Finding anomalies in databases

Fernando Berzal*, Juan-Carlos Cubero*, Nicolás Marín*, Matías Gámez**

* Department of Computer Science and AI. University of Granada.
Granada 18071 Spain
{fberzal|jc.cubero|nicm}@decsai.ugr.es
** Department of Business and Economics. University of Castilla - La Mancha
Albacete 02071 Spain
Matias.Gamez@uclm.es

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.

1 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[17], 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, 27, 25]. 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.

2 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 [21], 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 [19] or evaluating the rules according to other interestingness measures in order to check (at least, in some extent) their degree of surprisingness, i.e, if the rules convey new and useful information which could be viewed as unexpected [8, 9, 21, 6]. Some proposals [13, 25] even introduce 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 stating some restriction about the potential attributes which may appear in a relation [22], 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, 26].
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 itemsets. The support and confidence of an association rule $X \Rightarrow Y$ are defined as usual and they will be represented by $\text{supp}(X \Rightarrow Y)$ and $\text{conf}(X \Rightarrow Y)$, respectively. The usual minimum support and confident thresholds are denoted by $MinSupp$ and $MinConf$, respectively. A frequent rule is a rule with high support (greater than or equal to the support threshold $MinSupp$), while a confident rule is a rule with high confidence (greater than or equal to the confidence threshold $MinConf$). A *strong rule* is a classical association rule, i.e, a frequent and confident one.

[7, 20] try to find non-frequent but highly correlated itemsets, whereas [28] 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. [27] 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, 25, 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, 25] as follows:

$$\begin{gathered} X \Rightarrow Y \\ XZ \Rightarrow \neg Y \\ X \not\Rightarrow Z \end{gathered}$$

Here, $X \Rightarrow Y$ is a *common sense* rule (a strong rule). $XZ \Rightarrow \neg Y$ is the *exception*, where $\neg Y$ could be a concrete value E (the $\underline{E}$xception [25]). Finally, $X \not\Rightarrow Z$ is a *reference* rule. It should be noted that we have simplified the definition of exceptions since the authors use five [13] or more [25] 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 [24], 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`.

These exception rules indicate that there is some kind of interaction between two factors, X and Z, so that the presence of Z alters the usual behavior (Y) the population have when X is present.

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 experimentation, 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 respect to a variable $\mathbb{Y}$. 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 $\mathbb{Z}$ for both populations and, by performing an exception rule analysis, he could conclude that the interaction between two itemsets X and Z (where Z is the itemset corresponding to the values of $\mathbb{Z}$) 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.

3 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 itemsets containing discrete values.

Formally, we can give the following definition to anomalous association rules:

Definition 1. *Let X, Y, and A be arbitrary itemsets. We say that $X \rightsquigarrow A$ is an anomalous rule with respect to $X \Rightarrow Y$, where A denotes the Anomaly, if the following conditions hold:*

a) $X \Rightarrow Y$ *is a strong rule (frequent and confident)*
b) $X\neg Y \Rightarrow A$ *is a confident rule*
c) $XY \Rightarrow \neg A$ *is a confident rule*

It should be noted that, implicitly in the definition, we have used the common minimum support ($MinSupp$) and confidence ($MinConf$) 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, 25], 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 $\text{conf}(X \Rightarrow Y) = 0.6$, $\text{conf}(XZ \Rightarrow \neg Y) = \text{conf}(XZ \Rightarrow A) = 1$, and $\text{conf}(X \Rightarrow Z) = 0.2$. If we suppose that the itemset XY satisfies the support threshold and we use 0.6 as confidence threshold, then "$XZ \Rightarrow A$ is an exception to $X \Rightarrow Y$, with reference rule $X \Rightarrow \neg Z$". This exception is not highlighted as an anomaly using our approach because A is not always present when $X\neg Y$. In fact, $\text{conf}(X\neg Y \Rightarrow A)$ 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 $\text{conf}(X \Rightarrow Y) = 0.8$, $\text{conf}(XY \Rightarrow \neg A) = 0.75$, and $\text{conf}(X\neg Y \Rightarrow A) = 1$. No Z-value exists to originate an exception, but $X \rightsquigarrow A$ is clearly an anomaly with respect to $X \Rightarrow Y$.

X	Y	A_4	Z_3	$\cdots$
X	Y	A_1	Z_1	$\cdots$
X	Y	A_2	Z_2	$\cdots$
X	Y	A_1	Z_3	$\cdots$
X	Y	A_2	Z_1	$\cdots$
X	Y	A_3	Z_2	$\cdots$
X	Y_1	A_4	Z_3	$\cdots$
X	Y_2	A_4	Z_1	$\cdots$
X	Y_3	A	Z	$\cdots$
X	Y_4	A	Z	$\cdots$
		$\cdots$		

Fig. 1. A is an exception to $X \Rightarrow Y$ 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 (attributes in a relational database) is high, then the chance of finding spurious Z itemsets correlated with $\neg Y$ notably increases. As a consequence, the number of rules obtained can be really high (see [25, 23] 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

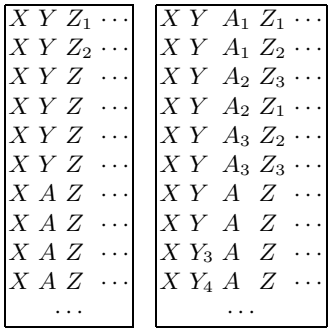

Fig. 2. $X \rightsquigarrow A$ is detected as an anomalous rule, even when no exception can be found through the Z-values.

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 $X\neg Y \Rightarrow A$ and $XY \Rightarrow \neg A$ are, the stronger the anomaly is. This fact could be useful in order to define a degree of strength associated to the anomaly.

4 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 itemsets.
- Generating the association rules derived from the previously-obtained itemsets.

The first subtask is the most time-consuming part and many efficient algorithms have been devised to solve it in the case of conventional association rules. For instance, Apriori-based algorithms are iterative [16]. Each iteration consists of two phases. The first phase, candidate generation, generates potentially frequent k-itemsets (C_k) from the previously obtained frequent (k-1)-itemsets (L_{k-1}). The second phase, support counting, scans the database to find the actual frequent k-itemsets (L_k). Apriori-based algorithms are based on the fact that that all subsets of a frequent itemset are 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 itemsets [11].

In the case of anomalous association rules, when we say that $X \rightsquigarrow A$ is an anomalous rule with respect to $X \Rightarrow Y$, that means that the itemset $X \cup \neg Y \cup A$ appears often when the rule $X \Rightarrow Y$ does not hold. Since it represents an anomaly, by definition, we cannot establish any minimum support threshold for $X \cup \neg Y \cup A$. 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 $X \rightsquigarrow A$ with respect to $X \Rightarrow Y$, let us denote by R the subset of the database that, containing X, does not verify the association rule $X \Rightarrow Y$. In other words, R will be the part of the database that does not verify the rule and might host an anomaly. The anomalous association rule confidence will be, therefore, given by the following expression:

$$conf_R(X \rightsquigarrow A) = \frac{supp_R(X \cup A)}{supp_R(X)}$$

When we write $supp_R(X)$, it actually represents $supp(X \cup \neg Y)$ in the complete database. Although this value is not usually computed when obtaining the itemsets, it can be easily computed as $supp(X) - supp(X \cup Y)$. Both values in this expression are always available after the conventional association rule mining process, since both X and $X \cup Y$ are frequent itemsets.

Applying the same reasoning, the following expression can be derived to represent the confidence of the anomaly $X \rightsquigarrow A$ with respect to $X \Rightarrow Y$:

$$conf_R(X \rightsquigarrow A) = \frac{supp(X \cup A) - supp(X \cup Y \cup A)}{supp(X) - supp(X \cup Y)}$$

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 $X \cup A$ and $X \cup Y \cup A$ might not be frequent, they are extensions of the frequent itemsets X and $X \cup Y$, respectively.

Since A will represent individual items, our problem reduces to being able to compute the support of $L \cup i$, for each frequent itemset 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 itemset extensions:

- Candidate generation: As in any Apriori-based algorithm, we generate potentially frequent k-itemsets from the frequent itemsets 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-itemset is obtained by scanning the database in order to obtain the actual frequent k-itemsets.
- 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 itemsets, an additional database scan can be used to compute the support for the extensions of the larger frequent itemsets.

Using a variation of an standard association rule mining algorithm as TBAR [4], nicknamed ATBAR (Anomaly TBAR), we can efficiently compute the support for each frequent itemset 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 itemset tree used by algorithms like TBAR [4]. The extended itemset tree stores the support for frequent itemset extensions as well as for all the frequent itemsets 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 itemset tree as an index to explore a database in the quest for anomalies.

As we will see in the following section, this data structure can be built with a relatively low overhead with respect to the cost of discovering just the frequent itemsets (which is what we need to discover standard association rules).

5 Experimental results

Table 1 presents three of the datasets from the UCI Machine Learning Repository we used to test ATBAR, as well as some information on the number of frequent patterns obtained for different minimum support thresholds. Since we were interested in detecting anomalies, we removed binary attributes from the original ADULT and CENSUS databases. In general, anomalies should be detected only when there are more than three alternatives, since in those situations is where they can provide insight

The experiments where performed using a 2.4GHz Celeron notebook with 512MB of main memory, running Windows XP, and using Borland InterBase 6.5 as database management system. The mining algorithms were implemented in Java and Sun Microsystems' JDK 1.4.1_03 runtime environment was employed in the tests.

Table 2 shows the time needed to mine all the frequent patterns in the case of traditional association rule mining and the frequent patterns plus their extensions in the case of anomalous association rule mining. The need to obtain the support for frequent itemset extensions obviously incurs in some overhead,

Database	Size metrics		Frequent itemsets					
Name	Dataset size	Items	Support	L[1]	L[2]	L[3]	L[4]	L[5]
NURSERY	12960 × 9	116640	10%	30	137	12	0	
			5%	30	389	226	19	0
ADULT'	48842 × 12	586104	10%	28	156	393	551	461
			5%	39	284	889	1558	1666
CENSUS'	299285 × 33	9876405	10%	60	1178	12244	82521	501141
			5%	83	1767	19917	146411	784886

Table 1. Datasets used in our experiments.

Database	Rule size	MinSupp	Anomalies	Associations	Overhead
NURSERY	3	5%	11282ms	11559ms	-2%
	4	5%	15514ms	16085ms	-4%
	5	5%	20153ms	17795ms	13%
ADULT'	3	10%	35954ms	34586ms	4%
		5%	45794ms	34824ms	31%
	4	10%	58756ms	44963ms	31%
		5%	92532ms	48466ms	91%
	5	10%	91543ms	56611ms	62%
		5%	147847ms	72749ms	103%
CENSUS'	3	10%	892363ms	663634ms	34%
		5%	993228ms	630306ms	58%
	4	10%	4829324ms	1832515ms	164%
		5%	5713505ms	1999355ms	186%

Table 2. Time required to obtain the support for all relevant itemsets.

Database	Rule size	MinSupp	Anomalies	Associations	Variation
NURSERY	3	5%	0	51	-100%
	4	5%	1	88	-99%
	5	5%	2	88	-98%
ADULT'	3	10%	3	512	-99%
		5%	6	932	-99%
	4	10%	41	1648	-98%
		5%	78	3642	-98%
	5	10%	174	3250	-95%
		5%	391	8631	-95%
CENSUS'	3	10%	167	19235	-99%
		5%	264	27388	-99%
	4	10%	5212	294579	-98%
		5%	7903	444021	-98%

Table 3. Number of rules obtained with 90% minimum confidence.

although this overhead is reasonable even for large datasets. Moreover, when looking for particular anomalies, the user usually is interested in particular attributes, which reduces the number of itemset extensions needed and paves the way for constraint-based anomalous rule mining.

Table 3 illustrates the rule set size you might expect when mining anomalous association rules. As this particular table illustrates, there are many less anomalous rules than association rules. In all the experiments we performed, the number of anomalous rules was less than one fifth the number of association rules with the same number of items involved. Usually, the number of anomalies is one order of magnitude less than the number of associations.

With respect to the particular experiments we performed, we found that there are almost no anomalies in the NURSERY dataset. This is a small dataset, about 13K tuples, from which you can derive a relatively small set of associations. In fact, using 90% as minimum confidence threshold, we discovered only these two anomalies using a 5% minimum support threshold:

```
if NURSERY:very_crit
   and HEALTH:priority
then CLASS:priority
when not CLASS:spec_prior
   (9 out of 9/864)

if PARENTS:great_pret
   and FINANCE:inconv
   and HEALTH:priority
then CLASS:priority
when not CLASS:spec_prior
   (66 out of 66/720)
```

In these anomalous rules, whose style closely corresponds to the prototypical anomalies discussed in Section 3, (α `out of` β`/`γ`)` means that α instances verify the anomaly out of β instances that verify the antecedent but not the usual consequent (i.e., the consequent in the association rule that hides the anomaly). Finally, γ refers to the number of instances in the database that verify the rule antecedent.

Both rules above highlight anomalous situations in nursery school applications. In rare cases, applications are not given the highest priority ('spec-prior'), even when that would be the usual outcome. Since application ranking requires objective explanations for disgruntled parents, further research might be needed to justify these ranking anomalies.

With respect to the ADULT database, we also found a small set of anomalies. One of them drew our attention almost immediately:

```
if WORKCLASS:Local-gov
then CAPGAIN:[99999.0,99999.0]
when not CAPGAIN:[0.0,20051.0]
   (7 out of 7/3136)
```

That is, local government employees do not usually experience a huge capital gain, but those who do view their capital substantially increased. Maybe they received an inheritance. This particular example illustrates how anomalies can be detected without the need to have data explaining them in the original dataset (as required in [25]).

Anomalous association rules do not have to correspond just to an alternative value for the attribute present in the association consequent, as happened above. The consequent can refer to different attributes, as the following anomaly from the CENSUS database shows:

```
if AAGE:[0.0,18.0]
then ACLSWKR:Private
when not ADTIND:[0.0,11.0]
   (4896 out of 5416/87518)
```

This rule states that children and teenagers sometimes belong to the class of workers knows as 'private wage and salary workers', specially when their industry code occupation is not in the [0,11] interval. Obviously, nobody would expect them to be government workers nor self-employed at such a young age, but that is an explanation the expert must find when reasoning about her discoveries.

In fact, if the anomaly derives from an association rule with several itemsets in its consequent, the rule might read something like this (again from the CENSUS database):

```
if AHGA:Children
then AMARITL:'Married-Spouse present'
when not ( ACLSWKR:'Not in universe'
           and AMARITL:'Never married' )
   (10 out of 10/70864)
```

which could be read as, if a person has children and has married someone but currently has no job, then his/her spouse must be alive (and possibly working to earn some money for the family).

The previous examples try to illustrate the kind of knowledge anomalous association rules provide. Their discovery process, which resembles the conventional association rule mining process, provides the basis upon which exploratory tools can be built to detect anomalies. Once detected, the discovery process should continue to find an explanation, which might exist in the original dataset or not.

6 Conclusions 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 Kowledge 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 classification model. *Machine Learning*, 54(1):67–92, 2004.
6. DR Carvalho, AA Freitas, and NFF Ebecken. A critical review of rule surprisingness 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. Han and 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 SIGMOD 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 Learning 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.
16. R. Srikant R. Agrawal. Fast algorithms for mining association rules. In *Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile*, 1994.
17. Sridhar Ramaswamy, Sameer Mahajan, and Abraham Silberschatz. On the discovery of interesting patterns in association rules. In *The VLDB Journal*, pages 368–379, 1998.
18. Sigal Sahar. Interestingness via what is not interesting. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 332–336, 1999.
19. Devavrat Shah, Laks V. S. Lakshmanan, Krithi Ramamritham, and S. Sudarshan. Interestingness and pruning of mined patterns. In *1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 1999.
20. J.-L. Sheng-Ma, Hellerstein. Mining mutually dependent patterns. In *Proceedings ICDM'01*, pages 409–416, 2001.
21. A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. On Knowledge And Data Engineering*, 8:970–974, 1996.
22. Ramakrishnan Srikant, Quoc Vu, and Rakesh Agrawal. Mining association rules with item constraints. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 67–73. AAAI Press, 14–17 1997.
23. Einoshin Suzuki. Scheduled discovery of exception rules. In *Discovery Science*, volume 1721, pages 184–195. Lecture Notes in Artificial Intelligence, 1999.
24. Einoshin Suzuki. In pursuit of interesting patterns with undirected discovery of exception rules. In *Progress Discovery Science*, volume 2281, pages 504–517. Lecture Notes in Artificial Intelligence, 2001.
25. Einoshin Suzuki. Undirected discovery of interesting exception rules. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(8):1065–1086, 2002.
26. Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29:293–313, 2003.
27. J. Yang, W. Wang, and P.S. Yu. Mining surprising periodic patterns. *Data Mining and Knowledge Discovery*, 9:1–28, 2004.
28. Ning Zhong, Yiyu Yao, and Muneaki Ohshima. Peculiarity oriented multidatabase mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):952–960, 2003.