A STUDY ON ROLE OF DATA MINING IN RESEARCH METHODOLOGY

Akanksha. A. Kherdikar-Kurlekar,
Lecturer, ASM's Institute of Computer Studies,
Pimpri, Pune, India

Anusuya .S,
Asst.Professor, ASM's Institute of Computer Studies,
Pimpri, Pune, India

ABSTRACT

The term "Research" pertains to a "Search of facts". It is the process which includes defining and redefining problems, formulating hypothesis, collecting, organizing and evaluating data; making deductions and reaching conclusions and at last presenting it in a detailed, accurate manner. Research Methodology is a way to systematically solve the research problem by logically adopting various steps to refer the logic behind the research process. The main aim of research is to find out the hidden truth and Information Technology provides such useful tools to make this research process easier, accurate and reliable. Data Mining is one such tool in researcher's tool bag which facilitates the research process. Data Mining and Knowledge Discovery in Databases (KDD) are rapidly evolving areas of research that are at the intersection of several disciplines, including statistics, databases, pattern recognition and optimization. The term data mining has been mostly used by statisticians, data analysts, and database communities and also by the researchers. This paper mainly focuses on concepts of data mining such as Classification, Clustering, Regression, Association, Estimation and Prediction that help the researchers in the areas of research process such as Literature Review, Collection, Analysis, Interpretation and Representation of data. Present study also concentrates on role of data mining in research methodology to reveal patterns and relationships from the piles of data using the software's like CART, SPSS, and MATLAB etc.

Keywords: Data Mining, MATLAB, SPSS, CART, Classification, clustering, collection of data, data analysis and Data sampling tool.

INTRODUCTION:

1.1 Data Mining:

Data is raw. It simply exists and has no significance beyond its existence; Information is data that has been given meaning by way of relational connection; Knowledge is the appropriate collection of information, such that it's intent is to be useful. Knowledge is a deterministic process. Data collection is the systematic recording of information; data analysis involves working to uncover patterns and trends in data sets; data interpretation involves explaining those patterns and trends.

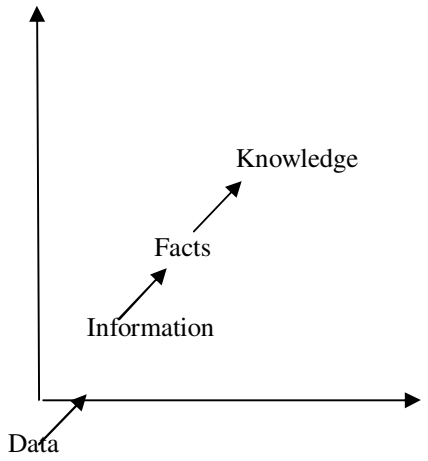


Fig.1: Levels of Data

Data preprocessing or data cleaning or data preparation is also a key part of data mining. Quality decisions and quality mining results come from quality data. Data are always dirty and are not ready for data mining in the real world. For example,

- Data need to be integrated from different sources;
- Data contain missing values. i.e. incomplete data;
- Data are noisy, i.e. contain outliers or errors, and inconsistent values (i.e. contain discrepancies in codes or names);
- Data are not at the right level of aggregation.

Data is gathered because it is needed for some operational purpose, e.g. inventory control or billing. And, once it has served that purpose, it languishes on tape or gets discarded. For learning to take place, data from many sources must first be gathered together and organized in a consistent and useful way. This is called Data Warehousing.

Data warehousing provides the enterprise with a memory. But, memory is of little use without intelligence. That is where data mining comes in. Intelligence allows us to comb through our memories noticing patterns, devising rules, coming up with new ideas to try, and making predictions about the

future. The data must be analyzed, understood, and turned into actionable information. Data mining provides tools and techniques that add intelligence to the data warehouse. Data mining provides the enterprise with intelligence.

Data mining consists of five major elements: i) Extract, transform, and load transaction data onto the data warehouse system. ii) Store and manage the data in a multidimensional database system. iii) Provide data access to business analysts and information technology professionals. iv) Analyze the data by application software. V) Present the data in a useful format, such as a graph or table.

1.1.1 The main data mining tasks:

The main tasks of data mining involve extracting meaningful new information from the data. Knowledge discovery (learning from data) comes in two flavors: directed (supervised) and undirected (unsupervised) learning from data.

The six main activities of data mining are:

Classification: - It is the task of generalizing known structure to apply to new data.

Clustering - It is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Regression - Attempts to find a function which models the data with the least error.

Association rule learning - Searches for relationships between variables.

Estimation: - Given some input data, coming up with a value for some unknown continuous variable such as income, height, or credit-card balance.

Prediction: - It is same as classification and estimation except that the records are classified according to some predicted future behavior or estimated future value.

The choice of a particular combination of data mining techniques to apply in a particular situation depends on both the nature of the data mining task to be accomplished and the nature of the available data.

1.2 Research Methodology:

Research is the systematic approach towards purposeful investigation. This needs formulating hypothesis, collection of data on relevant variables, analyzing and interpreting the results and reaching conclusions either in form of a solution or certain generalizations.

Research Methodology is a scientific and systematic way to solve research problems. A researcher has to design his methodology i.e. in addition to the knowledge of methods/techniques; he has to apply the methodology as well. The methodology may differ from problem to problem. Thus, the scope of research methodology is wider than research methods. In a way, research methodology deals with the research methods and takes into considerations the logic behind the methods, we use. In Research methodology collection of data is one of important step, once the data is collected, it must be analyzed properly to bring out the important features. This needs arranging of the data according to certain common features which leads to Classification, Tabulation and representation of Data.

1.3 Data mining in Research:

Collecting data is only one step in a scientific investigation, and scientific knowledge is much more than a simple compilation of data points. The world is full of observations that can be made, but not every observation constitutes a useful piece of data. All researchers make choices about which data are most relevant to their research and what to do with that data: how to turn a collection of measurements into a useful dataset through processing and analysis, and how to interpret those analyzed data in the context of what they already know. The thoughtful and systematic collection, analysis, and interpretation of data allow it to be developed into evidence that supports scientific ideas, arguments, and hypotheses. Thus Data Mining techniques can be useful for researchers for analyzing data as it extracts or mines the knowledge from large amount of data. The Present study serves cup of methods like classification and clustering for facilitating research work.

2. DATA MINING TOOLS:

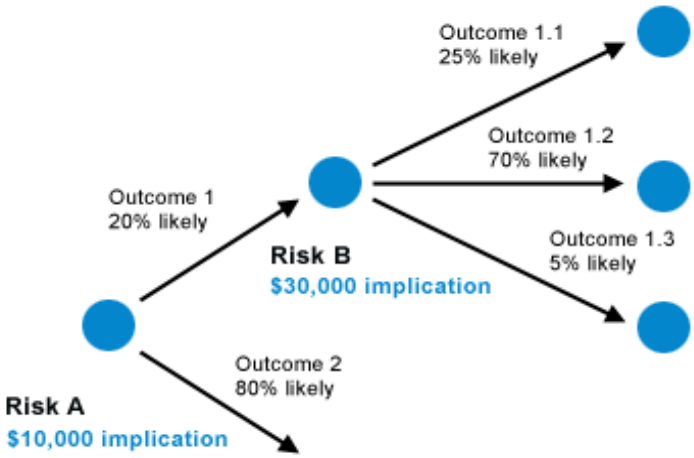
2 .1 Classifications:

It is the process of arranging the data in group and classes according to resemblances and similarities. This classification can be done with the help of data mining technique called as "Classification technique" which is used to predict group membership for data instances. For example, Prediction of whether the weather on a particular day will be "sunny", "rainy" or "cloudy" Popular classification techniques include decision trees and neural networks and Rules.

2.1.1 Decision Trees:

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. Decision trees can be used for segmentation of the original dataset (each segment would be one of the leaves of the tree) like segmentation of customers, products, and sales regions etc so that similar data can be grouped in one segment. The segmentation is done for the prediction of some important piece of information. The records that fall within each segment fall there because they have similarity with respect to the information being predicted - not just that they are similar - without similarity being well defined. These predictive segments that are derived from the decision tree also come with a description of the characteristics that define the predictive segment. Thus the decision trees that create them may be complex; the results can be presented in an easy to understand way that can be quite useful to the researcher. Decision tree can be used for problems ranging from credit card attrition prediction to time series prediction of the exchange rate of different international currencies. They can be used for Exploration of data, Data Preprocessing and also for Prediction.

For example: A company needs to place a large equipment order. Project manger think there is a 20 percent risk that their primary hardware supplier may not be able to provide all the equipment that is needed for a large order in a timely manner. This could be risk A.As a part of the risk response plan, Project manager decide to talk to a second vendor to see if they can help fulfill the equipment order on short notice. The vendor normally has the equipment in stock. However, he also discovers that there is a 25 percent possibility that there may be a disruption in their plant because of a potential strike. This is risk B. in this example two risks are interconnected so decision tree now comes into picture to determine the probability and impact of each risk combination.



So company should try to achieve outcome 1.3 because it has the smallest financial risk impact. If company doesn't think that it can achieve outcome 1.3, it should try for outcome 2. There is an 80 percent chance of hitting outcome 2.

Software's for decision tree - C4.5, GAtree, Mangrove, OC1, ODBCMINE, PC4.5, SMILES, Random forests from Leo Breiman and YaDT: Yet another Decision Tree builder and CART 5.0 decision-tree software, SPSS AnswerTree.

2.1.2 Neural Network:

True neural networks are biological systems (BRAIN) that detect patterns, make predictions and learn. The human brain is a very complex part of the human body, due mainly to the interactions and connectivity with other parts of our body, and the way it controls and defines every aspect of our being.

The purpose of a neural network is to learn to recognize patterns in data. Once the neural network has been trained on samples , it can make predictions by detecting similar patterns in future data.

Neural networks provide a range of powerful new techniques for solving problems in pattern recognition, data analysis, and control. They have several notable features including high processing speeds and the ability to learn the solution to a problem from a set of example.

Example: In order to make a prediction the neural network accepts the values for the predictors on what are called the input nodes. These become the values for those nodes those values are then multiplied by values that are stored in the links (sometimes called links and in some ways similar to the weights that were applied to predictors in the nearest neighbor method). These values are then added together at the node at the far right (the output node) a special threshold function is applied and the resulting number is the prediction. In this case if the resulting number is 0 the record is considered to be a good credit risk (no default) if the number is 1 the record is considered to be a bad credit risk (likely default).

A simplified version of the calculations is shown in below figure. Here the value age of 47 is normalized to fall between 0.0 and 1.0 and has the value 0.47 and the income is normalized to the value 0.65. This simplified neural network makes the prediction of no default for a 47 year old making $65,000. The links are weighted at 0.7 and 0.1 and the resulting value after multiplying the node values by the link weights is 0.39. The network has been trained to learn that an output value of 1.0 indicates default and that 0.0 indicate non-default. The output value calculated here (0.39) is closer to 0.0 than to 1.0 so the record is assigned a non-default prediction.

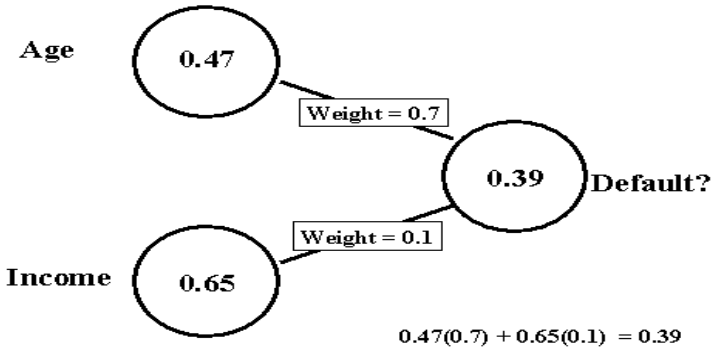


Figure: -The normalized input values are multiplied by the link weights and added together at the output.

Software for Neural Network: NuClass7, Sciengy RPF, Sharky Neural Network and Alyuda NeuroIntelligence, MATLAB Neural Net Toolbox, SPSS Neural Connection 2

2.1.3 Association rule learning:

Rule induction is one of the major forms of data mining and is the most common form of knowledge discovery in unsupervised learning systems. Rule induction on a data base can be a massive undertaking where all possible patterns are systematically pulled out of the data and then an accuracy and significance are added to them that tell the user how strong the pattern is and how likely it is to occur again. In general these rules are relatively simple such as for a market basket database of items scanned in a consumer market basket an interesting correlation in database can be found as:

- If bagels are purchased then cream cheese is purchased 90% of the time and this pattern occurs in 3% of all shopping baskets.
- If live plants are purchased from a hardware store then plant fertilizer is purchased 60% of the time and these two items are bought together in 6% of the shopping baskets.

The rules that are pulled from the database are extracted and ordered to be presented to the user based on the percentage of times that they are correct and how often they apply.

2.1.3.1 Prediction:

After the rules are created and their interestingness is measured there is a call for performing prediction with the rules. Each rule by itself can perform prediction - the consequent is the target and the accuracy of the rule is the accuracy of the prediction. But because rule induction systems produce many rules for a given antecedent or consequent there can be conflicting predictions with different accuracies. This is an opportunity for improving the overall performance of the systems by combining the rules. This can be done in a variety of ways by summing the accuracies as if they were weights or just by taking the prediction of the rule with the maximum accuracy.

Table 2 shows how a given consequent or antecedent can be part of many rules with different accuracies and coverage. From this example consider the prediction problem of trying to predict whether milk was purchased based solely on the other items that were in the shopping basket. If the shopping basket contained only bread then from the table we would guess that there was a 35% chance that milk was also purchased. If, however, bread and butter and eggs and cheese were purchased what would be the

prediction for milk then? 65% chance of milk because the relationship between butter and milk is the greatest at 65%? Or would all of the other items in the basket increase even further the chance of milk being purchased to well beyond 65%? Determining how to combine evidence from multiple rules is a key part of the algorithms for using rules for prediction.

Antecedent	Consequent	Accuracy	Coverage
bagels	cream cheese	80%	5%
bagels	orange juice	40%	3%
bagels	coffee	40%	2%
bagels	eggs	25%	2%
bread	milk	35%	30%
butter	milk	65%	20%
eggs	milk	35%	15%
cheese	milk	40%	8%

Table 2 Accuracy and Coverage in Rule Antecedents and Consequents

Software's for Association rule: - arules, Apriori, Apriori, FP-growth, Eclat and DIC implementations, ARtool, DM-II system, Magnum Opus Demo and Azmy SuperQuery, The LPA Data Mining Toolkit

2.2 Clustering:

Identifying groups of individuals or objects that are similar to each other but different from individuals in other groups can be intellectually satisfying, profitable, or sometimes both. One of the areas of management research is customer segmentation that forms clusters of customers who have similar buying habits or demographics. Similarly, based on scores on psychological inventories, we can cluster patients into subgroups that have similar response patterns. This may help in targeting appropriate treatment and studying typologies of diseases. Clustering is also a very popular method of microarray analysis to group co-expressed genes. Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. In educational research analysis, data for clustering can be students, parents, sex or test score. Clustering is an important method for understanding and utility of cluster in educational research. Cluster analysis in educational research can be used for data exploration, cluster confirmation and hypothesis testing. Data exploration is used when there is little information about which schools or students will be grouped together. It aims at discovering any meaningful clusters of units based on measures on a set of response variables. Cluster confirmation is used

for confirming the previously reported cluster results. Hypothesis testing is used for arranging cluster structure.

Software for Clustering :- Autoclass C, CLUTO, Databionic ESOM Tools, David Dowe Mixture Modeling page, MCLUST/EMCLUST, PermutMatrix, Snob, StarProbe and BayesiaLab, ClustanGraphics3.

3. DATA MINING SOFTWARE FOR RESEARCH:

3.1 Spss (Statistical Product And Service Solutions):

It is statistical package for beginning, intermediate and advance data analysis. It can perform highly complex data manipulation and analysis with simple instructions. It helps researcher in perform following functions;

- Data Entry
- Variables Calculation
- Select or weight cases in data
- Describing the data (frequencies, graphics, statistics)
- Perform statistical tests (t-test, chi-square test, correlation test)
- Perform statistical analysis (anova, regression, factor, discriminant, ...)
- Graphical presentation of data

3.1.1 Hypothesis Testing:

Hypothesis testing is a statistical procedure used to “accept” or “reject” the hypothesis based on sample information. If hypothesis is correct, it will fall in the confidence interval (known as supported). If hypothesis is incorrect, it will fall outside the confidence interval (known as not supported). IBM’s SPSS Statistics Base is easy to use and forms the foundation for many types of statistical analyses. The procedures within IBM SPSS Statistics Base used to get a quick look at data, formulate hypotheses for additional testing, and then carry out a number of statistical and analytic procedures to help clarify relationships between variables, create clusters, identify trends and make predictions.

The following tasks on hypotheses can be performed using IBM Statistics Base.

- Quickly access and analyze massive datasets.
- Easily prepare and manage data for analysis.
- Analyze data with a comprehensive range of statistical procedures.
- Easily build charts with sophisticated reporting capabilities.
- Discover new insights in data with tables, graphs, cubes and pivoting technology.
- Quickly build dialog boxes or let advanced users create customized dialog boxes.

3.2. Matlab:

It stands for MATrix LABoratory, is a state of the art mathematical software package, which can be used by the researcher. It provides interactive environment for data visualization, data analysis, and numeric computation.

- High-level language for technical computing.
- Development environment for managing code, files, and data.
- Interactive tools for iterative exploration, design, and problem solving.
- Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization, and numerical integration.
- 2-D and 3-D graphics functions for visualizing data.
- Tools for building custom graphical user interfaces.

4. DATA MINING SYSTEMS:

- Clementine by SPSS, Enterprise Miner from SAS Institute, Insightful Miner from Insightful Inc. and Intelligent Miner from IBM provides a wide range of data mining functions, including association mining, classification, regression, predictive modeling, deviation detection, clustering and sequential pattern analysis.
- Microsoft SQL Server 2005, is a database management system that incorporates multiple data mining functions smoothly in its relational database system and data warehouse environments
- Oracle Data Mining(ODM), an option to oracle Database 10g Enterprise Edition, provides several data mining functions, including association mining, classification, prediction, regression, clustering , and sequence similarity search and analysis.
- CART from Salford Systems provides decision trees for classification and regression trees for prediction.

5. CONCLUSION:

There are definite differences in the types of problems faced by the researcher, that are most conducive to each technique but the reality of real world data and its dynamic nature needs to be considered as an important factor in determining the methodologies/techniques. Hence researcher has to choose appropriate data mining technique by slicing and dicing research data. This paper brings out the approaches by analyzing, interpreting, grouping and representing the data of various nature to make it a useful information for researcher. Further studies will be focused on different aspects of data mining in research.

REFERENCES:

Han Jiawei and Michelime Kamber(2006), "Data Mining Concepts and Techniques", 2nd Edition, Morgan Kaufmann Publishers.

Bhattacharyya Dipak Kumar (2006), "Research Methodology", 2nd Edition, Excel Books Publication.

http://spss.co.in/spssstatistics.aspx

http://www.thearling.com

http://www.kdnuggets.com

http://linguistics.byu.edu/faculty/henrichsenl/researchmethods/RM_1_05.html

http://changingminds.org/explanations/research/measurement/types_data.htm