Measuring the accuracy and interest of association rules: A new framework

Fernando Berzal, Ignacio Blanco, Daniel Sánchez* and María-Amparo Vila
Department of Computer Science and Artificial Intelligence, University of Granada, E.T.S.I.I., Avda. Andalucia 38, 18071 Granada, Spain

Received 1 December 2001
Accepted 1 March 2002

Abstract. It has been pointed out that the usual framework to assess association rules, based on support and confidence as measures of importance and accuracy, has several drawbacks. In particular, the presence of items with very high support can lead to obtain many misleading rules, even in the order of 95% of the discovered rules in some of our experiments. In this paper we introduce a different framework, based on Shortliffe and Buchanan's certainty factors and the new concept of *very strong rules,* and we discuss some intuitive properties of the new framework. Both the theoretical properties and the experiments we have performed show that we can avoid the discovery of misleading rules, improving the manageability and quality of the results.

1. Introduction

Nowadays, the amount of data stored in databases increases in an impressive way. One of the main motivations for data storage is to have the opportunity to analyze them, in order to obtain useful knowledge. For this purpose, the area of Knowledge Discovery was born. Knowledge Discovery is concerned with finding novel, previously unknown and potentially useful knowledge in databases.

The main step of the knowledge discovery task is called data mining, and it is concerned with finding frequent patterns in data. One of the main problems in the field of Data Mining is how to assess the patterns that are found in data, such as association rules in T-sets [1]. We call T-set a set of transactions, where each transaction is a subset of items. Association rules are "implications" that relate the presence of items in the transactions of a T-set. More formally, given a set of items I and a T-set R on I, an association rule is an expression of the form $A \Rightarrow C$, with $A, C \subset I$, $A \cap C = \emptyset$, where A and C are called *antecedent* and *consequent* of the rule respectively.

The classical example of T-set is a set of market baskets, where each basket is a transaction that contains a subset of products (items). Rules extracted from market basket relate the presence of items in the same basket, for example "every basket that contains milk contains bread", noted $milk \Rightarrow bread$.

[*] Corresponding author. Tel.: +34 958 246397; Fax: +34 958 243317; E-mail: daniel@decsai.ugr.es.

1088-467X/02/$8.00

The usual measures to assess association rules are support and confidence, both based on the concept of support of an *itemset* (a subset of items). Given a set of items I and a T-set R on I, the support of an itemset $I_0 \subseteq I$ is

$$supp(I_0) = \frac{|\{\tau \in R \mid I_0 \subseteq \tau\}|}{|R|} \tag{1}$$

i.e., the probability that the itemset appears in a transaction of R. The support of the association rule $A \Rightarrow C$ in R is

$$Supp(A \Rightarrow C) = supp(A \cup C) \tag{2}$$

and its confidence is

$$Conf(A \Rightarrow C) = \frac{supp(A \cup C)}{supp(A)} = \frac{Supp(A \Rightarrow C)}{supp(A)}\ . \tag{3}$$

Support is the percentage of transactions where the rule holds. Confidence is the conditional probability of C with respect to A or, in other words, the relative cardinality of C with respect to A. The techniques for mining association rules attempt to discover rules whose support and confidence are greater than user-defined thresholds called *minsupp* and *minconf* respectively. These are called *strong rules*.

However, several authors have pointed out some drawbacks of this framework that lead to find many more rules than it should [3,6,9]. The following example is from ([3]): in the CENSUS database of 1990, the rule “past active duty in military ⇒ no service in Vietnam” has a very high confidence of 0.9. This rule suggests that knowing that a person served in military we should believe that he/she did not serve in Vietnam. However, the itemset “no service in Vietnam” has a support over 95%, so in fact the probability that a person did not serve in Vietnam *decreases* (from 95% to 90%) when we know he/she served in military, and hence the association is negative. Clearly, this rule is misleading.

In this paper we introduce a new framework to assess association rules in order to avoid to obtain misleading rules. In Section 2 we describe some drawbacks of the support/confidence framework. Section 3 contains some related work. Section 4 is devoted to describe our new proposal. Experiments and conclusions are summarized in Sections 5 and 6, respectively.

2. Drawbacks of the support/Confidence framework

2.1. Confidence

Confidence is an accuracy measure of a rule. In [5], Piatetsky-Shapiro suggested that any accuracy measure ACC should verify three specific properties in order to separate strong and weak rules (in the sense of assigning them high and low values respectively). The properties are the following:

P1 $ACC(A \Rightarrow C) = 0$ when $Supp(A \Rightarrow C) = supp(A)\ supp(C)$. This property claims that any accuracy measure must test the independence (though values other than 0 could be used, depending on the range of ACC).

P2 $ACC(A \Rightarrow C)$ monotonically increases with $Supp(A \Rightarrow C)$ when other parameters remain the same.

Table 1
(A) The T-set R_1. (B) Support of several itemsets in R_1

A

i_1	i_2	i_3	i_4
1	0	1	0
0	0	0	1
0	1	1	1
0	1	1	1
1	1	1	1
1	1	1	1

B

Itemset	Support
$\{i_1\}$	1/2
$\{i_2\}$	2/3
$\{i_1, i_2\}$	1/3

P3 $ACC(A \Rightarrow C)$ monotonically decreases with $supp(A)$ (or $supp(C)$) when other parameters remain the same.

Now we show that confidence does not verify all the properties:

Proposition 1. *Confidence does not verify the property* **P1***.*

Proof. Here is a counterexample: let $I_1 = \{i_1, i_2, i_3, i_4\}$ be a set of items, and let R_1 be the T-set on I_1 of table 1.A. Rows represent transactions, and columns represent items. A cell containing "1" means that the item/column is present in the transaction/row. Table 1B shows the support of three itemsets with items in I_1. Since $supp(\{i_1\})supp(\{i_2\}) = 1/3 = supp(\{i_1, i_2\})$, i_1 and i_2 are statistically independent and hence the confidence should be 0. However, $Conf(\{i_1\} \Rightarrow \{i_2\}) == \frac{1/3}{1/2} = 2/3 \neq 0$.

Proposition 2. *Confidence verifies the property* **P2***.*

Proof. Trivial regarding Eq. (3).

Proposition 3. *Confidence verifies the property* **P3** *only for* $supp(A)$*.*

Proof. It is easy to see for $supp(A)$ regarding Eq. (3). It is also trivial to see that **P3** does not hold with respect to $supp(C)$ since $supp(C)$ does not appear in Eq. (3), and $supp(A \cup C)$ and $supp(A)$ remain the same by the conditions of **P3**.

In summary, confidence is not able to detect statistical independence (**P1**) nor negative dependence between items (the examples in the introduction), because it does not take into account the support of the consequent.

2.2. Support

A common principle in association rule mining is "the greater the support, the better the itemset", but we think this is only true to some extent. Indeed, itemsets with very high support are a source of misleading rules because they appear in most of the transactions, and hence any itemset (despite its meaning) seems to be a good predictor of the presence of the high-support itemset.

An example is $\{i_3\}$ in Table 1A. It is easy to verify that any itemset involving only i_1 and i_2 is a perfect predictor of $\{i_3\}$ (any rule with $\{i_3\}$ in the consequent has total accuracy, that is, confidence is 1 for all such rules). Also, $Conf(\{i_4\} \Rightarrow \{i_3\}) = 0.8$, that is pretty high. But we cannot be sure that these associations hold in real world. In fact what holds most times is negative dependence or independence, as the examples in the introduction showed.

As we have seen, an accuracy measure verifying **P1 - P3** can solve the problem when $Conf(A \Rightarrow C) \leqslant supp(C)$ (i.e., negative dependence or independence). But when $supp(C)$ is very high and $Conf(A \Rightarrow C) > supp(C)$, we can obtain a high accuracy. However, there is a lack of variability in the presence of C in data that does not allow us to be sure about the rule. Fortunately, this situation can be detected by checking that $supp(C)$ is not very high, but no method to check this has been incorporated into the existing techniques to find association rules.

These problems lead to obtain much more rules than it should. Suppose we have a T-set R on a set of items I, from where a set of reliable rules S has been obtained. Think of adding an item i_{vf} to I and to include it in the transactions of R so that i_{vf} has a very high support. It is very likely that adding i_{vf} to the consequent of any rule, both support and accuracy of the rules don't change. The same can be expected for support if we add the item to the antecedent, so we can obtain in the order of three times more rules (the original set, and those obtained by adding i_{vf} to the antecedent, or to the consequent). But we must also consider that since i_{vf} is very frequent, almost any itemset could be a good predictor of the presence of i_{vf} in a transaction, so we may obtain in the order of $2^{|I|}$ more rules. For example, if $|I| = 10$ (without i_{vf}) and $|S| = 50$ (a modest case), by adding i_{vf} we could obtain 1124 misleading rules in the worst case! Even if we restrict ourselves to find rules with only one item in the consequent, we are talking about 1074 rules in the worst case.

The problem is clearly that the user is overwhelmed with a big amount of misleading rules. The situation gets worse exponentially if we add two or more items with very high support. Now think of mining a real database such as the CENSUS data employed in [3], where $|I| = 2166$ and there are many items with support above 95%. It is clear that a new framework to assess association rules is needed.

3. Related work

Several authors have proposed alternatives to confidence, see [5,3,9,10,8] among others. In this section we briefly describe two of them.

3.1. Conviction

Conviction was introduced in [3] to be

$$Conv(A \Rightarrow C) = \frac{supp(A)\ supp(\neg C)}{supp(A \cup \neg C)} \tag{4}$$

where $\neg C$ means the absence of C. Its domain is $(0, \infty)$, 1 meaning independence. Values in $(0, 1)$ mean negative dependence. In our opinion, the main drawback of this measure is that its range is not bounded, so it is not easy to compare the conviction of rules because differences between them are not meaningful and, much more important, it is difficult to define a conviction threshold (for example, some rules considered interesting in [3] have conviction values of 1.28, 2.94, 50 and ∞, this last meaning total accuracy). Also, from (4) it is easy to see that conviction does not verify property **P3** for supp(A).

3.2. Interest

In [9], the χ^2 test is used to find dependencies between items. However, the value of the χ^2 statistic is not suitable to measure the degree of dependence, so interest is used instead. The interest is defined as

$$Int(A \Rightarrow C) = \frac{Supp(A \Rightarrow C)}{supp(A)\ supp(C)} \tag{5}$$

Interest verifies **P1–P3**, the value 1 meaning independence. But as conviction, its range is not bounded so it has the same drawbacks. Moreover, interest is symmetric (i.e. the interest of $A \Rightarrow C$ and $C \Rightarrow A$ is the same), and this is not intuitive in most of the cases. Association rules require to measure the strength of implication in both directions, not only the degree of dependence.

4. A new framework to assess association rules

4.1. Measuring accuracy

To assess the accuracy of association rules we use Shortliffe and Buchanan's *certainty factors* [7] instead of confidence. Certainty factors were developed to represent uncertainty in the rules of the MICYN expert system, and they have been recognized as one of the best models in the development of rule-based expert systems (however, they have been also used in data mining [4,6]).

Definition 1. *We name certainty factor of $A \Rightarrow C$ to the value*

$$CF(A \Rightarrow C) = \frac{Conf(A \Rightarrow C) - supp(C)}{1 - supp(C)} \tag{6}$$

if $Conf(A \Rightarrow C) > supp(C)$, and

$$CF(A \Rightarrow C) = \frac{Conf(A \Rightarrow C) - supp(C)}{supp(C)} \tag{7}$$

if $Conf(A \Rightarrow C) < supp(C)$, and 0 otherwise.

The certainty factor is interpreted as a measure of *variation* of the probability that C is in a transaction when we consider only those transactions where A is. More specifically, a positive CF measures the decrease of the probability that C is not in a transaction, given that A is. A similar interpretation can be done for negative CFs.

By Eqs (6) and (7) it is clear that CFs take into account both the confidence of the rule and the support of C. Moreover, they verify properties **P1–P3**, as the following propositions show:

Proposition 4. *Certainty factors verify* **P1**.

Proof. If $Supp(A \Rightarrow C) = supp(A)\ supp(C)$ then $Conf(A \Rightarrow C) = supp(C)$ and then by definition $CF(A \Rightarrow C) = 0$.

By this property, CFs are an independence test. But CFs can also detect the kind of dependence. If there is a positive dependence between A and C then $Supp(A \Rightarrow C) > supp(A)\ supp(C)$, so $Conf(A \Rightarrow C) > supp(C)$ and hence $CF(A \Rightarrow C) > 0$. If there is a negative dependence then $Conf(A \Rightarrow C) < supp(C)$ and hence $CF(A \Rightarrow C) < 0$.

Proposition 5. *Certainty factors verify* **P2**.

Proof. Confidence verifies **P2** and, when confidence increases and $supp(C)$ remains the same, CF increases (see Eqs (6) and (7)). Hence, CF increases with $Supp(A \Rightarrow C)$ when other parameters remain the same.

Proposition 6. *Certainty factors verify* **P3**.

Proof. CF verifies **P3** for $supp(A)$ since confidence does (see the proof of **P2**). Now we shall prove **P3** for $supp(C)$.

- Suppose $CF(A \Rightarrow C) < 0$. Then, by (7) it is clear that if $supp(C)$ increases then $CF(A \Rightarrow C)$ decreases.
- Suppose $CF(A \Rightarrow C) > 0$. CF is a function of confidence and $supp(C)$. By the conditions of **P3** we assume that confidence remains the same (i.e., it is a constant). If we derive with respect to $supp(C)$ we obtain

  $$CF'(A \Rightarrow C) = \frac{Conf(A \Rightarrow C) - 1}{(1 - supp(C))^2}$$

  so $CF'(A \Rightarrow C) \leqslant 0$, and hence $CF(A \Rightarrow C)$ monotonically decreases with $supp(C)$.
- Let $Conf(A \Rightarrow C) = c_0$ with $0 < c_0 < 1$. Let us increase monotonically $supp(C)$ from 0 to 1. While $supp(C) < c_0$ it holds that $CF(A \Rightarrow C) > 0$ and monotonically decreases, as we have shown. When $supp(C)$ reaches c_0 then $CF(A \Rightarrow C) = 0$, so it keeps decreasing. Finally, when $supp(C) > c_0$ it holds that $CF(A \Rightarrow C) < 0$, so it has decreased, and it keeps decreasing as $supp(C)$ increases, as we have shown.

Hence, CF monotonically decreases with $supp(C)$ when other parameters remain the same.

Other interesting properties of CFs are the following:

Proposition 7.

- $CF(A \Rightarrow C) \leqslant Conf(A \Rightarrow C)$
- $CF(A \Rightarrow C) = Conf(A \Rightarrow C)$ *iff* $CF(A \Rightarrow C) = 1$ *and* $supp(C) < 1$.

Proof.

- If $CF(A \Rightarrow C) \leqslant 0$ then $CF(A \Rightarrow C) \leqslant Conf(A \Rightarrow C)$. If $CF(A \Rightarrow C) > 0$ then

  $$CF(A \Rightarrow C) = \frac{Conf(A \Rightarrow C) - supp(C)}{1 - supp(C)}$$

  Hence,

  $$Conf(A \Rightarrow C) = CF(A \Rightarrow C)(1 - supp(C)) + supp(C) =$$

  $$= CF(A \Rightarrow C) + (supp(C) - CF(A \Rightarrow C)supp(C)).$$

  Since $CF(A \Rightarrow C) \in (0,1]$, we obtain $CF(A \Rightarrow C)supp(C) \leqslant supp(C)$, so $(supp(C) - CF(A \Rightarrow C)supp(C)) > 0$ and

  $$Conf(A \Rightarrow C) = CF(A \Rightarrow C) + (supp(C) - CF(A \Rightarrow C)supp(C)) \geqslant$$

  $$\geqslant CF(A \Rightarrow C)$$

- Let $supp(C) < 1$. Then $CF(A \Rightarrow C) = Conf(A \Rightarrow C)$ iff $(supp(C) - CF(A \Rightarrow C)supp(C)) = 0$ iff $CF(A \Rightarrow C) = 1$.

Hence, using the same threshold, we shall obtain always less rules using CFs than using confidence. Indeed, we are avoiding the discovery of misleading rules that are not detected by confidence. At the same time, perfect rules are recognized by both measures. A particular case is $supp(C) = 1$, where confidence finds a perfect rule but CF doesn't, because there is independence between A and C, so $CF(A \Rightarrow C) = 0$.

The following propositions relate CFs to conviction and interest.

Proposition 8. *Let $CF(A \Rightarrow C) > 0$ and $supp(C) < 1$ and $supp(A) > 0$. Then*

$$CF(A \Rightarrow C) = 1 - \frac{1}{Conv(A \Rightarrow C)} \tag{8}$$

Proof.

$$Conv(A \Rightarrow C) = \frac{supp(A)supp(\neg C)}{supp(A \cup \neg C)} = \frac{supp(A)(1 - supp(C))}{supp(A) - supp(A \cup C)}$$

Hence

$$\frac{1}{Conv(A \Rightarrow C)} = \frac{supp(A) - supp(A \cup C)}{supp(A)(1 - supp(C))} = \frac{1 - Conf(A \Rightarrow C)}{1 - supp(C)}$$

and

$$\frac{-1}{Conv(A \Rightarrow C)} = \frac{Conf(A \Rightarrow C) - 1}{1 - supp(C)} = \frac{Conf(A \Rightarrow C) - supp(C) + supp(C) - 1}{1 - supp(C)}$$
$$= CF(A \Rightarrow C) + \frac{supp(C) - 1}{1 - supp(C)}$$

Then

$$\frac{1}{Conv(A \Rightarrow C)} = -CF(A \Rightarrow C) - \frac{supp(C) - 1}{1 - supp(C)} = 1 - CF(A \Rightarrow C)$$

Thus

$$CF(A \Rightarrow C) = 1 - \frac{1}{Conv(A \Rightarrow C)}$$

Proposition 9. *Let $CF(A \Rightarrow C) < 0$ and $supp(C) > 0$. Then*

$$CF(A \Rightarrow C) = Int(A \Rightarrow C) - 1 \tag{9}$$

Proof.

$$Int(A \Rightarrow C) = \frac{Supp(A \Rightarrow C)}{supp(A)supp(C)} = \frac{Conf(A \Rightarrow C)}{supp(C)} = \frac{Conf(A \Rightarrow C) - supp(C) + supp(C)}{supp(C)}$$
$$= CF(A \Rightarrow C) + 1$$

Hence

$$CF(A \Rightarrow C) = Int(A \Rightarrow C) - 1$$

Corollary 1. *Let* $CF(A \Rightarrow C) < 0$*. Then* $CF(A \Rightarrow C) = CF(C \Rightarrow A)$

Corollary 2. $CF(A \Rightarrow C)CF(C \Rightarrow A) > 0$

From these properties, it is immediate that negative certainty factors are a non-directional measure of the strength of negative dependence, while positive certainty factors take into account the direction of the association.

From now on, we will call a rule *strong* if its support and CF are greater than user-specified thresholds *minsupp* and *minCF* respectively. Let us remark that we are interested only in rules with positive CF, meaning positive dependence among items, and hence we shall assume $minCF > 0$.

Of course, it could be argued that negative associations with a high enough absolute value of CF are strong rules, since they relate the presence of A to the *absence* of C and that can be interesting, see [9]. However, if we are interested in that kind of rules it is better to consider X and $\neg X$ as itemsets in the search since otherwise, rules relating for example the absence of A to the presence of C (also interesting) cannot be discovered.

4.2. Solving the support drawback

A simple solution would be to use a maximum support threshold *maxsupp* to solve the support drawback, and to avoid reporting those rules involving itemsets with support above *maxsupp*. However, the user should provide the value for *maxsupp*. In order to avoid this we introduce the concept of *very strong rule*.

Definition 2. *The rule* $A \Rightarrow C$ *is very strong if both* $A \Rightarrow C$ *and* $\neg C \Rightarrow \neg A$ *are strong rules.*

The rationale behind this definition is that $A \Rightarrow C$ and $\neg C \Rightarrow \neg A$ are logically equivalent, so we should look for strong evidence of both rules to believe that they are interesting. This definition can help us to solve the support drawback since when $supp(C)$ (or $supp(A)$) is very high, $Supp(\neg C \Rightarrow \neg A)$ is very low, and hence the rule $\neg C \Rightarrow \neg A$ won't be strong and $A \Rightarrow C$ won't be very strong.

By definition, a very strong rule must verify:

1. Support conditions:
   (a) $Supp(A \Rightarrow C) > minsupp$
   (b) $Supp(\neg C \Rightarrow \neg A) > minsupp$
2. CF conditions:
   (a) $CF(A \Rightarrow C) > minCF$
   (b) $CF(\neg C \Rightarrow \neg A) > minCF$

So there are two new conditions for a rule to be interesting, 1(b) and 2(b). But, in practice, only one CF condition must be checked, as a result of the following proposition:

Proposition 10. *If* $CF(A \Rightarrow C) > 0$ *then* $CF(A \Rightarrow C) = CF(\neg C \Rightarrow \neg A)$.

Proof. We shall use the usual probability notation, i.e., $Conf(X \Rightarrow Y) = p(Y|X)$ and $supp(X) = p(X)$. By Bayes' Rule

$$p(A|\neg C) = \frac{p(\neg C|A)p(A)}{p(\neg C)} = \frac{(1 - p(C|A))\,p(A)}{p(\neg C)}$$

and

$$p(\neg A|\neg C) = 1 - p(A|\neg C) = 1 - \frac{(1 - p(C|A))\, p(A)}{p(\neg C)}$$

Let us use Eq. (6) to obtain $CF(\neg C \Rightarrow \neg A)$. If we obtain a positive value, that will be the CF of the rule, since in that case $Conf(\neg C \Rightarrow \neg A) > supp(\neg A)$ (i.e., $p(\neg A|\neg C) > p(\neg A)$). Otherwise, we should have used Eq. (7).

$$\begin{aligned} CF(\neg C \Rightarrow \neg A) &= \frac{p(\neg A|\neg C) - p(\neg A)}{1 - p(\neg A)} = \frac{1 - \frac{(1-p(C|A))p(A)}{p(\neg C)} - (1 - p(A))}{p(A)} \\ &= \frac{p(A) - \frac{(1-p(C|A))p(A)}{p(\neg C)}}{p(A)} = 1 - \frac{1 - p(C|A)}{p(\neg C)} = \frac{(1 - p(C)) - (1 - p(C|A))}{(1 - p(C))} \\ &= \frac{p(C|A) - p(C)}{1 - p(C)} = CF(A \Rightarrow C). \end{aligned}$$

Since $CF(A \Rightarrow C) > 0$, we have used the correct expression and we have shown that $CF(\neg C \Rightarrow \neg A) = CF(A \Rightarrow C)$.

This property is not only useful, but also intuitive. Conviction also verifies it, but confidence and interest don't, as the following propositions show:

Proposition 11. *Conviction verifies proposition 10.*

Proof. Immediate since conviction is related to positive CF by proposition 8.

Proposition 12. *Confidence does not verify proposition 10.*

Proof. Immediate since confidence does not take into account the probability of C.

Proposition 13. *Interest does not verify proposition 10.*

Proof. Immediate since interest is related to negative CF by proposition 9 and, in general, negative CFs don't verify the property.

Another interesting property is the following

Proposition 14.

- *Let $supp(A) + supp(C) > 1$. Then $A \Rightarrow C$ is very strong iff $A \Rightarrow C$ is strong.*
- *Let $supp(A) + supp(C) < 1$. Then $A \Rightarrow C$ is very strong iff $\neg C \Rightarrow \neg A$ is strong.*
- *Let $supp(A) + supp(C) = 1$. Then $A \Rightarrow C$ is strong iff $\neg C \Rightarrow \neg A$ is strong.*

Proof. As we have shown, the certainty factor of a rule and its counter-reciprocal is the same when $mincf \geqslant 0$. With respect to support, it is easy to verify that

$$Supp(\neg C \Rightarrow \neg A) = 1 - supp(C) - supp(A) + Supp(A \Rightarrow C) \tag{10}$$

Hence,

- If $supp(A) + supp(C) > 1$ then $Supp(\neg C \Rightarrow \neg A) > Supp(A \Rightarrow C)$, so $Supp(A \Rightarrow C) > minsupp$ implies $Supp(\neg C \Rightarrow \neg A) > minsupp$.
- If $supp(A) + supp(C) < 1$ then $Supp(\neg C \Rightarrow \neg A) < Supp(A \Rightarrow C)$, so $Supp(\neg C \Rightarrow \neg A) > minsupp$ implies $Supp(A \Rightarrow C) > minsupp$.
- If $supp(A) + supp(C) < 1$ then $Supp(\neg C \Rightarrow \neg A) = Supp(A \Rightarrow C)$, so $Supp(\neg C \Rightarrow \neg A) > minsupp$ iff $Supp(A \Rightarrow C) > minsupp$.

4.3. Implementation

One of the advantages of our new framework is that it is easy to incorporate it into existing algorithms. Most of them work in two steps:

Step 1. Find the itemsets whose support is greater than *minsupp* (called *frequent itemsets*). This step is the most computationally expensive.

Step 2. Obtain rules with accuracy greater than a given threshold from the frequent itemsets obtained in step 1, specifically the rule $A \Rightarrow C$ is obtained from the itemsets $A \cup C$ and A.

To find very strong rules, step 1 remains the same. In step 2 we obtain the CF of the rule from the rule confidence and $supp(C)$, both calculated in step 1 (since $A \cup C$ is frequent, A and C also are), and we verify the CF condition. Support condition 1(a) is ensured because $A \cup C$ is a frequent itemset. Support condition 1(b) is also easy to verify since

$$supp(\neg C \cup \neg A) = 1 - supp(C) - supp(A) + supp(A \cup C) \tag{11}$$

and $supp(C)$, $supp(A)$ and $supp(A \cup C)$ are available. An important feature of these modifications is that they keep both the time and space complexity of the algorithms.

Finally, let us remark that support condition 1(a) is usually employed to bound the search for frequent itemsets in step 1 (hence reducing both time and space employed). Further reduction can be obtained by also using 1(b). This can benefit from the following property:

Proposition 15. *If* $supp(A \cup C) > 1 - minsupp$ *then* $supp(\neg C \cup \neg A) < minsupp$.

Proof. If $supp(A \cup C) > 1 - minsupp$ then $1 - supp(A \cup C) < minsupp$. Clearly $supp(A \cup C) + supp(\neg C \cup \neg A) \leqslant 1$, so $supp(\neg C \cup \neg A) \leqslant 1 - supp(A \cup C) < minsupp$.

The last proposition also suggests very strong rules implicitly use a value *maxsupp* of at most 1-*minsupp*. A description about how to use 1.b. to use this to reduce time and space expended in the search can be found in [2].

5. Experiments

5.1. Experiments with the CENSUS database

To illustrate the problems we have discussed, and to show the performance of our proposals, we have performed some experiments with the CENSUS database. The database we have worked with was

Table 2
Some attributes from the CENSUS database

Name	Description
AMARITL	Marital status
AHGA	Education
ACLSWKR	Class of worker
AHSCOL	Enrrolled in edu. inst. last wk.
ARACE	Race
ASEX	Sex
PENATVTY	Country of birth

Table 3
Some rules obtained from the CENSUS database

#R	Rule	Conf.	CF
1	[AHGA = Children] ⇒ [ASEX = Male]	0.5	0.05
2	[AHGA = High School Graduate] ⇒ [ASEX = Female]	0.54	0.06
3	[ASEX = Male] ⇒ [ARACE = White]	0.84	0.03
4	[ASEX = Female] ⇒ [ARACE = White]	0.83	0
5	[AHGA = Children] ⇒ [ACLSWKR = Not in universe]	1	1
6	[AHGA = Children] ⇒ [AMARITL = Never married]	0.99	0.99

extracted by T. Lane and R. Kohavi using the Data Extraction System from the census bureau database, found at http://www.census.gov/ftp/pub/DES/www/welcome.html.

Specifically, we have worked with a test database containing 99762 instances, obtained from the original database by using MineSet's MIndUtil mineset-to-mlc utility.

The database contains 40 attributes, but we have employed only those in Table 2. Let us remark that in relational databases, the usual interpretation is that items take the form [attribute=value].

5.1.1. *A first experiment*

In a first step, we looked for rules involving items associated to all the attributes except PENATVTY. Also, we restricted the search to rules with only one item in both antecedent and consequent. Using $minsupp = 0.05$ and $minconf = 0.5$ we obtained 52 rules. Some of them are shown in Table 3.

Rules 1 and 2 are clearly misleading, and in fact the certainty factor of both rules is close to 0, meaning independence between antecedent and consequent. But their confidence is not low, as it might be. The reason is that, assuming independence between education and sex, the expected conditional probability is 0.5 because sex takes only two values and they are approximately equally distributed on data. Hence, only performing a comparison between the conditional and the a-priori probabilities of the consequent (as certainty factors do) we can detect that the accuracy of these rules is uninteresting.

Rules 3 and 4 are also misleading, and in this case the problem is the very high support of the item [ARACE = White] in the database (around 84%). Hence, almost any other item seems to be a good predictor of the presence of white people (there is 12 misleading rules of this kind). For rules 3 and 4 certainty factors provide a value around 0 meaning independence, that is, the intuitively correct result again. No rule with [ARACE = White] in the consequent is reported when using mincf = 0.5. From these rules, the highest certainty factor is 0.31, that of rule

[AMARITL = Married-civilian spouse present] ⇒ [ARACE = White]

Finally, rules 5 and 6 are expected rules. The first one claims that children are not involved in any class of work, and the second one states that children have never been married. As such, their confidence is close to 1, and we can see that the value of certainty factor is also close to 1. That shows how certainty factors are suitable to detect rules with very high accuracy.

Table 4
Some rules obtained from the CENSUS database (II)

#R	Rule	Conf.	CF	Sup.	Rec.
5	[AHGA = Children] ⇒ [ACLSWKR = Not in universe]	1	1	0.23	0.498
6	[AHGA = Children] ⇒ [AMARITL = Never married]	0.99	0.99	0.23	0.567
7	[AHGA = Children] ⇒ [AHSCOL = Not in universe]	0.99	0.99	0.23	0.063
8	[AMARITL = Married-c.sp.pres.] ⇒ [AHSCOL = Not in u.]	0.99	0.92	0.41	0.061
9	[AMARITL = Widowed] ⇒ [AHSCOL = Not in universe]	0.99	0.99	0.05	0.063
10	[AMARITL = Divorced] ⇒ [AHSCOL = Not in universe]	0.99	0.96	0.06	0.063
11	[AHGA = Bachelors degree] ⇒ [AHSCOL = Not in universe]	0.98	0.72	0.09	0.061
12	[AHGA = High School G.] ⇒ [AHSCOL = Not in universe]	0.98	0.71	0.23	0.059

Table 5
Some rules obtained from the CENSUS database (III)

#R	Rule	Conf.	CF
13	[AHGA = Children] ⇒ [PENATVTY = USA]	0.95	0.62
14	[ACLSWKR = Private] ⇒ [PENATVTY = USA]	0.85	−0.03
15	[ACLSWKR = Not in universe] ⇒ [PENATVTY = USA]	0.9	0.13
16	[PENATVTY = USA] ⇒ [ACLSWKR = Not in universe]	0.51	0.01
17	[ARACE = Black] ⇒ [PENATVTY = USA]	0.9	0.41
18	[ARACE = White] ⇒ [PENATVTY = USA]	0.9	0.17
19	[PENATVTY = USA] ⇒ [ARACE = White]	0.85	0.11
20	[AMARITL = Never married] ⇒ [PENATVTY = USA]	0.91	0.29
21	[AMARITL = Married-c.sp.present] ⇒ [PENATVTY = USA]	0.86	−0.02
22	[AMARITL = Divorced] ⇒ [PENATVTY = USA]	0.9	0.13
23	[AHGA = Bachelors degree] ⇒ [PENATVTY = USA]	0.87	−0.01
24	[AHGA = High School G.] ⇒ [PENATVTY = USA]	0.89	0.07
25	[AHGA = Some college but not degree] ⇒ [PENATVTY = USA]	0.9	0.17
26	[ASEX = Male] ⇒ [PENATVTY = USA]	0.88	0.01
27	[PENATVTY = USA] ⇒ [ASEX = Female]	0.51	0
28	[ASEX = Female] ⇒ [PENATVTY = USA]	0.88	0
29	[PENATVTY = USA] ⇒ [AHSCOL = Not in universe]	0.93	0
30	[AHSCOL = Not in universe] ⇒ [PENATVTY = USA]	0.88	0

The same experiment was repeated but using mincf = 0.5 instead of confidence. In this occasion, we obtained only 8 rules. They are shown in Table 4. The last column shows the support of the counter-reciprocal rule $\neg C \Rightarrow \neg A$.

The first two of them were rules 5 and 6, already shown in Table 3. They are indeed very strong rules for minsupp $\leqslant$ 0.23 (remark that the maximum support for a very strong rule is 0.5), so they are highly reliable.

The rest of the rules seems to be obtained because of the very high support of the item [AHSCOL = Not in universe] (around 93%). But it is not very clear to us whether someone is interesting or not.

Though the support of the consequent is very high, certainty factors don't find rules 7–12 uninteresting since the confidence is greater than the support of the consequent (see Section 2.2). Except rules 11 and 12, the rest are rules with a very high accuracy, almost perfect. However, it should be noted that these rules are not very strong rules when minsupp $\geqslant$ 0.07, so we are able to detect that the support of the consequent is very high, and hence we have a criterion that can help to make a decision about the importance of the rules.

In summary, using minsupp = 0.07 and looking for very strong rules with certainty factors, only rules 5 and 6 are found when mincf = 0.5, instead of around 50 using confidence with minconf = 0.5. Clearly, we are avoiding the discovery of misleading rules.

Table 6
Number of rules obtained by using confidence and CF

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.98	1
minconf $= \alpha$	1185	1078	1066	1034	971	839	706	494	96	0
mincf $= \alpha$	795	633	549	431	304	190	104	96	20	0

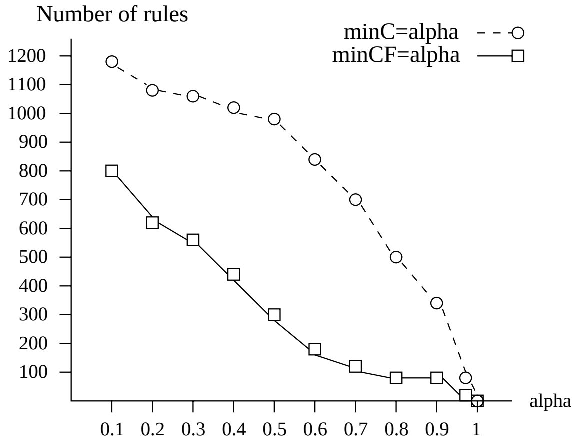

Fig. 1. Graphical representation of Table 6

5.1.2. Adding PENATVTY

We have repeated the experiment adding the items associated to PENATVTY. The values of the attribute PENATVTY are the different countries, but the distribution of the values gives [PENATVTY $=$ USA] a very high support of around 88%. As a consequence is again the case that, using confidence, almost any other item seems to be a good predictor that a person was born in USA. In our experiments, we obtained 70 rules (52 where those obtained in the previous experiment). Table 5 shows the 18 new rules obtained.

Among rules in Table 5, apart from rule 13, only rule 17 has a rather remarkable cf of 0.41. For the rest of rules we can see in general very high values of confidence and very low values of certainty factors, showing the independence that hold between items. This independence is intuitive in many cases. For example, it is difficult to think that being male/female and being born in USA are dependent (rules 26–28).

That's why only 9 rules are found when using mincf $= 0.5$: the 8 rules of the first experiment without PENATVTY, and rule 13. This rule makes some sense, though its CF is not an excellent value, and it is very strong for minsupp $\leqslant 0.1$. What is remarkable here is that using certainty factors, the number of rules obtained when adding an item with very high support ([PENATVTY $=$ USA] in this case) is almost the same, while the number of rules obtained by using confidence grows, because many misleading rules appear.

5.2. Other databases

Now we are interested in showing how using the ordinary support/confidence framework affects the number of rules, but considering rules with one or more items in the antecedent. Table 6 and Fig. 1 show the results of one of our experiments on a T-set, containing more than $225 \cdot 10^6$ transactions and 10 items, obtained from data about surgical operations in the University Hospital of Granada [6]. By using CFs,

Table 7
Number of rules obtained with and without items "blood" and "prosthesis"

mincf	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
With B/P	795	633	549	431	304	190	104	96	96	0
Without B/P	140	108	95	76	54	28	10	8	8	0

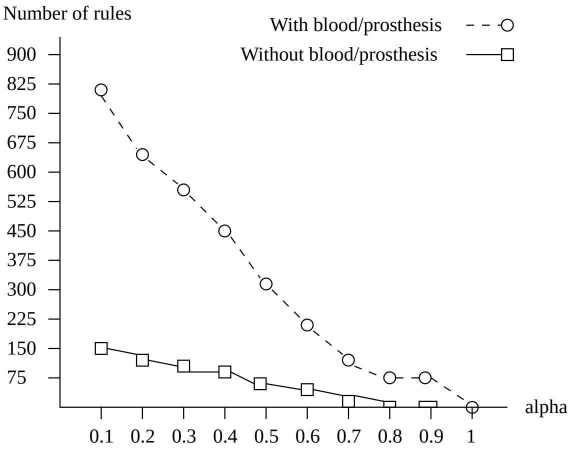

Fig. 2. Graphical representation of Table 7

rules with negative dependence or independence are discarded, and hence much fewer rules are obtained (we used a value *minsupp* $= 0.01$).

We also detected the influence of the items with very high support "blood" and "prosthesis". Their meaning is that blood transfusion and prosthesis are involved in the operation, respectively. Table 7 and Fig. 2 show how the number of rules is reduced if attributes "blood" and "prosthesis" are not considered.

For a minimum accuracy of 0.8, the number of rules has been reduced from 494 (using confidence) to 8. Among the discarded rules, 398 are rules with negative dependence or independence, 32 have "blood" in the consequent, 32 have "prosthesis" in the consequent, and 24 are obtained from the 8 very strong rules by adding to the antecedent "blood", or "prosthesis", or both. They were all misleading rules. Other experiments, involving the CENSUS database, are detailed in [2] and show similar results.

6. Conclusions

By their theoretical properties, very strong rules based on CFs are a suitable framework to discard misleading rules. Our experiments on real-world databases confirm this point. CFs are successfully used in expert systems where the task is predictive, and it is well-known that CFs of rules can be obtained from humans, so we think that CFs are meaningful and that improves the understanding and comparison of rules, and the definition of *mincf*. The concept of very strong rule is very intuitive, since it is based on the logical equivalence between a rule and its counter-reciprocal, and it captures the idea that, since both rules are equivalent, finding evidence of both in data enforces our belief that the rule is important. An additional advantage of the new framework is that the introduction of new support conditions can help to reduce the time and space expended in the first step of the discovery process, the search for frequent itemsets. We shall follow this research avenue in the future.

References

[1] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, in: *Proc. Of the 1993 ACM SIGMOD Conference*, 1993, pp. 207–216.

[2] F. Berzal, M. Delgado, D. Sánchez and M.A. Vila, Measuring the accuracy and importance of association rules. Technical Report CCIA-00-01-16, Department of Computer Science and Artificial Intelligence, University of Granada, 2000.

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

[4] L.M. Fu and E.H. Shortliffe, The application of certainty factors to neural computing for rule discovery, *IEEE Transactions on Neural Networks* **11**(3) (2000), 647–657.

[5] G. Piatetsky-Shapiro, Discovery, analysis, and presentation of strong rules, in: *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W. Frawley, eds, AAAI/MIT Press, 1991, pp. 229–238.

[6] D. Sánchez, *Adquisición de Relaciones Entre Atributos En Bases de Datos Relacionales* (*Translates to: Acquisition of Relationships Between Attributes in Relational Databases*) (*in Spanish*), PhD thesis, Department of Computer Science and Artificial Intelligence, University of Granada, December 1999.

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

[8] A. Silberschatz and A. Tuzhilin, On subjective measure of interestingness in knowledge discovery, in: *Proc. First Int'l Conf. Knowledge Discovery and Data Mining* (*KDD'95*), August 1995, pp. 275–281.

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

[10] P. Smyth and R.M. Goodman, Rule induction using information theory, in: *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W.J. Frawley, eds, AAAI/MIT Press, 1991.