A Review Study of various Data Mining Classification & Clustering Techniques

Karambeer Kaur
M.Tech Scholar
Om Institute of Technology & Management, Hisar

Mr Surender Singh
Assistant Professor
Om Institute of Technology & Management, Hisar

ABSTRACT

Data mining application includes a variety of methodologies that have been developed by commercial & research centers. This technique has been used for industrial, commercial and scientific purposes. It is most useful in an exploratory analysis scenario in which there are no prearranged notions about what will compose an "interesting" outcome. The WEKA contains a set of visualization tools & algorithms for data analysis and predictive modeling, together with graphical user interfaces for simple access to this functionality.

KEYWORDS: WEKA Software, Data Mining, Classification, Clustering

INTRODUCTION

The Data mining software application includes various methodologies that have been developed by both commercial and research centers. These techniques have been used for industrial, commercial and scientific purposes. For example, data mining has been used to analyze large datasets and establish a useful classification and patterns in the data sets. Agricultural, medical, education and biological research studies have used various techniques of data analysis, including, natural trees, statistical machine learning, classification, clustering and other analysis methods. The main objectives of our work are to investigate the performance analysis of different classification and clustering methods using the WEKA software for education dataset. In this thesis, we present the comparison of different classification and clustering techniques using Waikato Environment for the knowledge Analysis or in short WEKA and developed at the University of Waikato. The study in this thesis will focus on the use of data mining techniques or pervious analyzed data set. The data mining tool WEKA will be used. WEKA is the free software available under the GNU general public license. WEKA is the open source software which consists of a collection of machine learning algorithms for data mining tasks. [3]

Data mining is the process of finding of hidden information from a huge amount of data. Data mining analyzing the data from different source and convert it into meaningful information. Data mining is a new powerful technology that helps business to focus on important information like future trends, decision making, customer choice etc. A target dataset is prepared before applying the data mining algorithm. The common source of data is the data warehouse. Pre–processing is needed to analyze the data sets before applying the data mining. [2]

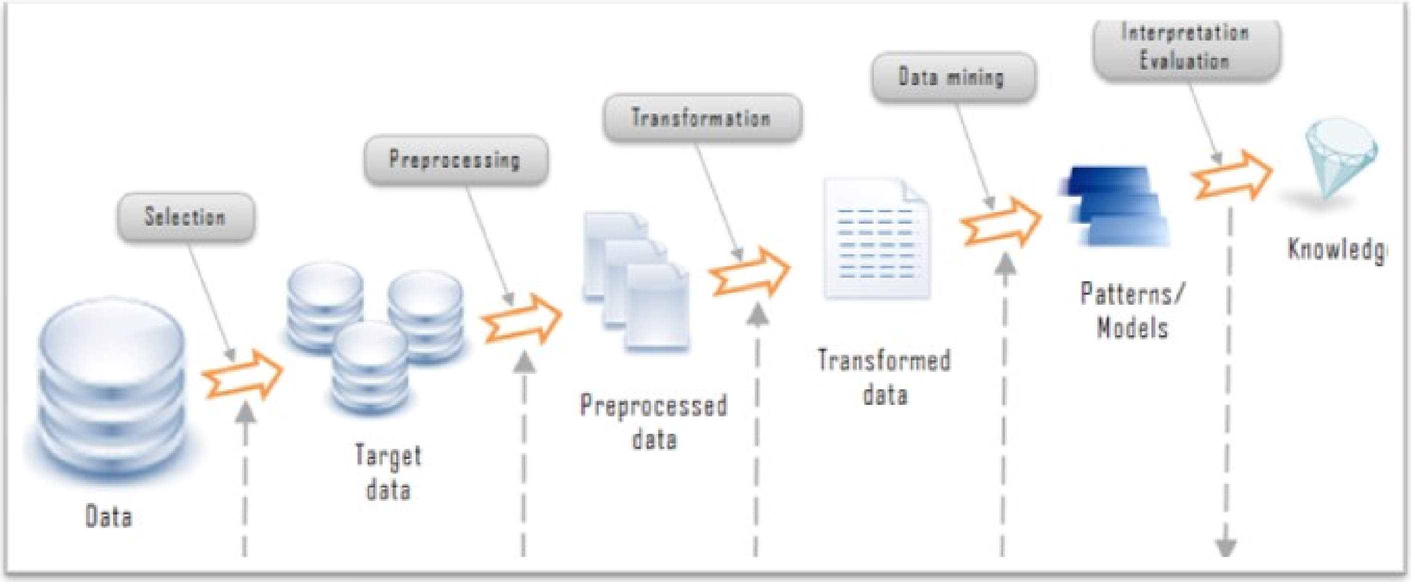
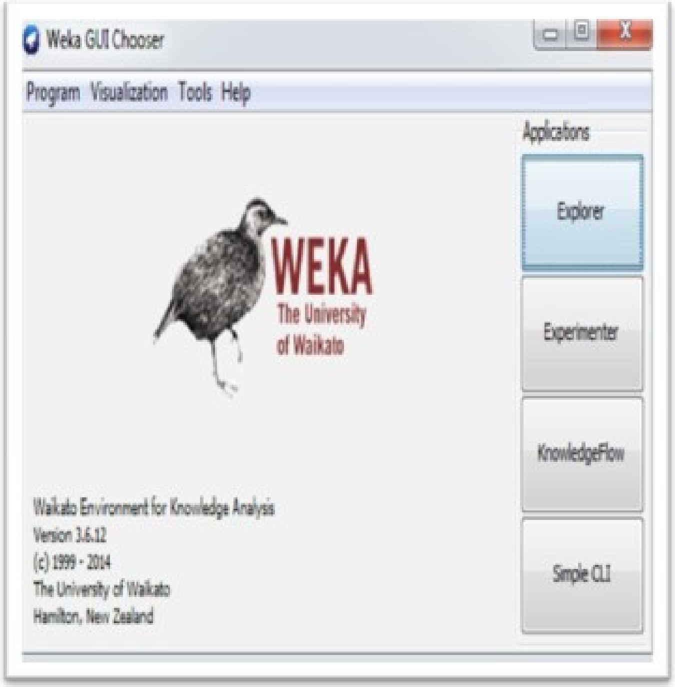


Figure 1: Knowledge Discovery Process (KDP)

Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. According to this definition, data is a set of facts that is somehow accessible in electronic form. The term "patterns" indicates models and regularities which can be observed within the data. Patterns have to be valid, i.e. they should be true on new data with some degree of certainty. A novel pattern is not previously known or trivially true. The potentially usefulness of patterns refers to the possibility that they lead to an action providing a benefit. A pattern is understandable if it is interpretable by a human user. At last KDD is a process, indicating that there are several steps that are repeated in several iterations. [5]

WEKA: A DATA MINING SOFTWARE

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to this functionality. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka), is now used in many different application areas, in particular for educational purposes and research. [6]



Waikato Environment for Knowledge Analysis
Version 3.6.12
(c) 1999 - 2014
The University of Waikato
Hamilton, New Zealand

Figure 2: Weka User Interface

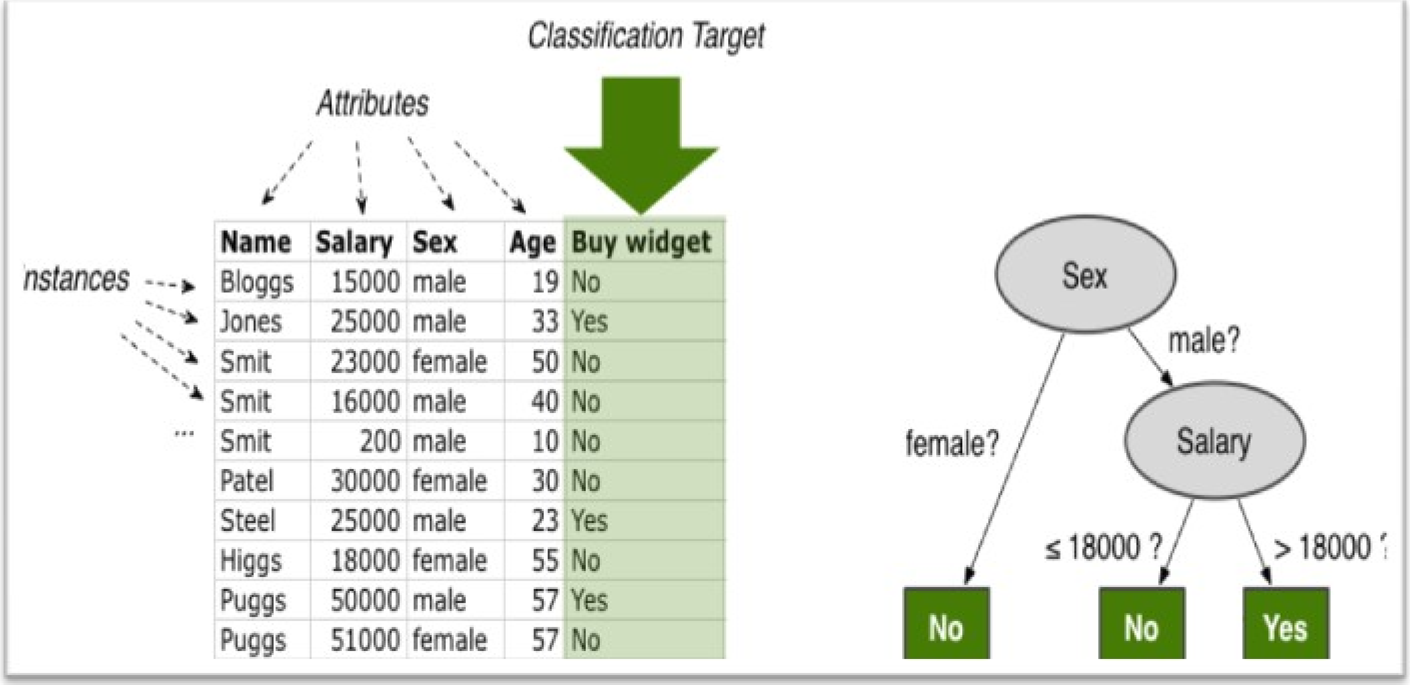
CLASSIFICATION & CLUSTERING TECHNIQUES

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. [4]

Types of classification models:

1. Classification by decision tree induction
2. Bayesian Classification
3. Neural Networks
4. Support Vector Machines (SVM)
5. Classification Based on Associations



Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

Figure 3: Example of Classification Technique

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. [7]

Types of clustering methods:

1. Partitioning Methods
2. Hierarchical Agglomerative (divisive) methods
3. Density based methods
4. Farthest First method
5. Grid-based methods
6. Model-based methods

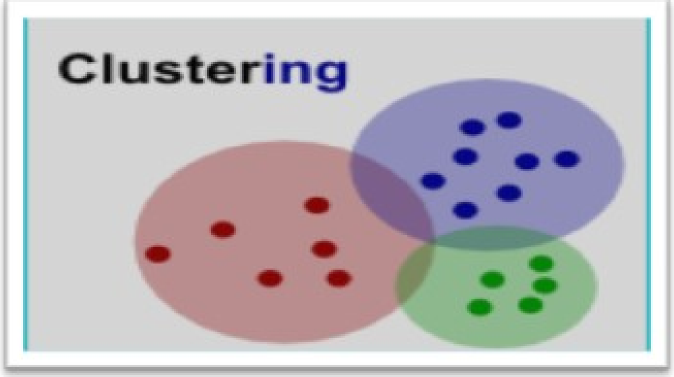


Figure 4: Example of Clustering

CONCLUSION

Data Mining is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome.

In practice, the two primary goals of Data Mining tend to be prediction and description. Prediction involves using some variables or fields in the data set to predict unknown or future values of other variables of interest. Data Mining (DM) represents a set of specific methods and algorithms aimed solely at extracting patterns from raw data. Data mining sometimes is also called knowledge discovery in databases (KDD). Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modeling of large data repositories. KDD is the organized process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. We can also find the existing relationships and patterns. Data mining combines machine learning, statistics and visualization techniques to discover and extract knowledge.

REFERENCES

[1] Nikita Jain, Vishal Srivastava" DATA MINING TECHNIQUES: A SURVEY PAPER" IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308

[2] M.Vijayakamal, Mulugu Narendhar "A NOVEL APPROACH FOR WEKA & STUDY ON DATA MINING TOOLS" International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 2, August 2012.

[3] Fateh Ahmadi " DATA MINING IN TEACHER EVALUATION SYSTEM USING WEKA" International Journal of Computer Applications (0975 – 8887) Volume 63– No.10, February 2013.

[4] Amandeep Kaur Man, Navneet Kaur" SURVEY PAPER ON CLUSTERING TECHNIQUES" International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013

[5] Kalyani M Raval" DATA MINING TECHNIQUES" International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 10, October 2012 ISSN: 2277 128X

[6] Swasti Singhal, Monika Jena" A STUDY ON WEKA TOOL FOR DATA PREPROCESSING, CLASSIFICATION AND CLUSTERING" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[7] Mrs. Bharati M. Ramageri" DATA MINING TECHNIQUES AND APPLICATIONS" Bharati M. Ramageri / Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305.

[8] Pallavi, Sunila Godara "A COMPARATIVE PERFORMANCE ANALYSIS OF CLUSTERING ALGORITHMS" Pallavi, Sunila Godara / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 1, Issue 3, pp.441-445

[9] Dr. Mohd Maqsood Ali" ROLE OF DATA MINING IN EDUCATION SECTOR" International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 2, Issue. 4, April 2013, pg.374 – 383

[10] S. Anupama Kumar and M. N. Vijayalakshmi "RELEVANCE OF DATA MINING TECHNIQUES IN EDUCATION SECTOR" International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013

[11] Priyanka Sharma" COMPARATIVE ANALYSIS OF VARIOUS DECISION TREE CLASSIFICATION ALGORITHMS USING WEKA" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 2 684 – 690

[12] Umamaheswari. K, S. Niraimathi" A STUDY ON STUDENT DATA ANALYSIS USING DATA MINING TECHNIQUES" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 8, August 2013 ISSN: 2277 128X

[13] Sunita Beniwal, Jitender Arora" CLASSIFICATION AND FEATURE SELECTION TECHNIQUES IN DATA MINING" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012 ISSN: 2278-0181.

[14] Aditi Mahajan, Anita Ganpati "PERFORMANCE EVALUATION OF RULE BASED CLASSIFICATION ALGORITHMS" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 10, October 2014.

[15] K. Wisaeng" A COMPARISON OF DIFFERENT CLASSIFICATION TECHNIQUES FOR BANK DIRECT MARKETING" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-4, September 2013.