

Interestingness Measures for Multi-Level Association Rules

R Vijaya Prakash[1*], Dr. A. Govardhan[2], Prof. SSVN. Sarma[3]
1. Dept. of Informatics, Kakatiya University, Warangal, India
2. Dept. Of Computer Science & Engineering, JNT University, Hyderabad, India
3. Dept. Of Computer Science & Engineering, Vaagdevi College of Engineering
*E-mail of the corresponding author: vijprak@hotmail.com

Abstract
Association rule mining is one technique that is widely used to obtain useful associations rules among sets of items. Much work has been done focusing on efficiency, effectiveness and redundancy. There has also been a focusing on the quality of rules from single level datasets with many interestingness measures proposed. However, there is a lack of interestingness measures developed for multi-level and cross-level Association rules. Single level measures do not take into account the hierarchy found in a multi-level dataset. This leaves the Support-Confidence approach, which does not consider for the hierarchy. In this paper we propose two approaches which measure multi-level association rules to help and evaluate their interestingness. These measures of diversity and peculiarity can be used to identify those rules from multi-level datasets that are potentially useful.
Keywords: Information Retrieval, Interestingness Measures, Association Rules, Multi-Level Datasets, Itemsets

1. Introduction

Interestingness measures are necessary to rank Association rule patterns. Each interestingness measure produces different results, and experts, Lenca *et.al* (2008) have different opinions of what constitutes a good rule. The interestingness of discovered association rules is an important and active area within data mining research (L. Geng & H.J. Hamilton 2006). The primary problem is the selection of interestingness measures for a given application domain. However, there is no formal agreement on a definition for what makes rules interesting. Association rule algorithms produce thousands of rules, many of which are redundant (K. McGarry 2005, Li. J. *et al* 2003). In order to filter the rules, the user generally supplies a minimum threshold for support and confidence. Support and confidence(R. Agrawal *et al* 1993, G. Dong & J. Li 1998, S. Lallich *et al* 2006) are basic measures of association rule *interestingness*. All of these measures were proposed for association rules derived from single level or flat datasets, which were most commonly transactional datasets. Today multi-level datasets are more common in many domains. With this increase in usage there is a big demand for techniques to discover multi-level and cross-level association rules and also techniques to measure interestingness of rules derived from multi-level datasets. J. Han & Y. Fu (1995, 1999), C.N. Zeigler *et al* (2005) proposed some approaches for multi-level and cross-level frequent itemset discovery have been proposed. However, multi-level datasets are often a source of numerous rules and in fact the rules can be so numerous it can be much more difficult to determine which ones are interesting (R. Agrawal *et al* 1993, G. Dong & J. Li 1998). Moreover, the existing interestingness measures for single level association rules can not accurately measure the interestingness of multi-level rules since they do not take into consideration the concept of the hierarchical structure that exists in multi-level datasets.

In this paper, we propose measures particularly for assessing the interestingness of multilevel association rules by examining the diversity and peculiarity among rules. These measures can be determined at rule discovery phase during post-processing to help users determine the interesting rules. Diversity of a data set is defined as when comparing two data sets, the one with more diverse rules is more interesting. Diversity will be used to compare two data sets to determine which data set contains rules that are more interesting.

The paper is organized as follows. Section 2 discusses related work. The theory, background and assumptions behind our proposed interestingness measures are presented in Section 3. Experiments and results are presented in Section 4. Lastly, Section 5 concludes the paper.

2. Related Work

For as long as association rule mining has been around, there has been a need to determine which rules are interesting. Originally this started with using the concepts of support and confidence (R. Agrawal *et al* 1993). Since then, many more measures have been proposed (G. Dong & J. Li 1998, L. Geng & H.J. Hamilton 2006, S. Lallich *et al* 2006). The Support-Confidence approach is appealing due to the anti monotonicity property of the support. However, the support component will ignore itemsets with a low support even though these itemsets may generate rules with a high confidence (S. Lallich *et al* 2006). Also, the Support-Confidence approach does

not necessarily ensure that the rules are truly interesting, especially when the confidence is equal to the marginal frequency of the consequent (S. Lallich *et al* 2006). Based on this argument, other measures for determine the interestingness of a rule are needed.

All of these existing measures fall into three categories; objective based measures (based on the raw data), subjective based (based on the raw data and the user) and semantic based measures (based on the semantic and explanations of the patterns) (L. Geng & H.J. Hamilton 2006). In the survey presented in (L. Geng & H.J. Hamilton 2006) there are nine criteria listed that can be used to determine if a pattern or rule is interesting. These nine criteria are; conciseness, coverage, reliability, peculiarity, diversity, novelty, surprisingness, utility and actionability or applicability. The first five criteria are considered to be objective, with the next two, novelty and surprisingness being considered to be subjective. The final two criteria are considered to be semantic.

Despite all the different measures, studies and works undertaken, there is no widely agreed upon formal definition of what interestingness is in the context of patterns and association rules (L. Geng & H.J. Hamilton 2006). More recently several surveys of interestingness measures have been presented (L. Geng & H.J. Hamilton 2006, S. Lallich *et al* 2006, P. Lenca *et al* 2007 & K. McGray 2008). In K. McGray (2008) survey evaluated the strengths and weaknesses of various measures from the point of view of the level or extent of user interaction. P. Lenca *et al* (2007) survey looked at classifying various interestingness measures into five formal and five experimental classes, along with eight evaluation properties. However, all of these surveys result in different outcomes over how useful, suitable etc., an interestingness measure is. Therefore the usefulness of a measure can be considered to be subjective.

All of these measures mentioned above are for rules derived from single level datasets. They work on items on a single level but do not have the capacity for comparing different levels or rules containing items from multiple levels simultaneously.

Hilderman R.J. and Hamilton H.J. (2001) established three primary principles that a good interestingness measure should satisfy :

- The *minimum value principle*, which states that a uniform distribution is the most uninteresting.
- The *maximum value principle*, which states the most uneven distribution is the most interesting.
- The *skewness principle*, which states that the interestingness measure for the most uneven distribution will decrease when then number of classes of tuples increases.
- The *permutation invariance principle*, which states that interestingness for diversity is unrelated to the order of the class and it is only determined by the distribution of counts.
- The *transfer principle*, which states that interestingness increases when a positive transfer is made from the count of one tuple to another whose count is greater.

Here in our work we propose to measure the interestingness of multi-level rules in terms of diversity and peculiarity (also known as distance). These measures were chosen as they are considered to be objective (rely on just the data).

3. Interestingness Measures for Multi-Level Association Rules

3.1. Diversity-Based Interestingness Measures

According to L. Geng & H.J. Hamilton 2006, a "pattern is diverse if its elements differ significantly from each other, while a set of patterns is diverse if the patterns in the set differ significantly from each other" (L. Geng & H.J. Hamilton 2006). Summaries can be measured using diversity-based interestingness measures. While there has been research on using diversity to measure summaries, there is little, if any, research that focuses on measuring the interestingness of association or classification rules (L. Geng & H.J. Hamilton 2006). Therefore, this study suggests that diversity can be used to measure association rule interestingness.

Diversity generally determined by two factors: 1) the proportional distribution of classes in the population, and 2) the number of classes. Consider two different sets of rules mined from a dataset. We can consider the rules that are more diverse to be more interesting. Additionally, we can consider a set of rules less interesting if the rules are less diverse. Another way of thinking about this is by considering that too many similar rules will convey less knowledge to a user.

The equations for variance and the Shannon, C.E (1948) measure are shown in the equation (1) and (2) These are only two measures based on diversity. There are fourteen other measures based on diversity that are available. Most measures are designed to be used with a specific application domain. We chose variance and Shannon because of their wide-spread use.

$$variance = \frac{\sum_{i=1}^{m}(p_i-\bar{q})^2}{m-1} \quad (1)$$

$$Claude\ Shannon = -\sum_{i=1}^{m} p_i log_2 p_1 \quad (2)$$

In the above equation for variance, p_i is the probability for class *i*, and is the average probability for all classes.

3.2. Peculiarity Based Interestingness Measures

Peculiarity is an objective measure that determines how far away one association rule is from others. The further away the rule is, the more peculiar. It is usually done through the use of a distance measure to determine how far apart rules are from each other. Peculiar rules are usually few in number (often generated from outlying data) and significantly different from the rest of the rule set. It is also possible that these peculiar rules can be interesting as they may be unknown. G. Dong & J. Li (1998) proposed peculiarity measure, which is neighborhood-based unexpected measure. In this proposal it is argued that a rule's interestingness is influenced by the rules that surround it in its neighborhood.

The measure is based on the idea of determining and measuring the symmetric difference between two rules, which forms the basis of the distance between them. From this G. Dong & J. Li (1998) proposed that unexpected confidence (where the confidence of a rule R is far from the average confidence of the rules in R's neighborhood) and sparsity (where the number of mined rules in a neighborhood is far less than that of all the potential rules for that neighborhood) could be determined, measured and used as interestingness measures (G. Dong & J. Li 1998, L. Geng & H.J. Hamilton 2006).

G. Dong & J. Li (1998) measure determines the symmetric difference was developed for single level datasets where each item was equally weighted. Thus the measure is actually a count of the number of items that are not common between the two rules. In a multi-level dataset, each item cannot be regarded as being equal due to the hierarchy. Thus the G. Dong & J. Li (1998) measure needs to be enhanced to be useful with these datasets. Here we will present an enhancement as part of our proposed work.

We believe it is possible to take the distance measure presented in (G. Dong & J. Li 1998) and enhance it for multi-level datasets. The original measure is a syntax-based distance metric in the following form:

$$P(R_1, R_2) = \delta_1 * |(X_1 \cup Y_1)\Delta(X_2 \cup Y_2)| + \delta_2 * |(X_1 \Delta X_2)| + \delta_3 |(Y_1 \Delta Y_2)| \quad (3)$$

The Δ operator denotes the symmetric difference between two item sets, thus X Δ Y is equivalent to (X –Y)∪(Y –X) δ_1, δ_2 and δ_3 are the weighting factors to be applied to different parts of the rule. Equation 3 measures the peculiarity of two rules by a weighted sum of the cardinalities of the symmetric difference between the two rule's antecedents, consequents and the rules themselves.

We propose an enhancement to this measure to allow it to handle a hierarchy. Under the existing measure, every item is unique and therefore none share any kind of 'syntax' similarity. However, we argue that the items 1-*-*-*, 1-1-*-*, 1-1-1-* and 1-1-1-1 (based on Figure 1) all have a relationship with each other. Thus they are not completely different and should have a 'syntax' similarity due to their relation through the dataset's hierarchy.

The greater the $P(R_1,R_2)$ value is, lower similarity and so the greater the distance between those two rules. Therefore, the further apart the relation is between two items, the greater the difference and distance. Thus if we have,

R_1 : 1 – 1 – 1 – * → 1 – * – * – *

R_2 : 1 – 1 – * – * → 1 – * – * – *

R_3 : 1 – 1 – 1 – 1 → 1 – * – * – *

We believe that the following should hold; P(R1,R3) < P(R2,R3) as 1-1-*-* and 1-1-1-1 are further removed from each other than 1-1-1-* and 1-1-1-1. The difference between any two hierarchically related items\ nodes must be less than 1. Thus, for the above rules, 1 > P(R2,R3) > P(R1,R2) > 0. In order to achieve this we modify Equation 3 by calculating the diversity of the symmetric difference between two rules instead of the cardinality of the symmetric difference. The cardinality of the symmetric difference measures the difference between two rules in terms of the number of different items in the rules. The diversity of the symmetric difference takes into consideration the hierarchical difference of the items in the symmetric difference to measure the difference of the two rules. We recite Equation 2 in terms of a set of items below, where S is a set containing n items:

$$PD(S) = \frac{\alpha \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} HRD(i,j)}{n(n-1)} + \frac{\beta \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} LD(i,j)}{n(n-1)} \quad (4)$$

Where HRD is The Hierarchical Relationship Distance between two items is defined as the ratio between the average number of levels between the two items and their common ancestor and the height of the tree. In general HRD is defined as width or horizontal distance, which is defined as

$$HRD(n_1, n_2) = \frac{(NLD(n_1,ca)+NLD(n_2,ca)}{2*TreeHeight} \quad (5)$$

LD is the Level Distance which measures the distance between two items in terms of their height. This is also called as height distance or vertical distance, which is defined as

$$LD(n_1, n_2) = \frac{NLD(n_1,n_2)}{(TreeHeight-1)} \quad (6)$$

NLD is Number of levels NLD(x,y) = |hierarchy level of x| - |hierarchy level y|.

N1 and n2 are two items, ca is common ancestor. The α, β are the user threshold values, where α+β=1. In our experiment we set these values as α=β=0.5.

Thus the neighbourhood-based distance measure between two rules shown in Equation 3 now becomes;

$$PM(R_1, R_2) = \delta_1 * PD((X_1 \cup Y_1)\Delta(X_2 \cup Y_2) + \delta_2 * PD(X_1\Delta X_2) + \delta_3 * PD(Y_1\Delta Y_2) \quad (7)$$

4. Experiments

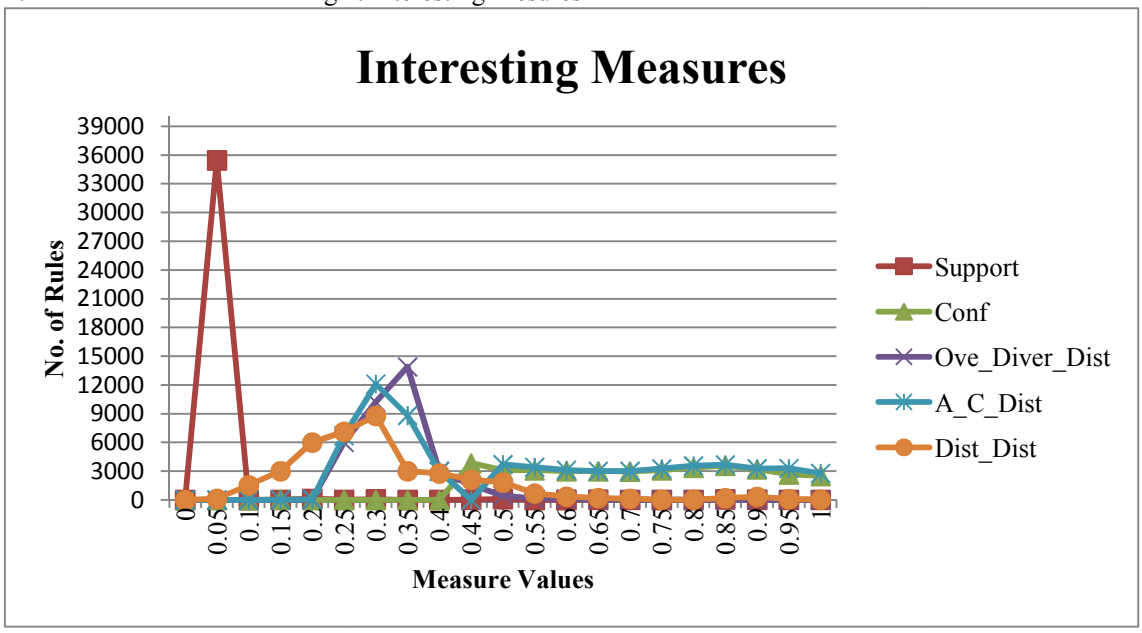
The dataset used for our experiments is a real world dataset, the BookCrossing dataset (obtained from http://www.informatik.uni-freiburg.de/ cziegler/BX/). From this dataset we built a multi-level transactional dataset that contains 91,550 user records and 970 leaf items, with 3 concept / hierarchy levels. To discover the frequent itemsets we use the MLT2 L1 algorithm proposed by J. Han & Y. Fu (1995, 1999) with each concept level having its own minimum support. From these frequent itemsets we then derive the frequent closed itemsets and generators using the CLOSE+ algorithm proposed by J. Pei *et al*. From this we then derive the non-redundant association rules using the MinMaxApprox (MMA) rule mining algorithm.

The confidence curve shows that the rules are spread out from 0.5 (which is the minimum confidence threshold) up to close to 1. The distribution of rules in this area is fairly consistant and even, ranging from as low as 2,181 rules for 0.95 to 1, to as high as 4,430 rules for 0.85 to 0.9. Using confidence to determine the interesting rules is more practical than support, but still leaves over 2,000 rules in the top bin.

The overall diversity curve shows that the majority of rules (23,665) here have an average overall diversity value of between 0.3 to 0.4. The curve however, also shows that there are some rules which have an overall diveristy value below the majority, in the range of 0.15 to 0.25 and some that are above the majority, in the range of 0.45 up to 0.7. The rules located above the majority are different to the rules that make up the majority and could be of interest as these rules have a high overall diversity.

The antecedent-consequent diversity curve is similar to that of the overall diversity. It has a similar spread of rules, but the antecedent-consequent diversity curve peaks earlier at 0.3 to 0.35 (where as the overall diversity curve peaks at 0.35 to 0.4), with 12,408 rules. The curve then drops down to a low number of rules at 0.45 to 0.5, before peaking again at 0.5 to 0.55, wih 2,564 rules. The shape of this curve with that of the overall diversity seems to show that the two diversity approaches are related. Using the antecedent-consequent diversity allows rules with differing antecedents and consequents to be discovered when support and confidence will not identify them.

Fig 1. Interesting Mesures



5. Conclusion

In this paper we proposed two interestingness measures for Multi Level Association rules. These proposed interestingness measures are diversity and peculiarity respectively. Diversity is a measure that compares items within a rule and peculiarity compares items in two rules to see how different they are. In our experiments we have shown how diversity and peculiarity distance can be used to identify potentially interesting rules which can't be identified using basic measurements support and confidence.

References

R. Agrawal, T. Imielinski and A. Swami. (1993), "Mining Association Rules between Sets of Items in Large Databases". In ACM SIGMOD International Conference on Management of Data (SIGMOD'93), pages 207–216, Washington D.C., USA.

G. Dong and J. Li. (1998), "Interestingness of Discovered Association Rules in terms of Neighbourhood-Based Unexpectedness". In Second Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'98), pages 72–86, Melbourne, Australia.

L. Geng and H. J. Hamilton. (2006), "Interestingness Measures for Data Mining: A Survey". ACM Computing Surveys (CSUR), Volume 38, pages 9.

J. Han and Y. Fu (1995). "Discovery of Multiple-Level Association Rules from Large Databases". In 21st International Conference on Very Large Databases (VLDB'95), pages 420–431, Zurich, Switzerland.

J. Han and Y. Fu (1999). "Mining Multiple-Level Association Rules in Large Databases". IEEE Transactions on Knowledge and Data Engineering, Volume 11, pages 798–805.

S. Lallich, O. Teytaud and E. Prudhomme (2006), "Association rule interestingness: measure and statistical validation. Quality Measures in Data Mining", Volume 43, pages 251–276.

P. Lenca, B. Vaillant, B. Meyer and S. Lallich. (2007) "Association rule interestingness: experimental and theoretical studies". Studies in Computational Intelligence, Volume 43, pages 51–76.

K. McGarry. (2005), "A Survey of Interestingness Measures for Knowledge Discovery". The Knowledge Engineering Review, Volume 20, pages 39–61.

C.-N. Ziegler, S. M. McNee, J. A. Konstan and G. Lausen (2005), "Improving Recommendation Lists Through Topic Diversification". In 14th International Conference on World Wide Web (WWW'05), pages 22–32, Chiba, Japan.

Lenca, P., Meyer, P., Vaillant, B., & Lallich, S. (2008). "On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid". *European Journal of Operations Research, 184*, 610-626.

Li, J., & Zhang, Y. (2003). "Direct Interesting Rule Generation". Proceedings of the Third IEEE International Conference on Data Mining (ICDM '03), 155-167.

Hilderman, R. J., & Hamilton, H. J. (2001). "Knowledge Discovery and Measures of Interest". Boston, MA: Kluwer Academic.

Shannon, C. E. (1948). "A mathematical theory of communication". The Bell System Technical Journal, 27, 379-423.