Anomaly detection using fuzzy association rules

M. Dolores Ruiz* and Maria J. Martin-Bautista
Department of Computer Science and A.I.,
CITIC-UGR,
University of Granada,
C/ Periodista Rafael Gómez, Granada, Spain
E-mail: mdruiz@decsai.ugr.es
E-mail: mbautis@decsai.ugr.es
*Corresponding author

Daniel Sánchez
Department of Computer Science and A.I.,
CITIC-UGR,
University of Granada,
C/ Periodista Rafael Gómez, Granada, Spain
European Centre for Soft Computing,
C/ Gonzalo Gutiérrez Quiró, Mieres, Spain
E-mail: daniel@decsai.ugr.es
E-mail: daniel.sanchez@softcomputing.es

Maria-Amparo Vila and Miguel Delgado
Department of Computer Science and A.I.,
CITIC-UGR,
University of Granada,
C/ Periodista Rafael Gómez, Granada, Spain
E-mail: vila@decsai.ugr.es
E-mail: mdelgado@decsai.ugr.es

Abstract: Data mining techniques are a very important tool for extracting useful knowledge from databases. Recently, some approaches have been developed for mining novel kinds of useful information, such as anomalous rules. These kinds of rules are a good technique for the recognition of normal and anomalous behaviour, that can be of interest in several area domains such as security systems, financial data analysis, network traffic flow, etc. The aim of this paper is to propose an association rule mining process for extracting the common and anomalous patterns in data that is affected by some kind of imprecision or uncertainty, obtaining information that will be meaningful and interesting for the user. This is done by mining fuzzy anomalous rules. We present a new approach for mining such rules, and we apply it to the case of detecting normal and anomalous patterns on credit data.

Keywords: data mining; fuzzy association rules; anomalous rules; anomaly detection; credit; electronic security; digital forensics.

Biographical notes: M. Dolores Ruiz received her Mathematics degree in 2005 and a European PhD in Computer Science in 2010, both from the University of Granada. She held non-permanent teaching positions in the Universities of Jaén and Granada. She is currently working at the Computer Science and Artificial Intelligence Department in the University of Granada (CITIC-UGR) and she belongs to the Approximate Reasoning and AI research group. She is a co-organiser of the FQAS’2013 conference. Her research areas of interest are: data mining, information retrieval, correlation statistical measures, sentence quantification and fuzzy sets theory. Her expertise involves knowledge extraction from databases involving uncertainty using association rules, exception rules, anomalous rules and gradual dependences, as well as, formal modelling for the representation and evaluation of association rules.

Maria J. Martin-Bautista is a Permanent Professor at the Department of Computer Science and Artificial Intelligence at the University of Granada, Spain, where she received her PhD in Computer Science in 2000. She is a member of the Intelligent Data Bases and Information Systems (IDBIS) research group. Her current research interests include data, text and web mining, intelligent information systems, data warehousing, information retrieval and knowledge representation with fuzzy logic. She has supervised several PhD theses and she has published more than 60 papers in journals and international conferences. She has leaded and participated in more than 15 R+D national and international projects and has supervised several research and technology transfer projects with companies. She is a member of the European Society for Fuzzy Logic and Technology.

Daniel Sánchez received his MS and PhD degrees in Computer Science, both from the University of Granada, Spain, in 1995 and 1999, respectively. Since 2001, he is an Associate Professor at the Department of Computer Science and Artificial Intelligence of the University of Granada. He has participated and is currently a member of the teams of more than ten research projects, leading currently a national project in the area of linguistic description of visual information. He has published more than 100 papers in international journals and conferences, 23 of them in indexed journals (SCI-JCR). He has co-edited a special issue in the International Journal of Intelligent Systems (IJIS) about ‘Fuzzy bags, cardinality, and quantification’. He has organised several workshops and special sessions about soft computing and applications in data mining and computer vision in national and international conferences, and has been advisor of four PhD theses in these fields.

Maria-Amparo Vila is a Full Professor of Computer Science and AI at the University of Granada since 1992, she was previously Associate Professor and Analyst Programmer of the Computer Centre at the same university. She has developed her research and teaching primarily in the area of databases and intelligent information systems, and its main lines of research are: treatment of imprecision in information systems using fuzzy logic, knowledge discovery in databases using techniques ‘Soft Computing’, and ubiquitous computing and knowledge mobilisation. She has been the advisor of 19 doctoral theses, she has been or is responsible for more than ten research projects and she has published numerous research papers, which highlights more than 80 papers in SCI journals. She has been responsible for the research groups ARAI and IDBIS in
Miguel Delgado is a Full Professor at the Department of Computer Science and AI at the University of Granada, Spain, where he received his PhD in Mathematics in 1975. He is the Head of the Approximate Reasoning and Artificial Intelligence research group. His main research lines are knowledge representation with uncertainty and imprecision management and ubiquitous computing systems. He has supervised several PhD theses and he has published more than 100 papers in journals. He has given more than 20 lectures in several universities all over the world on topics related to approximate reasoning, intelligent systems, data mining and knowledge mobilisation. He is a member of several Scientific Societies like European Association of Technology and Fuzzy Logic (EUSFLAT), International Fuzzy Systems Association (IFSA) and the Institute of Electrical and Electronic Engineers (IEEE).

This paper is a revised and expanded version of a paper entitled ‘Anomaly detection using fuzzy association rules’ presented at 9th ICGS3-13 Conference, London, 4–6 December 2013.

1 Introduction

Nowadays, the recognition of normal and anomalous patterns has gained much importance in research. Systems able to recognise and classify such type of patterns are very important in lot of areas such as in security systems, financial data analysis, network traffic flow, biological or chemical processes, etc. The classification of these two kinds of patterns had been made by using expert systems (Chemouil and Stern, 1992), neural networks (Didelet et al., 1994; Ziaii et al., 2012) or clustering techniques (Thiprungsri and Vasarhelyi, 2011; Cai et al., 2012).

Data mining techniques are also important in this area. Previous works try to discover the usual profiles of legitimate customer behaviour and then search the anomalies using different methodologies such as clustering (Fawcet and Provost., 1997). Association rules seem to be useful in this context too, since they allow identifying novel, useful and comprehensive knowledge. The kind of knowledge they try to extract is the appearance of a set of items together in most of the transactions in a database. An example of association rule is “most of transactions that contain hamburger also contain beer”, and it is usually noted ‘hamburger! Beer’. In particular, there exist few approaches dealing with the extraction of unusual or anomalous knowledge (Berzal et al., 2004; Delgado et al., 2011a) that might be useful for discovering both types of patterns: the normal and the anomalous one. In general, these approaches are able to manage rules that, being infrequent, provide a specific domain information usually delimited by an association rule.

It is also well known, that the majority of information that we manage is affected by imprecision, is ambiguous or uncertain in its nature. In these cases the theory of fuzzy Subsets proposed by Zadeh (1965) has been proved to be a very good tool. Some advantages of the use of fuzzy sets are, on the one hand, softening bounds and, on the other hand, giving a formal representation for the most semantically significant and
meaningful knowledge for the user. In particular, it has been used for mining fuzzy association rules in a lot of scenarios (Hüllermeier, 2011; Delgado et al., 2011b).

The main scope of this paper is then to apply such kind of ‘infrequent’ rules to the case of detecting anomalous patterns automatically, as well as the normal pattern to which the anomaly is related, that could help for obtaining the common customer behaviour (association rules) as well as the anomalous deviations (anomalous rules). For this purpose, we present a new approach that takes advantage of fuzzy techniques for mining rules in order to extract the normal and anomalous patterns in form of fuzzy rules. In addition we perform several experiments in financial data concerning bank credits.

The structure of the paper is the following: next section offers a brief description of background concepts about crisp and fuzzy association rules. In Section 3, we review the previous approach for defining anomalous rules and some related works in this area. Then, in Section 4, we develop our approach in order to extract fuzzy anomalous rules. Section 5 presents the algorithm for extracting these kinds of rules and its application to the real dataset German-Statlog about credits in a certain bank. Finally, Section 6 contains the conclusions and the future research.

2 Background concepts

2.1 Association rules

Association rules are one of the frequent used tools in data mining. Given a set $I$ (‘set of items’) and a database $D$ constituted by set of transactions, each one being a subset of $I$, association rules (Agrawal et al., 1993) are ‘implications’ of the form $A \rightarrow B$ that relate the presence of itemsets $A$ and $B$ in transactions of $D$, assuming $A, B \subseteq I$, $A \cap B = \emptyset$, $A, B \neq \emptyset$. The intensity of the above association rule is frequently measured by the ordinary measures of support and confidence proposed in Agrawal et al. (2002). The support of an itemset is defined as the probability that a transaction contains the itemset, i.e., $\text{supp}(A) = \frac{|\{t \in D \mid A \subseteq t\}|}{|D|}$.

The support of a rule is the percentage of transactions satisfying both parts of the rule [i.e., the joint probability $P(A \cup B)$] :

$$\text{Supp}(A \rightarrow B) = \text{supp}(A \cup B)$$

And the confidence of a rule measures the proportion of transactions that satisfying the antecedent, also satisfies the consequent [i.e., the conditional probability $P(B \mid A)$] :

$$\text{Conf}(A \rightarrow B) = \frac{\text{supp}(A \cup B)}{\text{supp}(A)}$$

Given the minimum thresholds $\text{minsupp}$ and $\text{minconf}$, that should be imposed by the user, we will say that $A \rightarrow B$ is frequent if $\text{Supp}(A \rightarrow B) \geq \text{minsupp}$, and confident if $\text{Conf}(A \rightarrow B) \geq \text{minconf}$.

Definition 1 (Berzal et al., 2002): An association rule $A \rightarrow B$ is strong if it exceeds the minimum thresholds $\text{minsupp}$ and $\text{minconf}$ imposed by the user, i.e., if $A \rightarrow B$ is frequent and confident.
There also exist many proposals imposing new quality measures for extracting semantically or even statistically different association rules (Geng and Hamilton, 2006; Delgado et al., 2010). In this line, an alternative framework was proposed in Berzal et al. (2002) where the accuracy is measured by means of Shortliffe and Buchanan’s certainty factors (Shortliffe and Buchanan, 1975), in the following way.

Definition 2 (Delgado et al., 2003): Let \( \text{supp}(B) \) be the support of the itemset \( B \), and let \( \text{Conf}(A \rightarrow B) \) be the confidence of the rule. The certainty factor of the rule, is defined as:

\[
CF(A \rightarrow B) = \begin{cases} 
\frac{\text{Conf}(A \rightarrow B) - \text{supp}(B)}{1 - \text{supp}(B)} & \text{if } \text{Conf}(A \rightarrow B) > \text{supp}(B) \\
\frac{\text{Conf}(A \rightarrow B) - \text{supp}(B)}{1 - \text{supp}(B)} & \text{if } \text{Conf}(A \rightarrow B) < \text{supp}(B) \\
0 & \text{otherwise.}
\end{cases}
\]

The certainty factor yields a value in the interval \([-1, 1]\) and measures how our belief that \( B \) is in a transaction changes when we are told that \( A \) is in that transaction. Positive values indicate that our belief increases, negative values mean that our belief decreases, and 0 means no change. Certainty factor has better properties than confidence, and helps to solve some of its drawbacks (Berzal et al., 2002; Delgado et al., 2003). In particular, it helps to reduce the number of rules obtained by filtering those rules corresponding to statistical independence or negative dependence.

Analogously, we will say that \( A \rightarrow B \) is certain if \( \text{Supp}(A \rightarrow B) \geq \text{minCF} \), where \( \text{minCF} \) is the minimum threshold for the certainty factor given by the user. The definition for strong rules can be reformulated when using CF as a rule which must be frequent and certain.

Definition 3 (Berzal et al., 2002): An association rule \( A \rightarrow B \) is very strong if both rules \( A \rightarrow B \) and \( -B \rightarrow -A \) are strong.

In addition, the certainty factor has the following property \( CF(A \rightarrow B) = CF(-B \rightarrow -A) \), which tell us that when using the certainty factor, a strong rule is also very strong (Berzal et al., 2002).

### 2.2 Fuzzy association rules

In Delgado et al. (2003), the model for association rules is extended in order to manage fuzzy values in databases. The approach is based on the definition of fuzzy transactions as fuzzy subsets of items.

Definition 4 (Delgado et al., 2003): Let \( I = \{i_1, \ldots, i_m\} \) be a finite set of items. A fuzzy transaction is a non-empty fuzzy subset \( \tilde{\tau} \subseteq I \).

For every item \( i \in I \) and every transaction \( \tilde{\tau} \), an item \( i \) will belong to \( \tilde{\tau} \) with grade \( \tilde{\tau}(i) \) where \( \tilde{\tau}(i) \) is a real number in the interval \([0, 1]\). Let \( A \subseteq I \) be an itemset. The membership grade of \( A \) to the fuzzy transaction \( \tilde{\tau} \) is defined as \( \tilde{\tau}(A) = \min_{i \in A} \tilde{\tau}(i) \).

According to Definition 4 a crisp transaction is a special case of fuzzy transaction where every item in the transaction has membership grade equal to 1 or 0 depending on if they are in the transaction or not.
Definition 5 (Delgado et al., 2003): Let $I$ be a set of items, $\tilde{D}$ a set of fuzzy transactions and $A, B \subseteq I$ two disjoint itemsets, i.e., $A \cap B = \emptyset$. A fuzzy association rule $A \rightarrow B$ is completely satisfied in $\tilde{D}$ if and only if, $\tilde{\tau}(A) \leq \tilde{\tau}(B)$ for all $\tilde{\tau} \in \tilde{D}$, that is, the membership grade of $B$ is higher than the membership grade of $A$ for all fuzzy transactions $\tilde{\tau}$ in $\tilde{D}$.

This definition holds the meaning of crisp association rules because if we need $A \subseteq \tilde{\tau}$ to be satisfied, we also need to satisfy $B \subseteq \tilde{\tau}$. In our case this can be translated to $\tilde{\tau}(A) \leq \tilde{\tau}(B)$. In this way, since a crisp transaction is a special case of fuzzy transaction, a crisp association rule will be a special case of fuzzy association rule.

Example 1: We consider the set of items $I = \{i_1, i_2, i_3, i_4, i_5\}$ and the set of fuzzy transactions given by Table 1.

<table>
<thead>
<tr>
<th></th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{\tau}_1$</td>
<td>1</td>
<td>0.2</td>
<td>1</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>$\tilde{\tau}_2$</td>
<td>1</td>
<td>1</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tilde{\tau}_3$</td>
<td>0.5</td>
<td>0.1</td>
<td>0.7</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>$\tilde{\tau}_4$</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$\tilde{\tau}_5$</td>
<td>0.4</td>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\tilde{\tau}_6$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

In particular, we can see that $\tilde{\tau}_6$ is a crisp transaction. Some membership grade could be: $\tilde{\tau}_1(\{i_1, i_4\}) = 0.9$, $\tilde{\tau}_1(\{i_2, i_3, i_4\}) = 0.2$ and $\tilde{\tau}_3(\{i_1, i_2\}) = 1$.

In order to measure the interest and accuracy of a fuzzy association rule, we employ a semantic approach based on the evaluation of quantified sentences, using the fuzzy quantifier $Q_M(x) = x$, as follows:

- the support of an itemset $I$ is the evaluation of the quantified sentence ‘$Q_M$ of $\tilde{D}$ are $\tilde{I}$’ where $\tilde{I}$ is a fuzzy set defined as $\tilde{\Gamma}_I(t) = \tilde{\tau}(I)$

- the support of the fuzzy association rule $A \rightarrow B$ in $\tilde{D}$, $FSupp(A \rightarrow B)$, is the evaluation of the quantified sentence ‘$Q_M$ of $\tilde{D}$ are $(\tilde{\Gamma}_A \cap \tilde{\Gamma}_B)$’

- the confidence of the fuzzy association rule $A \rightarrow B$ in $\tilde{D}$, $FConf(A \rightarrow B)$, is the evaluation of the quantified sentence ‘$Q_M$ of $\tilde{D}$ are $\tilde{\Gamma}_A \cap \tilde{\Gamma}_B$’

- the certainty factor is obtained from support and confidence using the equation in Definition 2.

We evaluate a quantified sentence of the form ‘$Q$ of $F$ are $G$’ by means of method GD, defined in Delgado et al. (2000) as:
Anomaly detection using fuzzy association rules

\[ GD_G(G/F) = \sum_{\alpha_i \in \Lambda(G/F)} \left( \alpha_i - \alpha_{i+1} \right) \left( \frac{|G \cap F|}{|F_i|} \right) \]  

(3)

where \( \Lambda(G/F) = \Lambda(G \cap F) \cup \Lambda(F) \), \( \Lambda(F) \) being the level set of \( F \), and \( \Lambda(G/F) = \{ \alpha_1, \ldots, \alpha_p \} \) with \( \alpha_i > \alpha_{i+1} \) for every \( i \in \{1, \ldots, p-1\} \), and considering \( \alpha_{p+1} = 0 \). The set \( F \) is assumed to be normalised. If not, \( F \) is normalised and the same normalisation factor is applied to \( G \setminus F \). Other important properties defining the semantics of this proposal are those of equations (4) and (5).

\[ FConf(A \rightarrow B) = 1 \text{ if and only if } \forall \tilde{t} \in \tilde{D}, \, \tilde{t}(A) \leq \tilde{t}(B) \]  

(4)

\[ FCF(A \rightarrow B) = 1 \text{ if and only if } FConf(A \rightarrow B) = 1. \]  

(5)

Example 2: Following the previous example we can find in \( \tilde{D} \) the following fuzzy rules where we have considered \( \Lambda = \{1, 0.8, 0.6, 0.4, 0.2\} \) as the set of \( \alpha \)-cuts.

<table>
<thead>
<tr>
<th>Rule</th>
<th>FSupp</th>
<th>FConf</th>
<th>FCF</th>
</tr>
</thead>
<tbody>
<tr>
<td>{i_1, i_2} \rightarrow {i_3}</td>
<td>0.167</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>{i_4} \rightarrow {i_5}</td>
<td>0.2</td>
<td>0.767</td>
<td>0.68</td>
</tr>
</tbody>
</table>

We want to remark that we have normalised the itemset \( \{i_4\} \) and applied the same normalisation factor to \( \{i_4, i_5\} \) in order to compute the FConf and FCF measures.

3 Anomalous association rules

An anomalous rule is an association rule that is verified when the common rule fails. In other words, it comes to the surface when the dominant effect produced by the strong rule is removed (Berzal et al., 2004). Table 2 shows its formal definition, where the more confident the rules \( X \land \lnot Y \rightarrow A \) and \( X \land Y \rightarrow \lnot A \) are, the stronger the anomaly is. In this approach, there is no imposition over the support of the anomalous and the reference rules.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Description</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X \rightarrow Y )</td>
<td>Common sense rule</td>
<td>(High supp and high conf)</td>
</tr>
<tr>
<td>( X \land \lnot Y \rightarrow A )</td>
<td>Anomalous rule</td>
<td>(High conf)</td>
</tr>
<tr>
<td>( X \land Y \rightarrow \lnot A )</td>
<td>Reference rule</td>
<td>(High conf)</td>
</tr>
</tbody>
</table>

An example of anomalous rule will be: “if a patient have symptoms \( X \) then he usually has the disease \( Y \); if not, he has the disease \( A \)”. Anomalous rules try to capture the deviation from the common sense rule (i.e., from the usual behaviour). In other words: when \( X \), then we have either \( Y \) (usually) or \( A \) (unusually). In this case \( A \), is the alternative behaviour when the usual fails.

It is worth to mention that the reference rule acts as a pruning criterion to reduce the high number of obtained anomalies. In the following, we present other works that are complementary to that of anomalous rules that extract other types of interesting information by means of infrequent rules.
3.1 Related works

The common denominator when mining association rules is their high support. Usually, the mining process, as, for instance, Apriori (Agrawal et al., 1996), uses a candidate generation function which exploits the downward closure property of support (also called anti-monotonicity) which guarantees that for a frequent itemset all its subsets are also frequent. The problem here is that anomalous rules are infrequent rules, and therefore such property cannot be used. In the literature we can find different approaches utilising infrequent rules for capturing a novel type of knowledge hidden in data.

**Peculiarity rules** are discovered from the data by searching the relevant data among the peculiar data (Zhong et al., 2001). Roughly speaking, peculiar data is given by the attributes which contain any peculiar value. A peculiar value will be recognised when it is very different from the rest of values of the attribute in the data set. Peculiarity rules are defined as a new type of association rule representing a kind of regularity hidden in a relatively small number of peculiar data.

**Infrequent rules** are rules mined without exceeding the minimum support threshold. They have been studied mainly for intrusion detection joint with exceptions (Yao et al., 2005; Zhou and Yau, 2007). There exists some approaches for mining them: in Sim et al. (2008) the known Lambda measure is modified for obtaining more interesting rules using some pruning techniques. In Zhou and Yau (2007), infrequent items are obtained first and then some measures are used for mining the infrequent rules. In particular, they used correlation and interest measures together with an incremental ratio of conditional probabilities associated to pairs of items. In Ding and Yau (2009), the authors extract infrequent rules using a new structure called co-occurrence transactional matrix instead of new interest measures.

**Exception rules** were first defined as rules that contradict the user’s common belief (Suzuki, 1996). In other words, for searching an exception rule we have to find an attribute that changes the consequent of a strong rule (Suzuki and M. Shimura, 1996; Hussain et al., 2000; Suzuki, 2004).

We can find two different ways of mining exception rules: direct or indirect techniques. The former are in most of the cases highly subjective as the set of user’s beliefs is compared to the set of mined rules (Silberschatz and Tuzhilin, 1996; Padmanabhan and Tuzhilin, 1998; Liu et al., 1999). The indirect techniques use the knowledge provided by a set of rules (usually strong rules) and then the exception rules are those that contradict or deviate this knowledge (Suzuki, 2004; Yao et al., 2005). Good surveys on this topic can be found in Duval et al. (2007), Tanier et al. (2008), and Delgado et al. (2011).

**Anomalous rules** are in appearance similar to exception rules, but semantically different. Anomalous association rule is an association rule that appears when the strong rule ‘fails’. In other words, it is an association rule that complement the usual behaviour represented by the strong rule (Berzal et al., 2004). Therefore, the anomalous rules will represent the unusual behaviour, having in general low support.
4 Our approach for mining fuzzy anomalous rules

Our approach for extracting fuzzy anomalous rules is based on two ideas:

1. To define fuzzy anomalous rules using the domain $\tilde{D}_X$.

2. To use the certainty factor instead of the confidence. The certainty factor reduces the number of common sense rules since it discards non-reliable rules and, as a consequence, the number of anomalous rules is also reduced.

In Delgado et al. (2011), there is an analysis of the reference rule taken in the approach of Berzal et al. This analysis concludes affirming that the increasing of $\text{Conf}(X \rightarrow Y \rightarrow \neg A)$ is higher as $\text{Supp}(X \rightarrow Y)$ increases. This leads to affirm that the reference rule condition depends on the following supports $\text{Supp}(X \rightarrow Y) = \text{supp}(X \cup Y)$ and $\text{supp}(X \cup Y \cup A)$. This gives reason to propose an alternative formulation for anomalous rules changing the reference rule for a stronger condition than the one given in Berzal et al. (2004), and Balderas et al. (2005). We present here the approach for mining fuzzy anomalous rules in a fuzzy database $\tilde{D}$.

Definition 1: Let $X$, $Y$ be two non-empty itemsets and $A$ an item in $\tilde{D}$. We define a fuzzy anomalous rule by the triple $(F_{csr}, Fanom, F_{ref})$ composed of three fuzzy rules satisfying the following conditions:

- $X \rightarrow Y$ is a frequent and certain fuzzy rule ($F_{csr}$)
- $\neg Y \rightarrow A$ is certain in $\tilde{D}_X$ ($Fanom$)
- $A \rightarrow \neg Y$ is certain in $\tilde{D}_X$ ($F_{ref}$).

where the fuzzy versions of support and certainty factor given in Section 2.2 are used here.

Our proposal is similar and logically equivalent to that of Berzal et al. for the crisp case [see Delgado et al. (2011) for more details] but it does not have the disadvantage that the confidence of the rule $X \cup U \rightarrow \neg A$ is affected by an increment when the support of $X \cup Y$ is high.

It can be proven that $\text{Conf}_X(A \rightarrow B) = \text{Conf}(X \cap A \rightarrow B)$, but this is not true when using the certainty factor. This is due to the appearance of the consequent’s support in $D$ or $DX$ in the computation of certainty factor:

\[
\text{CF}(X \land \neg Y \rightarrow X \land A) \neq \text{CF}_X(\neg Y \rightarrow A)
\]

\[
\text{CF}(X \land A \rightarrow X \land \neg Y) \neq \text{CF}_X(A \rightarrow \neg Y)
\]

because

\[
\text{supp}(X \land A) = \frac{|X \cap A|}{|D|} \neq \frac{|X \cap A|}{|X|} = \text{supp}_X(A).
\]
5 Experimental evaluation

We have proposed a new approach using the fuzzy version of the certainty factor for mining anomalous rules. Mining anomalies associated to a strong rule extracts the unusual and anomalous pattern associated to the common one which is represented by the strong rule. Since this is a preliminary work, we are developing an efficient algorithm for mining such kind of rules. The two main problems are:

1. finding the fuzzy rules which algorithmically is more expensive than mining crisp rules
2. the fuzzy anomalous rules are infrequent, and therefore the pruning strategies used in algorithms like Apriori cannot be used here.

To overcome these problems we propose to use the algorithm developed in Delgado et al. (2003) for the discovery of fuzzy association rules, and then to adapt it for mining anomalous rules.

For the experimental evaluation of the approach, we have chosen a benchmark data set in the financial area. The data set German-Statlog can be found in the UCI Machine Learning and it concerns data about credits and the clients having a credit in a German bank. It is composed of 1,000 transactions and 21 attributes, from which 18 are categorical or numerical, and 3 of them are continuous. In this case, since the database is crisp, we employ a set of linguistic labels defined by fuzzy sets on the continuous domains in order to have a meaningful meaning to the user. In Figure 1 the fuzzy linguistic labels used in this database are shown. However we cannot forget the possibility that there exists data that can be fuzzy. We have also considered for the experimentation 10 equidistributed \( \alpha \)-cuts in the unit interval. We highlight here some of the extracted rules, that we think they are in some dataset sense remarkable. In general, once the rules are obtained, an expert should clarify which ones are really interesting or meaningful.

**Figure 1** Representation of fuzzy linguistic labels for the continues attributes in German-Statlog dataset (see online version for colours)
“IF Purpose = retraining
THEN duration is low ($Supp = 0.008$ & $CF = 0.781$)
OR duration is normal (unusually with $CF_1 = 1$, $CF_2 = 1$)”.

“IF Purpose = domestic-app
THEN existing credits paid back duly till now
($Supp = 0.01$ & $CF = 0.645$)
OR Status of existing checking account = 0DM
(unusually with $CF_1 = 1$, $CF_2 = 1$)”.

6 Conclusions

Mining anomalous rules can be useful in several domains. We have proposed to use fuzzy anomalous rules for representing the anomaly patterns that deviate from the common one, extracting together the usual pattern and the associated anomaly. Our approach is also sustained in using the certainty factor as an alternative to confidence, achieving a smaller and a more accurate set of anomalies. We also provide here a preliminary algorithm for mining such kind of rules, but it could be improved by parallelising the process in each $\alpha$-cut. This will remain as a future work. We have proved our approach in a database about credits, obtaining a manageable set of interesting rules that should be analysed by an expert.

There also exist interesting approaches that extract new kind of information using other types of infrequent rules, e.g., exception rules, that can be useful in several domains such as in the security field. Developing new approaches handling with imprecision or uncertainty for these new types of rules could be an interesting line for future research.

Acknowledgements

The research reported in this paper was partially supported by the Andalusian Government (Junta de Andalucía) under project P11-TIC-7460; from the Spanish Ministry for Science and Innovation by the project grants TIN2009-14538-C02-01, TIN2009-08296 and TIN2012-30939; and the project FP7-SEC-2012-312651, funded from the European Union in the Seventh Framework Programme [FP7/2007-2013] under grant agreement No 312651.

References


Anomaly detection using fuzzy association rules


