

# Valency Based Weighted Association Rule Mining

Yun Sing Koh, Russel Pears, and Wai Yeap

School of Computing and Mathematical Sciences,  
Auckland University of Technology, New Zealand  
{ykoh, rpears, wyeap}@aut.ac.nz

**Abstract.** Association rule mining is an important data mining task that discovers relationships among items in a transaction database. Most approaches to association rule mining assume that all items within a dataset have a uniform distribution with respect to support. Therefore, weighted association rule mining (WARM) was introduced to provide a notion of importance to individual items. Previous approaches to the weighted association rule mining problem require users to assign weights to items. This is infeasible when millions of items are present in a dataset. In this paper we propose a method that is based on a novel Valency model that automatically infers item weights based on interactions between items. Our experimentation shows that the weighting scheme results in rules that better capture the natural variation that occurs in a dataset when compared to a miner that does not employ such a weighting scheme.

**Keywords:** weighted association rule mining, valency, principal components.

## 1 Introduction

Association rule mining was introduced by [1] and is widely used to derive meaningful rules that are statistically related. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in transaction databases. The relationships are not based on the inherent properties of the data themselves but rather based on the co-occurrence of the items within the database. The original motivation for seeking association rules came from the need to analyze supermarket transaction data also known as market basket analysis. An example of a common association rule is  $\{bread\} \rightarrow \{butter\}$ . This indicates that a customer buying bread would also buy butter. With traditional rule mining techniques even a modest sized dataset can produce thousands of rules, and as datasets get larger, the number of rules produced becomes unmanageable. This highlights a key problem in association rule mining; keeping the number of generated itemsets and rules in check, whilst identifying interesting rules amongst the plethora generated.

In the classical model of association rule mining, all items are treated with equal importance. In reality, most datasets are skewed with imbalanced data. By applying the classical model to these datasets, important but critical rules which occur infrequently may be missed. For example consider the rule:  $\{stiff\ neck, fever, aversion\ to\ light\} \rightarrow \{meningitis\}$ . Meningitis occurs relatively infrequently in a medical dataset, however if it is not detected early the consequences can be fatal.

Recent research [2,3,4,5] has used item weighting to identify such rules that rarely manifest but are nonetheless very important. For example, items in a market basket dataset may be weighted based on the profit they generate. However, most datasets do not come with preassigned weights, the weights must be manually assigned in a time consuming and error-prone fashion. Research in the area of weighted association rule mining has concentrated exclusively on formulating efficient algorithms for exploiting pre-assigned weights rather than deducing item weights from a given transactional database. We believe that it is possible to deduce the relative importance of items based on their interactions with each other. In application domains where user's input on item weights is either unavailable or impractical, an automated approach to assigning weights to items can contribute significantly to distinguishing high value rules from those with low value.

In this paper we make two contributions to the field of association rule mining. Firstly, we present a novel scheme that automates the process of assigning weights to items. The weights assignment process is underpinned by a "Valency model" that we propose. The model considers two factors: purity and connectivity. The purity of an item is determined by the number of items that it is associated with over the entire transactional database, whereas connectivity represents the strength of the interactions between items. We will elaborate on the Valency model later in the paper in section 3. Secondly, association rules produced by the Valency model are evaluated through a novel scheme based on Principal Components Analysis. The formulation of this interest measure was motivated by the fact that none of the popularly used interest measures such as Confidence and Lift was able to capture differences between rules with highly weighted items from those with lowly weighted ones.

The rest of the paper is organized as follows. In the next section, we look at previous work in the area of weighted association rule mining. In section 3 we give a formal definition of the weighted association rule mining problem. Section 4 describes our proposed Valency model while Section 5 presents the evaluation scheme used to assess the performance of the Valency model. Our experimental results are presented in Section 6. Finally we summarize our research contributions in Section 7 and outline directions for future work.

## 2 Background

The classical association rule mining scheme has thrived since its inception in [1] with application across a very wide range of domains. However, traditional Apriori-like approaches were not designed to deal with the rare items problem [6,7]. Items which are rare but have high confidence levels are unlikely to reach the minimum support threshold and are therefore pruned out. For example, Cohen [8] noted that in market basket analysis rules such as  $\{caviar\} \rightarrow \{vodka\}$  will not be generated by traditional association rule mining algorithms. This is because both caviar and vodka are expensive items which are not purchased frequently, and will thus not meet the support threshold.

Numerous algorithms have been proposed to overcome this problem. Many of these algorithms follow the classical framework but substitute an item's support with a weighted form of support. Each item is assigned a weight to represent the importance of individual

items, with items that are considered interesting having a larger weight. This approach is called *weighted association rule mining* (WARM) [2,4,5,9,10]. Sanjay et al. [9] introduced weighted support to association rule mining by assigning weights to both items and transactions. In their approach rules whose weighted support is larger than a given threshold are kept for candidate generation, much like in traditional Apriori [1]. A similar approach was adopted by [2], but they applied weights to items and did not weigh transactions. They also proposed two different ways to calculate the weight of an itemset, either as the sum of all the constituent items' weights or as the average of the weights. However, both of these approaches invalidated the downward closure property [11].

This led Tao et al. [10] to propose a "weighted downward closure property". In their approach, two types of weights were assigned, item weight and itemset weight. The goal of using weighted support is to make use of the weight in the mining process and prioritize the selection of targeted itemsets according to their perceived significance in the dataset, rather than by their frequency alone.

Yan and Li [5] working in the domain area of Web mining proposed that weights be assigned on the basis of the time taken by a user to view a web page. Unlike the previous approaches [2,4,9,10] that assumed a fixed weight for each item, Yan and Li [5] allowed their weights to vary according to the dynamics of the system, as pages became more popular (or less popular) the weights would increase (or decrease), as the case may be.

Recently Jian and Ming [12] introduced a system for incorporating weights for mining association rules in communication networks. They made use of a method based on a subjective judgements matrix to set weights for individual items. Inputs to the matrix were supplied by domain specialists in the area of communications networks.

Thus it can be seen in previous work that the weight assignment process relies on user's subjective judgements. The major issue with relying on subjective input is that rules generated only encapsulate known patterns, thus excluding the discovery of unexpected but nonetheless important rules. Another issue is that the reliance on domain specific information constrains the range of applicability to only those domains where such information is readily available. There is no published work that is known to the authors that addresses these two issues. This motivated us to formulate a generic solution for the weight assignment problem that can be deployed across different application domains.

### 3 The Weighted Association Rule Mining (WARM) Problem

Given a set of items,  $I = \{i_1, i_2, \dots, i_n\}$ , a transaction may be defined as a subset of  $I$  and a dataset as a set  $D$  of transactions. A set  $X$  of items is called an itemset. The support of  $X$ ,  $\text{sup}(X)$ , is the proportion of transactions containing  $X$  in the dataset. An *association rule* is an implication of the form  $X \rightarrow Y$ , where  $X \subset I$ ,  $Y \subset I$ , and  $X \cap Y = \emptyset$ . The rule  $X \rightarrow Y$  has *support* of  $s$  in the transaction set  $D$ , if  $s = \text{sup}(XY)$ . The rule  $X \rightarrow Y$  holds in the transaction set  $D$  with *confidence*  $c$  where  $c = \text{conf}(X \rightarrow Y) = \text{sup}(XY)/\text{sup}(X)$ . Given a transaction database  $D$ , a support threshold *minsup* and a confidence threshold *minconf*, the task of association rule mining is to generate all association rules that have support and confidence above the user-specified thresholds.

In weighted association rule mining a weight  $w_i$  is assigned to each item  $i$ , where  $-1 \leq w_i \leq 1$ , reflecting the relative importance of an item over other items that it is

associated with. The weighted support of an item  $i$  is  $w_i \text{sup}(i)$ . Similar to traditional association rule mining, a weighted support threshold and a confidence threshold is assigned to measure the strength of the association rules produced. The weight of a  $k$ -itemset,  $X$ , is given by:

$$\left( \sum_{i \in X} w_i \right) \text{sup}(X) \quad (1)$$

Here a  $k$ -itemset,  $X$ , is considered a frequent itemset if the weighted support of this itemset is greater than the user-defined minimum weighted support ( $w\text{minsup}$ ) threshold.

$$\left( \sum_{i \in X} w_i \right) \text{sup}(X) \geq w\text{minsup} \quad (2)$$

The weighted support of a rule  $X \rightarrow Y$  is:

$$\left( \sum_{i \in X \cup Y} w_i \right) \text{sup}(XY) \quad (3)$$

**Algorithm: Weighted Association Rule Mining (WARM)**

**Input:** Transaction database  $D$ , weighted minimum support  $w\text{minsup}$ , universe of items  $I$

**Output:** Weighted Frequent itemsets

```

 $L_k \leftarrow \{\{i\} | i \in I, \text{weight}(c) * \text{support}(c) > w\text{minsup}\}$ 
 $k \leftarrow 1$ 
while ( $|L_k| > 0$ ) do
   $k \leftarrow k + 1$ 
   $C_k \leftarrow \{x \cup y | x, y \in L_{k-1}, |x \cap y| = k - 2\}$ 
   $L_k \leftarrow \{c | c \in C_k, \text{weight}(c) * \text{support}(c) > w\text{minsup}\}$ 
 $L_k \leftarrow \bigcup_k L_k$ 

```

An association rule  $X \rightarrow Y$  is called an interesting rule if  $X \cup Y$  is a large itemset and the confidence of the rule is greater than or equal to a minimum confidence threshold. A general weighted association rule mining algorithm [10] is shown above. The algorithm requires a weighted minimum support to be provided. In this algorithm  $L_k$  represents the frequent itemsets also known as the large itemsets and  $C_k$  represents the candidate itemsets. Candidate itemsets whose weighted support exceeds the weighted minimum support are considered large itemsets and will be included in the rule generation phase.

Thus it can be seen that item weighting enables items with relatively low support to be considered interesting (large) and conversely, items which have relatively high support may turn out to be uninteresting (not large). This adds a new dimension to the classical association rule mining process and enables rules with high weights in their rule terms to be ranked ahead of others, thus reducing the burden on the end user in sifting through and identifying rules that are of the greatest value.

## 4 Valency Model

The Valency model is based on the intuitive notion that an item should be weighted based on the strength of its connections to other items as well as the number of items that it is connected with. We say that two items are connected if they have occurred together in at least one transaction. Items that appear often together when compared

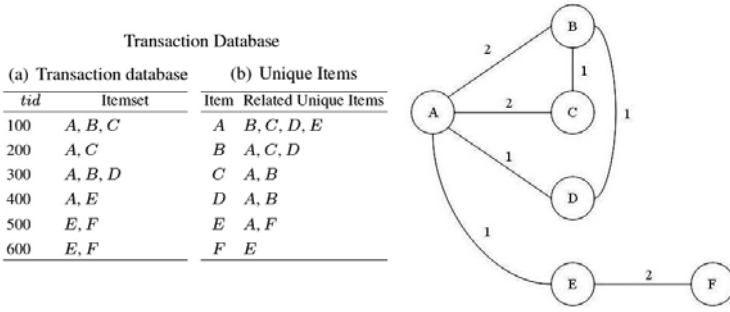


Fig. 1. Items Graph

to their individual support have a high degree of connectivity and are thus weighted higher. At the same time, an item that is contained in a small clique of items is said to have a high degree of purity and is given a proportionally higher weight. We will formally define the notions of *connectivity* and *purity* with the help of the following example.

Figure 1 is an example transaction dataset which can be represented as a graph whereby the nodes represent an item and the edges represent the support of the two items as an itemset. For example, the edge between node A and node B has a strength of 2, meaning that A and B occur together twice in the dataset. The Valency model we developed for our research is inspired by the Inverse Distance Weighting function (which was proposed by Shepard [13]). Inverse distance weighting is an interpolation technique which generates values for unknown points as a function of the values of a set of known points scattered throughout the dataset. In defining purity we took into account the importance of an item being featured in a rule term. In general, we prefer rules that have items that are not merely associated with each other strongly but also are distinctive in the sense that they appear with relatively few items. Such distinctive items add value to rules as they are more likely to capture genuine affinities, and possibly causal effects than an item that is selected only on the basis of a strong statistical correlation [14]. A strong statistical correlation between two items does not always indicate that a natural affinity exists between them. Consider, for example an item  $X$  having very high support that ends up being associated with many items  $Y, Z$ , etc merely because of the fact that it occurs in a very large fraction of the transactions, thus making the associations between  $(X, Y)$  and  $(X, Z)$  spurious, even though the correlations between  $X$  and  $Y$  on the one hand and  $X$  and  $Z$  on the other hand are high. Keeping these facts in mind, we formally define the purity,  $p$ , for a given item  $k$  as:

$$p_k = 1 - \frac{\log_2(|I_k|) + \log_2(|I_k|)^2}{\log_2(|U|)^3} \tag{4}$$

Where  $|U|$  represents the number of unique items in the dataset and  $|I_k|$  represents the number of unique items which are co-occurring with item  $k$ . Purity as defined in Equation 4 ensures that the maximum purity value of 1 is obtained when the number of items linked with the given item is 1, whereas the purity converges to the minimum

value of 0 as the number of linkages increases and become close to the number of items in the universal set of items. We chose to model purity with a non-linear logarithmic function as we wanted it to decrease sharply with the number of linkages. The  $\log(|U|)^3$  term in the denominator ensures that the rate of decrease in purity is sensitive to the size of the universal set. For databases with a larger number of items (larger  $|U|$ ) the gradient of descent is steeper when compared to databases with a smaller pool of items (smaller  $|U|$ ) and so a smaller number of items will acquire high purity values. The second contribution to an item's valency relies on how strongly it is connected to its neighboring items, or its *connectivity*. Given an item  $k$  which is connected to  $n$  items in its neighborhood, the connectivity,  $c_k$  is defined as:

$$c_k = \sum_i^n \frac{\text{count}(ik)}{\text{count}(k)} \quad (5)$$

We can now define the valency contained by an item  $k$ , denoted by  $v_k$  as the combination of both the purity and the connectivity components:

$$v_k = \beta.p_k + (1 - \beta). \sum_i^n \frac{\text{count}(ik)}{\text{count}(k)}.p_i \quad (6)$$

where  $\beta$  is a parameter that measures the relative contribution of the item  $k$  over the items that it is connected with in the database. The objective of the Valency model is to capture rules over small cliques of items such that items within a given clique have high purity and connectivity with other items contained within that clique. Since all items within a given clique are connected to each other, it follows from our definition of purity that all items within a clique have the same purity. Thus we can re-write the above equation as:

$$v_k = \beta.p_k + (1 - \beta).p_k. \sum_i^n \frac{\text{count}(ik)}{\text{count}(k)} \quad (7)$$

Thus, for a given item  $k$ , the relative contribution made to the valency by other items in the clique is dependent on both the value of the parameter  $\beta$  as well as the sum of the connectivity values from item  $k$ . We set the value of  $\beta$  as:

$$\beta = \frac{1}{n} \sum_i^n \frac{\text{count}(ik)}{\text{count}(k)} \quad (8)$$

With this setting of  $\beta$  we can re-write Equation 7 as:

$$v_k = \beta.p_k + n\beta(1 - \beta).p_k \quad (9)$$

With this setting of  $\beta$  we can see from the above expression that the relative contribution of the neighboring items of  $k$  over itself is  $1 - \beta$ , which means that as the value of  $\beta$  increases the item  $k$  itself assumes more importance in relation to its neighbors. We use the valency of an item as its weight. The weight calculation for an item is thus a computationally straightforward process as the weight for an item is independent of the

weights of other items. Also, the weight assignment process for items can be accomplished in the first pass through the dataset as the local neighborhoods for each item can be computed on the fly together with the reading of the dataset. In the next section we discuss our evaluation criteria for determining the quality of the rules obtained by applying the Valency model.

## 5 Rule Evaluation Methodology

A vast array of metrics for evaluating the quality of association rules have been proposed in the literature. Apart from the standard metrics of rule Support and Confidence, other measures such as Lift, Information Gain, Conviction, and Correlation have been used. The standard metrics are excellent at evaluating rules at the individual level in terms of the strength of correlation between terms and in assessing predictive accuracy. However, in the context of weighted association rule mining it is necessary that the contribution from each rule item is quantified and the contribution that it makes to the overall rule quality be assessed. Existing metrics tend to operate on the rule level rather than on the individual item level. This motivated us to investigate the use of Principal Components Analysis (PCA) to evaluate the quality of our weighted association rule miner.

PCA is a mathematical technique that has been widely used in the data mining arena. Basically, PCA takes a set of variables and finds a set of independent axes (the Principal Components) which explain all or most of the variation that occurs within the dataset. Its main application is in the area of classification and clustering where it is used as a pre-processing technique for dimensionality reduction. It has also been used before in association rule mining, but in a limited context where items are defined on a true numerical scale [15]. However, our use of PCA is quite different.

We concentrate solely on the right hand sides (RHSs) of rules as they encapsulate the actionable components of rules. Focussing on rule consequents allows us to test the degree of diversity amongst the actionable components discovered by the rule generator without the confounding effect of diversity amongst the left hand sides (LHSs) of rules. A set of rules with exactly the same RHS does not yield as much knowledge as rules that are diverse in their RHSs. For example, a set of rules *egg, bread* → *milk*; *butter, bread* → *milk*; and *tuna, egg* → *milk*, can be considered less interesting than rules with a greater diversity such as *diaper* → *baby food*; *ham* → *cheese*; and *chips* → *soda*. In a medical database containing information about a number of different diseases a rule generator that has poor coverage of the set of diseases (i.e. only identifies a small fraction of the diseases in the RHSs of the rules) is not as useful as one that has a better coverage with respect to the set of diseases.

We first apply PCA to the transaction dataset and obtain the Eigen vectors for the first two principal components. These vectors will be a linear function of the form:  $e_{k1}I_1 + e_{k2}I_2 + \dots + e_{kn}I_n$  where  $e_{kp}$  is the Eigen value for the  $p$ th item on the  $k$ th principal component ( $k$  is either 1 or 2). We process the rule set by removing LHS of each rule. This results in a collection of rule consequents (RHSs) containing duplicate entries as the LHS terms have been eliminated. After duplicate elimination we obtain a more compact representation of the rule set,  $R$ . We project the rule set  $R$  on its first two principal components and obtain a quantified version of the rule set, which we

denote by  $S$ . The set  $S$  contains a set of ordered pairs  $(X, Y)$  where  $X, Y$  are vectors representing principal components 1 and 2 respectively for each rule.

PCA enables us to capture the amount of variance that is explained by each rule term for each rule. The greater the amount of variance that is captured by a rule term, the higher the value of that term and the higher the contribution it makes to the rule as a whole. Thus PCA provides us with an independent method of evaluating the efficacy of our Valency model. If our Valency model is to outperform an unweighted association mining scheme such as Apriori then the delineation of the rules in PCA space produced by our Valency model should be better. In order to assess the quality of the delineation we applied the K-means clustering algorithm to the  $(X, Y)$  vectors and then visualized the clusters. We also quantified the degree of delineation by calculating a cluster purity measure along the axis that provided the better delineation, which happened to be the first principal axis (rather than the second) in most of the experiments that we carried out. In the next section we present the results of our experimentation with our Valency model.

## 6 Experimental Results

Our motivation in introducing the Valency model was to facilitate the automatic assignment of weights from a given transaction dataset without requiring additional information from users. As such we were interested in examining the impact of the weight assessment process in an environment where user input is not available, and this led us to compare our algorithm with the classical Apriori approach. Our experimentation was conducted in two steps, firstly a performance comparison with Apriori, and secondly an examination of the impact of key parameters of the Valency model. We used seven UCI datasets [16]. We also experimented with synthetic data for which we used the data generator proposed by [11]. Datasets  $D$ , were created with the following parameters: number of transactions  $|D|$ , average size of transactions  $|T|$ , number of unique items  $|I|$ , and number of large itemsets  $|L|$ .

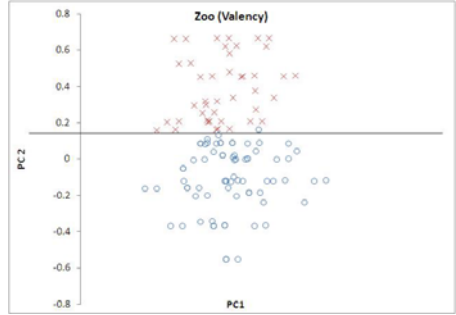
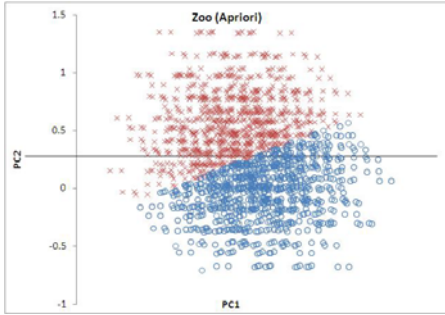
### 6.1 Principal Components Analysis of the Rule Bases

Table 1 shows the results of running both Apriori and our algorithm on the datasets mentioned above. Each row shows the dataset, the number of rules produced, the number of RHSs produced (bracketed), and the cluster purity obtained by clustering the resulting rule bases on the first two principal components. We see from the results that the effect of weighting is to produce a much more compact rule base as Valency's rule base, with the exception of Soybean, is much smaller than Apriori's. In order to keep the comparison fair we ran the two algorithms at minimum support thresholds so that they produced rules bases which had approximately the same support distributions. The compact nature of Valency's rule base vis-a-vis Apriori is due to the influence of the purity component that reduces the weighted support of an item sharply as the number of items that it interacts with increases. We verified this by substituting the non linear purity function with a linear one. The linear function did not punish highly connected items as severely as its non linear counterpart, thus resulting in a rule base that exploded in size and became very similar to that of Apriori in both qualitative and quantitative



**Table 1.** Clustering Results for First Two Principal Components

Dataset	Apriori		Valency		Improvement
	No. of Rules	Cluster Purity	No. of Rules	Cluster Purity	
Bridges	1875(15)	100	505(15)	100	0.0
Flag	486500(719)	86.6	119948(121)	94.2	8.8
Flare	244753(1052)	90.9	715(32)	100	10.0
Hepatitis	720633(2065)	87.9	45850(233)	96.6	9.9
Mushroom	61515(1064)	92.5	4005(134)	92.5	0.0
Soybean	188223(1211)	82.9	456753(1310)	99.7	20.3
Synthetic (T25I200D1K)	618579(2195)	89.7	266914(853)	98.4	9.7
Zoo	644890(3128)	89.7	5475(127)	99.2	10.6



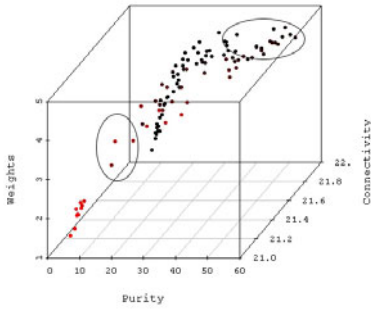
**Fig. 2.** PC1 and PC2 Clusters for Zoo Dataset based on Apriori

**Fig. 3.** PC1 and PC2 Clusters for Zoo Dataset based on Valency

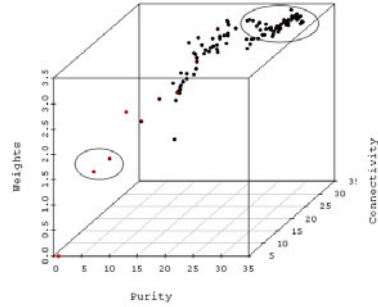
terms. The effect of the linear function was to dilute the effect of purity and hence the weighting scheme was not as effective in discriminating between different items, thus resulting in a larger proportion of items assuming higher purity and higher weight values.

The second interesting aspect of the results is that Valency produced a better set of clusters when compared to Apriori. The cluster purity improvement measure for Valency ranges from 0% for the relatively sparse Bridges and Mushroom datasets to 20.3% for the denser Soybean dataset. This improvement is due to the fact that the items that feature in Valency’s rules capture a higher proportion of variance that occurs over the underlying dataset in relation to Apriori. This result confirms our hypothesis that it is possible to automatically deduce weights from the interactions that occur between items in a transactional database. The result for the experimentation with various types of synthetic datasets were broadly similar to that of the real-world datasets. As the density of the dataset increased so did the improvement in cluster purity value between Valency and Apriori. We do not report on all synthetic datasets due to lack of space. Instead we present the result for one such dataset (T25I200D1K) that is representative of the experimentation that we conducted with synthetic data.

Figures 2 and 3 shows the clusters generated in PCA space for the Apriori and Valency schemes on the Zoo dataset. For the Zoo dataset the 2nd principal component produced the cleaner demarcation between the clusters. The figures show that Valency produces a visibly better separation of clusters around the point of intersection with the



**Fig. 4.** Weights, Purity, and Connectivity for Mushroom Dataset



**Fig. 5.** Weights, Purity, and Connectivity for Soybean Dataset

second principal axis. The other visualizations which we could not include due to lack of space, were similar except that the axis of greater separation was the first principal component for both of the algorithms.

## 6.2 Impact of Purity and Connectivity on Item Weight

In this part of the experimentation we investigated the simultaneous effects of purity and connectivity on item weight. For each of the datasets that we experimented with we plotted item weight versus purity and connectivity in a 3D representation in order to assess the simultaneous effects of purity and connectivity on weight.

Figures 4 and 5 show the results for two representative datasets, namely Mushroom and Soybean. It is clear from the plots that high weights only result if both purity and connectivity are high at the same time (the ovals on the north east corner of the cubes). We can also see the filtering effect of purity on weight. Items with high connectivity but low purity end up with lower weight (denoted by the ovals on the south west corners of the cubes). Algorithms such as Apriori (and all such un-weighted association rule mining algorithms) that do not discriminate on purity will tend to capture rules that contain items that occur with a large number of other items thus producing rules that are unlikely to be novel or useful to the decision maker.

## 6.3 Case Study Zoo Dataset

We compare the results on the zoo dataset based on the rule bases produced by the Apriori and the Valency schemes. Using Apriori, we found that the item *toothed = 1* occurs with 33/40 (or 82.5%) of the other items, and that it appears in 431079/644890 (or 66.8%) of the rules. The item *toothed = 1* thus serves to dilute the effects of all rules that it participates in, which happens to be the majority (66.8%) of the Apriori rule base. The subrule  $\{eggs = 0, legs = 4\} \rightarrow \{toothed = 1\}$  appeared in all of the top 20 rules when ranked by Lift. Considering that all three items within this subrule have low weights, and given the fact that all of Apriori's top 20 rules embed this subrule, it follows that the degree of diversity captured by Apriori's top ranked rules happens to be low. With the Valency scheme, we noticed that *toothed = 1* occurs with 1508/5474

(27.5%) of the other rules. When we ranked the rules by lift, we noticed that *toothed = 1* first appeared in rule 317 with a lift value of 2.06.

For Apriori, the items which appeared in the first 20 rules were: *eggs = 0*, *legs = 4*, *airborne = 0*, *fins = 0*, *hair = 0*, *toothed = 1*, *milk = 1*, *backbone = 1*, *breathes = 1*, *type = 1*, and *feathers = 0*. Out of the 11 items only 4 items are of high weight. The average weight for the items was 1.16. With the Valency scheme, the items that appeared in the first 20 rules were *backbone = 1*, *breathes = 1*, *milk = 0*, *venomous = 0*, *eggs = 1*, *fins = 0*, *tail = 1*, and *domestic = 0*. The average weight for the items was 1.34.

The above results illustrate the extent to which items that are not distinctive can dilute the efficacy of rules produced by an association miner that does not utilize item weighting. On the other hand the Valency model is more discriminative in its use of items such as *toothed = 1*. Whenever such items feature in its rule base, it tends to include items with higher weight to compensate, thus mitigating the effects of such lowly weighted items.

## 7 Conclusions and Future Work

In this paper, we propose a new item weighting scheme based on the Valency model. We fit the weights to items based on the notions of connectivity and purity. The valency weighting function consists of two different parts: weights of an item based on its strength of connections and weights of its neighboring items. We used PCA to investigate the degree of variation captured by the rule bases. Overall, the Valency model produces better rules than traditional Apriori. In terms of future work we would like to investigate the effects of not just the immediate neighbors on an item's weight but to also capture the effects of non-neighboring items that are not directly connected to the given item under consideration.

## References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
2. Cai, C.H., Fu, A.W.C., Cheng, C.H., Kwong, W.W.: Mining association rules with weighted items. In: IDEAS 1998: Proceedings of the 1998 International Symposium on Database Engineering & Applications, Washington, DC, USA, p. 68, IEEE Computer Society, Los Alamitos (1998)
3. Sun, K., Bai, F.: Mining weighted association rules without preassigned weights. *IEEE Trans. on Knowl. and Data Eng.* 20(4), 489–495 (2008)
4. Wang, W., Yang, J., Yu, P.S.: Efficient mining of weighted association rules (WAR). In: KDD 2000: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 270–274. ACM, New York (2000)
5. Yan, L., Li, C.: Incorporating pageview weight into an association-rule-based web recommendation system. In: Sattar, A., Kang, B.-h. (eds.) AI 2006. LNCS (LNAI), vol. 4304, pp. 577–586. Springer, Heidelberg (2006)
6. Liu, B., Hsu, W., Ma, Y.: Mining association rules with multiple minimum supports. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 337–341 (1999)

7. Koh, Y.S., Rountree, N.: Finding sporadic rules using apriori-inverse. In: Ho, T.-B., Cheung, D., Liu, H. (eds.) PAKDD 2005. LNCS (LNAI), vol. 3518, pp. 97–106. Springer, Heidelberg (2005)
8. Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Motwani, R., Ullman, J.D., Yang, C.: Finding interesting association rules without support pruning. *IEEE Transactions on Knowledge and Data Engineering* 13, 64–78 (2001)
9. Sanjay, R., Ranka, S., Tsur, S.: Weighted association rules: Model and algorithm (1997), <http://citeseer.ist.psu.edu/185924.html>
10. Tao, F., Murtagh, F., Farid, M.: Weighted association rule mining using weighted support and significance framework. In: KDD 2003: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 661–666. ACM, New York (2003)
11. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Bocca, J.B., Jarke, M., Zaniolo, C. (eds.) Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, pp. 487–499 (1994)
12. Jian, W., Ming, L.X.: An effective mining algorithm for weighted association rules in communication networks. *Journal of Computers* 3(4) (2008)
13. Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 23rd ACM national conference, pp. 517–524. ACM, New York (1968)
14. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson Inc., London (2006)
15. Faloutsos, C., Korn, F.A., Labrinidis, Y., Kotidis, A.K., Perkovic, D.: Quantifiable data mining using principal component analysis. Technical Report CS-TR-3754, Institute for Systems Research, University of Maryland, College Park, MD (1997)
16. Asuncion, A., Newman, D.: UCI machine learning repository (2007), <http://www.ics.uci.edu/~mllearn/MLRepository.html>