Multi-level mining of association rules from warehouse schema

Muhammad Usman and Muhammad Usman*

Dept. of Computing, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology Islamabad, Pakistan

Corresponding author: dr.usman@szabist-isb.edu.pk

Abstract

The integration of data mining techniques with data warehousing is becoming an interesting domain. The reason behind this popularity is the ability to extract knowledge from large data sets. However, in current available techniques, a big emphasis is put on solutions, where data mining plays a front end role to data warehousing for mining of data. Very little work is done, in order to apply data mining techniques in design of data warehouses. While techniques like data clustering have been implied on multidimensional data to enhance the knowledge discovery process, still a number of issues remain unresolved related to the multidimensional schema design. These issues include the manual process of selection of important facts and dimensions in high dimensional data environment, an activity which is a challenging job for human designers, where data is available in large volume having many related variables. In this research we propose a technique to select a subset of informative dimensions and fact variables, to start the mining process. Our experimental results after implementation of method on real world data set taken from UCI machine learning website show that the rules discovered from the schema were better in terms of importance as well as diversity, as compared to the rules discovered from typical data mining process used on the original data without schema imposed on it.

Keywords: Association rule mining; data warehouses; interestingness measures; knowledge discovery; multidimensional schema.

1. Introduction

Data mining provides the ability to analyze large data sets to discover informative knowledge, whereas data warehousing contains integrated data from different sources (Moses & Dian, 2015). The objectives of both technologies are same, and both can be used to take benefit from each other to strengthen the knowledge discovery process. Even though, both technologies have gone through independent development phases, yet there is a close match between the two in terms of objectives. For instance, both technologies are aimed at providing insights into the data at abstract level. Moreover, both technologies specifically targeted large volume of data. More recently, there has been a trend to integrate data mining with data warehousing for knowledge discovery process (Goil & Choudhary, 2001; Usman *et al*., 2009; Zhen & Minyi, 2001). Techniques exist in literature, where mining is achieved by constructing the data cubes or by exploiting the Star schema (Chung & Mangamuri, 2005; Ng *et al.*, 2002; Kamber *et al.*, 1997; Tjioe & Taniar, 2005). The techniques exploited data warehouse design to extract informative knowledge, but relied heavily on user's domain knowledge to construct the data warehouse design. In recent past, effort has been done in order to automate the process of data warehouse schema generation in order to apply data mining techniques on it. Such technique results in discovering knowledge better in terms of importance from data warehouse. However a number of issues exist in the approached presented by Usman *et al*. (2013).

In current approach, the technique ranks nominal and numeric variables in the data is set to construct the fact table for data warehouse design. The ranking procedure helps to select variables for the design of data warehouse, but it treats nominal and numeric variables separately. In most cases, data in different variables is inter-related, so this process does not consider the relationship, which exists within the two types of data.

Moreover, the technique makes use of agglomerative hierarchical clustering, to generate clusters at different levels. At each cluster level, authors create multidimensional schema on which mining process is applied. Although it helps to discover knowledge at different levels of hierarchy, manual generation of data warehouse schema against each cluster at different levels is a cumbersome process.

Finally, association rules are generated using the data warehouse schema, and advance diversity measures are applied for interestingness. This steps involve manual calculation of diversity values for each cluster, at all levels in the hierarchy, making it a lengthy procedure.

In this research, we make the following contributions. We provide a better technique to create the multidimensional schema. We convert all nominal data to equivalent numeric data and apply a single statistical method at each level of abstraction. The mining of association rules on this schema generates rules with higher importance than the previous approach (Usman *et al.*, 2013).

Moreover, we automate the data warehouse schema generation process. We construct schema for all required clusters at different levels by obtaining input against each. It helps to reduce the time required for schema design at all levels.

Finally, we create a tool to assist researchers in calculating diversity measures for rules sets for Rae, CON and Hill at once. It reduces the time required for calculation of diversity measures at all levels.

We have tested the model on real world dataset (automobile) from UCI machine learning repository (Asuncion & Newman, 2010). We generated rules for clusters at different levels and found that we achieve better importance of association rules, using our generated schema than the previous approach (Usman *et al*., 2013).

The rest of paper is organized as follows. Related work section presents an overview of the previous research relevant to the area. We discuss the actual model with the help of an example in Proposed Conceptual Model. In next section we present the application of our model to the real world case study. Finally, a summary of our research achievements and some future guidelines are presented in last section.

2. Related work

In this section we present previous work done in the area of rule mining over data warehouses. A number of techniques are proposed in literature for association rule mining over data warehouses.

Han *et al*. (1997) designed a methodology, which comprises of few algorithms. These algorithms are used to generate association rules using a data cube structure. The *m-D slicing algorithm* as developed by the authors uses thresholds called support and confidence in each dimension in the data cube. To find large item sets, the process filters the rules based on confidence values. The process generates a cube, if it is not present already. During the process of cube construction, another algorithm prunes records in order to find large item sets. This algorithm works on a smaller set at the end of generation of association rules. Authors have shown that the algorithm works better than *apriori algorithm*. Authors used traditional measures in order to evaluate the interestingness and it is not evident, as if the generated rules with better confidence and support values were actually meaningful for the business analysts or not.

Kamber *et al*. (1997) used meta rule guided mining on data cubes, in order to create a smaller subset of data as input for rule mining process. The mining process mines rules from data cubes containing aggregated data. In case a data cube is pre-computed and available for mining, two algorithms are proposed. *Multi-D-Slicing algorithm* withdraws 1-predicate sets and these are used for multi-dimensional slicing on the data cube. Second algorithm n-D cube search works on *p-D cells* of a particular *n-D cube*. The item set is selected if minimum support threshold is justified. In case the pre-computed data cube is not available, mining process creates a cube. Authors developed *n-D Cube* construction algorithm which works using meta rules. It uses a subset of an n-D cube after construction of data cube. *Multi-p-D algorithm* for construction of cube is used to generate *p-predicate meta rules* with *n-dimensions*. This process generates rules as guided by meta rules, but business analysts having limited technical knowledge find it difficult to use.

Messaoud *et al.* (2007) categorized the association rule mining over multi-dimensional data into three parts. Authors named association within a dimension as intra-dimensional association. They named the association of multiple dimensions as inter-dmiensional association. Authors defined another term called hybrid association, which combines previous two categories. Authors construct a tabular format using the data cube and find out frequent items. These frequent items are then used for association rule mining.

Extended association rules were first defined by Psaila & Lanzi (2000), which draw rules from more than one dimension. The process is started by selection of non-item attributes. In order to generate association rules using these attributes, the methodology creates SQL Queries.

The other part of the process is meant to mine rules from data available in aggregated format. Since the rules are generated using query processing, the data is not required to be loaded from the data warehouse. However business analysts are required to have technical knowledge to do this.

Inspired by Kamber *et al*. (1997), Messaoud *et al.* (2006) worked on data cubes to mine rules using meta rule guided approach. Authors used advanced aggregated measures to evaluate the rules. Such approach generates rules using a sub set of data cubes and it's not necessary to use th whole data set. So the process adopted decreases the time for mining of rules. Authors have used aggregated measures like *min, max, avg, sum* as an alternative to the COUNT measures. Authors used traditional *apriori algorithm* with bottom-up approach for mining of rules. Although execution time is reduced in this approach, and due to aggregated measures, better rules are generated; there is a lot of emphasis on meta-rules. Users must know the technical aspects behind meta-rules to adapt the technique effectively.

Usman *et al*. (2013) adopted a two step process for multi-dimensional mining. The first step creates a multi-dimensional schema by using limited set of dimension and facts. The second step generates rules using this schema. Authors used agglomerative hierarchical clustering, in order to draw clusters at different levels of hierarchy. At each cluster, authors created multi-dimensional schema using top-ranked variables. Nominal variables are ranked using information gain, whereas numerical variables are ranked using statistical method called principal component analysis (PCA). In this process, since both types of variables are ranked separately, the association between these variables is not used. This issue affects the ranking process and does not present a true picture of underlined data.

Our methodology is closely related to the model presented by Usman *et al*. (2013). It is evident that multi-level mining of association rules not only generates knowledge at different levels of hierarchy, but it also draws better knowledge in terms of association rules compared with the original data. However, there are some issues in the current approach as discussed below.

We observe that one of closely related technique creates multi-dimensional schema and achieves better importance than the original data. This technique ranks variables in the data set to select top variables for schema generation. However, this technique ranks nominal and numeric variables separately, thus does not consider the relationship among both types of data. The existing relationship between both types of data should be considered for ranking process. We propose the ranking procedure in the way that it ranks all data at once. We convert nominal variables to equivalent numeric variables and apply principal component analysis at once. We generate schema based on our ranked variables. We show that the rules generated using our schema based on the combined ranking of variables, generates more important rules than the approach adapted by Usman *et al*. (2013).

Moreover, the technique requires schema generation for each cluster at every level. The large number of clusters makes it a cumbersome process since it requires manual creation of schema in data warehouse for each cluster. We created a tool to automate the process of schema generation. The tool takes input of list of nominal and numerical variables with the required clusters. It outputs the schema for all clusters to create required data warehouse design.

In next section we provide key concepts and background knowledge of different techniques used in our proposed model.

3. Proposed conceptual model

In this section, we explain each step of our proposed methodology using an exemplary dataset. Figure 1 shows the proposed conceptual model. The proposed model prepares multidimensional schema using top-ranked variables. We generate association rules in order to show the effectiveness of our approach.

Suppose that there is a real world data set D of library subscribers. The data set has three numerical variables (age, no-of-books-issued, online-accesses) and includes four nominal variables (study-level, gender, study-subject, user-type. Let us assume that the data set contains x number of records. As the first step, we apply agglomerative hierarchical clustering to all data, in order to generate clusters at different levels of hierarchy.

After clustering is done, nominal data is converted to numeric data using Rosarios's approach (Rosario *et al.,* 2004). This process is repeated for each cluster at every level of the hierarchy. As a result, all clusters contain data in numerical format only. We rank variables in each cluster using the statistical technique called principal component analysis (PCA).

If all variables were used to define the split of C2 into

C21 and C22, then a variable having an impact on split will have greater variance in one of the child clusters and lesser in other clusters. So, for example, user type plays more significant role in cluster C21 compared with C22. This variable has the highest difference between factor loadings and thus stands first in the ranking.

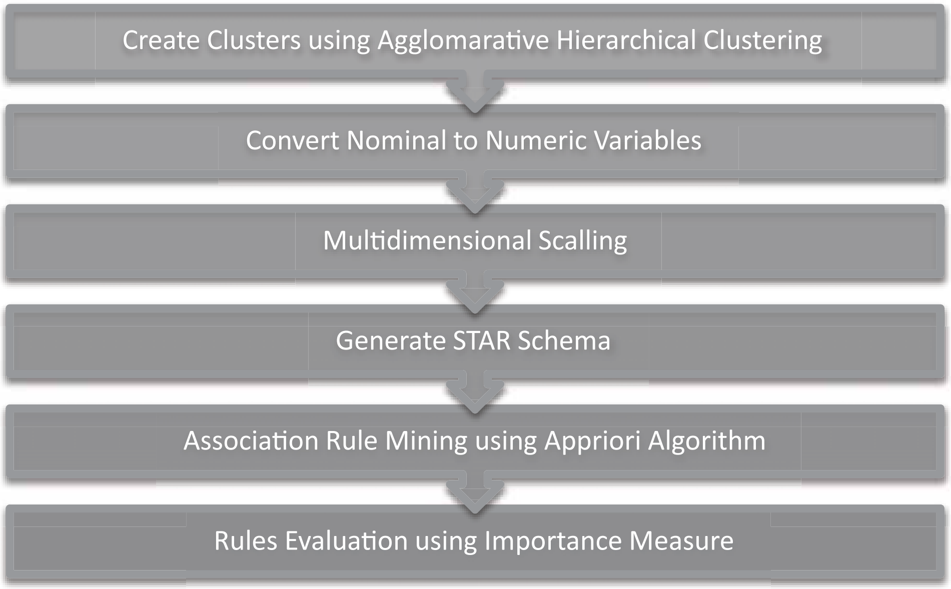

Create Clusters using Agglomarative Hierarchical Clustering
Convert Nominal to Numeric Variables
Multidimensional Scalling
Generate STAR Schema
Association Rule Mining using Appriori Algorithm
Rules Evaluation using Importance Measure

Fig. 1. Conceptual model for multi-level association rule mining

Table 1. Variables ranking in cluster c2

Variables	C21 Factor Loadings	C22 Factor Loadings	Comparison Results	Rank
age	0.997	0.627	0.370	Rank # 3
Online accesses	0.666	0.546	0.120	Rank # 6
No. of books issued	0.733	0.330	0.403	Rank # 2
Study level	0.334	0.102	0.232	Rank # 4
User type	0.974	0.310	0.664	Rank # 1
Study subject	0.849	0.700	0.149	Rank # 5

In our next step, we define natural groupings of all nominal variables. Using this step, we place distinct values in a single group, which are not far away from each other. After the variables are ranked, nominal variables are considered as dimensions and numeric variables are treated as facts for generation of STAR schema. The STAR schema is generated for each cluster by using facts and dimensions. The group information is also used to define the dimensional hierarchy. At each cluster, the dimensions have a group and value level. For example, if we take USER TYPE as a dimension then it has USER-TYPE (all) Level → USER-TYPE_GROUPS (GROUP) LEVEL → USER-TYPE_NAMES (VALUE) LEVEL.

We generated multidimensional schema using numeric and nominal variables in our demo application. Our application generates SQL Scripts in order to create dimensions, fact tables and all necessary relationships. We show the multidimensional schema generated for cluster C2 in figure 2. The schema contains dimensions and fact tables for cluster C2. In order to show the effectiveness of our approach, we used *apriori algorithm* to generated association rules using original data as well as the schema which is generated in our case. The results of ranking steps are used in rule mining step. Nominal variables, which are ranked higher than the other variables in each cluster are used as dimension in order to identify the association between them. Such association is significant due to high ranks. We believe that top-ranked variables picked through PCA provide more informative rules as PCA provides degree of variance, which in turn is effective on complete data set. Importance measures are used for measuring the interestingness as conventional measures are only suitable for transactional databases. The generated rules are shown in table 3 for cluster C2 In order to check the effectiveness of our approach, these rules are further compared with the rules generated using the original data in the table 2.

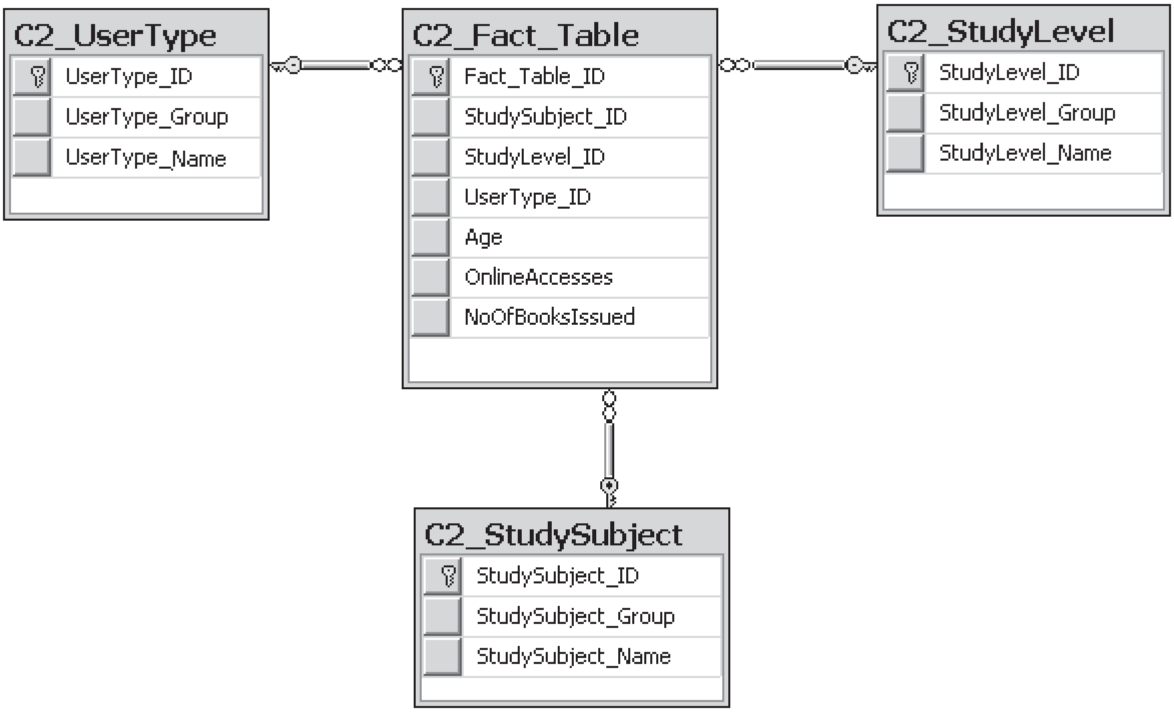


Fig. 2. Schema generated by the demo application for cluster C2

It is clear that the association between different dimensions can be determined using these rules. For instance, in table 2, results show that *StudySubject* will be “Agr. Science” if *StudyLevel* is PHD and *UserType* is “Researcher”. It reflects that there is a relationship between these variables since the importance value is positive. Similarly, other rules also suggest a relationship between different dimensions. On the other hand, since we have created groups of distinct values of nominal variables in our model, rules are shown for groups of values of dimensions. For example, rule 2 in table 3 suggests that all distinct values in Group 2 for *UserType* variable have “Micro Bio. Science” as *StudySubject*. It is clear that these rules are better in terms of importance as well as cover multiple values of dimensions at a time.

Table 2. Rule Generated for without multidimensional schema for c2

Without Multidimensional Schema		
No.	Rule	Importance
1	If StudyLevel [PHD] & UserType[Researcher] →StudySubject[Agr. Science]	1.03
2	If UserType[Student]→ StudySubject[Bio. Sciences]	1.01
3	If StudyLevel[MPHIL] & UserType[Student]→StudySubject[App. Science]	0.73

Table 3. Rule Generated for with multidimensional schema for c2

Without Multidimensional Schema		
No.	Rule	Importance
1	If StudyLevel [G1] & UserType[G1] →StudySubject[Agr. Science]	1.21
2	If UserType[G2]→ StudySubject[Micro. Bio. Sciences]	1.13
3	If StudyLevel[G1] & UserType[G3]→StudySubject[Computer Science]	1.00

We rank the rules using these measures in terms of diversity. Table 4 describes the diversity values for generated rules for cluster C1 in our example.

Table 4. Rule Evaluation using advanced diversity measures for cluster C1

	Without Multidimensional Schema			With Multidimensional Schema (% increase)		
Rule Set	Rae	CON	Hill	Rae	CON	Hill
R1-R2	0.23	0.21	-1.5	0.51	0.76	-0.11
R3-R4	0.17	0.06	-2.9	0.29	0.33	-1.2

We present the case study of automobile dataset retrieved from UCI machine learning website to show the effectiveness of our approach.

4. Case study on real world data sets

In order to validate and to find the effectiveness of the methodology we proposed, we proceeded with the case study approach. This approach allowed us to evaluate our methodology to discover diverse association rules from the multi-dimensional structure.

4.1. Case study – Automobile dataset

In order to validate the proposed model, we selected automobile data set available at UCI machine learning repository (Schlimmer, 1985). This dataset is a mixture of numeric and nominal variables. The analysts may be interested in both types of data depending upon the data set, so this data set stands useful to verify the methodology presented above.

The automobile dataset contains 26 variables. There are 11 nominal (categorical) variables and 15 numeric variables. This standard dataset describes the characteristics of an automobile. More details of the dataset are available at UCI machine learning website.

According to the first step, we applied agglomerative hierarchical clustering using HCE Explorer available online. We took clusters at different levels and exported their data for further usage. In order to rank variables within clusters, we used PCA technique. PCA was applied through IBM's statistics analysis tool called SPSS. We ranked variables within each cluster namely C1, C11 and C12. The reason for picking C1, C11, and C12 is the fact that most data is located in C1 cluster which further divides itself into C11, C12. Since PCA works only on numeric data, we converted nominal variables to corresponding numeric variables using Rosario's approach (Rosario *et al.*, 2004). All numeric data was then passed to PCA for ranking purposes. Figure 3 shows the ranking of numeric variables in the three clusters after performing the PCA.

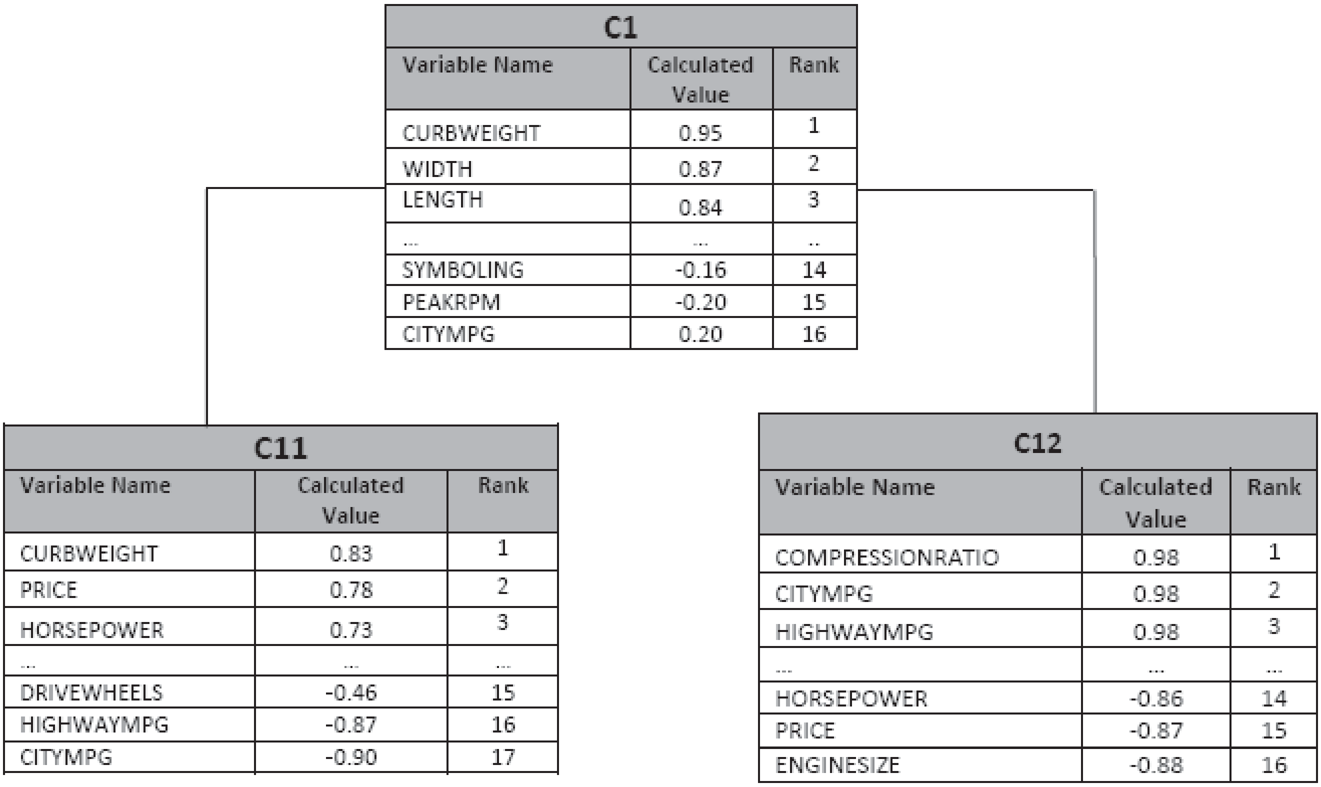
From the Figure 3, it is evident that the same variables have different ranks within the different clusters. In other words, every cluster contains a distinctive set of numeric variables on top of the list. For example, *Curb Weight*, *Width* and *Length* are top ranked variables in cluster C1. However, in the lower data abstraction level, for example, C11, *Curb Weight* maintains its position, but *Width and Height* don't maintain their rank. In case of C11, the top ranked variables include *Price* and *Horse Power*. The situation is more interesting in case of C12, where none of the top 3 ranked variables in C1 appear on top and instead go well at the bottom of the list.

The variable *Curb Weight* remained on top in C1 and C11, but was positioned 13th in C12. *Width* was ranked at 2nd position in C1 and it was ranked 4th in C11 whereas it went to 11th position in C12. *Length* variable was ranked at 3rd position in C1, and maintained its position in C11 as well, but it was ranked at 9th position in C12. We can conclude the cluster C1 is split into two clusters on the basis of *Price* and *Horse Power* mainly and on *Compression Ratio* marginally. Figure 6 shows the ranking of nominal variables in the three clusters after performing PCA.

In case of nominal variables, there is also a difference in ranking of variables at different levels. It can be seen from Figure 4 that top ranked variables, *Engine Type*, *Body Style* and *Engine Location* in C1 do not have same ranks in C11 and C12. The variable *Engine Type* is ranked at first position in C1. But this variable is ranked as 5th in C11. Same variable is ranked at 4th position in C12. *Body Style* was ranked at 2nd position in C1, and was found at 6th position in C12. Due to zero variance in C11, *Body Style* was not ranked at all. *Engine Location* was ranked at position 3 in C1. It was moved to 4th position in C11. Due to zero variance in C12, the variable was not ranked. The top ranked variable in C11 is *Make*, and it was ranked at 6th position in C1. The top ranked variable *Fuel Type* in C12 is ranked at 4th position in C1. It can be concluded that C1 cluster is split into two clusters on the basis of *Make*, *Fuel Type* and *Aspiration*.

In our prototype, we input the list of numeric and nominal variables separately. The prototype generates the SQL schema with required tables in order to perform rule mining process. All dimension tables are created using nominal variables list. The grouping obtained above is imported into the dimension tables and a generated script from the same application maps all data to dimension table groups. The output of this step is the multidimensional schema for all clusters. Figure 5 shows our demo application used for schema generation. The top tanked dimensions of C11 cluster are Engine Type, Engine Style and Engine Location. The Engine Type variable values are put into two groups. First group of Engine Type variable contains rotor, dohc, ohcv, l and ohc. The second group contains ohcf. It is clear from the group information that Engine Type values are naturally grouped together. Similarly other dimensions such as Engine Size, Engine Location etc also contain groupings which enlighten the semantic relationship present in these variables.

In order to assess the benefits of variables ranking at different levels, we applied association rule mining on top ranked variables to generate rules for clusters at different levels. The rules were compared with the set of rules generated by the technique adapted by Usman *et al*. (2013). The rules generated through our process have minimum .8 probability value and minimum importance value of 0.99. For cluster C11, our technique generates 8 rules. From the rules produced, we see that the low ranked variables have tendency to predict the *Make*, of an automobile. Similarly the *Number of Cylinders* and *Aspiration* can predict other *Make* when combined together.

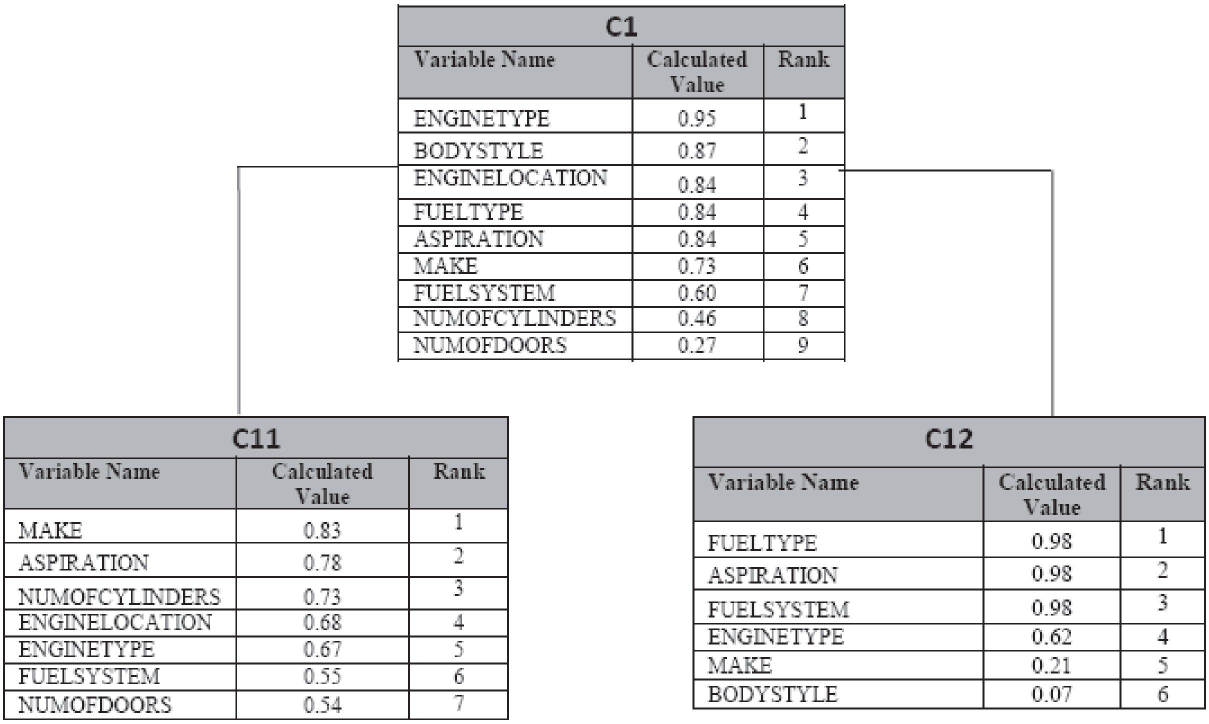


C1		
Variable Name	Calculated Value	Rank
CURBWEIGHT	0.95	1
WIDTH	0.87	2
LENGTH	0.84	3
...	...	..
SYMBOLING	-0.16	14
PEAKRPM	-0.20	15
CITYMPG	0.20	16

C11		
Variable Name	Calculated Value	Rank
CURBWEIGHT	0.83	1
PRICE	0.78	2
HORSEPOWER	0.73	3
...	...	...
DRIVEWHEELS	-0.46	15
HIGHWAYMPG	-0.87	16
CITYMPG	-0.90	17

C12		
Variable Name	Calculated Value	Rank
COMPRESSIONRATIO	0.98	1
CITYMPG	0.98	2
HIGHWAYMPG	0.98	3
...	...	...
HORSEPOWER	-0.86	14
PRICE	-0.87	15
ENGINESIZE	-0.88	16

Fig. 3. Ranking of numeric variables using PCA for clusters C1, C11 and C12

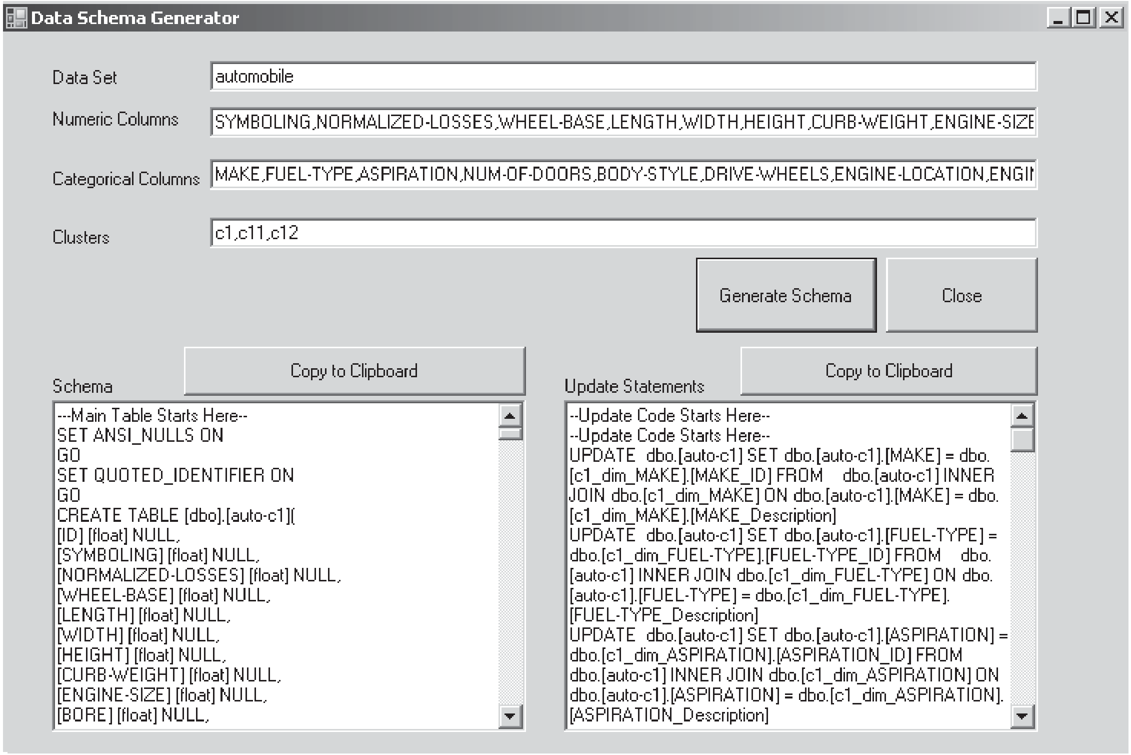


C1		
Variable Name	Calculated Value	Rank
ENGINETYPE	0.95	1
BODYSTYLE	0.87	2
ENGINELOCATION	0.84	3
FUELTYPE	0.84	4
ASPIRATION	0.84	5
MAKE	0.73	6
FUELSYSTEM	0.60	7
NUMOFCYLINDERS	0.46	8
NUMOFDOORS	0.27	9

C11		
Variable Name	Calculated Value	Rank
MAKE	0.83	1
ASPIRATION	0.78	2
NUMOFCYLINDERS	0.73	3
ENGINELOCATION	0.68	4
ENGINETYPE	0.67	5
FUELSYSTEM	0.55	6
NUMOFDOORS	0.54	7

C12		
Variable Name	Calculated Value	Rank
FUELTYPE	0.98	1
ASPIRATION	0.98	2
FUELSYSTEM	0.98	3
ENGINETYPE	0.62	4
MAKE	0.21	5
BODYSTYLE	0.07	6

Fig. 4. Ranking of nominal variables using PCA for clusters C1, C11 and C12



Data Set: automobile

Numeric Columns: SYMBOLING,NORMALIZED-LOSSES,WHEEL-BASE,LENGTH,WIDTH,HEIGHT,CURB-WEIGHT,ENGINE-SIZE

Categorical Columns: MAKE,FUEL-TYPE,ASPIRATION,NUM-OF-DOORS,BODY-STYLE,DRIVE-WHEELS,ENGINE-LOCATION,ENGI

Clusters: c1,c11,c12

Schema

```
---Main Table Starts Here--
SET ANSI_NULLS ON
GO
SET QUOTED_IDENTIFIER ON
GO
CREATE TABLE [dbo].[auto-c1](
[ID] [float] NULL,
[SYMBOLING] [float] NULL,
[NORMALIZED-LOSSES] [float] NULL,
[WHEEL-BASE] [float] NULL,
[LENGTH] [float] NULL,
[WIDTH] [float] NULL,
[HEIGHT] [float] NULL,
[CURB-WEIGHT] [float] NULL,
[ENGINE-SIZE] [float] NULL,
[BORE] [float] NULL,
```

Update Statements

```
--Update Code Starts Here--
--Update Code Starts Here--
UPDATE  dbo.[auto-c1] SET dbo.[auto-c1].[MAKE] = dbo.
[c1_dim_MAKE].[MAKE_ID] FROM    dbo.[auto-c1] INNER
JOIN dbo.[c1_dim_MAKE] ON dbo.[auto-c1].[MAKE] = dbo.
[c1_dim_MAKE].[MAKE_Description]
UPDATE  dbo.[auto-c1] SET dbo.[auto-c1].[FUEL-TYPE] =
dbo.[c1_dim_FUEL-TYPE].[FUEL-TYPE_ID] FROM    dbo.
[auto-c1] INNER JOIN dbo.[c1_dim_FUEL-TYPE] ON dbo.
[auto-c1].[FUEL-TYPE] = dbo.[c1_dim_FUEL-TYPE].
[FUEL-TYPE_Description]
UPDATE  dbo.[auto-c1] SET dbo.[auto-c1].[ASPIRATION] =
dbo.[c1_dim_ASPIRATION].[ASPIRATION_ID] FROM
dbo.[auto-c1] INNER JOIN dbo.[c1_dim_ASPIRATION] ON
dbo.[auto-c1].[ASPIRATION] = dbo.[c1_dim_ASPIRATION].
[ASPIRATION_Description]
```

Fig. 5. Demo application used for generation of multi-dimensional schema for the data sets

At this point, our data is available for rule mining process on both original clusters data as well as the respective multidimensional schema of the cluster. For each cluster, we picked top three nominal variables with respect to the ranking in order to extract association rules. Table 5 shows the 8 rules for cluster C11 generated with our schema.

Table 5. Rules generated from the proposed technique for cluster c11 with schema generated from proposed methodology

S.NO	Rule	Importance
1	NUM-OF-CYLINDERS Group = Group 1, ASPIRATION Group= Group-Others → MAKE Name = mercedes-benz	1.90
2	NUM-OF-CYLINDERS Group = Group 1 → MAKE Name = mercedes-benz	1.90
3	NUM-OF-CYLINDERS Group = Group 1 → MAKE Name = audi	1.64
4	NUM-OF-CYLINDERS Group = Group 1, ASPIRATION Group = Group-Others → MAKE Name = audi	1.31
5	NUM-OF-CYLINDERS Group = Group 1, ASPIRATION Group = Group-Others → MAKE Name = audi	1.17
6	NUM-OF-CYLINDERS Group = Group 1 → MAKE Name = mazda	1.14
7	NUM-OF-CYLINDERS Group = Group 1, ASPIRATION Group = Group-Others → MAKE Name = mazda	1.14
8	NUM-OF-CYLINDERS Group = Group-Others, ASPIRATION Group = Group-Others → MAKE Name = nissan	0.99

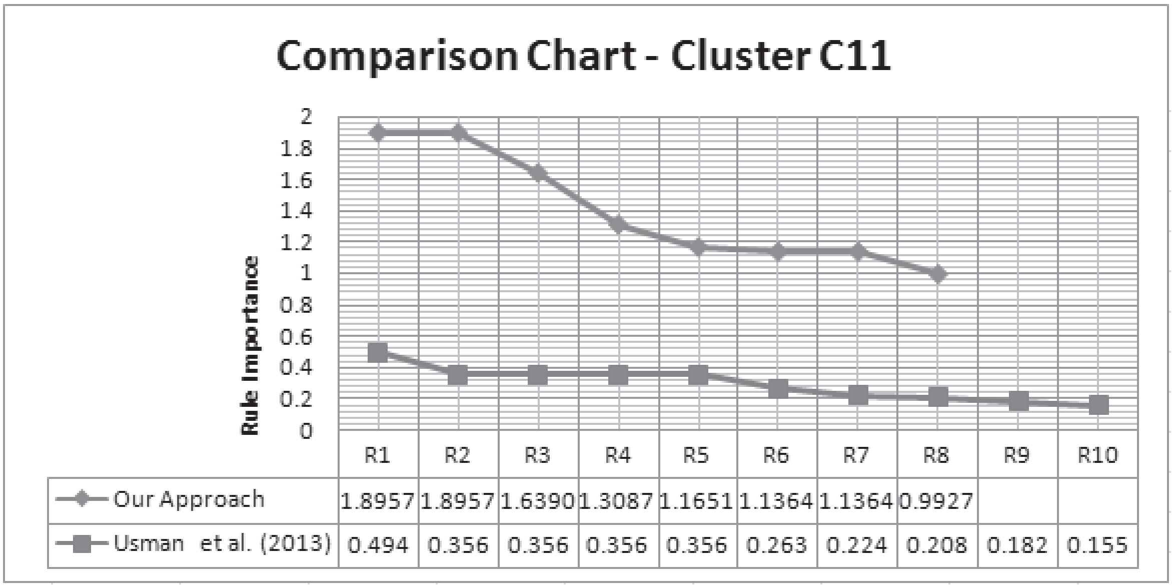


Fig. 6. Comparison of rules generated using schema from our approach and Usman *et al.* (2013) approach for cluster C11

We present a comparison of our rules with the set of rules generated for same cluster using the previous approach. Our top rule provides an importance of 1.9 as compared to the top rule of the set against which we are comparing, which has 0.49 importance value for the same cluster. This means that the rules generated by our technique are more important than the other technique. Moreover, our technique generated all rules having importance value greater than or equal to 0.9. Whereas the other technique generates rules set in which top rule has 0.49 importance value.

It is evident from Figure 6, that our technique generates more important rules than the technique provided by Usman *et al*. (2013). The most important rule generated from the technique provided by Usman *et al.* (2013) has a value 0.49 where as the lowest important rule in our approach has a greater value of 0.99. It clear indicates that our approach out performs the other technique.

We show the diversity calculations for all clusters using original data as well as our schema. The results are shown in Table 6. We present the results of Usman *et al.* (2013) in Table 7. We prepare rule sets using different rules in each cluster and draw the measures of interestingness (Rae, CON, Hill). The results of these measures using original data are given in ‘No Schema’ column. We present our results in ‘With Schema’ columns. It is clear that our schema values against all measures are better than the No Schema values. In order to show effectiveness in comparison with the approach discussed by Usman *et al.* (2013), we calculate the percentage increase in for without schema and with schema diversity for all measures.

We calculate the increase for all cluster for all measures and present the comparison of our approach and Usman *et al.* (2013) in Figures 7, 8 and 9. Figure 7 presents the comparison of diversity increase (%) for Rae measure. It is evident that the increase in diversity between non-schema and schema rule sets is very prominent than the approach adapted by Usman *et al.* (2013). Similarly our approach yields better increase for other measures (CON, Hill) as depicted in Figures 8 and 9.

Table 6. Rules interestingness comparison using objective diversity measures with %increase achieved

Clusters	Rule Set	No Schema Rae	With Schema Rae	No Schema CON	With Schema CON	NO Schema Hill	With Schema Hill
C1	R1-R6	0.139	**0.318**	0.168	**0.429**	-3.919	**-1.978**
	R1-R7	0.116	**0.295**	0.154	**0.424**	-4.701	**-2.162**
	R1-R8	0.1	**0.274**	0.142	**0.416**	-5.502	**-2.350**
	R1-R9	0.086	**0.257**	0.132	**0.408**	-6.321	**-2.526**
	R1-R10	0.076	**0.557**	0.122	**0.715**	-7.154	**-0.569**
C11	R1-R6	0.246	**0.353**	0.31	**0.477**	-2.784	**-1.611**
	R1-R7	0.206	**0.339**	0.28	**0.482**	-3.454	**-1.695**
	R1-R8	0.173	**0.332**	0.244	**0.490**	-4.27	**-1.735**
C12	R1-R6	0.178	**0.153**	0.196	**0**	-3.796	**-5**

Table 7. Rules interestingness comparison using objective diversity measures with % increase achieved (Usman et al. 2013)

Clusters	Rule Set	No schema Rae	With Schema Rae	No Schema CON	With Schema CON	NO Schema Hill	With Schema Hill
C1	R1-R6	0.139	**0.196**	0.168	**0.221**	-3.919	**-3.473**
	R1-R7	0.116	**0.179**	0.154	**0.233**	-4.701	**-3.823**
	R1-R8	0.1	**0.151**	0.142	**0.203**	-5.502	**-4.682**
	R1-R9	0.086	**0.142**	0.132	**0.195**	-6.321	**-5.446**
	R1-R10	0.076	**0.117**	0.122	**0.167**	-7.154	**-6.287**
C11	R1-R6	0.246	**0.291**	0.31	**0.396**	-2.784	**-2.079**
	R1-R7	0.206	**0.271**	0.28	**0.393**	-3.454	**-2.272**
	R1-R8	0.173	**0.22**	0.244	**0.473**	-4.27	**-3.126**
	R1-R9	0.152	**0.189**	0.224	**0.303**	-4.927	**-3.755**
	R1-R10	0.151	**0.186**	0.246	**0.317**	-4.965	**-3.791**
C12	R1-R6	0.178	**0.203**	0.196	**0.244**	-3.796	**-3.261**
	R1-R7	0.169	**0.186**	0.233	**0.256**	-3.976	**-3.574**
	R1-R8	0.161	**0.181**	0.253	**0.287**	-4.158	**-3.682**
	R1-R9	0.153	**0.174**	0.265	**0.294**	-4.343	**-3.791**
	R1-R10	0.146	**0.17**	0.271	**0.301**	-4.529	**-3.899**

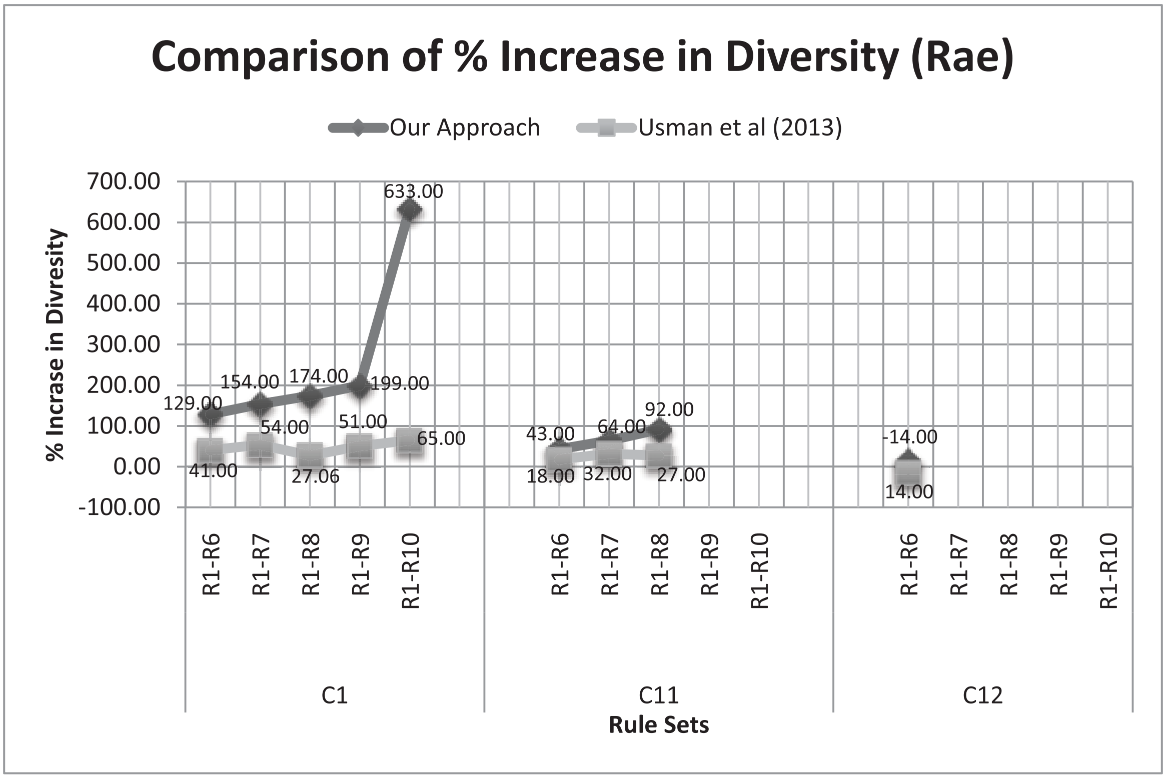


Fig. 7. Comparison of % increase in diversity (Rae) with Usman *et al.*(2013)

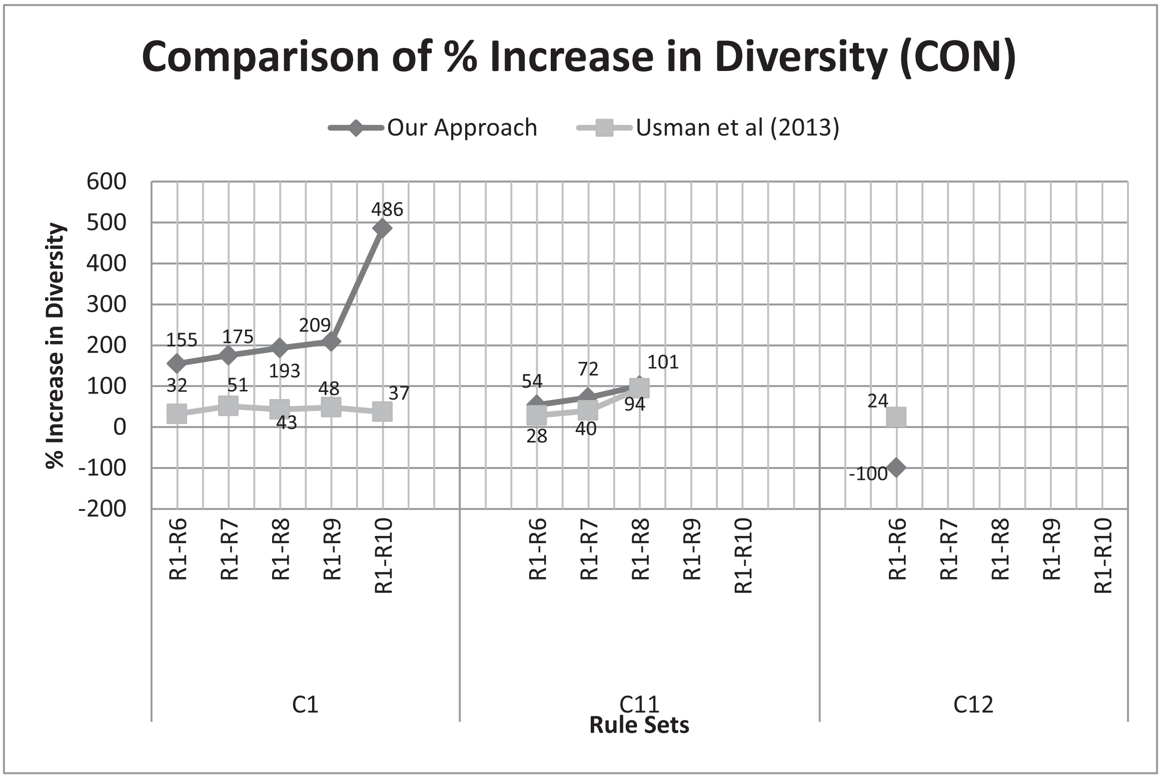


Fig. 8. Comparison of % increase in diversity (CON) with Usman *et al.*(2013)

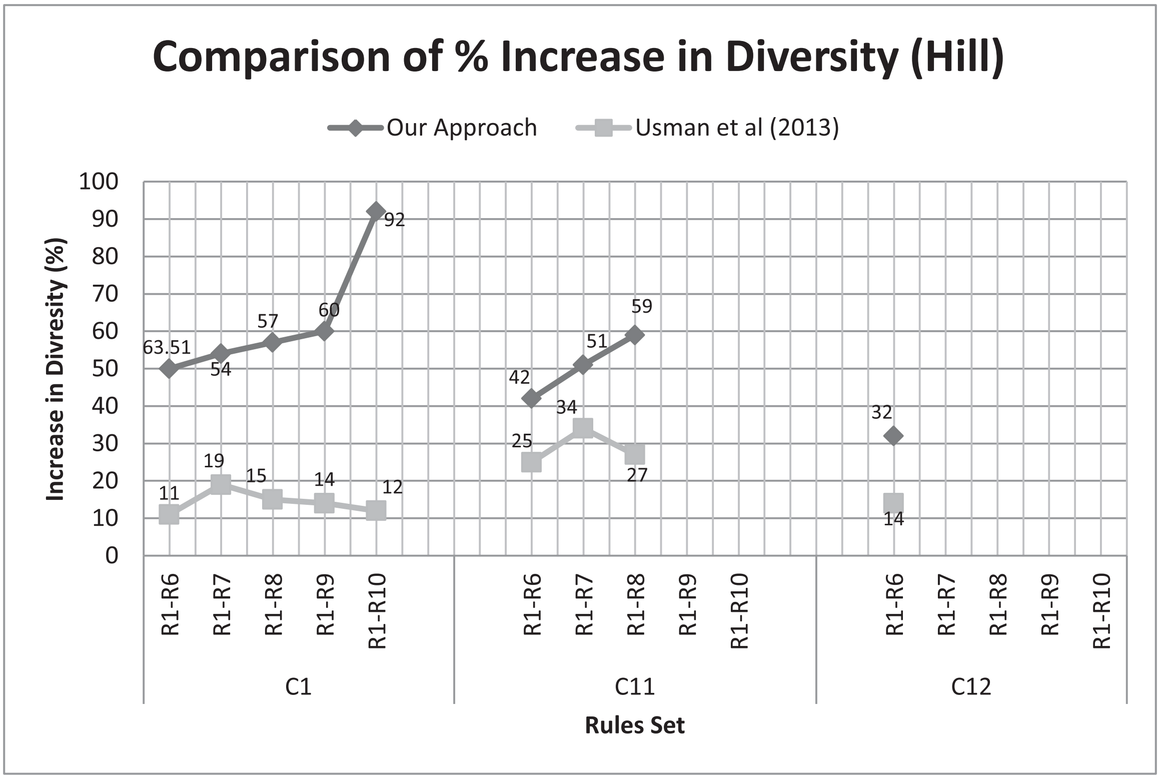


Fig. 9. Comparison of % increase in diversity (Hill) with Usman *et al.*(2013)

5. Conclusion

In this paper, we have presented a methodology to mine association rules using multidimensional schema. We apply hierarchical clustering at different levels of abstraction in the data and use statistical techniques to draw a sub set of variables containing more informative data. We generate multidimensional schema using the top ranked variables in the data and use this schema to generate association rules using *Apriori Algorithm*. We were able to achieve better importance of rules than a similar approach presented by Usman *et al*. (2013). We also show that rules generated using our approach are more diverse than the approach presented by Usman *et al*. (2013). We propose that different approaches might be tried instead of PCA for variable ranking to check for better results. It is also proposed the accuracy of rules may be checked in order to make sure that predications are accurate enough. Moreover, the generated rules might be presented in a better way using some graphical interface.

References

Asuncion, A. & Newman, D.J. (**2010).** UCI machine learning repository. http://archive.ics.uci.edu/ml.

Chung, S.M. & Mangamuri, M. (2005). Mining association rules from the star schema on a parallel NCR teradata database system. International Conference on Information Technology, Coding and Computing (ITCC'05). Dayton, OH, USA.

Goil, S. & Choudhary, A. (**2001).** Parismony: An infrastructure for parallel multidimensional analysis and data mining. Journal of Parallel and Distributed Computing, **61**:285–321.

Han, J., Kamber, M. & Chiang, J. (**1997).** Mining multi-dimensional association rules using data cubes. Technical report, Database Systems Research Laboratory, School of Science, Simon Fraser University, Burnaby, BC, Canada.

Kamber, M., Han, J. & Chiang, J.Y. (**1997).** Metarule-guided mining of multi-dimensional association rules using data cubes. KDD. Burnaby, BC, Canada.

Messaoud, R.B., Rabaséda, S.L., Boussaid, O. & Missaoui, R. (**2006).** Enhanced mining of association rules from data cubes. DOLAP '06 Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, New York, USA.

Messaoud, R.B., Loudcher, S., Missaoui, R. & Boussaid, O. (**2007).** "OLEMAR: an on-line environment for mining association rules in multidimensional data," Advances in Data Warehousing and Mining, IGI Global, Vol. 2, 2007, pp. 1-35. DOI: 10.4018/978-1-59904-960-1.ch001

Moses, D. & Deisy, C. (2015). A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, **42**(2):206-235.

Ng, E.K.K., Fu, A.W.C. & Wang, K. (2002). Mining association rules from stars. Proceedings of IEEE International Conference on Data Mining IEED-ICDM, Maebashi City, Japan, pp. 322-329.

Psaila, G. & Lanzi, P.L. (**2000).** Hierarchy-based mining of association rules in data warehouses. Proceedings of the 2000 ACM symposium on Applied computing, Como, Italy.

Rosario, G.E., Rundensteiner, E.A., Brown, D.C., Ward, M.O. &

Huang, S. (**2004).** Mapping nominal values to numbers for effective visualization. Information Visualization, **3**:80-95.

Schlimmer, J.C. (1985). "Automobile dataset" Retrieved 20 june, 2012, from http://archive.ics.uci.edu/ml/datasets/Automobile. University of California, Irvine, School of Information and Computer Sciences

Tjioe, H.C. & Taniar, D. (2005). Mining association rules in data warehouses. International Journal of Data Warehousing and Mining, **1**(3):28-62.

Usman, M., Asghar, S. & Fong, S. (2009). Conceptual model for combining enhanced OLAP and data mining systems. Fifth international joint conference on INC, IMS and IDC 09, Seoul, Korea, pp. 1958–1963.

Usman, M., Pears, R. & Fong, S. (2013). Discovering diverse association rules from multidimensional schema. Expert Systems with Applications, **40**(15):5975-5996.

Zhen, L. & Minyi, G. (**2001).** A proposal of integrating data mining and on-line analytical processing in data warehouse. Proceedings of the international conference on info-tech and info-net, Beijing, China, pp. 146–151.

***Submitted* :** 26/02/2015
Revised **:** 28/12/2015
Accepted **:** 28/12/2015

التعدين متعدد المستويات لقواعد الإرتباط من مخطط مستودعات البيانات

محمد عثمان و *محمد عثمان

قسم الحوسبة، معهد الشهيد ذو الفقار علي بوتو للعلوم والتكنولوجيا، إسلام أباد، باكستان

*المؤلف المراسل: dr.usman@szabist-isb.edu.pk

خلاصة

دمج تقنيات استخراج البيانات مع تخزين البيانات (مستودعات البيانات) أصبح مجالاً للإهتمام. والسبب وراء هذه الشعبية هي القدرة على استخراج المعرفة من مجموعات البيانات الكبيرة. ومع ذلك، في التقنيات المتاحة حاليا، يكون التركيز على الحلول، حيث يلعب استخراج البيانات دور الواجهة الأمامية لتخزين البيانات بغرض التنقيب. وقليل من البحث تناول تطبيق تقنيات التنقيب عن البيانات في تصميم مستودعات البيانات. وبينما تقنيات مثل تكتل البيانات تستخدم في البيانات متعددة الأبعاد لتعزيز عملية اكتشاف المعرفة، لا يزال عدد من القضايا لم تحل تتعلق بتصميم مخطط متعدد الأبعاد. وتشمل هذه القضايا عملية الإختيار اليدوي من الحقائق والأبعاد في بيئة البيانات متعددة الأبعاد وعالية الأهمية، وهو نشاط يمثل تحدياً للمصمم البشري، حيث تتوفر بيانات في حجم كبير ووجود العديد من المتغيرات ذات الصلة. في هذا البحث نقترح تقنية لتحديد مجموعة فرعية من الأبعاد المعلوماتية ومتغيرات الواقع، لبدء عملية التعدين. النتائج التجريبية بعد تطبيق الطريقة على مجموعة بيانات واقعية مأخوذة من موقع تعلم الآلة UCI تبين أن القواعد التي تم اكتشافها من المخطط كانت أفضل من حيث الأهمية، وكذلك التنوع، بالمقارنة مع قواعد اكتشفت من عملية استخراج البيانات المستخدمة عادة في البيانات الأصلية دون فرض مخطط عليها.

الكلمات المفتاحية: التنقيب عن قواعدالإرتباط؛ مستودعات البيانات؛ مقاييس الأهمية؛ اكتشاف المعرفة؛ مخطط متعدد الأبعاد.