Filtered-top-k Association Discovery

Geoffrey I Webb
Faculty of Information Technology, Monash University
Geoff.Webb@monash.edu

Abstract

Association mining has been one of the most intensively researched areas of data mining. However, direct uptake of the resulting technologies has been relatively low. This paper examines some of the reasons why the dominant paradigms in association mining have not lived up to their promise, and argues that a powerful alternative is provided by top-k techniques coupled with appropriate statistical and other filtering.

Introduction

Association mining is a fundamental data mining technique. Stated simply, association mining identifies items that are associated with one another in data.

Association mining has been extensively researched over the two decades since the seminal work of Agrawal et al (1). However, historically the main body of this research has concentrated on developing efficient techniques for finding frequent itemsets and has paid little attention to the questions of what types of association are useful to find and to how those types of association might be found. This paper argues that the dominant association mining paradigm, *frequent association mining*, has significant limitations and often discovers so many spurious associations that it is next to impossible to identify the potentially useful ones. It argues the case for *filtered-top-k association*

The role of association mining in data mining

Association mining complements other data mining techniques in a number of ways. First, Association mining avoids the problems associated with model selection. Most data mining techniques produce a single global model of the data. A problem with such a strategy is that there will often be many such models, all of which describe the available data equally well. The typical data mining system chooses between these models arbitrarily, without even notifying the user that these alternatives exist. However, while the system may have no reason for preferring one model over another, the user may. For example, two medical tests might be almost equally predictive in a given application. If so, the user is likely to prefer the model that uses that test that is cheaper or less invasive.

Association mining finds all local models that satisfy user-specified criteria. Thus, if two tests are both almost equally predictive, associations will be revealed that involve each test, and the user will be able to choose one that best suits the application at hand.

Second, a single model that is globally optimal may be locally suboptimal in specific regions of the problem space. By seeking local models, association mining can find models that are optimal in any given region. If there is no need for a global model, locally optimised models may be more effective.

How association mining differs from correlation analysis

Association mining is distinguished from existing statistical techniques for categorical association analysis in three respects.

One distinguishing feature is that association mining techniques scale to high-dimensional data. The standard statistical approach to categorical association analysis, log-linear analysis (2) has complexity that is exponential with respect to the number of variables. In contrast, association mining techniques can typically handle many thousands of variables.

Another distinguishing feature is that association mining concentrates on discovering relationships between values rather than variables. This is a non-trivial distinction. If someone is told that there is an association between gender and some medical condition, they are likely to immediately wish to know which gender is positively associated with the condition and which is not. Association mining goes directly to this question of interest. Further, association between values, rather than variables, can be more powerful (discover weaker relationships) when variables have more than two values. Statistical techniques may have difficulty detecting an association when there are many values for each variable and two values are strongly associated, but there are only weak interactions amongst the remaining values.

A final distinguishing feature is that association mining focuses on finding associations that are useful for the user whereas statistical techniques focus on controlling the risk of making false discoveries. In contexts where there are very large numbers of associations, it is critical to help users identify which are the most important for their immediate applications.

discovery, an approach to association mining that focuses on finding the most useful associations for a user's specific application. This approach is embodied in the *Magnum Opus* software, which by way of contrast to the relatively low uptake of frequent association mining, has enjoyed steady sales for over a decade and has been used in numerous scientific applications (a list of some published applications is maintained at http://www.giwebb.com/MOApplicationsfr.html).

Section 2 investigates some limitations of frequent association mining. Section 3 examines the filtered-top-k association discovery philosophy that underpins Magnum Opus, and explains how it addresses the core challenges of association mining. Section 4 focuses on the issue of statistical filtering, explaining why it is necessary and examining appropriate statistical filtering techniques. While this paper limits its scope to association mining, the issues it raises extend to pattern mining in general.

The Frequent Association Paradigm

The dominant approach to association mining is based on the Apriori approach (1). The key step is to identify *frequent itemsets*. These are sets of items that occur together frequently in the data. To this end, the user sets a value for *minimum support*. Sets of items that occur together at least as frequently as the specified minimum support are considered to be frequent itemsets. Association rules are generated from the identified frequent itemsets. Usually, all rules are generated whose support satisfies the minimum support constraint and that also satisfy further constraints such as minimum confidence or lift.

Terminology and Notation

A *database* is a collection of records.

A *record* is a set of items.

For tabular data, an *item* is an attribute-value pair that represents a value for a column of the database. For transactional data, an item is an object in the transaction.

An *itemset* is a set of items. Hence, each record is an itemset.

N is the number of records in a database.

A *rule* $A \rightarrow C$ comprises two parts, the *antecedent* or *left-hand-side*, A and the *consequent* or *right-hand-side*, C. A is an itemset and C is an item.

The *support* of an itemset I, sup(I), is a count of the number of records in the database of which I is a subset.

The *support* of a rule, sup($A \rightarrow C$) equals sup($A \cup \{C\}$).

The *confidence* of a rule, conf($A \rightarrow C$) equals sup($\{C\}$)/sup(A).

The *lift* of a rule, lift($A \rightarrow C$) equals sup($A \rightarrow C$)/[sup(A)×sup($\{C\}$)/N]

The *leverage* of a rule, lev($A \rightarrow C$) equals sup($A \rightarrow C$)−[sup(A)×sup($\{C\}$)/N]

There are three main reasons for the minimum support constraint. The first relates to computational efficiency. The search space is combinatorial, and quickly becomes too large to completely explore if there are large numbers of items. The minimum support constraint allows most of this search space to be ignored, leading to efficient computation.

The second reason for a minimum support constraint is to limit the number of associations that the user must consider. Just as the search space is combinatorial, so are the numbers of associations that can be generated. Unless strong constraints are placed on the associations to be presented to the user, many billions will be created. The typical user is unable to make use of such quantities of rules.

The third reason for a minimum support constraint is to seek to ensure that spurious associations are avoided. These are associations that appear to hold in the given database, but do so only by chance, and do not reflect true associations that will hold for future data.

Limitations of the Frequent Association Paradigm

The frequent association paradigm has a number of limitations.

The *vodka and caviar* problem (3) is that many high-value associations will be relatively infrequent, and hence not discovered when using minimum support. For example, in my experience with mining retail data, there will be very occasional transactions where the customer purchases large quantities of each of numerous different items. These are transactions where the shopper is, in

effect, acting on behalf of a consortium. For example, they might be buying camping equipment for the local scout group or purchasing supplies for an office party. Often these shoppers are relatively price insensitive, as they are not spending their own money. So not only do they buy in large quantities, but they often buy products at large profit margins. As a result, these transactions can be very profitable for the store, and it is desirable that data mining should reveal them. However, they are extremely infrequent, and will be overlooked using frequent association techniques.

Another problem is that minimum support provides a relatively crude mechanism for controlling the number of associations that are discovered. It is impossible to determine in advance what value for minimum support will result in a reasonable number of associations. Too low a value will result in far too many associations, and too high a value will result in too few. It typically takes considerable trial and error to find an appropriate value to use.

The frequent association paradigm cannot handle dense data. Originally developed for transaction data, it relies on each item being relatively infrequent. If many items occur in many transactions, then many combinations of them will also be frequent, and hence the number of frequent itemsets may greatly outnumber the original transactions. Instead of simplifying the problem, generating frequent itemsets as an intermediate step may actually complicate it.

Most importantly, however, minimum support may simply be irrelevant to whether an association is potentially interesting or not. If it is irrelevant then it cannot be useful to impose it as a constraint on the associations that are found. Specifically, minimum support cannot be set low enough to capture all valid rules. Nor can it be set high enough to exclude all spurious rules. To illustrate the first of these points, consider the following association rule discovered using the Magnum Opus software (4) by searching for the 100 statistically sound rules with the highest lift in the covtype dataset (5). This dataset contains 581,012 records. The rule discovered has support of only 6. It would be infeasible to discover this rule using minimum support techniques. However, using the techniques that are explained in the remainder of this paper, we can be confident that it represents a true association of potential interest.

2890<=Elevation<=3105 & cover type=6 → ST34=1 [Coverage=6; Support=6; Confidence =1.000; Lift=360.65]

To illustrate that minimum support cannot be set high enough to ensure that spurious associations are excluded, consider the highest support rule with respect to the covtype dataset (5).

ST15=0 → ST07=0 [Coverage=581,009; Support=580,904; Confidence=1.000; Lift=1.00]

This, and the next 197,183,685 rules with highest support, all turn out to represent negative associations, rather than the positive associations that they superficially appear to represent. In the case of this rule, variables ST15 and ST07 both assume the value 0 for almost all of the records in the data. Hence they occur together extremely frequently. However, all 105 occurrences of ST07=1 occur when ST15=0. Thus, to the contrary of what the rule suggests, ST07=0 is less likely when ST15=0 than otherwise.

Filtered-Top-k Association Discovery

The filtered-top-*k* association discovery paradigm provides a powerful alternative to the frequent association paradigm. Under this approach, the user specifies three parameters:

1. a measure of how potentially interesting an association is,
2. filters for discarding inappropriate associations, and
3. the number of associations to be discovered, *k*.

Any of the numerous measures of an association's worth (3, 6-17) may be used. Filters can be imposed such as a requirement that associations be non-redundant (18, 19), productive (20) or pass statistical evaluation (21). The system finds the *k* associations that optimise the specified measure within the constraints of the user-specified filters. This solves directly the problems of controlling the number of associations discovered and of focusing the results on associations that are likely to be interesting. It is often possible to derive very efficient search by using *k* together with the objective function and filters to constrain the search (16, 22-30). The result is that association mining can be performed efficiently, focusing on associations that are likely to be interesting to the user, without any need for a minimum support constraint.

This approach circumvents all of the above limitations of the frequent association paradigm. As there is no use of a minimum support constraint, there is no corresponding discontinuity in the objective function. Nor does the vodka and caviar problem occur. The objective function can capture the true value of associations without regard for whether they are frequent. Rather than using an indirect mechanism to control the number of association discovered, the user sets this parameter directly. As minimum support is not considered, the density of the data is not an issue. There is no enforced application of irrelevant constraints on the associations that may be found.

The problem of false discoveries

By its very nature, association mining is extremely susceptible to making *false discoveries*. These are associations that appear to hold in the sample data but in fact do not hold in the process that generated the data and hence will not hold in future data.

The enormous risk of false discoveries arises from the massive search that association mining entails. Consider the retail data. It contains 16,470 items. This gives rise to $2^{16,470}$ possible association rules. Admittedly, many of these are extraordinarily long, and are unlikely to be of potential interest. However, even if we restrict ourselves to association rules with no more than 4 items in the antecedent, this still gives rise to more than 10^{19} possible rules. While a $1/10^{19}$ probability event is extremely unlikely, if you look at 10^{19} separate events, it becomes likely that something as unlikely as a $1/10^{19}$ probability event will occur at least once. Hence, apparent associations that are extremely unlikely to appear by chance in a one-off investigation of a potential association are likely to appear by chance when very large numbers of potential associations are investigated.

It is not possible to use straightforward statistical analysis to control this problem. If a statistical test is applied with a standard significance level of 0.05 it means that there is only a 5% risk of making a false discovery each time it is applied. Consider what would happen if such a test was applied to

random data containing no associations. Of the 10^{19} possible rules, 5% might be expected to pass, resulting in more than 10^{17} rules being accepted. This clearly does not solve the problem.

Three techniques have been developed for controlling this risk. They are a *within search Bonferroni correction*, *holdout evaluation* and *randomisation testing that embeds the discovery process*.

Within-search Bonferroni correction

The within-search approach applies statistical tests to each rule considered during the search process (21). To overcome the multiple testing problem, a Bonferroni correction is applied. In the simplest form, this divides the desired critical value by the size of the search space. Thus, if searching the retail dataset for association rules containing up to four items in the antecedent and using a critical value of 0.05, the search space size is approximately 10^{16}, so a critical value is used of approximately $0.05/10^{16}$, which equals 5×10^{-18}.

The layered critical values approach provides a refinement to this simple method (31). This refined approach is motivated by the observation that greater numbers of significant associations tend to be found with fewer items in the antecedent. In consequence, it applies different critical values to each antecedent size, applying more relaxed critical values to associations with smaller antecedents. The critical value is divided by the number of different antecedent sizes to be investigated. Then the resulting value is divided by the size of the search space for the antecedent size to obtain the relevant critical value. For the retail data, searching for association rules with antecedents of up to four items, the resulting critical values are 9.2×10^{-11}, 5.6×10^{-15}, 1.0×10^{-18} and 2.5×10^{-18} for antecedent size 1, 2, 3 and 4, respectively[1]. With these critical values, applying Fisher exact tests for the null hypothesis that at least one item is independent of the others, 19,391 rules are found.

One advantage of the within-search approach is that any statistical test may be employed. Different null hypotheses may be best suited to different applications, and the most appropriate tests can be selected each time.

Another advantage of the within-search approach is that it supports top-k association mining techniques. Other approaches apply the statistical tests as a post process. As it is not possible to anticipate how many rules will pass post processing, it is not possible to ensure that any specific number of associations are accepted. In contrast, by applying the test within the search process, it is possible to search for the top-k associations that pass the test.

A potential limitation of the approach is that application of a computationally expensive test to each potential association that is considered can impose a high computational cost.

Holdout evaluation

The holdout evaluation approach divides the available data into two sets, the exploratory data and the holdout data (21). Associations are discovered using only the exploratory data. Statistical tests are then applied using the holdout data. It is still necessary to correct for multiple testing, but only

[1] In calculating the critical value for rules with one element in the antecedent, we count any two rule $\{x\} \rightarrow y$ and $\{y\} \rightarrow x$ as equivalent, as if one is a true association then so to must the other be. Hence the critical value for antecedent size 1 is (0.05/4)/([16470*(16470-1)]/2).

for the number of associations that were discovered, rather than for the size of the search space explored. Further, because all of the associations are tested at the same time, it is possible to apply more powerful corrections for the multiple testing problem than the Bonferroni correction, such as the Holm procedure (32).

Like the within-search approach, any statistical test may be applied with the holdout approach. Experimental results suggest that it is slightly more powerful than the within-search approach. That is, it tends to be able to find slightly more associations. However, it does not support top-k techniques, as it is not possible to anticipate how many rules will pass holdout evaluation.

Randomization testing that embeds the discovery process

Randomisation testing that embeds the discovery process (33, 34) operates by randomly shuffling the data in order to establish the null hypothesis. For example, by shuffling the columns of the data, one establishes the null hypotheses that the items are independent of one another. The association mining process is then applied to the shuffled data. This is done repeatedly and some statistic is measured each time, such as the maximum value for support, or the minimum p-value of a statistical test. These values, one for each run, are then analysed to identify a value v, the i^{th} percentile in the results, where i corresponds to the significance level that is sought. Returning to the results obtained by applying the association mining process to the unshuffled data, any association which achieves a more extreme value than v is accepted. This ensures that, if the null hypothesis holds, the risk of finding any associations is no more than the desired significance level.

This approach has a number of attractive features. First, because it looks at the distribution of results, it automatically takes account of correlations between associations and of the properties of the discovery system that may complicate the analysis required by other approaches. Second, it is conceptually straightforward to implement.

However, a limitation of this approach is that it utilises a fixed null hypothesis, that the data is drawn from a specific distribution. This is not necessarily the null hypothesis that we should want to consider in association mining. Specifically, it is credible that we will often want a null hypothesis that the data is drawn from any of a specific class of distributions.

Consider an association {pregnant} $\rightarrow$ oedema, that represents a well known association between being pregnant and suffering from oedema. We want our association mining systems to discover such associations. Now consider a random item, rand, that we might introduce to the data. Being random, it will not be associated with any other item. In consequence, it is difficult to imagine a scenario in which it would be desirable to find an association {pregnant, rand} $\rightarrow$ oedema. However, there is no one distribution that embodies the appropriate null hypothesis for assessing such associations. Hence, there is a risk that randomization techniques will 'discover' it.

Specifically, we want to test that none of the following four types of distribution holds:

- pregnant and rand are associated, but not oedema,
- pregnant and oedema are associated but not rand,
- rand and oedema are associated but not pregnant, or
- none of the three items are associated.

In other words, in many applications it will be desirable to apply statistical tests for the null hypothesis that *any one* item (or more) in an association is independent of the other items. No one distribution can establish this null hypothesis.

To illustrate why this null hypothesis is desirable, the following experiment was performed. The covtype data was divided into exploratory and holdout data sets, each containing half the data. Using the holdout evaluation method, the top 10,000 rules on lift were found that had no more than 4 items in the antecedent. These were then subjected to holdout evaluation using two null hypotheses. The first was that the antecedent and consequent were independent of one another. This is a weaker null hypothesis than the typical randomization testing null hypothesis that all items in the association are independent of one another. As a result, it is possible that it rejected more rules than would have been rejected under randomization testing. The second null hypothesis was that at least one of the items was independent of the remaining items[2]. Of the 10,000 rules found, the first test rejected 3,120. The second test rejected 9,512. In other words, 6,392 of the rules that passed the test for independence between the antecedent and the consequent failed the test for any one item being independent of the rest.

The highest lift rule that passed the second test and contained fewer than the maximum number of 4 items in the antecedent was

> {Horizontal_Distance_To_Roadways>2823, Horizontal_Distance_To_Fire_Points<1253, cover_type=7} → ST36=1
> [Coverage=302; Support=32; Confidence=0.106; Lift=521.73]

48 variants of this rule were rejected, each of which added an irrelevant item into the antecedent, such as the following.

> {Horizontal_Distance_To_Roadways>2823, 206<=Hillshade_9am<=227, Horizontal_Distance_To_Fire_Points<1253, cover_type=7} → ST36=1
> [Coverage=129; Support=11; Confidence=0.085; Lift=419.86]

In practice, if one tests only for the null hypothesis that all items are independent, or even that the antecedent and consequent are independent; the majority of association rules found can contain irrelevant items in the antecedent.

On the desirability of itemsets

Most association mining techniques find associations in the form of association rules.
A disadvantage of this representation is that a single association can result in multiple rules. If two items, *A* and *B*, are associated, two rules may be created, $\{A\} \rightarrow B$ and $\{B\} \rightarrow A$. If three items, *A*, *B* and *C* are associated with each other, up to nine rules may be created, including $\{A\} \rightarrow C$, $\{A, B\} \rightarrow C$ and $\{C\} \rightarrow A$. Four items associated with each other may result in as many as 28 distinct rules.

---

[2] Strictly speaking the first null hypothesis was $P(C \mid A) \leq P(C)$ and the second null hypothesis was $\exists x \subseteq A\ P(C \mid x) \leq P(C)$, where *A* is the antecedent and *C* the consequent of the rule.

bruises=t → ring-type=p
[Coverage=3376; Support=3184; Lift=1.93; p<4.94E-322]

ring-type=p → bruises=t
[Coverage=3968; Support=3184; Lift=1.93; p<4.94E-322]

stalk-surface-above-ring=s & ring-type=p → bruises=t
[Coverage=3664; Support=3040; Lift=2.00; p=6.32E-041]

stalk-surface-below-ring=s & ring-type=p → bruises=t
[Coverage=3472; Support=2848; Lift=1.97; p=9.66E-013]

stalk-surface-above-ring=s & stalk-surface-below-ring=s & ring-type=p → bruises=t
[Coverage=3328; Support=2776; Lift=2.01; p=0.0166]

stalk-surface-above-ring=s & stalk-surface-below-ring=s → ring-type=p
[Coverage=4156; Support=3328; Lift=1.64; p=5.89E-178]

stalk-surface-above-ring=s & stalk-surface-below-ring=s → bruises=t
[Coverage=4156; Support=2968; Lift=1.72; p=1.47E-156]

stalk-surface-above-ring=s → ring-type=p
[Coverage=5176; Support=3664; Lift=1.45; p<4.94E-322]

ring-type=p → stalk-surface-above-ring=s
[Coverage=3968; Support=3664; Lift=1.45; p<4.94E-322]

stalk-surface-below-ring=s & ring-type=p → stalk-surface-above-ring=s
[Coverage=3472; Support=3328; Lift=1.50; p=3.05E-072]

stalk-surface-above-ring=s & ring-type=p → stalk-surface-below-ring=s
[Coverage=3664; Support=3328; Lift=1.49; p=3.05E-072]

bruises=t → stalk-surface-above-ring=s
[Coverage=3376; Support=3232; Lift=1.50; p<4.94E-322]

stalk-surface-above-ring=s → bruises=t
[Coverage=5176; Support=3232; Lift=1.50; p<4.94E-322]

stalk-surface-below-ring=s → ring-type=p
[Coverage=4936; Support=3472; Lift=1.44; p<4.94E-322]

ring-type=p → stalk-surface-below-ring=s
[Coverage=3968; Support=3472; Lift=1.44; p<4.94E-322]

bruises=t & stalk-surface-below-ring=s → stalk-surface-above-ring=s
[Coverage=3040; Support=2968; Lift=1.53; p=1.56E-036]

stalk-surface-below-ring=s → stalk-surface-above-ring=s
[Coverage=4936; Support=4156; Lift=1.32; p<4.94E-322]

stalk-surface-above-ring=s → stalk-surface-below-ring=s
[Coverage=5176; Support=4156; Lift=1.32; p<4.94E-322]

bruises=t & stalk-surface-above-ring=s → stalk-surface-below-ring=s
[Coverage=3232; Support=2968; Lift=1.51; p=1.56E-036]

bruises=t → stalk-surface-below-ring=s
[Coverage=3376; Support=3040; Lift=1.48; p<4.94E-322]

stalk-surface-below-ring=s → bruises=t
[Coverage=4936; Support=3040; Lift=1.48; p<4.94E-322]

Table 1: 21 rules from the UCI mushroom dataset

Table 1 shows a set of 21 rules from the UCI mushroom dataset (35). These contrast with a single itemset that conveys the same core information much more succinctly.

{bruises=t, stalk-surface-above-ring=s, stalk-surface-below-ring=s, ring-type=p}
[Coverage=2776; Leverage=928.9; p<4.94E-322]

From the 21 rules it is extremely difficult to see that they arise from associations between 4 items. The single itemset conveys this much more concisely. It is certainly true that the 21 rules contain additional information about the precise relationships between subsets of the 4 items. However, it is likely that such detailed information will be more useful as a follow-up, once the underlying association is understood, than as a starting point for understanding the associations that underlie some set of data.

While there has been much research into efficient search for frequent itemsets (18, 36-41), most research into assessing the potential worth of associations has focused on association rules (3, 6, 10, 13, 18, 19, 42-44). I suspect that the reason for this lies in the difficulty of defining appropriate measures of interest for itemsets.

Most useful measures of interest for association rules relate to the degree to which the joint frequency of the antecedent and consequent deviate from the frequency that would be expected if they were independent. Such measures can be applied directly to itemsets comprising two items, treating one as antecedent and the other as consequent. However, once one gets beyond two items, it becomes less clear how best to measure interest. Simply measuring the extent to which the frequency of multiple items differs from the frequency that would be expected if they were independent does not provide satisfactory measures. This is because if two items are strongly associated, then the addition of any independent item to the itemset will also result in an itemset that is much more frequent than would be expected if all three items were independent.

One solution is to first evaluate the frequency that would be expected under each partition of the itemset into two disjoint subsets (45). For an itemset containing three items this results in three values. If seeking itemsets that represent positive associations, take the maximum of these values. If seeking itemsets that represent negative associations, take the minimum. Then measure the extent to which the frequency of the itemset differs from this maximum or minimum. For example, consider the itemset {bruises=t, stalk-surface-above-ring=s, stalk-surface-below-ring=s, ring-type=p}, presented above. {bruises=t, stalk-surface-above-ring=s, ring-type=p} occurs in 37.42% of records and stalk-surface-below-ring=s occurs in 60.76%. If they were independent they would be expected to occur in 22.736% of the 8124 cases, a total of 1847.1 times. This is the maximum expected frequency out of any partition of the four items. The frequency with which the itemset occurs is 2776 times. Subtracting the maximum expected frequency, 1847.1 from the actual frequency, 2276, one obtains the *leverage* of 928.9.

Statistical filtering can be applied in the same manner, assessing the probability that the observed frequency would occur if the elements of each binary partition of an itemset were independent of one another. The itemset is rejected if it fails this test of independence for *any* partition (45). In a similar vein, one can assess the significance of deviation between the full contingency table of an itemset and a maximum entropy model that takes account of interactions between subsets of the full itemset (46).

Searching for top-k itemsets in this manner has several desirable attributes relative to searching for association rules. First, as already discussed, itemsets can highlight the core underlying associations more readily than association rules. Second, the search space is much smaller, and hence the statistical corrections that must be applied are much smaller. This means that the resulting statistical tests are more powerful – substantially more associations can be revealed.

The above discussion assumes that we are searching for interesting itemsets without prior knowledge about the domain. If one has background knowledge, one might alternatively want to seek to identify itemsets whose frequencies differ significantly from those expected given the background knowledge (15, 47).

Conclusion

Association mining is a fundamental data mining task. It involves discovering local models of associations between items within data. It is distinguished from statistical correlation analysis by scalability to high-dimensional data, identifying interactions between values rather than variables, and by a concern for identifying useful associations rather than simply controlling the risk of making false discoveries.

Association mining complements other data mining techniques by finding all local models rather than a single global model. This empowers the user to select between alternative models on grounds that may be difficult to quantify and hence have a computational system take into account.

However, despite these desirable features, and despite association mining proving useful in other applications such as classification (48), clustering (49) and feature discovery (50); direct application of association mining has not enjoyed uptake commensurate to the research that has been invested into it. I argue that this is because much of the research has focused on how to find frequent itemsets efficiently, rather than focusing on what associations are useful to find. I argue that the dominant paradigm for association mining, frequent association mining, has serious limitations. An alternative, filtered-top-k association discovery, can circumvent many of these.

A further issue is that inherent in any attempt to identify associations is an extreme risk of false discoveries. These are apparent associations that are in fact only artefacts of the specific sample of data that has been collected. While randomization techniques have gained some popularity as an approach to guarding against false discoveries, they are limited in the null hypotheses that they support. It is usually important to guard against any item in an association being independent of the remaining items, but randomization tests cannot do this as no one distribution can establish a suitable null hypothesis. Application of appropriate statistical filters seems desirable in most association mining applications.

A final point that I raise is that it is often desirable to identify associations in the form of itemsets rather than rules. I promote the use of techniques that assess the potential value of an itemset by assessing the degree to which its frequency differs from the maximum (or in the case of negative associations, minimum) expected under any assumption of independence between binary partitions of the itemset.

The Magnum Opus association mining system, which embodies the filtered-top-k association discovery approach that I espouse, has enjoyed commercial success for more than a decade and has been used in numerous scientific applications. This success lends support to my arguments.

Acknowledgements

I am grateful to Wilhelmiina Hamalainen, Nikolaj Tatti, Jilles Vreeken, Mohammed Zaki and the anonymous reviewers for insightful comments and suggestions during the preparation of this paper.

References

1. Agrawal R, Srikant R: Fast algorithms for mining association rules. In: *Proceedings of the 20th International Conference on Very Large Databases VLDB '94,* p. 487-499 Eds. Jorge B. Bocca MJ, Carlo Z). Morgan Kaufmann: Santiago, Chile, 1994.
2. Agresti A: Categorical Data Analysis. p. Wiley-Interscience: New York, 2002.
3. Cohen E, Datar M, Fujiwara S, Gionis A, Indyk P*, et al.*: Finding interesting associations without support pruning. In: *Knowledge and Data Engineering,* p. 64-78 2001.
4. Webb GI: Magnum Opus. p. GI Webb & Associates: Melbourne, 2010.
5. Blackard JA, Dean DJ: Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*. 1999;**24**(3):131-151.
6. Piatetsky-Shapiro G: Discovery, analysis, and presentation of strong rules. In: *Knowledge Discovery in Databases,* p. 229-248 Eds. Piatetsky-Shapiro G, Frawley J). AAAI/MIT Press: Menlo Park, CA., 1991.
7. Silberschatz A, Tuzhilin A: What makes patterns interesting in knowledge discovery systems. In: *IEEE Transactions on Knowledge and Data Engineering,* p. 970-974 1996.
8. Bayardo JRJ, Agrawal R: Mining the most interesting rules. In: *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99),* p. 145-154 1999.
9. Freitas AA: On rule interestingness measures. In: *Knowledge-Based Systems,* p. 309-315 1999.
10. Klemettinen M, Mannila H, Ronkainen P, Toivonen H, Verkamo A: Finding interesting rules from large sets of discovered association rules. In: *Proceedings of the Third International Conference on Information and Knowledge Management,* p. 401-407 1999.
11. Liu B, Hsu W, Chen S, Ma Y: Analyzing the subjective interestingness of association rules. In: *IEEE Intelligent Systems,* p. 47-55 2000.
12. Sese J, Morishita S: Answering the Most Correlated N Association Rules Efficiently. In: *Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02),* p. 410-422 2002.
13. Tan P-N, Kumar V, Srivastava J: Selecting the right interestingness measure for association patterns. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002),* p. 491–502: Edmonton, Alberta, Canada, 2002.
14. Calders T, Goethals B: Non-Derivable itemset mining. In: *Data Mining and Knowledge Discovery,* p. 171-206 2007.
15. Jaroszewicz S, Scheffer T, Simovici D: Scalable pattern mining with Bayesian networks as background knowledge. *Data Mining and Knowledge Discovery*. 2009;**18**(1):56-100.
16. Hämäläinen W: Efficient discovery of the top-k optimal dependency rules with the Fisher's exact test of significance. In: *Proceedings of the Tenth IEEE International Conference on Data Mining,* p. 196-205 Eds. Webb GI, Liu B, Zhang C, Gunopulos D, Wu X)2010.
17. Wu T, Chen Y, Han J: Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*. 2010;**21**(3):371-397.
18. Bastide Y, Pasquier N, Taouil R, Stumme G, Lakhal L: Mining minimal non-redundant association rules using frequent closed itemsets. In: *First International Conference on Computational Logic - CL 2000,* p. 972-986. Springer-Verlag: Berlin, 2000.
19. Zaki MJ: Mining non-redundant association rules. In: *Data Mining and Knowledge Discovery,* p. 223-248 2004.
20. Webb GI: Discovering significant rules. In: *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD-2006,* p. 434 - 443 Eds. Ungar L, Craven M, Gunopulos D, Eliassi-Rad T). The Association for Computing Machinery2006.
21. Webb GI: Discovering significant patterns. In: *Machine Learning,* p. 1-33 2007.
22. Webb GI: OPUS: An efficient admissible algorithm for unordered search. In: *Journal of Artificial Intelligence Research,* p. 431-465 1995.

23. Webb GI: Efficient search for association rules. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000),* p. 99-107. The Association for Computing Machinery: Boston, MA, 2000.
24. Webb GI, Zhang S: Further pruning for efficient association rule discovery. In: *Proceedings of the 14th Australian Joint Conference on Artificial Intelligence,* p. 605-618. Springer: Berlin, 2001.
25. Han J, Wang J, Lu Y, Tzvetkov P: Mining top-k frequent closed patterns without minimum support. In: *International Conference on Data Mining,* p. 211-218 2002.
26. Webb GI, Zhang S: Removing trivial associations in association rule discovery. In: *Proceedings of the First International NAISO Congress on Autonomous Intelligent Systems,* p. NAISO Academic Press: Geelong, 2002.
27. Webb GI, Zhang S: K-Optimal rule discovery. In: *Data Mining and Knowledge Discovery,* p. 39-79 2005.
28. Wang J, Han J, Lu Y, Tzvetkov P: TFP: An efficient algorithm for mining top-k frequent closed itemsets. In: *IEEE Transactions on Knowledge and Data Engineering,* p. 652-664 2005.
29. Fu AWC, Renfrew WK, Tang J: Mining n-most interesting itemsets. In: *Proceedings of the 12th International Symposium on Foundations of Intelligent Systems,* p. 59-67 2000.
30. Pietracaprina A, Riondato M, Upfal E, Vandin F: Mining top-k frequent itemsets through progressive sampling. *Data Mining and Knowledge Discovery*. 2010;**21**(2):310-326.
31. Webb GI: Layered critical values: A powerful direct-adjustment approach to discovering significant patterns. In: *Machine Learning,* p. 307-323 2008.
32. Holm S: A simple sequentially rejective multiple test procedure. In: *Scandinavian Journal of Statistics,* p. 65-70 1979.
33. Megiddo N, Srikant R: Discovering predictive association rules. In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98),* p. 27-78. AAAI Press: Menlo Park, US, 1998.
34. Gionis A, Mannila H, Mielikinen T, Tsaparas P: Assessing data mining results via swap randomization. In: *ACM Transactions on Knowledge Discovery from Data (TKDD),* p. 2007.
35. Asuncion A, Newman DJ: UCI machine learning repository. p. University of California, Irvine, School of Information and Computer Sciences2010.
36. Han J, Pei J, Yin Y: Mining frequent patterns without candidate generation. In: *Proceedings 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD'00),* p. 1-12: Dallas, TX, 2000.
37. Pei J, Han J, Mao R: CLOSET: An efficient algorithm for mining frequent closed itemsets. In: *Proceedings 2000 ACM-SIGMOD International Workshop on Data Mining and Knowledge Discovery (DMKD'00),* p. 21-30: Dallas, TX, 2000.
38. Gouda K, Zaki MJ: Efficiently mining maximal frequent itemsets. In: *First IEEE International Conference on Data Mining,* p. 163-170: San Jose, CA., 2001.
39. Pasquier N, Bastide Y, Taouil R, Lakhal L: Discovering frequent closed itemsets for association rules. In: *Proceedings of the Seventh International Conference on Database Theory (ICDT'99),* p. 398-416: Jerusalem, Israel, 1999.
40. Zaki MJ, Hsiao CJ: CHARM: An efficient algorithm for closed itemset mining. In: *Proceedings of the Second SIAM International Conference on Data Mining,* p. 457-473 2002.
41. Savasere A, Omiecinski E, Navathe S: An efficient algorithm for mining association rules in large databases. In: *Proceedings of the 21st International Conference on Very Large Data Bases,* p. 432-444. Morgan Kaufmann: San Francisco, 1995.
42. Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI: Fast discovery of association rules. In: *Advances in Knowledge Discovery and Data Mining,* p. 307-328 Eds. Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R). AAAI Press: Menlo Park, CA., 1996.
43. Goethals B, Muhonen J, Toivonen H: Mining non-derivable association rules. In: *Proceedings of the Fifth SIAM International Conference on Data Mining SDM-05,* p. 239-249 2005.

44. Liu B, Hsu W, Ma Y: Identifying non-actionable association rules. In: *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001),* p. 329-334: San Francisco, CA, 2001.
45. Webb GI: Self-sufficient itemsets: An approach to screening potentially interesting associations between items. In: *Transactions on Knowledge Discovery from Data,* p. 3:1-3:20 2010.
46. Tatti N: Maximum entropy based significance of itemsets. *Knowledge and Information Systems*. 2008;**17**(1):57--77.
47. Tatti N, Mampaey M: Using background knowledge to rank itemsets. *Data Mining and Knowledge Discovery*. 2010;**21**(2):293-309.
48. Liu B, Hsu W, Ma Y: Integrating classification and association rule mining. In: *Proceedings of the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-98),* p. 80-86. AAAI: New York, 1998.
49. Guha S, Rastogi R, Shim K: Rock: A robust clustering algorithm for categorical attributes. *Information Systems*. 2000;**25**(5):345-366.
50. Flach P, Lavrac N: The role of feature construction in inductive rule learning. In: *Proceedings of the ICML2000 workshop on Attribute-Value and Relational Learning: crossing the boundaries,* p. 1–11 Eds. Raedt LD, Kramer S)2000.