FINDING ANOMALIES IN DATABASES

Mrs.K.Poongothai
Lecturer,
Dept. of Applied Science,
Selvam College of Technology

Mrs.M.Parimala
Lecturer,
Dept. of MCA,
M.Kumarasamy College of Engineering

Dr.S.Sathiyabama
Professor, Dept. of MCA,
K.S. Rangasamy College of Technology
Tiruchengode-637 215

Abstract - Association rules have become an important paradigm in knowledge discovery. Nevertheless, the huge number of rules which are usually obtained from standard datasets limits their applicability. In order to solve this problem, several solutions have been proposed, as the definition of subjective measures of interest for the rules or the use of more restrictive accuracy measures. Other approaches try to obtain different kinds of knowledge, referred to as peculiarities, infrequent rules, or exceptions. In general, the latter approaches are able to reduce the number of rules derived from the input dataset. This paper is focused on this topic. We introduce a new kind of rules, namely, anomalous rules, which can be viewed as association rules hidden by a dominant rule. We also develop an efficient algorithm to find all the anomalous rules existing in a database.

INTRODUCTION

Association rules have proved to be a practical tool in order to find tendencies in databases, and they have been extensively applied in areas such as market basket analysis and CRM (Customer Relationship Management). These practical applications have been made possible by the development of efficient algorithms to discover all the association rules in a database [11, 12, 4], as well as specialized parallel algorithms [1]. Related research on sequential patterns [2], associations varying over time [15], and associative classification models [5] have fostered the adoption of association rules in a wide range of data mining tasks.

Despite their proven applicability, association rules have serious drawbacks limiting their effective use. The main disadvantage stems from the large number of rules obtained even from small-sized databases, which may result in a second- order data mining problem. The existence of a large number of association rules makes them unmanageable for any human user, since she is overwhelmed with such a huge set of potentially useful relations. This disadvantage is a direct consequence of the type of knowledge the association rules try to extract, i.e., frequent and confident rules. Although it may be of interest in some application domains, where the expert tries to find unobserved frequent patters, it is not when we would like to extract hidden patterns. It has been noted that, in fact, the occurrence of a frequent event carries less information than the occurrence of a rare or hidden event. Therefore, it is often more interesting to find surprising non-frequent events than frequent ones [7, 12, 15]. In some sense, as mentioned in [7], the main cause behind the popularity of classical association rules is the possibility of building efficient algorithms to find all the rules which are present in a given database.

The crucial problem, then, is to determine which kind of events we are interested in, so that we can appropriately characterize them. Before we delve into the details, it should be stressed that the kinds of events we could be interested in are application-dependent. In other words, it depends on the type of knowledge we are looking for. For instance, we could be interested in finding infrequent rules for intrusion detection in computer systems, exceptions to classical associations for the detection of conflicting medicine therapies, or unusual short sequences of nucleotides in genome sequencing.

Our objective in this paper is to introduce a new kind of rule describing a type of knowledge we might me interested in, what we will call anomalous association rules henceforth. Anomalous association rules are confident rules representing homogeneous deviations from common behavior. This common behavior can be modeled by standard association rules and, therefore, it can be said that anomalous association rules are hidden by a dominant association rule. In the following section, we review some related work. We shall justify the need to define the concept of anomalous rule as something complementary to the study of exception rules. Section 3 contains the formal definition of anomalous association rules. Section 4 presents an efficient algorithm to detect this kind of rules. Finally, Section 5 discusses some experimental results.

MOTIVATION AND RELATED WORK

Several proposals have appeared in the data mining literature that try to reduce the number of associations obtained in a mining process, just to make them manageable by an expert. According to the terminology used in [6], we can distinguish between user-driven and data-driven approaches (also referred to as subjective and objective interestingness measures, respectively [15], although we prefer the first terminology).Let us remark that, once we have obtained the set of good rules (considered as such by any interestingness measure), we can apply filtering techniques such as eliminating redundant tuples [15] or evaluating the rules according to other interestingness measures in order to check (at least, in some extent) their degree of surprising ness, i.e., if the rules convey new and useful information which could be viewed as unexpected [8, 9, 11, 6]. Some proposals [13, 15] even intro- duce alternative interestingness measures which are strongly related to the kind of knowledge they try to extract. In user-driven approaches, an expert must intervene in some way: by stat- ing some restriction about the potential attributes which may appear in a reallocation [12], by imposing a hierarchical taxonomy [10], by indicating potential useful rules according to some prior knowledge [15], or just by eliminating non-interesting rules in a first step so that other rules can automatically removed in subsequent steps [18].On the other hand, data-driven approaches do not require the intervention of a human expert. They try to autonomously obtain more restrictive rules. This is mainly accomplished by two approaches:

a) Using interestingness measures differing from the usual support-confidence pair [14, 12].
b) Looking for other kinds of knowledge which are not even considered by classical association rule mining algorithms.

The latter approach pursues the objective of finding surprising rules in the sense that an informative rule has not necessary to be a frequent one. The work we present here is in line with this second data-driven approach. We shall introduce a new kind of association rules that we will call anomalous rules. Before we briefly review existing proposals in order to put our approach in context, we will describe the notation we will use henceforth. From now on, X,Y, Z, and A shall denote arbitrary item sets. The support and confidence of an association rule X) Y are defined as usual and they will be represented by supp(X) Y) and conf(X) Y), respectively. The usual minimum support and confident thresholds are denoted by M inSupp and M inc onf, respectively. A frequent rule is a rule with high support (greater than or equal to the sup- port threshold M inSupp), while a confident rule is a rule with high confidence (greater than or equal to the confidence threshold M inC onf). A strong rule is a classical association rule, i.e., a frequent and confident one [7] try to find non-frequent but highly correlated item sets, whereas [12] aims to obtain peculiarities defined as non-frequent but highly confident rules according to a nearness measure defined over each attribute, i.e., a peculiarity must be significantly far away from the rest of individuals. [9] finds unusual sequences, in the sense that items with low probability of occurrence are not expected to be together in several sequences. If so, a surprising sequence has been found. Another interesting approach [13, 10, 3] consists of looking for exceptions, in the sense that the presence of an attribute interacting with another may change the consequent in a strong association rule. The general form of an exception rule is introduced in [13, 15] as follows:

X) Y X Z): Y X 6) Z

Here, X) Y is a common sense rule (a strong rule). X Z): Y is the exception, where: Y could be a concrete value E (the Exception [12]). Finally, X 6) Z is a reference rule. It should be noted that we have simplified the definition of exceptions since the authors use five [13] or more [15] parameters which have to be settled beforehand, which could be viewed as a shortcoming of their discovery techniques. In general terms, the kind of knowledge these exceptions try to capture can be interpreted as follows:

X strongly implies Y (and not Z).
But, in conjunction with Z, X does not imply Y
(Maybe it implies another E)

For example [14], if X represents antibiotics, Y recovery, Z staphylococci, and E death, then the following rule might be discovered: with the help of antibiotics, the patient usually tends to recover, unless staphylococci appear; in such a case, antibiotics combined with staphylococci may lead to death. This is a very interesting kind of knowledge which cannot be detected by traditional association rules because the exceptions are hidden by a dominant rule. However, there are other exceptional associations which cannot be detected by applying the approach described above. For instance, in scientific experimental, it is usual to have two groups of individuals: one of them is given a placebo and the other one is treated with some real medicine. The scientist wants to discover if there are significant differences in both populations, perhaps with re spect to a variable Y. In those cases, where the change is significant, an ANOVA or contingency analysis is enough. Unfortunately, this is not always the case. What the scientist obtains is that both populations exhibit a similar behavior except in some rare cases. These infrequent events are the interesting ones for the scientist because they indicate that something happened to those individuals and the study must continue in order to determine the possible causes of this unusual change of behavior. In the ideal case, the scientist has recorded the values of a set of variables Z for both populations and, by performing an exception rule analysis, he could conclude that the interaction between two item sets X and Z (where Z is the item set corresponding to the values of Z) change the common behavior when X is present (and Z is not). However, the scientist does not always keep records of all the relevant variables for the experiment. He might not even be aware of which variables are really relevant. Therefore, in general, we cannot not derive any conclusion about the potential changes the medicine causes. In this case, the use of an alternative discovery mechanism is necessary. In the next section, we present such an alternative which might help our scientist to discover behavioral changes caused by the medicine he is testing

Defining Anomalous Association Rules

An anomalous association rule is an association rule that comes to the surface when we eliminate the dominant effect produced by a strong rule. In other words, it is an association rule that is verified when a common rule fails. In this paper, we will assume that rules are derived from item sets containing discrete values. Formally, we can give the following definition to anomalous association rules:

Definition 1. Let X, Y, and A be arbitrary item sets. We say that X A is an anomalous rule with respect to X) Y , where A denotes the Anomaly, if the following conditions hold:

a) X) Y is a strong rule (frequent and confident)
b) X :Y) A is a confident rule c) X Y): A is a confident rule

It should be noted that, implicitly in the definition, we have used the common minimum support (M inSupp) and confidence (M inC onf) thresholds, since they tell us which rules are frequent and confident, respectively. For the sake of simplicity, we have not explicitly mentioned them in the definition. A minimum support threshold is relevant to condition a), while the same minimum confidence threshold is used in conditions a), b), and c). The semantics this kind of rules tries to capture is the following:

X strongly implies Y, but in those cases where we do not obtain Y, then X confidently implies A

In other words: When X, then we have either Y (usually) or A (unusually) Therefore, anomalous association rules represent homogeneous deviations from the usual behavior. For instance, we could be interested in situations where a common rule holds:

If symptoms-X then disease-Y

Where the rule does not hold, we might discover an interesting anomaly:

if symptoms-X then disease-A

When not disease-Y

If we compare our definition with Hussain and Suzuki's [13, 12], we can see that they correspond to different semantics. Attending to our formal definition, our approximation does not require the existence of the conflictive itemset (what we called Z when describing Hussain and Suzuki's approach in the previous section). Furthermore, we impose that the majority of exceptions must correspond to the same consequent A in order to be considered an anomaly. In order to illustrate these differences, let us consider the relation shown in Figure 1, where we have selected those records containing X. From this dataset, we obtain conf(X)Y)=0:6, conf(X Z):Y)=conf(XZ)A)=1, and conf(X)Z)=0:2. If we suppose that the item set XY satisfies the support threshold and we use 0:6 as confidence threshold, then \X Z)A is an exception to X)Y , with reference rule X):Z". This exception is not highlighted as an anomaly using our approach because A is not always present when X: Y. In fact, conf(X:Y)A) is only 0:5, which is below the minimum confidence threshold 0:6. On the other hand, let us consider the relation in Figure 2, which shows two examples where an anomaly is not an exception. In the second example, we find that conf(X)Y) = 0:8, conf(X Y) :A) = 0:75, and conf(X :Y) A) = 1. No Z -value exists to originate an exception, but X

X	Y	A4	Z3	· · ·
X	Y	A1	Z1	· · ·
X	Y	A2	Z2	· · ·
X	Y	A1	Z3	· · ·
X	Y	A2	Z1	· · ·
X	Y	A3	Z2	· · ·
X	Y1	A4	Z3	· · ·
X	Y2	A4	Z1	· · ·
X	Y3	A	Z	· · ·
X	Y4	A	Z	· · ·
· · ·				

Fig.1. A is an exception to X Y when Z, but that anomaly is not confident enough to be considered an anomalous rule

The table in Figure 1 also shows that when the number of variables (at-tributes in a relational database) is high, then the chance of finding spurious Z item sets correlated with :Y notably increases. As a consequence, the number of rules obtained can be really high (see [15, 13] for empirical results). The semantics we have attributed to our anomalies is more restrictive than exceptions and, thus, when the expert is interested in this kind of knowledge, then he will obtain a more manageable number of rules to explore. Moreover, we do not require the existence of a Z explaining the exception. In particular, we have observed that users are usually interested in anomalies involving one item in their consequent. A more rational explanation of this fact might have psychological roots: As humans, we tend to find more problems when reasoning about negated facts. Since the anomaly introduces a negation in the

X	Y	Z1	· · ·
X	Y	Z2	· · ·
X	Y	Z	· · ·
X	Y	Z	· · ·
X	Y	Z	· · ·
X	Y	Z	· · ·

```
X A Z · · ·
X A Z · · ·
X A Z · · ·
X A Z · · ·
· · ·
X Y  A1 Z1 · · ·
X Y  A1 Z2 · · ·
X Y  A2 Z3 · · ·
 X Y  A2 Z1 · · ·
 X Y  A3 Z2 · · ·
X Y  A3 Z3 · · ·
X Y  A  Z  · · ·
X Y  A  Z  · · ·
X Y3 A  Z  · · ·
X Y4 A  Z  · · ·
· · ·
```

Fig.2. X→ A is detected as an anomalous rule, even when no exception can be found through the Z-values

In the Figure 2 A is detected as an anomalous rule, even when no exception can be found through the Z -values. rule antecedent, experts tend to look for `simple' understandable anomalies in order to detect unexpected facts. For instance, an expert physician might directly look for the anomalies related to common symptoms when these symptoms are not caused by the most probable cause (that is, the usual disease she would diagnose). The following section explores the implementation details associated to the discovery of such kind of anomalous association rules. Remark. It should be noted that, the more confident the rules X:Y) A and X Y):A are, the stronger the anomaly is. This fact could be useful in order to define a degree of strength associated to the anomaly.

DISCOVERING ANOMALOUS ASSOCIATION RULES

Given a database, mining conventional association rules consists of generating all the association rules whose support and confidence are greater than some user-specified minimum thresholds. We will use the traditional decomposition of the association rule mining process to obtain all the anomalous association rules existing in the database:

{Finding all the relevant item sets.
{Generating the association rules derived from the previously-obtained item- sets.

For instance, Apriori-based algorithms are iterative [8]. Each iteration consists of two phases. The first phase, candidate generation, generates potentially frequent k-item sets (Ck) from the previously obtained frequent (k-1)-item set (Lk−1). The second phase, support counting, scans the database to find the actual frequent k-item sets (Lk). Apriori-based algorithms are based on the fact that that all subsets of a frequent item set is also frequent. This allows for the generation of a reduced set of candidate itemsets. Nevertheless, it should be noted that the there is no actual need to build a candidate set of potentially frequent item sets [11]. In the case of anomalous association rules, when we say that X A is an anomalous rule with respect to X) Y , that means that the item set X[:Y[A appears often when the rule X) Y does not hold. Since it represents an anomaly, by definition, we cannot establish any minimum support threshold for X[:Y[A. In fact, an anomaly is not usually frequent in the whole database. Therefore, standard association rule mining algorithms cannot be used to detect anomalies without modification. Given an anomalous association rule X A with respect to X) Y, let us denote by R the subset of the database that, containing X, does not verify the association rule X)Y. In other words, R will be the part of the database that does not verify the rule and might host an anomaly. When we write suppR (X), it actually represents supp(X[:Y) in the complete database. Although this value is not usually computed when obtaining the item sets, it can be easily computed as supp(X) supp(X[Y). Both values in this expression are always available after the conventional association rule mining process, since both X and X [Y are frequent item sets. Applying the same reasoning, the following expression can be derived to represent the confidence of the anomaly X A with respect to X) Y :

ConfR (X A) = supp(X [A) − supp(X [Y [A)
Supp(X) − supp(X [Y)

Fortunately, when somebody is looking for anomalies, he is usually interested in anomalies involving individual items. We can exploit this fact by taking into account that, even when X [A and X [Y [A might not be frequent, they are extensions of the frequent item sets X and X [Y, respectively. Since A will represent individual items, our problem reduces to being able to compute the support of L [i, for each frequent item set L and item i potentially involved in an anomaly. Therefore, we can modify existing iterative association rule mining algorithms to efficiently obtain all the anomalies in the database by modifying the support counting phase to compute the support for frequent item set extensions:

{Candidate generation : As in any Apriori-based algorithm, we generate potentially frequent k-item sets from the frequent item sets of size k-1.
{Database scan: The database is read to collect the information needed to compute the rule confidence for potential anomalies. This phase involves two parallel tasks Candidate support counting: The frequency of each candidate k-item set is obtained by scanning the database in order to obtain the actual frequent k-item sets. Extension support counting: At the same time that candidate

support is computed, the frequency of each frequent k-1-itemset extension can also be obtained. Once we obtain the last set of frequent item sets, an additional database scan can be used to compute the support for the extensions of the larger frequent item sets. Using a variation of a standard association rule mining algorithm as TBAR [4], nicknamed ATBAR (Anomaly TBAR), we can efficiently compute the support for each frequent item set as well as the support for its extensions. In order to discover existing anomalies, a tree data structure is built to store all the support values needed to check potential anomalies. This tree is an extended version of the typical item set tree used by algorithms like TBAR [3]. The extended item set tree stores the support for frequent item set extensions as well as for all the frequent item sets themselves. Once we have these values, all anomalous association rules can be obtained by the proper traversal of this tree-shaped data structure. In interactive applications, the human user can also use the aforementioned extended item set tree as an index to explore a database in the quest for anomalies.

CONCLUSION AND FUTURE WORK

In this paper, we have studied situations where standard association rules do not provide the information the user seeks. Anomalous association rules have proved helpful in order to represent the kind of knowledge the user might be looking for when analyzing deviations from normal behavior. The normal behavior is modeled by conventional association rules, and the anomalous association rules are association rules which hold when the conventional rules fail. We have also developed an efficient algorithm to mine anomalies from databases. Our algorithm, ATBAR, is suitable for the discovery of anomalies in large databases. We intend to apply our technique to real problems involving datasets from the biomedical domain. Our approach could also prove useful in tasks such as fraud identification, intrusion detection systems and, in general, any application where the user is not really interested in the most common patterns, but in those patterns which differ from the norm.

REFERENCES

[1] R.Agrawal and J.Shafer. Parallel mining of association rules. IEEE Transactions on Knowledge and Data Engineering, 8(6):962{969, 1996.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Mining sequential patterns. In Philip S. Yu and Arbee S.P.Chen, editors, Eleventh International Conference on Data Engineering, pages 3{14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

[3] Yonatan Aumann and Yehuda Lindellt. A statistical theory for quantitative association rules. Journal of Intelligent Information Systems, 20(3):255{283, 2003.

[4] F. Berzal, J.C. Cubero, J.M. Marn, and J.M. Serrano. An efficient method for association rule mining in relational databases. Data and Knowledge Engineering, 37:47{84, 2001.

[5] F. Berzal, J.C. Cubero, Daniel Snchez, and J.M. Serrano. Art: A hybrid classifi- cation model. Machine Learning, 54(1):67{92, 2004.

[6] DR Carvalho, AA Freitas, and NFF Ebecken. A critical review of rule surprising-ness measures. In NFF Ebecken, CA Brebbia, and A Zanasi, editors, Proc. Data Mining IV Int. Conf. On Data Mining, pages 545{556. WIT Press, December 2003.

[7] Edith Cohen, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Jeffrey D. Ullman, and Cheng Yang. Finding interesting associations without support pruning. IEEE Transactions on Knowledge and Data Engineering, 13(1):64{78, 2001.

[8] AA Freitas. On Rule Interestingness Measures. Knowledge-Based Systems, 12(5-6):309{315, October 1999.

[9] Alex Alves Freitas. On objective measures of rule surprisingness. In Principles of Data Mining and Knowledge Discovery, pages 1{9, 1998.

[10] J.Hanand Y.Fu.Discovery of multiple-level association rules from large databases. In Proceedings of the VLDB Conference, pages 420{431, 1995.

[11] Jiawei Han, J. Pei, and Y.Yin. Mining frequent patterns without candidate generation. In Proceedings of the 1998 ACM SIGMOD international conference on Management of Data, pages 1{12, 2000.

[12] C. Hidber. Online association rule mining. In Proceedings of the 1999 ACM SIG-MOD international conference on Management of Data, pages 145{156, 1999.

[13] Farhad Hussain, Huan Liu, Einoshin Suzuki, and Hongjun Lu. Exception rule mining with a relative interestingness measure. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 86{97, 2000.

[14] Yves Kodratoff. Comparing machine learning and knowledge discovery in DataBases : An application to knowledge discovery in texts. In Machine Learn- ing and its Applications, volume 2049, pages 1{21. Lecture Notes in Computer Science, 2001.

[15] Bing Liu, Wynne Hsu, Shu Chen, and Yiming Ma. Analyzing the subjective interestingness of association rules. IEEE Intelligent Systems, pages 47{55, 2000.