Jan – Mar © 2019 Transactions; 6(21): 1105-1109

International Journal of Research and Applications

ISSN (online): 2349-0020 ISSN (print): 2394-4544 http://www.ijraonline.com/

Case Study

A Study on data mining methods and tools towards IOT and finance

Bolukonda Prashanth

Corresponding Author:
prashanth19bolukonda@gmail.com

DOI:
http://dx.doi.org/10.17812/IJRA.6.21(2)2019

Manuscript:
Received: 22nd Jan, 2019
Accepted: 25th Feb, 2019
Published: 17th Mar, 2019

Publisher:
Global Science Publishing Group, USA
http://www.globalsciencepg.org/

ABSTRACT

Data mining has as goal to extract knowledge from large databases. To extract this knowledge, a database may be considered as a large search space, and a mining algorithm as a search strategy. In general, a search space consists of an enormous number of elements, making an exhaustive search infeasible. Therefore, efficient search strategies are of vital importance. Search strategies based on genetic-based algorithms have been applied successfully in a wide range of applications. A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. In this paper, we discuss the suitability of genetic-based algorithms for data mining. We discuss the various application areas where genetic Algorithm plays evolutionary role with data mining technique and explain them in details.

Keywords: Genetic Algorithm, data mining, IOT.

Associate Professor, Dept., of Computer Science and Engineering,
Vaagdevi College of Engineering (Autonomous), Approved by AICTE,
Bollikunta, Warangal Urban (Dist.), Telangana State, India- 506005.

IJRA - Year of 2019 Transactions:
Month: January - March
Volume – 6, Issue – 21, Page No's:1105-1109
Subject Stream: Computers

Paper Communication: Author Direct

Paper Reference Id: IJRA-2019: 6(21)1105-1109

DOI: 10.17812/IJRA.6.21(2)2019

International Journal of Research and Applications
Jan - Mar © 2019 Transactions 6(21): 1105-1109

eISSN : 2349 – 0020
pISSN : 2394 – 4544
www.ijraonline.com

COMPUTERS CASE STUDY

A Study on data mining methods and tools towards IOT and finance

Bolukonda Prashanth
Associate Professor, Dept., of Computer Science and Engineering,
Vaagdevi College of Engineering(Autonomous), Approved by AICTE,
Bollikunta, Warangal Urban (Dist.), Telangana State, India- 506005.
prashanth19bolukonda@gmail.com

ABSTRACT

Data mining has as goal to extract knowledge from large databases. To extract this knowledge, a database may be considered as a large search space, and a mining algorithm as a search strategy. In general, a search space consists of an enormous number of elements, making an exhaustive search infeasible. Therefore, efficient search strategies are of vital importance. Search strategies based on genetic-based algorithms have been applied successfully in a wide range of applications. A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. In this paper, we discuss the suitability of genetic-based algorithms for data mining. We discuss the various application areas where genetic Algorithm plays evolutionary role with data mining technique and explain them in details.

Keywords: Genetic Algorithm, data mining, IOT.

1. INTRODUCTION

Data mining towards finance and accounting

Data mining tools become important in finance and accounting. Their classification and prediction abilities enable them to be used for the purposes of bankruptcy prediction, going concern status and financial distress prediction, management fraud detection, credit risk estimation, and corporate performance prediction. This study aims to provide a state-of-the-art review of the relative literature and to indicate relevant research opportunities.

Data Mining (DM) is a well honored field of Computer Science. It emerged in late 80's by using concepts and methods from the fields of Artificial Intelligence, Pattern Recognition, Database Systems and Statistics, DM aims to discover valid, complex and not obvious hidden information from large amounts of data. For this reason, another equivalent term for DM is Knowledge Discovery in Databases (KDD), which is equally often met in the literature.

Financial data are collected by many organizations like banks, stock exchange authorities, taxation authorities, big accounting and auditor offices specialized data bases, etc. and in some cases are publicly available. The application of DM techniques on financial data can contribute to the solution of classification and prediction problems and facilitate the decision making process. Typical examples of financial classification problems are corporate bankruptcy, credit risk estimation, going concern reporting, financial distress and corporate performance prediction.

Research on DM in finance and accounting and the application of its outcomes is a relatively new research field. The aim of the present study is to provide a state-of-the-art review about current research efforts on applying DM in finance and accounting. This review introduces the reader to specific topics concerning research objectives and methods employed.

In particular this study tries to address the following questions:

- What are the specific financial application areas to which DM methods have been applied?

- What DM methods have been applied and to what extent. Do these methods outperform previous more traditional methods?
- Over what kind of data do the methods operate? Are sample sizes satisfactory large? What are the applied feature selection methods?
- What are the relative performance metrics considerations?

Such a study helps the researcher to avoid overlapping efforts and benchmark his/her practices against new developments. Another aim of this study is to indicate fertile areas for further research work in the area.

2. THE LITERATURE REVIEW

For finding the studies concerning the application of DM techniques in finance and accounting we investigated the journals of four publishing houses: Elsevier, Emerald, Kluwer and Wiley. Relative articles have been found in the journals:

- Asia Pacific Financial Markets.
- Decision Support Systems,
- European Journal of Operational Research,
- Expert Systems with Applications,
- Intelligent Systems in Accounting, Finance & Management,
- International Journal of Accounting Information Systems,
- Journal of Forecasting,
- Knowledge Based Systems,
- Management Decision,
- Managerial Auditing Journal,
- Managerial Finance,
- Neural Networks, and
- Omega the International Journal of Management Science.

3. DATA MINING METHODS

The term Data Mining methods stands for a large number of algorithms, models and techniques derived from the osmosis of statistics, machine learning, databases and visualization. Several of these methods have been applied for examining financial data. Popular DM methods that will be mentioned in this study are Neural Networks, Genetic Algorithms, Decision Trees, Rough Set Theory, Case Base Reasoning and Mathematical Programming.

Neural Networks

Neural Networks (NN) is a mature technology with established theory and rec- organized applications areas. A NN consist of a number of neurons, i.e. interconnected processing units. Associated with each connection is a numerical value called "weight". Each neuron receives signals from connected neurons. If the combined input signal strength exceeds a threshold, then the neuron fires. The input value is transformed by the transfer function of the neuron.

The neurons are arranged into layers. A layered network consists of at least an input (first) and an output (last) layer. Between the input and output layer there may exist one or more hidden layers. Different kinds of NNs have a different number of layers. Self-organizing maps (SOM) have only an input and an output layer, whereas a back propagation NN has additionally one or more hidden layers.

After the network architecture is defined, the network must be trained. In back propagation networks a pattern is applied to the input layer and a final out- put is calculated at the output layer. The output is compared with the desired result and the errors are propagated backwards in the NN by tuning the weights of the connections. This process iterates until an acceptable error rate is reached. The back propagation NNs have become popular for prediction and classification problems.

SOM is a clustering and visualization method of unsupervised learning. For each input vector, only one output neuron will be activated. The winner's weight vector is updated to correspond with the input vectors. Thus, similar inputs will be mapped to the same or neighboring output neurons forming clusters. Two commonly used SOM topologies are the rectangular lattice, where each neuron has four neighbors and the hexagonal lattice where each neuron has six neighbors.

An important disadvantage of NNs is that they act as black boxes as it is difficult to humans to interpret the way NNs reach their decisions. However, algorithms have been proposed to extract comprehendible rules from NNs. Another criticism on NNs is that a number of parameters like the network topology must be defined empirically. It seems that NNs attract the interest

of most researchers in the area of our concern. Their structure and working principles enable them to deal with problems where an efficient algorithm based solution is not applicable. Since they learn from examples and generalize to new observations they can classify previously unseen patterns. They have the ability to deal with incomplete, ambiguous and noisy data. Unlike traditional statistical techniques they do not assume a priori about the data distribution properties, neither have they assumed independent input variables.

Genetic Algorithms

Genetic Algorithms (GA) apply ideas from the natural evolution where the fittest individuals survive. Rules concerning a problem are encoded as a set of strings each of which is composed of bits. These strings form a population. GA allows the strings with the highest fitness value to survive and proliferate renewing the population.

A chromosome is a particular string representing a point in the solution space. Population is a set of chromosomes. After the random creation of the initial population each chromosome is evaluated using a user-defined fitness function. The role of the fitness function is to evaluate the performance of the chromosome.

Three operators are applied to chromosomes.

- **Reproduction,** where the individuals self-proliferate by replicating themselves with a probability analogous to their fitness value.
- **Crossover,** where two chromosomes mutually exchange some bits creating new chromosomes
- **Mutation,** which operates on a single chromosome by changing one or more bits. The probability of mutation is very low.

Decision Trees

Decision Trees is a classification and prediction method, which successively divides observations into mutually exclusive subgroups. The method searches for the attribute that best separates the samples into individual classes. Subgroups are successively divided until the subgroups are too small or no significant statistical difference exists between candidate subsets. If the decision tree be-comes too large it is finally pruned.

Rough Set Theory

Rough Set Theory (RST) was introduced by Pawlak (1982). RST extents set theory with the notion of an element's possible membership in a set. Given a class C, the lower approximation of C consists of the samples that certainly be- long to C. The upper approximation of C consists of the samples that cannot be defined as not belonging to C. RST may be used to describe dependencies be- tween attributes, to evaluate significance of attributes, to deal with inconsistent data and to handle uncertainty (Dimitras etc. al.1999).

Case Base Reasoning

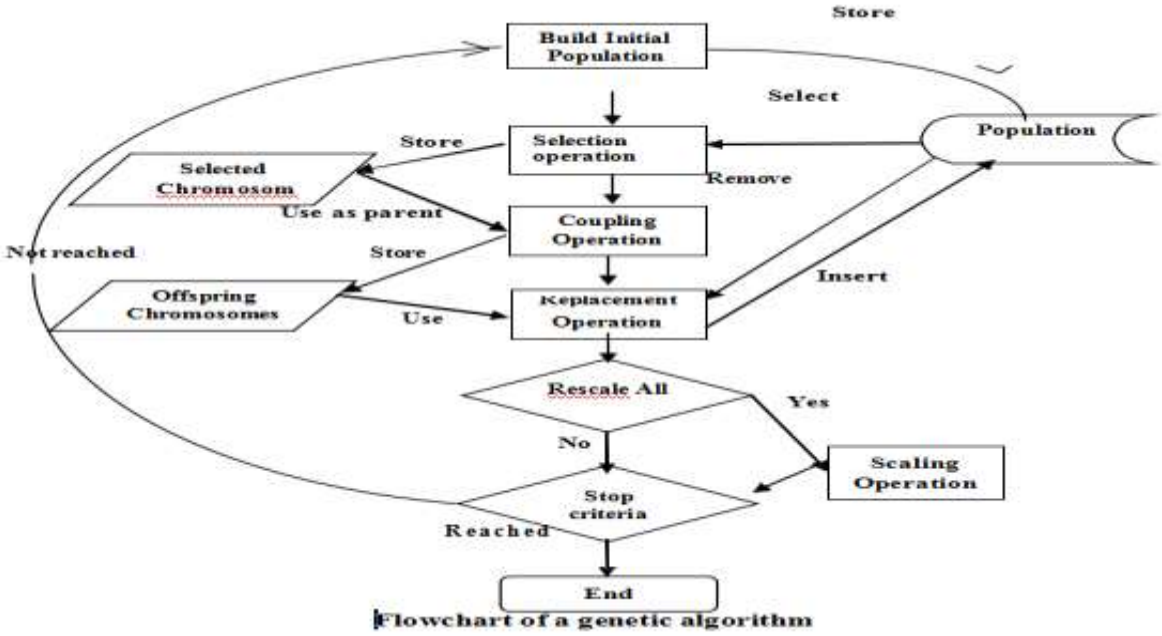
Case Base Reasoning (CBR) is a reasoning problem solving method. For solving a problem, CBR tries to retrieve a similar case from a case base. Key issues in CBR are the similarity measure and the retrieval of a similar case. Popular matching techniques are k-nearest neighbor (k-NN), inductive learning and knowledge guide. In its simplest version, k-NN assesses the similarity of two cases by calculating their Euclidean distance. This approach assumes that all features are equally relevant. Since this is not always the case, improved algorithms introducing weighted features have been proposed.

GENETIC ALGORITHM

A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. Genetic algorithms find application in bioinformatics, phylogenetic, computational science, engineering, economics, chemistry, manufacturing, mathematics, physics and other fields.

It has been shown to be efficient and powerful through many data mining applications that use optimization and classification. GAs can rapidly locate good solutions, in data mining even for difficult search spaces. GAs are used in various fields of Data mining to get the optimized solutions for the better performance of the data that are required in decision making and process the accurate result. There is also a greater scope of GA in data mining in future application to stimulate the data mining concepts.

Genetic algorithms are widely applicable to classification by means of inductive learning. GAs also provides a practical method for optimization of data preparation and data transformation steps. Hence GA can be used in a real analysis system to get the better solution.



Flowchart of a genetic algorithm

4. APPLICATION OF GENETIC ALGORITHM

a) **Genetic Algorithms is an effective tool to use in data mining and pattern recognition**.

There are two different methods to applying

GA in pattern recognition:

1. Use GA as a classifier directly in computation.
2. Use a GA to optimize the results i.e. as an optimizer to arrange the parameters in other classifiers.
Most applications of GAs in pattern recognition optimize some parameters in the classification process [4].

GAs has been applied to find an optimal set of feature weights that improve classification accuracy. First, a traditional feature extraction method such as Principal Component Analysis

(PCA) is applied, and then a classifier such as **k-NN** (**Nearest Neighbor Algorithm**) is used to calculate the fitness function for GA [6]. Combination of classifiers is another area that GAs have been used to optimize. GA is also used in selecting the prototypes in the case-based classification.

According to us second method of genetic algorithm to optimize the result from the dataset is more effective to compute the accurate values of observations of data by applying data mining techniques.

b) **Genetic Algorithm has a wide scope in business.**

There are large amount of data that has to be filtered to process the results for optimizing the business profits by using various data mining techniques.

There are many domains in business to which they can be applied:
Optimization: Give a business problem with certain variables and a well-defined definition of profit, a genetic algorithm can be used to automatically determine the optimal value for the variables that optimize the profit [1].
Prediction: Genetic algorithms have been used as Meta level operators that are used to help optimize other data mining algorithms. For instance, they have been used to find the optimal association rules in market-analysis.
Simulation: Sometimes a specific business problem is not well defined in terms what the profit is or whether one solution is better than the other. The business person instead just has large number of entities that they would like to simulate via simple interaction rules overtime.

5. CONCLUSION

The first and most important point is that genetic algorithms are intrinsically parallel. Most other

algorithms are serial and can only explore the solution space to a problem in one direction at a time since GAs have multiple offspring, they can explore the solution space in multiple directions at once. If one path turns out to be a dead end, they can easily eliminate it and continue work on more promising avenues, giving them a greater chance each run of finding the optimal solution. Genetic algorithms provide a comprehensive search methodology for machine learning and optimization. DM techniques have classification and prediction capabilities which can facilitate the decision making process in financial problems. The financial and prediction tasks in the collected literature address the topics of bankruptcy prediction, credit risk estimation, going concern reporting, financial distress, corporate performance prediction and management fraud. Bankruptcy prediction seems to be the most popular application area. The data mining methods employed in the collected literature include Neural Networks, Genetic Algorithms, Decision Trees, Rough Set Theory, Case Base Reasoning and Mathematical Programming. Most of the researches seem to prefer the Neural Network model.

REFERENCES

1) B. Back, J. Toivonen, H. Vanhatanta and A. Visa: "Comparing Numerical Data and Text Information from Annual Reports Using Self-organizing Maps", International Journal of Accounting Information Systems, 2(4): 249-269, (2001).
2) M.J. Beynon and M.J. Peel: "Variable Precision Rough Set Theory and Data Discretization: an Application to Corporate Failure Prediction", Omega the International Journal of Management Science, 29(6): 561-576, (2001).
3) T.G. Calderon and J.J. Cheh: "A Roadmap for Future Neural Networks Research in Auditing and Risk Assessment", International Journal of Accounting Information Systems, 3(4): 203-236, (2002).
4) A.I. Dimitras, R. Slowinski, R. Susmaga and C. Zopounidis: "Business Failure Prediction using Rough Sets", European Journal of Operational Research, 114(2): 263-280, (1998).
5) I.A.M. Fraser, D.J. Hatherly, and K.Z. Lin: "An empirical investigation of the use of analytical review by external auditors", The British Accounting Review, 29(1): 35-47, (1997).
6) H. Konno and H. Kobayashi: "Failure Discrimination and Rating of Enterprises by Semi-Definite Programming", Asia-Pacific Financial Markets, 7(3): 261-273, (2000).
7) Yeshwanth Rao Bhandayker, "Artificial Intelligence and Big Data for Computer Cyber Security Systems" in "Journal of Advances in Science and Technology", Vol. 12, Issue No. 24, November-2016 [ISSN: 2230-9659].
8) Sugandhi Maheshwaram, "A Comprehensive Review on the Implementation of Big Data Solutions" in "International Journal of Information Technology and Management", Vol. XI, Issue No. XVII, Nov-2016 [ISSN: 2249-4510].
9) Sugandhi Maheshwaram, "An Overview of Open Research Issues in Big Data Analytics" in "Journal of Advances in Science and Technology", Vol. 14, Issue No. 2, September-2017 [ISSN: 2230-9659].
10) Yeshwanth Rao Bhandayker, "Security Mechanisms for Providing Security to the Network" in "International Journal of Information Technology and Management", Vol. 12, Issue No. 1, Feb-2017, [ISSN: 2249-4510].
11) Yeshwanth Rao Bhandayker, "A Study on the Research Challenges and Trends of Cloud Computing" in "RESEARCH REVIEW International Journal of Multidisciplinary", Volume-04, Issue-02, and February-2019 [ISSN: 2455-3085].
12) Sriramoju Ajay Babu, Dr. S. Shoban Babu, "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications", Volume 1, Issue 1, and Jan-Mar 2014 [ISSN: 2349-0020].
13) Dr. Shoban Babu Sriramoju, Ramesh Gadde, "A Ranking Model Framework for Multiple Vertical Search Domains" in "International Journal of Research and Applications" Vol 1, Issue 1, Jan-Mar 2014 [ISSN: 2349-0020].
14) Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju, "Risk-Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management", Volume VI, Issue I, Feb 2014 [ISSN : 2249-4510].