Academic Guru Publication House

Advances in Cloud Computing Security
Techniques and Applications
Volume 1
Year: 2021

AG PH Books

A Survey on Intelligent Data Analysis: Issues and Challenges

Mousami V. Munot[1*], Kaustubh V. Sakhare[2]

[1]*Associate Professor, Department of Electronics & Telecommunication Engg., SCTR's Pune Institute of Computer Technology (PICT), Pune*

[2]*Technical Specialist, Lear Corporation, Pune*

Abstract

(IDA) that is Intelligent data analysis is a new topic that combines several disciplines, particularly AI as well as statistics, to analyse data sets automatically or semiautomatically in a variety of real-world applications. All three of these areas are mutually beneficial: Several statistical procedures depend on computers, especially for big data sets, yet computational power alone cannot replace statistical expertise. There has been a rise in an intelligent data analysis system. It is goal of such a work to address a broad variety of issues that might arise when analysing data, as well as to provide solutions. A real-world instance of such a risk assessment of the level crossing data is used to analyse a few of such issues and ideas.

Keywords: Data analysis, data mining, risk assessment of level crossing, rule extraction, neural networks, rule induction

1. INTRODUCTION

It is the activity of utilising AI as well as machine learning to evaluate and turn large datasets into an intelligent data insight that can subsequently be utilised to enhance services as well as investments that is known as data intelligence. Better business processes may be developed via the use of data

* ISBN No. 978-81-955340-6-7

the intelligence tools as well as strategies that assist decision makers well comprehend the data they've acquired.

There are five primary components to the data-driven intelligence process: descriptive data, prescriptive information, diagnostic information, as well as predictive information. These fields are concerned with figuring out how to make sense of data, coming up with new ideas, resolving problems, and looking back at the past to make predictions about the future. Cybersecurity, banking, health, as well as insurance, as well as law enforcement, are some of the most pressing fields in need of data intelligence. Using intelligent data capture technologies, printed documents or photographs may be transformed into useful data in various fields.

Business intelligence relies on the use of intelligent data as a foundational component. As a result of intelligent data processing, big datasets can be restructured into the valuable information which is relevant to the business performance, allowing organisations to understand patterns, make the informed choices as well as adapt to the new information; as well as advanced analytics can be incorporated to improve visualisations of prescriptive as well as predictive analytics in order to better understand the data as well as make better decisions.

It is a multidisciplinary topic of research that focuses on the extraction of usable information from data using methods from a wide range of domains, including AI which is Artificial Intelligence, elevated performance computing, pattern recognition, as well as statistics. Firms like, Strategic Data Intelligence, Data Visualization Intelligence and Global Data Intelligence provide platforms as well as solutions for data intelligence. The Data intelligence companies like these included.

Increased demand and supply for more advanced IDA approaches have been spurred by the explosive growth in the amount of real-time data generated by online, multimedia, as well as electronic commerce activities. The fundamental notion of analysing enormous volumes of data with detailed descriptions is both attractive and straightforward, but it is a substantial challenge and difficulty to implement in practise. To make the most of data acquired from certain huge and complicated sources, there has to be a plan in place.

In other words, IDA extract value from the data by finding out rules as well as knowledge from it. Although it's impossible to count the precise no. of IDA methods, their evolving patterns, which include (1) algorithm principle, (2) dataset size, (3) dataset type may be summarised.

2. ALGORITHM PRINCIPLE

The evolution of IDA's algorithm concept has showed a progression from basic to sophisticated. On the basis of probability theory or even Euclidean distance dependent similarity theory, earlier IDA algorithms were developed IDA principles became increasingly complicated over time when computational intelligence was included.

2.1. Probability Based Algorithm

The IDA methods depending on the probability theory are often used for classification and grouping because of the property of probability theory altogether. Prior probability as well as posteriori probability are used in the Naive Bayes Classifier (NBC) to categorise sample data. Classification is performed using C4.S, which calculates the sampling data's entropy gain, whereas clustering is carried out using that is Expectation Maximization (EM), which seeks to find the parameters' greatest probability estimates. Because they are easy to implement as well as perform well, IDA algorithms that is probability-based have become a popular choice.

[2] use the auxiliary feature approach to perform a 2nd feature selection following NBC to improve the accuracy of NBC with in large text categorization.

[3] combine the NBC with the Decision Tree in order to improve classification accuracy as well as reduce the phenomena of over-fitting.

[4] By using stochastic processes in feature selection step of NBC's, (RSNB) that is Randomly Selected Naive Bayes method avoids the local optimum difficulties that plague classic NBC.

[5] During plant monitoring, use EM towards the challenge of detecting change points for the multivariate data.

[6] To perform fault diagnosis based on data, they suggest a hybrid EM technique depending on forward and backward Kalman filtering.

[7] Boolean factor analysis based on High-dimensional may be done using two new EM methods.

2.2. Euclidean Distance Based Algorithm

Inside the context of the n-dimensional dataset, Euclidean distance between various components may be used to measure the similarity in between them in context of dataset. Euclidean distance IDA techniques that focus on finding cluster centres by minimising total number of mean-square errors, such as k-Means as well as (k-NN) which is k-nearest-neighbour algorithms, are also popular choices. With SVM model, the data is represented as points in the space, which are then remapped to the higher-dimensional space. This creates a gap as large as possible between the data points in each category, resulting in a distinct separation between them.

[8] Breast cancer tumours may be diagnosed using a combination of k-means clustering and a support vector machine (SVM).

[9] Reduce sensitivity to an initial cluster centre while increasing the capacity to cope with scattered data by modifying k-Means technique with just an evolutionary approach, [10] The repetitious training for the continuous input conditions may be eliminated by using a quick k-Means algorithm to the graphic processing.

3. DATA ANALYSIS TASKS AND TECHNIQUES

For example, a user's purpose may be to characterise the entire data set or to construct linkages between the subsets of the patterns within data set [11]. Purpose of Predictive modelling is to create predictions depending over the data's basic properties. There are various preset classes and real-valued prediction variables that may be used to model data. Predictive modelling may be performed using whatever supervised machine learning technique which builds a model based on past or current data. The model is taught how to correctly forecast the future by being provided a list of previously established facts with right replies. A few of the methodologies used to map discrete-valued sets of variables include, decision trees, even neural networks, K-nearest neighbour classifier, Bayesian classifiers, reasoning based on case, genetic programming, fuzzy sets, as well as rough sets. There are a number of strategies for mappings regular-valued target variables using regression, even neural networks, including radial basis functions. Clustering is a technique used to create hierarchies of events by grouping together those having similar features. It is possible to accomplish clustering using any of the unsupervised machine learning technique that does not have a preconceived set of data categories. There are certain pre-existing facts that model is given, through it produces categories of the data with comparable features. Methods for clustering include partitioning, even hierarchical as well as density-depend algorithms; including model-based algorithms. [12].

Using link analysis, one may discover the intrinsic connections between data points. This objective is accomplished by the identification of associations, the discovery of sequential patterns, and related activities involving the discovery of temporal sequences [11]. By anticipating the correlation of elements that might otherwise be obscure, such activities reveal samples as well as patterns. Counting all potential combinations of objects is the basis for link analysis approaches. Apriori and its variants are among the most often used algorithms. [13].

4. CHALLENGES FACING THE IDA IN BIG DATA ENVIRONMENT

People's need to get more out of their data has been reached historic heights in the big data era, making it difficult for IDA to keep up. There are four ways wherein big data environment presents problems to the IDA, that may be summarised as follows: (1) data management, (2) data collecting, (3) data analysis, and (4) application pattern are all aspects of the big data process.

a. Big Data Management

Massive data management technologies, such as (HDFS) that is Hadoop Distributed File System [14], are already available and quite mature. A good large data management system, on the other hand, does more than just store information correctly. There are three key issues in the big data management: managing life cycle of data, securing data, and managing costs.

- **Data Life Cycle Management**

The most difficult part of managing the life cycle of data is determining how long a piece of data should be kept in storage. The conventional wisdom is that data life cycle concludes with the completion of the data analysis. Despite this, data life cycle management is indeed no more a straightforward matter in the context of big data. Users' perspectives vary even while working with the same dataset, resulting in varying amounts of value gleaned from the information. Data lifecycles are also varied because of this. A medical record of patient , for instance, comes to an end whenever the patient is well enough to be discharged. Medical records, on the other hand, may be an invaluable source of information for a clinician interested in learning about a history of allergies of family of the Patient. In contrast, for just an epidemiologist looking into a specific pandemic, large medical records integration is essential and useful. Case studies like this one demonstrate how diverse data life cycle management may result in varied data value extractions.

b. Data Security Management

Another issue that the IDA must deal with is the complete lifecycle of data security management. Individuals in the big data environment are indeed worried about their data's privacy as well as security. Personal, corporate, as well as national secrets may all be compromised at any point in the data lifecycle with such an inclusive environment. Data encryption might protect data to a certain degree, but it can also slow down data processing if encryption is too complicated. Data fuzzification, in addition to encryption, is another option for data security, albeit it has the potential to damage data. For having data security maintenance, it is the critical to optimise the efficiency of data security IDA.

c. Cost Management

Another issue that the IDA must deal with is how much money it spends on operations. The expense of strengthening the IDA's performance should be kept in check in order to ensure sustainable development. Dispersed data analysis, for instance, speeds up data processing whereas increasing the cost of hardware as well as network transmission; implementing various life cycle management strengthens value extracted from the data; complicated data encryption strengthens data security during increasing the computation complexity; and a variety of other examples. In order to achieve a balance between value as well as cost, the primary goal of the cost management is being to reduce both the explicit but also hidden costs.

d. Data Collection

In additional increase in size, data with in the big data environment have demonstrated qualities such as multi-source as well as heterogeneous. It is difficult for the IDA to gather data from several diverse sources and combine or preprocess such a large amount of heterogeneous data.

- **Fusion of Multi-source Heterogeneous**

Data In big data environment, data sources can be ubiquitous, which gives rise to the data heterogeneousness. Apart from traditional numerical data, data also include text, images, sound, and other electrical signals. The key point is to develop a integrate analysis algorithm in a novel big data analysis framework to transform all the multi-mode heterogeneous data into a uniform format that can be dealt with by IDA.

- **Pre-processing of Messy Data**

The noise and redundancy contained in raw data may greatly influence IDA's speed, accuracy, and robustness [15]. The preprocessing of data is necessary. But in big data environment, traditional data pre-processing technologies cannot reach the real-time demand of the applications. Hence, the key point of the pre-processing of messy data is to directly do feature selection to the raw data with noise and redundancy, and find out the features that matter most, so that the real-time changing of analysis demand can be followed.

e. Data Analysis

Data analysis difficulties such as the identification of features for distributed data analysis, an unbalanced dataset, as well as big data modelling persist despite the advances achieved by IDA methods.

- Asymmetrical Dataset Feature Selection

Data preparation has picked the most essential characteristics from such multi-source heterogeneous dataset, yet this selection may result in an unbalanced dataset. The minority in an unbalanced dataset are always ignored as noises by traditional IDA techniques. However, in other situations, such as fault diagnosis, the knowledge contained in these outliers may be quite significant. For this reason, it is vital to build specialised feature selection methods for unbalanced datasets.

f. Application Pattern

There is a limited number of ways to use IDA in the traditional sense. Patterns for such cross-platform applications.

Data interchange standards and display of complicated data face additional hurdles in the context of big data.

- **Exchanging Data Standard**

As IDA's application pattern evolves, so does the amount of data that may move across various platforms. It is challenging for such cross-platform IDA programme to unify data storage formats as well as data structures during the data transmission. The expenses of standardising data storage formats as well as architectures are also increasing. RosettaNet, for example, is indeed a standard for industrial data transmission, but it hasn't been widely used because of issues with universality as well as usability. By defining the data exchange standard having strong relevance, universality as well as usability, it is possible to even further boost data mobility inside a large-scale data environment.

- **Visualization**

Data that is difficult to analyse In a (DSS) decision support system, data visualisation may provide a straightforward as well as consumer-friendly man-machine interface. Inside a big data context, correlations between data grow increasingly complicated because of the rising size, dimensionality, data sources, as well as heterogeneity of such data. The decision maker might well be better able to understand the IDA findings but also make more informed judgments if the data is shown. In spite of this, IDA products may become better known as a result of the display of complicated data.

5. CONCLUSION

There are several stages to data analysis: issue conceptualization; data quality assurance; model creation; interpretation as well as post-processing. An investigation of intelligent data analysis has been conducted in this study. Problems in actual applications are the primary source of the issues, thus solutions must be tailored to the situation at hand. It is essential that new data models as well as lDAs be customised for individual applications in order to extract the maximum amount of value and information that can be put to use. IDA researchers must engage more with industry and mix actual applications as well as theoretical investigations in order to address the issues that may arise in future.

REFERENCES

1. R. Nayak, Data Mining for Web-Enabled Electronic Business Applications, to be published in Architectural Issues of Web-Enabled Electronic Business, Shi Nansi Ed., Idea Publishing Group, April 2002.
2. Zhang, Wei, and Feng Gao, "An Improvement to Naive Bayes for Text Classification," Procedia Engineering, vol. 15, pp. 2160-2164, 2011.
3. Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., and Strachan, R., "Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks," Expert Systems with Applications, vol. 41(4), pp. 1937-1946,2014.
4. Liangxiao Jiang, Zhihua Cai, Harry Zhang, and Dianhong Wang, "Not so greedy: Randomly Selected Naive Bayes," Expert Systems with Applications, vol. 39(12), pp. 11022-11028,2012.
5. Keshavarz, M., and Huang, B., "Bayesian and Expectation Maximization methods for multivariate change point detection," Computers & Chemical Engineering, vol. 60, pp. 339-353,2014.
6. Mahmoud, M. S., and Khalid, H. M., "Expectation maximization approach to data-based fault diagnostics," Information Sciences, vol. 235, pp. 80-96,2013.

7. Frolov, A. A., Husek, D., and Polyakov, P. Y., "Two Expectation-Maximization algorithms for Boolean Factor Analysis," Neurocomputing, vol. 130, pp. 83-97,2013

8. Zheng, B., Yoon, S. W., and Lam, S. S., "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Systems with Applications, vol. 41(4), pp. 1476-1482,2014.

9. M.e. Naldi, and RJ.G.B. Campello, "Evolutionary k-means for distributed data sets," Neurocomputing, vol. 127, pp. 30-42,2014.

10. Lin, e. H., Chen, e. e., Lee, H. L., and Liao, 1. R., "Fast K-means algorithm based on a level histogram for image retrieval," Expert Systems with Applications, vol. 41(7), pp. 3276-3283,2014.

11. P. Cabena, P. Hadjinian, R. Stadler, J. Verhees & A. Zanasi, Discovering Data Mining from Concept to Implementation, Prentice Hall PTR, 1997.

12. J. Han & M. Kamber, Mastering Data Mining, San Francisco: Morgan Kaufmann, 2001.

13. R. Agrawal & R. Srikant, Fast Algorithms for Mining Association Rules, IBM Research Report RJ9839, IBM Almaden Research Center, 1994.

14. Kambatla, K., Kollias, G., Kumar, V., and Grama, A., "Trends in big data analytics," Journal of Parallel and Distributed Computing, in press.

15. Kwon, 0., and Sim, 1. M., "Effects of data set features on the performances of classification algorithms," Expert Systems with Applications, vol. 40(5), pp. 1847-1857,2013.