DataMine - Interactive Rule Discovery System

T. Imielinski A. Virmani
Department of Computer Science,
Rutgers University,
New Brunswick, NJ 08903
email: {imielins, avirmani}@cs.rutgers.edu

1 Introduction

DataMine is a statistical database mining system with strong emphasis on interactiveness and nice graphical representation of information produced. It also supports an offline mode of discovery, and provides an extensive API which allows users to write "mining applications" just as easily as routine database applications. The central idea is to perform discovery with a "human in the loop" guiding the system using his initial hypothesis and the feedback from the system. Users can pose a rule-query against a rulebase and the system can generate all rules matching their query. The rulebase could either be pregenerated (using offline mode) or could be realized in real-time as the discovery progresses.

Rules generated by the system are of the form:

Body −> Consequent

where Body is a conjunction of the elementary predicates of the form (A=a), where A is an attribute and a is a value from the attribute domain of A. Consequent is a single elementary predicate. Each rule can have several parameters like support, confidence, atypicality, color etc. (the definitions have been left out) which can also be used by the user in framing the rule query.

For continuous attributes, the system also allows the user some control in deciding how they are discretized. It also allows for the creation of extra attributes at run time which can then be used in queries like the rest.

2 System operation

DataMine operates in two modes – **Incremental mining** and **Background mining.** In the Incremental mode, the main aim of the system is to keep the user's interest alive by producing results in real time, and then refining them with some guidance from the user. The Rulebase is "hot wired" to the display interface which directs the user's attention towards the "hot tuples" as the rules are generated.

In the background, or offline mode, the system generates all possible rules within the constraints set by the user, and the rule-patterns he is looking for. The user can then, at a later time, query this rulebase, or sift through it using a smart browser, and pick out the rules interesting to him to use as hypotheses for refining further.

3 Platform

The mining engine has been written in C (about 15,000 lines of code) and the interface has been done in Tcl-Tk (about 20,000 lines of code). The system currently runs on Solaris 2.4 on a Sparc 2. It supports multiple input and output channels for data (Sybase, flat file database, etc.) and the architecture can incorporate others with almost no change in code.

4 Demonstration Description

We are currently experimenting with two large data sets - health care management data, and historical stock market data. We plan to use these data sets for our demonstration. We will demonstrate the process of "mining around" a rule - where the system can improve a user's first guess by extending it with extra predicates. We will also show how the system and the user can guide each other during the discovery process and demonstrate the relative edge provided by this "human in the loop" paradigm, which allows the user much finer grain control over the depth and direction of the mining process.

5 Acknowledgments

We would like to thank Chuan Li, Joe Orost, Vishal Desai, and Sarojini Davis for their help in the ongoing development of DataMine.

SIGMOD ' 95,San Jose , CA USA
© 1995 ACM 0-89791-731-6/95/0005..$3.50