Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining

Pirjo Moen

Department of Computer Science
P.O. Box 26, FIN-00014 University of Helsinki, Finland
Pirjo.Moen@cs.Helsinki.FI, http://www.cs.Helsinki.FI/Pirjo.Moen/

PhD Thesis, Series of Publications A, Report A-2000-1
Helsinki, February 2000, vi + 190 + 9 pages
ISSN 1238-8645, ISBN 951-45-9102-X

Abstract

In data mining and knowledge discovery, similarity between objects is one of the central concepts. A measure of similarity can be user-defined, but an important problem is defining similarity on the basis of data. In this thesis we consider three kinds of similarity notions: similarity between binary attributes, similarity between event sequences, and similarity between event types occurring in sequences.

Traditional approaches for defining similarity between two attributes typically consider only the values of those two attributes, not the values of any other attributes in the relation. Such similarity measures are often useful, but unfortunately they cannot describe all important types of similarity. Therefore, we introduce a new attribute similarity measure that takes into account the values of other attributes in the relation. The behavior of the different measures of attribute similarity is demonstrated by giving empirical results on two real-life data sets.

We also present a simple model for defining similarity between event sequences. This model is based on the idea that a similarity notion should reflect how much work is needed in transforming an event sequence into another. We formalize this notion as an edit distance between sequences. Then we show how the resulting measure of distance can be efficiently computed using a form of dynamic programming, and also give some experimental results on two real-life data sets.

As the third case of similarity notions, we study how similarity between types of events occurring in sequences could be defined. Intuitively, two event types are similar if they occur in similar contexts. We show different possibilities for how a context of an event can be extracted from a sequence. Then we discuss ways of defining similarity between two event types by using sets of the contexts of all their occurrences in given sequences. Results of experiments on the event type similarity with different measures are described on both synthetic and real-life data sets.

Computing Reviews (1998) Categories and Subject Descriptors:
H.2.8 Database Applications: Data mining
H.3.3 Information Search and Retrieval
I.2.6 Learning

General Terms:
Algorithms, Experimentation, Theory

Additional Key Words and Phrases:
Data mining, Knowledge discovery, Similarity, Distance, Relational data, Binary attributes, Event sequences, Event types in sequences