Constrained Data Mining

Brad Morantz
Science Applications International Corporation, USA

INTRODUCTION

Mining a large data set can be time consuming, and without constraints, the process could generate sets or rules that are invalid or redundant. Some methods, for example clustering, are effective, but can be extremely time consuming for large data sets. As the set grows in size, the processing time grows exponentially.

In other situations, without guidance via constraints, the data mining process might find morsels that have no relevance to the topic or are trivial and hence worthless. The knowledge extracted must be comprehensible to experts in the field. (Pazzani, 1997) With time-ordered data, finding things that are in reverse chronological order might produce an impossible rule. Certain actions always precede others. Some things happen together while others are mutually exclusive. Sometimes there are maximum or minimum values that can not be violated. Must the observation fit all of the requirements or just most. And how many is "most?"

Constraints attenuate the amount of output (Hipp & Guntzer, 2002). By doing a first-stage constrained mining, that is, going through the data and finding records that fulfill certain requirements before the next processing stage, time can be saved and the quality of the results improved. The second stage also might contain constraints to further refine the output. Constraints help to focus the search or mining process and attenuate the computational time. This has been empirically proven to improve cluster purity. (Wagstaff & Cardie, 2000)(Hipp & Guntzer, 2002)

The theory behind these results is that the constraints help guide the clustering, showing where to connect, and which ones to avoid. The application of user-provided knowledge, in the form of constraints, reduces the hypothesis space and can reduce the processing time and improve the learning quality.

BACKGROUND

Data mining has been defined as the process of using historical data to discover regular patterns in order to improve future decisions. (Mitchell, 1999) The goal is to extract usable knowledge from data. (Pazzani, 1997) It is sometimes called knowledge discovery from databases (KDD), machine learning, or advanced data analysis. (Mitchell, 1999)

Due to improvements in technology, the amount of data collected has grown substantially. The quantities are so large that proper mining of a database can be extremely time consuming, if not impossible, or it can generate poor quality answers or muddy or meaningless patterns. Without some guidance, it is similar to the example of a monkey on a typewriter: Every now and then, a real word is created, but the vast majority of the results is totally worthless. Some things just happen at the same time, yet there exists no theory to correlate the two, as in the proverbial case of skirt length and stock prices.

Some of the methods of deriving knowledge from a set of examples are: association rules, decision trees, inductive logic programming, ratio rules, and clustering, as well as the standard statistical procedures. Some also use neural networks for pattern recognition or genetic algorithms (evolutionary computing). Semi-supervised learning, a similar field, combines supervised learning with self-organizing or unsupervised training to gain knowledge (Zhu, 2006) (Chappelle et al., 2006). The similarity is that both constrained data mining and semi-supervised learning utilize the a-priori knowledge to help the overall learning process.

Unsupervised and unrestricted mining can present problems. Most clustering, rule generation, and decision tree methods have order O much greater than N, so as the amount of data increases, the time required to generate clusters increases at an even faster rate. Additionally, the size of the clusters could increase, making it harder to find valuable patterns. Without constraints, the clustering might generate rules or patterns that have no significance or correlation to the problem at hand. As the number of attributes grows, the complexity and the number of patterns, rules, or clusters grows exponentially, becoming unmanageable and overly complex. (Perng et al,2002)

A constraint is a restriction; a limitation. By adding constraints, one guides the search and limits the results by applying boundaries of acceptability. This is done when retrieving the data to search (i.e. using SQL) and/or during the data mining process. The former reduces the amount of data that will be organized and processed in the mining by removing extraneous and unacceptable regions. The latter is what directly focuses the process to the desired results.

MAIN FOCUS

Constrained Data Mining Applications

Constrained data mining has been said to be the "best division of labor," where the computer does the number crunching and the human provides the focus of attention and direction of the search by providing search constraints. (Han et al, 1999) Constraints do two things: 1) They limit where the algorithm can look; and 2) they give hints about where to look. (Davidson & Ravi, 2005) As a constraint is a guide to direct the search, combining knowledge with inductive logic programming is a type of constraint, and that knowledge directs the search and limits the results. This combination is extremely effective. (Muggleton, 1999)

If every possible pattern is selected and the constraints tested afterwards, then the search space becomes large and the time required to perform this becomes excessive. (Boulicaut & Jeudy, 2005) The constraints must be in place during the search. They can be as simple as thresholds on rule quality measure support or confidence, or more complicated logic to formulate various conditions. (Hipp & Guntzer, 2002)

In mining with a structured query language (SQL), the constraint can be a predicate for association rules. (Ng et al, 1998) In this case, the rule has a constraint limiting which records to select. This can either be the total job or produce data for a next stage of refinement. For example, in a large database of bank transactions, one could specify only records of ACH transactions that occurred during the first half of this year. This reduces the search space for the next process.

A typical search would be:

```
select * where year = 2006 and where month < 7
```

It might be necessary that two certain types always cluster together (must-link), or the opposite, that they may never be in the same cluster (cannot-link). (Ravi & Davidson, 2005) In clustering (except fuzzy clustering), elements either are or are not in the same cluster. (Boulicaut & Jeudy, 2005) Application of this to the above example could further require that the transactions must have occurred on the first business day of the week (must-link), even further attenuating the dataset. It could be even further restricted by adding a cannot-link rule such as not including a national holiday. In the U.S.A., this rule would reduce the search space by a little over 10 percent. The rule would be similar to:

```
select * where day = monday and day <8 and where
day \=holiday
```

If mining with a decision tree, pruning is an effective way of applying constraints. This has the effect of pruning the clustering dendogram (clustering tree). If none of the elements on the branch meet the constraints, then the entire branch can be pruned. (Boulicaut & Jeudy, 2005) In Ravi and Davidson's study of image location for a robot, the savings from pruning were between 50 percent and 80 percent. There was also a typical improvement of a 15 percent reduction in distortion in the clusters, and the class label purity improved. Applying this to the banking example, any branch that had a Monday national holiday could be deleted. This would save about five weeks a year, or about 10 percent.

The Constraints

Types of constraints:

1. **Knowledge-based:** what type of relationships are desired, association between records, classification, prediction, or unusual repetitive patterns
2. **Data-based:** range of values, dates or times, relative values
3. **Rules:** time order, relationships, acceptable patterns