Knowledge Discovery in Fuzzy Databases Using Attribute-Oriented Induction

Rafal A. Angryk, Frederick E. Petry
Electrical Engineering and Computer Science Department
Tulane University
New Orleans, LA 70118
USA
{angryk, fep}@eecs.tulane.edu

Abstract

In this paper we analyze an attribute-oriented data induction technique for discovery of generalized knowledge from large data repositories. We employ a fuzzy relational database as the medium carrying the original information, where the lack of precise information about an entity can be reflected via multiple attribute values, and the classical equivalence relation is replaced with relation of the fuzzy proximity. Following a well-known approach for exact data generalization in the ordinary databases [1], we propose three ways in which the original methodology can be successfully implemented in the environment of fuzzy databases. During our investigation we point out both the advantages and the disadvantages of the developed tactics when applied to mine knowledge from fuzzy tuples.

1. Introduction

The majority of current works on data mining describes the construction or application of algorithms performing complex analyses of stored data. Despite the predominant attention on this phase of analysis, because of the extensive volume of data in databases, techniques allowing conversion of raw data into condensed representations has become a practical necessity in many data-mining projects [2-3].

Attribute-Oriented Induction (AOI) [1, 4-12] is a descriptive database mining technique allowing such a transformation. It is an iterative process of grouping of data, enabling hierarchical transformation of similar itemsets stored originally in a database at the low (primitive) level, into more abstract conceptual representations. It allows compression of the original data set (i.e. initial relation) into a generalized relation, which provides

concise and summarative information about the massive set of task-relevant data.

To take advantage of computationally expensive analyses in practice it is often indispensable to start by pruning and compressing the voluminous sets of the original data. Continuous processing of the original data is excessively time consuming and might be expendable, if we are actually interested only in information on abstraction levels much higher than directly reflected by the technical details stored usually in large databases (e.g. serial numbers, time of transactions with precision in seconds, detailed GPS locations, etc.). Simultaneously, the data itself represents information at multiple levels (e.g. Tulane University represents an academic institutions of Louisiana, the university is located in the West South, which is a part of the North American educational system, etc.), and is naturally suitable for generalization (i.e. transformation to a preferred level of abstraction).

Despite the attractive myth of fully automatic data-mining applications, detailed knowledge about the analyzed areas remains indispensable in avoiding many fundamental pitfalls of data mining. As appropriately pointed out in [6], there are actually three foundations of effective data mining projects: (1) the set of data relevant to a given data mining task, (2) the expected form of knowledge to be discovered and (3) the background knowledge, which usually supports the whole process of knowledge acquisition.

Generalization of database records in the AOI approach is performed on an attribute-by-attribute basis, applying a separate concept hierarchy for each of the generalized attributes included in the relation of task-relevant data. The concept hierarchy, which in the original AOI approach is considered to be a part of background knowledge (which makes it the third element of the above mentioned primitives), is treated as an indispensable and crucial element of this data mining technique. Here we investigate the character of knowledge available in the similarity and proximity relations of fuzzy databases and analyze possible ways of its application to the generalization of the information originally stored in fuzzy tuples.

In the next sections we introduce attribute-oriented induction, and briefly characterize crisp and fuzzy approaches to the data generalization; we will also discuss the unique features of fuzzy database schemas that were utilized in our approach to attribute-oriented generalization. In the third part we will present three techniques allowing convenient generalization of records stored in fuzzy databases. The increase in efficiency of these methods over the originally proposed solutions is achieved by taking full advantage of the knowledge about generalized domains stored implicitly in fuzzy database models. Then we will propose a method that allows