Fuzzy summarization of data using fuzzy cardinalities

Patrick Bosc, Didier Dubois, Olivier Pivert, Henri Prade, Martine De Calmès

Fuzzy Summarization of Data Using Fuzzy Cardinalities

[a]Patrick Bosc, [b]Didier Dubois, [a]Olivier Pivert, [b]Henri Prade, [b]Martine de Calmès

[a]IRISA-ENSSAT
Technopole Anticipa BP 447
22305 Lannion Cedex – France

[b]IRIT
Université Paul Sabatier – 118 route de Narbonne
31062 Toulouse Cedex 4 – France

Abstract

The paper investigates the following knowledge extraction method from a database. A set of a small number of attributes of interest are chosen. The idea is to figure out to what extent one of the attributes (the "output attribute" B) is influenced by the others (called "input attributes" A^i). Each of the input attribute domain are supposed to be equipped with a fuzzy partition made of a relatively small number (2 to 5) of fuzzy sets A^i_j or B_k respectively. These fuzzy partitions are supposed to be meaningful for the user and/or to be in agreement with the way the attribute values are scattered. For a given tuple (A^1_j,..., A^n_j, B_k), we compute the fuzzy cardinality of the set of items of the database which are A^1_j and ... and A^n_j and B_k , and of the set of items which are A^1_j and ... and A^n_j. From which we can easily obtain the fuzzy-valued confidence and support of the association rule "if x is A^1_j and ... and A^n_j, then x is B_k". In order to obtain more interesting association rules, the B_k are not chosen beforehand, but on the basis of a possibilistic case-based reasoning machinery which provides the fuzzy set of possible output values for a class of situations described in terms of labels A^i_j of the fuzzy partitions of the input attributes. Thus a set of fuzzy association rules summarizing a database can be obtained.

Keywords:
Fuzzy query, fuzzy cardinality, fuzzy prediction, fuzzy association rule.

1 Introduction

Querying a database in order to retrieve items satisfying some requirements, and extracting knowledge from a database are two types of operations which are often complementary for the users. Indeed in a querying session, a user may be interested in knowing what are the possible values of an attribute (e.g. the price) for some class of items specified in terms of other attributes (e.g. the size, the location of houses to be rent), or in knowing if there are many items available in the database which satisfy the requirements he has in mind). The two types of operations can be related at the processing level also. This will be illustrated in the following, where we are more particularly interested in summarizing a part of a database in terms of fuzzy association rules, which will be built by using flexible querying and fuzzy prediction techniques.

During the last years, the number and volume of databases have tremendously increased and the need for extracting some "condensed" information has received attention. The related research area, called knowledge discovery in databases (KDD) or data mining, aims at the discovery of useful information from large collections of data. The discovered knowledge can be rules describing properties of data, frequently occurring patterns, clusterings of the objects in the database, etc. Among the recent works, a great deal of attention has been paid to the discovery of a specific type of rules called association rules [1]. Association rules are of the type "when the properties A and B are satisfied in the data, then property C is also satisfied". Let us give a simple formalization of the problem.

Given a schema $R = \{A_1, ..., A_n\}$ of attributes with respective domains $D_1, ..., D_n$, a relation r can be represented by a Boolean matrix in which a row corresponds to a tuple and a column to an attribute value. An association rule about r is an expression of the form: $X \Rightarrow B$, where $X \subseteq R$ and $B \in R \setminus X$. The intuitive meaning of the rule is that if a row of the matrix r has a 1 in each column of X, then the row tends to have a 1 also in column B. This semantics is captured by the measures of frequency (or support) and confidence. Given $W \subseteq R$, we denote by sup(W, r) the frequency of

W in r: the fraction of rows that have a 1 in each column of W. The frequency of the rule $X \Rightarrow B$ in r is defined to be $sup(X \cup \{B\}, r)$ and the confidence of the rule is $sup(X \cup \{B\}, r)/sup(X, r)$.

In the discovery of association rules, the task is to find all rules $X \Rightarrow B$ such that the frequency of the rule is at least equal to a given threshold σ and the confidence of the rule reaches at least another threshold θ. In other words, one wants to obtain rules that are sufficiently frequent and valid.

As stated by R. Yager [2], the use of fuzzy logic seems particularly interesting in the context of knowledge discovery inasmuch as it allows to express properties about the current content of a database as statements of the natural language, thus providing knowledge that can be easily understood by nonexperts. In this paper, we propose an extension of the notion of an association rule, based on the aggregation of sufficiently close data into fuzzy sets and on the use of fuzzy cardinalities.

The paper is organized into five sections. Section 2 presents the notion of a fuzzy association rule based on fuzzy cardinalities. Section 3 describes an aggregation procedure which is performed on a non fuzzy relational database. It aims at replacing the initial database by a fuzzy relational database which summarizes it. This summarization is made in terms of labels of fuzzy partitions of the attribute domains which are supposed to be given and to be meaningful for the user. This aggregation procedure is augmented with the computation of the fuzzy cardinalities of the sets of items to which a given fuzzy label in a cell of the new relation is applicable. Section 4 describes how it should be possible to generate linguistic summaries of interest (in the form of fuzzy association rules) from the summarized fuzzy relational database built in the previous section and the associated fuzzy cardinalities. Section 5 shows how a possibilistic case-based prediction method can help to choose and use a meaningful partition for the output attribute domain. Section 6 reports some experiment results.

2 Fuzzy association rules based on fuzzy cardinalities

The starting point is to aggregate sufficiently close data into fuzzy sets, on the basis of fuzzy partitions which are meaningful for a user. Another key idea is to view fuzzy sets as a way for describing the different possible labellings that can be made by a user for borderline data.

2.1 Notations

Let R be a (non fuzzy) relation involving attributes A, B, C, ... In fact, we assume that the user is interested in possible summaries (association rules) involving a given subset of attributes. We only consider the projection of R on this subset, and for notational simplicity, we use a 3 element subset of attribute, say A, B and C, in the following, which is sufficiently general for discussing the main issues.

Let (a_i, b_j, c_k) denote a tuple of R(A, B, C) projected on attributes A, B and C. Let D_A, D_B, D_C be the attribute domains. We assume that each domain is equipped with a fuzzy partition $(A_1, A_2, ..., A_{na})$, $(B_1, B_2, ..., B_{nb})$, $(C_1, C_2, ..., C_{nc})$ respectively. Each fuzzy set in a partition is assumed to be normalized. Each partition is ordered, and a fuzzy set, say A_i, can only overlap with its predecessor A_{i-1} or its successor A_{i+1} (when they exist).

We further assume that a finite scale (with m + 1 levels) is used for assessing the membership degrees, namely $1 = \sigma_1 > ... > \sigma_m > 0$. Each level corresponds to a different possible understanding of A_r as the crisp level cut $(A_r)_{\sigma_i}$. The use of a finite scale then greatly facilitates the computation of fuzzy cardinalities, as it is shown in the following, without being a serious limitation in practice.

2.2 Principle

The approach corresponds to a straightforward extension of the usual definition of an association rule. Its principle is the following: the validity of the rule $(A, A_r) \Rightarrow (B, B_s)$ depends on the number of tuples which are A_r on the one hand and on the number of tuples which are A_r and B_s on the other hand.

For instance, using scalar cardinalities, the validity of the rule $(A, A_r) \Rightarrow (B, B_s)$ can be defined as:

$$|A_r \cap B_s| / |A_r| = \Sigma_{t \in R} \min(\mu_{A_r}(t), \mu_{B_s}(t)) / \Sigma_{t \in R} \mu_{A_r}(t)$$

which is nothing but a straightforward extension of the usual definition of the confidence.

Remark. It is worth noticing that the principle consisting in: i) rewriting the data by means of a more general vocabulary, and ii) trying to discover properties in the rewritten database, has also been advocated in a "non-fuzzy context". For instance, in [3], the authors use a hierarchy of (Boolean) concepts in order to "rewrite" a relation so as to discover different types of rules. See also [4] where the notion of a generalized association rule is introduced. Nevertheless, the fact of using linguistic labels (i.e., fuzzy sets) to rewrite the data allows to discover more robust rules. The key point, with a fuzzy partition, is that the borderline data are taken into account in each of the two classes, which decreases the sensitivity to the boundaries.

The general principle described above has been advocated by Yager [2], and more recently by Kacprzyk [5]. These authors use scalar relative cardinalities to compute the degree of validity of a summary (i.e., of a fuzzy association rule). Scalar fuzzy cardinality, which amounts to the addition of membership degrees, considers a collection of several elements with small membership grades whose sum is 1, as equivalent to one element with full membership for instance; this might be debatable or even misleading from a user point of view. In this paper, we present an alternative fuzzy approach based on the use of fuzzy cardinalities. This approach constitutes an extension of those described in [6, 7].

3 Summarizing a relation

From the relation R (restricted to A, B and C), we build a new relation R_{su} (for "R summarized") by a procedure involving two main steps which are now described. The idea is to perform a kind of information compression.

3.1 The labelling step

For each tuple $(a_i\ b_j\ c_k)$, we replace it by one or several tuples of fuzzy sets (A_r, B_s, C_t) subject to the constraint:

$$A_r(a_i) > 0,\ B_s(b_j) > 0,\ C_t(c_k) > 0.$$

Thus $(a_i\ b_j\ c_k)$ may be replaced by one tuple $(A_r\ B_s\ C_t)$ if all the three degrees of membership are equal to 1, or by several (up to $2^3 = 8$) in case one or several of the element(s) in the tuple belong to two fuzzy sets. For instance, if $A_r(a_i) = 1$, $B_s(b_j) = 0.8$, $B_{s+1}(b_j) = 0.2$, $C_{t-1}(c_k) = 0.6$, $C_t(c_k) = 0.4$, we give birth to the tuples:

$$(A_r\ 0.8/B_s\ 0.6/C_{t-1}),\ (A_r\ 0.8/B_s\ 0.4/C_t),$$
$$(A_r\ 0.2/B_{s+1}\ 0.6/C_{t-1}),\ (A_r\ 0.2/B_{s+1}\ 0.4/C_t)$$

where we keep track of the membership degrees (A_r stands for $1/A_r$). This corresponds to all the possible "readings" of the tuple (a_i, b_j, c_k) in terms of the vocabulary provided by the fuzzy partitions. In a data mining context, it is not necessary to store the summarized relation R_{su}. The only additional data that have to be stored are the fuzzy cardinalities, whose computation is described in the following subsection.

3.2 Fusion step and computation of fuzzy cardinalities

We want to know how many tuples from R are A_r, are B_s, are C_t, are A_r and B_s, ..., are A_r and B_s and C_t, and this, for all the fuzzy labels. In order to have a more accurate representation of the relation, fuzzy cardinalities are used instead of scalar ones. It is then necessary to compute the different cardinalities related to each linguistic label and to the diverse conjunctive combinations of these labels.

All the tuples of the form $(x/A_r\ y/B_s\ z/C_t)$ which are identical with respect to the three labels are fused into one tuple $(A_r\ B_s\ C_t)$ of R_{su}. At the same time, we compute the cardinalities F_{Ar}, F_{Bs}, F_{Ct}, F_{ArBs}, F_{ArCt}, F_{BsCt}, F_{ArBsCt} where F_{Ar} (resp. F_{Bs}, F_{Ct}, F_{ArBs}, F_{ArCt}, F_{BsCt}, F_{ArBsCt}) is a fuzzy set defined on the integers {0, 1, ...} which represents the fuzzy number of tuples which are somewhat A_r (resp. B_s, C_t, A_r and B_s, A_r and C_t, B_s and C_t, A_r and B_s and C_t) and which are fused into the considered tuple (for all the combinations of labels appearing in at least one tuple of R_{su}).

Each cardinality is computed incrementally in the following way. At the beginning F_{Ar} = 1/0. Let:

$$F_{Ar} = 1/0 + ... + 1/n\text{-}1 + 1/n + \lambda_1/(n+1) + ... + \lambda_k/(n+k) + 0/(n+k+1) + ...$$

be the current value of the fuzzy cardinality F_{Ar} with $1 > \lambda_1 \geq ... \geq \lambda_k > \lambda_{k+1} = 0$ and $n \geq 0$, $k \geq 0$. Let us recall that this expression represents a cardinality that possibly equals at least n to degree 1 and possibly equals at least (n+k) to degree λ_k [8].

Let us consider a new tuple whose A -value rewrites A_r. Let x' be the degree attached to A_r in this tuple. F_{Ar} must then be modified. If x' = 1, F_{Ar} becomes:

$$1/0 + ... + 1/n + 1/(n+1) + \lambda_1/(n+2) + ... + \lambda_k/(n+k+1) + 0/(n+k+2) + ...$$

If x' < 1 , there are two cases. Either $\exists$ i, x' = λ_i or not. If $\exists$ i, x' = $\lambda_i > \lambda_{i+1}$ then F_{Ar} is modified into:

$$1/0 + ... + 1/n\text{-}1 + 1/n + \lambda_1/(n+1) + ... + \lambda_i/(n+i) + \lambda_i/(n+i+1) + ... + \lambda_k/(n+k+1) + 0/(n+k+2) + ...$$

Otherwise, $\exists$ j, $\lambda_j > x' > \lambda_{j+1}$ (we assume that $\lambda_0 = 1$) and F_{Ar} becomes:

$$1/0 + ... + 1/n\text{-}1 + 1/n + \lambda_1/(n+1) + ... + \lambda_j/(n+j) + x'/(n+j+1) + \lambda_{j+1}/(n+j+2) + ... + \lambda_k/(n+k+1) + 0/(n+k+2) + ...$$

Note that the fuzzy cardinalities computed this way are such that $\forall$i, j, $\lambda_i \neq 0$, $\lambda_j \neq 0$, i > j $\Rightarrow \lambda_i \geq \lambda_j$. If, for the computation of F_{Ar} (resp. F_{Bs} et F_{Ct}), one takes into account the value x' (resp. y' the degree related to B_s, and z' the degree related to C_t), the computation of F_{ArBs} (resp. F_{ArCt}, F_{BsCt} and F_{ArBsCt}), takes into account the value min(x', y') (resp. min(x', z'), min(y', z'), min(x', y', z')), thus reflecting the fact that the tuple to fuse is both A_r and B_s (resp. A_r and C_t, B_s and C_t, A_r and B_s and C_t).

Let us notice that the maximum number of tuples that can be obtained in R_{su} is na $*$ nb $*$ nc, i.e. the product of the numbers of labels appearing in the considered partitions. Thus, the "summarized" relation can be significantly smaller than the original relation R for large relations.

4 Computing the validity of a summary

Fuzzy cardinalities computed as explained in Section 3 can be used to evaluate the validity of fuzzy association rules. In this approach, we assume that the user indicates the attributes A, B or C that he/she is interested in. Two general forms of rules can be thought of. The first type of rule follows the pattern "the tuples of R are A_r and B_s". This can be seen as a kind of degenerated rule with no attribute involved in the antecedent. The validity of such a rule corresponds to the extent to which the set of labels $\{A_r, B_s\}$ is frequent. The second type of rule corresponds to the pattern "the tuples which are A_r in R are also B_s". The computation of the validity of the rule (in terms of a fuzzy cardinality) is discussed in the following.

4.1 Computing fuzzy relative cardinalities

The two types of rules presented above relate to proportions. It will thus be necessary to compute fuzzy relative cardinalities. Let us denote by F^R_{Ar} (resp. F^R_{ArBs}, F^R_{ArBsCt}) the fuzzy proportion of the tuples of R which are "A_r" (resp. "A_r and B_s", "A_r and B_s and C_t"). This fuzzy number is obtained by dividing each (more or less) possible cardinality appearing in F_{Ar} (resp. F_{ArBs}, F_{ArBsCt}) by the number of tuples in R.

For instance, the fuzzy relative cardinality F^R_{ArBs} representing the fuzzy proportion of elements in R which are both A_r and B_s is obtained by replacing $F_{ArBs} = 1/0 + ... + 1/n + ... + \lambda_i/(n+i) + ...$ by $F^R_{ArBs} = 1/0 + ... + 1/(n/K) + ... + \lambda_i/((n+i)/K) + ...$ where K denotes the number of tuples in R. This means that the proportion of elements which are A_r and B_s at level λ_i is at least equal to (n+i)/K. Thus, the support of the fuzzy set F^R_{ArBs} is included in the unit interval.

4.2 Frequent sets of linguistic labels

Let us first consider the case where the user does not specify any fuzzy quantifier. From the tuples in R_{su}, it is possible to produce summaries of the form (A_r, B_s, C_t, F^R_{ArBsCt}) or relative to projections of R such as (A_r, B_s, F^R_{ArBs}). As in the nonfuzzy case, one produces the frequency, which is now a fuzzy number.

Of course, these summaries will be given to the user in a linguistic form expressing the variability of the cardinality (the fuzzy cardinality is then described by some proportions obtained for different levels of possibility). Let us recall that these levels of possibility corresponds to the more or less "elastic" interpretations that can be associated with the linguistic labels involved in the summary.

4.3 Fuzzy association rules

Let us now consider a fuzzy association rule of the form: "the tuples in R which are A_r are also B_s". We have available the fuzzy number of tuples in R which are A_r and B_s on the one hand, and the fuzzy number of tuples which are A_r on the other hand. Several approaches can be thought of, for instance:

- by analogy with the nonfuzzy case and with the approach based on scalar cardinality of fuzzy sets, it is possible to compute the fuzzy quotient $\rho = F_{ArBs} / F_{Ar}$, enforcing the constraint that the fuzzy number F_{ArBs} restricts a value which is less than or equal to the value restricted by F_{Ar} (see [9]). However, this quotient of such two fuzzy numbers may lead to a too imprecise result.

- a more interesting approach consists in determining how the proportion of tuples which are $(A_r)_\alpha$ and $(B_s)_\alpha$ with respect to those which are $(A_r)_\alpha$ changes when α varies, i.e., depends or not a lot on the interpretation of the linguistic labels (according to whether they are reduced to their cores or extended to their supports). Then, the computation involves three steps:

  1) for each α appearing in F_{Ar} or F_{ArBs}
     1.1) determine $|(A_r)_\alpha|$ and $|(A_rB_s)_\alpha|$ (which can be directly read on the fuzzy cardinalities computed as explained in section III.B);
     1.2) compute
     $$c_\alpha = |(A_rB_s)_\alpha| / |(A_r)_\alpha|$$
     corresponding to the confidence value of the rule when the α-level cut is used to interpret the labels;

  2) compute ρ' as the convex hull of the fuzzy number: $... + \alpha/c_\alpha + ... + 1/c_1$.

  Given a confidence threshold θ specified by the user, it is then possible to determine the highest degree α such that $\rho'_\alpha \geq \theta$. The association rule can then be expressed in a linguistic way, inasmuch as the α-cuts of A_r and B_s can be expressed in a linguistic way too. This approach in terms of α-level cuts of the cardinalities is in the spirit of [10] although these authors compute scalar cardinalities as weighted averages of cardinalities of α-cuts.

5 Case-based prediction

An order to determine meaningful clusters of values in the domain of the attribute involved in the conclusion part of the rules, a possibilistic case-based prediction method is used. Case-based reasoning, in general, assumes the following implicit principle: "*similar situations may give similar outcomes*". Thus, a similarity relation S between problem descriptions or situations, and a similarity measure T between outcomes are needed. This implicit CBR-principle can be expressed in the framework of fuzzy rules as, "*the more similar are the values of the situation attributes in the sense of S, the more possible the similarity of the values of the outcome attributes in the sense of T*" [11]. Given a situation s_0 associated to an unknown outcome t_0 and a current case (s, t), this principle enables us to conclude on the *possibility* of t_0 being equal to a value similar to t. This acknowledges the fact that, often in practice, a database may contain cases which are rather similar with respect to the problem description attributes, but which may be distinct with respect to outcome attribute(s). This emphasizes that case-based reasoning can only lead to cautious conclusions.

This can be modelled in terms of the possibility rule [12] "the more similar s and s_0, the more *possible* t and t_0 are similar". Then the fuzzy set F of possible values t' for t_0 with respect to case (s, t) is given by

$$F_{t_0}(t') = \min(S(s, s_0), T(t, t')).$$

However, here, we are interested in computing the set values which are close or equal to the value of attribute C for the tuples whose values of attributes A and B obey to some flexible requirements A_r and B_s. This means that $S(s, s_0)$ is replaced by $\min(A_r(a_i), B_s(b_j))$ and $T(t, t')$ by $T(c_k, c')$ where (a_i, b_j, c_k) is a tuple of the base.

As it can be seen, what is obtained is the fuzzy set $T(c_k, \cdot)$ of values c' which are T-similar to c_k, whose possibility level is upper bounded by the global degree $\min(A(a_i), B(b_j))$. The max-based aggregation of the various contributions, obtained from the evaluation of each tuple in the base, acknowledges the fact that each new comparison may suggest new possible values for C. Thus, we obtain the following fuzzy prediction set P of possible values c' :

$$P(c') = \max_{(ai, bj, ck) \in R} \min(\min(A(a_i), B(b_j)), T(c_k, c'))$$

P is the fuzzy set of possible values of attribute C for the items in the base for which the attributes values of A and B are restricted by A_i and B_j respectively. The similarity relation T on the domain of attribute C is used here for providing an understanding of the precise attribute values c_k in the tuples of the base as interchangeable with any value close to them. This contributes to make P smoother.

6 Experiment results

The experiment was carried out on the IRIT Platform for Experimentation and Research in the Treatment of Information (PRETI) accessible at http://www.irit.fr/PRETI. PRETI has a database containing the description of more than 600 houses which can be rent for vacations. They are described in terms of about 20 attributes. The results of the experiment are illustrated in two parts.

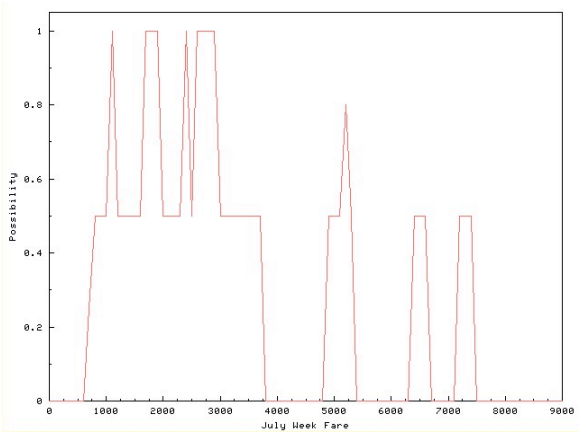

Figure 1.

Fuzzy ranges for **July Week Fare (FF)** taken from the distribution of Fig. 1	**Cumulated number of houses with matching degree**
(600, 1000, 1200, 1500)	1/1., 4/0.4
(1400, 1600, 2000, 2300)	4/1., 13/0.4, 17/0.2, 18/0.
(2200, 2300, 3000, 3100)	9/1., 15/0.4, 16/0.2, 17/0.
(3000, 3100, 3700, 3800)	6/0.4, 8/0.2
(4800, 5100, 5300, 5400)	1/0.8, 2/0.4
(6300, 6400, 6600, 6700)	1/0.4
(7100, 7200, 7400, 7500)	1/0.4

Table 1.

First, we consider a query focusing on houses whose comfort is about 2 stars ('about' means that 2 stars has possibility 1, while 1 star and 3 stars have possibilty 1/2), and which are close to the sea ('close' means less than 5km with possibility 1, and a linearly decreasing possibility between 5km and 10km, possibility is 0 for distances greater or equal to 10km). Figure 1 exhibits the result of the possibilistic prediction for the July weekly fares (still in French Francs!) of the houses corresponding to this fuzzy query. The fuzzy similarity relation which is used for the fare has a peak corresponding to perfect identity and a support between –200 and +200 FF. Table 1 provides a complementary information by indicating the fuzzy cardinality of the fuzzy sets of houses close to the sea, with about a 2-star comfort, and having a price around one of the peaks of the distribution given in Figure 1. Note that if a peak is too large it may lead to slice it into several ranges of prices. The fuzzy cardinality is presented by giving the cardinality followed after the '/' by the level of the corresponding level cut. When the cardinality remains unchanged for several level cuts, only the highest level cut appears. As expected the extreme prices correspond to very small numbers of houses. It can be also noticed that the cardinality number may have a substantial increase below .6, this mainly corresponds to the impact of taking into consideration 1 star or 3 stars houses also with only a .5 possibility degree. Indeed, generally speaking, the interest of using fuzzy cardinality is to show the variations induced by the relaxation of flexible requirements.

Table 2 provides examples of fuzzy associations rules obtained by the method described in this paper. Here, for simplicity, we use a crisp 2-part partition for comfort, namely { {1 star, 2 stars},{3 stars, 4 stars} }, and a 3-part fuzzy partition for the distance to the sea (close, not too far, far). 'Close', 'not too far', 'far' are represented by the following trapezoids (0, 0, 5, 10), (5, 10, 30, 35), (30, 35, 100, 100) respectively. We have only kept the rules corresponding to more than 10 hou-

ses in the database. The fuzzy sets used for the July week fare are given under the form of trapezoids and are taken from fuzzy predictions like the one of Figure 1. These predictions are not given due to the lack of space. The cardinality of the data base which is used for computing the support is 444, which is the number of data for which we have a complete and precise information about the three considered attributes in the base. The number before the '/' is the support, or the confidence, and the corresponding level cutting is given after. Again it enables us to show if some rules are sensitive to changes of the scope of the fuzzy sets used in the partitioning of the attribute domains.

Comfort	Dist. to sea	July week fare	Support	Confidence
{1,2}	close	(1400, 1500, 1900, 2000)	0.05/1., 0.06/0.8, 0.07/0.4	0.44/1., 0.46/0.6, 0.45/0.4, 0.5/0.2, 0.51/0.001
		(2200, 2300, 2900, 3100)		0.44/1., 0.40/0.8, 0.39/0.6, 0.38/0.4, 0.34/0.2, 0.33/0.001
	not too far	(1000, 1200, 1900, 1900)	0.14/1., 0.16/0.8, 0.17/0.6, 0.19/0.4, 0.21/0.2	0.47/1., 0.49/0.8, 0.5/0.4, 0.51/0.2
		(1900, 1900, 2600, 2800)		0.36/1., 0.37/0.8, 0.38/0.6, 0.37/0.4
		(3000, 3100, 3300, 3500)		0.16/1., 0.14/0.8, 0.12/0.4, 0.11/0.2
	far	(1000, 1100, 2000, 2100)	0.39/1., 0.4/0.8, 0.41/0.6, 0.42/0.4, 0.43/0.2	0.66/1.
		(2000, 2100, 2900, 3000)		0.30/1.
{3,4}	close			
	not too far	(1900, 2100, 2700, 2800)	0.07/1., 0.08/0.6, 0.09/0.4, 0.1/0.01	0.32/1., 0.36/0.8, 0.35/0.6, 0.33/0.4, 0.34/0.2, 0.33/0.001
		(2700, 2800, 3500, 3600)		0.20/1., 0.18/0.8, 0.24/0.6, 0.28/0.4, 0.27/0.2, 0.26/0.001
		(3500, 3600, 4100, 4200)		0.30/1., 0.27/0.8, 0.24/0.6, 0.25/0.4, 0.24/0.2
	far	1500, 1700, 2600, 2700	0.23/1., 0.24/0.4	0.39/1., 0.37/0.4, 0.38/0.2
		2600, 2700, 3400, 3500		0.40/1., 0.41/0.6, 0.42/0.4

Table 2.

7 Conclusion

This paper has outlined a new approach to the linguistic summarization of data bases. The basic ideas are i) to use fuzzy partitions of attribute domains which are meaningful for the user (since fuzzy partitions are more compatible with a linguistic labelling), ii) to perform a "soft compression" of the data base and then to exploit it for evaluating potential summaries, iii) this evaluation is made by computing fuzzy cardinalities which account for the possible variations of the interpretation of the labels, and iv) to use a fuzzy prediction tool for selecting fuzzy sets of interest in the output domain. What is obtained are the (fuzzily known) proportions of elements satisfying fuzzy specifications.

References

[1] R. Agrawal, T. Imielinski, and A. Swami (1993). Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD'93*, 207-216, 1993.

[2] R.R. Yager (1996). Database discovery using fuzzy sets. *Int. J. of Intelligent Syst.*, 11(9), 691-712, 1996.

[3] J. Han and Y. Fu (1995). Discovery of multiple-level association rules from large databases. *Proc. VLDB'95*, 420-431, 1995.

[4] R. Srikant and R. Agrawal (1995). Mining generalized association rules. In *Proc. VLDB'95*, 407-419, 1995.

[5] J. Kacprzyk (1999). Fuzzy logic for linguistic summarization of databases. *Proc. FUZZ-IEEE'99*, 813-818, 1999.

[6] D. Dubois and H. Prade (2000). Fuzzy sets in data summaries – Outline of a new approach. *Proc. IPMU'00*, 1035-1040, 2000.

[7] P. Bosc, D. Dubois, O. Pivert, and H. Prade (2000). Résumés de données et ensembles flous – Principe d'une nouvelle approche", *Actes des Rencontres Francophones sur la Logique Floue & Applic.*, p.333-340, 2000.

[8] D. Dubois and H. Prade (1985). Fuzzy cardinality and the modeling of imprecise quantification. *Fuzzy Sets and Systems*, 16, 199-230, 1985.

[9] D. Dubois and H. Prade (1988). *Possibility Theory*, Plenum Press, 1988.

[10] M.J. Martin-Bautista, D. Sanchez, M.A. Vila, and H. Larsen (2001). Measuring effectiveness in fuzzy information retrieval. In *Flexible Query Answering Systems – Recent Advances*, H. Larsen, J. Kacpryk, S. Zadrozny, T. Andreasen, H. Christiansen, eds., Physica Verlag, 396-402, 2001.

[11] D. Dubois, E. Hüllermeier, H. Prade (2000) Flexible control of case-based prediction in the framework of possibility theory. In: *Advances in Case-Based Reasoning* (Proc.5th Eureopean Worshop, EWCBR 2000, 6-9 sept.), Trento, Italie. (E.Banzieri, L.Portinal, Eds., LNCS n°1898), Springer-Verlag, Berlin Heidelberg , 61-73.

[12] D. Dubois, H. Prade (1996) What are fuzzy rules and how to use them. *Fuzzy Set and Systems*, 84, 169-185.