Intern. J. Fuzzy Mathematical Archive
Vol. 6, No. 2, 2015,133-138
ISSN: 2320 –3242 (P), 2320 –3250 (online)
Published on 22 January 2015
www.researchmathsci.org

International Journal of
Fuzzy Mathematical
Archive

Interesting Set of Association Rules

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

[1]Department of Computer Science and Engineering, Hindustan University, Chennai India. e-mail: kselvaviji@gmail.com
Corresponding Author

[2]Department of Information Technology, Hindustan University, Chennai, India
e-mail: krkumar@hindustanuniv.edu

Received 8 November 2014; accepted 30 November 2014

Abstract. The rate of growth of data in an information system is exponential every year, hence data base of an information system become huge, and the number of association rules extracted from these data bases is too many. A challenging question in front of knowledge finder is the extraction of the actual knowledge present from these set of association rules. Prior to the knowledge extraction from a set of association rules, determination of their interest plays a vital role. Since quality rules can only lead to extract implicit knowledge. Interesting association rules may mined directly by data mining tools on applying interestingness measures while mining, or their interestingness determined by quality measures after mining. In this work, we developed a method to determine an interesting set of association rules from a set of mined rules by determining homogeneity coefficient (HC). The range of HC varies from 0 to 1. HC value of a measure on a rule close to 1, leads interesting set of association rule and the knowledge extracted from this set of rules consistent with actual knowledge present.

Keywords: Data mining, Association rule, Interestingness Measures, variability coefficient, homogeneity coefficient.

AMS Mathematics Subject Classification (2010): 68P99

1. Introduction

An approximate measure of the right thing is better than the exact measure of the wrong thing. Hence we may assume approximate measure on interesting rule will lead to better knowledge in the process knowledge discovery in data (KDD). Cluster analysis is a class of techniques used to classify objects or cases into relatively homogenous groups called clusters [2, 7, 8]. Objects in each cluster tend to be similar to each other and dissimilar to objects in other clusters. This is an approach to 'let the rules speak for themselves' by means of transactions. Application of clustering techniques might improve the understandability of mined rules by bringing together 'similar' rules into the same cluster. It may be easier to infer item behavior from rule clusters than from a rule list. This is because consecutive rules in a rule list may not have any relationship to each other. This can confound the user thus making the interpretation difficult. Clustering differs from grouping, in that there is no preconceived notion of the structure or the number of groups that may exist in the data [2]. The idea here is to look for a 'natural'

structure in the data on the basis of which clusters are evolved. Researchers have used clustering and grouping as strategies to improve the understandability of rules. 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 [5]. We may assume that rule means always an association rule in this study. Statistically, variability is defined as the deviation from base point. Variability may calculate by range, mean, variance, deviations and coefficient of variation (CV). In our previous work [12] we ranked ARs by the value of CV. 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. While the Coefficient of Variation [3], defined as s/M, is often used to compare two standard deviations when their means differ substantially, it, too, is inadequate for present purposes: because s is not always smaller than the mean, it is possible for CV to be greater than 1-lack of a natural ceiling which, as in the case of s^2 and s, makes a definitive interpretation of the size of CV impossible. Because of this drawback of CV, we proposed a clustering technique using variability coefficient (VC) and homogeneity coefficient [11]. We assume that quality remains the same as interest in this study.

2. Related works

Goktas and Isci [4] reviewed some common measures used to measure the association between two rules; the degree of association will determine the interestingness of ARs. Most of the measures used to determine the quality of association rules are build with mean and variance. Lent et al. [10] have introduced the notion of a 'clustered' AR. A clustered AR is a rule that is formed by combining similar, 'adjacent' association rules to form a few general rules. Wang et al. [14] maximizes certain interestingness criteria during the merging process. Toivonen et al. [13] proposed another approach; Distance between two rules is defined as the number of transactions in which the two rules with the same consequents differ. Gupta et al. [5] have proposed a normalized distance function called conditional market-basket probability (CMPB) distance. This distance function tends to group all those rules that 'cover' the same set of transactions. Gupta et al. [5] state "rules involving different items but serving equal purposes were found to be close good neighbors" [5]. Thus, their approach is able to capture some amount of customer purchasing behavior. One of the limitations of both the schemes is the arbitrariness of the distance measures used for rule clustering [1]. Moreover, they do not develop any framework to concisely describe the generated rule clusters.

3. Problem statement

When it comes to quality of an association rule, how the quality of a rule is measured, to determine if it is useful, interesting, important etc. But there is no formal definition of quality and/or interestingness [7]. Currently there is a collection of different measures available which is partly due to the traditional methods of support and confidence being considered insufficient [10]. Most of the quality measures defined in terms of mean and variance are not able find the actual degree of association due to the impossibility of CV [12], when variance is greater than mean. The degree of homogeneity is a group exhibits

on some measure and the difference in homogeneity is the group exhibits across two or more measures. These issues assume particular relevance when the interest lies in deciding whether to subdivide the set of ARs on the basis of the information at hand.

3.1. Materials and methods

Interestingness of set association rules will be calculated in this work by variability coefficient (VC) or by coefficient of homogeneity (HC). By fixing the threshold on VC or HC, we will cluster the association rule and make decision on the necessity of further division.

3.1.1. Variability coefficient

Variability consists of the differences in magnitude that exist in a set of occurrences of some measure. If at least one occurrence differs in magnitude from the others, the set of rules exhibits variability; if no difference occurs, then the set of rule does not exhibit variability. When only one occurrence differs in size from the others, the set exhibits minimum variability; and the greater the total difference in magnitude among the occurrences, the greater the variability exhibited by the set of rules. If variability is seen in this light, then its measure can be formulated as the sum of the observed differences among occurrences of a measure divided by the maximum possible sum of the differences. This is known as variability coefficient and express by the equation 1

Variability Coefficient (VC) = OV/MPV (1)

where: OV = Observed variation, MPV = Maximum possible variation

The value of VC always lies between 0 and 1. Since there is no variation in set of rule scores the OV become zero hence it is clear that VC become zero(by equation 1) In case of maximum variation among the rules scores, OV is equal to MPV hence VC become 1 in this case.

The observed variability (OV) is the sum of the absolute differences among occurrences of the measure at hand. A matrix arrangement of the differences among a group of scores is helpful in visualizing the calculations used to derive OV. Statistically it is the fact that, the maximum sum of differences in a set of scores will occur if half the scores have the lowest value contained in the set and the other half carry the highest value. For a comparison matrix of a data set half of which consists of one uniform value and half of which consists of a different uniform value, only comparisons of the two different values will yield nonzero remainders.

The derivation of MPV in (1) is based on the following reasoning: the maximum sum of differences in a set of scores will occur if half the scores have the lowest value contained in the set and the other half carry the highest value. For an even number of cases, the number of such comparisons is the number of scores in the group's lower half multiplied by the number of scores in the group's upper half, that is $\left(\frac{N}{2}\right)\left(\frac{N}{2}\right)$ and thus, the number of non-zero comparisons will equal the square of half the cases in the data set that is, $\left(\frac{N}{2}\right)^2$. The highest possible variability will consist of the product of this square and the sum of the comparisons of the two values. Thus, for a group of scores consisting of an even number of cases, MPV can be calculated as follows equation 2:

$$MPV = \left(\frac{N}{2}\right)^2 R \quad (2)$$

Where, N = group size and R = the range, that is, the difference between the highest and lowest scores. For a group of scores consisting of an odd number of cases, MPV can be calculated by equation 3: $MPV = \left(\frac{N-1}{2}\right)\left(\frac{N+1}{2}\right)R$ (3)

3.1.2. Homogeneity coefficient

A coefficient of homogeneity (HC) can be defined as the complement of VC; hence it is calculated by equation 4:$HC = 1 - VC$ (4)
Since the VC value lies between 0 and 1, the HC value also lies between 0 and 1.

3.1.3. Calculation of HC and VC

Let us consider a relational data base R, and a set of association rules R_1, R_2, R_3,…, R_n on R with rule score $x_1, x_2, x_3, \ldots, x_n$respectively. OV in equation 1 is the sum of absolute differences among the occurrence of the rules which is calculated by the equation 5. A matrix arrangement of the differences among a group of scores is helpful in visualizing the calculations used to derive OV. For the set of rules, the matrix is displayed in Table 1. The scores in Table 1appear vertically along the table's left as well as horizontally along its top. For each row, the cells represent the difference between the score on the left column and the other scores in the set. Each score on the horizontal list is subtracted from each of the scores on the vertical list and the remainder for each subtraction is recorded as an absolute value in the intersecting cell. If no difference emerges, a 0 is recorded. $OV = \sum|x_i - x_j| \; foralli, j$ (5)
The derivation of MPV in (1) is based on the following reasoning: the maximum sum of differences in a set of scores will occur if half the scores have the lowest value contained in the set and the other half carry the highest value. Let the least and highest score in table 1 be named$y_i and y_j$ respectively. The MPV calculation is represented in table 2.

3.2. Implementation

Let us consider a relational data base R, and a set of association rules R_1, R_2, R_3, …, R_{10} on R with scores 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, respectively. A matrix arrangement of the differences among rule scores is helpful in visualizing the calculations used to derive OV. For the above set of rules, the matrix is displayed in Table 3. The OV value is calculated by equation 5 and its value for the above set of rules is 820. The derivation of MPV for the above set of rules is displayed in table 4. The highest variation will occur if the data take the following values: 35, 35, 35, 35, 35, 75, 75, 75, 75, 75 and the MPV = 1000 (by equation. 2) by applying OV and MPV value in equation 1,$VC = \frac{820}{1000} = 0.82$. And the HC value is given by equation 4. $HC = 1 - VC = 1 - 0.82 = 0.28$

Scores	x_1	x_2	x_3	…	x_i.	…	x_n
x_1	0	$\|x_1 - x_2\|$	$\|x_1 - x_3\|$		$\|x_1 - x_i\|$		$\|x_1 - x_n\|$
x_2		0	$\|x_2 - x_3\|$		$\|x_2 - x_i\|$		$\|x_2 - x_n\|$
x_3			0		$\|x_3 - x_i\|$		$\|x_3 - x_n\|$
.				.			.
x_i					0		$\|x_i - x_n\|$
.						.	.
x_n							0

Table 1: Matrix arrangement of differences in rule scores

Scores	y_i	y_i	y_i	…	y_j.	…	y_j
y_i	0	0	0		$\lvert y_i - y_j \rvert$		$\lvert y_i - y_j \rvert$
y_i		0	0		$\lvert y_i - y_j \rvert$		$\lvert y_i - y_j \rvert$
y_i			0		$\lvert y_i - y_j \rvert$		$\lvert y_i - y_j \rvert$
.				.			.
y_j					0		$\lvert y_i - y_j \rvert$
.						.	.
y_j							0

Table 2: Matrix arrangement for MPV calculation

According to the user knowledge expectation the set of rules generated from the relational data base R using data mining tools. For this set of rules VC and HC value calculated as above, based on the values of VC and HC we may conclude the interesting set of rules

	30	35	40	45	50	55	60	65	70	75
30	0	5	10	15	20	25	30	35	40	45
35		0	5	10	15	20	25	30	35	40
40			0	5	10	15	20	25	30	35
45				0	5	10	15	20	25	30
50					0	5	10	15	20	25
55						0	5	10	15	20
60							0	5	10	15
65								0	5	10
70									0	5
75										0

Table 3: Matrix arrangement of differences in a group scores

	30	30	30	30	30	75	75	75	75	75
30	0	0	0	0	0	45	45	45	45	45
30		0	0	0	0	45	45	45	45	45
30			0	0	0	45	45	45	45	45
30				0	0	45	45	45	45	45
30					0	45	45	45	45	45
75						0	0	0	0	0
75							0	0	0	0
75								0	0	0
75									0	0
75										0

Table 4: Matrix arrangement for MPV calculation

4. Conclusion and future work

In this paper, we have presented a method by determining the variability coefficient, the value of VC is not depending on mean and variance, and hence the drawback on coefficient of variation will be eliminated. VC close to 1 means the set of rules exhibit more variations, and the rules produces more knowledge and they do not consistent with actual knowledge due to the over whelming. This supports the Geng and Hamilton [9] conclusion presented on their survey. That is more the variation means that less homogeneity. Hence less homogeneity set of rules may divide further to make homogeneous set of rules. This work directs, when interest lies further subdividing of data in hand possibilities. Implementing on big data sets by the way of algorithm may enhance this work.

REFERENCES

1. G.Adomavicius and A.Tuzhilin, Expert-driven validation of rule-based user models inpersonalization applications, *Data Mining and Knowledge Discovery,* 5 (2001) 33-58.
2. M.R.Anderberg, Cluster analysis for applications, New York: Academic Press, 1973.
3. F.E.Croxton, D.J.Crowden and S. Klein, Applied General Statistics, 3rd Edn., Prentice-Hall, New York, 1967.
4. A. Goktas and Ö. Isci., A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation, *MetodološkiZvezki*, 8 (2011) 17-37.
5. G. K. Gupta, A. Strehl and J. Ghosh, Distance based clustering of association rules, Proc. Intelligent Engineering Systems through Artificial Neural Networks (ANNIE 1999) 9 (1999) 759–764.
6. J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd Ed. Elsevier Inc., 261-272, 2006.
7. A.K.Jain, M.N.Murty and P.J. Flynn, Data clustering: A review, ACM *Comput. Survey*, 31(1999) 264–323.
8. L.Kaufman, and P.J Rousseeuw, Finding groups in data: An introduction to cluster analysis, New York, Wiley, 1990.
9. G.Liquing and H.J.Hamilton, Interestingness measures for data mining: A survey, *ACM Comput. Surveys*, 38 (2006).
10. B.Lent, A.N. Swami and J.Widom, Clustering association rules, In *ICDE*, 220-231, 1997.
11. M.Martinez-Pons, Coefficient of variation, *J. Mathem. Statistics*, 9 (2013) 62-64. 2013.
12. K.Selvarangam, and K.Ramesh Kumar, Selecting perfect interestingness measures by coefficient of variation based ranking algorithm, *J. Computer science.*, 10 (2014) 1672 – 1679.
13. H.Toivonen, M. Klemettinen, P. Ronkainen, K.Hatonen and H.Mannila, Pruning and grouping discovered association rules, Proc. Mlnet Workshop on Statistics, Machine Learning and Knowledge Discovery in Databases, Herakhion, Crete, Greece, 1999.
14. K.Wang, H.W.Tay Soon, and B.Liu, Interestingness-based interval merger for numeric association rules. Proc. 4th Int. Conf. on Data Mining and Knowledge Discovery (KDD 98) New York: AAAI Press, 1998, 121-128.