Discovering hidden association rules

Marco-Antonio Balderas[†], Fernando Berzal[⋆], Juan-Carlos Cubero[⋆], Eduardo Eisman[†], Nicolás Marín[⋆]
Department of Computer Science and AI
University of Granada
Granada 18071 Spain
[⋆]{fberzal|jc.cubero|nicm}@decsai.ugr.es,[†]{mbald|eeisman}@correo.ugr.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.

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].

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 be 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 supp$(X \Rightarrow Y)$ and conf$(X \Rightarrow Y)$, respectively. The usual minimum support and confidence 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 *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 <u>A</u>nomaly, 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

In order to emphasize the involved consequents, we will also used the notation $X \rightsquigarrow A|\neg Y$, which can be read as:
"X is associated with A when Y is not present"

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 conf$(X \Rightarrow Y) = 0.6$, conf$(XZ \Rightarrow \neg Y) =$ conf$(XZ \Rightarrow A) = 1$, and 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, 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 conf$(X \Rightarrow Y) = 0.8$, conf$(XY \Rightarrow \neg A) = 0.75$, and conf$(X\neg Y \Rightarrow A) = 1$. No Z-value exists to originate an exception, but $X \rightsquigarrow A|\neg Y$ is clearly an anomaly.

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

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	$\cdots$
		$\ldots$		

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

$\neg Y$ 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.

X	Y	Z_1	$\cdots$
X	Y	Z_2	$\cdots$
X	Y	Z	$\cdots$
X	Y	Z	$\cdots$
X	Y	Z	$\cdots$
X	Y	Z	$\cdots$
X	A	Z	$\cdots$
X	A	Z	$\cdots$
X	A	Z	$\cdots$
X	A	Z	$\cdots$
		$\ldots$	

X	Y	A_1	Z_1	$\cdots$
X	Y	A_1	Z_2	$\cdots$
X	Y	A_2	Z_3	$\cdots$
X	Y	A_2	Z_1	$\cdots$
X	Y	A_3	Z_2	$\cdots$
X	Y	A_3	Z_3	$\cdots$
X	Y	A	Z	$\cdots$
X	Y	A	Z	$\cdots$
X	Y_3	A	Z	$\cdots$
X	Y_4	A	Z	$\cdots$
		$\ldots$		

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

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 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.

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 $X \rightsquigarrow A|\neg Y$ is an anomalous rule, 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 a minimum support threshold for $X \cup \neg Y \cup A$, in the same sense than a strong rule. In fact, an anomaly is not usually very frequent in the whole database. Therefore, standard association rule mining algorithms, exploiting the classical *Apriori* support pruning, cannot be used to detect anomalies without modification.

Given an anomalous association rule $X \rightsquigarrow A|\neg Y$, 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:

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

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 $X \rightsquigarrow A|\neg Y$:

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

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.

5. Pruning and summarizing rules

Deriving anomalous association rules without imposing some constraints is meaningless. We introduce some general criteria which can be divided into two groups: *a priori* and *a posteriori*.

A priori pruning criteria. (Restrictions imposed before proceeding to the construction of the itemset tree)

- Do not allow an attribute with only two different values to appear in the anomalous consequent part of the rule. In general, attributes appearing in the anomalous consequents, should have at least three or four distinct values.
- Null values should not appear in the anomalous consequent part of a rule, but they could appear in the strong part. A strong rule with a null consequent but a non-null anomalous consequent could provide useful information to the user.

A posteriori pruning criteria. (Criteria imposed once the set of anomalous rules is construted)

- Eliminate those rules sharing the same strong and anomalous consequent, and having more antecedents. In this case, the simplest rule is included and the others are pruned.

  If there exists an anomalous rule $X \rightsquigarrow A|\neg Y$, then every anomalous rule $XH \rightsquigarrow A|\neg Y$ is pruned.
- Do not allow anomalies supported by just one or two records. Thus, a support threshold for the anomaly should be considered. A minimum support of three records might be choosen.

  If supp$(XA) < 3$, then $X \rightsquigarrow A|\neg Y$ is pruned.

			Confidence 90%				Confidence 75%			
DataBase	Ant. Size	MinSupp	Anom. Prun.	Anom.	Assoc.	Reduct.	Anom. Prun.	Anom.	Assoc.	Reduct.
HEPATITIS	1	10%	4	61	131	97%	57	229	398	86%
		5%	4	63	137	97%	70	253	427	84%
		1%	4	63	238	98%	70	398	561	88%
	2	10%	11	901	1639	99%	222	3029	3820	94%
		5%	11	1806	3249	99%	310	7017	7352	96%
		1%	11	1806	12406	100%	310	13496	18836	98%
BREAST-CANCER	1	10%	0	0	9	100%	1	2	43	98%
		5%	0	2	12	100%	2	5	61	97%
		1%	0	2	24	100%	2	35	89	98%
	2	10%	3	11	62	95%	27	50	265	90%
		5%	3	55	146	98%	44	146	485	91%
		1%	3	85	736	100%	50	574	1423	96%
WISCONSIN-BREAST-CANCER	1	10%	1	2	29	97%	1	2	80	99%
		5%	1	13	43	98%	4	19	117	97%
		1%	1	47	70	99%	15	121	170	91%
	2	10%	7	63	183	96%	33	100	427	92%
		5%	16	163	313	95%	71	248	688	90%
		1%	19	600	936	98%	117	1634	1811	94%
POSTOPERATIVE	1	10%	0	0	14	100%	3	5	29	90%
		5%	0	0	14	100%	3	6	30	90%
		1%	0	0	43	100%	3	6	59	95%
	2	10%	0	11	87	100%	2	37	206	99%
		5%	0	11	123	100%	2	57	310	99%
		1%	0	11	586	100%	2	64	792	100%
CONTRACEPTIVE	1	10%	0	0	32	100%	3	3	76	96%
		5%	0	0	34	100%	3	3	84	96%
		1%	0	0	36	100%	3	3	87	97%
	2	10%	4	7	132	97%	9	34	253	96%
		5%	16	32	311	95%	49	131	612	92%
		1%	17	65	527	97%	106	314	1114	90%
PIMA DIABETES	1	10%	0	0	36	100%	0	0	49	100%
		5%	0	0	36	100%	0	0	49	100%
		1%	0	0	36	100%	0	0	49	100%
	2	10%	0	2	45	100%	4	12	54	93%
		5%	1	25	185	99%	17	124	232	93%
		1%	1	141	543	100%	77	691	834	91%

Table 1. Number of rules obtained after pruning

These pruning methods should be applied to eliminate spurious and trivial anomalous rules. The application of these simple criteria can dramatically decrease the number of outputs as Table 1 shows (see description of Table 1 in next section). Once the reduced set of rules is obtained, summarizing and ranking measures could also be applied. Such measures should be applied once the whole set of pruned rules are discovered. The particular measures used for a particular problem might depend on specific domain knowledge. Some criteria are:

Summarizing criteria help us to merge several rules into a single one.

For instance, we can merge several rules with the same pair of strong and anomalous consequents in the following way:

> All the anomalous rules $X_i \rightsquigarrow A|\neg Y$, could be merged into one single rule $(\vee_i X_i) \rightsquigarrow A|\neg Y$, where $\vee$ stands for the logical or.

This summarizing method is aimed at presenting a simple set of rules to the user. Obviously, the confidence and support values can not be merged and, therefore, the individual rules should still be stored in case the user wanted to analyze them.

Let us note that the greater the number of different X_i are merged, the more confident we are that the negative association between Y and A, is not related to those X_i. For instance, Y could stand for *Less than 18 years* and A for *Has the car licence*.

On the other hand, the first a posteriori pruning method we introduced before could be rewritten as a summarizing one, but following Occam's razor we prefer to consider the simplest rule, and thus eliminate (not summarize) unnecessarily complex rules.

Ranking measures give a numerical value to the interest of each rule. Some examples are:

- If an anomalous rule involves the same numerical attribute in the strong and in the anomalous consequent part, then a ranking measure could give more importance to those rules where such intervals are not closed, because such rule would detect very opposite behaviors.
- The more confident the rules $X\neg Y \Rightarrow A$ and $XY \Rightarrow \neg A$ are, the stronger the $X \rightsquigarrow A|\neg Y$ anomaly is. This fact could be useful in order to define a degree of strength associated to the anomaly.

6. Experimental results

Table 1 presents some results obtained with ATBAR using datasets from the UCI Machine Learning Repository (we focused our experimentation on medical datasets). As motivated in Section 3, we only consider associations with one consequent value. Numerical attributes are a priori clustered in 5 intervals by using a classical equi-depth partitioning algorithm. `Ant.Size` represents the number of antecedents. We restrict our experimentation to the case of one and two antecedents. `MinSupp` is the support threshold (as a percentage) for the strong rule. `Confidence` is the confidence of the strong rule (as well as the confidence of the anomaly), as stated in definition 1. `Anom` is the number of anomalous rules. `Anom.Prun.` is the number of pruned rules obtained by using the basic methods introduced in Section 5 with four distinct values in each attribute (we do not apply any ranking measure or summarizing criteria). `Assoc` is the number of association rules satisfying the support (row) and confidence (column) thresholds. `Reduct` is the reduction percentage of `Anom Pruned` with respect to `Assoc`. It is worth mentioning that this percentage is included only as a reference to the problem complexity, because anomalies and associations are not the same concept.

The need to obtain the support for frequent itemset extensions obviously incurs in some overhead, although it is reasonable even for large datasets. The overhead in time is about 20% in the experiments we have performed.

7. 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. Our approach could 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.

We intend to apply our technique to huge datasets as well as to contrast the results with experts in order to evaluate the false positive rate and analyze summarizing criteria in depth, so more rules can be pruned.

Acknowledgments

Marco-Antonio Balderas was supported by a scholarship grant from Mexican National Research Council on Science and Technology and the ”Universidad Autonoma de Tamaulipas” (CONACyT-PROMEP).
Eduardo Eisman was supported by a scholarship grant from the University of Granada in Spain.
During the preparation of this paper, Fernando Berzal was a visiting research scientist at the research group led by prof. Jiawei Han at the University of Illinois at Urbana-Champaign.

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] R. Agrawal and R. Srikant. Mining sequential patterns. In P. S. Yu and A. S. P. Chen, editors, *Eleventh International Conference on Data Engineering*, pages 3–14, Taipei, Taiwan, 1995. IEEE Computer Society Press.

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

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

[5] F. Berzal, J. Cubero, D. Sanchez, and J. Serrano. Art: A hybrid classification model. *Machine Learning*, 54(1):67–92, 2004.

[6] D. Carvalho, A. Freitas, and N. Ebecken. A critical review of rule surprisingness measures. In N. Ebecken, C. Brebbia, and A. Zanasi, editors, *Proc. Data Mining IV - Int. Conf. on Data Mining*, pages 545–556. WIT Press, December 2003.

[7] E. Cohen, M. Datar, S. Fujiwara, A. Gionis, P. Indyk, R. Motwani, J. D. Ullman, and C. Yang. Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):64–78, 2001.

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

[9] A. A. 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] J. 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] F. Hussain, H. Liu, E. Suzuki, and H. Lu. Exception rule mining with a relative interestingness measure. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 86–97, 2000.

[14] Y. 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] B. Liu, W. Hsu, S. Chen, and Y. Ma. Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, pages 47–55, 2000.

[16] R. S. R. Agrawal. Fast algorithms for mining association rules. In *Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile*, 1994.

[17] S. Ramaswamy, S. Mahajan, and A. Silberschatz. On the discovery of interesting patterns in association rules. In *The VLDB Journal*, pages 368–379, 1998.

[18] S. 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] D. Shah, L. V. S. Lakshmanan, K. 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] R. Srikant, Q. Vu, and R. Agrawal. Mining association rules with item constraints. In D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD*, pages 67–73. AAAI Press, 14–17 1997.

[23] E. Suzuki. Scheduled discovery of exception rules. In *Discovery Science*, volume 1721, pages 184–195. Lecture Notes in Artificial Intelligence, 1999.

[24] E. 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] E. Suzuki. Undirected discovery of interesting exception rules. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(8):1065–1086, 2002.

[26] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right objective measure for association analysis. *Information Systems*, 29:293–313, 2003.

[27] J. Yang, W. Wang, and P. Yu. Mining surprising periodic patterns. *Data Mining and Knowledge Discovery*, 9:1–28, 2004.

[28] N. Zhong, Y. Yao, and M. Ohshima. Peculiarity oriented multidatabase mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):952–960, 2003.