Academic Guru Publication House

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

AG
PH
Books

A Survey on Intelligence Data Analysis: Issues and Challenges

Dr. Pankaj Saxena[1*]

[1]*Professor, R.B.S. Management Technical Campus, Agra (India) , E-mail : pankajrbsmtc@gmail.com*

Abstract

Many real-world applications now need automated or semi-automatic analysis of the data, which has given rise to the new area of (IDA) "intelligent data analysis," which combines several disciplines, including artificial intelligence and statistics in particular. As a whole, they work well together: Computing power is not a replacement for statistical understanding when it comes to statistical procedures for huge data sets. As a result, data analysis systems are becoming more sophisticated. When analysing data, there are a broad variety of issues that might arise. This article discusses some of these issues and provides possible solutions. Using data from a real-world level crossing risk assessment scenario, we investigate some of these issues and ideas.

Keywords: analysis of data, mining of data, level crossing risk evaluation, extraction of rules, artificial neural networks, induction of rules.

1. INTRODUCTION

Data intelligence refers to the process of using AI and machine learning to analyse and transform massive information into intelligent data insights which can be used to improve services and investments. The use of data intelligence tools and methods may assist in the development of improved business processes via the development of a better understanding of acquired information.

The data-driven intelligence process consists of five main components: descriptive data, prescriptive information, diagnostic information, and predictive information. In these fields, you'll learn how to

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

analyse data to comprehend it, acquire new knowledge, and come up with solutions to problems. Cybersecurity, finance, health, and insurance, as well as law enforcement, are some of the most pressing fields in need of data intelligence. An intelligent data capture system may be used to turn print documents or images into usable data in these situations.

Big data and business intelligence are strongly dependent on the utilisation of intelligent data. An intelligent data processing technique that restructured and enhanced the enormous datasets utilised by AI is capable of helping humans discover patterns, develop well-informed judgments, and adapt to changing circumstances. In order to improve the visualisation of prescriptive and predictive analytics, advanced analytics approaches are also used.

AI, high-performance computing (HPC), pattern recognition (PCR), and statistics (STAT) are all employed in IDA, which is an interdisciplinary study focused on extracting useable data. Strategic Data Intelligence, Global Data Intelligence, and the like all offer platforms and solutions for the analysis of large amounts of raw data.

There has been an increase in the need for more advanced IDA approaches as a consequence of the expansion in online and multimedia activities, as well as electronic commerce and others [1]. Analyzing vast volumes of data with detailed descriptions seems tempting on the surface, but in practise it is quite tough. There must be a plan in place to properly use the vast amounts of data that are generated by such massive and complicated datasets.

IDA derives value from data by uncovering patterns and rules in the data. It is impossible to count the number of IDA algorithms in the world, but the trend of their development may be characterised in three ways: (a) algorithm principle, (b) magnitude of dataset, and (c) kind of dataset.

2. ALGORITHM PRINCIPLE

IDA's algorithm has progressed from a basic to a complicated state during the course of its development. Early IDA algorithms were built on a basis of classical probability theory and a distance-based similarity theory based on Euclidean geometry. Computational intelligence was included into the IDA later, making its principles more sophisticated.

2.1. Probability Based Algorithm

The IDA methods based on probability probability theory are often used for the classification and grouping because of the property of probability theory itself. Prior probability and posteriori probability are used by the Naive Bayes Classifier (NBC) to classify sample data. Classification is performed by C4.S using the sample data entropy gain, while clustering is performed by Expectation Maximization (EM) using the maximum likelihood estimate of the parameters. They are extensively utilised because of their ease of implementation and high performance.

Second feature selection may increase the accuracy of NBC in huge text classification using the auxiliary feature strategy. Rework over-fitting and boost classification accuracy by using Decision Trees and neural networks. The Randomly Selected Naive Bayes (RSNB) technique overcomes the local optimum problems that afflict standard NBC through using stochastic processes in the NBC's feature stage of the selection process.

EM may be used to identify change points in multivariate data in plant monitoring. The authors [6] recommend adopting an EM hybrid approach based on forward-backward Kalman filtering for data-driven fault identification. [7] do high-dimensional Boolean factor analysis using two novel EM approaches.

2.2. Euclidean Distance Based Algorithm

It's possible to visualise the similarity between distinct components in an n-dimensional dataset by comparing the Euclidean distances between the dataset's n-dimensional vectors. Euclidean distance IDA techniques that focus on finding the cluster centres by minimising the total number of mean-square errors, such as the k-Means and the k-nearest-neighbor (k-NN) algorithms, are widely used. Example points in space are represented in a higher-dimensional context by use of the Support Vector Machine (SVM) model. These points are then translated to a higher-dimensional space in order to create two hyperplanes that are as broad as feasible.

[8] identify benign and malignant breast cancer tumours using a combination k-Means and SVM algorithm.

By incorporating an evolutionary approach into the k-Means method, we may lessen our dependence on the original cluster centres while simultaneously improving our capacity to handle scattered data. The repetitious training for continuous input condition may be eliminated by using a quick k-Means algorithm to graphic processing [10].

3. DATA ANALYSIS TASKS AND TECHNIQUES

Predictive modelling, clustering, and link analysis may all be part of a data analysis process, depending on the end user's goals and interests [11]. Making predictions based on fundamental aspects of the data is the purpose of predictive modelling. Using a mathematical model, data must be mapped to one of several established classes or to a real-valued forecasting variable. Predictive modelling may be done using any supervised machine learning approach that trains a model based on previous or current data. In order to train the model, we give it a set of previously known information and ask it to predict the future with the correct answers. Neural networks, decision trees, Bayesian classifiers, K-nearest neighbour classifiers, case-based reasoning, genetic algorithms, and rough and fuzzy sets are some of the approaches used to map discrete-valued target variables. Variables with continuous values may be mapped in many ways, including regression, induction trees, neural networks, and radial basis functions.Clustering is a technique used to create hierarchies of events by grouping together those having similar features. Clustering may be accomplished using any unsupervised machine learning approach for which the incoming data set does not include a preset set of data categories. There are certain pre-existing facts that the model is given, from which it produces categories of data with comparable features. These include partitioning, hierarchical strategies based on density and model-based methods [12]. "

Using link analysis, one may discover the intrinsic connections between data points. Achieving this objective is made possible by tasks such as discovering associations, identifying sequential patterns, and performing other sequence-discovery activities [11]. They reveal samples and patterns by forecasting correlations between elements that are not evident. When it comes to link analysis, it's all about counting all conceivable combinations. Apriori and its variants [13] are among the most often utilised algorithms.

4. DIFFICULTIES THE IDA FACE IN A BIG DATA ENVIRONMENT

There are significant roadblocks for IDA in the age of big data, when people's desire to get the most out of their data has never been greater. In a big data world, the IDA is confronted with four perspectives: (a) Large data management, (b) data gathering, (c) data analysis, and (d) application pattern are all aspects of the data lifecycle.

4.1. Management of Big Data

Hadoop Distributed File Solution (HDFS), for example, is a rather established system for managing enormous amounts of data. A good large data management system, on the other hand, does more than just store information correctly. Managing data lifecycles, data security, and costs in the context of big data management are all important considerations for getting the most out of data.

- **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 the data life cycle ends when the analysis of the data is done. Data lifecycle management is no longer a straightforward subject in the big data era. Even for an identical dataset, the value extracted from data by various users is varied. Data lifecycles are also varied because of this. When a patient has fully healed, their medical record's life cycle comes to an end from the patient's perspective. However, if a clinician is interested in learning more about a patient's family history of allergies, the medical record might provide valuable information. Medical records integration is critical for epidemiologists looking to learn more about an outbreak. If you maintain a dataset differently, you could get different information out of it, as shown in this scenario.

4.2. Data Security Management

The IDA has another challenge: managing data security throughout its lifecycle. People in a big data environment are always worried about their data's privacy and security. Personal, corporate, and national secrets may be compromised at any point in the data lifecycle in such an open environment. Data encryption may help protect data, but it can also limit the speed at which it can be processed if the encryption is very complicated. Data fuzzification, in addition to encryption, is another option for data security, albeit it has the potential to distort the data. Data security management's primary issue is to optimise the performance of IDA in the context of data security.

4.3. Cost Management

The IDA is also dealing with the difficulty of keeping costs down. The expense of improving IDA performance must be kept under control in order to ensure long-term sustainability. It is possible to increase the value derived from data by using diverse life cycle management techniques, such as distributed data analysis, complicated data encryption, and a range of other life cycle management alternatives, all while raising the costs of hardware and network transmission. The major purpose of cost management is to lower both explicit and implicit costs as long as the value offered is adequate for a specific application and a balance between value and cost can be established.

4.4. Data Collection

Multi-source and heterogeneous big data features have emerged in addition to their rising size. Integrating and pre-processing huge multi-source heterogeneous data sets is a challenge in data collection.

- Multi-source Heterogeneous Fusion

The great variety of data sources accessible in a big data context results in data heterogeneity.. The data may also include additional elements like as text, images, audio, and other electrical impulses. A new big data framework is essential to integrating all of the heterogeneous multi-mode data into a format that IDA can process.

- Pre-processing of Messy Data

Noise and redundancy in raw data may have a significant impact on IDA's performance, accuracy, and resiliency [15]. We can't avoid preparing the data in some way. There is a problem, however, with typical data preparation solutions in a large data context. When dealing with noisy and redundant data, the primary goal of pre-processing is to rapidly execute feature selection to identify the most important characteristics so that real-time changing analysis requirements may be followed in a timely manner.

4.5. Analysis of Data

Many IDA approaches, including as feature selection for imbalanced datasets, distributed analysis of data, and big data modelling, have made great progress in data analysis, but there are still many challenges to be faced.

- Imbalanced Feature Selection

Dataset Data preparation may result in an unbalanced dataset, even if the most significant characteristics are picked from the multi-source heterogeneous dataset. Because traditional IDA algorithms focus on broad generalisation, the unbalanced dataset's minority are ignored and treated as random noise. However, in other situations, such as fault diagnosis, the knowledge contained in these outliers may be quite significant. The unbalanced dataset necessitates the development of feature selection methods that are tailored to the problem at hand.

4.6. Application Pattern

Traditional IDA use cases tend to follow a set pattern. Patterns for cross-platform applications.

Data sharing standards and complicated data visualisations face additional problems in the big data age.

- Data Exchange Standard

As IDA's application pattern evolves, so does the amount of data that may move across platforms. For cross-platform IDA applications, it's difficult to unify the many storage formats and data structures which are employed in the data interchange. Standardizing data storage formats and systems will need further funding. Data interchange standards exist in certain industries, such as RosettaNet, but they haven't gained much traction because of issues with universality and usability. Therefore, in the big data era, developing an exchange standard with high pertinency, universality, and usability will become a major challenge.

- Visualization

Complicated Data In a decision support system, data visualisation may provide a straightforward and user-friendly man-machine interface (DSS). In a big data context, correlations between data grow increasingly complicated because of the rising size, dimensionality, data sources, and heterogeneity of the data. The decision maker may be better able to understand the IDA findings and make informed judgments if the data is shown. However, the ability to easily understand complicated data via the use of IDA tools may help them become more well known.

5. CONCLUSION

Data analysis is an iterative and interactive process that includes problem conceptualization, data quality assurance, model construction, and interpreting and post-processing of the results..Using intelligent data analysis, this article looked at the underlying concerns and challenges. Issues in real-world applications provide the most difficult difficulties, and the best solutions are those that focus on solving those problems. Data models and lDAs should be tailored to specific applications in order to get the most value and actionable insights from them. IDA researchers should interact more with industry and mix actual applications and theoretical research in order to solve the issues that may arise in the next years.

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.