Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 1996 Proceedings

Americas Conference on Information Systems (AMCIS)

8-16-1996

Understandable Database Mining In Imprecise Domains

Lawrence J. Mazlack
University of California, mazlack@uc.edu

Understandable Database Mining In Imprecise Domains

Lawrence J. Mazlack
Computer Science Division, EECS Department
University of California
Berkeley, California 94720-1776
mazlack@uc.edu

An unsupervised database mining methodology is under development. A particular goal is that the process be understandable. This methodology tries to identify relationships having the most information value through a progressive reduction of cognitive dissonance. This work is dependent on soft computing tools.

INTRODUCTION

Database mining seeks to discover noteworthy, unrecognized associations between database items. This can be thought of as discovering 'interesting' pattern.

Databases have significant amounts of stored data. Much of the data is implicitly or explicitly imprecise. The data is valuable because it was collected to explicitly support enterprise activities. There could well be undiscovered, valuable relationships. The issue is best recognizing them.

Database mining supports the creation of knowledge from collected data. Instead of the modelling and implementation of existing, explicit human knowledge; data mining elicits knowledge that is implicit in the databases. In a sense, the implicit knowledge is knowledge not yet available for use.

Much of the existing database mining work has used supervised discovery. However, supervised search limits the results as it determines in advance the subjects that are of interest. This is counter-intuitive to the general goal of conducting discovery to find unexpected, interesting things. In keeping with the more ambitious database discovery goals, this work considers unsupervised discovery.

One problem with unsupervised search is combinatoric explosion. To consider fully the interrelationships between all the attributes is computationally prohibitive. Possible heuristic help comes from the realization that many of the combinations do not appear to have much information value. This leads to this work's controlling heuristic focusing on increasing coherence and decreasing dissonance.

The heuristic assumption is that reducing cognitive dissonance increases useful information. The speculation is that database exploration can be accomplished through a progressive reduction of cognitive dissonance.

BACKGROUND

A database is a collection of data from which different facts can be efficiently retrieved. Numerous databases have significant amounts of stored data. Although simple statistical techniques for data analysis have been developed, advanced intelligent data analysis techniques are not yet mature. As a result, there is a growing gap between data generation and data understanding. At the same time, there is a growing realization and expectation that data, intelligently analyzed and presented, should be a valuable resource.

Knowledge And Understanding

There are several ways to classify database mining work. One way is by the amount of semantic domain knowledge used. Supervised methods may use semantic knowledge. Semantic knowledge is used to improve efficiency and guide discovery. Unsupervised methods use non-semantic heuristics. Heuristics are often from the following areas: information theory [Agrawal, 1993] [Smyth, 1991], fuzzy sets [Zadeh, 1972], rough sets [Ziarko, 1991], and statistics [Langley, 1992].

Fully recognizing authentically interesting information from what is discovered requires domain knowledge. However, embedding domain knowledge in a context sensitive structure is difficult. Conversely, it is also difficult to non-semantically recognize the interesting. Various techniques [Klemettienen, 1994] [Hoschka, 1991] [Piatetsky-Shapiro, 1994] have been proposed. In general, non-semantic definitions of interestingness need more development.

Uncertainty

Lack of crispness is an issue. Most mining techniques implicitly assume that the data is clean and that the data can eventually be clustered using a precise metric. However, the reality is that data is often imperfect, that some values are inherently imprecise. Another concern is that the group boundaries may not be crisp.

Another source of uncertainty is that partitions can and should be formed on competing taxonomical structures. For example, if we were mining against hospital data, we might discover interesting information about both patients and services provided.

One class of solutions considers impreciseness to be an impediment to correctness and uses a variety of techniques to reduce data variability. Techniques include filtering, thresholds, statistics, and partial matching [Chan, 1989] [Chen, 1985] [Dzeroski, 1991] [Mathews, 1993].

The other approach considers some data to be inherently non-crisp and seeks to work with the data while retaining its non-crispness. For other purposes, there is a considerable history in using fuzzy techniques [Zadeh, 1976] [Bezdek, 1992]. Other complementary approaches include Dempster-Shafer Theory [Shafer, 1976] and Rough Sets [Pawlak, 1991].

Intelligibility

Human understanding of results is not a general goal of machine learning. However, it is of importance in database mining. This is because one class of ultimate 'customers' of mining products are humans. A parallel can be drawn with knowledge-based systems (KBS). One reason that KBS originally came into use was that KBS rules and procedures were humanly comprehensible and explainable. Similarly, human decision makers generally require comprehensible decision making material.

Database mining people must ultimately deal with comprehensibility, as well as with other ill-defined notions such as interestingness or usefulness. They indeed sometimes attempt to give precise definitions to these notions. However, many of these definitions are ad hoc. There is a need [Kodratoff, 1995] for a definition of comprehensibility that could be grounded by psychological studies.

Mechanizing a complex process may not require comprehensible intermediate results. But, the whole process, as well as the decisions it helps to make, requires a high level of comprehensibility. One aspect negatively affecting comprehensibility is what physiologists and software engineers call cognitive dissonance. In simplified terms, this means that comprehensible things are those that are simply stated; the less complex, the more comprehensible.

Increasing Cohesion

Reducing database disorder may increase comprehension through increasing cohesion. But, before computationally decreasing something, it is necessary to measure what is being decreased. Because disorder metrics measure a heuristic, they are by their very nature, implicitly imprecise. They are often explicitly imprecise as well.

Many different disorder measures have been suggested. Many of these have focused on an attribute's fit with a classification scheme. Possibly the most common metric is Quinlan's [1986] entropy, that in turn is based on Shannon's information theory. Perhaps, the broadest term is uncertainty. Some of these metrics are total uncertainty [Pal, 1993], dissonance or conflict [Yager, 1983], confusion [Hohle, 1981], non-specifity [Higashi, 1983] [Dubois, 1985], discord [Klir, 1990], global uncertainty [Lamata, 1987], total uncertainty [Klir, 1990].

Approaching the problem using a cohesion enhancement paradigm has the benefit of avoiding the intractability of combinatorial complexity that arises in attempting to discover relationships between all elements. Cohesive information is also easily understood. The most cohesive groups of data can be thought of as information 'nuggets'. This is suggested by Figure 2.1 showing the reduction of a data collection to a cohesive information 'nugget'.

CONCEPTUAL OVERVIEW

The underlying research question is the extent to which data mining methods must necessarily employ special domain knowledge. The speculation is that increasing cohesion will aid in database mining.

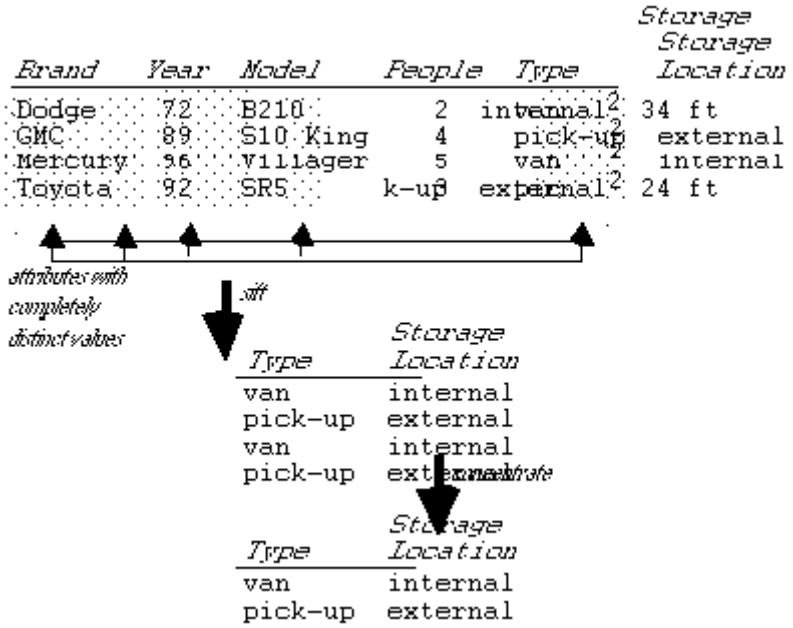

Figure 2.1. Simple example illustrating discovery through cohesive data concentration.

Rationale

There are a number of different techniques that can be used to mine a database. Many of them use some form of supervised search; i.e., before and/or during mining, they receive direction as to the target. For example, in a medical database, the question could be identifying anomalous physician services.

However, supervised search limits the results because it is necessary to determine in advance the subjects that are of interest. This is almost counter-intuitive to the general goal of conducting database mining of finding unexpected, interesting things. Heraclitus observed "If you do not expect the unexpected, you will not find it."

On the other hand, unsupervised search has a combinatorial explosion problem. In a typical database, there may be hundreds of attributes. To consider the interrelationships between all of them is prohibitive. Possible heuristic help comes from the realization that many of the combinations do not appear to have much information value.

The main metric in this effort is the dissonance within a data partition. This is opposed to the more common machine learning approaches that attempt to discover metrics to best discriminate between individual data items as with Quinlan's [1986] entropy disorder metric. Quinlan sought to determine the attribute sequence that most efficiently discriminates among the database items. The sequences are often captured in classification trees that do not have an intelligibility goal.

This work suggests a methodology of unsupervised search that is based on identifying groups of data that have the most information value. The speculation is that discovery can be accomplished through a progressive reduction of cognitive dissonance. The heuristic assumption is that reducing cognitive dissonance increases information. This work is not particularly interested in capturing the partitioning sequence, although it may be a process artifact.

This work has the specific motivation of providing information intelligible to the human user. The view is that intelligible results are essentially cohesive. These are results that have limited cognitive dissonance.

Partitioning

A major issue is whether it is better to attempt to functionally combine all items of a table into a single disorder measure; or, to focus on a cohesion metric of individual attributes. There are many possible ways of measuring cohesion. Focusing on the coherence of only a single attribute may be satisfactory. Or, it might be better to consider combinations of attributes. Or, there may be other coherence metrics that may be successfully applied. Several different methods will be explored during this investigation. The most straight forward, comprehensible, and computationally attainable is to use the distinctiveness of the values within one or more attribute.

Partitioning On Crisp Data

If a table of data T is made up of elements ti,j where i represents the row (or tuple) of data and j represents an attribute of the database, T is partitioned by placing the rows into different partitions. The partitions are constructed so that the coherence of the resulting partitions is greater than the coherence of the initial data table.

Partitions can be formed using the distinctness of attributes with crisp data attributes. T can be partitioned on the distinctiveness of attributes so that each partition only contains only a single value for a particular attribute. For example, the table in Figure 3.1 is split into two sub-partitions on ti,2. T could have also been partitioned into two different sub-partitions on ti,6. However, partitioning on ti,2 also partitions on ti,7; thus, the partitioning is accomplished on two attributes as opposed to one. This presents a possibly useful partitioning heuristic of partitioning on the maximum count of attributes. Notice three partitions could have been formed using ti,1. The preferable count of partitions (i.e., more, less, some count) is a research question.

Partitioning On Non-Crisp Data

Partitioning may be extended to include non-crisp values by using linguistic values. One possible formation of linguistic variables is on granularized, ordered values. While it is more common to form linguistic variables on scalar data; it may also be useful to form linguistic variables on non-scalar, ordered values. Fuzzy membership [Zadeh, 1972] will be used as a tool in granularization.

In an ordered sequence of linguistic variables, a possible partitioning heuristic is to form the partition in a linguistic variable ordered between two other linguistic variables.

The approach's utility becomes clear when considering data tables. It surfaces when maximizing the count of attributes that are used to form a partition. For example, given Figure 3.2, the partition on the maximum count of attributes is on ti,1&2&3. This partitions on one attribute with crisp values and two attributes with non-crisp values. If only attributes with crisp values were used, the partitioning would be on ti,4&5.

Figure 3.1 Partitions formed on crisp data attributes. Partitions formed on ti,2 and ti,7.

Figure 3.2 Partitions formed on both crisp and non-crisp data attributes. Partitioning on ti,1, ti,2, ti,3.

Forming partitions on multiple attributes in a large database on multiple attributes is not computationally simple. When adding granularized semantic variables to the mix, the problem becomes computationally more complex. Help may be available from the operations research discipline.

EPILOGUE

The goal of this work is to conduct understandable database mining. The research quest is to determine which, if any dissonance reduction techniques hold promise in discovering interesting information as a product of database mining. Approximate reasoning techniques to address inherently non-crisp numbers and imprecise partitions will be used. This is being done to

BIBLIOGRAPHY

Due to space considerations, the bibliographic sources have not been shown. They are available from the author.