A Fuzzy Rule Mining Approach involving Absent Items*

M. Delgado[1] M.D. Ruiz[1] D. Sánchez[1,2] J.M. Serrano[3]

[1]Dept. Computer Science and A.I. University of Granada
[2]European Centre for Soft Computing
[3]Dept. of Computer Science. University of Jaén

Abstract

In this paper we present how to extract fuzzy association rules involving both the presence and the absence of items using a fuzzy rule mining procedure introduced by the authors in previous works. The rule mining procedure is based on the GUHA logical model, fuzzified via a recently proposed representation of gradualness. We present some results obtained with real datasets.

Keywords: Data mining, Fuzzy association rules, absence of items, negative rules.

1. Introduction

Fuzzy association rules have been developed as a powerful tool for obtaining meaningful and interesting information from databases. Fuzzy rules offer a good representation of gradual concepts, providing comprehensive results for the user. We have recently proposed a formal model for mining fuzzy rules [10] intended to preserve the main properties of crisp rules and to provide a formal framework for analyzing the accuracy measures and their properties. The formalization uses the basic notions of four fold table (4ft for short) and 4ft-quantifier which represents the measure for evaluating association rules, and the representation of fuzziness using levels [19]. The main idea is to apply a crisp mining method in every considered level from a fixed set and then to summarize the obtained results.

In the setting of association rule mining, several authors have motivated the task of mining for new and useful relations considering not only the presence but also the absence of items [5, 21, 22, 30]. The formalization of this idea has been made by means of a negation operator which represents the complement of the occurrence of an item in a transactional database, that is, its absence. Formally, in these situations the set of items I has both positive $(i_1, \ldots, i_m)$ and negative $(\neg i_1, \ldots, \neg i_m)$ items, where $\neg i_k$ means that i_k is not present in a transaction. Therefore, a *negative association rule* is a rule that contains a negative item (i.e. a rule for which either its antecedent and/or its consequent can contain a conjunction of both present and absent items). The rationale behind this is that a rule like $sugar \Rightarrow \neg sweetener$ (those purchasing sugar are not purchasing sweetener) in market baskets can be as useful as a rule like $burguer \Rightarrow ketchup$. We humans make use of this kind of rules in our everyday life for making decisions and/or for pure understanding of what's going on. The same can be claimed in the case of fuzzy rules.

In this paper we extend the fuzzy rule mining framework developed in [10] for mining fuzzy rules that also involve absent items. An algorithm following the model philosophy is presented and we perform several experiments in some real fuzzy databases, analyzing the performance of our proposal and comparing it with Han and Beheshti approach [13, 14], which is the one closer to ours, in the sense that it mines fuzzy rules involving negative items. Our objective is to show the feasibility of the approach.

The paper is organized as follows: Section 2 introduces some preliminary concepts about rule mining and reviews the existent approaches dealing with the absence of items from the crisp to the fuzzy case. In section 3 we include a brief overview of our proposal for mining fuzzy rules as well as different ways of representing the obtained results. Section 4 applies our approach for mining rules involving both present and absent items. It also contains the algorithmic implementation and the results obtained in some real fuzzy databases. We finish with some conclusions and possible lines for future research.

2. Background Concepts and Related Work

2.1. Association Rules

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 [1] are "implications" of the form $A \to B$ that relate the presence of itemsets A and B in transactions of D, assuming $A, B \subseteq I$, $A \cap B = \emptyset$ and $A, B \neq \emptyset$.

The ordinary measures proposed in [1] to assess association rules are *confidence* (the conditional probability $P(B|A)$) and *support* (the joint probability $P(A \cup B)$). An alternative framework was proposed in [4] where the accuracy is measured by

*The research reported in this paper was partially supported by the Andalusian Government (Junta de Andalucía) under project P07-TIC03175 and from the Spanish Ministry for Science and Innovation by the project grants TIN2006-15041-C04-01 and TIN2009-08296.

means of Shortliffe and Buchanan's certainty factors [23], defined as:

$$\frac{\text{Conf}(A \to B) - \text{supp}(B)}{1 - \text{supp}(B)}$$

when $\text{Conf}(A \to B) > \text{supp}(B)$, and

$$\frac{\text{Conf}(A \to B) - \text{supp}(B)}{\text{supp}(B)}$$

when $\text{Conf}(A \to B) < \text{supp}(B)$, and 0 otherwise. 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 [4, 6], like eliminating rules corresponding to statistical independence or negative dependence.

2.2. Fuzzy Association Rules

Different approaches and semantics have been proposed for fuzzy association rules, for instance [15, 6, 11]. In [6], 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 1. [6] 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[1] $\tilde{\tau}(i)$ 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 1 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 2. [6] Let I be a set of items, $\tilde{D}$ a set of fuzzy transactions and $A, B \subset I$ two disjoint itemsets, i.e. $A \cap B = \emptyset$. A fuzzy association rule $A \to B$ 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 $\tilde{D}$.

On the basis of this definition, the rule $A \Rightarrow B$ holds iff $A \subseteq \tilde{\tau}$ implies $B \subseteq \tilde{\tau}$. 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.

[1] For sake of simplicity we note $\tilde{\tau}(i)$ as $\mu_{\tilde{\tau}}(i)$ where $\mu_{\tilde{\tau}} : I \longrightarrow [0, 1]$ is the membership function associated to the fuzzy set $\tilde{\tau}$ on the referential $I = \{$ set of items $\}$.

2.3. Related Work dealing with Absent Items

We can distinguish several approaches for taking into account both the presence and the absence of items. We can classify them in two different groups. The first class contains those approaches that formalize the concept of absence by means of the complementary or the negation of items. In this case the set of items I contains both positive and negative items $(i_1, \neg i_1, \ldots, i_m, \neg i_m)$ where $\neg i_k$ means that i_k is not present in a transaction. The approaches in the second class are those that search a group of rules with a predefined meaning considering for that the absence of items too. In this class we can stress some important types of rules like exception rules (a set of three or two rules where the second one involves the absence of an item meaning that the second rule is an exception to the first one) [25] or anomalous rules [7] (a set of three rules meaning that the second rule represents an anomalous behavior that deviates from the usual one, represented by the first rule).

Some problems arise when considering the absence of items: the density of data is higher, the complexity may increase significantly in terms of the number of data items and some prune strategies used to restrict the search space and to guarantee the efficiency in classical AR mining algorithms cannot be used. This is due to the fact that the absence of items does not fulfil the upward closure property of frequent itemsets. For this reason, the authors have addressed the task of mining negative association rules using several perspectives.

Some of them obtain negative associations using a predefined taxonomy or graph-based structure, but this is not available in all cases. In this group we can find a novel approach presented by Savasere et al. [21, 22] where the domain knowledge is given in the form of a taxonomy which is then used to mine the negative associations, and the approach of Yuan et al. [30] which employs a hierarchical graph-structured taxonomy containing classification information about the similarity between items.

Another alternative is that of considering a measure stronger than confidence. These approaches obtain a smaller set of negative rules. Wu et al. [28] use a principle introduced by Piatetsky-Shapiro for discarding uninteresting rules, and with a minimum interestingness threshold for a better pruning of the frequent items generated. Yan et al. [29] mine positive and negative fuzzy rules using the support-confidence framework and they incorporate in the process a new measure called *degree of implication* for measuring the relative fraction of transactions which are not negative examples of the rule. Teng et al. [26] focus in searching substitution rules (which contain negative items) employing the chi-square and the negative correlation measures.

The last alternative that we have found in literature consists on considering only those negative

items whose positive is frequent in the mining process. With this imposition the set of candidates to be in a rule is pruned, and therefore the number of extracted rules is substantially reduced. To our knowledge this type of imposition was first used in [26] where the negative rules are mined among the set of frequent positive items, measuring their negative correlation. This idea has been also employed in [2] by Antonie et al. where the rho correlation measure is employed to divide the mining process in two parts: if two items X, Y are positively correlated, rules $X \to Y$ or $\neg X \to \neg Y$ can be found, and if X, Y are negatively correlated they search for rules $\neg X \to Y$ or $X \to \neg Y$.

Han and Beheshti formalized this idea in [13, 14] to mine fuzzy positive and negative rules by defining a valid negative rule $A \to \neg B$ such as the one that fulfils the following conditions[2]:

- $A \cup B = \emptyset$,
- $\text{supp}(A) \geq minsupp$,
- $\text{supp}(B) \geq minsupp$,
- $\text{Supp}(A \to \neg B) \geq minsupp$,
- $\text{Conf}(A \to \neg B) \geq minconf$.

Wang et al. [27] also propose some pruning strategies by defining two new measures based on the fulfilment or not of (1)∼(5). Nevertheless, in this work the authors propose an algorithm that uses X and Y and their negations directly, where I only contains positive items. In this case, the mined rules involve the disjunction of items, since for instance, if $X = i_1 \wedge i_2$ then $\neg X = \neg i_1 \vee \neg i_2$. We have to be careful in these situations where the disjunction of items appears, as the user may not be interested in such kind of rules.

3. Formal Model for Mining Fuzzy Rules

This section is devoted to present a summary of the formal model developed in [8, 10]. This model allows us to mine fuzzy rules in a straightforward way, extending the accuracy measures from the crisp case. Its formalization basically underlies in two concepts: the representation by levels (RL) associated to a fuzzy property (called *restriction level representation* in previous works [19] and closely related to gradual sets [12]) and the four fold table associated to the items A and B noted by $\mathcal{M} = 4ft(A, B, D)$, where D denotes the database and $X, Y \in I$ are two itemsets that could be pairs of the form $\langle attribute, value \rangle$.

A RL associated to a fuzzy property A in a universe X is defined as a pair (Λ_A, ρ_A) where $\Lambda_A = \{\alpha_1, \ldots, \alpha_m\}$ is a finite set of levels verifying that $1 = \alpha_1 > \cdots > \alpha_m > \alpha_{m+1} = 0$ and $\rho_A : \Lambda_A \to \mathcal{P}(X)$ is a function which applies each α_i into a crisp realization in this level [19]. The set of crisp representatives of A is the set $\Omega_A = \{\rho_A(\alpha) \,|\, \alpha \in \Lambda_A\}$. The values of Λ_A can be interpreted as values of possibility for a possibility measure defined for all $\rho_A(\alpha_i) \in \Omega_A$ as $Pos(\rho_A(\alpha_i)) = \alpha_i$. Following this interpretation we define the associated probability distribution $m : \Omega_A \to [0, 1]$ as in equation (1) which give us information about how representative is each crisp set of the property A in Ω_A.

$$m_A(Y) = \sum_{\alpha_i \,|\, Y = \rho(\alpha_i)} \alpha_i - \alpha_{i+1}. \qquad (1)$$

For each $Y \in \Omega_A$, the value $m_A(Y)$ represents the proportion to which the available evidence supports claim that the property A is represented by Y. From this point of view, a RL can be seen as a random set, i.e., a basic probability assignment in the sense of the theory of evidence on sets, *plus a structure indicating dependencies between the possible representations of different properties*.

The four fold table associated to the itemsets involved in a rule $A \to B$ is constructed from the database D by the number of transactions in D satisfying the four possible combinations between A and B using the logic connectors $\wedge$ (conjunction) and $\neg$ (negation). So, the four fold table is determined by four natural numbers $\mathcal{M} = 4ft(A, B, D) = \{a, b, c, d\}$ where a is the number of rows of D satisfying $A \wedge B$, b the number of rows satisfying $A \wedge \neg B$, c represents those satisfying $\neg A \wedge B$ and d those satisfying the last possibility $\neg A \wedge \neg B$ [18, 9]. Note that $|D| = a + b + c + d = n$ is the total number of transactions of D. The validity of an association rule is assessed by using M by means of the so-called 4ft-quantifier $\approx$ which is an operator (interestingness measure) that measures the strength of the association between itemsets A and B according to some validity criteria. In particular the known measures of support and confidence are 4ft-quantifiers defined as follows:

$$\text{Supp}(A \to B) = \approx_S (a, b, c, d) = \frac{a}{a + b + c + d}$$
$$\text{Conf}(A \to B) = \approx_C (a, b, c, d) = \frac{a}{a + b}.$$

Using these two models we have proposed [10] a framework that enables us to extend the interestingness measures for their validation from the crisp to the fuzzy case. Let us consider $\tilde{D}$ a fuzzy database (where an item is satisfied with a certain degree in the unit interval in each transaction), and A, B two itemsets in $\tilde{D}$. The main idea consists in considering a fixed set of levels and defining the associated four fold table for each level [10].

Let us consider the fuzzy sets associated to A and B as $\tilde{\Gamma}_A(\tilde{\tau}) = \tilde{\tau}(A)$ and $\tilde{\Gamma}_B(\tilde{\tau}) = \tilde{\tau}(B)$ respectively, where $\tilde{\tau}$ represents a transaction of $\tilde{D}$ and $\tilde{\tau}(A) = \min_{i \in A} \tilde{\tau}(i)$. The obtained fuzzy sets could be represented by $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$, $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$. Analogously to the crisp case, we consider the four possible conjunctions between A and B and their negations. For

[2] Definitions for the rules $\neg A \to B$ and $\neg A \to \neg B$ are analogous.

each level, $\rho_A(\alpha)$ is a crisp set, so we can compute its cardinality (the set of transactions that A has membership greater or equal than α) in the usual way. In the same manner we can proceed with the cardinality of $\rho_{A\wedge B}(\alpha), \ldots, \rho_{\neg A\wedge\neg B}(\alpha)$. These computations produce the frequencies a, b, c and d in each level, giving as a result a four fold table for each level $\mathcal{M}_{\alpha_i} = 4ft(\tilde{\Gamma}_A, \tilde{\Gamma}_B, \tilde{D}, \alpha_i) = (a_i, b_i, c_i, d_i)$ defined as follows [10]:

$\mathcal{M}_{\alpha_i}$	$\tilde{\Gamma}_B$	$\tilde{\Gamma}_{\neg B}$
$\tilde{\Gamma}_A$	a_i	b_i
$\tilde{\Gamma}_{\neg A}$	c_i	d_i

where a_i, b_i, c_i and d_i are non negative integers such that $a_i = |\rho_{\tilde{A}\wedge\tilde{B}}(\alpha_i)|$, $b_i = |\rho_{\tilde{A}\wedge\neg\tilde{B}}(\alpha_i)|$ and analogously with c_i and d_i. Note that $a_i + b_i + c_i + d_i = n = \left|\tilde{D}\right|$. Using $\mathcal{M}_{\alpha_i}$ and the probability distribution of equation (1) we extend the accuracy measures for fuzzy rules from the crisp case [10]:

$$\sum_{\alpha_i\in\Lambda_{\tilde{A}}\cup\Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\approx (a_i, b_i, c_i, d_i)\right). \quad (2)$$

In [10] we prove that this model is a good generalization of the crisp case. In particular we define the fuzzy support and confidence of a fuzzy rule as:

$$\text{FSupp}(A \to B) = \sum_{\alpha_i\in\Lambda_{\tilde{A}}\cup\Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\frac{a_i}{n}\right)$$

$$\text{FConf}(A \to B) = \sum_{\alpha_i\in\Lambda_{\tilde{A}}\cup\Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\frac{a_i}{a_i + b_i}\right).$$

Previous definition has the inconvenience of presenting indeterminations of the form "$\frac{0}{0}$" when $|\rho_{\tilde{A}}(\alpha_i)| = a_i + b_i = 0$. This happens when there does not exist transactions satisfying at the same time the antecedent and the consequent. So, to preserve the definition of fuzzy rule (see Definition 2) we will take the value 1 for that indetermination.

Several works have pointed out some drawbacks of the support/confidence framework to assess association rules (see for instance [24]). In general, the same problems occur when dealing with fuzzy rules. In this paper we shall employ the certainty factor for the validation of association rules [4].

$$\approx_{CF}(a, b, c, d) = \begin{cases} \dfrac{ad - bc}{(a + b)(b + d)} & \text{if } ad > bc \\ 0 & \text{if } ad = bc \\ \dfrac{ad - bc}{(a + b)(a + c)} & \text{if } ad < bc. \end{cases}$$

which is conveniently generalized to the fuzzy case, called $\text{FCF}(A \to B)$, as follows:

$$\sum_{\alpha_i\in\Lambda_{\tilde{A}}\cup\Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\approx_{CF} (a_i, b_i, c_i, d_i)\right) \quad (3)$$

where the representations of both itemsets must be normalized to fulfil the definition of fuzzy rule.

3.1. Representation of the results

The proposed formalization for fuzzy rules by levels allows the user to choose several options in order to manage the obtained results. In the previous section we have dealt with the standard option, i.e. we evaluate the accuracy of fuzzy rules by a crisp number given by the value of the associated $4ft$-quantifier (equation (2)). This approach is equal to achieve a crisp mining in each level and then summarize it by the accuracy measure of equation (2) extracting only those rules that satisfy the imposed thresholds for the fuzzy measure.

But other kinds of results could be obtained by considering a different perspective: (1) obtaining a crisp result for each level or only for the level of interest and (2) summarizing the results using a gradual number[3].

Let $\tilde{D}_1$ be the fuzzy database of Table 1. We can analyze rules at different levels, for example by centering our interests in those rules satisfying the thresholds for support and confidence/CF in levels greater than 0.6. If we look at Table 2 we can see that rule $\{i_1, i_3\} \to \{i_4\}$ does not have a high value in levels $0.6, 0.8$ and 1 so, it would not be interesting for the user. Rule $\{i_1, i_2\} \to \{i_3\}$ would be extracted at levels 0.6 and 0.8 but not at level 1.

	i_1	i_2	i_3	i_4	i_5
$\tilde{\tau}_1$	1	0.2	1	0.9	0.9
$\tilde{\tau}_2$	1	1	0.8	0	0
$\tilde{\tau}_3$	0.5	0.1	0.7	0.6	0
$\tilde{\tau}_4$	0.6	0	0	0.5	0.5
$\tilde{\tau}_5$	0.4	0.1	0.6	0	0
$\tilde{\tau}_6$	0	1	0	0	0

Table 1: Set of fuzzy transactions $\tilde{D}_1$

α	**Supp**	**Conf**	**CF**	**Supp**	**Conf**	**CF**
1	0	0	0	0	0	-1
0.8	0.166	0.5	0.4	0.166	1	1
0.6	0.166	0.5	0.25	0.166	1	1
0.4	0.5	0.75	0.5	0.166	1	1
0.2	0.5	0.75	0.5	0.333	1	1
	0.266	0.5	0.33	0.167	0.8	0.6

Table 2: Measurement for rules $\{i_1, i_3\} \to \{i_4\}$ and $\{i_1, i_2\} \to \{i_3\}$ respectively seen by levels, and their summary values.

For each rule, the assignment of support (resp. certainty factor) values to levels gives a gradual number. For a fuzzy rule $A \to B$:

$$\text{Supp}_{\alpha_1}(A \to B)/\alpha_1 + \ldots + \text{Supp}_{\alpha_m}(A \to B)/\alpha_m$$
$$\text{CF}_{\alpha_1}(A \to B)/\alpha_1 + \ldots + \text{CF}_{\alpha_m}(A \to B)/\alpha_m$$

[3] A gradual number is a level representation $(\Lambda, \mathcal{R})$ where Λ is a RL-set and $\mathcal{R} : (0, 1] \to \mathbb{R}$).

where $\text{Supp}_\alpha(A \to B)$ and $\text{CF}_\alpha(A \to B)$ represent respectively the support and the CF of $A \to B$ at level α. This option provides to the user the measurement values of each rule at each level.

Another interesting issue of our method is that we can summarize the obtained set of fuzzy association rules in terms of the random set view (evidence). We obtain all crisp rules in each level and then we reduce the number of existent rules by discarding those less interesting according to the associated thresholds at each level. So, the resulting set of association rules in terms of evidence could be showed by the following expression:

$$\begin{aligned} \mathit{Ruleset} = &\{i_1 \to i_3, i_2 \to i_4\}/0.6+ \\ &+ \{i_1 \to i_3, i_2 \to \neg i_4, i_1 \to \neg i_2\}/0.5+ \\ &+ \{i_1 \to i_2\}/0.4 \end{aligned} \quad (4)$$

where, for instance, (4) shows that rules $i_1 \to i_3$ and $i_2 \to i_4$ exceed the imposed thresholds with evidence 0.6. It could be also helpful for the user to rank the obtained rules at each level by the relevance measure used, for instance by certainty factor, obtaining in this case that rule $i_1 \to i_3$ is more certain that $i_2 \to i_4$ with evidence 0.6. And of course, if the user is interested in knowing the associated values for the accuracy measures they could be showed at the same time by using the following modification of the previous expression:

$$\begin{aligned} &\{i_1 \to i_3(0.2, 0.9), i_2 \to i_4(0.1, 0.8)\}/0.6+ \\ &+\{i_1 \to i_3(0.3, 0.9), i_2 \to \neg i_4(0.1, 0.6), \\ &i_1 \to i_2(0.15, 0.7)\}/0.5 + \{i_1 \to i_2(0.25, 0.9)\}/0.4 \end{aligned}$$

where the numbers in parentheses indicate in this case the values (support, CF).

Notice that the previous expressions are not actually fuzzy sets, but still can be very helpful in order to interpret the set of results, as we relate the relevance degree and the basic probability assignment for each rule or set of rules.

4. Application in Managing the Absence of Items

In most of works, the absence of items has been formalized by considering the negation operator $\neg$ which represents the complement of the occurrence of an item, that is, its absence. In the proposed model we considered the logic operators $\wedge$ and $\neg$ between items. In particular, the negation has an active role as it has been used for computing the four frequencies involved in the four fold tables $\mathcal{M}_\alpha$.

Let $\tilde{D}$ be a set of fuzzy transactions and $I = \{i_1, \ldots, i_m, \neg i_1, \ldots, \neg i_m\}$ be a set of items in $\tilde{D}$. We want to remark that when dealing with a conjunction of two or more items, being negative at least one of them, the subsequent level set may not correspond necessarily to a fuzzy set[4].

[4]We perform the operations in each level. In particular when considering the negation, the resulted representation may not correspond to a fuzzy set (see [19])

Employing our approach it is easy to prove that the obtained measures for negative items are the ones that we expect. Let $A, B \subset I$ be two itemsets with associated $4ft$-table for each level $\mathcal{M}_{\alpha_i} = 4ft(\tilde{\Gamma}_A, \tilde{\Gamma}_B, \tilde{D}) = \langle a_i, b_i, c_i, d_i \rangle$, then $4ft(\tilde{\Gamma}_{\neg A}, \tilde{\Gamma}_B, \tilde{D}) = \langle c_i, d_i, a_i, b_i \rangle$:

$\mathcal{M}_{\alpha_i}$	B	$\neg B$
A	a_i	b_i
$\neg A$	c_i	d_i

	B	$\neg B$
$\neg A$	c_i	d_i
A	a_i	b_i

$$\begin{aligned} \text{supp}_{\alpha_i}(\neg A) &= \frac{c_i + d_i}{n} = \frac{n - (a_i + b_i)}{n} \\ &= 1 - \text{supp}_{\alpha_i}(A) \end{aligned} \quad (5)$$

$$\text{Conf}_{\alpha_i}(c_i, d_i, a_i, b_i) = \frac{c_i}{c_i + d_i} = \text{Conf}_{\alpha_i}(\neg A \to B) \quad (6)$$

In consequence, it is not necessary to compute extra frequencies when the negation of items is involved.

Finally, contrary to the fuzzy sets case, it is not possible to find itemsets containing both items i and $\neg i$ whilst, at the same time, both items may appear in the representation of a transaction (but always in different levels).

4.1. Algorithm

Implementation is one of the crucial stages in the data mining area. Nowadays, research about new and fast algorithms for large databases is constantly growing. Our proposal consists in extracting fuzzy rules by means of parallelizing a particular crisp mining process. Then, any crisp mining algorithm could be used for mining fuzzy association rules by applying it in every level in a straightforward way. Afterwards, we summarize the obtained results by means of one of the several methods proposed in Section 3.1.

The proposed algorithm is a variation of Apriori, modified for dealing with a set of items represented by means of BitSets [17]. The general framework for mining fuzzy rules is described in algorithm 4.1 whose main steps are:

Steps 1.1 and 1.2 are conjunctly processed in order to read the database only once. For computing the frequencies of itemsets we use the *cardinality* function implemented in the BitSet java class. Then the frequencies in $\mathcal{M}_\alpha$ are computed using the cardinalities of the antecedent A, the consequent B and their conjunction $A \wedge B$ without increasing the computational cost of the algorithm:

$$\begin{aligned} &a = card(A \wedge B), \\ &b = (a + b) - a = card(A) - card(A \wedge B), \\ &c = (a + c) - a = card(B) - card(A \wedge B), \\ &d = (a + b + c + d) - a - b - c = |\tilde{D}| - a - b - c. \end{aligned}$$

In our experiments, we have considered support and certainty factor but other types of measures can also

Algorithm 1 : Mining Fuzzy Rules

Input: $\tilde{D}$, set of RLs, *minsupp*, *minCF*
Output: Set of fuzzy association rules.

1. Database Preprocessing.
1.1 Transform $\tilde{D}$ into k boolean databases (k being the number of RLs).
1.2 Store the database into k vectors of BitSets.
2. Mining Process.
For each RL
2.1 Search the set of candidates
If i_k is a frequent item, add it to candidates.
Store BitSet vector indexes of candidates (C)
2.2 Search frequent l-itemsets using C.
2.3 Form the rules using C
Compute their associated 4ft $\mathcal{M}_\alpha$.
Store rules and $\mathcal{M}_\alpha$ in a file.
3. Summarizing the obtained results.
3.1 Read all the found rules in every level α.
3.2 Compute the fuzzy measures using $\mathcal{M}_\alpha$.

be implemented immediately by means of the appropriate quantifiers representing the type of association the user wants to extract from the data.

The Algorithm 4.1 has three different parts. The overall complexity of our approach will depend on the complexity of each part. In the Apriori case the complexity of the second part is $O(n2^i)$, where n is the number of transactions and i the number of items (twice the number of positive items). Our algorithm repeats the mining process for each level. If we note with k the number of RLs, we have a theoretic complexity of $O(kn2^i)$ but we can reduce it to $O(n2^i)$ by a suitable parallelization, using one separate thread for each level. The third step is the most time consuming as it depends on the number of rules in each level.

Concerning space, the size of memory requirements for standard databases is acceptable. For a database with 61810 transactions and 33 items the memory occupied by the vector of BitSets is around 1 MB which must be multiplied by the number of levels considered.

4.2. Experiments and Results

The main problem of existing approaches is that the absence of an item tends to be always frequent as its occurrence usually has low support. Most of times the considered approach is that proposed by Han and Beheshti [14] which consists in taking only negative items whose positive is frequent [14].

In order to see if this condition is adequate or not we have taken into account two different approaches, one that only considers those items in $I' = \{i_1, \ldots, i_m, \neg i_1, \ldots, \neg i_m\}$ that are frequent and the other one following the approach developed in [14] where the negative items are chosen as candidates when its positive is frequent. For the former, we use the Algorithm 4.1 considering that $i_k \in I'$ and for the latter, we consider that $i_k \in I$ where $I = \{i_1, \ldots, i_k\}$ and we also impose in step 2.1 that if i_k is frequent, we add i_k and $\neg i_k$ to the set of candidates.

It is worth to mention that in the mining rule process a rule of the form $i_k \rightarrow \neg i_k$ cannot be found as $i_k \wedge \neg i_k$ is the null itemset and its cardinality is equal to zero and consequently its support too (remember that in a given level it is impossible that both i_k and $\neg i_k$ appear simultaneously).

In our experiments we took a fixed set of levels $\Lambda = \{1, 0.9, \ldots, 0.2, 0.1\}$ with just ten elements which is sufficient to obtain reliable fuzzy rules. For testing the performance of our algorithm we have carried out several experiments over the databases in table 3. The first one has been used in [20] to obtain information about olive agriculture in the south of Spain. The second database is Auto-mpg from the known UCI machine learning repository [3], where we have fuzzified continuous attributes using the following set of linguistic labels: *low, medium* and *high*. The *Forest Cov-type* databases are reduced versions originating from the database used in [16] where we have excluded the binary attributes and we have fuzzified the remaining attributes using two different methods: equi-depth intervals and the K-means clustering for reducing their numeric domains to linguistic ones employing the labels *low, medium, high*. The last database is a random choice of Soil database transactions in order to obtain a fuzzy database with a higher number of transactions.

Table 3: Databases employed in the experiments.

Fuzzy Database	Size	Pos. Items	Total Items
Soil	541	33	66
Auto-mpg	398	39	78
Forest Cov-type (equidepth)	581012	37	74
Forest Cov-type (k-means)	581012	37	74
Synthetic from Soil DB	99811	33	66

To illustrate the complexity in time we discussed before and to show the performance of both proposals, we have carried out an extensive battery of experiments over the mentioned databases in a 2.13GHz Intel Core i3 notebook with 4GB of main memory, running Windows 7.

We have run diverse experiments in order to study (1) the suitability of using confidence or certainty factor measures when mining negative rules, (2) to see the differences between the use of the imposition proposed by Han and Beheshti [13] for the confidence and the certainty factor, and (3) to study the time consumed in the third part of the algorithm in function on the number of mined rules.

We have executed the algorithm imposing only one item in both parts of the rule (antecedent and

	ms	$mc \backslash mcf$	rules	s	rules	s	rules	s
Conf WR	.01	0.8	1106	10	1408	13	1106	10
		0.9	1022	7	1402	12	1020	8
	.05	0.8	1083	10	1377	13	1083	10
		0.9	996	8	1371	16	994	8
CF WR	.01	0.8	26	1	9	1	31	1
		0.9	22	1	7	1	27	1
	.05	0.8	24	1	9	1	29	1
		0.9	21	1	7	1	26	1
Conf R	.01	0.8	72	1	13	1	72	1
		0.9	54	1	7	1	52	1
	.05	0.8	45	1	13	1	44	1
		0.9	28	1	7	1	26	1
CF R	.01	0.8	26	1	6	1	31	1
		0.9	22	1	4	1	27	1
	.05	0.8	24	1	6	1	29	1
		0.9	20	1	4	1	25	1

Table 4: Fuzzy rules obtained in `Soil` (left), `Auto-mpg` (middle) and `Synthetic Soil` (right).

Approach	ms	$mc \backslash mcf$	rules	sec	rules	sec
Conf without restriction	.01	0.8	1232	8	1297	9
		0.9	1228	8	1228	8
	.05	0.8	1232	8	1297	8
		0.9	1228	8	1228	9
CF without restriction	.01	0.8	4	1	8	1
		0.9	4	1	7	1
	.05	0.8	4	1	8	1
		0.9	4	1	7	1
Conf with restriction	.01	0.8	10	1	23	1
		0.9	6	1	6	1
	.05	0.8	10	1	23	1
		0.9	6	1	6	1
CF with restriction	.01	0.8	4	1	8	1
		0.9	4	1	7	1
	.05	0.8	4	1	8	1
		0.9	4	1	7	1

Table 5: Fuzzy rules obtained in `Forest equidepth` (left) and `Forest k-means` (right).

consequent), and restraining the confidence and the certainty factor in all databases to several threshold values as we show in Tables 4 and 5. We refer to the imposition proposed by Han and Beheshti [13] when we mention "with restriction (R)" in previous tables (WR, without restriction, is used for representing our approach). We observed a high decrease in the number of extracted rules when considering the certainty factor.

Relative to the second point, when carrying the same experiments imposing the condition proposed by Han and Beheshti [13], i.e. the absence of items must have their positive frequent, we found that this imposition reduces the total number of mined rules but it can leave behind some interesting rules, and sometimes by only imposing a measure stronger than confidence it is enough to obtain a manageable set of rules. Summarizing the results to answer the second point, we observed that it would be convenient to use some pruning strategy (like the one in [13]) in order to obtain a manageable set of rules in some cases, but in other cases it would be enough to use a strong accuracy measure with appropriate properties like the certainty factor.

Regarding the execution time of our approach, we observed that the most consuming time part is the third one. As it summarizes the results it depends

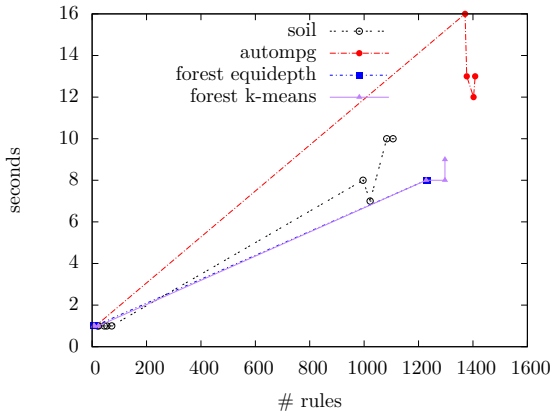

Figure 1: Relation between time (s) and number of extracted rules with different threshold impositions.

on the total number of extracted rules. In figure 1 we show some of the times resulted in our experimentation regarding the number of rules obtained in the summary. When the number of rules to be processed exceeds 100 000 in every level, the time is highly increased. As, in real world cases, the user is usually interested in obtaining a reduced set of rules easier to manage, we think that the obtained times are reasonable.

5. Conclusions and Future Works

Mining association rules involving absent items, as well as mining fuzzy association rules, are extensions of the association rule mining task that have been shown to be very interesting in the literature. We have proposed a novel approach for mining fuzzy association rules involving absent items, taking advantage of the representation by levels in that crisp properties and measures can be easily translated to the fuzzy case, and experimenting in both real and synthetic fuzzy databases. Our objective has been to show that the approach is feasible, and our first obtained results suggest that this is the case, specially in those cases where the user can guide the mining process, for example, by specifying the associations he is interested in. We have also compared our approach with imposition-based models, concluding that the more items appear in one of the rule sides, the more the imposition drastically reduces the number of extracted rules.

In this respect, we plan to consider stronger measures, different impositions or distinct methods to obtain a suitable set of rules involving the "negation" of items. Another promising line is to study a predefined group of rules involving the absence of items having significant semantics for the user. Some efforts have been already done when mining exceptions or anomalous rules [7]. Finally, we are already applying our techniques on databases with real data, in order to show that the approach is useful in practice.

References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining associations between sets of items in massive databases. In *ACM-SIGMOD*, pages 207–216, 1993.

[2] M.L. Antonie and O.R. Zaïane. Mining positive and negative association rules: An approach for confined rules. In *European PKDD Conference*, pages 27–38, 2004.

[3] A. Asuncion and D.J. Newman. UCI machine learning repository, 2007.

[4] F. Berzal, M. Delgado, D. Sánchez, and M.A. Vila. Measuring accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3):221–235, 2002.

[5] S. Brin, R. Motwani, J.D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Record*, 26(2):255–264, 1997.

[6] M. Delgado, N. Marín, D. Sánchez, and M.A. Vila. Fuzzy association rules: General model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2):214–225, 2003.

[7] M. Delgado, M.D. Ruiz, and D. Sánchez. New approaches for discovering exception and anomalous rules. *Int. J. of Uncert., Fuzziness and Knowledge-Based Sys., Submitted*.

[8] M. Delgado, M.D. Ruiz, and D. Sánchez. A restriction level approach for the representation and evaluation of fuzzy association rules. In *Proc. IFSA-EUSFLAT*, pages 1583–1588, Lisbon, Portugal, 2009.

[9] M. Delgado, M.D. Ruiz, and D. Sánchez. Studying interest measures for association rules through a logical model. *Int. J. Uncert., Fuzz. and Knowledge-Based Sys.*, 18(1):87–106, 2010.

[10] M. Delgado, M.D. Ruiz, D. Sánchez, and J.M. Serrano. A formal model for mining fuzzy rules using the RL representation theory. *Information Sciences, Submitted*, 2011.

[11] D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2):167–192, 2006.

[12] D. Dubois and H. Prade. Gradual elements in a fuzzy set. *Soft Computing*, 12:165–175, 2008.

[13] J. Han and M. Beheshti. Discovering both positive and negative fuzzy association rules in large transaction databases. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 10(3):287–294, 2006.

[14] J. Han and M. Beheshti. Mining fuzzy association rules: Interestingness measure and algorithm. In *IEEE International Conference on Granular Computing*, pages 659–662, 2006.

[15] Chan-Man Kuok, Ada Fu, and Man Hon Wong. Mining fuzzy association rules in databases. *SIGMOD Record*, 27(1):41–46, 1998.

[16] H. Liu, F.Hussain, C.L. Tan, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6:393–423, 2002.

[17] E. Louie and T.Y. Lin. Finding association rules using fast bit computation: Machine-oriented modeling. *LNAI*, 1932:486–494, 2000.

[18] J. Rauch and M. Šimunek. Mining for 4ft association rules. *Lecture Notes in Artificial Intelligence*, 1967:268–272, 2000.

[19] D. Sánchez, M. Delgado, M.A. Vila, and J. Chamorro-Martínez. On a level-based representation of fuzziness. *Fuzzy Sets and Systems, Submitted*.

[20] D. Sánchez, J.M. Serrano, M.A. Vila, M. Delgado, G. Calero, J. Sánchez, and V.M. Aranda. Building a fuzzy logic information network and a decision-support system for olive cultivation in Andalusia. *Spanish Journal of Agricultural Research*, 6:252–263, 2008.

[21] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st Conference on Very Large Databases*, pages 432–444, Zürich, Switzerland, 1995.

[22] A. Savasere, E. Omiecinski, and S. Navathe. Mining for strong negative associations in a large database of customer transactions. In *ICDE*, pages 494–502. IEEE Computer Society, 1998.

[23] E. Shortliffe and B. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379, 1975.

[24] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68, 1998.

[25] E. Suzuki. Discovering interesting exception rules with rule pair. In *Proc. Workshop on Advances in Inductive Rule Learning at PKDD-04*, pages 163–178, 2004.

[26] W. Teng, M. Hsieh, and M. Chen. On the mining of substitution rules for statistically dependent items. In *Proc. of the ICDM*, pages 442–449, 2002.

[27] H. Wang, X. Zhang, and G. Chen. Mining a complete set of both positive and negative association rules from large databases. In *Proc. PAKDD 2008 Lecture Notes in Artificial Intelligence, 5012*, pages 777–784, 2008.

[28] X. Wu, C. Zhang, and S. Zhang. Mining both positive and negative association rules. In *Proceedings of ICML*, pages 658–665, 2002.

[29] P. Yan, G. Chen, C. Cornelis, M. De Cock, and E. Kerre. Mining positive and negative fuzzy association rules. *Knowledge-Based Intelligent Information and Engineering Systems (KES'2004)*, 3213:270–276, 2004.

[30] X. Yuan, B. Buckles, Z. Yuan, and J. Zhang. Mining negative association rules. In *Proceedings of ISCC*, pages 623–629, 2002.