ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

Knowledge Discovery Process through Data Mining: A Technical Approach for Data Analysis

Sk.Abid Hussain, Ch.Venkateswarlu

Assistant Professor, Dept. of MCA, NEC-NELLORE, AP, India

Assistant Professor, Dept. of MCA, NEC-NELLORE, AP, India

ABSTRACT: Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and simplest analytical step in data mining is to describe the data - summarize its statistical attributes (such as means and standard deviations), visually review it using charts and graphs, and look for potentially meaningful links among variables. In the Data Mining Process, collecting, exploring and selecting the right data are critically important.

Knowledge Discovery demonstrates intelligent computing at its best, and is the most desirable and interesting end-product of Information Technology. To be able to discover and to extract knowledge from data is a task that many researchers and practitioners are endeavouring to accomplish. There is a lot of hidden knowledge waiting to be discovered – this is the challenge created by today's abundance of data. Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns from large datasets.

KEYWORDS: Data mining, knowledge discovery, machine learning, datasets

I. INTRODUCTION

Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns (implicit or explicit) which are the essence of useful knowledge. Advances in data gathering storage and distribution have created a need for computational tools and techniques to aid in data analysis. Data Mining and Knowledge Discovery in Databases is a rapidly growing area of research and application that builds on techniques and theories from many fields including statistics databases pattern recognition and learning data visualization uncertainty modelling data warehousing and OLAP optimization and high performance computing. KDD is concerned with issues of scalability, the multi-step knowledge discovery process for extracting useful patterns and models from raw data stores (including data cleaning and noise modelling) and issues of making discovered patterns understandable.

II. LITERATURE SURVEY

Knowledge Discovery includes: Theory and Foundational Issues: Data and knowledge representation; modelling of structured textual and multimedia data; uncertainty management; metrics of interestingness and utility of discovered knowledge; algorithmic complexity efficiency and scalability issues in data mining; statistics over massive data sets. Data Mining Methods: including classification clustering probabilistic modeling prediction and estimation dependency analysis search and optimization. Algorithms for data mining including spatial textual and multimedia data (e.g. the Web) scalability to large databases parallel and distributed data mining techniques and automated discovery agents.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

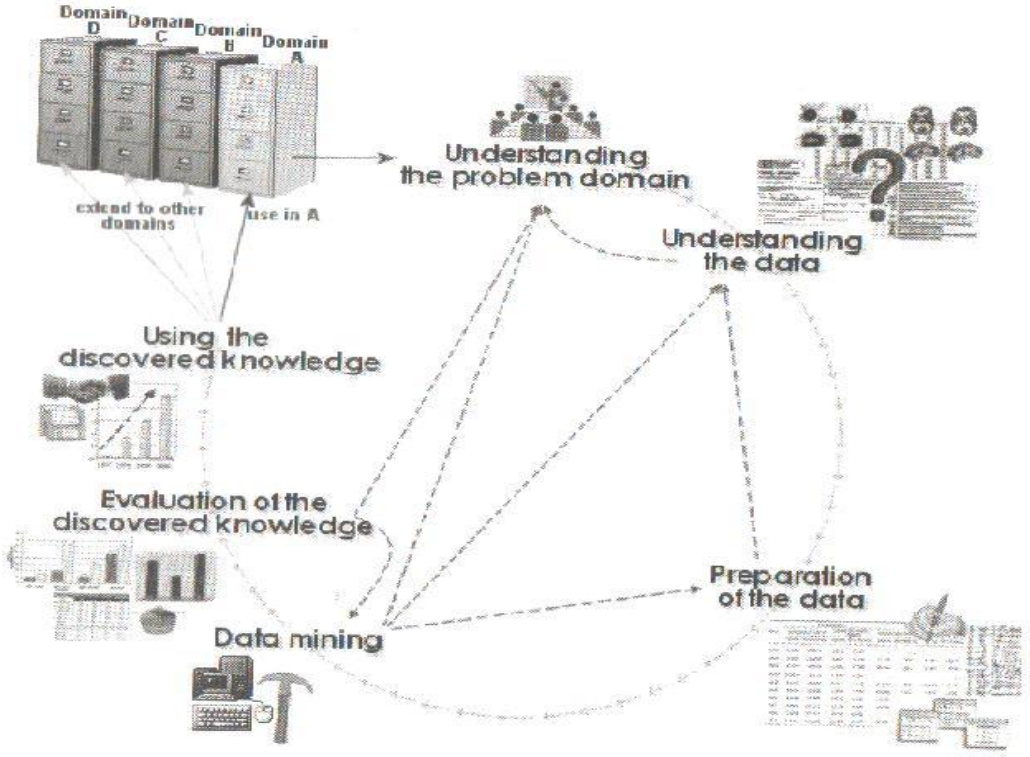
Vol. 2, Issue 7, July 2014

III. THE KDD PROCESS

The knowledge discovery process is iterative and interactive, consisting of several steps. The process starts with determining the KDD goals, and "ends" with the implementation of the discovered knowledge. As a result, changes would have to be made in the application domain (such as offering different features to mobile phone users in order to reduce churning). This closes the loop, and the effects are then measured on the new data repositories, and the KDD process is launched again. The following are the steps that are used:

III.I. Developing an understanding of the application domain. This is the initial preparatory step. It prepares the scene for understanding what should be done with the many decisions (about transformation, algorithms, representation, etc.).

III.II. Selecting and creating a data set on which discovery will be performed. Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includes finding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledge discovery into one data set, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail.



III.III. Pre-processing and cleansing. In this stage, data reliability is enhanced. It includes data clearing, such as handling missing values and removal of noise or outliers. Several methods are explained in the handbook, from doing nothing to becoming the major part (in terms of time consumed) of a KDD process in certain projects. It may involve complex statistical methods, or using specific Data Mining algorithm in this context.

III.IV. Data transformation. In this stage, the generation of better data for the data mining is prepared and developed. Methods here include dimension reduction (such as feature selection and extraction, and record sampling), and attribute

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

transformation (such as Discretization of numerical attributes and functional transformation). This step is often crucial for the success of the entire KDD project, but it is usually very project-specific.

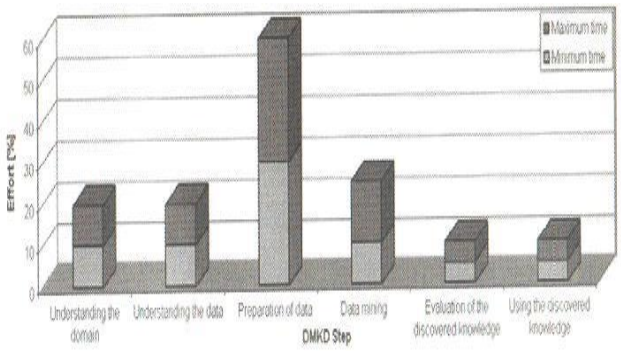
III.V. Choosing the appropriate Data Mining task. We are now ready to decide on which type of Data Mining to use, for example, classification, regression, or clustering. This mostly depends on the KDD goals, and also on the previous steps. There are two major goals in Data Mining: prediction and description. Prediction is often referred to as supervised Data Mining, while descriptive Data Mining includes the unsupervised and visualization aspects of Data Mining.

III.VI. Choosing the Data Mining algorithm. Having the strategy, we now decide on the tactics. This stage includes selecting the specific method to be used for searching patterns (including multiple inducers). For example, in considering precision versus understand ability, the former is better with neural networks, while the latter is better with decision trees. For each strategy of meta-learning there are several possibilities of how it can be accomplished.

III.VII. Employing the Data Mining algorithm. Finally the implementation of the Data Mining algorithm is reached. In this step we might need to employ the algorithm several times until a satisfied result is obtained, for instance by tuning the algorithm's control parameters, such as the minimum number of instances in a single leaf of a decision tree.

III.VIII. Evaluation. In this stage we evaluate and interpret the mined patterns (rules, reliability etc.), with respect to the goals defined in the first step.

The following figure presents a summary corresponding to the relative effort spent on each of the DMKD steps.



IV. DATA MINING METHODOLOGY

It should be clear from the above that data mining is not a single technique; any method that will help to get more information out of data is useful. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most methods commonly used for data mining can be classified into the following groups.

Statistical Methods:Historically, statistical work has focused mainly on testing of preconceived hypotheses and on fitting models to data. Statistical approaches usually rely on an explicit underlying probability model. In addition, it is generally assumed that these methods will be used by statisticians, and hence human intervention is required for the generation of candidate hypotheses and models.

Case-Based Reasoning: Case-based reasoning (CBR) is a technology that tries to solve a given problem by making direct use of past experiences and solutions. A case is usually a specific

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

problem that has been previously encountered and solved. Given a particular new problem, case-based reasoning examines the set of stored cases and finds similar ones. If similar cases exist, their solution is applied to the new problem, and the problem is added to the case base for future reference.

Neural Networks: Neural networks (NN) are a class of systems modeled after the human brain. As the human brain consists of millions of neurons that are interconnected by synapses, neural
networks are formed from large numbers of simulated neurons, connected to each other in a manner similar to brain neurons. Like in the human brain, the strength of neuron interconnections may change (or be changed by the learning algorithm) in response to a presented stimulus or an obtained output, which enables the
network to "learn".

Decision Trees:A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item. Depending on the outcome of the test, one chooses a certain branch. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). When a terminal node is reached, a decision is made. Decision trees can also be interpreted as a special form of a rule set, characterized by their hierarchical organization of rules.

Rule Induction*:*Rules state a statistical correlation between the occurrence of certain attributes in a data item, or between certain data items in a data set. The general form of an association rule is *Xl* ^ ... ^ *Xn => Y [C, S]*, meaning that the attributes *Xl,...,Xn* predict *Y* with a confidence *C* and a significance *S*.

Bayesian Belief Networks: Bayesian belief networks (BBN) are graphical representations of probability distributions, derived from co-occurrence counts in the set of data items. Specifically, a BBN is a directed, acyclic graph, where the nodes represent attribute variables and the edges represent probabilistic dependencies between the attribute variables. Associated with each node are conditional probability distributions that describe the relationships between the node and its parents.
Genetic algorithms / Evolutionary Programming: Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed
in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other.

Fuzzy Sets: Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today's databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable.

Rough Sets:A rough set is defined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain nonmember of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly (but not certainly) amember of the set.

V. CONCLUSION AND FUTURE WORK

Knowledge discovery can be broadly defined as the automated discovery of novel and useful information from commercial databases. Data mining is one step at the core of the knowledge
discovery process, dealing with the extraction of patterns and relationships from large amounts of data. Today, most enterprises are actively collecting and storing large databases. Many of them have recognized the potential value of these data as an information source for making business decisions.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 7, July 2014

REFERENCES

1. Arbel, R. and Rokach, L., Classifier evaluation under limited resources, Pattern Recognition Letters, 27(14): 1619–1631, 2006,
2. Cohen S., Rokach L., Maimon O., Decision Tree Instance Space Decomposition with Grouped Gain-Ratio, Information Science, Volume 177, Issue 17, pp. 3592-3612, 2007.
3. Hastie, T. and Tibshirani, R. and Friedman, J. and Franklin, J., The elements of statistical learning: data mining, inference and prediction, The Mathematical Intelligencer, 27(2): 83–85, 2005.
4. Han, J. and Kamber, M., Data mining: concepts and techniques, Morgan Kaufmann, 2006. H. Kriege, K. M. Borgwardt, P. Krger, A. Pryakhin, M. Schubert and Arthur Zimek, Future trends in data mining, Data Mining and Knowledge Discovery, 15(1):87-97, 2007.
5. Larose, D.T., Discovering knowledge in data: an introduction to data mining, JohnWiley and Sons, 2005. Maimon O., and Rokach, L. Data Mining by Attribute Decomposition with semiconductors manufacturing case study, in Data Mining for Design and Manufacturing: Methods and Applications, D. Braha (ed.), Kluwer Academic Publishers, pp. 311–336, 2001.
6. Maimon O. and Rokach L., "Improving supervised learning by feature decomposition", Proceedings of the Second International Symposium on Foundations of Information and Knowledge Systems, Lecture Notes in Computer Science, Springer, pp. 178-196, 2002.
7. Maimon, O. and Rokach, L., Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications, Series in Machine Perception and Artificial Intelligence - Vol. 61, World Scientific Publishing, ISBN:981-256-079-3, 2005.

BIOGRAPHY

ABID HUSSAIN SHAIK is a Assistant Professor in the Department of Master of Computer Applications, Narayana Engineering College, Nellore. He received Master of Computer Application (MCA) degree in 2011 from JNTUA, India. His research interests are Data Mining, Computer Networks, and Test Designing etc.

Venkateswarlu CH is a Assistant Professor in the Department of Master of Computer Applications, Narayana Engineering College, Nellore. His research interests are Data Mining, Computer Networks, and Test Designing etc.