Chapter XXIII

Fuzzy Classification on Relational Databases

Andreas Meier
University of Fribourg, Switzerland

Günter Schindler
Galexis AG, Switzerland

Nicolas Werro
University of Fribourg, Switzerland

ABSTRACT

In practice, information systems are based on very large data collections mostly stored in relational databases. As a result of information overload, it has become increasingly difficult to analyze huge amounts of data and to generate appropriate management decisions. Furthermore, data are often imprecise because they do not accurately represent the world or because they are themselves imperfect. For these reasons, a context model with fuzzy classes is proposed to extend relational database systems. More precisely, fuzzy classes and linguistic variables and terms, together with appropriate membership functions, are added to the database schema. The fuzzy classification query language (fCQL) allows the user to formulate unsharp queries that are then transformed into appropriate SQL statements using the fCQL toolkit so that no migration of the raw data is needed. In addition to the context model with fuzzy classes, fCQL and its implementation are presented here, illustrated by concrete examples.

MOTIVATION

The fuzzy logic theory proposed by Zadeh (1965) is based on intuitive reasoning and takes into account human subjectivity and imprecision. Unlike statistical data mining techniques such as cluster or regression analysis, fuzzy logic enables the use of nonnumerical values and introduces the notion of linguistic variables (Zadeh, 1975a, 1975b, 1975c). Using linguistic terms and variables hides
Fuzzy Classification on Relational Databases

the complexity of the domain and permits a more intuitive and human-oriented querying process in different application domains.

Information systems are based on huge data collections, mostly stored in relational databases (Berson, Smith, & Thearling, 2000). Due to this information overload, it is becoming difficult to analyze huge amounts of data and to generate business decisions (Edmunds & Morris, 2000). To address this, the fCQL (fuzzy classification query language) toolkit for classification, analysis, and decision support has been developed (Meier, Savary, Schindler, & Veryha, 2001; Meier, Werro, Albrecht, & Sarakinos, 2005; Werro, Meier, Mezger, & Schindler, 2005). The toolkit reduces the complexity of business data and extracts valuable hidden information through a fuzzy classification. The main advantage of a fuzzy classification compared to a classical one is that an element is not limited to a single class but can be assigned to several classes. Furthermore, each element has one or more membership degrees that illustrate to what extent this element belongs to the classes it has been assigned to. This notion of membership not only provides a better description of the elements, it also helps to reveal both the potential and the possible weaknesses of the elements under consideration.

Another important issue, considering the size and the security needs of the data collection, is that no modification of the underlying databases or migration of the existing data has to be undertaken. The fuzzy classification is achieved by an extension of the relational database schema and, in this way, can directly operate on the underlying databases. The fCQL toolkit provides graphic support for defining fuzzy classes, linguistic variables, terms, and membership functions. In addition, it transforms fuzzy queries into SQL (structured query language) statements allowing business managers to formulate and analyze unsharp queries at the linguistic level (Meier & Werro, 2006; Meier et al., 2005; Werro, Stormer, & Meier, 2005).

The remainder of this chapter is organized as follows. The section “Conceptual Basis” contains the conceptual aspects; it extends the relational database schema by a context model, a context-based relational algebra, fuzzy classes, and a hierarchical decomposition. The fuzzy classification query language is presented together with a discussion about related research work. “Implementation” illustrates implementation aspects of the fCQL toolkit, namely, architectural issues, the user interface, and fuzzy classification meta-tables. The section titled “Applications” demonstrates successful project results gained in customer relationship management. Finally, the last section gives the conclusion and discusses further research questions.

CONCEPTUAL BASIS

Context Model and Fuzzy Classification

For the classification of tuples in a relational database, two approaches can be distinguished: sharp and fuzzy. In the case of sharp classification, each assignment of tuples to a class is dichotomous; that is, the membership function of a tuple to a class is 1 for inclusion and 0 for exclusion.

A fuzzy approach allows the membership function to take values between 0 and 1. A fuzzy classification therefore assigns to each tuple a degree and expresses how strongly it belongs to that class. More precisely, if \( X \) is a set then the set \( \Lambda = \{ (x, \mu_A(x)) | x \in X \} \) is called a fuzzy set with the membership function \( \mu_A : X \rightarrow [0,1] \).

Figure 1 illustrates examples of handling sharp and fuzzy sets in a relational database. If customers are sharply classified regarding their turnover, SQL allows precise queries and results. For example, all customers with a turnover between 500 and 1,000 Euro can be evaluated. If the customers are divided into sharp sets, a sharp judgment of customers takes place at the border of this class: Customers with a low turnover have spent between 0 and 499 Euro, and customers with a high turnover between 500 and 1,000 Euro. This will lead