International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-6, Issue-4, July-2018
DOI: 10.21276/ijircst.2018.6.4.4

Regression and Correlation Analysis of Different Interestingness Measures for Mining Association Rules

Mir Md Jahangir Kabir, Tansif Anzar

***Abstract*— Association Rule Mining is the significant way to extract knowledge from data sets. The association among the instance of a dataset can measured with Interestingness Measures (IM) metrics. IM define how much interesting the extract knowledge is. Researchers have proved that the classical Support-Confidence metrics can't extract the real knowledge and they have been proposing different IM. From a user perspective it's really tough to select the minimal and best measures from them. From our experiment, the correlation among the various IM such as Support, Confidence, Lift, Cosine, Jaccard, Leverage etc. are evaluated in different popular data sets. In this paper our contribution is to find the correlation among the IM with different ranges in different types of data sets which were not applied in past researches. This study also identified that the correlation varies from data set to data set and proposed a solution based on multiple criterion that will help the users to select the minimal and best from a large number of IM.**

***Index Terms*—Association rules, correlation, interestingness measures, regression analysis.**

I. INTRODUCTION

A large number of measures are available for evaluating an association rule, some of those are conflicting and others get similar ranking [1]. If two interestingness measures have same ranking, then the rules generated by using these measures are same or in other words the rules are redundant [2].

By following this idea this study focusing on finding the similarities and correlation and regression analysis among the measures. Finally, this research introduces different ranges based on the correlation values of different measures. To conduct this research different data sets such as balance scale, Monks, Nursery etc. are used for getting regression and correlation values of different interestingness measures.

In this research study, the following tasks are performed 1) All the examined measures are theoretically defined and their characteristics are explained, 2) Interestingness measures which will be evaluated are applied for different datasets, 3) Regression analysis and correlation among the measures are processed and 4) Measures are grouped into different ranges based on correlation values.

The following sections are organized as follows: Interestingness measures which are evaluated for different data sets are described in section two. Experimental set of this research and the data sets which are used for this experiments are explained in section three. Result analysis is shown through section four which are followed by conclusion and future work part of section five.

Manuscript received July 15,2018

Mir Md. Jahangir Kabir, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh, (e-mail: mmjahangir.kabir@gamil.com).

Tansif Anzar, Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh,

II. INTERESTINGNESS MEASURES

In this section the significant interestingness measures will be explained. These measures are used one different data sets for evaluating association rules. The following measures such as support, confidence, cosine, imbalance ratio, jaccard, leverage, lift are described.

Support [3]:

The frequency of an item set X is defined by the support value of X [4]–[6]. If a transaction t contains by the dataset D and X is a subset of that transaction t, then the support value of X is defined by the following equation,

$$supp(X) = \frac{|\{t \in D; X \subseteq t\}|}{|D|} = Prob(X)$$

Support value is used to define how frequently the items occur in a data set D [7]. The ranges of support value is 0 to 1. Now if we want to find the support value of an association rule of X=>Y, then the support value of X and Y is divided by the total number of transactions.

$$supp(X \Rightarrow Y) = supp(X \cup Y) = Prob(X \cap Y)$$

Confidence [3]:

Another interestingness measure is used to evaluate an association rule, named confidence. It gives the

Copyright © 2018. Innovative Research Publications. All Rights Reserved

probability of a rule of checking conditions of the consequent for a transaction which contains the antecedent [8]. The ranges of a confidence varies from 0 to 1.

$$conf(X => Y) = \frac{supp(X => Y)}{supp(X)} = \frac{supp(X \cup Y)}{supp(X)} = \frac{Prob(X \cap Y)}{P(X)}$$

Cosine [1]:

It gives the probability of a rule of checking conditions of the consequent for a transaction which contains the checking of antecedent. The ranges of a cosine varies from 0 to 1.

$$cosine(X => Y) = \frac{supp(X \cup Y)}{\sqrt{supp(X)\,supp(Y)}} = \frac{Prob(X \cap Y)}{\sqrt{Prob(X)Prob(Y)}}$$

Imbalance Ratio [1] :

The presence of L.H.S and R.H.S in a transaction and the degree of imbalance ratio between these two events is evaluated by a measure, called imbalance ratio (IR). The similarity and dissimilarity of conditional probabilities are defined by the value of imbalance ratio 0 and 1, respectively.

$$IB(X => Y) = \frac{|supp(X) - supp(Y)|}{supp(X) + supp(Y) - supp(X)\,supp(Y)}$$

Jaccard Co efficient [1]:

The similarity between two occurrences is defined by the Jaccard coefficient which is shown by the following equation. The ranges of this measure varies from -1 to 1. The similarity, dissimilarity and independence of an event are defined by the value 1, -1 and 0, respectively.

$$jaccard(X => Y) = \frac{supp(X \cup Y)}{supp(X) + supp(Y) - supp(X \cup Y)}$$

$$= \frac{Prob(X \cap Y)}{Prob(X) + Prob(Y) - Prob(X \cap Y)}$$

Leverage [9]:

It measures the difference between occurring together of events and their independent occurring. The ranges of this measure varies from -1 to 1. The similarity, dissimilarity and independence of an event are defined by the value 1, -1 and 0, respectively.

$$leverage(X => Y) = supp(X => Y) - supp(X)(Y)$$
$$= Prob(X \cap Y) - Prob(X)P(Y)$$

Lift [10]:

The ratio of two events X and Y occurred together divided by their independent support value is given by this measure. The strong relation between the events depends on the large value of lift.

$$lift(X => Y) = lift(Y => X) = \frac{conf(X => Y)}{supp(Y)}$$
$$= \frac{conf(Y => X)}{supp(X)} = \frac{Prob(X \cap Y)}{Prob(X)P(Y)}$$

For this study Spearman rank correlation approach is applied for analyzing the correlation between two variables. This approach is denoted by the mathematical symbol rho ρ, is a non-parametric measure for evaluating the degree of correlation among the interestingness measures (IM) [11].

III. Experimental Setup

To conduct the experiments the following configurations are used. For this research, the most popular algorithm named Apriori is used for rule generation, Spearman method is used for correlation analysis and linear regression is used for the checking of linear dependency among the interestingness measures. The algorithm is coded by using Python 3.6.

Fourth generation intel core processors along with 4 GB RAM, mobile intel HM87 express and intel HD graphics 4600 are used for performing the experiments. The following data sets such as Balance Scale, Monk's Problems and Nursery are used for conducting this study. These data sets are taken from the UCI machine learning repository. The parameters of these data sets are shown through the following table.

Table 1: Parameters of different data sets

Parameters	Name of the Data Sets		
	Balance Scale	Monk's Problems	Nursery
No of records	625	431	12,960
No of attributes	23	19	32

IV. Experimental Results

In this section linear regression and correlation are shown and analysed through the experimental results. The selected measures are applied on different data sets to get the results. Through linear regression the users can predict one interestingness measures knowing others.

A. Linear Regression Analysis

For linear regression analysis selected measures are applied on different data sets. Here we show how one variable is linearly dependent on others. Fig 1-8 shows the regression analysis of different measures.

Copyright © 2018. Innovative Research Publications. All Rights Reserved

International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-6, Issue-4, July-2018
DOI: 10.21276/ijircst.2018.6.4.4

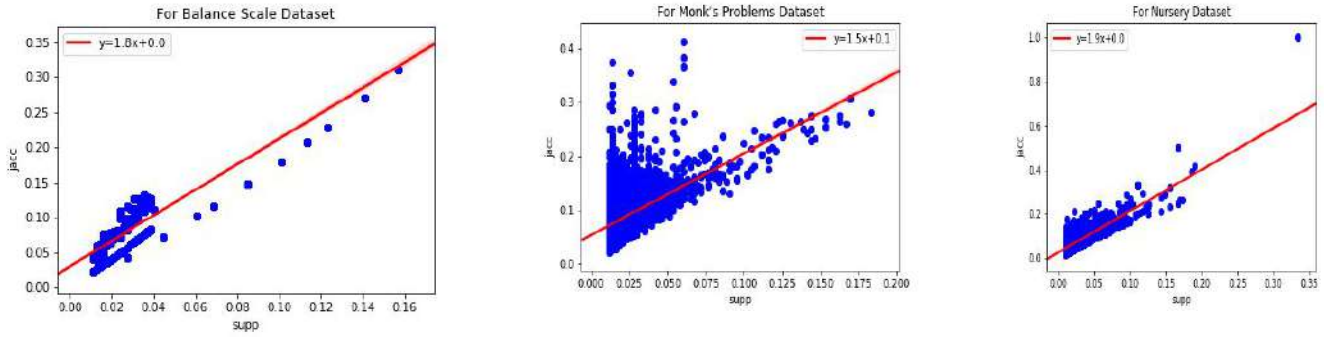

Fig 1. Regression analysis of Support Vs Jaccard for three data sets

In Fig. 1 linear relation among Support and Jaccard are established. If support is increased one unit then jaccard is increased about 1.5 times on average for three datasets.

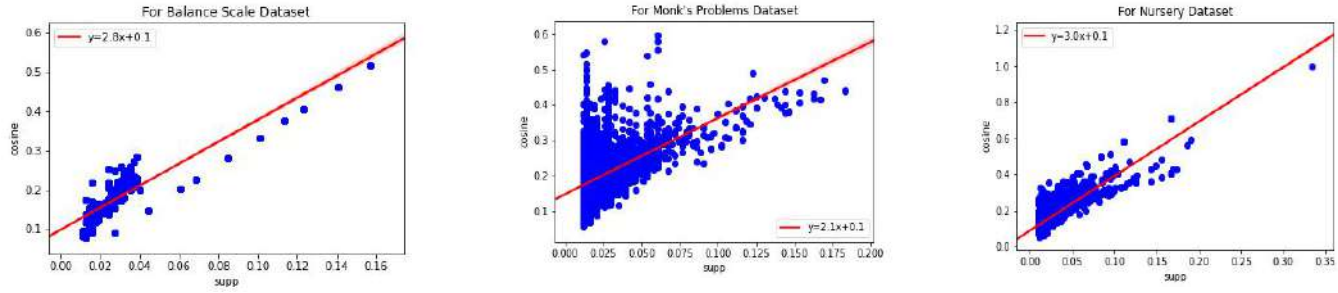

Fig 2. Regression analysis of Support Vs Cosine for three data sets

In Fig. 2 linear relation among Support and Cosine are established. If support is increased one unit then cosine is increased about 2.5 times on average for three datasets**.**

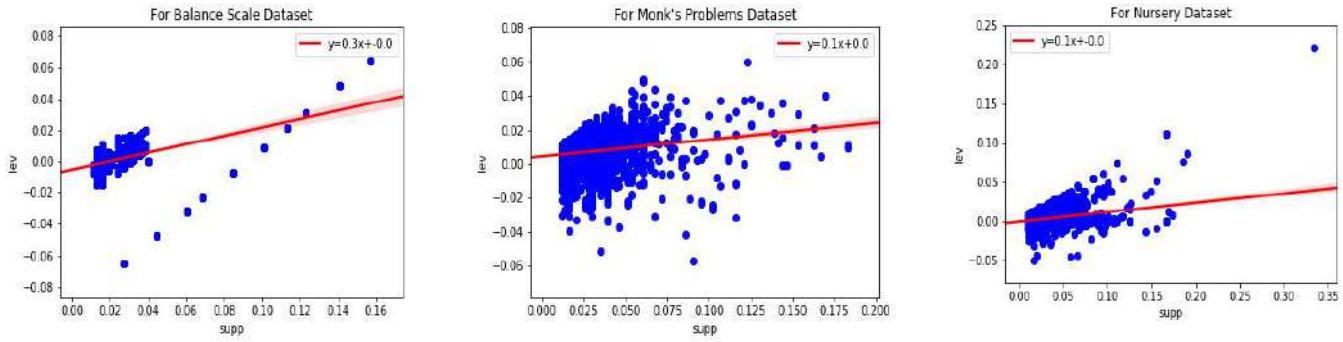

Fig 3. Regression analysis of Support Vs Leverage for three data sets

In Fig.3 if support is increased one unit then leverage is increased about 0.1 times on average for three datasets. So, there exist almost no linear relationship between them.

Copyright © 2018. Innovative Research Publications. All Rights Reserved

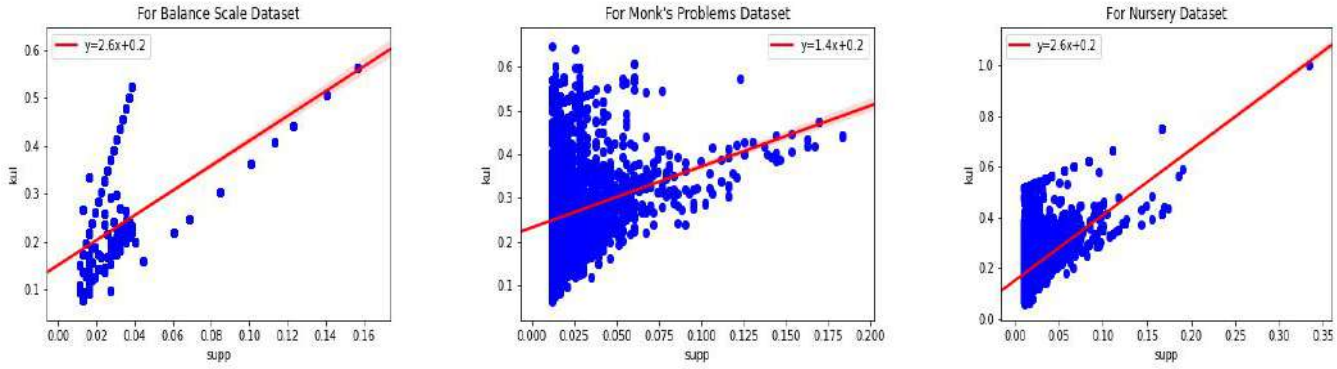

Fig 4. Regression analysis of Support Vs Kulczynski for three data sets

In Fig.4 linear relation among Support and Kulczynski are established. If support is increased one unit then kulczynski is increased about more than 2 times on average for three datasets.

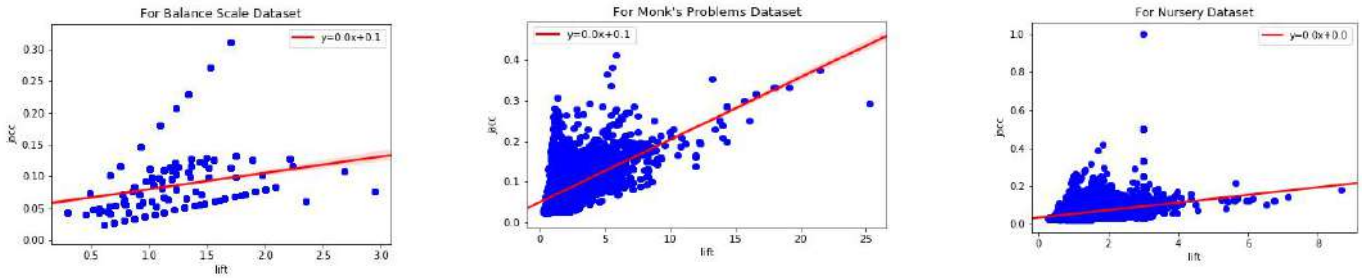

Fig 5. Regression analysis of Lift Vs Jaccard for three data sets

Fig.5 shows that Lift and Jaccard are almost not linearly related.

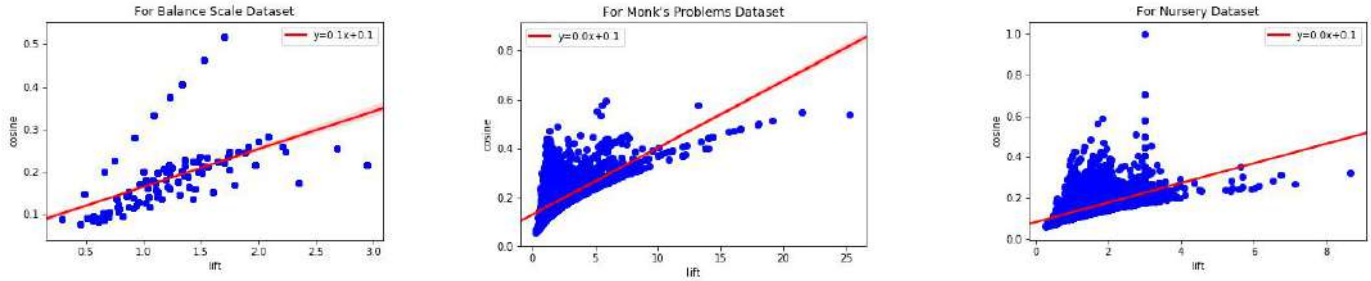

Fig 6. Regression analysis of Lift Vs Cosine for three data sets

Fig. 6 shows that Lift and Cosine are almost not linearly related.

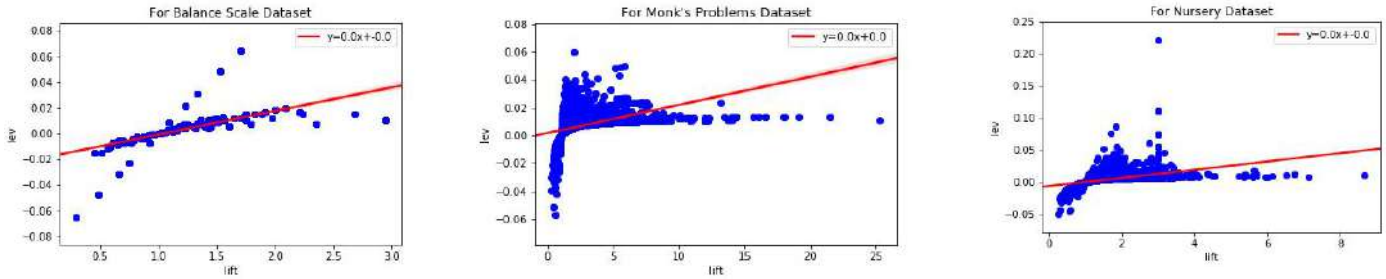

Fig 7. Regression analysis of Lift Vs Leverage for three data sets

In Fig.7 it shows that Lift and Leverage are almost not linearly related

Copyright © 2018. Innovative Research Publications. All Rights Reserved

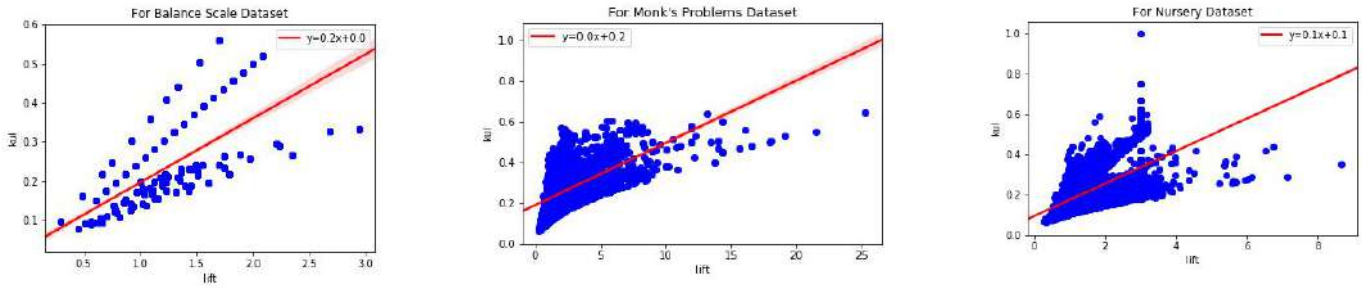

Fig 8. Regression analysis of Lift Vs Kulczynski for three data sets

In Fig.8 it shows that Lift and Kulczynski are almost not linearly related.

A. Correlation Analysis

In this section, the correlation among different measures are shown through experimental results. The correlated values are grouped into different ranges. If the correlated values in between the range 0.75-1.00, then this range is classified into group A. Similarly, others ranges are classified into different groups which are shown through the following tables.

Table 2: Correlation values are grouped into different ranges

Correlation Group							
A	**B**	**C**	**D**	**-A**	**-B**	**-C**	**-D**
0.75-1.00	**0.50-0.74**	**0.25-0.49**	**0.1-0.24**	**-(0.75-1.00)**	**-(0.50-0.74)**	**-(0.25-0.49)**	**-(0.1-0.24)**

To find the correlation among different measures this study applied Spearman rank correlation methods. In this experiment, the correlated measures are shown through the following tables.

Table 3: Correlation between support and different interestingness measures for different data sets

		Datasets		
Support Vs	**Interesting Measures**	**Balance**	**Monk**	**Nursery**
	Confidence	**A**	**D**	**D**
	Cosine	**B**	**C**	**B**
	IR	**-B**	**-D**	**-D**
	Jaccard	**A**	**C**	**B**
	Kul	**C**	**D**	**C**
	Leverage	**D**	**D**	**D**
	Lift	**D**	**-D**	**D**

In Balance Scale Dataset, support and confidence are strongly correlated. So user can use one of them. In Monk Data set,Support and Lift are negatively co related. So user have to use both of them.In Nursery Data set,Support and IR are negatively co related. So user have to use both of them.

Table 4: Correlation between confidence and different interestingness measures for different data sets

		Datasets		
Confidence Vs	**Interesting Measures**	**Balance**	**Monk**	**Nursery**
	Support	**A**	**D**	**D**
	Cosine	**C**	**B**	**C**
	IR	**-B**	**B**	**-B**
	Jaccard	**A**	**A**	**A**
	Kul	**D**	**D**	**B**
	Leverage	**D**	**B**	**B**
	Lift	**D**	**C**	**B**

In Balance Scale Dataset, Confidence, Support and Jaccard are strongly co related. So user can use one of them. In Monk Dataset, Confidence and Jaccard are strongly co related. So user can use one of them. Similarly in Nursery Dataset, Confidence and Jaccard are strongly co related. So user can use one of them.

Copyright © 2018. Innovative Research Publications. All Rights Reserved

Table 5: Correlation between cosine and different interestingness measures for different data sets

		Datasets		
	Interesting Measures	Balance	Monk	Nursery
Cosine Vs	Support	B	C	B
	Confidence	C	B	C
	IR	-D	-D	-D
	Jaccard	A	A	A
	Kul	A	A	B
	Leverage	B	A	B
	Lift	B	C	B

In Balance Scale Dataset, Cosine, Jaccard, and Kul are strongly correlated. So user can use any of them to mine an association rule. In Monk Dataset, Cosine, Jaccard, Kul and Leverage strongly co related. So user can use one of them. In Nursery Dataset, Cosine and Jaccard are strongly correlated. So user can chose one of them. Through this way, users could avoid the generation of redundant association rules.

Table 6: Correlation between imbalance ratio (IR) and different interestingness measures for different data sets

		Datasets		
	Interesting Measures	Balance	Monk	Nursery
IR Vs	Support	-B	-D	-D
	Confidence	-B	-B	-B
	Cosine	-D	-D	-D
	Jaccard	-B	-B	-B
	Kul	C	C	B
	Leverage	C	-D	D
	Lift	C	-D	D

In Balance Scale Dataset, IR and Support are negatively co related. So users have to use both of them. This is also true for Monk and Nursery data set. Most of the measures are either negatively or loosely correlated. So users should use all of the measures for evaluating the association rules.

Table 7: Correlation between jaccard and different interestingness measures for different data sets

		Datasets		
	Interesting Measures	Balance	Monk	Nursery
Jaccard Vs	Support	A	C	B
	Confidence	A	A	A
	Cosine	A	A	B
	IR	-B	-B	-B
	Kul	C	C	D
	Leverage	C	B	C
	Lift	C	B	C

In Balance Scale Dataset, Jaccard, Support, Confidence and Cosine are strongly correlated. So user can use one of them for mining an association rule. Jaccard, Confidence and Cosine are strongly correlated in Monk Dataset. So user can use one of them. On the other hand, Jaccard and Confidence are strongly correlated for Nursery data set.

Table 8: Correlation between kul and different interestingness measures for different data sets

		Datasets		
	Interesting Measures	Balance	Monk	Nursery
Kul Vs	Support	C	D	C
	Confidence	D	D	-D
	Cosine	A	A	B
	IR	C	C	B
	Jaccard	C	C	D
	Leverage	B	B	C
	Lift	B	B	C

In Balance Scale Dataset, Kul and Cosine are strongly correlated. So user can use one of them. This is also true for Monk data set. In Nursery Dataset, Kul and Confidence are negatively co related. So user have to use both of them.

Table 9: Correlation between leverage and different interestingness measures for different data sets

		Datasets		
	Interesting Measures	Balance	Monk	Nursery
Leverage Vs	Support	D	D	D
	Confidence	D	B	C
	Cosine	B	-A	B
	IR	C	D	D
	Jaccard	C	B	C
	Kul	B	A	C
	Lift	A	A	A

In Balance Scale Dataset, Leverage and Lift are strongly correlated. So user can use any one of them. In Monk Dataset, Leverage, Kul and Lift are strongly correlated. So user can use one of them among the three measures. In Nursery Dataset, Leverage and Lift are strongly co related.

Table 10: Correlation between lift and different

Copyright © 2018. Innovative Research Publications. All Rights Reserved

interestingness measures for different data sets

		Datasets		
	Interesting Measures	**Balance**	**Monk**	**Nursery**
Lift Vs	**Support**	**D**	**-D**	**D**
	Confidence	**D**	**C**	**C**
	Cosine	**B**	**B**	**B**
	IR	**C**	**-D**	**D**
	Jaccard	**C**	**B**	**C**
	Kul	**B**	**B**	**C**
	Leverage	**A**	**A**	**A**

In this table Lift and Leverage are strongly correlated for all three data sets. So user can use any one of them for the recommended data sets.

V. Conclusion and Future works

In this article we have worked with seven interestingness measures (IM) of association rules. First of all those measures are theoretically defined and their characteristics and nature are explained. After that those selected IM are applied in different data sets for analysing regression and correlation among those measures. From the analysis it can conclude that linear regression and correlation do not provide the same result every time and varies from data sets to data sets. We tried to group them in different ranges from correlation values. From this, a user can select an IM from a group and don't need to use the rest of the measures if they are correlated of that group for finding association rules. In future researchers could use more data sets from different domain and metadata will be given priority as the users are involved in decision making.

References

[1] P. N. Tan, V. Kumar, and J. Srivastava, "Selecting the right objective measure for association analysis," Inf. Syst., vol. 29, no. 4, pp. 293–313, 2004.

[2] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton, "Behavior-based clustering and analysis of interestingness measures for association rule mining," Data Min. Knowl. Discov., vol. 28, no. 4, pp. 1004–1045, 2014.

[3] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.

[4] D. Martin, A. Rosete, J. Alcala-Fdez, and F. Herrera, "A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules," IEEE Trans. Evol. Comput., vol. 18, no. 1, pp. 54–69, 2014.

[5] M. M. J. Kabir, S. Xu, B. H. Kang, and Z. Zhao, "A new evolutionary algorithm for extracting a reduced set of interesting association rules," in 22nd International Conference On Neural Information Processing, 2015, pp. 133–142.

[6] M. M. J. Kabir, S. Xu, B. H. O. Kang, and Z. Zhao, "Association Rule Mining for Both Frequent and Infrequent Items Using Particle Swarm Optimization Algorithm," Int. J. Comput. Sci. Eng., vol. 6, no. 07, pp. 221–231, 2014.

[7] M. M. J. Kabir, S. Xu, B. H. Kang, and Z. Zhao, "Comparative analysis of genetic based approach and apriori algorithm for mining maximal frequent item sets," in IEEE Congress on Evolutionary Computation, 2015, pp. 39–45.

[8] M. M. J. Kabir, S. Xu, B. H. Kang, and Z. Zhao, "A new multiple seeds based genetic algorithm for discovering a set of interesting Boolean association rules," Expert Syst. Appl., vol. 74, pp. 55–69, 2017.

[9] G. Piatetsky-Shapiro, "Discovery, analysis, and presentation of strong rules," in Knowledge Discovery in Databases, Menlo Park, Calif, USA: AAAI/MIT Press, 1991, pp. 229–248.

[10] S. Birn, R. Motwani, J. . Ullman, and S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data," in Proceeding of the ACM SIGMOD, 1997, pp. 255–264.

[11] M. M. Mukaka, "Statistics corner: A guide to appropriate use of correlation coefficient in medical research," Malawi Med. J., vol. 24, no. 3, pp. 69–71, 2012.

Author's Profile

Mir Md Jahangir Kabir is an Associate Professor of Computer Science and Engineering Department, Rajshahi University of Engineering and Technology, Bangladesh. He received B.Sc., M.Sc. and PhD degrees, from Rajshahi University of Engineering and Technology, Bangladesh, University of Stuttgart, Germany and University of Tasmania, Australia in 2004, 2009, and 2016 respectively. After working as a Lecturer (from 2004), he was an Assistant Professor (from 2010) in the Dept. of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. He joined as an Associate Professor (from 2017) in the same department of that University. He received an Overseas Postgraduate Research Award from the Australian government in 2013 to research in PhD. His research interests include the theory and applications of Data Mining, Genetic Algorithm, Machine Learning and Artificial Intelligence.

Tansif Anzar is a final year student and pursuing undergraduate thesis under the supervision of Associate Professor Dr. Mir Md Jahangir Kabir at the Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Bangladesh. He won several national prizes and shortlisted for different international competitions. He was awarded Excellence Award from his department for his outstanding performance in the year of 2017. His field of interest include Data Mining, Data Analytics and Business Intelligence.

Copyright © 2018. Innovative Research Publications. All Rights Reserved