World Journal of Advanced Engineering Technology and Sciences

eISSN: 2582-8266

Cross Ref DOI: 10.30574/wjaets

Journal homepage: https://wjaets.com/

(Review Article)

A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach

Omari Firas *

Jordan University of Science and Technology, IRBID, Jordan.

World Journal of Advanced Engineering Technology and Sciences, 2023, 08(01), 009–014

Publication history: Received on 07 November 2022; revised on 23 December 2022; accepted on 26 December 2022

Article DOI: https://doi.org/10.30574/wjaets.2023.8.1.0147

Abstract

Data analytics has emerged as one of the most advanced technologies in recent times. However, the successful implementation of analytics is still a great challenge since they suffer from technical barriers and have a lack of structured approaches for performing analytics. Data mining models are considered as a potential tool for solving problems related to data analytics. Data mining is a process used for extracting the relevant attributes from raw data, which is further processed using the mechanism of knowledge discovery for support decision making. Formal concept analysis (FCA) provides a robust platform for knowledge discovery and helps in the successful adoption of data mining for handling big data. Several mining techniques powered by FCA are discussed by the researchers. However, the analysis of FCA suggests that the effectiveness of FCA for big data needs, a deeper investigation in order to expand its application horizon. In this context, this research emphasizes the application of FCA for developing an effective strategy through a combination of SEMMA and CRISP models for handling big data by integrating knowledge discovery with data mining.

Keywords: Formal concept analysis; Knowledge discovery; Data mining; Big data; CRISP-DM model; SEMMA model

1. Introduction

1.1. Background of the study

The value of DA - Data Analytics has been broadly utilized by various academic and industrial communities. Big Data and Data Mining concepts transform our lives in all dimensions from businesses to consumers. Therefore, this transformation takes place at each level from different microeconomics to macroeconomics, where the consumers and manufacturers connect with each other in order to realize emerging forms of value that lead to societies and economic function. Further, data analytics and big data are considered ever-growing technology that attracts more attention in an industrial view and scientific literature with various news media. Most of the companies using these technologies believe that analytics and management deliver values and indicate basic competitive source benefits in the industry. However, due to the lack of DA insights in the system, most companies are not aware of the potential veiled behind certain unexploited available data that creates with the advanced information methods. Nevertheless, changing the real-time data into a data-driven one becomes challenging [1]. Moreover, identifying the activities of the human is considered an important factor in the customized context-aware systems design and development applied in living applications, further this enables the appropriate assistance. Normally the system will gather information with the behavior analysis to identify the activities processed and the continuous behaviors that are considered vital in order to complete the activity performed. At the same time, it is important to pick the appropriate feedback in order to help the smart environment residents. However, identifying human activities using the sensor-based smart home is considered a challenging task, especially in the scientific community due to the human-object interactions, different behavioral

* Corresponding author: Omari Firas; Email: firasomari2015@gmail.com
Jordan University of Science and Technology, IRBID, Jordan.

patterns, and reliable data. Further, the difficulty during the recognition of activities increases due to multiple residents, especially in a smart environment. In these scenarios, multiple inferences should be employed to the similar sensors at the same place and exact time [2].

1.2. Formal Concept Analysis (FCA)

FCA - Formal Concept Analysis is considered a powerful and efficient tool support system. Recently, several tools on Big Data exist and failed to identify the issues occurring between the current tools and appropriate processing that should be exhibited in order to address the FCA's current needs. It is significant for researchers to find scalable and efficient solutions, especially for FCA that enable the proper prototyping for executing FCA cross-tables even on a large-scale [3]. The topic detection task is completely targeted at exploring the main important topics in order to be addressed by a whole range of documents and the topics are outlined as self-contained, cohesive, and thematically similar. Hence, this particular task is widely examined from the clustering and probabilistic methods. The FCA concept is an exploratory method utilized for data analysis [4] and these FCA-based techniques are used for topic detection that is employed in the literature by providing the stability concept for the selection of topic. Certainly, the FCA-based concept has the ability to overcome the issues of clustering and probabilistic methods. With the increasing capability to accumulate, analyze, and store the data generated with increasing frequency, the data science field started to grow rapidly. Similarly, due to the rapid increase in the organization using data analysis allows accumulating of larger amounts of data. Therefore, these analyses keep on increasing which becomes a team activity, instead of work performed by a single data scientist. During this process, the growth in the utilization of big data has exceeded the knowledge of supporting teams that are required to do big data tasks. The algorithms utilized in the research help develop insightful analysis, frameworks, tools, and techniques that allow teams to process effectively to perform big data projects. Still, operations research, business intelligence, and software development are discussed in existing research that provides proper insights on big data process utilization [10].

1.3. Data mining and knowledge discovery

The data mining concepts and knowledge discovery techniques used in recent research have been discussed with different degrees of success. Both DM - Data Mining & KDD - knowledge discovery in databases, DM, and KDD terms are utilized to define the techniques, tools, and research utilized in order to extract appropriate and useful information from a large volume of information or data. The process which has been completely executed in the data extraction process is known as the KDD process. The existing research such as CRISP-DM - Cross-Industry Standard Process for DM - Data Mining [11] that is used in DM development projects. Therefore, these techniques define the activities in order to develop a DM project and each activity consists of tasks. These tasks generate output and required inputs are detailed and these methods are used to resolve the issues with CRISP-DM in existing DM development projects.

2. Literature Review

In [5], fuzzy-concept has been utilized for minimizing the number of concepts in FCA of data with the help of fuzzy attributes. Therefore, the weight of FFC - Fuzzy Formal concepts is computed based on Shannon entropy. Further, the weight of FFC is minimized at selected computed weight granulation. The results are acquired from the proposed method with interval-valued FFC and Levenshtein distance technique has been procured with better results, however with minimum computational complexity.

In [6], the k-nearest neighbor algorithm - a DM technique has been examined to obtain smart data which are high-quality data and the issues are discussed along with the data preprocessing methods in order to eradicate the noise, missing values, and redundant data in order to speed up the process of the techniques. Various data preprocessing methods are discussed based on the k-NN techniques. Therefore, relevant k-NN-based data preprocessing methods are picked and experimented with under Apache spark and have processed empirical analysis with the collected behavior from datasets, the enable non-experts and practitioners in the field to define the Smart Data-preprocessing methods with big datasets. However, in most of the cases, the accuracy was minimized and the reduction offered is so high due to redundant information.

In [7]~~,~~ the attribute reduction technique has been utilized for formal decision contexts and a simplified discernibility matrix used in the research is not required to develop the formal concepts. It was indicated that the computation time and storage space are less compared to the original technique. Moreover, approximation algorithms are used for acquiring a less redact that takes place in a formal decision context in accordance with the designed graph theory. Eventually, various experiments are executed to validate the effectiveness of the utilized method. The results obtained from larger datasets with the attributes procured better performance in both storage speed and space. However, the reduction approach in formal decision contexts is taking considerable time because of discernibility matrix construction and computational complexity.

In [8], state-of-the-art principles and DM methods are reviewed with missing information or process outliers in different industrial applications. The data preprocessing methods are compared with the normalization and data cleaning. The robust statistical process modeling methods have been utilized on the PCA basis. The different statistical DM methods are examined with various process characteristics such as dynamics, non-linear, and non-Gaussian.

In [9], the interval pattern structures are used on categorizing the work into three including bi-clusters that have been utilized in recommender systems with effective techniques and appropriate semantics lacking; pattern-based classifier defines numerical patterns, closed patterns, and generators that enable with the association rules initialization while processing it so that can be utilized in supervised classification tasks; *k*-anonymity with the projections that enables the closed patterns and the presented generators are utilized in the datasets for securing the data and critical problems exhibited on the web.

In [11], an overview of knowledge discovery methods, Data Mining, and process models have been utilized in the study and a brief description of the KDD process has been elaborated in this research with the special features, pros, and cons. Other than that, the presented DM techniques that focus on various tasks have been defined in each method and interpret the entire KDD process. The results are compared with the new DM and KDD process which is known as a refined DM process for developing any kind of KDD and DM, this technique helps to develop specific tasks carried out with the analyzed techniques.

The work mentioned in [12] proposed a domain driven data mining approach for developing an actionable knowledge discovery framework. The primary objective of this research is to enhance the implementation of the learned rules. To achieve this aim, the proposed approach analyzes the actions that have to be performed in the business tasks and to determine an appropriate sequence for performing the tasks. In addition, the constraints required for performing the tasks and the data used in these tasks are also investigated. It was observed from the results that the knowledge obtained from the rule learning aspects or from the domain knowledge is not sufficient for performing the tasks and require a deeper analysis.

Significance of the study

Recently, data mining has shown impressive effectiveness in handling big data and has achieved success when exploited in tasks like data analytics. Moreover, the insertion of attention mechanisms in data mining have made it feasible to consider and extract the related important data rather than extracting the available information which are invaluable for big data analytics. These characteristics have motivated to develop a data mining based framework and explore its robustness and efficacy for performing effective big data analytics. Furthermore the excellence of FCA in knowledge discovery has motivated this research to integrate FCA with data mining for effectively handling big data.

The main contributions of this study are summarized in below points

- The study presents a comprehensive analysis of FCA based knowledge discovery for data analytics and explores its role in handling big data.
- The study identifies the main challenges associated with the adoption of data mining and KD and attempts to overcome the problem of conventional techniques which fail to analyze big data without using preprocessing techniques.
- The study introduces a novel concept for modeling the FCA based knowledge discovery process and intends to contribute to the existing research works.

3. Problem Statement

Data mining and knowledge discovery (KD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases. Data mining helps in the generation and extraction of valuable non-trivial knowledge and data from unstructured and structured information. Data mining and KD is used in the visualization, categorization and explication of information and hence are considered as the key elements for enhancing the effectiveness of business processes and main enablers of competitive benefits. However, the success rate of data mining is very low and the value of data in most of the organizations are not available. Various researchers have discussed the enhancement of knowledge discovery and data mining using Formal Concept Analysis (FCA). It can be inferred from existing works that the main challenges associated with the adoption of data mining and KD are the lack of availability of structured techniques for data analytics. Several data mining models powered by FCA have been proposed and there has been a trivial interest in this domain. However, an analysis of current techniques have suggested that the performance of FCA is not satisfactory and even appropriate FCA based techniques cannot be successfully applied for large scale applications. In addition, most of the existing works have used FCA to deal with numerical data, and conventional techniques cannot analyze this data without using preprocessing techniques [12].

The KD model acts as a data centric model which acts as an iterative and interactive data analytics process. However, while data-related tasks are detailed, there is a lack of business perspective in this outline model. Most of the existing KD approaches do not focus on performing data mining for extra large scale datasets. This is mainly because these approaches suffer from high overhead problems while dealing with large scale data processing. This restricts the adaptability of the KD approaches for applications requiring larger datasets. When a software model is used, the KD process becomes an important step in the software development framework that encompasses project management and development processes. Hence, it is essential to develop an efficient model for data mining and KD which is suitable for a dynamic environment and performs well for the volatile objectives which are used in the exploratory analysis. Else, the classical software development approaches might become too rigid and cumbersome to follow. The work presented in [1] shows that despite the availability of several research works, there is a lack of an effective research approach which can overcome the intrinsic limitations of traditional process models and achieve better performance in agile technologies. There is a great demand for a holistic approach to enhance the data mining and KD process and offer better support to analytics project management. Also, it is important to reduce the risk of analytics project failures by accompanying better technologies which can help in designing better processes and provide better understanding of the potential impact of analytics models.

In this context, this research aims to provide a comprehensive analysis of Knowledge Discovery and data mining to address the limitations inferred from existing works.

Aims and Objectives

The preliminary aim of the proposed research is to develop a data mining and knowledge discovery approach for handling big data using formal concept analysis (FCA). The objectives considered in this research are as follows:

Research Objectives

- To explore the application of interactive knowledge discovery and data mining with numeric formal concept analysis.
- To develop an effective strategy for handling big data by integrating knowledge discovery with data mining.
- To improve the efficiency and scalability of the proposed approach in order to make it suitable for larger data sets and to improve the robustness to missing and incorrect items in data, visualization and exploration of the mining results

4. Research Methodology

This research aims to integrate a data mining and knowledge discovery approach for handling big data applications. This research aims to integrate two effective KDD models namely SEMMA and CRISP–DM models. It can be inferred from existing literary works that most of the organizations and business processes do not have a fixed and appropriate methodology for improving the business related tasks. This clearly indicates that the conventional big data processing tools are not quite effective and there is always a lack of a potential robust model which can meet the requirements of the analytical projects in a more efficient manner. Correspondingly, there is also a great demand for the implementation of domain specific methodologies which also necessitates the need for a standard KD process. The increase in the adoption of own methodologies indicates the significance of data analytics models which also includes data mining tasks. The experimental outcome suggests that the process models employed by the data analysts and business management people intend to research more on the methodologies available and identify their own steps based on their requirements. Considering this fact, this research aims to develop a novel methodology by integrating two effective KDD models for effectively handling big data.

4.1. Overview of the proposed technologies

The SEMMA process model consists of 5 main stages namely, sampling, exploring, modifying, modeling, and assessing. These stages focus on modeling the development aspects of data mining. It can be considered as a fundamental plan for modeling the KDD process and it works perfectly with all tasks of KDD in parallel. The workflow stages of the SEMMA model is linked to the SAS Enterprise Miner software and hence is considered as one of the top most processing models used for data analytics. Though the SEMMA model can be used extensively, it is challenging to perform iterations and carryout interactions between multiple tasks in the original model. In order to overcome this drawback of the SEMMA model, this research intends to combine it with the CRISP–DM (CRoss Industry Standard Process for DM).

CRISP-DM is based on a waterfall life cycle model which has a hierarchical process which helps in providing guidance on how to perform a task. In general, there are six important stages which constitute the top level of hierarchy, which are; understanding the business, understanding the data, data preparation, modeling, performance evaluation, and model deployment. These phases incorporate generic tasks wherein the inputs and corresponding outputs are defined clearly. The generic tasks cover the entire data mining process and are suitable for all data mining applications. This process is more stable since it is applicable for all novel and unknown technologies and processes. Though the process involved in the waterfall life cycle is linear and sequential, the CRISP-DM considers an iterative process and feedback loops. It is one of the excellent and well documented process models and its clear and comprehensive documentation plays an important role in improving the functioning of the standard knowledge discovery process models. The CRISP-DM model incorporates all updated elements from the previous models and these elements are considered as the reference for performing all future processes. There are two main explicit steps involved in the CRISP-DM model i.e., business understanding and data understanding. These two steps are considered as an edge for making any data mining task successful since they help in gaining more insights into the objectives of business processes and the data availability.

4.2. Proposed workflow

There are two main stages involved in this process for handling big data which are data discovery and model deployment.

4.2.1. The discovery phase

This phase is based on the process of the SEMMA model which includes the following steps:

Analyzing the requests

This phase is similar to the understanding of the business requirements, process, and goals.

Data preparation

The data preparation mainly focuses on collecting data from different sources and converting into relevant and valid input for the data analytics model (here the SEMMA and CRISP-DM model) to process big data. This step is more time consuming and is one of the critical steps involved in the knowledge discovery and data mining process.

Data exploration

In this step, the model performs initial data analytics using modern data visualization tools which refines the business data such as business queries, and applies a suitable analytical approach and identifies the missing data in order to make data suitable for data mining tasks.

Data modeling

The data is modeled using analytical and machine learning algorithms to identify the relationship between the data to answer the business related queries.

4.2.2. The deployment phase

This stage is related to the implementation of the proposed model which can learn from the data discovery phase and put into action using the automated and suitable data models. This step is critical since the entire phase is dependent on the deployment phase. Once the model is deployed and is approved, any strategic and operational decisions can be made and the performance is evaluated. The obtained feedback is used for further deployment.

Limitations of the study

Despite the fact that the proposed approach aims to address the limitations of the conventional techniques, there are certain limitations of this study which are as follows:

- The study does not focus on the implementation of modern big data tools such as Apache Spark or Apache Cassandra for processing large scale big data. This might require additional data processing tools and might increase the computational cost.
- The study is limited to the analysis of FCA based knowledge discovery for handling big data without emphasizing on the issues such as data quality, storage, data validation and the impact of data accumulation from different sources.

5. Conclusion

This study presents a suitable data mining and knowledge discovery approach for handling big data using the concept of formal concept analysis. This research aims to develop an effective strategy for handling big data by integrating knowledge discovery with data mining. An integrated approach combining the SEMMA and CRISP-DM model is proposed for achieving this aim. It is believed that the proposed approach will improve the efficiency and scalability of the proposed approach and can provide a robust approach for handling big data.

Compliance with ethical standards

Acknowledgments

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Disclosure of conflict of interest

The author has no relevant financial or non-financial interests to disclose.

References

[1] Rotondo, A., & Quilligan, F. (2020). Evolution paths for knowledge discovery and data mining process models. SN Computer Science, 1(2), 1-19.

[2] Hao, J., Bouzouane, A., & Gaboury, S. (2018). Recognizing multi-resident activities in non-intrusive sensor-based smart homes by formal concept analysis. Neurocomputing, 318, 75-89.

[3] Tamburri, D. A. (2020). Design principles for the General Data Protection Regulation (GDPR): A formal concept analysis and its evaluation. Information Systems, 91, 101469.

[4] Castellanos, A., Cigarrán, J., & García-Serrano, A. (2017). Formal concept analysis for topic detection: a clustering quality experimental analysis. Information Systems, 66, 24-42.

[5] Singh, P. K., Cherukuri, A. K., & Li, J. (2017). Concepts reduction in formal concept analysis with fuzzy setting using Shannon entropy. International Journal of Machine Learning and Cybernetics, 8(1), 179-189.

[6] Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors' algorithm to obtain quality data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(2), e1289.

[7] Chen, J., Mi, J., Xie, B., & Lin, Y. (2019). A fast attribute reduction method for large formal decision contexts. International journal of approximate reasoning, 106, 1-17.

[8] Zhu, J., Ge, Z., Song, Z., & GAO, F. (2018). Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. Annual Reviews in Control, 46, 107-133.

[9] Kaytoue, M. (2010). Mining Patterns in Numerical Data with Formal Concept Analysis.

[10] Saltz, J. S. (2015, October). The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 2066-2071). IEEE.

[11] Mariscal, G., Marban, O., & Fernandez, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review, 25(2), 137-166.

[12] Fatima, F., Talib, R., Hanif, M. K., & Awais, M. (2020). A paradigm-shifting from domain-driven data mining frameworks to process-based domain-driven data mining-actionable knowledge discovery framework. IEEE Access, 8, 210763-210774.