Research Report

Using clustering techniques to find Association Rules for continuous data

Authors: [redacted]

Student ID: [redacted]

Contents

1 Introduction

Association rules are a useful tool for describing patterns and relationship within a body of data. They are effective in data mining since they aptly describe possible cause $\Rightarrow$ effect relationships which are prevalent in the real world, and hence enable us to discover more of such relationships, which may not be otherwise apparent, from mining data.

Association rules naturally appropriate well to boolean data attributes, and can be relatively easily extended to categorical data. However, modelling association rules from richer data attributes such as numerical or continuous attributes proves a greater challenge. Clustering seems to be a natural solution to bridging this gap, though not without its own trade-offs.

This report will propose, describe and analyse several new approaches for applying clustering to the problem of generating association rules from continuous (or numeric) data attributes. Clustering is considered as a new approach to statistical property association rules. Additional phases of clustering is considered when mining massive volumes of data.

2 Mining Statistical Property Association Rules by Clustering

Clustering may be considered for application to the problem of mining statistical property association rules.

2.1 Background

Association rules, when applied to continuous attributes, have been described in two varieties:

- *Interval* based association rules[1] which are concerned with identifying ranges/intervals of values of one attribute which when present imply particular values (or intervals) for another attribute.
  For example, "people between 16 and 18 years of age are (likely to be) enrolled in secondary education".

- *Statistical property* based association rules[2] which are concerned with identifying values (or intervals) of an attribute which identify a population subset which is statistically distinct from the broader population.
  For example, "people between 16 and 18 years of age earn on average \$18.50/hour, compared with a population average of \$37.20/hour".

The attribute which identifies a population-subset of interest can be:

- *Categorical* by specifying (or excluding) a value.
  For example, "people who own a car".

- *Quantitative* by specifying a value or interval.
  For example, "people between 16 and 18 years of age".

This proposed approach focuses on mining statistical property based association rules, where each population subset is identified by a quantitative attribute interval. The paper[2] describes these as $Quantitative \implies Quantitative$ rules:

$$\text{range of } X \implies \text{mean of } Y \tag{1}$$

The *window* algorithm for mining statistical property rules based on quantitative attributes involves first finding the population Y average, then sorting the data entries by X, and then searching for intervals of Y which have significantly different means to the

population. This search consists of scanning (along X) for the first entry with Y above mean to initialise an interval, then continuing to expand this interval as long as expansion maintains a mean Y above the population mean. Once an interval is found to end, it is accepted or rejected as a rule based on a statistical Z-test. If accepted, this interval is also recursively searched for any rules it may contain, treating the interval's data and mean as a new population. The entire process is repeated for below-mean rules. The window algorithm has complexity $O(n)$, plus $O(nlog(n))$ for sorting.

2.2 Concept

The new proposed approach consists of clustering the data entries on $[X, Y]$. From each cluster, an interval and rule may be generated. A cluster may be accepted or rejected (as before) by performing a Z-test on the cluster mean against population mean.

The initial appeal for considering this approach is performance. Clustering algorithms do not require pre-sorting of the data, which is window's largest performance consideration. Some clustering algorithms have complexity $O(nlog(n))$ which when unsorted is on par with the existing method, while others can even be $O(n)$[3]. Given that real-world data is often not pre-sorted (sorted/indexed by attributes other than those subject to analysis), and sorting before processing could impose significant storage constraints, this is a noteworthy benefit for some applications should this approach be realised feasible. Of course, the performance may turn out to be far worse with clustering, but some other possible benefits below may be considered.

From this scheme several appealing extensions can be made. For one, several clustering algorithms either intrinsically calculate the cluster's mean, or allow it to be easily tracked from one cluster to the next. Whilst this is also possible with the window algorithm, clustering algorithms may do so on both (or more) attributes at once. Hence, a cluster may be Z-tested on both (or more) attributes in a single clustering pass, allowing a rule to be formed against either attribute, of which an interval can be derived from the cluster.

Hierarchical clustering algorithms may also be considered, with the benefit that like the recursive window algorithm, they may also discover rules within rules. In this case, a candidate cluster may be Z-tested against its parent cluster. Divisive hierarchical clustering algorithms may also be used with possible benefits through performance and pruning. Pruning is of particular interest in the data mining domain, since the number of insights presented to the user must be minimised to ensure that they are accessible, rather than swamped out. Minimum thresholds may be established for support, along with other interestingness metrics, to enable the pruning of less useful statistical associations.

2.3 Analysis

The above listed benefits may be promising, however they must be carefully be weighted up. A major drawback when using clustering to mine statistical property association rules is the non-deterministic nature of many such algorithms, particularly those which provide the best performance. Because of this, the usefulness of these methods cannot be guaranteed until experimentally verified. Additionally, as much as there is the possibility of a performance benefit, many clustering algorithms perform far worse than $O(nlog(n))$ leaving them no better than the window algorithm.

The method for deriving rule intervals from clusters must also be analysed. Clusters may also not be symmetrical, so although the cluster mean is often known this does not necessarily lead to a simple derivation of the interval. However, it may suffice as an approximation[4], or optimisation may be considered to determine an interval of a desired level of accuracy.

There is also a concern that clustering algorithms may not be an appropriate fit for discovering statistically distinct subsets of the data. Many clustering algorithms impose a particular shape on the data; for example, the BIRCH *(Balanced Iterative Reducing and Clustering using Hierarchies)*[5] algorithm is ideal for identifying spherical clusters, due to its use of diameter to control cluster boundaries. This contrasts to the window algorithm which uses rectangular regions and has been shown to be highly effective, however spherical (or other) shapes imposed by an algorithm may translate well to identifying normally distributed sub-populations. Whilst these characteristics are not prohibitive, they must be thoroughly considered when assessing a clustering algorithm's suitability for identifying statistically distinct subsets. Experimental validation here would be beneficial.

Clustering algorithms may however provide a different and complimentary set of insights to that of the window algorithm. They may also be effective for datasets on which the window algorithm might not be as effective, which would enable statistical property association rules to be utilised in unconsidered problem domains.

3 Combining Pre- and Post-association Clustering for Large Datasets

Clustering may be applied both before and after association to drastically reduce the number of association rules generated.

3.1 Background

A significant challenge in the area of data mining is the need to adequately simplify and filter results down to a handful of meaningful insights. Data mining techniques may mine many candidate association rules or other forms of insights, but if these are presented directly to the user this may be quite an overwhelming experience. It is not realistic for a user to pour over endless observations in an attempt to find a few meaningful ones, and so algorithms are required to distinguish more interesting rules from less interesting ones, and aggregate related or near-equivalent observations[2]. This also inherently increases the likelihood of generating rules which are interesting, or of significance to the user. Clustering is one possible solution towards this end.

Data mining operations are often also required to mine vast quantities of data, within budgets of time constraints and computational resources. Some mining algorithms are particularly processing intensive, and even other more efficient ones will struggle at some point. It is desirable where possible to summarise, aggregate or collate the data before such operations. Clustering provides numerous avenues for doing so.

We propose that when mining overwhelmingly large datasets, clustering be considered as both a pre- and post-association operation. This is particularly relevant for overly sparse, variant or heterogeneous data sources where a significant volume of influential factors may be present, or where patterns and associations may be more hidden due to more subtle data variance.

3.2 Concept

This strategy consists of clustering the transactions prior to generating association rules, such as by *fuzzy clustering*[6]. Initial clusters are then associated by generated cluster rules, from which association rules are then generated. Rather than supplying these rules directly as a result, these are then further clustered.

At both stages there are several clustering algorithms which could be used. The algorithms chosen depend in part on whether the data being processed is purely quantitative or if the rules imply a categorical attribute value. If both are required, each would likely

need to be done in a separate KDD process. For the second stage, hierarchical clustering would bring additional benefits in providing navigable results as a mesh of related (child-parent) cluster-based rules.

A variation on the fuzzy clustering based method is to cluster the cluster rules before using them for generating association rules. However, this may be of limited effectiveness due to being applied almost directly to already clustered data.

3.3 Analysis

Due to the number of clustering algorithms available and their unique trade-offs, along with their various suitabilities for different data types, distributions and sources, choosing the appropriate clustering algorithm for both stages is both at times a challenging task and also highly domain specific.

A key concern is that the initial clustering might result in few enough association rules, so that applying clustering later may be destructive to the results. Because of this, there should be a minimum number of transactions before this strategy is feasible, which would likely require experimentation to determine. However, even very large data sets have a considerable chance of converging early into a small number of clusters, leaving very few associations for the second clustering stage. This could be addressed by allowing the user to specify a threshold of the desired number of end clusters/rules. This would have the added benefit of allowing the clustering algorithm to continue until an accessibly small number of rules remain, though this could also be detrimental due to the risk of merging unrelated rules. Another approach would be to implement a user interface that allows drill-down analysis of the generated rules, either alone or in combination with a desired top threshold. Drill-down analysis would be easily facilitated by a hierarchical clustering algorithm, and could perhaps even be mapped visually for the user allowing them to engage with what might otherwise have been a less tractable or intuitive set of insights.

4 Future Work

The mining of statistical property association rules by clustering appears to be a very new idea, and has a lot of scope for exploration. Particular avenues for future research include the appropriateness of fit of various clustering algorithms against discovering statistically distinct intervals, specifically whether they fit a cluster well to a specific interval as discovered by the window algorithm. Or likewise, whether clustering algorithms reveal different yet also meaningful intervals.

Due to the more flexible nature of clustering and the variety of algorithms available, mining statistical properties may also be considered in greater than 2 dimensions, which may be of some use.

Regarding bidirectional association detection, i.e.: Z-testing the mean on both dimensions, it is necessary to investigate whether it is meaningful. Specifically, whether:

$$A \implies \text{distinct mean } B \qquad \text{implies} \qquad B \implies \text{distinct mean } A \tag{2}$$

holds, and whether this relation is exclusive, or whether any other such relation exists.

The combination of applying clustering both pre- and post-association requires some further investigation. There may be substantial need for such an approach, however even if so there would certainly be severe degradation in result quality if the technique were misappropriated, in particular if the technique is applied to too small a data set. This also imposes restrictions on the feasibility and flexibility of testing such an approach, since real data of a larger size will be not only harder ot obtain, but will also be more difficult to manually verify.

Ultimately combining pre- and post-association clustering also needs to be explored on a variety of problem domains, due to the variety of different combinations of algorithms available, and since these algorithms will span various problem domains.

5 Conclusion

Clustering is a versatile tool for summarising, combining and correlating data sets, and has the potential to be a very useful tool for complimenting association rule based data mining. Since both of these domains are relatively diverse, there are many possible configurations for introducing clustering into the association rule discovery process. Additionally, there are many known clustering algorithms available, and doubtless more will be discovered in the near future. This gives a wide range of flexibility in both how clustering is administered, and in the types, natures and sources of quantitative data being mined.

The combination of clustering and association is versatile and can enable forms of knowledge discovery from data (KDD) which have not been previously possible. This has been the case to date, and there are doubtless more applications of this technology waiting to be discovered.

References

[1] Takeshi Fukuda, Yasuhido Morimoto, Shinichi Morishita, and Takeshi Tokuyama. Mining optimized association rules for numeric attributes. In *Proceedings of the fifteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 182–191. ACM, 1996.

[2] Yonatan Aumann and Yehuda Lindell. A statistical theory for quantitative association rules. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 261–270. ACM, 1999.

[3] Marjan Kuchaki Rafsanjani, Zahra Asghari Varzaneh, and Nasibeh Emami Chukanlo. A survey of hierarchial clustering algorithms. *The Journal of Mathematics and Computer Science*, 2012.

[4] Carlos Ordonez. A model for association rules based on clustering. In *SAC*, pages 545–546, 2005.

[5] Moses Charikar, Chandra Chekuri, Tomás Feder, and Rajeev Motwani. Incremental clustering and dynamic information retrieval. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, pages 626–635, New York, NY, USA, 1997. ACM.

[6] David M Blei. Hierarchical clustering. *Lecture Slides, February*, 2008.