In the proceeding of the Seventh International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems IEA/AIE-94

AN EXPERIMENTAL STUDY OF DISCOVERY IN LARGE TEMPORAL DATABASES

Ibrahim F. Imam
Center for Artificial Intelligence
George Mason University
Fairfax, VA. 22030
iimam@aic.gmu.edu

ABSTRACT

Time is a common factor in most of the large databases. The time factor in such data can be utilized during the knowledge discovery process to overcome some limitations, such as the size of the data, of many learning systems. The paper presents an experiment for discovering knowledge in large temporal databases. The method divides the learning space into subspaces each corresponds to one time interval. The process of discovery is performed on the data available in each subspace, for a given set of decision classes, using the learning system AQ15. The method determines a set of attributes relevant to the discovery process using the important score (IS) system. The experiment is done on international TRADE data; the data is concerned with the imports and exports between the US and the world. The preliminary results show that strong relationships in the original data can be discovered over different subsets of the data.

Key words: machine learning, knowledge discovery in databases, temporal database.

1. INTRODUCTION

Most of the real world databases are continuously accumulated over the time. That increases the need for strong techniques to discover different kinds of knowledge from the data. Many learning systems can not perform well with very large databases due to different limitations. To overcome this problem, some discovery systems refine the database before the discovery process begins. The refinement operations reduce complexity, size, or noise in the data. Examples of such techniques include domain quantization, data reduction, partial classification, noise pruning, data optimization, etc. (e.g. [CS93]; [FI92]; [KMK91]; [Qui90]; [SS92]; [SW84]; [UFS91]; [Zia91]). An attractive approach to simplifying the discovery process involves splitting the learning space into a set of subspaces. This can be done by taking one attribute out and generating as many subspaces, using the remaining attributes values, as the legal number of (quantized) values of that attribute. Each subspace has a distinct subset of the data and has a relation to the other subspaces. The discovery is done on each dataset then the obtained knowledge is accumulated and analyzed. An advantage of this approach is that it simplifies the discovery process and provides accurate knowledge. Also, it allows additional means of discovery by considering the change in the knowledge over different subspaces, for example, discovering interesting patterns, relevant attributes, domains relationships, and their changes over the different subspaces.

This paper presents an experiment on discovering different types of knowledge in the TRADE database. The Trade data is concerned with the import and export of goods between the US and the world. The experiment uses the time factor to partition the space of the discovery. It uses the learning system AQ15 [MMH*86] and the INLEN system [MKK*92a,b] to discover common pattern in the data. The Importance Score program (IS) (IMK93) is used to obtain the relevant attributes and to drive other types of information.

A case study is done to provide relationships between the different decision classes of the monthly change in the trade value, and the sets of legal values of the two most relevant

attributes (the market share change and the change in the percentage of the share of a product to its parent class). The results show a very high correlation, which was approved by an expert. The importance score is used again to observe relations between the attributes over the first 12 months of the study. This knowledge is examined against data of the last 5 months.

2. THE DISCOVERY TOOLS

Machine learning systems play a great role in discovering interesting knowledge in databases due to their capabilities in performing different types of inferences (e.g. [BGS*91]; [GMT*91]; [MKK*92a,b]; [Sch92]). The experiment presented uses the symbolic learning system AQ15 and the feature selection program IS to perform the discovery. The rest of this section presents a brief description on those systems.

2.1. The AQ15 rule learning program

AQ15 learns decision rules for a given set of decision classes from examples of decisions, using the STAR methodology [Mic83]. The simplest algorithm based on this methodology, called AQ, starts with a "seed" example of a given decision class, and generates a set of the most general conjunctive descriptions of the seed that do not cover examples of other decision classes (alternative decision rules for the seed example). Such a set is called the "star" of the seed example. The algorithm selects from the star a description that optimizes a criterion reflecting the needs of the problem domain. If the criterion is not defined, the program uses a default criterion that selects the description that covers the largest number of positive examples (to minimize the total number of rules needed) and, with the second priority, involves the smallest number of attributes (to minimize the number of attributes needed for arriving at a decision).

If the selected description does not cover all examples of a given decision class, a new seed is selected from uncovered examples, and the process continues until a complete class description is generated. The algorithm can work with few examples or with many examples, and can optimize the description to a variety of easily-modifiable hypothesis quality criteria.

The learned descriptions are represented in the form of a set of decision rules, expressed in an attributional logic calculus, called *variable-valued logic 1* or *VL1* [Mic73]. A distinctive feature of this representation is that it employs, in addition to standard logic operators, the internal disjunction operator (a disjunction of values of the same attribute in a condition) and the "range" operator (to express conditions involving a range of discrete or continuous values). These operators help to simplify rules involving multivalued discrete attributes; the second operator is also used for creating logical expressions involving continuous attributes.

AQ15 can generate decision rules that represent either *characteristic* or *discriminant* concept descriptions, depending on the settings of its parameters [Mic83]. A characteristic description states properties that are true for all objects in the concept. The simplest characteristic concept description is in the form of a single conjunctive rule (in general, it can be a set of such rules). The most desirable is the maximal characteristic description, that is a rule with the *longest* condition part, i.e., stating as many common properties of objects of the given class as can be determined. A discriminant description states properties that discriminate a given concept from a fixed set of other concepts. The most desirable is the minimal discriminant descriptions, that is a rule with the *shortest* condition part. For example, to distinguish a given set of tables from a set of chairs, one may only need to indicate that tables "have a large flat top." A characteristic description of the tables would include also properties such as "have four legs, have no back, have four corners, etc." Discriminant descriptions are much shorter than characteristic descriptions.

Another option provided in AQ15 controls the relationship among the generated descriptions ("rulesets" or "covers") of different decision classes. In the "IC" (Intersecting Covers) mode, rulesets of different classes may logically intersect over areas of the description space in which there are no training examples. In the "DC" (Disjoint Covers) mode, descriptions of different classes are logically disjoint. The DC mode descriptions are usually more complex, both in the number of rules and the number of conditions. There is also a "DL" mode (a Decision List mode, also called "VL" mode--for variable-valued logic mode), in which the program generates rulesets that are linearly ordered. To assign a decision to an example using such rulesets, the program evaluates them

in order. If ruleset *i* is satisfied by the example, then the decision is made, otherwise, the program proceeds to the evaluation of the ruleset *i+1*. In IC and DC modes, rulesets can be evaluated in any order.

2.2. Measuring the Attribute's Importance

The importance of an attribute depends on how this attribute's values can classify different decision classes. To determine such a measurements, the method uses the Importance Score (IS) method [IV94]. The method was developed as a feature selection tool that searches for the minimum set of attributes, the most important ones, which produce a better predictive accuracy. The method assigns a score to each attribute based on its dependence with the description of the decision classes (which are learned by any of the AQ systems in this case). The attributes are ordered descendingly by their scores and a greedy-like search is performed to select the best attribute set. A dynamic threshold is used to start the search with some set of attributes, then to add a new attribute to the selected set. The method uses an arbitrary (user-defined) tolerance to determine when to end the search.

Given n decision classes $C_1,..,C_n$, and m attributes $A_1,..,A_m$. Assume also for each attribute A_j there are Ec_{ij} examples that match rules containing A_j, and belong to class C_i (i=1,..,n; j=1,..,m), the importance score for A_j is calculated as follows:

$$IS(A_j) = \left(\sum_{i=1}^{n} Ec_{ij} \right) \qquad (1)$$

The search for the optimal attribute set is directed by a dynamic threshold which takes initial value between zero and one. The initial threshold is chosen such that the number of attributes with higher importance scores is at least equal to the square root of the total number of attribute (e.g. for 7 attributes, the first threshold should be less than the highest three importance scores). Any attribute has importance score greater than that threshold is considered relevant, and the data is modified to include only the relevant attributes. AQ15 is ran against the modified data and the output rules are tested against the testing examples. Then, the method reduces the threshold to include another attribute. In the case of having more than one attribute with the same score, the attribute to be selected is the one which exists in rules with large number of complexes. The change in the predictive accuracy of the learned rules is reported after each iteration. The search ends whenever this change is greater than a given tolerance. For simplicity, the tolerance is set to 1%, and for each subset of data, the number of iterations was very small (less than 4; generally with high tolerance it always less than half the number of attributes). The program is modified to prune rules that are learned from few number of examples (less than 5% of the data).

Example: consider that we have three decision classes C1, C2, and C3, four attributes x1, x2, x3, and x4, and the following sets of rules:

```
{C1  <=  [x1=2] & [x2=2]
         (t : 34, u : 29)
 C1  <=  [x1=3] & [x3=1 v 3] & [x4=1]
         (t : 4, u : 4) }
{C2  <=  [x1=1 v 2] & [x2=3 v 4]
         (t : 21, u : 19)
 C2  <=  [x1=3] & [x3=1 v 2] & [x4=2]
         (t : 11, u : 9) }
{C3  <=  [x1=1] & [x2=1]
         (t : 26, u : 26)
 C3  <=  [x1=4] & [x3=2 v 3] & [x4=3]
         (t : 6, u : 2) }
```

The importance score of the given attributes are: IS(x1)=102; IS(x2)=81; IS(x3)=21; IS(x4)=21.

3. DESCRIPTION OF THE METHOD

The method temporally divides the data into subsets where each subset contains all the records that belong to one and only one time interval. The time interval here is chosen to be a one month period. Attributes with real domains are mapped into new attributes with equivalent but quantized domains. The new domains consist of seven to ten symbolic values, each corresponding to one interval of the original domain. All of the data subsets are used in parallel by AQ15 for learning sets of rules or patterns in the data. The rules are used by the IS program to either prune the weak patterns, or to provide relevant attributes for rediscovery or for visualizing the importance of different attributes over the time. Figure 1 shows a general description of the method.

The experiment included some data refinement operations to adapt the data for the discovery.

These operations are used in the following phases: 1) temporal classification; 2) domain quantization; 3) example formulation; 4) pruning irrelevant attributes.

The final phase starts after learning concept descriptions of the decision classes and discovering which attributes are irrelevant by the IS program. The set of irrelevant attributes are

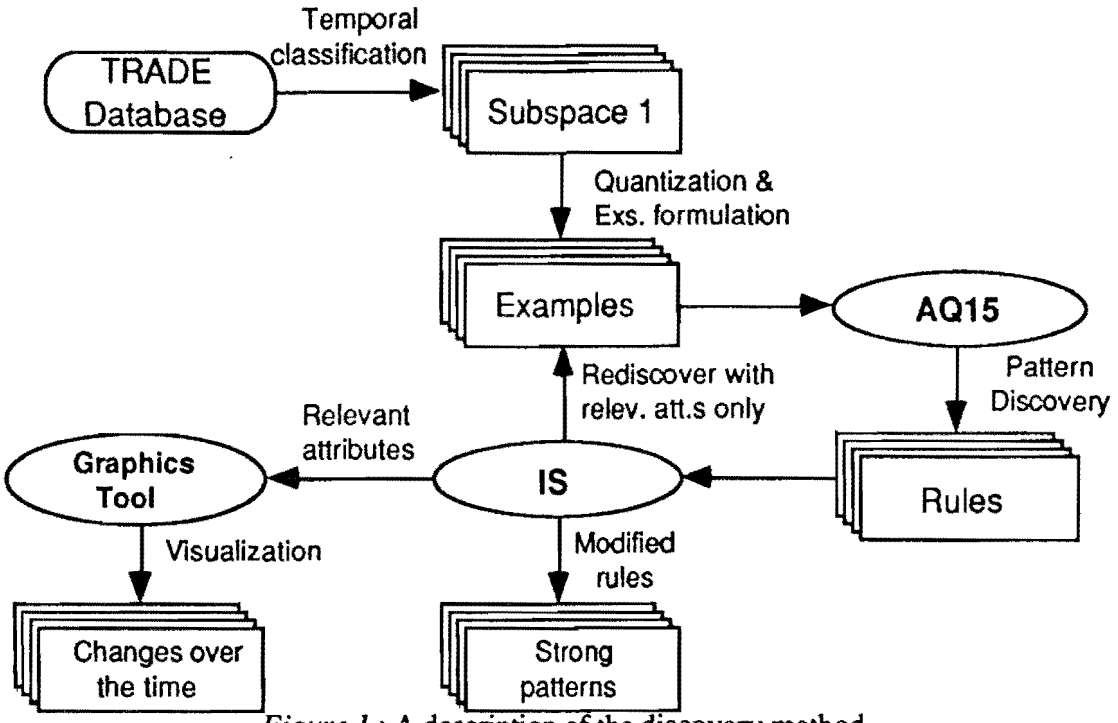


Figure 1 : A description of the discovery method.

In the temporal classification phase, the data is divided into subsets where each subset represents data from one month. As a result of this process, the year and month attributes are removed from the data.

The data quantization phase is concerned with mapping uniformly the domain of any attribute, with real values, into a set of continuous intervals. The number of intervals is chosen to be between seven and ten. The union of these intervals is equivalent to the whole attribute domain. An experiment was done to test the quantization with a uniform distribution on the number of data records in each interval. The method did not fit with more than one month. This phase also includes mapping the domains of discrete attributes into a one-to-one symbolic domains.

The third phase of the data refinement process maps the data from each month into a set of decision classes and training examples. This phase also includes the generation of an input file that follows the syntax of AQ family input files.

removed from the original data, and the learning process is repeated only with the relevant attributes.

4. THE INTERNATIONAL TRADE DATA

The international TRADE database is concerned with the different import and export values between the United States and the world. The experiment is done on data of 17 month period, the 12 months of 1989 and the first five of 1990. An analysis and some experiments were done on small and random portion of this data [KMK91]. The data is sorted by destination, by product, and by date. The products are organized into a hierarchy of about 150 groups. A product may be defined as crop, device, animal, etc. The data contains a total number of 12580 records. Each record has 39 attributes values. The data from the first 12 months (1989) are used for knowledge discovery, while the data from the last five months (1990's first five months) are used for testing and comparison.

The TRADE data has the following attributes
1) Country—three groups (Japan, China, and the rest of the world);
2) Dir—indicates the state of the transaction, ("I" for import and "X" for export);
3&4) Year and Month—the year and the month where the transaction took place;
5) Commodity—the transaction product code;
6) Trade value—trade value in millions of US dollars;
7) Change for period—current period compared to previous one in percent;
8) Change for Year ago—current period compared to same period, previous year, in percent;
9) Change vs. period average—ratio of current period vs average period change for previous 60 months in percent;
10) Change vs. yearly average—ratio of current year vs average yearly change for previous 60 months in percent;
11) Market share—percent. of this commodity traded to or from the partner in question;
12) Market share change—this period's market share compared to previous period, change in percent;
13) Market share forecast—forecast value for current period's market share;
14) Market share low confidence bound—95% confidence range of market-share forecast;
15) Market share high confidence bound—95% confidence range of market-share forecast;
16) Market share forecast actual ratio—forecast value divided by actual value;
17) Parent share—percentage of commodity to class traded in this commodity;
18) Parent share change—this period's parent-share vs. previous period's parent share, in percent;
19) Parent share forecast—forecast value for current period's parent share;
20) Parent share low confidence bound—95% confidence range of parent-share forecast;
21) Parent share high confidence bound—95% confidence range of parent-share forecast;
22) Parent share forecast actual ratio—forecast value divided by actual value;
23) Trade value change—difference between this periods and last period's trade value;
24) Box_chng—Measure of forecast accuracy by box jenkins;
25) Box_fcst—the predicted trade value using box jenkins program;
26) Box_jlc_bnd—95% confidence range of box-jenkins forecast;
27) Box_jhc_bnd—95% confidence range of box-jenkins forecast;
28) Tv_pct—commodity percentage of all trade with the country;
29) Ytd_cum—sum of trade_values from January to current;
30) Ytd_ann—extrapolate whole year from ytd_cumulative;
31) Prv_ytd_cum—ytd cumulative for same period, previous year;
32) Prv_ytd_ann—extrapolation from prev-year;
33) Ytd_v_pytd—difference in percent between ytd and last year's;
34) High_val—maximum actual trade value prior to current period;
35) High_val_y—year of previous high_value;
36) High_val_p—period of previous high_value;
37) Low_val—minimum actual trade value prior to current period;
38) Low_val_y—year of previous low_value;
39) Low_val_p—period of previous low_value.

5. THE EXPERIMENT

In this experiment, the attribute "change for period" (x7) represents the monthly change in the trade value. This attribute measures the ratio comparing the trade value from the past period (month) to the current one. The goal of the discovery is to learn information that characterizes the different classes of that change. The decision classes are given as follows: (C1) Not-applicable, which contains any example where the value of x7 is undefined; (C2) Major-loss, where, for a given case, the value of x7 is less than -30 (in other words, the trade value from the current month is more than 30% less than the previous month); (C3) Slight-change, where the x7 value is between -30 and 30 (the trade value of the current month is close to the value of the previous month); (C4) Moderate-increase, where the value is between 30 and 100; otherwise the case belongs to (C5) Exceptional-increase.
The following is an example of strong rules in march 1989 of the class exceptional increase:

[cmdty = 11 v 14X v 23X v 42 v 42X v 423X v 433X v 436X v 44X v 441X v 451X v 52] & [ms_chng = F v G]

The rule shows that for a certain set of commdities, and with high percentage of the

market share change (75,100], the trade value is exceptionally increased.

The IS program is used to determine the relevant attributes to the monthly change in the trade value. Table 1 shows the attributes *most relevant* (relevant in more than one month).

Another strong and simple rule from the same month is learned for the moderate increase class:

[cmdty = 14 v 15 v 15X v 17 v 18X v 22X v 234X v 33 v 421X v 424X v 43X v 431X v 432X v 435X v 437X v 452X v 454X v 51] & [trd_val = B v C v D v G] &

[yr_chng <> A v H v I] & [ms_chng = F] & [mktsh_hcf_bnd = A v B v C v J] &

[high_val = B v C v E v G]

The interpretation of this rule is:

The change of the trade value is moderatly increased from the past month if 1) the comdity is grains, fruit and vegetables, coffe and coca, other foods, vegetable material, unwrought metals, textiles, machinery, engines, hard tools, business machines, scientific instruments, radio receivers, other consumer goods, diamonds; 2) the trade value within any of the ranges (0,20] or (45, 150] million dollars; 3) the change for year ago within the range (-50,12000]; 4) the percentage of the market share change to the privious period is in the range (75,99]; 5) the market share high confidence bound is less than or equal to 25 or unknown; 6) the maximum trade value prior to the current period is in the range of (0,60], (200,1000], or (4000,10000].

There is one attribute, number 6, relevant to the monthly change in 11 months. There are 5 attributes relevant in 4 to 6 months. Finally, there are 4 attributes relevant in two or three months. Three attributes (commodity, market share change, and parent share change) play a very important role in describing the learned patterns over the whole data set. The rest of this section introduces a brief description of the relationships between the two most important attributes and the decision classes. These relationships are obtained from the simple and strong patterns by counting the number of times each attribute's value appear in the rules. A pattern is strong if it is learned from large number of examples. A pattern is simple if it can be described by very few attributes.

The class "Not applicable" covers cases in which measuring the change of the trade value in the current month compared to the previous one is not possible (usually when there was no exchange of that product during the previous month, creating a divide by zero situation). It has a simple relationship with the market share change, and the parent share change. This relation indicates that the value is undefined whenever the value of one or both of these two attributes are undefined.

Since the attributes commodity, trade value, change over the past year, market share change, and parent share change are the most important attributes in describing any pattern. So it was very necessary to study the relation between some of them individually and the decision

Attribute	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Commodity	x	x	x	x	x	x	x	x	x	x	x	x
Trade value	x	x	x	x		x	x	x	x	x	x	x
Change from a year ago		x	x			x	x					
Change vs. period average	x	x	x	x	x	x	x	x	x	x	x	x
Change vs. yearly average	x				x		x		x			
Market share						x	x			x		
Market share change	x	x	x	x	x	x	x	x	x	x	x	x
Parent share change	x	x	x	x	x	x	x	x	x	x	x	x
Parent share low conf. bound					x		x					
Parent share high conf. bound			x			x				x		
Box_chng	x				x		x			x	x	
Ytd_ann	x									x		
High_val		x	x			x	x		x	x		
High_val_y		x		x	x			x				

Table 1: The relevant attributes to the monthly change in the trade value.

There are four attributes that are relevant throughout the year, attributes 5, 9, 12, and 18.

classes, while the learned rules represent the relation between sets of these attributes and the decision classes. The analysis is done on the two most important attributes, market share change and parent share change.

The market share change and the parent share change have similar symbolic domain which is mapped by one function. The domain of the change is mapped into 10 different intervals. Each interval is mapped into one symbol. These symbols are: A - for any value less than or equal to (-50); B = (-50, -20]; C = (-20, 0); D = 0; E = (0, 50]; F = (50, 200]; G = (200, 1500]; H = (1500, 5000]; I - for values greater than 5000, and J = N/A. Figure 2 shows the relation between the domain of the decision classes and the domain of both attributes. These relationships were manually obtained from a random sample of four months by calculating the frequency of different values in the rules.

Other relationships are discovered by monitoring the change of the importance score for each attribute over the different periods of time and comparing it with other attributes. If attribute A has the highest importance score, then the values of this attribute play the most important role in classifying any example into one of the given classes. For example, the values of commodity (5), market share change (12) and parent share change (18) play important roles in classifying any new example as to whether it indicates loss or increase in the trade value.

Figure 3 shows a relation between the change in the market share change and the parent share change. Note that, only one attribute takes score 1 each month. In Figure 3, the score of each attribute is divided over the maximum number of covered examples. In January the market share change is more important in describing the discovered knowledge than the parent share

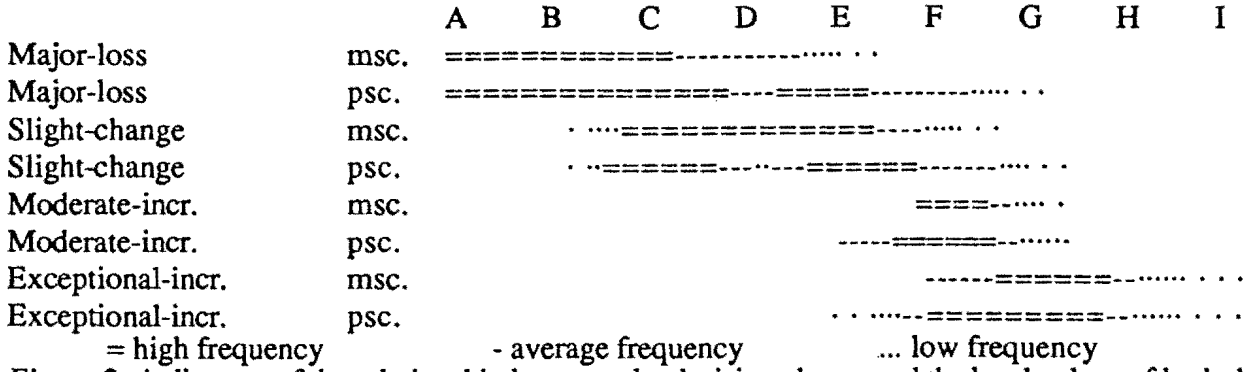
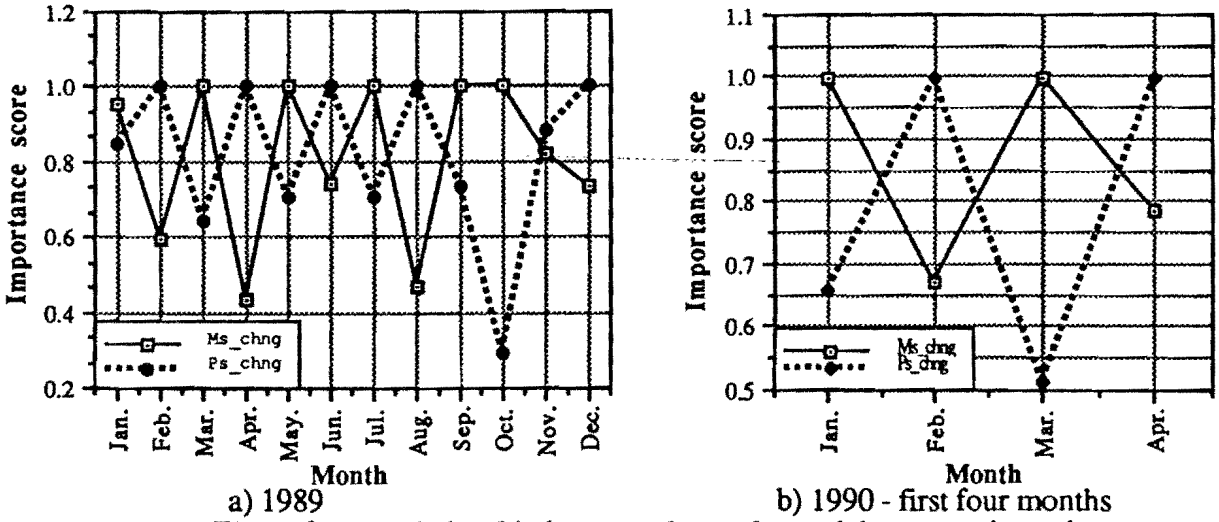


Figure 2: A diagram of the relationship between the decision classes and the legal values of both the market share change (msc) and parent share change (psc).

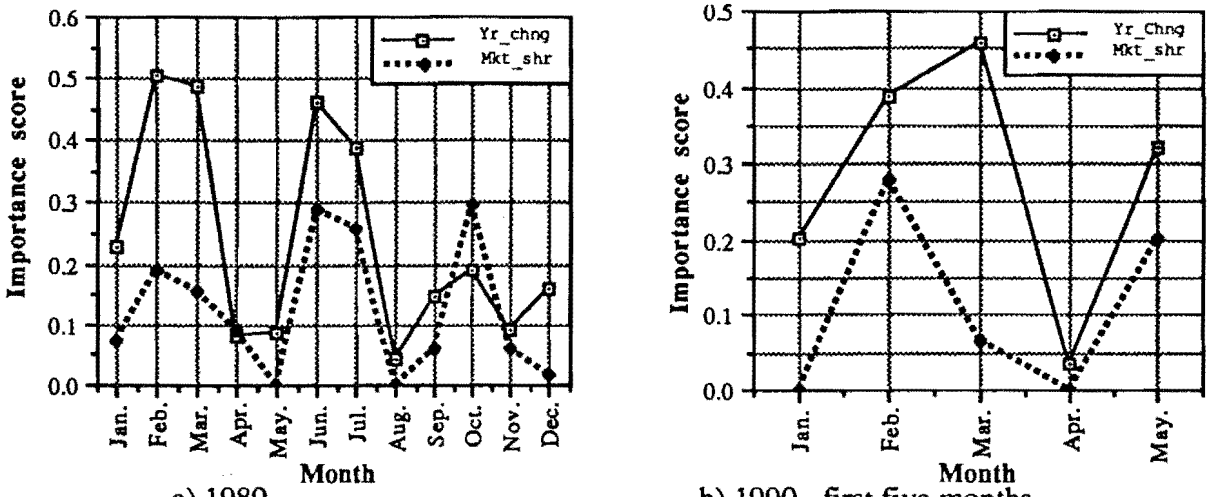


a) 1989 b) 1990 - first four months

Figure 3: The relationship between the market and the parent share change

change, but they are in a series of switching back and forth until September. In October and December, while one is decreasing the other is not decreasing.
Such information can help in anticipating changes in the trade value. For example, to achieve an exceptional increase, the values of the market share change and the parent share change are expected to be high. Moreover, the commodities are likely to be within a set of products such as crude metals, unwrought metals and others.
Figure 4 shows another temporal relationship between two other attributes the change from a year ago and the market share value. The two attributes *generally* behave in the same direction (i.e. they increase and decrease together).

score program (IS) [IMK93]. Any attribute relevant in more than one month is considered relevant to the whole data set. Other interesting relationships are discovered between some of the relevant attributes (individually) and the decision classes. These relationships are represented by a simple diagrams. Also, the change of the attribute relevance to the decision classes is visualized for 12 months, and compared with the other five months.
An advantage of this method is that it discovers relationships between the attributes in many ways such as attribute dependence, temporal change, classification, and common patterns. Another advantage is that it uses the learned knowledge as a feedback about the data so that the data can be refined another time and used for



a) 1989 b) 1990 - first five months

Figure 4: The relationship between the change from a year ago and the market share value.

6. CONCLUSION

The paper presents an experiment for knowledge discovery in large temporal databases. The method temporally divides the learning space into subsets, each corresponds to one time period. The experiment is applied on an international trade database between the US and the world for a period of 17 months. Discoveries of common or interesting patterns are done in parallel using the learning system AQ15 [MMH86] and the INLEN system [MKK*92a,b]. These patterns describe five different classes of the monthly change in the trade value. The method determines a set of attributes relevant to the monthly change. The relevant attributes are obtained by the importance

relearning.
A direction for future work is to use some of the capabilities of the INLEN system so one can test the learned rules to determine which elements of the rules are more significant. Also, INLEN can use the learned rules as a knowledge base for an expert system. Also, the IS program can be improved to provide the relationships between the relevant attributes and the decision classes.

ACKNOWLEDGMENTS

The author thanks Ryszard S. Michalski for helpful suggestions and comments, and thanks also Ken Kaufman and Mark Maloof for their careful review.

This research was conducted in the Center for Artificial Intelligence at George Mason University. The Center's research is supported in part by the National Science Foundation under grant No. IRI-9020266, in part by the Defense Advanced Research Projects Agency under the grant No. N00014-91-J-1854, administered by the Office of Naval Research, and the grant No. F49620-92-J-0549, administered by the Air Force Office of Scientific Research, and in part by the Office of Naval Research under grant No. N00014-91-J-1351.

REFERENCES

[BGS*91] Bergadano, F., Giordana, A., Saitta, L., Brancadori, F., and De Marchi, D., "Integrated Learning in a Real Domain", *Knowledge Discovery In Databases*, Piatetsky-Shapiro, G., Frawley, W., (Eds.), AAAI Press, 1991.

[CS93] Chan, P. and Stolfo, S., "Toward Parallel and distributed Learning by Meta-Learning", *Proceeding of the AAAI-93 Workshop on Knowledge Discovery in Databases*, Washington D.C., July 11-12, 1993.

[FI92] Fayyad, U.M., and Irani, K.B., "On the Handling of Continous-Valued Attributes in Decision Tree Generation", *in Journal of Machine Learning*, Vol. 8, No. 1, pp. 87-102, 1992.

[GMT*91] Gonzalez, A.J., Myler, H.R., Towhidnejad, M., McKenzie, F.D., Kladke, R.R., and Laureano, R., "Automated Extraction of Knowledge from Computer-Aided Design Databases", *Knowledge Discovery In Databases*, Piatetsky-Shapiro, G., Frawley, W., (Eds.), AAAI Press, 1991.

[IMK93] Imam, I.F., Michalski, R.S., and Kerschberg, L. "Discovering Attribute Dependence in Databases by Integrating Symbolic Learning and Statistical Analysis Techniques", *Proceeding of the AAAI-93 Workshop on Knowledge Discovery in Databases*, Washington D.C., July 11-12, 1993.

[IV94] Imam, I.F., Vafaie, H., "An Empirical Comparison Between Global and Greedy-Like Search for Feature Selection", *in the proceeding of the 7th Florida AI Research Symposium*, Florida, 1994.

[KMK91] Kaufman, K.A., Michalski, R.S., and Kerschberg, L., (b) "An Architecture for Integrating Machine Learning and Discovery Programs Into a Data Analysis System", *Proceeding of the AAAI-91 workshop on Knowledge Discovery in Databases*, Anaheim, CA, July 1991.

[Mic73] Michalski, R.S., "AQVAL/1-Computer Implementation of a Variable-Valued Logic System VL1 and Examples of its Application to Pattern Recognition" in *Proceeding of the First International Joint Conference on Pattern Recognition*, Washington, DC, pp. 3-17, October 30- November, 1973.

[Mic83] Michalski, R.S., "A Theory and Methodology of Inductive Learning", *Artificial Intelligence*, Vol. 20, pp. 111-116, 1983.

[MMH*86] Michalski, R.S., Mozetic, I., Hong, J., and Lavrac, N., "The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains," *Proceedings of AAAI-86*, pp. 1041-1045, Philadelphia, PA: , 1986.

[MKK*92a] Michalski, R.S., Kerschberg, L., Kaufman, K.A., and Ribeiro, J.S., (a) "Searcing for Knowledge in Large Databases", *Proceedings of the First International Conference on Expert Systems and Development*, Cairo, Egypt, 1992.

[MKK*92b] Michalski, R.S., Kerschberg, L., Kaufman, K.A., and Ribeiro, J.S., (b) "Mining for Knowledge in Databases: The INLEN Architecture, Initial Implementation and First Results, "*Intelligent Information Systems: Integrating Artificial Intelligence and Database Technologies*, Vol. 1, No. 1, August 1992.

[Qui90] Quinlan, J. R. "Probabilistic decision trees," in Y. Kodratoff and R.S. Michalski (Eds.), *Machine Learning: An Artificial Intelligence Approach, Vol. III*, San Mateo, CA, Morgan Kaufmann Publishers, (pp. 63-111), June, 1990.

[SS92] Scheines, R., & Spirtes, P., "Finding Latent Variable Models in Large Databases", *International Journal of Intelligent Systems*, Yager, R.R. (Ed.), pp. 609-621, Vol. 7, No. 7, Sep., 1992.

[Sch93] Schlimmer, J., "Using Learned Dependencies to Automatically Construct Sufficient and Sensible Editing Views", *Proceeding of the AAAI-93 Workshop on Knowledge Discovery in Databases*, Washington D.C., July 11-12, 1993.

[SW84] Shoshani, A., & Wong, H.K.T., "Characteristics of Scientific Databases", *Proceedings of the 10th International Conference on Very Large DataBases*, Singapore, August, 1984.

[UFS91] Uthurusamy, R., Fayyad, U., & Spangler, S., "Learning Useful Rules from Inconclusive Data", *Knowledge Discovery In Databases*, Shapiro, G., Frawley, W., (Eds.), AAAI Press, 1991.

[Zia91] Ziarko, W., "The Discovery, Analysis, and Representation of Data Dependencies in Databases", *Knowledge Discovery In Databases*, Shapiro, G., Frawley, W., (Eds.), AAAI Press, 1991.