Measuring Diversity of Associations Rules Extracted from A Data Warehouse

Muhammad Usman

***Abstract*—Knowledge discovery is a series of steps to extract useful information from data sets containing data in large volume. Nowadays, data sources contain large number of dimensions and data size is getting increased as a result. Data is archived in data warehouses now in an aggregate form. Data mining techniques to extract knowledge from datasets are now being applied in data warehouse. Interesting patterns are extracted in the form of association rules from data warehouses whereas interestingness measures are used to evaluate these patterns. The techniques available for evaluation of association rules were originally developed for transactional databases. In this research work, we enhance our previous methodology which extracts association rules in a data warehouse environment at multiple levels of abstraction. In this work we evaluate these association rules using advanced measures of interestingness particularly targeting the diversity measures. We have applied 9 measures of interestingness on association rules generated in the data warehouse and shown our results for diversity. Results further suggest that there is a strong correlation between some of these measures at cluster level. A future study can be conducted to deduce a linear model for prediction of diversity measures at lower levels in the hierarchy.**

***Index Terms*—Association rule mining, interestingness, data warehouse, multi-level mining.**

I. Introduction

Knowledge discovery is a series of steps to extract interesting information from a data set. Data is available in large volumes and high number of dimensions now-a-days. Techniques exist which extract knowledge from large data sets including data warehouses. In knowledge discovery process, patterns can be extracted in the form of association rules. Different evaluation measures have been adapted by researchers to evaluate the interestingness of the knowledge. Such evaluation is necessary as business analysts are often interested in hidden and interesting patterns in the data.

In some of previous studies the patterns are evaluated using statistical approaches [1], [2]. In some of the techniques, authors have relied on conventional measures like Support and Count etc. Moreover, a number of researchers have used advanced measures of interestingness like Rae, Con, Hill, Lift and Loveigner for evaluation purposes. However, few issues remain un-addressed when mining in a multi-dimensional environment.

While mining in a data warehouse environment, association rules are evaluated mainly using conventional measures like support and confidence. Some authors have started to use diversity measures. Few techniques are available which use Rae, CON and Hill in a data warehouse environment. There is a need to use other diversity measures like variance, Simpson, Mchintosh, Lorentz, Shutz and Whittaker to check the diversity of extracted patterns.

Moreover, some studies have been conducted to find correlation between different measures in a dataset. However, these studies are not performed on multiple data sets and cannot be generalized. The correlation between different measures is also not known when it comes to mining in a data warehouse environment. Moreover, in a multi-level environment, if such correlation is found, a linear model can be used for prediction of these measures in levels under a certain level in the hierarchy.

In this research work, we have enhanced our previous work done in [3]. The previous work included a model which extracted patterns in a data warehouse environment at multiple levels of abstraction. We evaluated the extracted patterns using diversity measures including Rae, Con and Hill. In this enhanced version, we have applied 9 Diversity Measures called Rae, Con, Hill, Variance, Simpson, Mchintosh, Lorentz, Shutz and Whittaker. We have evaluated the extracted rules using these measures and presented out results. We have also investigated the relationship between these measures at cluster level using Pearson's Correlation. It is found that some measures have a direct or indirect strong relationship. In future, we intend to create a linear model using information at cluster level for all diversity measures. The linear model can further be used to predict diversity measures using other diversity measures at other levels in the hierarchy.

The proceeding sections of the paper are organized in the following way: We present the literature review of methods/techniques used for evaluation of association rules in the past in section 2. Section 3 elucidates the model used in this approach with an exemplary dataset. A case study is presented in Section 4 which is performed on real world dataset called Forest Cover Type taken from UCI machine learning website. In Section 5, concluding remarks are presented with future guidelines.

II. Related Work

In this section, we review the previous work done in the area of advanced evaluation of association.

Measures of interestingness allow the business analysts to evaluate the extracted knowledge from a dataset. These measures of interestingness are divided into two categories.

Manuscript received February 12, 2018; revised May 4, 2018.

Muhammad Usman is with Pakistan Scientific and Technological Information Center, QAU Campus, Islamabad, Pakistan (e-mail: usmiusman@gmail.com).

doi: 10.18178/ijke.2018.4.1.101

Objective measures of interestingness are based upon structure of extracted patterns [4]. Subjective measures of interestingness are based upon class of users who examine the mining process [5]. In this section we review the previous work done in the area of advanced evaluation of association rules in a multi-dimensional environment using interestingness measures.

The approach presented in [6] for exploration of unexpected patterns from datasets includes an algorithm called ZoomUAR which is belief driven based upon the user's domain knowledge. This approach is tested on an in-house data set containing 36 different attributes. The comparison of the approach has been made with well known Apriori Algorithm. The results suggest that the proposed algorithm doesn't generate very strong rules in terms of confidence of the rule; however the generated rules are interestingness as they were unexpected from the user's expectation. Authors conclude that the method discovers only interesting rules compared to obvious and irrelevant rules generated by Apriori. Authors plan to implement the algorithm on variety of datasets.

In order to develop a new criteria for measuring interestingness authors in [7] reviewed objective and subject interestingness measures. Authors developed a third type called impartial interestingness criteria. This approach requires very little domain knowledge that a naive user can provide to eliminate the rules which are not interesting. The authors investigated the possibilities to include the interestingness measure in the mining process and found it to be useful.

As an alternate, the authors in [1] worked on indentifying interesting patterns by adapting a skewness based approach. Authors investigated three measures of interestingness for navigation rules. The approach examines the unexpectedness of navigation rules, axis shift in the navigation path, and generalization of association rules. The methodology has been tested using five datasets, and is reported to be effective. According to the authors, none of the previous measures work for navigation multi-dimensional data cubes.

Authors in [2] affirmed that the extracted rules must be post-processed after extraction for usability. Moreover, selection of right interestingness measures is very important. In this study RQAT tool has been used with 40 different interestingness measures. The correlation between different measures has been calculated using correlated graphs from the tool. These graphs helped to indentify the clusters of similar measures.

Chandanan and Shukla [8] explain that measuring the interestingness of a rule is an important area. The authors define the interestingness that covers the aspects like Conciseness, Generality/Coverage, Reliability, Peculiarity, Diversity, Novelty, Surprisingness, Utility, and Actionability. Conciseness tells how much easy a rule or rule set is easy to understand. Generality defines the coverage of a rule set for a dataset. If a condition or a relationship defined by rules occurs very often, then the rule is said to be more reliable. Peculiarity defines rules which are rarely found in the dataset. Diversity measure defines the rules which are significantly different from each other in the resultant set. Novel rules are not known beforehand and can't be extracted without user involvement. If a rule contradicts with the user's existing knowledge, the surprisingness is high for that rule. If a rule contributes to the achievement of a goal, it is more utilized. If a rule helps in making decision on future actions, the rule is said to be actionable.

As a step towards usage of advanced measures, authors in [9] worked on extraction of diverse-frequent patterns from transactional databases. Authors proposed a new interestingness measure called DiverseRank to rank frequent items on the basis of Diversity. A real world dataset has been used to test the effectiveness of the approach. It was found that the resultant patterns from the approach are different from the support-based approach. However, it is not clear as if the approach will be suitable for large size databases.

In another case study, authors in [10] emphasize the generic measures like support and confidence cannot be used to validate the association rules extracted in a multi-level mining environment. Moreover, such measures are also not able to extract interesting information for the users. Authors used diversity and peculiarity measures for interestingness evaluation in their work. BookCrossing dataset has been used to test the measures. Results are reported for both diversity and peculiarity measures for the same dataset. The results acknowledge that the measures can be used to identify the potential interesting rules from the datasets while mining at multiple levels in the concept hierarchy. The work done in this research only focuses on transactional databases.

Advanced measures of interestingness were used by [11]. The methodology proposed by the authors' extracted patterns from data cubes using Meta rule guided mining. The authors used advanced aggregated measures for measuring the interestingness of patterns. Lift and Loveinger measures are used to evaluate the interestingness of the extracted patterns. The advanced evaluation process worked as a module in the *OLEMAR* (Online Environment for Mining Association Rules) framework presented by the authors. The framework included visualization support for the visualizing the extracted patterns in an effective way. Usage of advanced measures of interestingness allows the possibility to un-ambiguously evaluate the extracted patterns.

Authors in [12] proposed the use of sixteen measures of interestingness which work in multi dimensional environment and are based on summarized data instead of typical transaction data. These measures are based upon diversity, dispersion, dominance and inequality. Authors provide results of these measures based upon a sample dataset. The results show that the order in which some of the measures rank summaries is highly correlated. This results in two groups of measures in which summaries are ranked in similar fashion. Authors emphasize that highly ranked summaries are a starting point for evluation of extracted knowledge. Authors further emphasize that the usage of all measures is tested against datasets of variety of population.

In another research study, authors in [13] worked on ranking summaries in data sets. Authors deduced that Rae,

Con and Hill measures satisfied the principals for ranking summaries. Moreover, these measures can further be used to evaluate the interestingness of the extracted knowledge. Authors worked on classification of diversity measures into groups by using Correlation. However, it is not clear that this classification will fit to all data sets or not.

The measures proposed by [13] were used by [14] in their research work which extracted patterns from a multi-dimensional schema. The authors applied Rae, Con and Hill measures in order to evaluate the extracted association rules for interestingness. It was found that the rules extracted from the technique were more diverse as compared to the rules extracted from the original data.

In a similar but theoretical work, [15] presented a conceptual model for mining of association rules at multiple levels of abstraction. The model was intended to create a multi-dimensional schema at different levels in the hierarchy which could be used to extract patterns in the form of association rules. Authors suggested that these patterns should be evaluated using advanced measures of interestingness, since these patterns are extracted in a multi-dimensional environment. The implementation of the model, however, was not included in the research work.

The model was implemented by the same authors in [3], which included two case studies on real world data sets. Authors extracted patterns using the proposed model in previous study and evaluated these patterns using advanced measures namely Rae, Con and Hill. The authors compared the increase in diversity (while mining using their model and actual data) with another study done on similar pattern. According to authors, the increase in diversity in their case was more than the previous approach. Same authors extended the concept by adding a visualization component in [16]. The evaluation of association was done using same measures that is Rae, Con and Hill in the enhanced version as well.

Although objective measures are used mainly in evaluation of patterns for interestingness, [5] emphasize that it is important to use subjective measures of interestingness of patterns along with the objective measures. From subjective point of view, a pattern may be unexpected or actionable and thus it can be interesting. Authors suggest that a pattern can be actionable and unexpected at the same time. Similarly, a pattern can be unexpected and non-actionable at the same time. The authors conclude that these two types of interestingness are in general, independent of each other.

It is clear from previous studies that usage of interestingness measures has been recently increased while mining in data warehouse environment. Interestingness of association rules in a data warehouse environment is being considered as a challenging task. Conciseness, Generality/Coverage, Reliability, Peculiarity, Diversity, Novelty, Surprisingness, Utility, and Actionability have been introduced theoretically. In some approaches, authors have used single or multiple statistical methods to address the issue of interestingness. Similarly, some authors have continued to use conventional approaches to measure the interestingness which were developed for transactional databases. Some authors have used advanced measures like Lift and Loveinger. In some other approaches diversity measures like Rae, Con and Hill are used. From previous studies, it is clear that there is a potential requirement of implementation of interestingness measures in association rule mining over data warehouses. Moreover, if mining is done in a multi-level environment, and correlation between different measures is found, the study leads to development of a linear model for prediction of diversity measures at each level in the hierarchy.

III. Measuring Diversity of Patterns Extracted From a Data Warehouse

It is evident from the previous studies that most authors believe in evaluation of extracted knowledge through interestingness measures. However since previously knowledge extraction was done in transactional databases, therefore some authors continued to use same measures in multi-dimensional environment. We believe that since the data is available in aggregate form in data warehouse environment, only the measures which are valid for data warehouse environment should be used. So there is a strong need to use advanced measures of interestingness for evaluation of extracted in data warehouse environment. It will also be interesting to see if there is correlation between different measures at cluster level while mining in multi-dimensional environment.

In [12] authors identified sixteen measures which can be used to rank the interestingness of summaries generated from a database. The measures are based uon the diversity criteria. The authors checked the effectiveness of these measures on a generic database. This work was further extended by [13] who added three more measures to the list. We now briefly explain 9 measures of interestingness which we intend to use for pattern evaluation in the data warehouse environment.

$$\text{Rae} = \sum_{i=0}^{m} \frac{\text{ni}(\text{ni}-1)}{\text{N}(\text{N}-1)}$$

$$\text{CON} = \sqrt{\frac{\left(\sum_{i=1}^{m} Pi^2\right) - \bar{q}}{1-\bar{q}}}$$

$$\text{Hill} = 1\text{-}\ \frac{1}{\sqrt{\sum_{\text{i=1}}^{\text{m}} \text{Pi}^3}}$$

$$\text{Variance} = \sqrt{\frac{\left(\sum_{i=1}^{m} (Pi-\bar{q})^2\right)}{m-1}}$$

$$\text{Simpson} = \sum_{i=1}^{m} Pi^2$$

$$\text{Mchintosh} = \frac{\text{N}-\sqrt{\sum_{\text{i=1}}^{\text{m}} ni^3}}{\text{N}-\sqrt{\text{N}}}$$

$$\text{Lorentzs} = \bar{q} \sum_{i=1}^{m} (m - i + 1)Pi$$

$$\text{Shutz} = \frac{\sum_{i=1}^{m} \mid Pi - \bar{q} \mid}{2\text{m}\bar{q}}$$

$$\text{Whittaker} = 1 - (0.5 \sum_{i=1}^{m} \mid Pi - \bar{q} \mid)$$

The measures described above were used to rank summaries in a generic database, however since the measures were used to work with summarized information, we believe that these measures can be used to deal with the interestingness measurement in data warehouse environment. In our case, knowledge is extracted in the form of association rules in a data warehouse environment. We now provide some detail on the model that we intend to use for our previous research work.

In our model, the dataset is taken in original form for mining purposes. The process is started by applying Agglomerative Hierarchical Clustering which uses numeric variables during the process. As a result, clusters of data are obtained at each level in the hierarchy. For instance, the whole data set is divided into two clusters C1 and C2 and at next level C1 is divided into C11 and C12. In order to apply pick top ranked variables at cluster level, we use PCA (Principal Component Analysis). We convert nominal variables to numeric variables using Rossarios's Approach [17]. At the end of this process, all clusters contain data in numerical format only. PCA is applied at each cluster level to rank variables. We group natural categories together within nominal variables to treat related categories as one group. In next step, a multi-dimensional schema (STAR schema) using top ranked variables. Nominal variables are taken as dimensions where as numeric variables are treated as facts in the schema. Data from each cluster is obtained within the data warehouse and Apriori Algorithm is applied on data warehouse to obtain association rules.

We now take an exemplary data set D related to Movies, Concerts and TV shows data taken from the IMDB database. The data set contains 154 records. Numeric variables include No. of Votes, Gross Income, Rating and Cast Count. The nominal variables include Genres, Company and Production Status. All steps are applied as described before and patterns are extracted in the form of association rules. Rules generated from Cluster C1 are shown in Table I.

We applied Rae, CON, Hill, Variance, Simpson, Mchintosh, Lorentz, Shutz and Whittaker measures proposed by authors in [12]. These measures are categorized as diversity measures. The values of these measures are shown in Table II below. The values in Table II show that the values of different measures have almost similar relationship with each other in both Rule Sets. However, nothing can be established as a fact without a reasonable number of rule sets. We proceed to a case study having large number of dimensions which results in a large number of rule sets. These rule sets can be used to confirm if there is a relationship between these measures across the rule sets.

The next section represents the implementation of diversity measures on a real world dataset taken from UCI machine learning website.

TABLE I: LIST OF RULES EXTRACTED – CLUSTER C1

S. No	Rules	Importance
1	If ProductionStatus[Group2]→Company [Paramount]	1.17
2	If Genres [Group1] & ProductionStatus [Group3]→Company [20th Century Fox]	1.06
3	If ProductionStatus [Group1] & Genres [Group2]→ Company [Paramount]	0.91

TABLE II: RESULTS OF DIVERSITY MEASURES AGAINST DIFFERENT RULE SETS IN CLUSTER C1

Rule Set	Rae	CON	Hill	Variance	Simpson	Mchintosh	Lorentz	Shutz	Whittaker
R1-R2	0.92	0.96	-0.09	0.01	0.89	-3315.5	6316.9	331.9	-3117.0
R3-R4	0.54	0.56	-1.43	0.03	0.49	-2735.1	4321.7	2657.0	-2615.9

IV. CASE STUDY

In this section, a case study on Adults dataset is presented which his taken from UCI Machine Learning Repository[18]. There are 13 variables in this data set including both nominal and numeric variables. The data set contains 8 nominal and 5 numeric variables. The data set consists of 48,482 rows in total. Initially, the dataset had missing values which were removed and 61% of the whole dataset (comprising of 30,162 rows) is used in the case study. For pattern extraction, all steps are applied as described before and patterns are extracted in the form of association rules. Rules generated from Cluster C1 are shown in Table III.

We now calculate the values of 9 diversity measures namely Rae, Con, Hill, Variance, Simpson, Mchintosh, Lorentz, Shutz and Whittaker. For the purpose of calculating these measures, we pass the support values to our prototype application and calculate all measures at once. The results are given in the Table III. Rule sets are given on left side, and diversity values against each measure are provided on right side. We focus on the values given at the end of the table where rule set size is increased. These values are shown in bold in the Table IV.

In order to find the relationship between these measures for the whole cluster, we created three chats. We kept Rae, Con, Simpson and Variance in a separate set as these val-

ues were very low as compared to the other measures. It can be observed from Fig. 1, that the relationship between all of these measures remains similar in each rule set. For instance, the value of Rae and Simpson measures is similar in every rule set. Moreover, Con measure has a value in between the value of Rae and Simpson in the respective rule set. Interestingly, the values of these measures drop as the rule set size increases. However, the inter-relationship between these values remains almost same in a rule set.

We created another comparison chart to monitor the relationship between Lorentz and Shutz measure. These measures are compared separetly, as the values are higher than other measures. The results of these measures are shown in Fig. 2. The results show that the relationship between both measures is almost similar across the rule sets. The chart shows that the value of Shutz remains lower than Lorentz across all rule sets in almost similar ratio.

We see that the given charts (Fig. 1, Fig. 2 and Fig. 3) depict that there is a relationship between different measures, and interestingly the proportion of relationship appears same in all rule sets. The results are passed to the statistical analysis tool called SPSS (Statistical Package for Social Science) to calculate Correlation Coefficients between different diversity measures.

The Correlation Coefficient has value between -1 and 1. If Correlation Coefficient is positive and higher than 0.50, there is a strong positive relationship between two variables. Positive relationship means that two variables are directly proportion to each other. Contrary to this, if the Correlation Coefficient is negative and higher than -0.50, there is a strong negative relationship between two variables. Negative relationship means that two variables are in-directly or inversely proportional to each other.

TABLE III: RULE GENERATED FROM OUR MODEL ALONG WITH IMPORTANCE VALUES

Rule	Importance
1) Occ. Group =*Group3*, Sex Group=*Group-Others*→Work Class = *Local-gov*	0.95
2) Occ. Group = *Group 3* → Work Class = *Local-gov*	0.95
3) Occ. Group =*Group3*, Sex Group =*Group-Others*→ Work Class = *Local-gov*	0.83
4) Occ. Group =*Group-Others*, Sex Group=*Group-Others*→Work Class = *Private*	0.11
5) Occ. Group = *Group-Others* → Work Class =*Private*	0.11
6) Occ. Group = *Group 3* → Work Class =*Private*	0.11
7) Occ. Group =*Group 3*,Sex Group =*Group-Others*→ Work Class =*Private*	0.11
8) Occ. Group =*Group 3*,Sex Group =*Group-Others*→ Work Class =*Private*	0.10
9) Occ. Group = *Group-Others* → Work Class =*Private*	0.09
10) Occ. Group =*Group-Others*, Sex Group =*Group-Others*→ Work Class =*Private*	0.09

Occ.* = Occupation

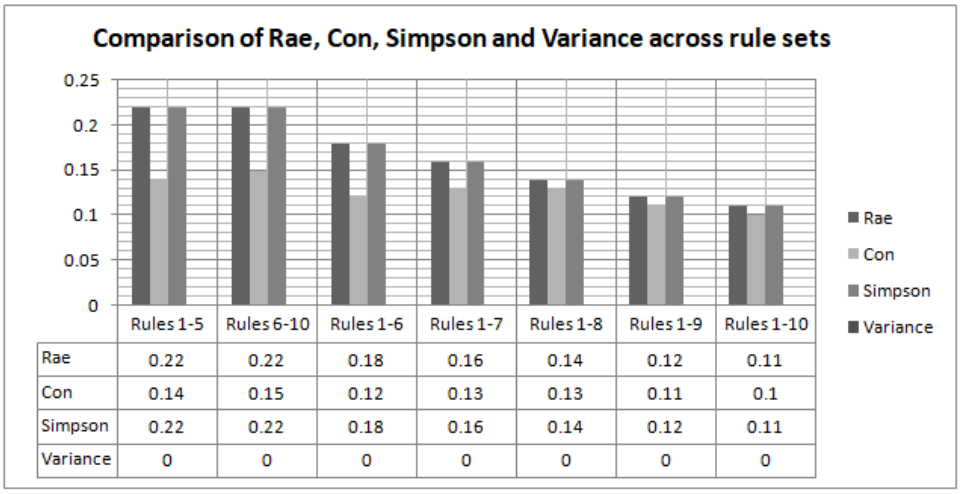
TABLE IV: RESULTS OF DIVERSITY MEASURES USING DIFFERENT RULE SETS IN CLUSTER C1

Rule Set size	Rules	Rae	Con	Hill	Variance	Simpson	Mchintosh	Lorentz	Shutz	Whittake
2	Rules 1-2	0.53	0.23	-0.86	0.03	0.53	-2876.74	5021.44	2695.00	-2694.00
	Rules 3-4	0.51	0.15	-0.94	0.01	0.51	-1835.39	3307.56	1766.50	-1765.50
	Rules 5-6	0.50	0.04	-0.99	0.00	0.50	-2934.48	5475.89	2892.00	-2891.00
	Rules 7-8	0.50	0.01	-1.00	0.00	0.50	-1302.32	2414.00	1277.50	-1276.50
	Rules 9-10	0.50	0.07	-0.99	0.00	0.50	-2217.42	4124.56	2175.00	-2174.00
3	Rules 1-3	0.37	0.23	-1.59	0.02	0.37	-2577.45	6442.21	3447.00	-3446.00
	Rules 4-6	0.34	0.11	-1.89	0.00	0.34	-2701.07	7201.79	3906.50	-3905.50
	Rules 7-9	0.36	0.21	-1.64	0.02	0.36	-1799.00	4453.79	2438.00	-2437.00
4	Rules 1-4	0.27	0.17	-2.53	0.01	0.27	-2454.51	8269.74	4461.50	-4460.50
	Rules 5-8	0.29	0.22	-2.31	0.01	0.29	-2426.31	7857.32	4169.50	-4168.50
5	**Rules 1-5**	**0.22**	**0.14**	**-3.49**	**0.00**	**0.22**	**-2599.03**	**10915.00**	**5967.00**	**-5966.00**
	Rules 6-10	**0.22**	**0.15**	**-3.42**	**0.00**	**0.22**	**-2134.52**	**8890.38**	**4839.00**	**-4838.00**
6	**Rules 1-6**	**0.18**	**0.12**	**-4.48**	**0.00**	**0.18**	**-2634.06**	**13385.59**	**7353.50**	**-7352.50**
7	**Rules 1-7**	**0.16**	**0.13**	**-5.16**	**0.00**	**0.16**	**-2525.57**	**14638.97**	**7996.50**	**-7995.50**
8	**Rules 1-8**	**0.14**	**0.13**	**-5.84**	**0.00**	**0.14**	**-2433.14**	**15853.79**	**8631.00**	**-8630.00**
9	**Rules 1-9**	**0.12**	**0.11**	**-6.83**	**0.00**	**0.12**	**-2420.71**	**17912.82**	**9791.50**	**-9790.50**
10	**Rules 1-10**	**0.11**	**0.10**	**-7.78**	**0.00**	**0.11**	**-2384.38**	**19739.57**	**10806.0**	**-10805.00**

The results of Correlation Analysis performed in SPSS are given in Table V. All measures are given on top as well as on the left hand side of the table. The intersection of a top measure with the measure on left has value of Correla-

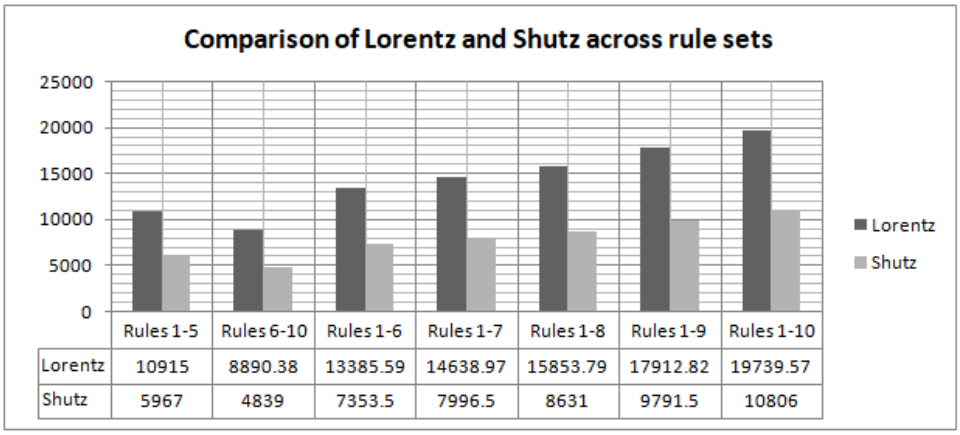
tion Coefficient between the two. For example 3rd cell in the 3rd row has Correlation Coefficient (0.917) of Rae measure with Hill measure.

Some of the values have Correlation Coefficient lie between 0.50 and -0.50. Moreover, some values of Correlation Coefficient are greater than 0.50 or greater than -0.50. These show a strong relationship. These values are highlighted in the Table III. For example Rae measure has a strong Correlation with Hill, Simpson, Lorentz, Shutz and Whittaker with C.C values 0.917, 1.0, -0.916, -0.917 and -0.917 respectively. Con measure has a strong Correlation with Variance only with a C.C value 0.783. Hill measure has strong Correlation with Rae, Simpson, Lorentz, Shutz and Whittaker with C.C. values 0.917, 0.917, -0.985, -0.986 and -0.986 respectively. Similarly Correlation between other different diversity measures is given in the same table.



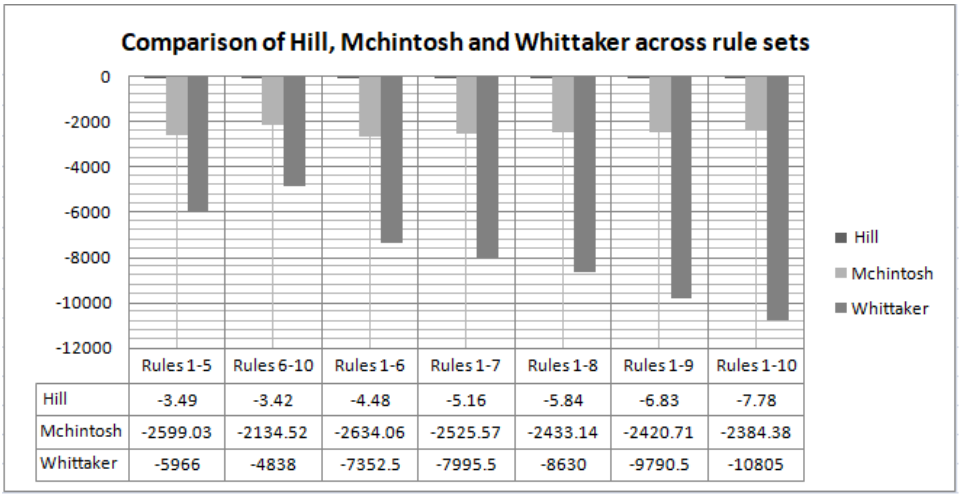
	Rules 1-5	Rules 6-10	Rules 1-6	Rules 1-7	Rules 1-8	Rules 1-9	Rules 1-10
Rae	0.22	0.22	0.18	0.16	0.14	0.12	0.11
Con	0.14	0.15	0.12	0.13	0.13	0.11	0.1
Simpson	0.22	0.22	0.18	0.16	0.14	0.12	0.11
Variance	0	0	0	0	0	0	0

Fig. 1. Comparison of Rae, Con, Simpson and Variance measures across rule sets in cluster C1.



	Rules 1-5	Rules 6-10	Rules 1-6	Rules 1-7	Rules 1-8	Rules 1-9	Rules 1-10
Lorentz	10915	8890.38	13385.59	14638.97	15853.79	17912.82	19739.57
Shutz	5967	4839	7353.5	7996.5	8631	9791.5	10806

Fig. 2. Comparison of Lorentz and Shutz measures across rule sets in cluster C1.



	Rules 1-5	Rules 6-10	Rules 1-6	Rules 1-7	Rules 1-8	Rules 1-9	Rules 1-10
Hill	-3.49	-3.42	-4.48	-5.16	-5.84	-6.83	-7.78
Mchintosh	-2599.03	-2134.52	-2634.06	-2525.57	-2433.14	-2420.71	-2384.38
Whittaker	-5966	-4838	-7352.5	-7995.5	-8630	-9790.5	-10805

Fig. 3. Comparison of Hill, Mchintosh and Whittaker measures across rule sets in cluster C1.

In this case study, we have calculated 9 diversity measures using the support values calculated from the as-

sociation rules extracted from the data warehouse. Moreover, we have shown that some of these measures are strongly correlated with each other. This relationship looked to be significant through charts. However, we have performed Correlation Analysis in order to confirm the correlation. Further analysis of the results was not done due to time limitations. We believe that this study can be extended in few directions. Firstly, this correlation can be checked at different levels in the hierarchy for each cluster data. An analysis can be performed in the results to see how correlation behaves at different levels. Secondly, Correlation Analysis only confirms that there is a relationship between measures. If Correlation is strong, we can extend the study to develop a linear regression model. Such model will help to predict different diversity measure values using some known diversity measure values. For example, if we know the values of Hill and Mchintosh, the linear model can provide us a value of Rae measure. Another possible direction is to perform this predictive linear model for un-limited number of hierarchies/levels.

TABLE V: PEARSON'S CORRELATIONS COEFFICIENT OF DIVERSITY MEASURES IN CLUSTER C1

Pearson's Correlation Coefficients of Diversity Measures									
	Rae	**Con**	**Hill**	**Variance**	**Simpson**	**Mchintosh**	**Lorentz**	**Shutz**	**Whittaker**
Rae	1	-.073	**.917****	.436	**1.000****	.211	**-.916****	**-.917****	**.917****
Con	-.073	1	.110	**.783****	-.073	-.232	-.069	-.070	.070
Hill	**.917****	.110	1	.466	**.917****	.173	**-.985****	**-.986****	**.986****
Variance	.436	**.783****	.466	1	.436	-.068	-.441	-.441	.441
Simpson	**1.000****	-.073	**.917****	.436	1	.211	**-.916****	**-.917****	**.917****
Mchintosh	.211	-.232	.173	-.068	.211	1	-.332	-.328	.328
Lorentz	**-.916****	-.069	**-.985****	-.441	**-.916****	-.332	1	**1.000****	**-1.000****
Shutz	**-.917****	-.070	**-.986****	-.441	**-.917****	-.328	**1.000****	1	**-1.000****
Whittaker	**.917****	.070	**.986****	.441	**.917****	.328	**-1.000****	**-1.0****	1

**. Correlation is significant at the 0.01 level (2-tailed).

V. CONCLUSION

In this paper we have applied 9 diversity measures on patterns extracted from a data warehouse environment. We have seen that the diversity measures have a strong correlation with each other in all rule sets in a cluster. Since we mine in a multi-level environment, we used statistical method called Correlation Coefficient to confirm this relationship. In future, Linear Regression can be utilized to create a linear model between diversity measures. Such model can help to predict diversity values at different levels using values of the diversity in the above levels in the hierarchy.

REFERENCES

[1] N. Kumar, A. Gangopadhyay, S. Bapna, G. Karabatis, and Z. Chen, "Measuring interestingness of discovered skewed patterns in data cubes," *Decision Support Systems*, vol. 46, pp. 429-439, 2008.

[2] X. H. Huynh, "Interestingness measures for association rules in a KDD process: Postprocessing of rules with ARQAT tool," *Université de Nantes*, 2006.

[3] M. Usman, "Multi-level mining of *association* rules from warehouse schema," *Kuwait Journal of Science*, vol. 44, 2017.

[4] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[5] A. Silberschatz and A. Tuzhilin, "On subjective measures of interestingness in *knowledge* discovery," *KDD*, pp. 275-281.

[6] B. Padmanabhan and A. Tuzhilin, "Unexpectedness as a measure of interestingness in knowledge discovery," *Decision Support Systems*, vol. 27, pp. 303-318, 1999.

[7] S. Sahar, "What is interesting: studies on interestingness in knowledge discovery," Ph.D thesis, School of Computer Science, Tel-Aviv University, 2003.

[8] A. Chandanan and M. Shukla, "Data mining for qualitative dataset Using *association* rules: A review," *International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARCSEE)*, vol. 2, pp. pp. 231-238, 2013.

[9] S. Srivastava, R. U. Kiran, and P. K. Reddy, "Discovering diverse-frequent patterns in transactional databases," in *Proc. of the 17th International Conference on Management of Data*, 2011, p. 14.

[10] G. Shaw, Y. Xu, and S. Geva, "Interestingness measures for multi-level association rules," in *Proc. of ADCS 2009*, pp. 27-34, 2009.

[11] R. B. Messaoud, S. L. Rabaséda, R. Missaoui, and O. Boussaid, "OLEMAR: An online environment for mining association rules in multidimensional data," in *Data Mining and Knowledge Discovery Technologies*, ed. IGI Global, pp. 1-35.

[12] R. J. Hilderman and H. J. Hamilton, "Heuristic measures of interestingness," in *European Conference on Principles of Data Mining and Knowledge Discovery*, 1999, pp. 232-241.

[13] N. Zbidi, S. Faiz, and M. Limam, "On mining summaries by objective measures of interestingness," *Machine Learning*, vol. 62, pp. 175-198, 2006.

[14] M. Usman, R. Pears, and A. C. M. Fong, "Discovering diverse association rules from *multidimensional* schema," *Expert Systems with Applications*, vol. 40, pp. 5975-5996, 2013.

[15] M. Usman, M. Usman, and W. Ahmad, "A conceptual model for multi-level mining and visualization of association rules," in *Proc.* of *2014 Ninth International Conference on Digital Information Management (ICDIM)*, pp. 175-181.

[16] M. Usman and M. Usman, "Multi-Level mining and visualization of informative association rules," *J. Inf. Sci. Eng.*, vol. 32, pp. 1061-1078, 2016.

[17] G. E. Rosario, E. A. Rundensteiner, D. C. Brown, M. O. Ward, and S. Huang, "Mapping nominal values to numbers for *effective* visualization," *Information Visualization*, vol. 3, pp. 80-95, 2004.

[18] A. Asuncion and D. Newman, *UCI Machine Learning Repository*, 2007.

Muhammad Usman was born in Attock, Punjab, Pakistan (1982). He did masters in science (Computer Science) (MSCS) from SZABIST, Islamabad (2002), masters in computer science (MCS) from International Islamic University (IIU), Islamabad (2006) and B.Sc(Mathematics, Statistics) from University of the Punjab, Lahore(2002).

He is currently working as database administrator at PASTIC National Center, Islamabad, Pakistan since Sep. 2017. He worked as web manager at the same office since Apr. 2010. Previously he worked in capacity as a project manager and software engineer in different organizations for 5 years. He is also working as freelancer at freelancer.com since 2008. He has completed 760+ web development projects and currently ranked at 451 out of 26.0 million all over the world. He has majorly worked in publishing research-centric data over the internet with different tools and technologies. He has recently published a book titled "Predictive Analysis on Large Data for Actionable Knowledge: Emerging Research and Opportunities" (IGI Global, 2018). Apart from this, he has published his work in international conferences and journals in the past. The research work majorly includes prediction and extraction of patterns from large datasets.

Muhammad Usman is involved in editorial committee for Pakistan Journal for Computing and Information Science (PJCIS) for primary recommendations of machine learning research papers for the journal. He is also a member of High School Summer Science Program (HSSSRP) for High School Students in Pakistan for mentoring/evaluation of student projects. He received intel appreciation award for evaluation of projects in 2015. He has trained over 1200 researchers and scientists on Statistical Package for Social Science (SPSS) during the research training activities at PASTIC National Center, Islamabad.