Knowledge Discovery through Data Mining: An Econometric Perspective

Dr. Somnath Pruthi

Assistant Professor in Economics, Mukand Lal National College, Yamunanagar, Haryana, India

Abstract—*In order to determine how data mining, the misnomer of Knowledge Discovery in Databases (KDD) is taken the concept of econometric techniques in this vast emerging field of computer science. This paper reviews the data mining techniques from the econometric perspective. The paper throws light on how other basic econometric techniques have become intrinsic to the subject in question.*

Keywords— Knowledge Discovery in Databases, Data Mining, Econometrics.

I. INTRODUCTION TO DATA MINING

In today's world of information the worth of collecting data that reflect your business or scientific activities to achieve competitive advantage is widely recognized. However, the bottleneck of turning this data in your success is the difficulty of extracting the knowledge about the system that you study from the collected data. Almost all fields of life whether it is finance, marketing, engineering research, medical science, population study, education or science all have some methods to collect and record the data but are handicapped to turn this data into useful information for the purpose of decision making. This scenario usually leads to making decisions based on the intuition of the analyst and thus endangering its outcome. The recent encroachment of data collection technology such as bar code readers, sensors in scientific and industrial domain have led to the generation of huge amount of data. This incredible growth of data and databases has enforced the requirement of intelligent tools and techniques that can turn such data into useful information and knowledge. The research in databases and information technology has given rise to an approach to store and analyze this precious data for further decision making which is called Data Mining. Data Mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. It is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

II. STEPS

Three steps involved are:

A. Exploration
B. Pattern identification
C. Deployment

A. Exploration

In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.

B. Pattern Identification

Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.

C. Deployment

Patterns are deployed for desired outcome.

Various econometrics techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities

determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

Clustering

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

Types of clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

Prediction/ Forecasting

Regression technique can be adapted for prediction. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Association rule

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Neural networks

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks

- Back Propagation

III. CONCLUSION

From the last few years, econometrics tools led Data Mining techniques have been used mainly in basic and applied research. Also, due to the growth of many sophisticated softwares by IT sector ranging from well established Enterprise Miner by SAS and Intelligent Miner by IBM, CLEMENTINE by SPSS, Poly analyst by Megaputer and many others, now, huge amount of data can

be processed and analyzed with the economizing of time. So, it can be concluded that the decision making in every field of life has become rational. However, the use of Data Mining techniques for knowledge discovery and decision making is subject to verification. Again, for such kind of verification, econometrics criteria can be used which generates future scope of research.

REFERENCES

[1] Brachman, R.J.; Anand, T. (1996): *The Process of Knowledge Discovery in Databases*. In Advances in Knowledge Discovery & Data Mining, Fayyad, U.M. - Piatetsky-Shapiro, G. - Smyth, P. - Uthurusamy, R., Eds. AAAI/MIT Press, Cambridge, Massachusetts.

[2] Chen, M.S.; Han, J.; Yu, P.S. (1996*): Data Mining: An Overview from a Database Perspective.* IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol.8, No.6, pp.866-883.

[3] Fayyad, U.M. (1996): *Data Mining and Knowledge Discovery: Making Sense Out of Data.* IEEE EXPERT, Vol.11, No.5, pp. 20-25.

[4] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. (1996): *From Data Mining to Knowledge Discovery: An Overview.* In Advances in Knowledge Discovery & Data Mining, Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R., Eds. AAAI/MIT Press, Cambridge, Massachusetts.

[5] Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. (1996): *The KDD Process for Extracting Useful Knowledge from Volumes of Data.* COMMUNICATIONS OF THE ACM, Vol.39, No.11, pp. 27-34.

[6] Freitas, A.A. (1997): *Generic, Set-Oriented Primitives to Support Data-Parallel Knowledge Discovery in Relational Database Systems.* Ph.D. Thesis, University of Essex, UK.

[7] Freitas, A.A.; Lavington, S.H. (1998): *Mining Very Large Databases with Parallel Processing.* Kluwer Academic Publishers, 1998, chapter Knowledge Discovery Paradigms. Table of contents on: http://www.ppgia.pucpr.br/~alex/book.html

[8] Hedberg, S. R. (1996): *Searching for the mother lode: tales of the first data miners.* IEEE EXPERT, Vol.11, No.5, pp. 4-7.

[9] Kohavi, R.; John, G. (1998): *The Wrapper Approach.* Book Chapter in Feature Selection for Knowledge Discovery and Data Mining. (Kluwer International Series in Engineering and Computer Science), Huan Liu and Hiroshi Motoda, editors.

[10] Mannila, H. (1997): *Methods and Problems in Data Mining.* In the proceedings of International Conference on Database Theory, Afrati, F. - Kolaitis, P., Delphi, Springer-Verlag.

[11] Mark, B. (1996): *Data mining - Here we go again?* IEEE EXPERT, Vol. 11, No.5.

[12] Simoudis, E. (1996): *Reality Check for Data Mining.* IEEE EXPERT, Vol.11, No.5

[13] Weiss, S.M.; Indurkhya, N. (1998): Predictive Data Mining. Morgan Kaufmann Publishers, Inc., San Francisco.

[14] Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R.Legates, J. O'Donnell, and C.M. Rowe, "Statistics for the evaluation and comparison of models." J. Geophy. Research 90 (1985), pp. 995-9005.