Constrained Data Mining

Brad Morantz
Imagery Technology & Systems Division
Science Applications International Corporation
101 N. Wilmot Rd.
Tucson, AZ 85711
U.S.A.
email: bradscientist@gmail.com

Constrained Data Mining

Brad Morantz, Science Applications International Corporation, Tucson, Ariz., U.S.A.

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
4. Statistical – at what levels are the results interesting (Han et al, 1999)
5. Linking – must-link and cannot-link (Davidson & Ravi, 2005)

Examples of constraints:

1. Borders
2. Relationships
3. Syntax
4. Chronology
5. Magnitude
6. Phase
7. Frequency

Sources of constraints:

1. A priori knowledge
2. Goals
3. Previous mining of the data set
4. The analyst
5. Insight into the data, problem, or situation
6. Time
7. User needs
8. Customer information or instruction
9. Domain knowledge

Attribute relationships and dependencies can provide needed constraints. Expert and/or domain knowledge, along with user preferences, can provide additional rules. (Perng et al, 2002) Common or domain knowledge as well as item taxonomies can add insight and direct which attributes should be included or excluded, how they should be used in the grouping, and the relationships between the two. (Perng et al, 2002)

When observations occur over time, some data occurs at a time when it could not be clustered with another item. For example, if A always occurs before B, then one should only look at records where this is the case. Furthermore, there can also be a maximum time between the two observations; if too much time occurs between A and B, then they are not related. This is a mathematical relationship, setting a window of allowable time. The max-gap strategy sets a limit for the maximum time that can be between two elements. (Boulicaut & Jeudy, 2005) An example is:

IF ((time(A) < time(B)) AND (time(B) – time(A) < limit))

Domain knowledge can also cause groupings. Some examples are: knowing that only one gender uses a certain item, most cars have four tires, or that something is only sold in pairs. These domain knowledge rules help to further divide the search space. They produce limitations on the data.

Another source of constraints is the threshold of how many times or what percentage the discovered pattern has occurred in the data set. (Wojciechowski & Zakrewicz, 2002) Confidence and support can create statistical rules that reduce the search space and apply probabilities to found rules. Using ratio rules is a further refinement to this approach, as it predicts what will be found based upon past experience. (Korn et al 2000)

Using the Constraints

User interactive constraints can be implemented by using a data mining query language generating a class of conditions that will extract a subset of the original database. (Goethals & van der Bussche, 2003) Some miners push the constraints into the mining process to attenuate the number of candidates at each level of the search. (Perng et al, 2002) This might suffice or further processing might be done to refine the data down to the desired results. Our research has shown that if properly done, constraints in the initial phase not only reduce computation time, but also produce clearer and more concise rules, with clusters that are statistically well defined. Setting the constraints also reduces computation time in clustering by reducing the number of distance calculations required. Using constraints in clustering increases accuracy while it reduces the number of iterations, which translates to faster processing times. (Davidson & Ravi, 2005)

Some researchers claim that sometimes it is more effective to apply the constraints after the querying process, rather than within the initial mining phase. (Goethals & van der Bussche, 2003) Using constraints too early in the process can have shortcomings or loss of information, and therefore, some recommend applying constraints later. (Hipp & Guntzer, 2002) The theory is that something might be excluded that should not have been.

Ratio rules (Korn et al, 2000) allow sorting the data into ranges of ratios, learning ratio patterns, or finding anomalies and outliers. Ratio rules allow a better understanding of the relationships within the data. These ratios are naturally occurring, in everyday life. A person wears, one skirt, one blouse, and zero or one hat. The numbers do not vary from that. In many things in life, ratios exist, and employing them in data mining will improve the results.

Ratio rules offer the analyst many opportunities to craft additional constraints that will improve the mining process. These rules can be used to look at anomalies and see why they are different, to eliminate the anomalies and only consider average or expected performances, split the database into ratio ranges, or only look within certain specified ranges. Generating ratio rules requires a variance-covariance matrix, which can be time- and resource-consuming.

In a time-ordered database, it is possible to relate the changes and put them into chronological order. The changes then become the first derivative over time and can present new information. This new database is typically smaller than the original one. Mining the changes database can reveal even more information.

In looking for relationships, one can look in both directions, both to see A occur and then look for B, or after finding B, look for the A that could have been the cause. After finding an A, the search is now only for the mating B, and nothing else, reducing even more the search space and processing time. A maximum time interval makes this a more powerful search criteria. This further reduces the search space.

User desires, a guide from the user, person, or group commissioning the data mining, is an important source. Why spend time and energy getting something that is not wanted. This kind of constraint is often overlooked but is very important. The project description will indicate which items

or relationships are not of interest, are not novel, or are not of concern. The user, with this domain knowledge, can direct the search.

Domain knowledge, with possible linking to other databases, will allow formulating a maximum or minimum value of interrelation that can serve as a limit and indicate a possibility. For example, if a person takes the bus home, and we have in our database the bus schedule, then we have some expected times and limits for the arrival time at home. Furthermore, if we knew how far the person lived from the bus stop, and how fast the person walked, we could calculate a window of acceptable times.

Calculations can be done on the data to generate the level of constraint. Some of this can be done in the query, and more complex mathematics can be implemented in post processing. Our research has used IDL for exploratory work and then implemented in Fortran 95 for automated processing. Fortran 95, with its matrix manipulations and built-in array functions, allowed fast programming. The question is not which constraints to use, but rather what is available. In applying what knowledge is in hand, the process is more focused and produces more useful information.

FUTURE TRENDS

An automated way to implement the constraints simply into the system without having to hard code them into the actual algorithm (as the author did in his research) would simplify the process and facilitate acceptance into the data mining community. Maybe a table where rules can be entered, or possibly a specialized programming language, would facilitate this process. The closer that it can approach natural language, the more likely it is to be accepted. The entire process is in its infancy and, over time, will increase performance both in speed and accuracy. One of the most important aspects will be the ability to uncover the hidden nuggets of information or relatioships that are both important and unexpected.

With the advent of multi-core processors, the algorithms for searching and clustering the data will change in a way to fully (or at least somewhat) utilize this increased computing power. More calculations and comparisons will be done on the fly as the data is being processed. This increased computing power will not only process faster, but will allow for more intelligent and computationally intensive constraints and directed searches.

CONCLUSION

Large attenuation in processing times has been proven by applying constraints that have been derived from user or domain knowledge, attribute relationships and dependencies, customer goals, and insights into given problems or situations. (Perng et al, 2002). In applying these rules, the quality of the results improves, and the results are quite often more meaningful.

REFERENCES

Boulicaut, J-P., & Jeudy, B., (2005). Constraint-Based Data Mining, *The Data Mining and Knowledge Discovery Handbook 2005*, Maimon, O. & Rokach, L. Eds Springer-Verlag 399-416

Chappelle,O., Scholkopf, B., & Zein, A. (2006), *Semi Supervised Learning*, MIT Press, Cambridge MA

Davidson, I., & Ravi, S. (2005). Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results, *Proceedings of PKDD 2005 9th European Conference on Principles and Practice of Knowledge Discovery in Database,* Jorge, A., P., & Gama, J, Eds. Springer, 59-70

Davidson, I., & Ravi, S. (2005). Clustering Under Constraints: Feasibility Issues and the k-Means Algorithm, *Papers Presented in the 2005 SIAM International Data Mining Conference*

Goethals, B., & van der Bussche, J. (2000). On Supporting Interactive Constrained Association Rule Mining, *Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery*; Vol. 1874, pages 307–316, Springer-Verlag;

Han, J., Lakshmanan, L., & Ng, R. (1999). Constraint-Based Multidimensional Data Mining, *IEEE Computer*, 32(8), August 1999

Hipp, J., & Guntzer, U. (). Is Pushing Constraints Deeply Into the Mining Algorithm Really What We Want? An Alternative Approach to Association Rule Mining, *SIGKDD*, 4(1), 50-55

Korn, F., Kotidis, Y., Faloutsos, C., & Labrinidis, A., (2000). Quantifiable Data Mining Using Ratio Rules, *The International Journal of Very Large Data Bases*, 8(3-4), 254-266

Mitchell, T. (1999). Machine Learning and Data Mining, *Communications of the ACM*, 42(11), November 1999, 30-36

Muggleton, S. (1999) Scientific Knowledge Discovery Using Inductive Logic Programming, *Communications of the ACM*, 42(11), November 1999, 42-46

Ng, R., Laks, V., Lakshmanan, S., Han, J., & Pang, A. (1998) Exploratory Mining and Pruning Optimizations of Constrained Association Rules, *Proceedings of ACM SIGMOD International Conference on Management of Data,* Haas, L, & Tiwary, A. Eds., ACM Press, 13-24

Pazzani, M., Mani, S.,& Shankle, W. (1997). Comprehensible Knowledge-Discovery in Databases, *Proceedings from the 1997 Cognitive Science Society Conference, 596-601.* Lawrence Erlbaum

Perng, C., Wang, H., Ma, S., & Hellerstein, J. (2002). Discovery in Multi-Attribute Data with User-Defined Constraints, *SIGKDD Explorations,* 4(1), 56-64

Srikant, R. & Vu, Q. (1997) Mining Association Rules with Item Constraints, *Proceedings of the 3rd International Conference of Knowledge Discovery And Data Mining,* Heckerman, D., Mannila, H., Pregibon, D., & Uthurusamy, R. Eds, AAAI Press, 67-73

Wagstaff, K., & Cardie, C. (2000). Clustering with Instance Level Constraints, *ICML*, in Davidson & Ravi

Wojciechowski, M., & Zakrzewicz, M, (2002). Dataset Filtering Techniques in Constraint-Based Frequent Pattern Mining. *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery, ACM Lecture Notes in Computer Science* 2447, 77-91

Zhu, X. (2006) *Semi-supervised Learning Literature Survey*, doctoral thesis, http://www.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html

KEY TERMS AND THEIR DEFINITIONS

Constraint: a restriction, a forced guideline, a specific confining rule; in data mining, it is an exacting, specific rule defining specific intents and areas.

Cluster: a collection of like items, similar along specified dimensions. Cluster analysis is a process where items or data points are sorted by their similarities and separated by their differences.

Tree building: a mathematical process if generating a decision tree that is data-driven and can be used for classification.

Anomaly: an irregular value, or one that is beyond the limits of acceptability; in statistics, a value further from the mean than a specified distance, typically 2 or 3 sigma.

Fortran 95: a mathematical programming language, one of the first computer languages ever developed; the latest versions are optimized for matrix operations and multiprocessing.

Association rule: a data mining term for a rule that describes items in a basket that are dependent on other items also in the basket.

Ratio rule: a data mining term for a quantity relationship between attributes.