Implementation of Knime-Data Mining Tool

Lalit Kataria
RKSD (PG) College, Kaithal
India

Abstract—Today data mining is one of the emerging research area in the field of computer science. Data mining is used for many business application finically sector etc. thus to implement the concept of data mining there are many tool available which provide the functionally of data mining algorithms .Thus in this paper we present the implementation of some data mining techniques through KNIME tool by configuring KNIME nodes KNIME tool provides the different nodes like file reader node, parathion node, decision leaner tree, decision predictor node scorer, color manager node all these nodes can work independent, But by using the output of a node as an input in other node one can implemented the data mining techniques like clustering, Association or some predicative models. KNIME also provide the functionality to represent the different cluster in the form of group which has different colors.

In This paper we have also implemented the K-means Algorithm through K-means Node which is used for clustering K-means,& Scorer etc.

Keywords—KNime, Predictors Node, K Node, Clustering K means Node, scorar Node.

I. INTRODUCTION

Data mining is the science of extracting useful information from large amount of data sets and databases. Data mining is also a analysis, which observe datasets to find unsuspected relationship and summarize the data in a novel way after that data are both understandable and useful. In fact data mining is use for patter reconigation the relation and summaries, which is derived from data after the data mining is called patterns and rules or models. Data mining is possible on the data which is already collected thus data mining simply stated that “Extracting or mining” knowledge from large data amounts.[10][13][14]. It is also known as “knowledge mining from data” or “knowledge mining” So mining is the term which characterizing the process that finds small sets of precious nuggets from a great deal of raw material. So it can be said, that “Data mining is the process of discovering the interesting knowledge from large amount of data stores in databases, data ware house and other information repositories” [12].Thus mining is gold from rocks. Data mining is possible on any kind of data repository and database, it may be rational database, data ware house, transactional databases and advance databases like specific application oriented database.[11]. Data mining extracts new information from data. Data mining tools do more than query and analysis tools, OLAP tools, or statistical techniques like an analysis of variance to name just a few examples. Understanding the kinds of questions data mining tools can answer is the best way to appreciate how they differ from other approaches.[8] Other query and analysis tools can respond to questions such as, "Do sales of Product X increase in November," or "Do sales of Product X decrease when there is a promotion on Product Y?" In contrast, one can use a data-mining tool to ask, "What are the factors that determine sales of Product X?"[7][9].

II. RELATED WORK

Michael R. Berthold this paper presents modular architecture it is easy to designate specific nodes to be run on separate machines. The meta-node abstraction provides a mechanism to encapsulate workflows and to assign them to dedicated severs for distributed processing, resulting in a significant acceleration of the workflow execution. But to accommodate the increasing availability of multi-core machines, also the support for shared memory parallelism becomes increasingly important. Knime will offer a unified framework to parallelize data-parallel operations as well as the distribution of operations on a cluster or a GRID [1][3]. Dominik Morent this paper demonstrated the new PMML preprocessing support introduced in the KNIME 2.4 release. In addition to adding support for PMML-preprocessing it is now also possible to modify (parts of) existing PMML codes modeled by KNIME workflows. We concluded by showing how meta learning schemes such as boosting can also be modeled in KNIME and how this can be extended to create PMML code for ensembles of models.[5]
STEPS OF DATA MINING

![Diagram of data mining process]

**Figure 1:** Various steps that are involved in mining data[9].

**Data Integration:** First of all the data are collected and integrated from all the different sources.

**Data Selection:** In this section it may not all collected data in the first step. So in this step select those data sets which are use full for data mining.[5][4]

**Data Cleaning:** The collected data are not clean and may contain errors, missing values, noisy or inconsistent data. So it is required to apply different techniques to get rid of such anomalies.

**Data Transformation:** The data even after cleaning are not ready for mining as needed to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.[2].

**Data Mining:** Now data is ready to apply data mining techniques on the data to discover the interesting patterns. Techniques like clustering and association analysis are among the many different techniques used for data mining.

**Pattern Evaluation and Knowledge Presentation:** This step involves visualization, transformation, removing redundant patterns etc. Decisions / Use of Discovered Knowledge: This step helps user to make use of the knowledge acquired to take better decisions. [9].

### III. KNIME: Data Mining Tool

KNIME was developed (and will continue to be expanded) by the Chair for Bioinformatics and Information Mining at the University of Konstanz, Germany. The group headed by Michael Berthold also uses KNIME for teaching and research at the University. Quite a number of new data analysis methods developed at the chair are integrated in KNIME. The KNIME base version already incorporates hundreds of processing nodes for data I/O, preprocessing and cleansing, modeling, analysis and data mining as well as various interactive views, such as scatter plots, parallel coordinates and others. It integrates all analysis modules of the well known Weka data mining environment and additional plug-in allow R-scripts to be run, offering access to a vast library of statistical routines. KNIME is based on the Eclipse platform and, through its modular API, easily extensible. When desired, custom nodes and types can be implemented in KNIME within hours thus extending KNIME to comprehend and provide first-tier support for highly domain-specific data. This modularity and extensibility permits KNIME to be employed in commercial production environments as well as teaching and research prototyping settings.[6]

### IV. IMPLEMENTATION

![KNIME Nodes Configuration]

**Figure 2:** KNIME Nodes Configuration
This figure shows a decision tree predictor node can be used as a predictor which is based on the model. It is also necessary that before using the decision predictor node one must use the decision learner node. Thus the output of the decision learner node works as an input to the decision predictor node. This node comes under the data mining node category. The decision tree as given in the model port along with the classified data. The tree can be expanded and collapsed with the plus/minus signs. This is also used for Hilite points. Number of patterns for Hilite determines the maximum number of patterns the tree will store to support Hilting.

Figure 3: Data Classification

Figure shows the data points in three colors Red, Green, and Blue. These colors are assigned to the data points according to the criteria which is taken when scatter node is configure. The dots display the colors which are defined by the color manager node. Thus the colors of the nodes depend on the value of data attributes.

Figure 4: Decision Tree & Predicator Nodes

Scatter plot node also has the feature to display the data in the decision tree form. Each tree has further has two sub trees the value of sub trees depends on the values of root node as shown in figure. The root node has the values between 36 to 100 further it sub trees has values 35 to 38 and 36 to 65. Thus this feature is really help to split the data into different parts according to the values and helpful in decision making. By using the decision tree structure no of partition and direction of partition easily can be detected.
V. CONCLUSIONS

The current study implemented the data mining techniques through the KNIME, a data mining tool. KNIME tool provides the different nodes like file reader node, parathion node, decision leaner tree, decision predictor node scorer, colour manager node all these nodes can work independent, but by using the output of a node as an input in other node one can implemented the data mining techniques like clustering, Association or some predicative models. KNIME also provide the functionality to represent the different cluster in the form of group which has different colours. In the current also implemented the K-means Algorithm through K-means Node which is used for clustering. This non-hierarchical method initially takes the number of components from the data equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Then it examines each component from data warehouse and assigns it to one of the clusters depending on the minimum distance. The centurion’s position is recalculated every time a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

ACKNOWLEDGMENT

I don’t have enough words to describe the profound gratitude and sense of indebtedness which I feel to express towards my supervisor MR VINOD KUMAR SRIVASTVA, Assistant Professor, Department of Computer Science & Engg, HARYANA COLLEGE OF TECHNOLOGY & MANAGEMENT for his invaluable guidance, persistent and useful suggestions, moral support and for making an environment conducive during the course of investigation reported in the present dissertation. Without his constant help and keen interest, it would have been difficult for me to sustain efforts for its completion. I am also grateful to Dr PK SURI, Chairperson CSE Department and my respected parents for all possible encouragement and inspiration from time to time given in this research.

REFERENCES

[3]. Michael R. Berthold; “KNIME – The Konstanz Information Miner”; SIGKDD Explorations; Volume 11, Issue 1
[4]. ALAA EL-HALEES” MINING STUDENTS DATA TO ANALYZE LEARNING BEHAVIOR: A CASE STUDY”.
[5]. Dominik Morent” Comprehensive PMML Preprocessing in KNIME”;