M., Takeshi, T.: Mining optimized association rules for numeric attributes. In: Symposium on Principles of Database Systems. pp. 182–191 (1996)

[20] Takeshi, F., Yasukiko, M., Shinichi, M., Takeshi, T.: Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization. In: International Conference on Management of Data. pp. 13–23 (1996)

[21] Yonatan, A., Yehuda, L.: A statistical theory for quantitative association rules. In: International Conference on Knowledge Discovery and Data Mining. pp. 261–270 (1999)