Non-Redundant Rare Itemset Generation

Yun Sing Koh[1] **Russel Pears**[2]

School of Computing Science and Mathematics
Auckland University of Technology, New Zealand,
Email: ykoh@aut.ac.nz[1], rpears@aut.ac.nz[2]

Abstract

Rare itemsets are likely to be of great interest because they often relate to high-impact transactions which may give rise to rules of great practical significance. Research into the rare association rule mining problem has gained momentum in the recent past. In this paper, we propose a novel approach that captures such rare rules while ensuring that redundant rules are eliminated. Extensive testing on real-world datasets from the UCI repository confirm that our approach outperforms both the Apriori-Inverse(Koh et al. 2006) and Relative Support (Yun et al. 2003) algorithms.

Keywords: Rare Association Rule Mining, Apriori-Inverse, Non-Redundant Itemset

1 Introduction

Association rule mining (Agrawal et al. 1993) is used to find common or frequent patterns within datasets. In the classical association rule mining process, all frequent itemsets are found, where an itemset is said to be frequent if it appears above a minimum frequency threshold s, called minimum support. Association rules are then derived from frequent items and are represented in the form $A \rightarrow B$ where AB is a frequent itemset. Strong association rules are those that meet the minimum confidence c threshold (the percentage of transactions containing A that also contain B).

The minimum support threshold is used as a noise filter to eliminate itemsets that do not appear often within the dataset. This threshold has to be sufficiently strong to reduce frequent itemsets to a manageable level. However, in some data mining applications relatively infrequent associations are likely to be of great interest as they relate to rare but crucial cases. Application domains that benefit from rare association mining include the diagnosis of rare diseases, the prediction of telecommunication equipment failure, and the identification of associations between infrequently purchased supermarket items.

For example in a supermarket transactional dataset, most purchasing behavior follows a very regular and predictable pattern,which is related to daily household items such as bread, butter, and milk. However, there also exists behavior which is uncharacteristic in respect to the volume of items sold when compared to the staple items mentioned earlier. Such behavioral patterns are potentially useful to the retailer as they could involve associations between items that are highly profitable. However, because of the relative infrequency with which such associations manifest, traditional frequent association mining techniques would be unable to capture the patterns involved. This is due to the combinatorial explosion in the number of candidate itemsets generated by setting the minimum support to a low enough value to capture the rare associations. Such an explosion in the search space renders traditional algorithm such as Apriori unusable. This problem was first highlighted by Cohen et al. (2000) , showing that association between expensive items such as vodka and caviar are likely to be infrequent but interesting due to their high value.

Current rare itemset generation techniques suffer from three issues. Firstly, their rule generators produce a mix of both frequent and infrequent rules (Liu et al. 1999, Szathmary et al. 2007). Furthermore, the rule bases they produce contain rules that could be inferred from other rules, thus making them redundant. In our experimentation we noticed the occurrence of such redundant rules when we did a comparative analysis with the Apriori-Inverse rule generator. Thirdly, they generate rules having only infrequent items in their rule terms, which represents only a sub class of rare rules (Koh et al. 2006). Such rule generators do not produce rare rules that consist of frequent items in their rule terms. Such rare rules are valuable as they represent scenarios where individual rule terms in rule antecedents are frequent on their own but rare in combination with each other. Such rules capture very specific but hard to detect detect events in their rule consequents, on account of their rarity. Up until now there has been no effective solution to the problems referred to earlier and this research represents an attempt to address each of these issues.

In this paper we introduce a novel approach called Non-Redundant Itemset Generation (N-RIG) which seeks to capture rare patterns by using an efficient pruning strategy, without the need for pre-processing the dataset by partitioning it. The non-redundant generator ensures that only itemsets that lead to an improvement in rule prediction accuracy are ever considered for rule generation. Pruning of redundant items is achieved by the introduction of a constraint that we introduce, known as Cumulative Productive Confidence (CPC).

The remainder of the paper is organized as follows. Section 2 provides a review of related research in the area of rare association rule mining. In Section 3 we give a brief overview of our new approach to finding rare association rules. In Section 4 we discuss the redundant itemset problem. In Section 5 we introduce the notion of an adaptive support threshold which further enhances the quality of the rules pro-

duced. Experimental results of applying the method on several real-world datasets is presented in Section 6. The paper concludes in Section 7 with a summary of the contributions made in this research.

2 Related Work

The efficient detection of rare association rules with low support but high confidence is a difficult data mining problem. To find such rules with traditional approaches such as the Apriori would require the minimum support (minsup) threshold to be set very low, resulting in large computational overhead while producing a large rule base, parts of which contain redundant rules. As a specific example of the problem, consider the association mining problem where we want to determine if there is an association between buying a food processor and buying a cooking pan (Liu et al. 1999). The problem is that both items are rarely purchased in a supermarket. Thus, even if the two items are almost always purchased together when either one of them is purchased, this association may not be found. Modifying the minsup threshold to take into account the importance of the items is one way to ensure that rare items remain in consideration. To find this association minsup must be set low. However setting this threshold low would cause a combinatorial explosion in the number of itemsets generated. Frequently occurring items will be associated with one another in an enormous number of ways simply because the items are so common that they cannot help but appear together. This is known as the rare item problem (Liu et al. 1999). It means that the application of Apriori-like approaches are unlikely to yield rules that indicate rare events of potentially dramatic consequence.

Liu et al. (1999) note that some individual items can have such low support that they cannot contribute to rules generated by Apriori, even though they may participate in rules that have very high confidence. They overcome this problem with a technique called MSApriori whereby each item in the database can have a minimum item support (MIS) given by the user. By tailoring the MIS value for different items, a higher minimum support is tolerated for rules that involve frequent items and a lower minimum support for rules that involve less frequent items. Yun et al. (2003) proposed the RSAA algorithm to generate rules in which significant rare itemsets take part, without the need for any user specified thresholds. This technique uses relative support measure, RSup in place of support. The RSup measure serves to decrease the support threshold for items that have low frequency and to increase the support threshold for items that have high frequency. In common with Apriori and MSApriori, RSAA is exhaustive in its generation of rules, and will generate rules which are not rare (i.e. rules with high support and high confidence).

Szathmary et al. (2007) presented an approach for rare itemset mining from a dataset that splits the problem into two tasks. The first task, the traversal of the frequent zone in the space, is addressed by two different algorithms, a naive one, Apriori-Rare, which relies on Apriori and hence enumerates all frequent itemsets; and MRG-Exp, which limits the considerations to frequent generators only. They consider computation of the rare itemsets that approaches them starting from the bottom of the itemset lattice and then moving upwards through the frequent zone. They defined a positive and the negative border of the frequent itemsets, and a negative lower border and the positive lower border of the rare itemsets, respectively. An itemset is a maximal frequent itemset (MFI) if it is frequent but all its proper supersets are rare. An itemset is a minimal rare itemset (mRI) if it is rare but all its proper subsets are frequent. If the minimum-allowable relative support value is set close to zero, MRG-Exp takes a similar amount of time to that taken by Apriori to generate low-support rules due to the need for sifting through the high-support rules.

Koh et al. (2006) proposed an approach to find rare rules with candidate itemsets that fall below a maxsup (maximum support) level but above a minimum absolute support value. They introduced an algorithm called Apriori-Inverse to find sporadic rules efficiently: for example, a rare association of two common symptoms indicating a rare disease. They used a maximum support threshold to prune out any items that may be frequent. They then use a minimum absolute support (minabssup) threshold value derived from an inverted Fisher's exact test (Weisstein 2005) to prune out noise. At the low levels of co-occurrences of candidate itemsets that need to be evaluated to generate rare rules, there is a possibility that such co-occurrences happen purely by chance and are not statistically significant. The Fisher test provided a statistically rigorous method of evaluating significance of co-occurrences and was thus an integral part of their approach. The main drawback of this method is that it cannot detect rare rules that embed frequent items in their rule terms.

Koh & Pears (2008) proposed a pre-processing mechanism, based on transaction clustering to generate rare association rules. The basic concept underlying transaction clustering stems from the concept of large items as defined by traditional association rule mining algorithms. In their approach, they partition the dataset and then run the Apriori-Inverse algorithm on each of the clusters found. They showed that pre-processing the dataset by clustering improves rule quality by as each cluster is able to express its own associations without interference or contamination from other sub groupings that have different patterns of relationship. The rare rules produced by each cluster were shown to be more informative than the rare rules found from direct association rule mining on the original unpartitioned dataset.

We have based our approach on Apriori-Inverse. Thus we use the minabssup threshold based on Fisher's exact test to filter out chance co-occurrences. However we differ from Apriori-Inverse in that we use the maxsup threshold only in the rule generation phase and not the candidate itemset phase. The rationale for this is explained in Section 4. In the next three sections we present our approach for rare association rule generation.

3 Our Approach

This section presents the key concepts governing the Non-Redundant Rare Itemset Generation (N-RIG) approach. The focus of our approach is to find rare rules that contain rule terms that may by themselves be frequent whilst preveting the generation of redundant rules. Our approach is adapted from the Apriori algorithm. Similar to Apriori, our approach is set in two phases, the candidate generation and the rule generation phase. We discuss the candidate generation phase below. The candidate generation phase itself consists of two steps.

In the first step, we allow itemsets that are above a minabssup threshold which we adopt from (Koh et al. 2006) and which fulfil our Cumulative Productive Confidence (CPC) measure to be extended. The minabssup threshold is calculated for every itemset and is used to eliminate noise and is used instead of a

fixed minimum support threshold. The CPC measure is used to eliminate *redundant* itemsets. We define a redundant itemset(I) as one that gives rise to a rule that can be inferred from a rule covered by some subset of itemset I. In the next section, we discuss the CPC measure in detail.

In the second step, use a maximum support threshold to prune the set candidate itemsets further. Here we prune out all itemsets that have support above the maximum support threshold. This is needed as we are only interested in rare itemsets.

Algorithm 1 N-RIG algorithm

Input: Transaction Database D, universe of items I, maximum support (maxsup) value
Output: Non-redundant Rare Itemsets
$N \leftarrow |D|$
$k \leftarrow 1$
$R_k \leftarrow \{\{i\}|i \in \text{dom } Idx, \text{count}(\{i\}) > 1\}$
while $R_k \neq \emptyset$ **do**
  $k \leftarrow k + 1$
  $C_k \leftarrow \{x \cup y | x, y \in R_{k-1}, |x \cap y| = k - 2\}$
  $R_k \leftarrow \{c | c \in C_k, \text{supp}(c) > \text{minabssup}, \text{CPC}(c) > 0\}$
end while
$R_k \leftarrow \{c | c \in R_k, \text{supp}(c) < \text{maxsup}\}$
return $\bigcup_{t=2}^{k-1} R_t$

4 The Redundant Itemset Problem

Despite the fact that Apriori-Inverse outperforms its rivals on performance and rule quality, there exists two areas where its performance can be improved. Firstly, it is possible that Apriori-Inverse generates rules that are redundant. In its itemset generation phase Apriori-Inverse combines itemsets A and B as long as they pass the Fisher test (Weisstein 2005). While the Fisher test does an excellent job of filtering out itemsets that co-occur together by chance, it does not guarantee rule minimality in the rule base that it generates.

Consider the following example with a dataset containing 50 transactions. Suppose that we have 3 itemsets A,B and C with support 20, 30 and 25 respectively. If $\text{supp}(AB) = 18$ and if AB co-occurs with every transaction with C, then we have $\text{supp}(AC) = 18$. With these statistics the Fisher test determines that items A and B do not occur by coincidence, thus Apriori-Inverse will record itemset AB as a candidate itemset for rule generation. If the minimum confidence threshold is set to 0.8 then the rule $A \rightarrow B$ will be generated as the rule confidence at $18/20 = 0.9$ exceeds the confidence threshold set.

Since $\text{supp}(AC) = 18$ it follows that this itemset too will pass the Fisher test, thus producing AC as another itemset. In the next level of itemset generation Apriori-Inverse will consider the generation of ABC from the candidate pairs AB and AC. Now $\text{supp}(ABC) = \text{supp}(AB)$ since A always co-occurs with C and hence it follows that ABC will also pass the Fisher test. This in turn leads to the following rule:

$AC \rightarrow B$ This rule too meets the confidence threshold as its confidence is:

$$\frac{\text{supp}(ABC)}{\text{supp}(AC)} = \frac{\text{supp}(AB)}{\text{supp}(AC)} = 1$$

since $\text{supp}(AB) = \text{supp}(AC)$.

However, it is clear that Rule 2 is redundant in the presence of Rule 1. Rule 1 captures the minimal conditions required to predict the occurrence of B given A. This example illustrates that Apriori-Inverse is vulnerable to the redundant rule generation problem. Whilst it is possible to apply a post rule generation filter to remove such redundant rules, a more efficient approach would ensure that itemsets such as ABC are never generated in the first place. The generation of itemset ABC has the potential to lead to even more redundancy as all pairs of itemsets such as (ABC, ABD) with a common prefix of AB propagates the redundancy of ABC with other items such as D, leading to many more itemsets such as $ABCD$ that give rise to redundant rules. This is clearly undesirable since the candidate generation phase is the performance bottleneck in the association rule mining process.

Our approach avoids this problem by pruning itemsets such as ABC from the set of candidates, thus ensuring that redundancy is eliminated at its source. We use an improvement measure called CPC, that ensures that any given itemset will only be extended if its extension produces an increase in the CPC measure over the improvement value when the itemset itself was being formed. In section 4.1 we show that no itemset that passes the improvement test will be redundant.

The second issue with Apriori-Inverse is that it uses a fixed threshold for determining rarity. The use of a fixed threshold inhibits the discovery of rules for items whose support is above the threshold but who co-occur together with support less than the threshold set. Consider for example items X and Y with support 0.2 and 0.3 respectively. Suppose that the support of XY is 0.08, and the maximum support threshold is set at 0.10. Apriori Inverse only combines items that meet the maximum support threshold constraint and thus X and Y will not be combined together, although their combination gives rise to a rare rule with support 0.08. This simple example illustrates that it would be desirable to explore items in the neighborhood of the maximum support threshold with a view to expanding Apriori-Inverse's rule base to capture rare rules that contain one or more terms that are frequent. Such types of rules are of interest in many types of applications. Such applications include disease diagnosis where certain symptoms occur on their own commonly but whose co-occurrence points to a specific disease that occurs rarely in the population

We now turn our attention to the issue of preventing the occurrence of redundant rules.

4.1 Redundancy Removal

As mentioned above Apriori-Inverse is vulnerable to the problem of redundant rules. Such redundant rules contain terms in the rule antecedent that do not contribute to an increase in the rule confidence. Such rules not only increase the size of the rule base unnecessarily, but also tend to mislead the decision maker into thinking that certain terms need to be satisfied in the antecedent when in reality they do not.

We first give a formal definition of rule redundancy from Bayardo (Bayardo 1998). Consider a generic rule $X \rightarrow Y$. The improvement, I, of such a rule is:

$$\text{I}(X \rightarrow Y) = \text{conf}(X \rightarrow Y) - (\text{conf}(Z \rightarrow Y)), \max(Z \subset X)$$

A redundant rule can now be defined as one whose improvement is less than zero. We prevent the occurrence of such rules by defining a metric called *Cumulative Productive Confidence* (CPC) that measures whether an extension to a given itemset will ensure that all rules that can be produced as a result of the extension have greater confidence than the rules produced with the original non extended version the itemset.

Suppose that we have itemsets X and Y that have passed the Fisher test. Itemset X will be merged with

Y and extended to XY if it satisfies the condition below:

$$CPC(XY) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} - \frac{\text{supp}(Y)}{\arg\min_{W \subset Y} \text{supp}(W)} > 0$$

Theorem 1 below offers a formal proof that the CPC measure inhibits the production of redundant rules.

Theorem 1. *All rules produced from an extension of an itemset that satisfies the CPC constraint defined above will be non redundant.*

Proof. Itemsets X and Y to be merged need to have a common prefix so we will represent X as AB and Y as AC. We now have $X \cup Y = ABC$. For ABC to be a legitimate itemset, itemset $Z = BC$ must also exist and have passed the Fisher test, since ABC can also be produced by $Y \cup Z$ and thus we cannot have ABC without Z passing the Fisher test as well. In order to produce ABC, it then follows that we must have

$$CPC(X, Y) > 0 \tag{1}$$

$$CPC(X, Z) > 0 \tag{2}$$

$$CPC(Y, Z) > 0 \tag{3}$$

From 1 we have:

$$\frac{\text{supp}(AB \cup AC)}{\text{supp}(AB)} - \frac{\text{supp}(AC)}{\arg\min_{W \subset AC} \text{supp}(W)} > 0$$

This implies that:

$$\frac{\text{supp}(ABC}{\text{supp}(AB)} - \frac{\text{supp}(AC)}{\text{supp}(A)} > 0 \tag{4}$$

since

$$\frac{\text{supp}(AC)}{\text{supp}(A)} \geq \frac{\text{supp}(AC)}{\arg\min_{W \subset AC} \text{supp}(W)}$$

From 4 it follows that

$$\text{conf}(AB \rightarrow C) > \text{conf}(A \rightarrow C)$$

By substituting C instead of A in 4 above we also have: $\text{conf}(AB \rightarrow C) > \text{conf}(C \rightarrow A)$. Thus the rule $AB \rightarrow C$ is non redundant. From 2 we have:

$$\frac{\text{supp}(BC \cup AC)}{\text{supp}(BC)} - \frac{\text{supp}(AC)}{\arg\min_{W \subset AC} \text{supp}(W)} > 0$$

From this we derive:

$$\frac{\text{supp}(BCA}{\text{supp}(BC)} - \frac{\text{supp}(AC)}{\text{supp}(A)} > 0 \tag{5}$$

As with 4 above we have:

$$\frac{\text{supp}(AC)}{\text{supp}(A)} \geq \frac{\text{supp}(AC)}{\arg\min_{W \subset AC} \text{supp}(W)}$$

From 5 it follows that:

$$\text{conf}(BC \rightarrow A) > \text{conf}(A \rightarrow C)$$

By substituting C instead of A in 5 above we also have:

$$\text{conf}(BC \rightarrow C) > (\text{conf}(C \rightarrow A)$$

We thus have the rule $BC \rightarrow C$ non redundant and lastly using 3 above we have:

$$\frac{\text{supp}(AC \cup BC)}{\text{supp}(AC)} - \frac{\text{supp}(BC)}{\arg\min_{W \subset AC} \text{supp}(W)} > 0$$

which leads to: $\text{conf}(AC \rightarrow B) > \text{conf}(A \rightarrow C)$and $\text{conf}(AC \rightarrow B) > \text{conf}(C \rightarrow A)$ which means that rule $BC \rightarrow C$ is also redundant. We have thus shown that all rules produced by the extension are non redundant and this proves the theorem.

5 Adaptive Thresholding

Although N-RIG dispenses with a maximum support threshold during itemset generation it still uses such a threshold during the rule gneration phase to esnure that only rare rules are generated. However, the use of such a threshold can have undesirable effects if its value is set arbitrarily. For example, by setting a threshold at 0.10 on a dense dataset, we would be letting through more itemsets when compared to setting the threshold at the same value on a sparse dataset. To find a suitable cut off point we use an adaptive threshold based on a modified version of a hill climbing algorithm. We inspect the support of the candidate itemset. Using a list of itemsets sorted in ascending order of support, we compare the support of itemset x to the support of itemset $x + 1$. If the difference of the support is less than $k\%$, the new support threshold is set as $\text{supp}(x + 1)$. The process is repeated until the difference between two consecutive itemsets is more than $k\%$, and we consider that we have reached a partition in the itemset support distribution that defines a suitable threshold value.

In the next section, we present the results from our approach and compare them with those produced by the Apriori-Inverse algorithm.

6 Evaluation and Results

In this section, we compare the performances of the standard Apriori-Inverse and RSAA algorithms with our proposed algorithm. The experiments were performed on a Windows Vista machine with Intel Duo Core having 3.0GHz CPU and 2.68 GB of RAM. Testing of the algorithms was carried out on 5 different datasets from the UCI Machine Learning Repository (Newman et al. 1998). Table 1 represents the summary of the results found using Apriori-Inverse, N-RIG, and RSAA algorithms. For Apriori-Inverse and N-RIG we set the maximum support threshold (maxsup) to 0.10 for all datasets. In all of the experiments, we set the minimum confidence threshold to 0.90. For a comparison have reported the number itemsets found by RSAA which fell below the 0.10 threshold.

We compare the time taken to produce the rare itemsets. Overall our approach generated more itemsets when compared to Apriori-Inverse. On the average, we generated 1229 itemsets as compared to Apriori-Inverse which generated an average of 682. The N-RIG approach is not merely confined to generating itemsets that contain only infrequent items, unlike Apriori-Inverse. This explains the difference in the overall number of itemsets produced between the two approaches. Despite the greater effort expended by N-RIG in expanding the scope of rare itemsets produced its run time compares well with that of Apriori-Inverse. The number of rare itemsets produced by RSAA was consistently higher than with the other two algorithms. In line with the greater number of itemsets produced RSAA runtime were also much higher than with the other two algorithms. In the case of the Soybean dataset, RSAA performed very

Table 1: Summary of Experimental Results

Dataset	Apriori-Inverse		N-RIG		RSAA	
	Rare Itemset	Time (s)	Rare Itemset	Time (s)	Rare Itemset	Time (s)
Flag	72	0.51	260	7.60	1210	98.2
Hepatitis	31	0.08	51	2.42	398	19.53
Soybean-Large	135	0.45	1289	88.16	6226	2388.67
Audiology	123	0.51	239	12.30	N/A	N/A
Mushroom	3051	55.76	4305	349.00	5804	360.56

poorly with respect to the runtime. As for the Audiology dataset, RSAA did not terminate after two hours and hence we decided to exclude it from the comparision.

6.1 Comparative Analysis

Table 2 shows clearly that both Apriori-Inverse and N-RIG both produce rules with high lift with the top 20 Lift values being identical for the larger datasets, Soybean and Mushroom. However, the lift values for RSAA was significantly smaller for these datasets. For the two smaller datasets, Flag and Hepatitis Apriori-Inverse generated just 3 and 7 rules respectively and so a meaningful comparison with N-RIG was not possible. RSAA produced mixed results for two smaller datasets, giving a higher lift for Hepatitis while producing a lower lift value for the Flag dataset.

Table 2: Rule Lift across selected UCI datasets

Dataset	Apriori-Inverse	N-RIG	RSAA
Flag	-	12.3	5.5
Hepatitis	-	6.70	11.0
Audiology	100	34.7	N/A
Soybean-Large	51.2	51.2	15.4
Mushroom	1015.5	1015.5	6.3

Table 3 illustrates a clear difference in behavior between the algorithms. While the top 20 rule support and overall rule support values are broadly similar for the Apriori-Inverse and N-RIG algorithms, the rule term support for N-RIG was significantly higher, particularly for the Soybean and Mushroom datasets. As shown in Table 3 the average antecedent rule support for the Mushroom dataset at 2.6% is a factor of 26 times higher than the corresponding value for Apriori Inverse. The same trend holds true for the smaller datasets, albeit on a smaller scale. We chose to exclude RSAA from further analysis as its lift values were smaller than with the other two approaches.

These results suggests that N-RIG is better able to capture rare rules where individual terms are frequent. As pointed out in Section 1 such rules are of great practical significance.

Such rules manifest with N-RIG as it is not restricted by a maximum support constraint in its candidate itemset generation phase, unlike Apriori-Inverse, thus enabling the former to produce rule terms of higher support than the latter. We next examine some of the rules discovered by N-RIG which Apriori-Inverse was unable to generate.

6.1.1 Mushroom Dataset

N-RIG produced a number of very rare and very high lift rules involving different combinations of the same terms appearing together. One such rule is given below with support 0.1% and Lift of 1015.5. Such extremely rare rules in the occurrence of a relatively dense dataset such as Mushroom tends to boost the

Table 3: Rule Support and Rule Term Support Comparison

Dataset	Apriori-Inverse			
	Support (Top 20 Rules)	Antecedent Rule Support	Consequent Rule Support	Support (Entire Rule Base)
Flag	2.1%	5.2%	2.1%	4.1%
Hepatitis	3.9%	3.9%	3.9%	7.1%
Audiology	1.0%	1.2%	3.3%	1.2%
Soybean-Large	2.0%	2.0%	2.3%	5.2%
Mushroom	0.1%	0.1%	0.3%	0.4%
Dataset	N-RIG			
	Support (Top 20 Rules)	Antecedent Rule Support	Consequent Rule Support	Support (Entire Rule Base)
Flag	2.1%	6.1%	25.4%	5.6%
Hepatitis	3.9%	4.3%	4.3%	16.1%
Audiology	1.0%	2.7%	8.1%	2.7%
Soybean-Large	2.6%	3.9%	22.8%	3.9%
Mushroom	0.3%	2.6%	32.5%	2.5%

Lift value to such heights and the Lift factor taken by itself is not indicative of the rule interestingness. Indeed, the other two rules given below appear to be more interesting, despite the fact that their Lift values are much smaller.

population:c, stalk-color-above-ring:y → stalk-color-below-ring:y, stalk-surface-above-ring:y, veil-color:y

NRIG was also able to discover subclasses of the two varieties of Mushrooms, namely the edible and poisonous species. Two such rules are given below.

stalk-color-above-ring:n → edible:p, stalk-shape:e, stalk-surface-below-ring:k

The above rule, with Lift 6.3, is interesting as it covers only 5.3% of the dataset and it applies to a subclass of poisonous mushrooms that cover only 11% of the total poisonous variety.

gill-attachment:a → cap-color:n, edible:e

The above rule (with Lift 5.9) is even more interesting as it covers only 2.3% of the dataset and applies to a subclass of edible mushrooms that cover only 4.9% of the edible variety.

6.1.2 Audiology Dataset

The Audiology dataset also produced some rare rules of interest. Two such examples are given below:

history_dizziness:t, history_fluctuating:t → class:possible_menieres

The above rule, with support 2.2% and Lift 25 identifies a histories of dizziness and fluctuating hearing levels as being strongly associated with a disorder of the inner ear that can affect hearing and balance.

class:conductive_fixation → tymp:as, ar_c:absent

This rule, having support 2.6 % and Lift 18.2 indicates that hearing disorder conductive fixation occurs in the absence of a condition coded as "tymp:as ar_c".

6.1.3 Adapative Threshold

Table 4: Results for Adaptive Threshold

Dataset	N-RIG with Adpative Threshold		
	Rare Itemset	Increment from normal N-RIG	Time (s)
Flag	393	39.6%	8.0
Hepatitis	4	-92.1%	2.39
Soybean-Large	2775	115.3%	88.95
Audiology	308	28.9%	12.7
Mushroom	5005	16.2%	351.25

Table 4 displays the effect of using an adaptive threshold with N-RIG. Out of the five datasets tested, four of the datasets produced more rare itemsets when compared to the arbitrary threshold set at 0.10 as given in Table 1. This denotes that the partition was set at a higher level than the 0.10 support value, whereas for one of the datasets the adaptive threshold value did not reach the 0.10 mark. This is due to the fact that the adaptive threshold value is dependent on the dataset that is being analysed.

7 Conclusion

In this research we have shown that the Non Redundant Itemset Generation (N-RIG) approach produces rules of practical significance that could not be discovered efficiently by the two other methods that we compared our approach with. By dispensing with an arbitrary maximum support during the candidate generation phase and replacing it with the Cumulative Productive Confidence measure we were able to generate rare rules with high frequency terms whilst keeping run time down to reasonable bounds.

References

Agrawal, R., Imielinski, T. & Swami, A. (1993), Mining association rules between sets of items in large databases, *in* P. Buneman & S. Jajodia, eds, 'Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data', pp. 207 – 216.

Bayardo, R. (1998), Efficiently mining long patterns from databases, *in* 'SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data', ACM Press, New York, NY, USA, pp. 85–93.

Cohen, E., Datar, M., Fujiwara, S., Gionis, A., Indyk, P., Ullman, J. D., Yang, C., Motwani, R. & Motwani, R. (2000), Finding interesting associations without support pruning, *in* 'ICDE '00: Proceedings of the 16th International Conference on Data Engineering', IEEE Computer Society, Washington, DC, USA, p. 489.

Koh, Y. S. & Pears, R. (2008), Rare association rule mining via transaction clustering, *in* J. F. Roddick, J. Li, P. Christen & P. J. Kennedy, eds, 'Seventh Australasian Data Mining Conference (AusDM 2008)', Vol. 87 of *CRPIT*, ACS, Glenelg, South Australia, pp. 87–94.

Koh, Y. S., Rountree, N. & O'Keefe, R. (2006), 'Finding non-coincidental sporadic rules using apriori-inverse', *International Journal of Data Warehousing and Mining* **2**(2), 38–54.

Liu, B., Hsu, W. & Ma, Y. (1999), 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.

Newman, D., Hettich, S., Blake, C. & Merz, C. (1998), 'UCI repository of machine learning databases', `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Szathmary, L., Napoli, A. & Valtchev, P. (2007), Towards rare itemset mining, *in* 'ICTAI (1)', IEEE Computer Society, pp. 305–312.

Weisstein, E. (2005), 'Fisher's exact test', MathWorld–A Wolfram Web Resource. `http://mathworld.wolfram.com/FishersExactTest.html`.

Yun, H., Ha, D., Hwang, B. & Ryu, K. H. (2003), 'Mining association rules on significant rare data using relative support', *The Journal of Systems and Software* **67**(3), 181 – 191.