Interestingness Measures for Multi-Level Association Rules

Gavin Shaw
School of Information Technology
Queensland University of Technology
Brisbane QLD Australia
g4.shaw@student.qut.edu.au

Yue Xu
School of Information Technology
Queensland University of Technology
Brisbane QLD Australia
yue.xu@qut.edu.au

Shlomo Geva
School of Information Technology
Queensland University of Technology
Brisbane QLD Australia
s.geva@qut.edu.au

Abstract *Association rule mining is one technique that is widely used when querying databases, especially those that are transactional, in order to obtain useful associations or correlations 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, with multi-level datasets now being common there is a lack of interestingness measures developed for multi-level and cross-level 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 the hierarchy anyway and has other drawbacks, as one of the few measures available.*

In this paper we propose two approaches which measure multi-level association rules to help evaluate their interestingness. These measures of diversity and peculiarity can be used to help identify those rules from multi-level datasets that are potentially useful.

Keywords Information Retrieval, Interestingness Measures, Association Rules, Multi-Level Datasets

1 Introduction

Association rule mining was first introduced in [1] and since then has become both an important and widespread tool in use. It allows associations between a set of items in large datasets to be discovered and often a huge amount of associations are found. Thus in order for a user to be able to handle the discovered rules it is necessary to be able to screen / measure the rules so that only those that are interesting are presented to the user. This is the role interestingness measures play. In an effort to help discover the interesting rules, work has focused on measuring rules in various ways from both objective and subjective points of view [3] [8]. The most common measure is the support-confidence approach [1] [2] [6], but there are numerous other measures [2] [3] [6] to name a few. 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. Some approaches for multi-level and cross-level frequent itemset discovery (the first step in rule mining) have been proposed [4] [5] [10]. 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 [1] [2]. 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 as our contribution we propose measures particularly for asessing the interestingness of multi-level association rules by examining the diversity and distance among rules. These measures can be determined during rule discovery phase for use during post-processing to help users determine the interesting rules. To the authors' best knowledge, this paper is the first attempt to investigate the interestingness measures focused on multi-level datasets.

The paper is organised 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.

Proceedings of the 14th Australasian Document Computing Symposium, Sydney, Australia, 4 December 2009.

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 [1]. Since then, many more measures have been proposed [2] [3] [6]. The Support-Confidence approach is appealing due to the antimonotonicity 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 (which is used to indicate the level of interestingness) [6]. 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 [6]. Based on this argument, other measures for determing the interestingness of a rule is needed.

Broadly speaking, 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) [3].

In the survey presented in [3] 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 [3]. More recently several surveys of interestingness measures have been presented [3] [6] [7] [8]. One survey [8] evaluated the strengths and weaknesses of various measures from the point of view of the level or extent of user interaction. Another survey [7] 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. Our research has found that up to now, little work has been done when it comes to interestingness measures for multi-level datasets that can handle items from muliple levels within one rule or rule set.

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 Concepts and Calculations of The Proposed Interestingness Measures

In this section we present the key parts of the theory and background and formula behind our proposed measures. We also present the assumptions we have made for our measures.

3.1 Assumptions and Definitions

Here we outline the assumptions we have made. Figure 1 depicts an example of the general structure of a multi-level dataset. As shown, there is a tree-like hierarchical structure to the concepts or items involved in the dataset. Thus items at the bottom are descendant from higher level items. An item at a higher level can contain multiple lower level items.

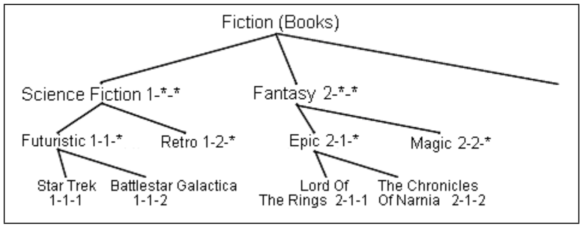

Figure 1: Example of a multi-level dataset.

With this hierarchy we have made the two following assumptions.

1. That each step in the hierarchy tree is of equal length / weight. Thus the step from 1-*-* to 1-1-* is of equal distance to the step from 2-*-* to 2-1-* or 1-1-* to 1-1-1.

2. That the order of sibling items is not important and the order could be changed (along with any descendants) without any effect.

3. That each concept/item has an ancestor concept/item (except for the root) so that no concepts/items or group(s) of concepts/items are isolated from the rest of the hierarchy.

Before presenting our proposed measures we firstly define several terms and formula used.

- n_1 and n_2: represent two items / concepts in the multi-level dataset.

- *ca*: (common ancestor) is the closest item that is an ancestor to both n_1 and n_2.

- *TreeHeight*: is the maximum number of items on a path in the multi-level dataset (not counting root) from the root to a item located at the lowest concept level.

- *h*: represents the entire multi-level dataset hierarchy.

- *Hierarchy level of an item*: the hierarchy level of the root is 1. The hierarchy level of an item in the dataset is larger than the level of its direct parent by 1.
- *Number of Levels Difference*:

  $$NLD(x,y) = |\, hierarchy\ level\ of\ x - hierarchy\ level\ of\ y\,| \quad (1)$$

  is the number of hierarchy levels difference between items *x* and *y*.

3.2 Diversity

Here we define a diversity measure for multi-level association rules which takes items' structural information into consideration. The diversity defined here is a measure of the difference or distance between the items within a rule, based on their positions in the hierarchy. Two different aspects of the items in a rule are considered to measure the diversity of the rule.

1. Hierarchical relationship distance (HRD) between items.
2. Concept level distance (LD) between items.

We propose that the diversity of a rule can be measured using two different approaches. The first, measures the overall diversity of a rule by combining the items in the antecedent with the items in the consequent into a single set. If the items within this combined itemset are very different, then the rule will have a high overall diversity, regardless of whether the items were from the antecedent or consequent.

Let *R* be a rule with *n* items and D_{OR} denotes the overall diversity of *R*, the diversity of *R* can be determined as follows:

$$D_{OR} = \frac{\alpha_1 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} HRD(i,j)}{n(n-1)} + \frac{\beta_1 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} LD(i,j)}{n(n-1)} \quad (2)$$

The second, measures the diversity between the items in the antecedent and those in the consequent. Those rules which have a high difference between their antecedent itemsets and consequent itemsets will have a high antecedent-consequent diversity. However, this approach does not consider the difference between items within the antecedent and/or consequent like the overall diversity approach.

Let *R* be a rule $R : A \rightarrow C$, with *n* items in *A* and *m* items in *C* and D_{ACR} denotes the antecedent to consequent diversity of *R*, the diversity of *R* can be determined as follows:

$$D_{ACR} = \frac{\alpha_2 \sum_{i=1}^{n-1} \sum_{j=1}^{m} HRD(i,j)}{n(n-1)} + \frac{\beta_2 \sum_{i=1}^{n-1} \sum_{j=1}^{m} LD(i,j)}{n(n-1)} \quad (3)$$

Where α and β are weighting factors such that $\alpha + \beta = 1$. The values of α and β need to be determined experimentally and for our experiments are both set at 0.5. Equation 2 & 3 consists of two parts, the average hierarchical relationship distance and the concept distance among the items in the rule, respectively. In the following subsections we will define the two aspects in detail.

3.2.1 Hierarchical Relationship Distance

The HRD of two items measures how close two items are in terms of a hierarchical relationship from a common ancestor item (or root). The further apart they are in a hierarchical relation; that is the greater the number of concept levels difference between two items and their common ancestor, the more diverse the two items are and the more diverse the rule will be.

Here for the HRD component, diversity is inversely related to the closeness of items in terms of a hierarchical relationship. The closer the two items are, the less diverse they are. The further / more distant the relationship, the more diverse. For maximum HRD diversity the two items need to have no common ancestor and both be located at the lowest concept level in the dataset.

HRD focuses on measuring the horizontal (or width) distance between two items. Usually the greater the horizontal distance, the greater the distance to a common ancestor and therefore the more diverse the two items are. Due to the second assumption, we can not measure the horizontal distance without also utilising the vertical (height) distance.

Thus to determine the Hierarchical Relationship Distance (HRD) component of the diversity the following is proposed:

$$HRD(n_1, n_2) = \frac{(NLD(n_1, ca) + NLD(n_2, ca))}{2 \times TreeHeight} \quad (4)$$

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. Thus if two items share a direct parent, the HRD value of the two items becomes the lowest value which is $1/TreeHeight$, while if the two items have no common ancestor or their common ancestor is the root, the HRD values of the two items can score high. Maximum HRD value, which is 1, is achieved when the two items have no common ancestor (or the common ancestor is the root) and both items are at the lowest concept level possible in the hierarchy. If n_1 and n_2 are the same item, then HRD becomes $1/TreeHeight$.

3.2.2 Concept Level Distance

This aspect is based on the hierarchical levels of the two items. The idea is that the more levels between the two items, the more diverse they will be. Thus two items on the same hierarchy level are not very diverse, but two items on different levels are more diverse as they have different degrees of specificity or abstractness.

LD differs from HRD in that HRD measures the distance from a common ancestor item (or root), whereas LD measures the distance between the two items themselves. LD focuses on measuring the distance between two items in terms of their height (vertical) difference (HRD considers the width (horizontal) distance).

Thus, we propose to use the ratio between the level difference (NLD) of two items and the height of the tree (eg. the maximum level difference) to measure the Level Distance of the two items as defined as follows:

$$LD(n_1, n_2) = \frac{NLD(n_1, n_2)}{(TreeHeight - 1)} \tag{5}$$

This means that two items on the same concept level will have a LD of 0, while an item at the highest concept level and another at the lowest concept lvel will have an LD of 1, as they are as far apart as possible in the given hierarchy.

3.3 Peculiarity

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. One proposal for measuring peculiarity is the neighbourhood-based unexpectednedd measure first proposed in [2]. In this proposal it is argued that a rule's interestingness is influenced by the rules that surround it in its neighbourhood.

The measure is based on the idea of determining and mesuring the symmetric difference between two rules, which forms the basis of the distance between them. From this it was proposed [2] that unexpected confidence (where the confidence of a rule *R* is far from the average confidence of the rules in *R's* neighbourhood) and sparsity (where the number of mined rules in a neighbourhood is far less than that of all the potential rules for that neighbourhood) could be determined, measured and used as interestingness measures [2] [3].

This measure [2] for determing the symmetric difference was developed for single level datasets where each item was equally weighted. Thus the mesure 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 measure proposed in [2] 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 [2] and enhance it for multi-level datasets. The original measure is a syntax-based distance metric in the following form:

$$\begin{aligned} P(R_1, R_2) = {} & \delta_1 \times |(X_1 \cup Y_1)\Theta(X_2 \cup Y_2)| + \\ & \delta_2 \times |X_1 \Theta X_2| + \delta_3 \times |Y_1 \Theta Y_2| \end{aligned} \tag{6}$$

The Θ operator denotes the symmetric difference between two item sets, thus $X\Theta Y$ is equivalent to $X - Y \cup Y - X$. δ_1, δ_2 and δ_3 are the weighting factors to be applied to different parts of the rule. Equation 6 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, the greater the difference (thus 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,
$R_1 : 1-1-1-* \Rightarrow 1-*-*-*$
$R_2 : 1-1-*-* \Rightarrow 1-*-*-*$
$R_3 : 1-1-1-1 \Rightarrow 1-*-*-*$
We believe that the following should hold; $P(R_1, R_3) < P(R_2, R_3)$ 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) $1 > P(R_2, R_3) > P(R_1, R_2) > 0$. In order to achieve this we modify Equation 6 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:

$$\begin{aligned} PD(S) \quad = \quad & \frac{\alpha \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} HRD(i,j)}{n(n-1)} + \\ & \frac{\beta \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} LD(i,j)}{n(n-1)} \end{aligned} \tag{7}$$

Thus the neighbourhood-based distance measure between two rules shown in Equation 6 now becomes;

$$\begin{aligned} PM(R_1, R_2) = {} & \delta_1 \times PD((X_1 \cup Y_1)\Theta(X_2 \cup Y_2)) + \\ & \delta_2 \times PD(X_1 \Theta X_2) + \delta_3 \times PD(Y_1 \Theta Y_2) \end{aligned} \tag{8}$$

Let *RS* be the ruleset of $\{R_1, R_2, ..., R_n\}$ then the average distance of a rule R_i to the ruleset *RS* can be determined by:

$$PM_{ave} = \frac{\sum^{n}_{\forall R_j \in RS \ and \ j \neq i} PM(R_i, R_j)}{|RS| - 1} \tag{9}$$

4 Experimental Results

In this section we present experimental results of our proposed interestingness measures being used for association rule discovery from a multi-level dataset.

4.1 Dataset and Setup

The dataset used for our experiments is a real world dataset, the BookCrossing dataset (obtained from http://www.informatik.uni-freiburg.de/ cziegler/BX/) [10]. From this dataset we built a multi-level transactional dataset that contains 92,005 user records and 960 leaf items, with 3 concept / hierarchy levels.

To discover the frequent itemsets we use the MLT2_L1 algorithm proposed in [4] [5] 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 in [9]. From this we then derive the non-redundant association rules using the MinMaxApprox (MMA) rule mining algorithm [9].

4.2 Results

For the experiment we simply use the previously mentioned rule mining algorithm to extract the rules from the multi-level dataset. For this experiment we assign a reducing minimum support threshold to each level. The minimum supports are set to 10% for the first hierarchy level, 7.5% for the second and 5% for the thrid level (the lowest). During the rule extraction process we determine the diversity and peculiarity distance of the rules that meet the confidence threshold. With two measures known for each rule, we are also able to determine the minimum, maximum and average diversity and peculiarity distance for the rule set.

4.2.1 Statistical Analysis

Firstly, we compare the distribution curves of the proposed measures (diversity and distance) against the distribution curves of support and confidence for the rule set. The distribution curves are shown in Figure 2. The value of each measure ranges from 0 to 1. The values of the distance measure are based on the minimum distance (in this case 33,903.7) being equal to 0 and the maximum distance (in this case being 53,862.5) being equal to 1. The range between these two has been uniformly divided into 20 bins.

As Figure 2 shows, the support curve shows that the majority of association rules only have a support of between 0.05 and 0.1. Thus for this dataset distinguishing interesting rules based on their support would be difficult as the vast majority have very similar support values. This would mean the more interesting or important rules would be lost.

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.

Lastly, the distance curve shows the largest spread of rules across a curve. There are rules which have a low distance from the rule set (0 to 0.1 which corresponds to a distance of 33,903.7 to 35,899.56) up to higher distances (such as 0.7 and above which corresponds to a distance of 47,874.88 to 53,862.52). The distance curve peaks at 0.3 to 0.35 (which is a distance of between 39,891.35 and 40,889.29). Using the distance curve to find interesting rules allows those that are close to the ruleset (small distance away) or those that are much further away (greater distance) to be discovered.

Next, we look at the trends of the various measures when compared against the proposed diversity and distance measures.

Figure 3 shows the trend of the average support, average confidence, average antecedent-consequent diversity and average distance values against that of overall diversity. As can be seen the average support remains fairly constant. There is tendancy for the support to increase for those rules with a high overall diversity. Even so, this shows that support does not always agree with overall diversity, so an overall diversity measure can be useful to find a different set of interesting rules.

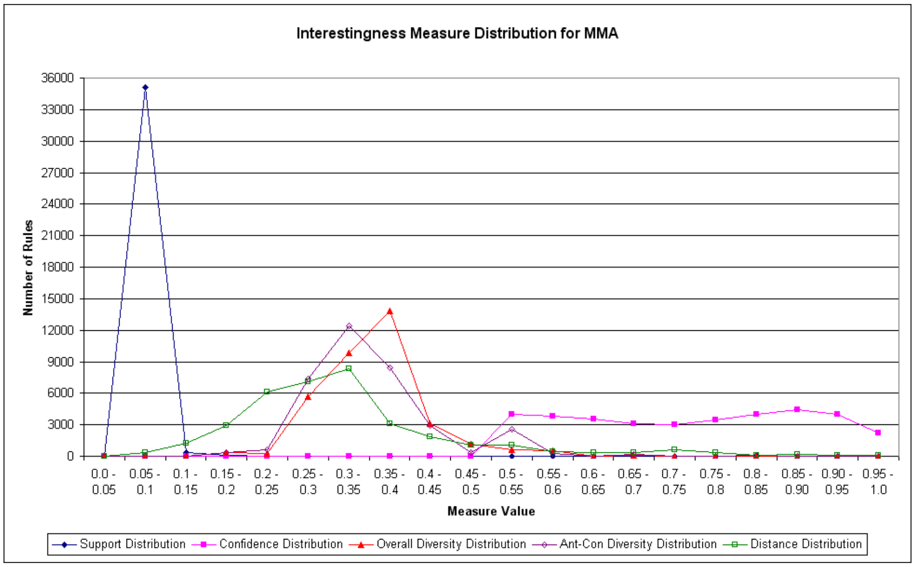

Figure 2: Distribution curves for the proposed interestingness measures, support and confidence.

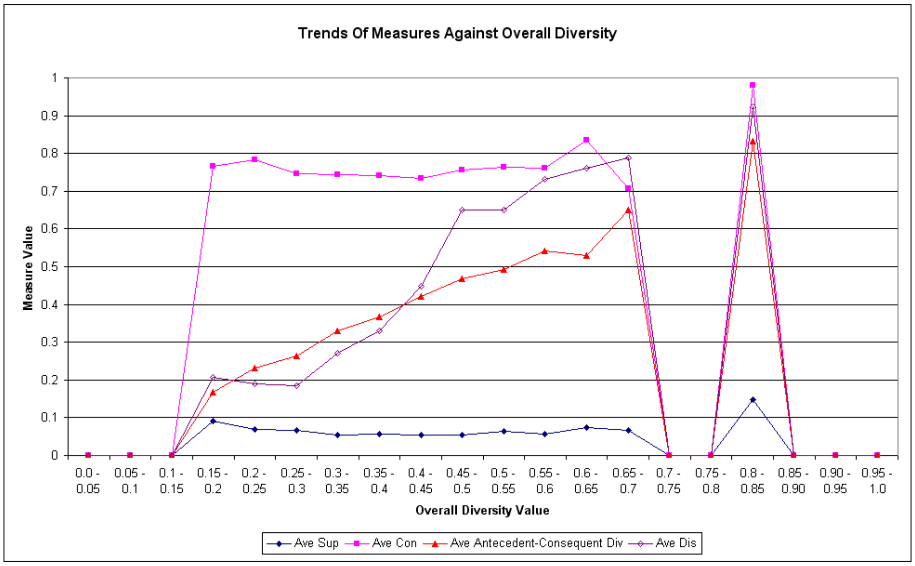

Figure 3: Trends of measures against the proposed overall diversity measure.

The confidence in Figure 3 is also fairly constant (usually varying by less than 0.1) until the end. Again this shows that the confidence will not always discover those rules that are more diverse overall.

The average antecedent-consequent diversity tends to have a consistant upward trend as the overall diversity increases. This shows that both the overall diversity and antecedent-consequent diversity are related/linked (which is not unexpected). It is quite possible that the greatest degree of diversity for a rule comes from comparing the items in the antecedent against those in the consequent and not from comparing the items within just the antecedent and/or consequent.

The distance has an overall upwards trend, although it is not a constant rate nor consistant (as there is a small decrease from 0.2 to 0.3). This, along with the trend of the average overall diversity (which shows a consistant upwards trend as the distance increases) in Figure 5 would indicate that potentially the more overall diverse rules have a higher distance from the rest of the rule set and therefore are further away. This would also imply that those rules with a higher distance are usually more diverse overall as well.

Figure 4 shows the trends of average support, average confidence, average overall diversity and average distance against that of antecedent-consequent diversity. Like in Figure 3, the support remains fairly constant regardless of the antecedent-consequent diversity value.

The confidence tends to decrease as the antecedent-consequent diversity increases, so the more diverse rules will not always be picked up by confidence.

The overall diversity tends to increase as antecedent-consequent diversity increases (similar

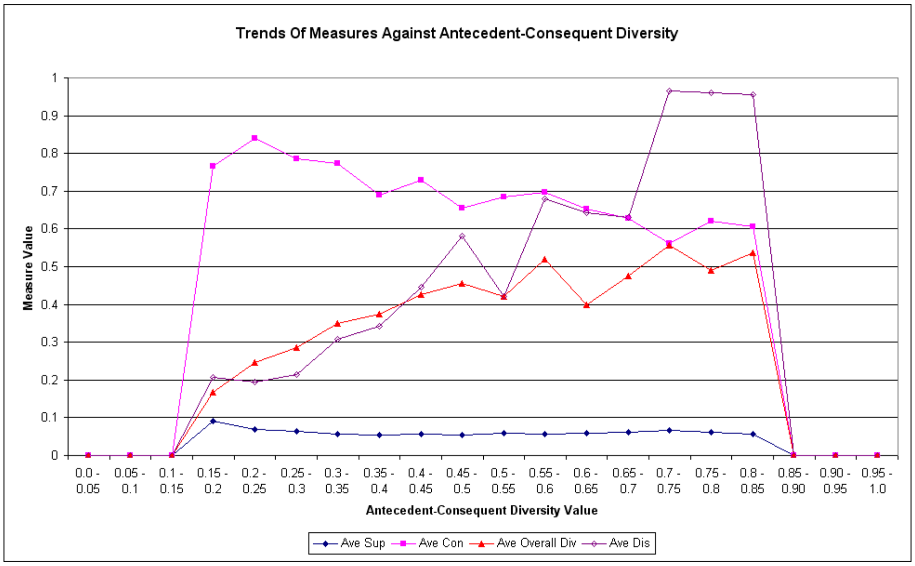

Figure 4: Trends of measures against the proposed antecedent-consequent diversity measure.

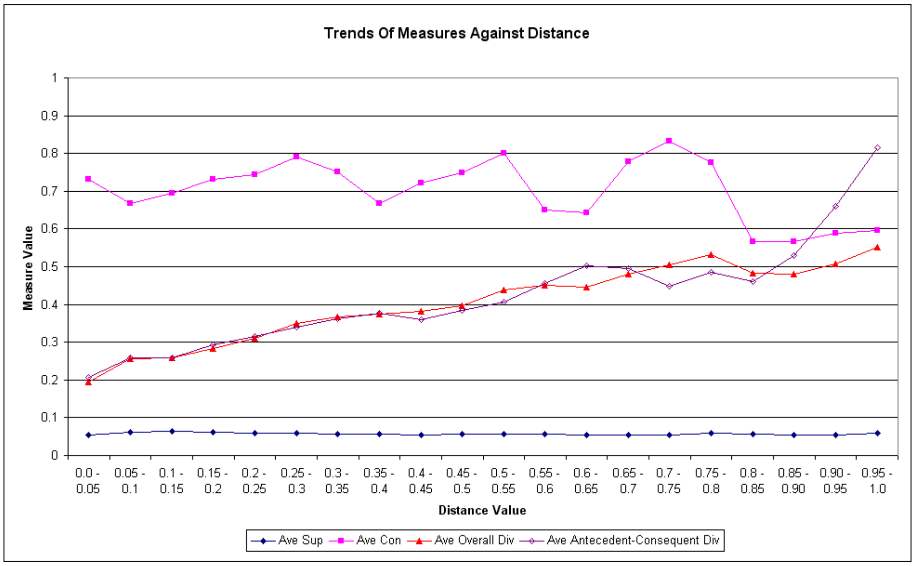

Figure 5: Trends of measures against the proposed distance measure.

to Figure 3). So again the biggest diversity in a rule is often in the difference between the antecedent and consequent.

The distance also tends to increase (gradually at first for lower antecedent-consequent diversity values). There is a big jump in the distance trend when the antecedent-consequent diversity increases from 0.65 to 0.75. The highest distance values are achieved when the antecedent-consequent diversity reaches its highest values (this is also shown in Figure 5).

Figure 5 shows the trend of the average support, average confidence, average overall diversity and average antecedent-consequent diversity values against that of distance. As shown, the support remains very constant regardless of the distance. This shows that support can not be used to discover rules that have a low or high peculiarity distance.

Like the average overall diversity, the average antecedent-consequent diversity also trends upwards as the distance increases. The rate of rise is similar to that of the overall diversity initially at lower distance values, but becomes much steeper at high distance values. This shows that it seems the most distant rules also have the highest diversity between their antecedent and consequent as Figure 5 shows the average antecedent-consequent diversity to be over 0.8 for the rules with the highest distance from the rest of the rule set.

The confidence trend in Figure 5 also shows that confidence will not always discover those rules far away from the rule set (0.8 / 49,870.76 and above), as at these distances the confidence values are at their lowest points. For rules with a low distance value, confidence may also not be the best measure as at these values

(0 / 33,903.7 to 0.1 / 35,899.58) confidence values are not at their highest. The highest confidence value(s) occur when the distance is 47,874.86 to 48,872.82 (0.7 to 0.75).

4.2.2 Examples of Proposed Measures

If we look closer at the discovered rules we find the following examples that show how diversity and peculiarity distance can be useful in identifying potentially interesting rules that would not normally be identified as such. (Note that the hypen breaks the concept levels, while a comma indicates a new item).

Example 1:

R_1=BookClubs-Lit.&Fiction-Pop.Fiction → Subjects-Lit.&Fiction-General

Supp 12.228% Conf 81.5% OverallDiv 0.5

R_2=BookClubs-Lit.&Fiction-Pop.Fiction → Subjects-Mystery&Thrillers

Supp 7.9% Conf 52.67% OverallDiv 0.67

R_1 has a higher support and confidence than R_2, but R_2 has a higher overall diversity. If we used either the support or confidence measure then R_1 would always be chosen as the more interesting rule. However, our proposed overall diversity measure indicates that R_2 is more interesting due to its diversity score, which can be attributed to its more general consequent.

Example 2:

R_3=Subjects-Biographies&Memoirs-General, Subjects-Lit.&Fiction-Authors(A..Z) → BookClubs-Lit.&Fiction

Supp 5.59% Conf 60.9% Ant-ConDiv 0.67

R_3 has low support and reasonbly low confidence, but it has high antecedent-consequent diversity (the average is 0.35). If we use support or confidence this rule will probably not be chosen as interesting as its support value is lower than the average support value for this rule set (5.8%) and its confidence is relatively low and is also lower than the average confidence of the rule set (74.4%). However, if we use antecedent-consequent diversity, then it will be selected as it has a high value. Hence this rule may be of interest because of the diversity between its antecedent and consequent itemsets, which come from different branches of the hierarchy.

Example 3:

R_4=BookClubs, Subjects-Lit.&Fiction-WorldLit. → Subjects-Lit.&Fiction-GenreFiction, Subjects-Mystery&Thrillers

Supp 6.7% Conf 57.7% Dist 50,311.4

R_4 has a noticably higher than average distance and is much further away from the rule set. This may be of interest to a user. But if support and confidence are used, this rule is considered to not be of interest due to their low values.

5 Conclusion

In this paper we have proposed two interestingness measures for association rules derived from multi-level datasets. These proposed interestingness measures are diversity and peculiarity (distance) 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 that normally would not be considered as interesting using the traditional support and confidence approach.

Acknowledgements Computational resources and services used in this work were provided by the HPC and Research Support Unit, Queensland University of Technology, Brisbane, Australia.

References

[1] R. Agrawal, T. Imielinski and A. Swami. 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, May 1993.

[2] G. Dong and J. Li. 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, April 1998.

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

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

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

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

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

[8] K. McGarry. A Survey of Interestingness Measures for Knowledge Discovery. *The Knowledge Engineering Review*, Volume 20, pages 39–61, 2005.

[9] N. Pasquier, R. Taouil, Y. Bastide and G. Stumme. Generating a Condensed Representation for Association Rules. *Journal of Intelligent Information Systems*, Volume 24, pages 29–60, 2005.

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