Journal of Electronic Research and Application

Review Article

Data Mining Techniques and its Uses in Different Fields: A Review Paper

Gaurav Dhawan

Department of Computer, Government Brijindra College, Faridkot, Punjab, India

Abstract: The paper introduced the data mining and issues related to it. Data mining is a technique by which we can extract useful knowledge from urge set of data. Data mining tasks used to perform various operations and used to solve various problems related to data mining. Data warehouse is the collection of different method and techniques used to extract useful information from raw data. Genetic algorithm is based on Darwin's theory in which low standard chromosomes are removed from the population due to their inability to survive the process of selection. The high standard chromosomes survive and are mixed by recombination to form more appropriate individuals. In this urge amount of data is used to predict future result by following several steps.

Keywords: *data mining; data warehouse; genetic algorithm; chromosomes*

Publication date: July 2018

Publication online: 30th July 2018

Corresponding Author: Gaurav Dhawan, dhawangaurav200@gmail.com

0 Introduction

Data mining is the major field in which we deal with extracting useful knowledge from urge raw data. Various issues related to data mining are discussed later. Uses of it in various fields are discussed at last and data mining with genetic algorithm is explained. Section I gives a brief introduction about the paper and Section II gives brief introduction about data mining. Section III describes the data mining tasks. Section IV provides the working of general-purpose computing on graphics processing units and CUDA. Section V describes the data mining issues and Section VI explains genetic algorithm in data mining. Section VII explains various data techniques and Section VIII explain uses of data mining in various fields. Section IX includes the conclusion of the paper that comes out with useful method used in data mining.

1 Data mining

Huge set of data is a data until we extract the information from it; the extracted information is so-called data mining. Obtaining useful information from huge amount of data is a part of data mining. It also includes data cleaning, data integration, data transformation, data mining, pattern evaluation, and data presentation. There are many applications in which data mining can be used, for example, market analysis in which we identify the best product, identify customer purchasing behavior and so on, fraud detection in which we detect fraud telephone calls, etc. Figure 1.

- **User interface:** It is the module of data mining system that helps the communication between user and the data mining system.
- **Data integration:** It is data preprocessing technique that merges the data from multiple different data source into coherent data store.
- **Data cleaning:** It is technique used for removing noise data and corrects the inconsistency in data.
- **Data selection:** It is a process in which analysis task is retrieved from database.
- **Cluster:** It is referred to group of similar kind of the objects.

2 Data mining tasks

Tasks mean kind of pattern that can be mined. There are two types of tasks performed by data mining:

- **2.1 Descriptive:** It deals with general properties of data, for example, class, mining of frequent patterns, clusters, etc.
- **2.2** classification is a process of finding a model that describes the data classes or concept while prediction uses to predict missing or unknown data.

3 Data mining issues

There are three major issues in data mining as mention below:

- **3.1 Mining methodology and user interaction:** Extracting different kind of knowledge from huge database and handling with incomplete data.
- **3.2 Performance issues:** In the efficiency and scalability of data are major issues.
- **3.3 Diverse data types:** Handling of relational and complex type of data and mining of data from heterogeneous database.

4 Data warehousing

It is a collection of different methods, techniques, and tools used to conduct data analyses which help in performing decision-making and improving information resources by knowledge workers. It also involves data cleaning, data integration, etc., type of tasks performed in data warehouse. There are two types of approaches in data warehouse, that is:

- **4.1 Query-driven approach or traditional approach:** It integrates heterogeneous database. Major drawback of this approach is complexity and expensiveness.
- **4.2 Updated driven approach:** Information from multiple heterogeneous data is collected in advance and store in it. It major advantage is its high performance.

5 Data mining using genetic algorithm

Genetic algorithm is the type of evolutionary algorithm that uses techniques inspired by nature such as inheritance, mutation, selection, and crossover[1,2]. A group of chromosomes in population is created randomly[3]. The chromosomes in the population go for selection process. The evaluation function is given by the programmer and gives the chromosomes fitness based on how well they perform at the given task[4]. Two chromosomes are then selected based on their fitness, the higher the fitness, higher, and the chance of being selected[5]. These individuals then perform crossover to create offspring, after which the offspring is mutated randomly[6]. This will go on until a suitable solution has been found or a certain number of generations have passed or according to the programmer needs[7-10]. Urge number of data is used to predict future value. Hence, we can use genetic algorithm to predict future value using data stored in database[11].

6 Data mining techniques

Data mining means collecting relevant information from unstructured data. The purpose of a predictive

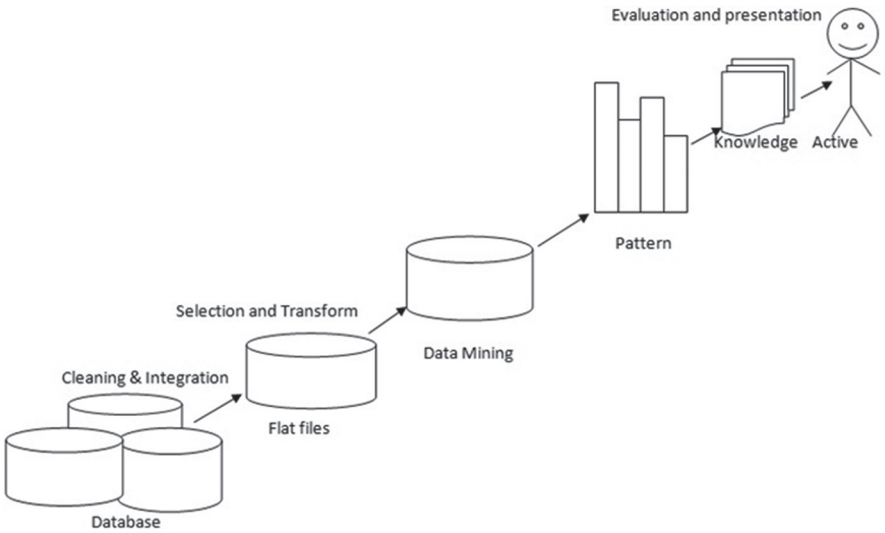

Figure 1. Data mining

model is to allow the data miner to predict an unknown or often future value of a specific variable [11,12].

- **6.1 Classification**: When the desired output for given input is known or supervised learning. Decision tree, neural network, and classification rule etc, (IF- Then) are used to achieve these goals.
- **6.2 Regression:** It is used to map a data item to a real-valued prediction variable. In other words, it can be adapted for prediction. In this target, values are known. We can predict the child behavior knowing on family history.
- **6.3 Time series analysis:** It is the process of using statistical techniques to model and explain a time-dependent series of data points. Time series forecasting is one of the method used in model to generate predictions for future events based on known past events [13].
- **6.4 Prediction:** It is one of a data mining techniques that discover relationship between dependent and independent variables.
- **6.5 Clustering:** It is a collection of similar data object. Dissimilar object is another cluster. It is wayfinding similarities between data according to their characteristic, for example, image processing, pattern recognition, and city planning [14].
- **6.6 Summarization:** Summarization is abstraction of data. It is a set of relevant task and gives an overview of data. For example, long-distance race can be summarized total minutes, seconds, and height.
- **6.7 Association rule**: It is the most popular data mining techniques and fined most frequent item set.
- **6.8 Sequence discovery:** Uncovers relationships among data. It is a set of object each associated with its own timeline of events, for example, scientific experiment, natural disaster, and analysis of DNA sequence.

7 Data mining used in various fields

Various fields adapted data mining technologies due to fast access of data and valuable information from a large amount of data. Data mining application area includes marketing, telecommunication, fraud detection, finance, and education sector, medical, and so on. Some of the main applications listed below:

- **7.1 Education sector:** Data mining in education sector is new emerging field called "education data mining." Using these term, we can know the performance of student, dropout student, and student behavior, which subject selected in the course. Data mining in higher education is a recent research field and use student's data to analyze their learning behavior to predict the results [12].
- **7.2 Market basket analysis:** This method is based on shopping database. Is this market basket analysis is finding the products that customers frequently purchase together. The stores can use this information to predict the item which are more sold and making them more visible and accessible for customers at the time of shopping [13].
- **7.3 Telecommunication:** The telecommunications field implement data mining technology due to telecommunication industry have the large amounts of data and have a very large customer, and rapidly changing and highly competitive environment. Telecommunication companies' use data mining technique to improve their marketing efforts, detection of fraud, and better management of telecommunication networks [14].
- **7.4 Cloud computing:** In cloud computing, data mining will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage. It uses the internet services that rely on clouds of servers to handle tasks [15].
- **7.5 Bioinformatics:** It generates a large amount of biological data. The importance of this new field of inquiry will grow as we continue to generate and integrate large quantities of genomic, proteomic, and other data [16].
- **7.6 Banking and finance:** Data mining has been used extensively in the banking and financial markets. In the banking field, data mining is used to predict credit card fraud, to estimate risk, and to analyze the trend and profitability. In the financial markets, data mining technique such as neural networks used in stock forecasting, price prediction, and so on [17].
- **7.7 Agriculture:** Data mining than emerging in agriculture field for crop yield analysis with respect to four parameters, namely year, rainfall, production, and area of sowing. Yield prediction is a very important agricultural problem that remains to be solved based on the available data. The yield prediction problem can be solved by employing data mining techniques such as K means, K nearest neighbor, artificial neural network, and support vector machine [11].

- **7.8 Earthquake prediction:** Predict the earthquake from the satellite maps. Earthquake is the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault in the interior. There are two basic categories of earthquake predictions: Forecasts (months to years in advance) and short-term predictions (hours or days in advance) [12].

8 Conclusion and future work

This paper provides a general idea of data mining, data techniques, and data mining in various fields. The main objectives of data mining techniques are to discover the knowledge from active data. These applications use classification, prediction, clustering, association techniques, and so on. Genetic algorithm is one of the best evolutionary algorithms used in data mining. In future work, we review other various evolutionary algorithms and its significance's in data mining and implementation of other evolutionary algorithms in data mining.

References

[1] Molga M, Smutnicki C. Test Functions for Optimization Needs; 2005. Available from: http://www.zsd.ict.pwr.wroc.pl/files/docs/functions.pdf. [Last retrieved on 2013 Jun 05]

[2] Ghoseiri K, Ghannadpour SF. Multi-objective vehicle routing problem with time windows using goal programming and genetic algorithm. Appl Soft Comput 2010;10:1096-107.

[3] Yang S, Cheng H, Wang F. Genetic algorithms with immigrants and memory schemes for dynamic shortest path routing problems in mobile ad hoc networks. Syst Man Cybern 2010;40:52-63.

[4] Aguilar OA, Huegel JC. Inverse kinematics solution for robotic manipulators using a cuda-based parallel genetic algorithm. In: Advances in Artificial Intelligence. Berlin: Springer; 2011. p. 490-503.

[5] Debattisti S, Marlat N, Mussi L, Cagnoni S. Implementation of a Simple Genetic Algorithm within the Cuda Architecture. In: The Genetic and Evolutionary Computation Conference; 2009.

[6] Sinha SS, Singh S. In: Speedup Genetic Algorithm Using Gpgpu. Vol. 5. IEEE International Conference on Communication Systems and Network Technologies, Gwalior, India; 2015. p. 138.

[7] Jaros J. Multi-Gpu Island-Based Genetic Algorithm for Solving the Knapsack Problem. In: Evolutionary Computation (CEC), 2012 IEEE Congress on. IEEE; 2012. p. 1-8.

[8] Pospichal P, Jaros J, Schwarz J. Parallel Genetic Algorithm on the Cuda Architecture. In: Applications of Evolutionary Computation. Berlin: Springer; 2010. p. 442-51.

[9] Rodriguez-Maya NE, Graff M, Flores JJ. Performance Classification of Genetic Algorithms on Continuous Optimization Problems. In: Nature-Inspired Computation and Machine Learning. Berlin: Springer; 2014. p. 1-12.

[10] Yue M, Hu T, Hu T, Guo X. The Research of Parameters of Genetic Algorithm and Comparison with Particle Swarm Optimization and Shuffled Frog-Leaping Algorithm. In: Power Electronics and Intelligent Transportation System (PEITS), 2009 2nd International Conference on. IEEE; 2009. p. 77-80.

[11] Hasan M. Genetic algorithm and its application to big data analysis. Int J Sci Eng Res 2014;5:2229-5518.

[12] Ahlawat A, Suri B. Improving Classification in Data Mining using Hybrid algorithm. In: 978-1-4673-6984-8/16/$31.00 © IEEE; 2016.

[13] Agarwal S, Pandey GN, Tiwari MD. Data mining in education: Data classification and decision tree approach. Int J e-Educ e-Bus e-Manag e-Learn 2012;2:2.

[14] Sharma SP. Use of data mining in various field: A survey paper. IOSR J Comput Eng 2014;16:18-21.

[15] Galván P. Educational Evaluation and Prediction of School Performance through Data Mining and Genetic Algorithms. In: FTC 2016 - Future Technologies Conference 2016. San Francisco, United States; 2016.

[16] Ramageri BM. Data mining techniques and applications. Indian J Comput Sci Eng 2010;1:301-5.

[17] Koukouvinos C. Genetic Algorithm and Data Mining Techniques for Design Selection in Databases. In: 2013 International Conference on Availability, Reliability and Security, ARES 2013.