Proceedings of the 6th WSEAS Int. Conf. on FUZZY SYSTEMS, Lisbon, Portugal, June 16-18, 2005 (pp34-39)

Algorithm of the Inverse Confidence of Data Mining Based on the Techniques of Association Rules and Fuzzy Logic

ANDERSON ARAÚJO CASANOVA, SOFIANE LABIDI
Laboratory of Intelligent Systems
Federal University of Maranhão
Campus do Bacanga, São Luis, Maranhão.
BRAZIL

Abstract: - This article proposes an algorithm for data mining that presents a new measure for assistance in the extraction of knowledge. The algorithm uses association rules to extract rules from the databases and fuzzy logic for the classification and comparison of the collected rules.
Key-words: data mining, association rules, fuzzy logic, similarity and algorithm of the inverse confidence.

1 Introduction

Data Mining is the extraction of the predictive information in great databases [10]. The mining algorithms are based on association rules that look for patterns that possess a minimum of frequency in the database. In that way, a very large or very small number of rules can be generated, depending on the parameters used in the search for that information. Besides that, the existent algorithms work with two measures that do not extract all the possible information from the analyzed data.

This article proposes an algorithm, idealized by its own authors, of the search of patterns, using the concept of association rules and fuzzy logic for the classification of data, in that the user formulates hypotheses and it executes tests in the database aided by ACI – the Algorithm of Inverse Confidence, to validate or to refute such hypotheses. The algorithm makes the extraction of data guided by the user and presents two classic measures. It proposes a new measure that presents a rule, taken from among five possible ones, to assist in the collection of knowledge. With the assistance of the fuzzy logic, besides judging the rule, it can also be classified as a bad, regular, good, great or excellent rule the algorithm seeks for similar rules in agreement with a pre-established criterion, making for a more efficient classification of the information as well decreasing the loss of information that would help in the formation of rules. And finally, with the aid of the fuzzy logic, the ACI presents the output of the extracted knowledge at a level of natural language.

Sections 2, 3 and 4 present the concepts of association rules, fuzzy logic and the rules of fuzzy association; section 5 demonstrates the operation of ACI; section 6 displays the obtained results when the algorithm is applied to the database of HUUFMA - Federal University of Maranhão Academic Hospital; section 7 compares ACI with other classic algorithms and the proposal of a new measure with the measure already in use; section 8 presents the conclusion and future works for the improvement of ACI and section 9 with all bibliographical references.

2 Association Rules

The mining of rules is the main function in data mining. The rules express the relationships among the attributes [6]. The representation of knowledge using rules supplies a pattern of behavior through which various knowledge types may be expressed. Association rule is an expression in the way:

$$x \rightarrow y,$$

where x and y are sets of items. The meaning of such a rule is that transactions from the database that contains x tend to contain y. The set of items represented by x is called the antecedent of the rule, and the set of items represented by y is called the consequent of the rule.

A set of items are referred to as ***itemsets***. To each rule two factors are associated: *support* and *confidence*. Support corresponds to the percentage of times that x and y occurs simultaneously throughout the total database records. C*onfidence* corresponds to the percentage of times that x and y occurs simultaneously throughout all the records that contain x. The task of mining the association rules consists of two stages [2]:

- Find all the frequent ***itemsets***;

- Generate association rules from the frequent ***itemsets***.

3 Fuzzy Logic

The theory of the fuzzy sets was proposed by Lotfi Zadeh in 1965 [1, 2, 3, and 11]. The fuzzy logic allows the creation of expertise systems using linguistic variables to create a base of rules [1]. Fuzzy logic deals with imprecise mathematical information usually used in human communication, allowing one to infer an approximate answer for a subject based on knowledge that is inexact, incomplete or not totally reliable. While in Boolean logic, usually used in computation, only two possible values are defined 0 or 1, fuzzy logic is multi-valued (in other words, there is a set of possible values) [2, 8]. The use of fuzzy techniques is considered one of the key systems in data mining because of its likeness to human representation of knowledge. [6]

The characteristic function of a classic set (Boolean) can assume only the values 0 or 1, determining in that way which are the members and the non-members of that set. That function can be generalized so that it can assume values at a certain interval, and the assumed value indicates the degree of pertinence of the element to the set in question. That function is called a pertinence function. The pertinence function of a fuzzy set is denoted by μA, this way:

$$\mu A : X \rightarrow [0,1].$$

The pertinence function maps the elements of a classic set X in real numbers in the interval [0,1]. In this way, a fuzzy set is characterized by a pertinence function μA(x), that associates a real number in the interval [0,1] to each element of the set. This way, the value of μA(x) represents the degree of pertinence of the element x in set A. The larger the value of μA(x) is, the larger the degree of pertinence of x in set A becomes [2].

4 Rules of Fuzzy Association

Quantitative association rules require the creation of appropriate intervals for each attribute. However, those intervals may frequently not be concise and meaningful enough for human experts to discover non-trivial knowledge. Fuzzy sets can be used to represent the intervals, thereby generating rules of fuzzy association. The assignment of significant linguistic terms to the fuzzy sets makes the rules more comprehensible [6].

5 Algorithm of Inverse Confidence

The objective of ACI can be described in this way: provided

- A set of transactions;
- The minimum amount of the sought ***itemset*** (xy);
- A *support* minimum (supmin);
- A *confidence* minimum (confmin);
- The ***itemset*** for which we want to seek, classify and find similar ***itemset***.

To obtain all the association rules that possess:

- *xy* ***itemset***;
- *Confidence* > *confmin*;
- *Support* > *supmin*;
- to find the relationship of y for one or all of the values of x (*Inverse Confidence*);
- the classification of an ***itemset*** in a rule (bad, regular, good, great and excellent);
- and the similarity amongst ***itemsets*** (not similar, a bit similar, almost similar and similar).

ACI sweeps the database in search of all the occurrences of the ***itemset,*** and it determines the support and the confidence of the proposed ***itemset***. If the amount of the sought ***itemset*** is smaller than the *xy* value and the support smaller than the *supmin* value and the confidence smaller than the *confmin* value then the program discards the sought ***itemset*** and it asks the user for a new ***itemset***. But if the parameters of xy, *supmin* and *confmin* are satisfied the program finds the support, the confidence, and the inverse confidence.

The *Inverse Confidence – CI* corresponds to the percentage of times that x and y occur simultaneously throughout all the records that contain y. *CI* of $x \rightarrow y$ is the confidence of the ***itemset*** $x \rightarrow y$. Just as *CI* of $y \rightarrow x$ is the confidence of $x \rightarrow y$. The classification of an ***itemset*** with base in *CI* is obtained without the need of doing another sweep in the records inverting the researched ***itemset***. The values that *CI* assumes and the respective rules for an ***itemset*** are demonstrated in table 1.

Table 1. Values and Rules of *CI*.

Values of CI	Rules
100%	The only y exists in all occurrences of *x*.
>50% e <100%	The sum of the occurrences of *y* when it occurs in *x* is greater than the sum of the other possibilities of x.
50%	The sum of the occurrences of *y* when it occurs in *x* is the same as the sum of the occurrences of *y* for the other possibilities of *x*.
>0% e < 50%	The sum of the occurrences of *y* when it occurs in *x* it is less than the sum of the occurrences of *y* for the other possibilities of *x*.
0%	Occurrences of *y* do not exist in *x*.

ACI is based on CI to discover the relationship between the numbers of occurrences of consequent *y* for the antecedent *x*.

5.1 Classification of a Rule

Using the concept of fuzzy sets to classify the result of CI in an ***itemset*** as a bad, regular, good, great or excellent rule, the pertinence functions for the respective sets ($\mu_{Bad}(CI)$, $\mu_{Regular}(CI)$, $\mu_{Good}(CI)$, $\mu_{Great}(CI)$ e $\mu_{Excellent}(CI)$) are exhibited in table 2.

Table 2. Pertinence Function for the sets of Bad, regular, good, great and excellent rules.

1	When	$CI < 15$
$\mu_{Bad}(CI) = (25 - CI)/10$	When	$15 \le CI < 25$
0	When	$CI \ge 25$
0	When	$10 \ge CI > 45$
$\mu_{Regular}(CI) = (CI - 15)/10$	When	$15 < CI < 25$
$\mu_{Regular}(CI) = (45 - CI)/10$	When	$35 < CI < 45$
1	When	$25 \le CI < 45$
0	When	$35 > CI \ge 45$
$\mu_{Good}(CI) = (CI - 35)/10$	When	$35 < CI < 45$
$\mu_{Good}(CI) = (65 - CI)/10$	When	$55 < CI < 65$
1	When	$45 \le CI < 55$
0	When	$55 \ge CI \ge 85$
$\mu_{Great}(CI) = (CI - 55)/10$	When	$55 < CI < 65$
$\mu_{Great}(CI) = (85 - CI)/10$	When	$75 < CI < 85$
1	When	$65 \le CI < 75$
0	When	$CI \le 75$
$\mu_{Excellent}(CI) = (CI - 75)/10$	When	$75 < CI < 85$
1	When	$CI \ge 85$

The set x that assumes the values of CI from 0 to 100% is mapped through the pertinence function, in fuzzy sets that assume the values of [0,1]. A graphical representation of the fuzzy sets for CI is presented in figure 1.

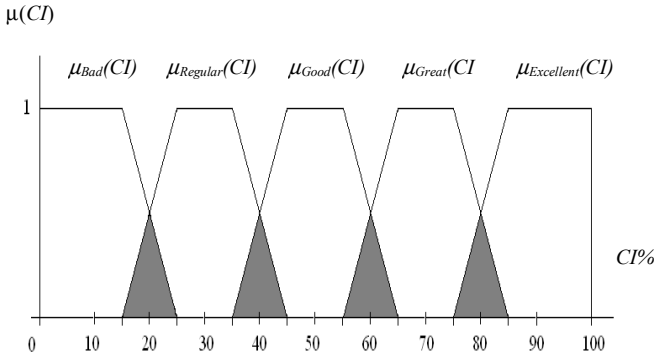

Figure 1. Graphical representation of the fuzzy set $\mu_{Bad}(CI)$, $\mu_{Regular}(CI)$, $\mu_{Good}(CI)$, $\mu_{Great}(CI)$ e $\mu_{Excellent}(CI)$.

5.2 Similarity among Rules

After an ***itemset*** is classified the algorithm seeks other ***itemsets*** that contains an approximate value to the first ***itemset*** proposed. The pertinence function that judges whether other ***itemsets*** are similar ($\mu_{Similarity}(D)$) to the initially proposed ***itemset*** is, only for ***itemsets*** that have

$$\left| CI_{\ i} - CI_{\ i+1} \right| < IV \qquad (1)$$

we calculated

$$D = \left| CI_{\ i} - CI_{\ i+1} \right| \qquad (2)$$

and

$$\mu_{\text{Similarity}}(D) = \left| \frac{D}{IV} - 1 \right| \qquad (3).$$

Where:

- $CI_{\ i}$ is the value of the *inverse confidence* for the proposed ***itemset***;
- $CI_{\ i+1}$ is the value of the *inverse confidence* for the following ***itemset***;
- D is the module of difference between $CI_{\ i}$ and $CI_{\ i+1}$;
- *IV* – Interval Value is the value that an ***itemset*** may distance itself from another ***itemset*** to be classified as not similar.

We defined $D \ge 0$ because of what can occur from $CI_{\ i+1} > CI_{\ i}$.

According to [5] the Cantor K set is a closed subset of the interval [0, 1], obtained as a complement of a meeting of intervals, in other words, we can redefine I.V. as a fuzzy subset [0, 1] inside of the fuzzy set of classification for ***itemsets.*** The function ($\mu_{Similarity}(D)$) may be adjustable; it is enough to alter the variable I.V. In this way the fuzzy subset is dimensioned so that it can increase

or decreased the interval that makes an ***itemset*** similar or not similar to another ***itemset***. The user defines the criterion for the adjustment of *IV*. After the definition for the interval of similarity, *IV*, the algorithm applies (1) to certify that the ***itemset*** is within the interval stipulated by the user. If the difference between CI_i and CI_{i+1} of the ***itemset*** is greater than or equal to the value of *IV*, then the ***itemsets*** are classified as not similar. If (1) is satisfied, the algorithm finds the value of D in equation (2), and then calculates the similarity function that is the quotient of D for *IV* minus 1. In this way we can find ***itemsets*** similar to the sought ***itemset*** or demonstrate the non-similarity among ***itemsets***.

Table 3 shows the classification for two ***itemset***s in agreement with the result of the **($\mu_{Similarity}$ *(D)*)** function presented by the algorithm.

Table 3. Classification for the fuzzy set similarity.

$(\mu_{Similaridade}(D)) = 1$	Similar
$0,5 <= (\mu_{Similaridade}(D)) < 1$	Quase Similar
$0 < (\mu_{Similaridade}(D)) < 0,5$	Pouco Similar
$(\mu_{Similaridade}(D)) = 0$	Não Similar

6 Application of the Algorithm

The algorithm was applied to the database of the HUUFMA's surgical center. The tables were structured in Microsoft Access®. The HUUFMA's database contained 29562 records dated from 22/02/1998. Figure 2 presents the logical model of data from the application.

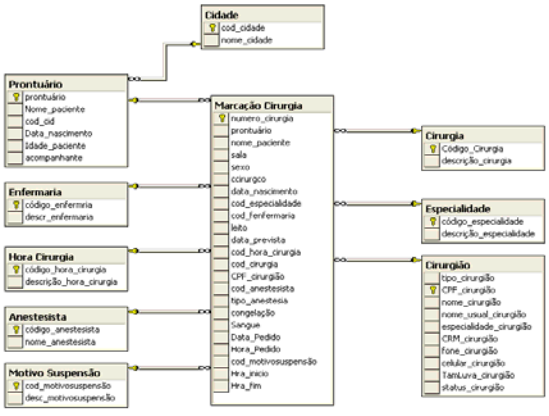

Figure 2. Logical model of data from HUUFMA's surgical center.

The following fields were used in the test: "sex", "clinic", "type of anesthetizes" and "reason for cancellation". The algorithm was executed several times for different ***itemsets***.

6.1 Preliminary results

As the objective was to test which rules would be extracted through the measure of *CI*, support values and low confidence were used. The I.V. variable that defines the similarity interval was defined, in an aleatory way, with a value of 20. Listed below are the results of the test for some ***itemsets***.

Itemset 1:

Sex (Male) ⇒ Anesthesia (Local)
Support = 4.99%, Confidence = 8.80% and Inverse Confidence = 63.48%

The use of the local anesthesia is more frequent in the male sex with considerable margin of difference. A value of *CI* greater than 50% shows that local anesthesia is applied more often in the male sex than in the female sex. When we apply the pertinence function we get the value 0,8; in other words, ***itemset*** 1 is closer to being classified as a great rule than just a good rule.

***Itemset 2*:**

Clinic (General Surgery) ⇒ Reason for Cancellation (Two)
Support = 2.56%, Confidence = 7.77%, Inverse Confidence = 40.15%

The sum of the occurrences for other motives of surgery cancellation occur more in general surgery than motive two, which stand for "Patient not admitted". Reason two is the most found occurrence of cancellation in General Surgery, because it contains a larger *CI* value than other specialties and reason of cancellation number two. Applying the pertinence function to ***itemset*** 2 gives us the value of 0.52, in other words, ***itemset*** 2 can be classified as a good rule, but is on the threshold of also being classified as a regular rule.

***Itemset 3*:**

Sex (Female) ⇒ Surgery (Thyroidectomy)
Support = 0.75%, Confidence = 1.73%, Inverse Confidence = 87.74%

Thyroid surgeries are more frequent in the female sex than in the male sex [9]. The CI value for this ***itemset*** reached quite a high value that could easily be turned into a rule with a low margin of error. According to the pertinence function ***itemset 3*** is classified as an excellent rule.

Applying the **($\mu_{Similarity}$ *(D)*)** function for the results gathered from ***itemsets 1***, ***2*** and ***3***, we obtain the following information: the ***itemsets*** were

classified as "not similar" because the difference between the rules is larger than the proposed similarity interval.

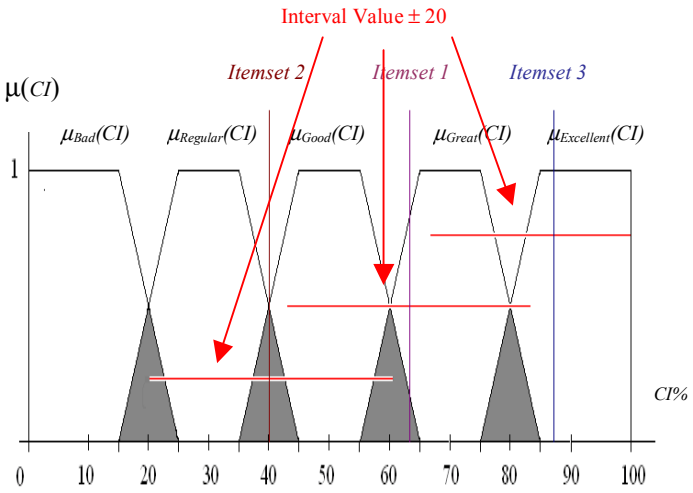

Figure 3. Classification of the researched *itemsets*.

Figure 3 shows the classification of the tested ***itemsets***. The lines in red, traverse to the *CI* axis, exhibit the distance of the interval value used to test the similarity between ***itemsets 1***, ***2*** and ***3***. It may be observed that none of the tested ***itemsets*** actually reaches one of the I.V. lines of another ***itemset*** so that it was classified by the **($\mu_{Similarity}$ *(D)*)** as you can see in Figure 3.

7 Related Works

7.1 APRIORI Algorithm

In the face of existent mining algorithms, that use association rules techniques, Table 4 makes a comparison between the proposed algorithm and a classic algorithm when it comes to association rules, Apriori [2, 4].

Table 4. Comparison between ACI and Apriori.

Algorithms / Characteristics	Apriori	ACI
Does it use an ***itemset*** defined by the user in the search for knowledge?	No	Yes
Does it present the tendency of Y in relation to X without needing to invert the ***itemset*** for a new search?	No	Yes
Does it contain a measure that has a number of fixed rules?	No	Yes
Does it use fuzzy logic to explain the extracted knowledge at a level of natural language?	No	Yes

Observations:

a. To generate rules of the interest for the user, based on a new measure that are not dealt with by other algorithms and that could be important and helpful in the extraction of knowledge;

b. To have a quantity of fixed rules making the analysis of results more concise and efficient, instead of having an indefinite set of rules;

c. To aid in the grouping of rules by fuzzy logic, with a foundation of similarity, avoiding the loss of information in cases where the rules are within a close range of values, but with different classifications.

7.2 *Lift* Measure

Lift [7 and 12] measures how much a rule improves the forecast of a result rather than by just simply assuming the result. The improvement is defined mathematically as the observed frequency of a rule divided by its expected frequency, given the frequencies of each one of its items.

In comparison to *CI* the following observations may be mentioned:

d. Lift measures how much a rule improves a forecast while CI classifies an itemset;

e. Lift works with the itemset $x \rightarrow y$ while CI works with itemset $y \rightarrow x$. This means that the two measures are important, because they each extract different types of information.

f. Lift depends on confidence to present a piece of information about a rule, in other words it depends on another measure. CI does not depend on any other measure.

8 Conclusion and Future Works

The proposed algorithm uses a practical technique for the extraction of knowledge, and it brings collaboration with a new measure that helps in the search for rules. Just like support and confidence, the greater the value of CI the greater the chance the extracted knowledge has of becoming a rule.

ACI has the following characteristics: (1) the measure of CI does not generate a very big set of rules, instead ACI tries to aggregate the maximum amount of ***itemsets*** with the same rule, because it possesses five fixed rules that can be used for several researched ***itemsets***, (2) it researches

only ***itemsets*** of interest to the user, (3) it can generate rules of $X \rightarrow Y$ as well as $Y \rightarrow X$, in the latter when the support and confidence values are low, which is the largest differential of ACI in relation to other algorithms, (4) it applies fuzzy logic to the classification of rules proposed by the user, in a way that allows it to take advantage of rules with proximity of values classified by ACI.

The result obtained by ACI for the ***Itemset 3***: Sex (Female) $\Rightarrow$ Surgery (Thyroidectomy) demonstrated a high value for CI that is proven in the medical section of the bibliography [9]. In this way we attest to the proven efficiency of the algorithm in the extraction of rules. ACI could be used at HUUFMA for the extraction of knowledge not only in the area of health as well as its various sections, and other areas of knowledge.

We are currently working on the improvement of the algorithm by focusing on the following items:

- To test and to validate the algorithm using another database that presents a different context than the current database;
- To compare the results of ACI with the results of other algorithms;
- To perfect ACI with other technologies for greater interaction and inter operation among various databases;
- To define a criterion for the fuzzy set of classification of the rules, in other words, attribute meaningful linguistic terms to the fuzzy set, as defined by the similarity fuzzy set presented in Table 3.

9 Bibliographical References

[1] BAUCHSPIESS, A., **Introdução aos Sistemas Inteligentes**. **Aplicações em Engenharia de Redes Neurais Artificiais, Lógica Fuzzy e Sistemas Neuro-Fuzzy.** http://www.ene.unb.br/adolfo/ISI Brasília. 2004.

[2] ESCOVAR, E. **Algoritmo SSDM para Mineração de Dados Semanticamente Similares**. 2004. Dissertação (Mestrado) – Departamento de Computação (DC), UFSCar, São Carlos. 2004.

[3] HELLMANN, M. **Fuzzy Logic Introduction**. Université de Rennes. 2001.

[4] KANTARDZIC, M., **Data Mining Concepts, Models, Methods, and Algorithms**. Ed Wiley. 2003.

[5] LIMA, E., Lages., **Curso de Análise**. Vol 1. Instituto de Matemática Pura e Aplicada. 1976.

[6] MITRA, S., ACHARYA, T. **DATA MINING: Multimedia, Soft Computing and Bioinformatics.** John wiley & Sons. 2003.

[7] PASSARI, A., **Exploração de Dados Atomizados para Previsão de Vendas no Varejo Utilizando Redes Neurais**. Dissertação (Mestrado) Departamento de Administração – USP. 2003

[8] REYNOLDS, Keith, M. **Fuzzy Logic knowledge Bases in Integrated Landscape Assessment: Examples and Possibilities.** United States Department of Agriculture. General Technical Report. 2001.

[9] VALENTI, P., **Medicina Interna: Compendio Práctico de Patologia Médica y Terapéutica Clínica**. Séptima Edición. Editorial Marin. Barcelona. 1967.

[10] WANG, J. **Data Mining – Opportunities and Challenges**. Idea Group Publishing. 2003.

URL's

[11] Fuzzy Logic and Fuzzy Expert Systems http://www-2.cs.cmu.edu/Groups/AI/html/faqs/ai/fuzzy/part1/faq-doc-2.html

[12] Applications of Data Mining – Association Rules http://www.comp.rgu.ac.uk/staff/smc/teaching/data mining/apriori-lab/