Finding Fuzzy Approximate Dependencies within STULONG Data

Fernando Berzal, Juan-Carlos Cubero, Daniel Sanchez, Jose M. Serrano[1], and M.-Amparo Vila

Dept. of Computer Sciences and Artificial Intelligence,
University of Granada,
18071 Granada, Spain
{berzal, JC.Cubero, daniel, jmserrano, vila}@decsai.ugr.es
http://frontdb.ugr.es

Abstract. Discovery and analysis of previously unknown and potentially useful relations between data in a given set of objects give us the basis for the establishment of, for example, decision-aid systems. A very interesting issue is the modelling of real problems by means of fuzzy logic, because real data can be affected by imprecision or uncertainty. But also, uncertainty can be introduced in a pre-processing step, in order to reduce granularity in data. According to this, existing KDD techniques must be extended to manage such type of information. This work is devoted to the extraction of fuzzy approximate dependencies from a selected part of the 2003 ECML discovery challenge data set. Several interesting results are provided and discussed.

1 Introduction

KDD techniques allow us to obtain previously unknown and potentially useful relations between attributes in a given set of objects, commonly stored in a database. An example of this is medical data. Hospital records can be analyzed in order to provide to medical experts invaluable information that could be used to describe diseases symptoms or to predict patient behaviors, for example.

Usually, real problems data can be affected by an imprecision or uncertainty degree, and existing KDD techniques must be extended in order to manage such type of information. The theory of fuzzy subsets [15] is a helpful tool to reach this goal. In others occasions, uncertainty can be applied to crisp data in order to reduce information granularity. Similarity relations can be established over categorical values, or linguistic labels, representing possibility distributions, can be defined over a numeric domain.

The 2003 ECML discovery challenge STULONG data set provides information about the twenty years lasting longitudinal study of the risk factors of the

[1] Corresponding author. Phone: +34 958 240466, Fax: +34 958 243317

atherosclerosis in the population of 1417 middle aged men. The study was realized at the 2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC), under the supervision of Prof. F. Boudík, MD, ScD, with collaboration of M. Tomeckov, MD, PhD and Ass. Prof. J. Bultas, MD, PhD. The data were transferred to the electronic form by the European Centre of Medical Informatics, Statistics and Epidemiology of Charles University and Academy of Sciences (head. Prof. RNDr. J. Zvrov, DrSc). The data resource is on the web pages `http://euromise.vse.cz/challenge2003`. At present time the data analysis is supported by the grant of the Ministry of Education CR Nr LN 00B 107.

In this work, STULONG data is analyzed in order to extract fuzzy approximate dependencies from it. The paper is organized as follows. Section 2 is devoted to the explanation of our methodologies as well as a brief description of the pre-processing step over data. In section 3, some of the analytical questions proposed in the challenge are answered by means of the described techniques, and an interpretation of the results is given. Section 4 concludes the paper.

2 Materials and methods

A very common issue in database analysis is the study of existing relations between data stored in a database. Mainly, we can distinguish two basic types of relations. On the one hand, there can be implicit or hidden relations between attribute values, not clear at a first moment. One of the most known examples of this are association rules, defined in [1]. Association rules are "implications" that relate the presence of itemsets (set of items) in a given set of transactions (a T-set). A classical example consider that items are things we can buy in a market, and transactions are market basket containing several items. These rules take the form of, for example, "80% of people that buy milk, buy also flour".

On the other hand, we can find explicit relations between data (i.e., we can affirm that a job class determines a salary class, or that a given postal code determines the city). We can say that there exists a functional dependence between attributes. Formally, let $R = \{At_1, \ldots, At_m\}$ be a set of attributes and let r be a table with attributes in R such that $|r| = n$. Also, let $X, Y \subset R$ with $X \cap Y = \emptyset$, and let $dom(X) = \{x_1, \ldots, x_K\}$ and $dom(Y) = \{y_1, \ldots, y_M\}$ be the values of X and Y appearing in r. A functional dependence $X \to Y$ holds in R if and only if for every instance r of R

$$\forall t, s \in r \;\; if\; t[X] = s[X] \; then \; t[Y] = s[Y] \tag{1}$$

2.1 Approximate dependencies

Approximate dependencies [6, 11, 12] can be roughly defined as functional dependencies with exceptions. The definition of approximate dependence is then a matter of how to define exceptions, and how to measure the accuracy of the dependence (see [5]). We shall follow the approach introduced in [13, 7, 4], where we applied the same methodology employed in mining for association rules to the discovery of approximate dependencies. The idea is that it is interesting to measure not only the accuracy of the dependence (as other existing approaches do [12, 11, 10]) but also its support, in order to see the empirical evidence associated to the dependence in data. This way, we can avoid to obtain trivial dependencies.

To assess the dependencies, we apply the same measures of interest and accuracy introduced in [1], that is support and confidence. As discussed in [13, 2, 3], confidence is used in order to compute a accuracy measure based on certainty factors (see [14] for the definition, and [2, 3] for the explanation). Formally, we obtain the certainty factor of a rule as follows,

$$CF(X \Rightarrow Y) = \frac{(Conf(X \Rightarrow Y)) - S(Y)}{1 - S(Y)} \tag{2}$$

if $Conf(X \Rightarrow Y) > S(Y)$, and

$$CF(X \Rightarrow Y) = \frac{(Conf(X \Rightarrow Y)) - S(Y)}{S(Y)} \tag{3}$$

if $Conf(X \Rightarrow Y) < S(Y)$, and 0 otherwise.

Certainty factors take values in $[-1, 1]$, indicating the extent to which our belief that the consequent is true varies when the antecedent is also true. It ranges from 1, meaning maximum increment (i.e., when A is true then B is true) to -1, meaning maximum decrement.

Returning to our definition of AD, the idea is that, since a functional dependence "$X \rightarrow Y$" can be seen as a rule that relates the equality of attribute values in pairs of tuples (see equation (1)), and association rules relate the presence of items in transactions, we can represent approximate dependencies as association rules by using the following interpretations of the concepts of item and transaction:

- An item is an object associated to an attribute of R. For every attribute $At_k \in R$ we note it_{At_k} the associated item.
- We introduce an itemset I_X to be

$$I_X = \{it_{At_k} \mid At_k \in X\}$$

- T_r is a T-set that, for each pair of tuples $< t, s > \in r \times r$ contains a transaction $ts \in T_r$ verifying

$$it_{At_k} \in ts \;\Leftrightarrow\; t[At_k] = s[At_k]$$

It is obvious that $|T_r| = |r \times r| = n^2$.

Then, an approximate dependence $X \rightarrow Y$ in the relation r is an association rule $I_X \Rightarrow I_Y$ in T_r (see [7, 4]). The support and certainty factor of $I_X \Rightarrow I_Y$ measure the interest and accuracy of the dependence $X \rightarrow Y$. In particular, the following property holds:

Proposition 1 ([4]).
If $CF(X \rightarrow Y) = 1$ then $X \rightarrow Y$ is a functional dependence.

2.2 Fuzzy Association Rules

In [9], 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. Let $I = \{i_1, \ldots, i_m\}$ be a set of items and T be a set of fuzzy transactions, where each fuzzy transaction is a fuzzy subset of I. Let $\tilde{\tau} \in T$ be a fuzzy transaction, we note $\tilde{\tau}(i_k)$ the membership degree of i_k in $\tilde{\tau}$. A fuzzy association rule is an implication of the form $A \Rightarrow C$ such that $A, C \subset R$ and $A \cap C = \emptyset$. A and C are called antecedent and consequent respectively.

For example, in the particular case of relational databases, let $R = \{At_1, ..., At_m\}$ be a set of attributes, and let $Lab(At_k) = \{a_{k_1}, ..., a_{k_n}\}$ be a set of linguistic labels defined on $dom(At_k)$ $\forall At_k \in R$. Let r be a relation with attributes in R. Then, a fuzzy transaction can be obtained from each $t \in R$ as follows:

$$\tilde{\tau}_t = \sum_{k \in \{1, \ldots, m\}} a_{k_i}(t[At_k])/a_{k_i}$$

where each item is a pair (At_k, a_{k_i}) representing *'At_k is a_{k_i}'*.

It is immediate that the set of transactions where a given item appears is a fuzzy set. We call it *representation* of the item. For item i_k in T we have the following fuzzy subset of T:

$$\tilde{\Gamma}_{i_k} = \sum_{\tilde{\tau} \in T} \tilde{\tau}(i_k)/\tilde{\tau} \qquad (4)$$

This representation can be extended to itemsets as follows: let $I_0 \in R$ be an itemset, its representation is the following subset of T:

$$\tilde{\Gamma}_{I_0} = \bigcap_{i \in I_0} \tilde{\Gamma}_i = min_{i \in I_0} \tilde{\Gamma}_i \qquad (5)$$

In order to measure the interest and accuracy of a fuzzy association rule, we must use approximate reasoning tools, because of the imprecision that affects fuzzy transactions and, consequently, the representation of itemsets. In [9], a semantic approach is proposed based on the evaluation of quantified sentences (see [16]). Let Q be a fuzzy coherent quantifier.

Definition 1 ([9]). *The support of an itemset is equal to the result of evaluating the quantified sentence Q of T are $\tilde{\Gamma}_{I_0}$.*

Definition 2 ([9]). *The support of the fuzzy association rule $A \Rightarrow C$ in the FT-set T, $Supp(A \Rightarrow C)$, is the evaluation of the quantified sentence Q of T are $\tilde{\Gamma}_{A \cup C} = Q$ of T are $(\tilde{\Gamma}_A \cap \tilde{\Gamma}_C)$.*

Definition 3 ([9]). *The confidence of the fuzzy association rule $A \Rightarrow C$ in the FT-set T, $Supp(A \Rightarrow C)$, is the evaluation of the quantified sentence Q of $\tilde{\Gamma}_A$ are $\tilde{\Gamma}_C$.*

We choose the quantifier Q_M, defined by $Q_M(x) = x$, since it verifies the conditions we request for a quantifier and it has a valuable property: the values obtained by using it in definitions 1, 2 and 3 in the case of crisp transactions, are the ordinary measures of support and confidence in the crisp case. This way, the proposed method is a generalization of the ordinary association rule assessment framework in the crisp case.

2.3 Fuzzy Approximate Dependencies

As seen in [5], it is possible to extend the concept of functional dependence in several ways by smoothing some of the elements of the rule in equation 1. We want to consider as much cases as we can, integrating both approximate dependencies (exceptions) and fuzzy dependencies. For that purpose, in addition to allowing exceptions, we have considered the relaxation of several elements of the definition of functional dependencies. In particular we consider membership degrees associated to pairs (attribute, value) as in the case of fuzzy association rules, and also fuzzy similarity relations to smooth the equality of the rule in equation 1. We consider S_{At_i} a fuzzy similarity relation in $dom(At_i)$. To be more precise, relations are assumed to be max-min transitive, i.e.

$$S_{At_k}(x_i, x_j) \geq \bigvee_{l=1}^{n} min(S_{At_k}(x_i, x_l), S_{At_k}(x_l, x_j)), \forall x_i, x_j \in dom(At_k) \quad (6)$$

We shall define fuzzy approximate dependencies in a relation as fuzzy association rules on a special FT-set obtained from that relation, in the same way that approximate dependencies are defined as association rules on a special T-set.

Let $I_R = \{it_{At_k} | At_k \in R\}$ be the set of items associated to the set of attributes R. We define a FT-set T'_r associated to table r with attributes in R as follows: for each pair of rows $< \tilde{t}, \tilde{s} >$ in $r \times r$ we have a fuzzy transaction $\widetilde{ts}$ in T'_r defined as

$$\widetilde{ts}(it_{At_k}) = min(\mu_{\tilde{t}}(At_k), \mu_{\tilde{s}}(At_k), S_{At_k}(\tilde{t}(At_k), \tilde{s}(At_k)))\ \forall it_{At_k} \in T'_r \quad (7)$$

This way, the membership degree of a certain item it_{At_k} in the transaction associated to tuples $\tilde{t}$ and $\tilde{s}$ takes into account the membership degree of the value of At_k in each tuple and the similarity between them. This value represents the degree to which tuples $\tilde{t}$ and $\tilde{s}$ agree in At_k, i.e., the kind of items that are related by the rule in equation 1. On this basis, we define fuzzy approximate dependencies as follows:

Definition 4. *Let $X, Y \subseteq R$ with $X \cap Y = \emptyset$ and $X, Y \neq \emptyset$. The fuzzy approximate dependence $X \rightarrow Y$ in r is defined as the fuzzy association rule $I_X \Rightarrow I_Y$ in T'_r following ([9]).*

The support and certainty factor of $I_X \Rightarrow I_Y$ are calculated from T'_r as explained in section 2.2, and they are employed to measure the importance and accuracy of $X \to Y$.

Following [9], the FAD $X \to Y$ holds with total accuracy (certainty factor $CF(X \to Y) = 1$) in a relation r iff $\widetilde{ts}(I_X) \leq \widetilde{ts}(I_Y)\ \forall \widetilde{ts} \in T'_r$ (let us remember that $\widetilde{ts}(I_X) = \min_{At_k \in X} \widetilde{ts}(it_{At_k})\ \forall X \subseteq R$). Moreover, since fuzzy association rules generalize crisp association rules, FAD's generalize AD's.

Additional properties and an efficient algorithm for computing FAD's are to appear in a separate paper.

2.4 STULONG Data Pre-processing

STULONG data consists of raw data matrices. Before the analysis, some attributes had to be transformed. Numeric attributes are problematic when extracting association rules or approximate dependencies. Results can be better described if linguistic labels are defined over a numeric domain. According to this, a set of three equi-depth intervals was computed for attributes $SYST1$, $DIAST1$, $SYST2$, $DIAST2$, $TRIC$, $SUBSC$, $CHLST$, and $TRIGL$ in *Entry* table. Also, attribute BMI, representing Body Mass Index, was computed from attributes $VYSKA$ and $VAHA$, as specified in challenge instructions. Also, as suggested, the attribute was categorized in only two intervals, *Overweight* if BMI is greater or equal 25, and $Thin$ otherwise.

The original intervals were smoothed to obtain fuzzy sets. We set the overlapping between two intervals at level 0 (no intersection). The main problem was then that values near to the bounds had to be obviated, and the number of these values should be minimal. In order to accomplish this, we assumed that the data distribution was normal, and computed 47th and 53th percentiles. We took as label amplitude the 5% percent of the distance between the percentiles, hence reducing the number of lost values.

A fuzzy similarity relation was defined for each categorical attribute. Similarity relations are most suitable for describing analogical data over discrete domains, in addition to fuzzy sets. Table 1 shows these relations, defined according to a semantic criteria. As relations for $PIVOMN$, $VINOMN$ and $LIHMN$ are the same, the latter two tables are omitted.

3 Analytical Questions and Results

We restrained to only one of STULONG data matrices. In this paper we try to solve the analytical questions related to *Entry* table. This table contains information about 1417 patients. Each of them was described by a total of 244 attributes, 64 of them codified in categories corresponding to intervals or transformation over other attributes. The set of attributes is divided into several groups, according to their semantics. Additional information can be found in the challenge website. Also, as indicated, patients were classified into three basic groups:

- Normal Group (attribute KONSKUP having values 1 or 2).
- Risk Group (attribute KONSKUP having values 3 or 4).
- Pathologic Group (value 5 for attribute KONSKUP).

In the following subsections, we discuss the obtained results. In order to optimize the number of useful dependencies, we took minimum certainty factor $minCF = 0.4$ as threshold.

3.1 Relations involving social factors

First, we must remark that many dependencies were found with a high CF and attribute $STAV$ (status) in the consequent. Due to the fact that there exists a big unbalance in attribute values (about 1200 of 1417 men are married), these dependencies give us little information about the overall relations. Hence, attribute $STAV$ might be omitted in the analysis.

Relations between social factors and physical activities Some of the obtained dependencies that we have found to be interesting are the following,

$$[VZDELANI] \rightarrow [AKTPOZAM], supp\ 17.47\%, CF\ 0.53$$
$$[ZODPOV] \rightarrow [AKTPOZAM], supp\ 19.81\%, CF\ 0.51$$

with similar CF values for all groups. They reveal a possible relation between reached education ($VZDELANI$) and responsibility in job ($ZODPOV$) with physical activity after job ($AKTPOZAM$). An intuitive relation can be expected between education or job category and healthy habits. Also, for patients in Normal and Risk groups, additional relations between $VZDELANI$ and $ZODPOV$ with time taken to get to work ($DOPRATRV$) are found,

$$[VZDELANI] \rightarrow [DOPRATRV], supp\ 14.78\%, CF\ 0.47$$
$$[ZODPOV] \rightarrow [DOPRATRV], supp\ 18.63\%, CF\ 0.44$$

Relations between social factors and smoking Strong relations appear involving reached education or responsibility in job and the time an ex-smoker has not been smoking ($BYVKURAK$). The CF values are higher in Risk groups, maybe indicating, as in the previous paragraph, that people with a high social status tend to live more healthily. This hypothesis is based on the clear unbalance of the data distribution towards ex-smokers.

$$[VZDELANI] \rightarrow [BYVKURAK], supp\ 21.36\%, CF\ 0.74$$
$$[ZODPOV] \rightarrow [BYVKURAK], supp\ 27.81\%, CF\ 0.74$$

Relations between social factors and alcohol consumption On the one hand, we have an attribute, $PIVO7$ (beer 7^o consumption), with a very high support (near 100%). This results in a set of dependencies with a high CF (greater than 0.9), and hence, nearly functional dependencies. Unfortunately, these dependencies appear to be of little use, and attribute $PIVO7$ will not be considered in the following. On the other hand, any other strong relation involving the rest of the attributes was found.

Relations between social factors and body mass index (BMI) Similar results are obtained for the three groups of patients, and they reveal relatively high dependencies with BMI as consequent, as the following,

$[ROKVSTUP] \rightarrow [BMI], \mathit{supp}\ 17.44\%, CF\ 0.53$

$[VZDELANI] \rightarrow [BMI], \mathit{supp}\ 17.04\%, CF\ 0.51$

$[ZODPOV] \rightarrow [BMI], \mathit{supp}\ 19.77\%, CF\ 0.51$

Relations between social factors and blood pressure Having into account that they hold only in Normal group, the found dependencies with higher CF values were the following,

$[ROKVSTUP] \rightarrow [DIAST1], \mathit{supp}\ 15.48\%, CF\ 0.49$

$[ROKVSTUP] \rightarrow [DIAST2], \mathit{supp}\ 15.17\%, CF\ 0.48$

$[VZDELANI] \rightarrow [DIAST1], \mathit{supp}\ 17.43\%, CF\ 0.47$

$[VZDELANI] \rightarrow [DIAST2], \mathit{supp}\ 17.32\%, CF\ 0.47$

$[ZODPOV] \rightarrow [DIAST1], \mathit{supp}\ 18.87\%, CF\ 0.45$

$[ZODPOV] \rightarrow [DIAST2], \mathit{supp}\ 19.10\%, CF\ 0.46$

Relations between social factors and cholesterol levels Unfortunately, no special relation appears to be found involving social factors and cholesterol levels by means of our methodology.

3.2 Relations involving physical activities

The next set of analytical questions searched for relations between physical activities and other characteristics as the following ones.

Relations between physical activities and smoking In general, several dependencies with a high CF (up to 0.76) and attribute $BYVKURAK$ as consequent can be obtained, see subsection 3.1. But, to our knowledge, we found more remarkable the following dependencies, related to Pathologic group,

$[KOURENI] \rightarrow [AKTPOZAM], \mathit{supp}\ 14.82\%, CF\ 0.58$

$[DOBAKOUR] \rightarrow [AKTPOZAM], \mathit{supp}\ 21.45\%, CF\ 0.53$

A relatively high relation is present involving intensity of smoking ($KOURENI$) and the time the patient has been smoking ($DOBAKOUR$) with physical activity after job. It looks reasonable that a relation like this occurs, as people with damaged lungs can see reduced physical activities.

Relations between physical activities and alcohol consumption According to our results, there are no significant differences between patients groups relating physical activities and alcohol consumption, excepting that no interesting dependencies with $DOPRATRV$ as consequent appear for people in Pathologic group. Some of the obtained dependencies, relating drinking of alcohol with physical activities after job and time employed to get to work, are shown in the following,

$[ALKOHOL] \rightarrow [AKTPOZAM], \textit{supp}\ 24.41\%, CF\ 0.45$
$[PIVOMN] \rightarrow [AKTPOZAM], \textit{supp}\ 22.70\%, CF\ 0.47$
$[VINOMN] \rightarrow [AKTPOZAM], \textit{supp}\ 23.34\%, CF\ 0.46$
$[LIHMN] \rightarrow [AKTPOZAM], \textit{supp}\ 21.53\%, CF\ 0.47$
$[ALKOHOL] \rightarrow [DOPRATRV], \textit{supp}\ 22.82\%, CF\ 0.41$
$[PIVOMN] \rightarrow [DOPRATRV], \textit{supp}\ 21.21\%, CF\ 0.43$
$[VINOMN] \rightarrow [DOPRATRV], \textit{supp}\ 21.47\%, CF\ 0.41$
$[LIHMN] \rightarrow [DOPRATRV], \textit{supp}\ 20.00\%, CF\ 0.43$

Relations between physical activities and BMI Nearly no differences can be found between the obtained dependencies for the three groups. In Pathologic group, CF values are a bit higher (up to 0.56), but in general, interpretations would remain the same. Some examples are the following ones,

$[TELAKTZA] \rightarrow [BMI], \textit{supp}\ 16.22\%, CF\ 0.51$
$[DOPRAVA] \rightarrow [BMI], \textit{supp}\ 24.93\%, CF\ 0.44$
$[DOPRATRV] \rightarrow [BMI], \textit{supp}\ 24.82\%, CF\ 0.47$
$[TELAKTZA, AKTPOZAM] \rightarrow [BMI], \textit{supp}\ 10.22\%, CF\ 0.56$
$[AKTPOZAM, DOPRAVA] \rightarrow [BMI], \textit{supp}\ 15.80\%, CF\ 0.50$
$[AKTPOZAM, DOPRATRV] \rightarrow [BMI], \textit{supp}\ 16.10\%, CF\ 0.55$

Relations between physical activities and blood pressure The main difference between groups is the one related to the number of obtained dependencies. It is higher in Normal group, involving dependencies with $DIAST1$ or $DIAST2$ as consequents, that do not appear for the remaining groups. Dependencies with $AKTPOZAM$ or $DOPRATRV$ as consequents have similar CF values for all groups.

$[AKTPOZAM, DOPRAVA] \rightarrow [DIAST1], \textit{supp}\ 11.99\%, CF\ 0.51$
$[AKTPOZAM, DOPRAVA] \rightarrow [DIAST2], \textit{supp}\ 12.04\%, CF\ 0.51$

$$\begin{array}{l}
[AKTPOZAM, DOPRATRV] \rightarrow [DIAST1], \textit{supp}\ 15.30\%, CF\ 0.47 \\
[AKTPOZAM, DOPRATRV] \rightarrow [DIAST2], \textit{supp}\ 15.48\%, CF\ 0.48 \\
[SYST1, DIAST1] \rightarrow [AKTPOZAM], \textit{supp}\ 15.02\%, CF\ 0.45 \\
[SYST1, SYST2] \rightarrow [AKTPOZAM], \textit{supp}\ 15.30\%, CF\ 0.45 \\
[SYST1, DIAST2] \rightarrow [AKTPOZAM], \textit{supp}\ 15.26\%, CF\ 0.46 \\
[DIAST1, SYST2] \rightarrow [AKTPOZAM], \textit{supp}\ 15.64\%, CF\ 0.45 \\
[SYST2, DIAST2] \rightarrow [AKTPOZAM], \textit{supp}\ 16.77\%, CF\ 0.45 \\
[DOPRAVA, DOPRATRV] \rightarrow [DIAST1], \textit{supp}\ 13.10\%, CF\ 0.52 \\
[DOPRAVA, DOPRATRV] \rightarrow [DIAST2], \textit{supp}\ 12.88\%, CF\ 0.51 \\
[SYST1, DIAST1] \rightarrow [DOPRATRV], \textit{supp}\ 14.69\%, CF\ 0.44 \\
[SYST1, SYST2] \rightarrow [DOPRATRV], \textit{supp}\ 14.97\%, CF\ 0.44 \\
[SYST1, DIAST2] \rightarrow [DOPRATRV], \textit{supp}\ 14.72\%, CF\ 0.44 \\
[DIAST1, SYST2] \rightarrow [DOPRATRV], \textit{supp}\ 15.29\%, CF\ 0.44 \\
[SYST2, DIAST2] \rightarrow [DOPRATRV], \textit{supp}\ 16.19\%, CF\ 0.43
\end{array}$$

Relations between physical activities and cholesterol levels Relations between cholesterol ($CHLST$) and triglycerides ($TRIGL$), as antecedents, and $AKTPOZAM$ and $DOPRATRV$, as consequent, appear specially in Risk group. No dependencies with $CHLST$ as consequent appear in Normal group, and no dependencies with $DOPRATRV$ as consequent appear in Pathologic group.

$$\begin{array}{l}
[CHLST] \rightarrow [AKTPOZAM], \textit{supp}\ 19.02\%, CF\ 0.47 \\
[TRIGL] \rightarrow [AKTPOZAM], \textit{supp}\ 14.62\%, CF\ 0.50 \\
[CHLST] \rightarrow [DOPRATRV], \textit{supp}\ 17.63\%, CF\ 0.43 \\
[TRIGL] \rightarrow [DOPRATRV], \textit{supp}\ 13.58\%, CF\ 0.46
\end{array}$$

3.3 Relations involving alcohol consumption

As discussed before, attribute $PIVO7$ had a very high support and, for this reason, dependencies involving it would be misleading. Hence, $PIVO7$ was not considered for the analysis.

Relations between alcohol consumption and smoking Similar results were obtained for the three groups, although with significant higher CF in Risk group (from 0.5 to 0.7). The dependencies related drinking with the time an ex-smoker has not been smoking, in the following way,

$$\begin{array}{l}
[ALKOHOL] \rightarrow [BYVKURAK], \textit{supp}\ 33.17\%, CF\ 0.69 \\
[PIVOMN] \rightarrow [BYVKURAK], \textit{supp}\ 29.87\%, CF\ 0.69 \\
[VINOMN] \rightarrow [BYVKURAK], \textit{supp}\ 31.49\%, CF\ 0.70 \\
[LIHMN] \rightarrow [BYVKURAK], \textit{supp}\ 29.20\%, CF\ 0.71
\end{array}$$

Relations between alcohol consumption and BMI Our results reveal that a strong relation exists between alcohol consumption and BMI, possibly overweight, as it was expected. The number of dependencies and their CF values are high, specially in Pathologic group, as we can see (only dependencies with one antecedent are shown),

$[ALKOHOL] \rightarrow [BMI], supp\ 22.74\%, CF\ 0.50$
$[PIVO10] \rightarrow [BMI], supp\ 30.32\%, CF\ 0.41$
$[VINO] \rightarrow [BMI], supp\ 29.69\%, CF\ 0.42$
$[LIHOV] \rightarrow [BMI], supp\ 29.85\%, CF\ 0.42$
$[PIVOMN] \rightarrow [BMI], supp\ 18.93\%, CF\ 0.52$
$[VINOMN] \rightarrow [BMI], supp\ 22.10\%, CF\ 0.50$
$[LIHMN] \rightarrow [BMI], supp\ 22.08\%, CF\ 0.49$

Relations between alcohol consumption and blood pressure In this case, dependencies were only found for population in Normal Group, and they reveal relatively high relations between alcohol drinks consumption and diastolic blood pressure, as our results show,

$[ALKOHOL] \rightarrow [DIAST1], supp\ 21.19\%, CF\ 0.43$
$[PIVOMN] \rightarrow [DIAST1], supp\ 22.51\%, CF\ 0.45$
$[VINOMN] \rightarrow [DIAST1], supp\ 22.74\%, CF\ 0.43$
$[LIHMN] \rightarrow [DIAST1], supp\ 21.26\%, CF\ 0.44$
$[ALKOHOL] \rightarrow [DIAST2], supp\ 21.00\%, CF\ 0.43$
$[PIVOMN] \rightarrow [DIAST2], supp\ 22.21\%, CF\ 0.44$
$[VINOMN] \rightarrow [DIAST2], supp\ 22.34\%, CF\ 0.42$
$[LIHMN] \rightarrow [DIAST2], supp\ 21.10\%, CF\ 0.43$

Relations between alcohol consumption and cholesterol levels No interesting dependencies were found relating alcohol consumption and cholesterol levels. We suggest a local study involving relations between attribute values by means of association rules.

3.4 Relations involving skin folds and BMI

Finally, to our knowledge, no correlation can be found by means of our methodology between skin folds and BMI. Our dependencies only show a directional relation from $TRIC$ (skinfold above musculus triceps) and $SUBSC$ (skinfold above musculus subscapularis) to BMI, as the following,

$[TRIC] \rightarrow [BMI], supp\ 15.85\%, CF\ 0.54$
$[SUBSC] \rightarrow [BMI], supp\ 17.28\%, CF\ 0.58$

The obtained results are similar in the three groups.

4 Concluding Remarks

STULONG data is analyzed by means of fuzzy approximate dependencies. Using fuzzy subsets, some attributes are transformed in an attempt of providing more comprehensible information. A set of analytical questions is solved according to the obtained results. Our methodology generates a set of dependencies that could be applied to datasets description or data prediction. Despite of the fact that the results appear to be interesting to our knowledge, the aid of medical experts would be desirable in order to give an useful interpretation. Further efforts would be devoted to develop user-guided tools in order to obtain better knowledge and hence take better decisions.

References

1. Agrawal R., Imielinski T., Swami A. (1993). Mining Association Rules between Sets of Items in Large Databases. Proc. of the 1993 ACM SIGMOD Conf., Washington DC, USA.
2. Berzal F., Blanco I., Sánchez D., and Vila M. (2001). A new framework to assess association rules. In Hoffmann, F., editor, *Advances in Intelligent Data Analysis. Fourth International Symposium, IDA'01. Lecture Notes in Computer Science 2189*, pages 95–104. Springer-Verlag.
3. Berzal F., Blanco I., Sánchez D., and Vila M. (2002). Measuring the Accuracy and Interest of Association rules: A New Framework. Intelligent Data Analysis 6 pp. 221-235.
4. Blanco I., Martín-Bautista M.J., Sánchez D., Vila, M.A. (2000). On the support of dependencies in relational databases: Strong approximate dependencies. Data Mining and Knowledge Discovery. Submitted.
5. Bosc P., Lietard L., Pivert O. (1997). Functional Dependencies Revisited Under Graduality and Imprecision. Annual Meeting of NAFIPS, pp. 57-62.
6. Bra P.D. and Paredaens J. (1983). Horizontal decompositions for handling exceptions to functional dependencies. Advances in Database Theory, 2:123-144.
7. Delgado M., Martín-Bautista M.J., Sánchez D., Vila M.A. (2000). Mining strong approximate dependencies from relational databases. In Proceedings of IPMU'2000.
8. Delgado M., Sánchez D., Vila M.A. (2000). Fuzzy cardinality based evaluation of quantified sentences. International Journal of Approximate Reasoning, vol. 23, pp. 23-66.
9. Delgado M., Marín N., Sánchez D., Vila M.A. (2003). Fuzzy Association Rules: General Model and Applications. IEEE Transactions on Fuzzy Systems 11(2), pp. 214-225.
10. Huhtala Y., Karkkainen J., Porkka P., Toivonen H. (1998). Efficient Discovery of Functional and Approximate Dependencies using Partitions. Proc. of the 14th Int. Conference on Data Engineering, pp. 392-401.
11. Kivinen J., Mannila H. (1995). Approximate Dependency Inference from Relations. Theoretical Computer Science 149(1), pp. 129-149.
12. Pfahringer, B. and Kramer, S. (1995). Compression-based evaluation of partial determinations. In *Proc. First Int'l Conf. Knowledge Discovery and Data Mining (KDD'95)*, pages 234–239.
13. Sánchez, D. (1999). Adquisición de relaciones entre atributos en bases de datos relacionales. Ph. D. Thesis (in Spanish). University of Granada.

14. Shortliffe, E. and Buchanan, B. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23:351-379.
15. Zadeh L.A. (1965). Fuzzy Sets. Information and Control, 8, pp: 338-353.
16. Zadeh L.A. (1983). A computational approach to fuzzy quantifiers in natural languages. Computing and Mathematics with Applications, vol. 9, no. 1, pp. 149-184.

Table 1. Fuzzy similarity relations over categorical attributes in *Entry* table

STAV	2	3	4	5
1	0.4	0.2	0.4	0.0
2		0.2	0.6	0.0
3			0.2	0.0
4				0.0

VZDELANI	2	3	4	5
1	0.5	0.4	0.3	0.0
2		0.4	0.3	0.0
3			0.3	0.0
4				0.0

ZODPOV	2	3	4	5	6
1	0.3	0.2	0.1	0.1	0.0
2		0.2	0.1	0.1	0.0
3			0.1	0.1	0.0
4				0.3	0.0
5					0.0

TELAKTZA	2	3	4	5
1	0.3	0.3	0.2	0.0
2		0.4	0.2	0.0
3			0.2	0.0
4				0.0

AKTPOZAM	2	3	4
1	0.3	0.2	0.0
2		0.2	0.0
3			0.0

DOPRAVA	2	3	4	9
1	0.4	0.3	0.3	0.0
2		0.3	0.3	0.0
3			0.4	0.0
4				0.0

DOPRATRV	6	7	8	9
5	0.3	0.3	0.3	0.0
6		0.3	0.3	0.0
7			0.3	0.0
8				0.0

KOURENI	2	3	4	5	6	13
1	0.1	0.1	0.1	0.1	0.1	0.0
2		0.4	0.4	0.4	0.4	0.0
3			0.4	0.4	0.4	0.0
4				0.4	0.4	0.0
5					0.4	0.0
6						0.0

DOBAKOUR	8	9	10
7	0.4	0.4	0.4
8		0.4	0.4
9			0.4

PIVOMN	2	3	10	*Empty*
1	0.1	0.1	0.0	0.0
2		0.1	0.0	0.0
3			0.0	0.0
10				0.0