International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in; editorijoer@gmail.com

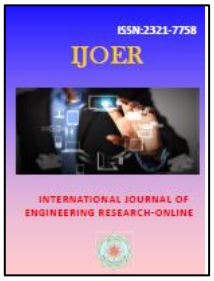
Vol.5., Issue.2, 2017
March-April

RESEARCH ARTICLE

ISSN: 2321-7758

REAL TIME COMPUTATION AND EVALUATION ON RAW DATA USING STATISTICAL TOOLS

RISHAB JASROTIA[1], NACHIKET NATEKAR[1], ARPIT JAISWAL[1], KEERTI KHARATMOL[2]
[1]B.E Student, [2]Assistant Professor
Department of Computer Engineering, KC College, Kopri, Thane
jasrotiarishab@gmail.com;nachinatekar@gmail.com;Arpitarpit39@gmail.com;Keertikc13@gmail.com



ABSTRACT
This paper introduces a approach which going to implement the evolutionary learning algorithms for Data Mining issues of various types such as classification, regression, etc. It includes different pre-processing methods allowing it perform complete analysis of raw data.

Keywords—Data Mining, Machine Learning, Classification, Regression

I. INTRODUCTION

Machine learning [1] and Data mining are the areas of research whose development is due to the new technologies in data analysis techniques, growth in database industries and the needs of market for extracting knowledge from large data sets. The machine learning has two approaches mainly symbolic approach and statistical approach.

Symbolic approaches: Inductive learning methods, such as decision trees, rules or logical representation.

Statistical approaches: Statistical or patter-matching methods, including k-nearest neighbor or instance-based learning, Bayesian classifiers and support vector machines.

Due to the major growth in the needs for techniques that are able to extract the important Knowledge from data instances, data mining (DM) and knowledge discovery in databases (KDD) the evolutionary tools are necessary for extracting such data.

Evolutionary Algorithms (EA) [2], [3], [4] consist the different various heuristics, which can solve the problem related to real-time computation. They may be use at different levels of abstraction, but they are working on whole populations of possible solution for given task.

Data mining is an pattern matching technique to find undetected relationships of datasets. Data mining (DM) often

involves the analysis of the data which is been stored in data warehouse. In this paper we are working over classification, regression and etc.

The goal of this paper is to create an application using the java technology, which implement the multiple evolutionary based methods for data mining problems. By using this application the effective solution will be found.

II. WHY USE DATA MINING?

Data mining (DM) is the way to extract the information from large database. There are two major reasons to use the data mining in rapidly needs. These are:

- Large amount of data and less information related to it.
- The need to extract the useful information from the datasets.

III. HISTORY OF DATA MINING

The term “Data mining” was introduced in the mid-90s, but it was the evolution with long history. The Data mining roots are linked to three

major Fields: classical statistics, artificial intelligence and machine learning.

A. Classical Statistics

Statistics are the main foundation of major technologies on which the data mining is designed, e.g. regression [5], cluster analysis [6], standard distribution and deviations, standard variance. All of these are used to study data and their relationships.

B. Artifical Intelligence

Artificial intelligence (AI) is built upon the heuristics as opposed to the statistics; the main concept of AI is to apply the human concepts to solve the statistical problems. Certain AI concepts which were adopted for query optimization modules use in Relational Database Managements System (RDBMS).

C. Machine Learning

Machine Learning (ML) is the combination of statistics and Artificial Intelligence. It can be considered an expansion of AI, because it blends Ai heuristics with the advance statistical analysis. Machine learning tries to let the computer based programs to learn about the data by learning, such that programs make different decision based on the qualities of the studied data, using statistics for concepts and adding more advanced technique.

IV. EXISTING SYSTEM

The existing data mining tools are such as Oracle, Microsoft data mining tools are mainly developed using the limited numbers of predefined algorithms.

Such types of tools are not capable to handle or predict the different outputs for a particular instance, such system are limited to the specific database or data types

A. Issues of existing system

1) Security issues, is an important issue with the data collection that is shared in current system.
2) Good user interface is not provided to ease of user understandability.
3) The issues of mining methodologies not implemented properly.
4) Data sources issues; the support to major data source is limited.
5) Dealing with large dataset is difficult.

V. PROPOSED SYSTEM

Now we have understood the existing system and have figured out that there are a lot of problems in it. So, what we propose is a solution of using the combination of multiple different clustering and classification algorithm for a better mining and prediction.

The proposed system has mainly divided into three modules; Data module, Educational module and Mining module.

A. Data module

This module consist set of tools that can be used to build, export or import the data sets. The fundamental approach of data module is to manipulate and transform the raw data so that we can get information content. This module will contain following sub modules:

1) Import data: This module will allow the user to import the data sets from different data types such as CSV, ARFF, XML and SQL database using the JDBC. The after import will be converted to uniform data type such as ".dat" extension file.
2) Export data: This module is opposite to previous one. It converts the data from ".dat" format to different data type's format.
3) Visualization: This module focus to represent or visualize the data.
4) Edit data: This module allows the user to edit the row sets from the data.
5) Data partition: This zone allows making the partition of data. It focus on k-fold cross validation.

B. Educational module

The educational module is an research module which can be used for educational purpose for students who wants to do experiment on data Mining (DM) algorithms.. This module is for learning the different parameters of algorithm. In this sense, educational module is simplified version of research tool, where only some relevant type of algorithm is available.

C. Mining Module

Many Mining tools have been developed in the last few years. Some of these are libraries that allow the implementation of new algorithms such as: Association rules, learning classifier system [7], etc.

This module is a User Interface (UI) that will allow the design of Mining experiments that can be used to solve various problems of classification, association and regression. This module will allow making configuration of algorithms attributes base on needs. The result will be reflected either in form of XML representation or java file. This java file can be run on any machine which has Java Virtual Machine (JVM) environment.

The Multiple algorithms can be used for experiment on the data set.

VI. CONCLUSION

In this paper, we have described a tool that provides the technique for analysis of Data mining problems using the evolutionary approach. This paper allows focusing on the analysis of new learning methods in comparison to existing system. This enables researchers apply machine learning algorithms in new way.

We have tried to show the different modules for the system and distinguished the three main parts: a module for data preparation, a module for mining purpose, and a module for educational purpose.

At the moment, we are developing the data management module and mining module with a new set of evolutionary algorithms.

REFERENCES

[1]. Goldberg, D.E. & Holland, J.H. Machine Learning (1988) 3: 95.

[2]. Rodrigo C. Barros, Christian V. Quevedo, Renata De Paris, Márcio P.Basgalupp, "Clustering Molecular Dynamics trajectories with a univariate estimation of distribution algorithm", Evolutionary Computation (CEC) 2015 IEEE Congress on, pp. 2058-2065, 2015.

[3]. Qasim Zeeshan Ahmed, Sajid Ahmed, Mohamed-Slim Alouini, Sonia Aïssa, "Minimizing the Symbol-Error-Rate for Amplify-and-Forward Relaying Systems Using Evolutionary Algorithms", Communications IEEE Transactions on, vol. 63, pp. 390-400, 2015, ISSN 0090-6778.

[4]. Ayangleima Laishram, Swati Vipsita, "Bi-clustering of gene expression microarray using coarse grained Parallel Genetic Algorithm(CgPGA) with migration", India Conference (INDICON) 2015 Annual IEEE, pp.1-6, 2015, ISSN 2325-9418.

[5]. Li-Li Wei, Chong-Zhao Han, "A Robust Method for Detecting Regression Change Points", Fuzzy Systems and Knowledge Discovery 2007. FSKD 2007. Fourth International Conference on, vol. 1, pp. 468-471, 2007.

[6]. Jiang-Hong Ma, Guo-Jun Wang, "A Case Study on the RCMD Method and Fuzzy C-Regression Models for Mining Regression Classes", Computer Science and Software Engineering 2008 International Conference on, vol. 1, pp. 915-918, 2008.

[7]. Xiangyang Li and Nong Ye, "A supervised clustering and classification algorithm for mining data with mixed variables," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol.36, no. 2, pp. 396-406, March 2006.