Journal of Computer Science 10 (9): 1672-1679, 2014
ISSN: 1549-3636

doi:10.3844/jcssp.2014.1672.1679 Published Online 10 (9) 2014 (http://www.thescipub.com/jcs.toc)

SELECTING PERFECT INTERESTINGNESS MEASURES BY COEFFICIENT OF VARIATION BASED RANKING ALGORITHM

[1]Selvarangam, K. and [2]K. Ramesh Kumar

[1]Department of Computer Science and Engineering, Hindustan University, Chennai, India
[2]Department of Information Technology, Hindustan University, Chennai, India

Received 2014-02-17; Revised 2014-02-24; Accepted 2014-04-17

ABSTRACT

Ranking interestingness measure is an active and essential research domain in the process of knowledge discovery from the extracted rules. Since various measures proposed by many researchers in various situations increases the list of measures and these are not able to use as a common measures to evaluate the rules, knowledge finders are not able to identify a perfect measure to ensure the actual knowledge on database. In this study, we presented about a ranking method to identify a perfect measure, which also reduces the number of measures. Ranking will be done by increasing order of Coefficient of Variation (CV) and not applicable measures are eliminated. Also we introduced heuristic association measures, U cost, S cost, R cost, T combined cost and ranked with existing measures using CV based ranking algorithm, our measures are placed in better position on ranking, compared with the existing measures.

Keywords: Pattern, Interestingness Measures, Association Rule, Coefficient of Variation

1. INTRODUCTION

The process of Knowledge Discovery in Data (KDD) includes a collection of components to identify or to extract the new patterns from the real data. The components in a knowledge discovery system may differ from each other, but some of the principle functions of knowledge discovery systems are control, data interface, focus, pattern extraction, evaluation and knowledge base. Interest and utility are considered as two important aspects in the process KDD. The evaluation metric, evaluates the interest and utility of the extracted pattern. Hence analyzing the interestingness of a pattern plays a vital role in KDD. Han and Kamber (2006) stated that all the patterns mined are not interesting or whatever the pattern mined by data mining tools are not interesting. To analyze the interestingness of a pattern various interestingness measures are proposed and analyzed by the researchers. In statistical aspect, there are many association measures available to measure the dependency between the variables. Segal *et al*. (2013) applied the association measures on their work. But all the association measures are not going to produce interestingness rules due to the over whelming of data and by existence of the redundant rules on mined patterns.

An association rule is an implication of the form $A \rightarrow B$ where $A \subset I$, $B \subset I$, $A \cap B = \emptyset$ and I is the item set. In this study, we represent given Data set, in terms of association rule, that is, the association rule $A \rightarrow B$ represented as a 2×2 contingency table as shown in the **Table 1** by the number of transactions supporting or not supporting the item sets A and B.

We will use the following notation thorughtout the study such as number of transactions supporting A and B, by the alphabet ‘a’, number of transactions supporting A but not B, by the alphabet ‘b’, number of transactions not supporting A but supporting B by the alphabet ‘c’ and number of transactions not supporting both A and B by the alphabet‘d’. Therefore the **Table 1** will be modified as shown in **Table 2**. Let N be the total number of transactions on the given data set, sum of a, b, c and d always equals to N.

Coefficient of Variation (CV) of a distribution with mean μ and variance σ^2 is defined as σ/μ. The coefficient of variation is usually used as a measure of precision for

Corresponding Author: Selvarangam, K., Department of Computer Science and Engineering, Hindustan University, Chennai, India

the dispersion of data set and is also often used to compare numerical distributions measured on different scales.

Statistically, Population CV is an ideal device for comparing the variation in two series of data which is measured in two different units (For example a comparison of variation in height with variation in weight). And the same population CV may be used to compare the dispersion of series measured in different units and also that series with same units, but running at different levels of magnitude. Similarly, the population CVs has been used to evaluate results from different experiments involving the same units of measure, possibly conducted by different persons.

Statistically, it is the fact that lower the CV leads, less deviation among the variables and higher the CV leads there will be more deviation among the variables. The CV predicts wrong deviation, when the variables having negative values or the mean of the variables become zero. And we know that if we measure temperature by Celsius and Fahrenheit units, the variation between Celsius and Fahrenheit units remains the same. Martinez Pons (2013) stated that, coefficient of variation used to compare two standard deviation when their mean differs substantially and its value become larger, when variance become greater than the mean and in this case size of CV is impossible. Hence lower the CV of a measure produce more interesting rules. This fact is the back bone of our algorithm. This study is organized as follows; section 2 describes the previous approaches on ranking association measures. In section 3 we listed the difficulties and draw backs of the existing ranking methods. Interestingness measures, its related properties and basic definitions are presented in section 4. Algorithm for selecting right interestingness measure is presented and implemented in section 5. Results of algorithm are discussed in section 6. Finally, future enhancement and conclusion are given in section 7.

Table 1. 2×2 contingency Table

A→B	B	$\bar{B}$	
A	n (AB)	n($A\bar{B}$)	n (A)
$\bar{A}$	n($\bar{A}$B)	n($\bar{A}\bar{B}$)	n($\bar{A}$)
	n (B)	n($\bar{B}$)	N

Table 2. 2×2 representation of association rule A→B

A→B	B	$\bar{B}$	
A	a	b	n (A)
$\bar{A}$	c	d	n($\bar{A}$)
	n (B)	n($\bar{B}$)	N

2. RELATED WORK

In this section, we review the selection of right measure to produce the Interesting patterns. Various methods and technique are implemented till now regarding the selection of good measure. Anandhavalli *et al.* (2010) ranked association rules mined by fast association rule mining Algorithm. They listed the mined association rules by the support and confidence and preceded the top most confident rules for ranking. The relative interestingness between the rules is calculated by applying entropy and variation.

Goktas and Ici (2011) compared most commonly used statistical association measures like Kendall's tau b, tau c Somers's d, Pearson coefficient and Spearman correlation with respect to large dimensional doubly ordered tables. All these measures are showing the less association, when compared with the actual degree of association present. Azevedo and Jorge (2007) proved on various data sets, the measure conviction clustered close to the top performing best rules by voting method. But it yields uninteresting results for the best rules in case of metric as relative measure. Tan *et al*. (2004) listed that in data mining literature, there exists more than forty association measures and they are producing different ranking. Geng and Hamilton (2006) confirmed the same on their survey. Azevedo and Jorge (2007) stated that the combination of different association measures may yield more interesting rules. Lallich *et al*. (2007) showed that the careful choice of interest measure and retaining significant rules lead more knowledge to the user. Nizal *et al*. (2010) confirmed the same, but they stated that, significant rules should be verified statistically using chi square test. Uma and Muneeswaran (2013) proved that through ranking the most relevant items will be retrieved effectively on a database.

3. PROBLEM STATEMENT

In the data mining literature, support, confidence and lift (interest) are the basic measures. Most of the existing measures are equivalent to these measures or the derived one of these. Support s of a rule A→B is the percentage of transactions containing A∪B in D. The rule A→B has confidence c if c% of transactions in D that contains A also contains B and the lift of the rule A→B is the ratio between the support of the rule A→B and product of probabilities of A and B in D.

Generally new measures are equivalent to the existing measures or statistically defined one. In this study, we proposed some measures based on both equivalent and statistical defined measures. The basic

measures in the pattern evaluation are support, confidence and lift. But each one of these has some draw backs. Jalali-Heravi and Zaïane, (2010) stated that in case of choosing large minimum support leads only to the rules, that contain obvious knowledge and missing the expectation case that are interesting. Whereas, choosing a low minimum support produces so many rules which could be redundant and noisy. Confidence is also not a perfect measure since it produces confident association between the statistically independent items. Similarly, lift also leads to wrong perdition in correlation that is in case of negative correlation it shows positive correlation, because lift is not depending on the null records.

The association rule mining algorithms has the advantage of allowing an unsupervised extraction of rules and of illustrating implicative tendency in data: It has the advantage of producing prohibitive number of rules. In the rule evaluation, we are facing main difficulty: That is how to extract the most interesting rule from the large amount of discovered rules. And the proposal of many interestingness measures in the literature leads to another difficulty that is, how to choose the interestingness measures that are adapted to its goal and its data, to detect the most interesting rules.

To reduce the above said difficulties, we proposed a ranking Algorithm based on CV of the measures calculated for the top most extracted association rules. Our algorithm ranks the given set of measures by eliminating the measures which are not suitable for the set of association rules.

4. INTERESTINGNESS MEASURES

Hiep (2010) stated that patterns are transformed into value by the interestingness measures. Jeyachidra and Punithavalli (2014) developed their feature selection algorithm DWFS-CK by using the interesting measure Gini Index. The interestingness of a measure depends on both data structure and on the decision maker's goal. Mcgarry (2005) classified these measures as objective and subjective in nature. Coverage, support, accuracy are criterias of objective and unexpectedness, actionable, novel are criteria under subjective nature. Geng and Hamilton (2006) added semantic as additional nature. Also they extended criteria with conciseness, reliability, peculiarity, diversity and utility. Defining the Interestingness of a measure is complex, but we may define the interestingness of measure by the above stated criteria. Some measures may be relevant with some context but not with others. Hence the ranking may be different on different data sets.

4.1. Properties of Interestingness Measures

Mustafa and Khan (2005) defined quality metric (measure) should possess minimality, formality, usability, accuracy, validity and reliability. Geng and Hamilton (2006) proposed that, a good measure M should have the following properties:

- P1: M = 0 if A and B are statistically Independent
- P2: M monotonically increases with P(A, B) when P(A) and P (B) remains the same
- P3: M monotonically decreases with P (A, B) when P(A) and P (B) remains the same
- P4: M is symmetric under variable permutation
- P5: M should have row and column scaling invariance
- P6: M is invariant under Inversion
- P7: M should null invariance
- P8: M becomes -M if either rows and columns are permuted
- P9: M increases as the total number of records increases
- P10: The threshold is easy to fix

Tan *et al*. (2004) listed 21 measures, later Geng and Hamilton (2006) extended the list with 38 measures. They proved that no measures satisfy all the properties listed above and no two measures produce same ranking.

4.2. Probability Based Objective Measures

Pecina and Schlesinger (2006) listed that there are nearly 82 association measures statistically; most of the measures are derived measures of joint probability and conditional probabilities. That is, in data mining literature, support and confidence are basic measures expressed in terms of probability. Most of the existing association measures are derived or equivalent to the basic measures. We discuss some basic derived measures and their ranking on different data sets.

4.2.1. Pointwise Mutual Information (PMI)

PMI will express numerically the association between item sets A and B and it is defined as the logarithmic value of the basic measure lift. Higher the PMI value indicates nearing perfect association, if there is no association between A and B, then P(AUB) = P(A)P(B) then lift is equal to one. That is PMI becomes zero. This is a symmetric measure Equation 1 and 2:

$$\text{Lift} = \frac{P(A \cup B)}{P(A)P(B)} = \frac{\frac{a}{N}}{\frac{(a+b)}{N}\frac{(a+c)}{N}} = \frac{Na}{(a+b)(a+c)} \quad (1)$$

$$\text{Point wise Mutual Information} = \log(\text{lift}) \quad (2)$$

4.2.2. Normalized Expectation (NE)

We define the Normalized Expectation (NE) as the existence of the rule A→B by knowing the presence of remaining items. The underlying concept is based on conditional probability defined in the Equation 3:

$$P\left(\frac{A}{B}\right) = \frac{P(A \cup B)}{P(B)} \quad (3)$$

where, P(A∪B) the joint probability, is mass function between A, B and P(B) is the marginal probability mass function B. We are interested in finding the set of all conditional probabilities measuring expectation of measuring A occurring, knowing that occurrence of B in N transactions. One way to find the above probabilities, we defined that is, one average event defining the conditional part of the probability (i.e., P(B)). The Fair Point Expectation (FPE) realizes this normalization. The FPE is theoretically defined as the average point of expectation embedding every particular point of expectation, thus reducing n particular point of expectation into just one average point. Basically, the fair point expectation is the arithmetic mean of all joint probabilities. We have only two events A, B so P(B) in Equation 3 which is replaced by the arithmetic mean of marginal probabilities of A and B. Now the normalized expectation is expressed by Equation 4 and 5:

$$\text{NormalizedExecptation} = \frac{2P(A \cup B)}{P(A) + P(B)} \quad (4)$$

$$= \frac{\frac{2a}{N}}{\frac{(a+b)}{N} + \frac{(a+c)}{N}} = \frac{2a}{2a+b+c} \quad (5)$$

4.2.3. Mutual Expectation (ME)

The product of normalized expectation and the support is called Mutual Expectation (ME) and it is calculated by the Equation 6:

$$\text{MutualExpectation} = \text{NEXSupport} = \frac{2a^2}{N(2a+b+c)} \quad (6)$$

4.2.4. Expected Frequency (EF)

Expected Frequency (EF) is the ratio of the product of number records of A and B to the total number of records N Equation 7:

$$\text{ExpectatedFrequency} = \frac{(a+b)(a+c)}{N} \quad (7)$$

4.2.5. Interestingness Factor (IF)

Interestingness Factor (IF) will express numerically, the deviation of the support from the statistical independence. It will be calculated by the Equation 8. Higher the IF value, there is more association and in case of there is no association it will lead the IF value to zero:

$$\text{Interestingness Factor} = P(A \cup B) - P(A)P(B) = \frac{a}{N} - \left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right) = \frac{Na - (a+b)(a+c)}{N^2} \quad (8)$$

4.2.6. Support Error (SE)

We define, numerical deviation of support from expected frequency as Support Error (SE) and it is calculated by Equation 9:

$$\text{Support error} = \frac{\text{Expectedfrequency} - a}{N} \quad (9)$$

4.2.7. U Cost, S Cost and R Cost

Pecina and Schlesinger (2006) listed U Cost, S Cost and R Cost are heuristic association measures, which is used to find the association between bigrams (between two variables). These measures are defined by the following Equation 10 to 12:

$$\text{U Cost} = \log\left(1 + \frac{\min(b,c) + a}{\max(b,c) + a}\right) \quad (10)$$

$$\text{S Cost} = \log\left(1 + \frac{\min(b,c)}{a+1}\right)^{-1/2} \quad (11)$$

$$\text{R Cost} = \log\left(1 + \frac{a}{a+b}\right) x \log\left(1 + \frac{a}{a+c}\right) \quad (12)$$

4.2.8. T Combined Cost

This measure also a heuristic association measures used in many researches, listed by Pecina and Schlesinger (2006), is defined by Equation 13:

$$\text{T Combined Cost} = \sqrt{UxSxR} \quad (13)$$

4.3. Assumptions and Definitions

We should assume the following throughout the paper, Association rules are mined from the transaction data base and the numbers of transactions on different data bases are nearly equal. And we should have the following definitions on a set of association rules.

Two or more number of measures are said to be consistent, if their correlation between the ranks is greater than or equal to some positive threshold.

A Measure M is called a not applicable measure, if its mean is zero.

Coefficient of variation (CV) of a distribution with mean μ and variance σ^2 is defined as σ/μ. That is:

$$\text{Coefficient of variation} = \frac{\sigma}{\mu} \quad (14)$$

A set of measures are said to be Equivalent, if their coefficient of variation remains same.

A measure M1 is earlier than M2, if CV (M1)<CV (M2).

5. COEFFICIENT OF VARIATION BASED RANKING

Liu *et al*. (2000) ranked the association rules by the existing domain knowledge of the user. The patterns (rules) may have different rank because their rank strongly depends on the choice of the Interestingness measure. Geng and Hamilton (2006) stated on their survey that the selection of interestingness measures can be done either by ranking or by clustering. Both ranking and clustering can be done by either based on data set or based on measures. Lallich *et al* (2007) stated that interesting measures should have less variation. In this study we measure the mined pattern using objective measures listed in **Table 3** and then by calculating the coefficient of variation we are grouping the measures suitable to the mined pattern as applicable measures, rest of them as Not Applicable measures (NA). Sharma *et al*. (2011) used logistic regression to find the variation between the measures.

5.1. CV based Ranking Algorithm

The top most set of association rules, A→B mined from a data set of the form 2×2 Contingency tables C_1, C_2, C_3, …, C_i, C_{i+1},…C_m and set of measures are given as input. Association rules mined are converted as numerical equivalent by given set of measures and listed as k column vectors. The collection of k column vectors represented as measurement matrix M and the order of matrix is given by mxk (number of association rules by number of measures). Each column in the measurement matrix is numerical equivalent of top most association rules with respect to the data set. Mean value $\bar{X}_k$ for each column k is calculated. The measures of columns whose mean value is zero are listed as set of not applicable measures. Rest of the measures are considered as applicable measures. For columns having applicable measures, standard deviation (σ) will be calculated. Applying mean $\bar{X}_k$ and standard deviation (σ) value in the Equation. 14 will yield, Coefficient of Variation (CV) value to the respective measures. These measures are arranged by the increasing order of CV. Thus we obtained the ranking of measures from most suitable to least. That is the measure having less CV leads to perfect measure.

Algorithm: CV Based Ranking Algorithm

Input: Association rules of the form 2×2 contingency table and set of measures M_1, M_2, M_3…, M_k, $M_{k+1,...}M_n$
Output:

- Ascending order of Applicable measures.
- Set of Measures not applicable.

Algorithm:

1. Get set of 2×2 contingency tables C_1, C_2, C_3, …, C_i, $C_{i+1,...}$Cm
2. Get set of measures M = { M_1, M_2, M_3…, M_k, $M_{k+1,...}M_n$ }
3. For i = 1 to m and for k = 1 to n Compute $M_k(C_i)$
4. Represent $M_k(C_i)$ as a Matrix M = {M_{ik}}, where i = 1 to m and k = 1 to n.
5. Find Mean of each column k, A(k),
6. List K values for which A(K) = 0
7. Remove the columns having A(k) = 0
8. List the Measures having A(k) = 0
9. Calculate Coefficient of variation CV_k for each column k, for k = 1 to n
10. Sort ascending order of CV_k
11. End

Output:

- Ascending order of Applicable measures.
- Set of not applicable measures.

5.2. Algorithm Implementation

We implemented the algorithm on five different sets of randomly generated (Generated by using IBM Quest

data set generator) association rules of 10,000 transactions each, with set of measures listed in **Table 3**. We generated the set of association rules (Twenty rules each) as follows: D_1 having random support of a, b, c and d. D_2, such that a+d is greater than b+c. That means with less deviation.D_3 with half of rules as in D_1 and another half of rules as in D_3. D_4 without null records (i.e., d = 0). D_5 satisfying as a+d increases b+c decreases. That is, with more variation. Ranking produced by the algorithm is presented in **Table 4**.

6. DISCUSSION

Tan *et al*. (2004) listed that, before deciding right measure to a particular domain the user must analyze several key factors, in this continuation, our algorithm decides perfect measures for a data set based on CV value. Geng and Hamilton (2006) suggested a promising method to find the interestingness using automatic selection or combining appropriate measures. Khan and Sheel (2013) also stated the importance of auto selection on their computing system for analysis of DNA sequences using OPTSDNA algorithm. Our algorithm ensures the automatic selection of measures. Hiep (2010) stated number of interestingness measures which may be reduced by considering a common measure on two or more measures. Also the interestingness of a measure can be calculated by the participating measures on a measure.

Table 3. Probability based objective measures

Measure	Formula
Pointwise Mutual Information	log(lift)
Normalized Expectation	$\frac{2a}{2a+b+c}$
Mutual Expectation	Normalized Expectation x Support
Expected Frequency	$\frac{(a+b)(a+c)}{N}$
Interestingness Factor	$\frac{Na-(a+b)(a+c)}{N^2}$
Support Error	$\frac{Expectedfrequency-a}{N}$
U Cost	$\log\left(1+\frac{\min(b,c)+a}{\max(b,c)+a}\right)$
S Cost	$\log\left(1+\frac{\min(b,c)}{a+1}\right)^{-1/2}$
R Cost	$\log\left(1+\frac{a}{a+b}\right)x\log\left(1+\frac{a}{a+c}\right)$
T Combined Cost	$\sqrt{UxSxR}$

Table 4. Ranking of objective measures

	D_1	D_2	D_3	D_4	D_5
1. PMI	NA	9	NA	NA	NA
2. NE	4	2	4	3	2
3. ME	7	7	7	6	6
4. EF	5	8	5	2	4
5. IS Factor	NA	6	NA	NA	NA
6. SE	NA	NA	NA	7	NA
7. U Cost	2	1	2	NA	NA
8. S Cost	1	5	1	1	1
9. R Cost	6	3	6	5	5
10. TCC	3	4	3	4	3

(NA-Not Applicable measures)

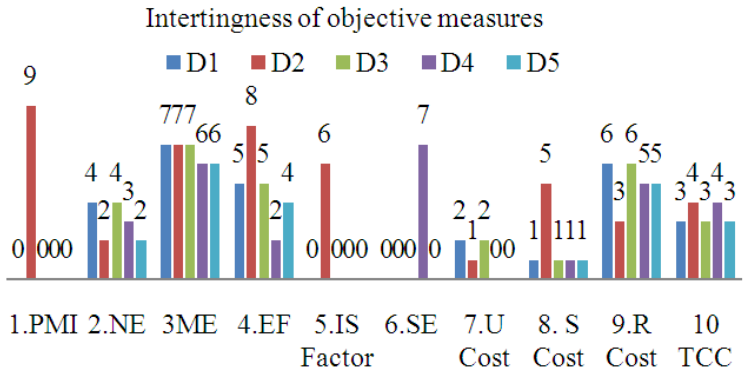

Fig. 1. Interestingness of objective measures

Equivalently our ranking on set of measures by variation on their numerical equivalent will suggest suitable measures in descending order. Topmost measures on ranking will be the perfect one. Since ranking done by eliminating not applicable measures, user's time and complexity on selecting measures will be reduced. Ranking represented in **Table 4** conclude that the measure S cost is having least ranking except on D_2 and it has less variation among the rules Refer to **Fig. 1**. Hence we may conclude that S Cost is the perfect measure to our set of rules. Our work is consistent with Geng and Hamilton's promising method stated above. And this may direct researchers to find common objective measure.

7. CONCLUSION

This study proposes a new approach for helping the user to identify perfect interesting measures. In particular, produces a list of measures by decending order of suitability with respect to the variability on a data base. Also it eliminates a set of measures not applicable, which reduces the list of measures considerably. We have taken a randomly generated set of association rules having equal number of records. This can be extended as a future work on set of association rules mined from the data base having unequal number of records and for more number of measures. In case of more number of measures, it is possible to get an equivalent set of measures, which will help the data mining community to reduce the number of measures in the literature. Finally, higher the variation on rules may make poor performance to our algorithm and in this case, coefficient of variant may be used.

8. REFERENCES

Anandhavalli, M., M.K.Ghose and K. Gauthaman, 2010. Interestingness measure for mining spatial gene expression data using association rule. J. Comput., 2: 110-114.

Azevedo, P.J. and A.M. Jorge, 2007. Comparing rule measures for predictive association rules. Proceedings of the 18th European Conference on Machine Learning, Sept. 17-21, Springer-Verlag, Warsaw, Poland, pp: 510-517. DOI: 10.1007/978-3-540-74958-5_47

Geng, L. and H.J. Hamilton, 2006. Interes-tingness Measures for Data Mining: A Survey. ACM Comput. Surveys. DOI: 10.1145/1132960.1132963

Goktas, A. and Ö. Isci, 2011. A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. Adv. Methodol. Stat., 8: 17-37.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Elsevier Inc., ISBN: 10: 1558609016, pp: 261-272.

Hiep, H.X., 2010. Interestingness measures for association rules in a KDD process: Post processing of rules with ARQAT tool. PhD Thesis, Nantes University, Nantes, France.

Jalali-Heravi, M. and O.R. Zaïane, 2010. A study on interestingness measures for associative classifier. Proceeding of the ACM Symposium on Applied Computing, (SAC '10), ACM, Switzerland, pp: 1039-1046. DOI: 10.1145/1774088.1774306

Jeyachidra and Punithavalli, 2014. Distinguishability based weighted feature selection using column wise k neighborhood for the classification of gene microarray dataset. Am. J. Applied Sci., 11: 1-7 DOI: 10.3844/ajassp.2014.1.7

Khan, M.I. and C. Sheel, 2013. An efficient distributed bioinformatics computing system for dna sequence analysis on encoding system. Am. J. Bioinformat., 2: 15-23. DOI: 10.3844/ajbsp.2013.15.23

Lallich, S., O. Teytaud and E. Prudhomme, 2007. Association Rule Interestingness: Measure and Statistical Validation. In: Quality measures in Data Mining, Guillet, F. and H.J. Hamilton (Eds.)., Springer, Berlin, ISBN-10: 3540449116, pp: 251-275.

Liu, B., W. Hsu, S. Chen and Y. Ma, 2000. Analyzing the subjective interestingness of association rules. IEEE, Intell. Syst. Applic., 15: 47-55. DOI:10.1016/j.ces.2004.07.070

Martinez-Pons, M., 2013. Coefficient of variation. J. Math. Stat, 9: 62-64. DOI: 10.3844/jmssp.2013.62.64

Mcgarry, K., 2005. Survey of interestingness measures for knowledge discovery. Knowl. Eng. Rev., 20: 39-61. DOI: 10.1017/S0269888905000408

Mustafa, K. and R.A. Khan, 2005. Quality metric development framework. J. Comput. Sci., 1: 437-444. DOI: 10.3844/jcssp.2005.437.444

Nizal, I., M. Shaharanee, F. Hadzic and T.S. Dillon, 2010. Interestingness measures for association rules based on statistical validity. Knowledge Based Syst., 24: 386-392.

Pecina, P. and P. Schlesinger, 2006. Combining association measures for collocation extraction. Proceedings of the COLING/ACL on Main Conference Poster Sessions, (CPS' 06), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 651-658.

Segal, G., A. Brom and E. Ramati, 2013. The "new settlers": Results of a bacteriological survey during the first 6-months operation period of an internal medicine ward in a tertiary hospital. J. Infect. Dis., 9: 136-141. DOI: 10.3844/ajidsp.2013.136.141

Sharma, D., D. McGee and B.M.G. Kibria, 2011. Measures of explained variation and the base-rate problem for logistic regression. Am. J. Biostatist., 2: 11-19. DOI: 10.3844/amjbsp.2011.11.19

Tan, P.N., V. Kumar and J. Srivastava, 2004. Selecting the right objective Measure for association analysis. Inform. Syst., 29: 293-313. DOI: 10.1016/S0306-4379(03)00072-3

Uma, R. and K. Muneeswaran, 2013. Efficacious geospatial information retrieval using density probabilistic document correlation approach. J. Comput. Sci., 9: 83-93. DOI: 10.3844/jcssp.2013.83.93