Knowledge-Based Systems 192 (2020) 105441

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

An analysis on new hybrid parameter selection model performance over big data set☆

Masurah Mohamad [a,b,g], Ali Selamat [a,b,c,d,*], Ondrej Krejcar [d], Hamido Fujita [e,f], Tao Wu [h]

[a] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] MagicX (Media & Games Center of Excellence), Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[c] Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur, Malaysia
[d] Faculty of Informatics and Management, University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
[e] Fac. of Software and Information Science, Iwate Prefectural University, 020-0193, Iwate, Japan
[f] Andalusian Research Institute DaSCI (Data Science and Computational Intelligence), University of Granada, Spain
[g] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Cawangan Perak, Kampus Tapah, 35400 Tapah, Malaysia
[h] Shanghai Jiao Tong University, School of Medicine, China

ARTICLE INFO

Article history:
Received 6 March 2019
Received in revised form 23 December 2019
Accepted 26 December 2019
Available online 30 December 2019

Keywords:
Big data
Parameter selection
Analysis tool
Decision
Hybrid method

ABSTRACT

Parameter selection or attribute selection is one of the crucial tasks in the data analysis process. Incorrect selection of the important attribute might generate imprecise or event for a wrong decision. It is an advantage if the decision-maker could select and apply the best model that helps in identifying the best-optimized attribute set — in the decision analysis process. Recently, many data scientists from various application areas are attracted to investigate and analyze the advantages and disadvantages of big data. One of the issues is, analyzing large volumes and variety of data in a big data environment is very challenging to the data scientists when there is a lack of a suitable model or no appropriate model to be implemented and used as a guideline. Hence, this paper proposes an alternative parameterization model that is able to generate the most optimized attribute set without requiring a high cost to learn, to use, and to maintain. The model is based on two integrated models that are combined with correlation-based feature selection, best-first search algorithm, soft set, and rough set theories which were compliments to each other as a parameter selection method. Experimental have shown that the proposed model has significantly shown as an alternative model in a big data analysis process.

1. Introduction

The data analysis process is the most crucial tasks in any application field. This process involved several tasks such as data pre-processing, data extraction, and data selection that will assist the decision-maker in getting the best solution for the specified problem. The parameterization process is a process of identifying the optimal set of a parameter by using any specified tools, mathematical formulations, or any modeling techniques [1–3]. This process involved preparing the data from raw and unstructured form until cleaned, formatted, and optimized data. An ineffective parameterization process will affect the decision-making process and return either wrong or inefficient solutions. Several factors might affect the parameterization process to become ineffective. The size and characteristics of the data are two main factors that will downgrade the efficiency of the parameterization process. The large size of data might contain a complex type of data set which means that the data set has multiple types of criteria and also has an imbalance, uncertain, and inconsistent data values [4]. Complex data sets are difficult to be analyzed, especially when using unsuitable methods and instruments. Many application fields such as healthcare [5], financial [6], transportation [7,8], and engineering problems [9] had conducted various research works in solving this issue. For instance, Wang et al. [10] had investigated on the feature selection methods that able to deal with bioinformatics data set, Pramanik et al. [11] discussed on the architecture and technologies of big data in healthcare system and Shen et al. [12] proposing the hybrid approach to diagnose the financial performance of life insurance companies. From these experiments, the researchers found that, the complexities of the data did influence the decision-making process.

Instead of considering the data, the researchers also need to select appropriate methods in conducting the decision-making

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.105441.

* Corresponding author at: School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.
E-mail addresses: masur480@uitm.edu.my (M. Mohamad), aselamat@utm.my (A. Selamat), ondrej.krejcar@uhk.cz (O. Krejcar), HFujita-799@acm.org (H. Fujita), mazuiwu88@vip.163.com (T. Wu).

https://doi.org/10.1016/j.knosys.2019.105441

process and for example, applying a fuzzy and rough set to handle uncertainties and non-linear problems [13,14], implementing neural networks to analyze complex data [11] and incorporating support vector machine (SVM) with other methods in dealing with high dimension data [15,16]. Different approaches, models, frameworks, or formulations had been proposed whereby, each of them is considered different kinds of problems or issues that need to be solved. Some of the works were highlighting the performance of the proposed methods or models, new definitions, new approaches for solving the capability of the hardware and software used in the decision-making process. All of these previous works had contributed to the focused area and will never end because the issue of data always emerge and become more complex.

Besides, many companies around the world are also actively searching and proposing solutions to deal with big data problems. Companies such as Oracle, IBM, and Google have introduced various kinds of technologies and applications in solving big data problems. Many products and techniques had been introduced to store, analyze, and visualize big data. For example, Google has initiated some of the services such as Google BigQuery, Google Cloud Datalab, and Google Cloud Dataproc that help customers in using and analyzing big data. Google had proven that all the provided services help the decision-makers to improve the decision-making process (https://cloud.google.com/why-google-cloud/big-data/). Oracle company also had contributed to a big data revolution by delivering service that could deal with traditional and new data sets, especially big data. Oracle provides a platform that could integrate, manage, and analyze big data in practical ways (https://www.oracle.com/big-data/index.html). Moreover, some of the researchers tend to investigate the parallel processing approach in handling big data [17].

Inspired by these remarkable existing works, this paper proposed an alternative way in selecting the most optimized parameter in a big data set using two phases hybrid parameter selection model. This proposed model was focused on handling high dimensions of data at the first level of parameter selection, while uncertainty data and inconsistency data will be identified and eliminated at level two of the hybrid model. Based on previous experimental works results [4,18], a combination of correlation feature-based selection (CFS) with best first search methods have been selected to execute the first level of parameter selection while at the second level, soft set (SS) and rough set (RS) parameter selection methods were used together to identify the uncertainty and inconsistency values in the data sets. This proposed work aims to provide an alternative approach to the decision-maker in conducting the decision analysis process using a cost-effective and time-saving model. This model requires no high-performance processor or ample memory storage when extracting, selecting, or analyzing any complex data sets. This proposed model will be implemented in the data pre-processing task to generate an optimized data set that will be used in the decision-making process before the decision is being made. This proposed model is suitable to be implemented in any decision-making field, such as classification, clustering, and prediction, as this model promotes a useful data pre-processing task.

The paper is organized as follows: Section 1, we briefly discussed the current issue related to the proposed work. In Section 2, we provided some of the essential works that related to the proposed model. The methodology of how the proposed model was conducted is discussed in Section 3. In Section 4, we proved the proposed work with an explanation of the data and results of the experimental works. Finally, in Section 5, we conclude the overall works and highlighted some of significant outcomes from the research.

2. Related works

Several topics related to the highlighted issue, such as big data, correlation feature-based selection (CFS), soft set (SS), and rough set (RS) parameter selection, are discussed in the following sections.

2.1. Big data

Since the world is facing with big data era, all information system components related to data such as technology and procedure were indirectly affected. Big data is characterized as high volume, high velocity, high variety, high value, and high veracity of the information that requires efficient and effective data processing tools [19,20]. The volume of big data refers to the data size of terabytes to zettabytes, while velocity means data in motions where the response rate of data from milliseconds to seconds when it is in the streaming process. High variety indicates the data contains many forms such as structured, unstructured, text, number, or multimedia. Big data with high value means the data consists of a different range of values such as from free to costly meanwhile, and data veracity indicates that the big data has very high uncertainty and inconsistency data. According to IBM, big data cannot be processed by the traditional relational database. Big data comes from various kinds of sources such as sensors, devices, video or audio, transactional applications, web, log files, and social media. These data are generated in real-time and on a huge scale (https://www.ibm.com/analytics/hadoop/big-data-analytics).

Big data had become a phenomenal and challenge to all data experts, either the database provider, data engineer, data analyst, or other related community, to come out with solutions on how to deal with this issue [21–24]. Big data can be categorized into four phases: (i) data generation, (ii) data acquisition, (iii) data storage, and (iv) data analysis [19]. Most of the solutions had been proposed to increase the capability and performance of the existing software, hardware, approaches, or algorithms in dealing with big data [25]. Some of the popular technologies related to big data are cloud computing, Internet of Things (IoT), Hadoop, NoSQL, and MapReduce. Several architectures such as reasoning, information extraction, and ontology alignment, several big data models such as BigTable, Cassandra and MongoDB, and plenty of methodologies, for instance cloud computing, granular computing, and quantum computing have been proposed to handle and implements big data sets. Companies from many industries, especially from financial, marketing, and retail, take advantage of big data. Big data do provide lots of beneficial information that might help these companies to produce new products and services. By using proper analytically methods and tools, profit and productivity can be increased while the performance can be improved. According to [25], current research works are focusing on big data storage and processing big data area. For example, the researchers were only focusing on three techniques, prediction, clustering, and classification. Only a few research works that focused on proposing the enhancement or new big data pre-processing area. As a conclusion, big data brings a broad issue to be explored.

2.2. Correlation-based Feature Selection

Correlation-based feature selection (CFS) is a multivariate feature selection method that was proposed by Hall in 1999. It is one of the feature selection methods in the filters category. This method will filter the attribute values based on the correlation heuristic function. CFS ranks and selects the values of the attributes by looking at the values that lean to the class

and has a correlation between the other attributes. CFS also eliminates the uncorrelated values with the class and the repetitious highly correlated attribute values [26,27]. CFS will rank and selects the values of the optimized attributes by two sequence phases, where phase one will calculate the correlation values between attributes and attributes and between attributes and class. Meanwhile, phase two will identify the most relevant attributes by looking at the attribute space using several heuristic search strategies such as best first search [28]. The formula to identify the most correlate attribute in the data set is shown in Eq. (1) [29].

$$cr_{zc} = \frac{f\overline{cr_{zi}}}{\sqrt{f + f(f-1)\overline{cr_{ii}}}} \quad (1)$$

where cr_{zc} is the heuristic value of a subset attribute for f number of attributes, $\overline{cr_{zi}}$ represents the average value of correlations between the attributes and the class, $\overline{cr_{ii}}$ holds the average value of inter-correlation between attribute pairs. The attribute set, which has the highest heuristic value, will be selected in the data reduction process. Then, the result of the reduction process will be selected as an optimized attribute set and to be used for the next analysis process.

The main advantage of CFS is that it requires less computational complexity compared to wrappers and other approaches. However, the performance of the learning algorithm is not promising compared to wrappers and other embedded approaches. Thus, many researchers had taken the initiative to improvise and to enhance the capability of CFS by integrating it with other feature selection methods. CFS has been widely implemented to deal many applications such as to solve issue of high dimensional data [30], medical [29], security [26] and bio-computing [28]. Recently, it is reported that CFS assisted the decision-makers in increasing decision-making performance by optimizing the capability of the existing decision analysis methods [30].

2.3. Soft set parameter selection

Soft set (SS) parameter selection is another filtering method that used mathematical formulation to select the most optimum attribute values in the data set. SS applies probability to measure the most optimal attribute sets and to eliminate the fuzzy, uncertainty, and inconsistency attribute values [31]. The soft set was initiated by Molodtsov in 1999 and was improvised and enhanced by many researchers to optimize the capability of soft set in helping the decision-maker in making good decisions [32]. Some researchers claimed that the soft set parameter selection method is better than a rough set parameter selection method in the process of identifying the most optimal and sub-optimal attribute values in the decision analysis process. Also, some researchers have claimed that the probability formulation used by the soft set parameter selection method much simpler than a rough set parameter selection method. The soft set parameter selection method had been implemented in many application areas and was proven successful [14,33].

The soft set is a mapping set from the parameter to the crisp subset of the universe. The following definition is the basic notion of soft set theory initiated by Molodtsov [34]. Other examples and proportions of soft set theory can refer to [34–36].

Definition 2.1. U is defined as a non-empty initial universe of objects. Then E is defined as a set of parameters concerning objects in U. Let $P(U)$ be the power set of U, and $A \subset E$.

A pair (F, A) is called a soft set over U, where F is a mapping given by $F : A \rightarrow (U)$. In other words, a soft set over U is a parameterized family of subsets of the universe U.

The soft set also was suitable to be integrated with other mathematical theories or models. Several experimental works had been conducted to validate the capability of the soft set parameter selection method. However, this method was unsuccessfully identifying the most optimal and sub-optimal attribute values when dealing with a large volume of data. Soft set (SS) suffered from the high computational complexity problem and required a large size of computer memory to execute the analysis process. Also, a soft set parameter selection usually resulted in producing the same number of attributes values as the original number of attributes of the selected data set [18,37,38].

2.4. Rough set parameter selection

Rough set (RS) parameter selection method is another filtering method that implements mathematical formulation. It was proposed to eliminate the ambiguous data based on the theory initiated by Pawlak in 1997 [39] using the probability concept. RS was focused on solving the fuzzy, uncertainty, and inconsistency problems that found in any kind of data [39]. Lately, researchers prefer to apply the RS parameter selection method in handling high dimensional data problems [40,41]. The detailed on Rough set theory definition and formulation can be referred to in many research works such as in [39].

The efficiency of RS parameter selection had been proven and being applied by many application fields to solve the complex problems [42,43] in health science and finance, in classification and prediction, or even as an optimization method to other decision analysis methods. Many researchers have taken initiatives on RS capability and its beneficence and try to improvise the RS by extending and integrating the original RS into a new RS concept. Some of the researchers tend to enhance the RS formulation by hybridizing the RS with other theories such as [44,45]. Some of the researchers used it as a supported method for the sake of improving the limitations faced by the other means. Some of the researchers improvised the rough set theory by extending it with new formulations such as the Dominance-based Rough Set Approach (DRSA) [46], which was proposed to handle ordinal data set and monotonic relationship that has been widely implemented.

The basic concept of a rough set can be defined as follows:

Definition 2.2. If the universe set U is a non-empty finite set, and σ is an equivalence relation on U. Then, (U, σ) is called an approximation space. If X is a subset of U, X either can be written or not as a union of the equivalence classes of U. X is definable if it can be written as the union of some equivalence classes of U, or else it is not definable. If X is not definable, it can be approximated into two definable subsets called lower and upper approximations of X as shown below [47].

$$\underline{app}(X) = \bigcup \{[x]_\sigma : [x]_\sigma \subseteq X\},$$

$$\overline{app}(X) = \bigcup \{[x]_\sigma : [x]_\sigma \cap X \neq \emptyset\}.$$

A rough set is comprised of $(\underline{app}(X), \overline{app}(X))$. Boundary region is when the set $\overline{app}(X)$ - $\underline{app}(X)$. Therefore, if $\underline{app}(X) = \overline{app}(X)$, X is definable. If $\overline{app}(X)$ - $\underline{app}(X)$, then X is an empty set.

For a set of X, $\underline{app}(X)$ is the greatest definable set contained in X, whereas $\overline{app}(X)$ is the least definable set containing X.

3. Methodology

This section explains the whole framework and approach that were used in constructing the hybrid parameter selection model. It comprised of five main processes starting with data pre-processing, data decomposition, feature selection, and result

generation and data recommendation. The proposed model aims to solve two main issues: (i) high dimensional data and (ii) uncertainty and inconsistency data. This model could become a comprehensive guideline for selecting the most optimal data set that will be used in the big data analysis and result generation tasks. Fig. 1 indicates the overall view of the proposed model.

The focused area of the proposed model is at data decomposition and feature selection phases. The model is started with a data preparation phase. This phase is used to clean, formatting, normalizing and randomizing data. The data preparation phase is essential to ensure the data that will go through the data decomposition and feature selection processes is cleaned and following the necessary parameterization tools' format. The cleaning and formatting processes are depending on the characteristics and the composition of the selected data itself. Data normalization and data randomization are executed to prepare the data into a small range of data values such as from 0–1 and in random order. The output of the data preparation phase or Phase 1 is a cleaned, formatted, normalized, and randomized data set.

The second phase of the proposed model is the data decomposition phase. The purpose of the data decomposition phase is to decompose the data into several parts of the group. However, the data will be decomposed only when the size of the data is too large to be processed by a single processing method. Therefore, each of the data set needs to go through this phase for the size identification and data reduction processes. There are two conditions to be tested, (i) either the data is more than 10,000 and (ii) either the data is less than 10,000. 10,000 representing the number of instances and also the number of features. This condition has been tested and applied in previous work [4], which is inspired by the speculative data decomposition technique proposed in [48]. For condition 1, if the size of data is more than 10,000, either instance or attribute, this data needs to go through the splitting process (SP). The splitting process usually being implemented in the parallel processing task to increase the processing speed and to decrease the processing time. The splitting process will start with the instance splitting process, then followed by the attribute splitting process.

The process of data decomposition is defined as follows:

Let X be defined as the number of groups and Y as several data.

$$X = (Y/10,000) \tag{2}$$

If X contains remainder, then

$$X = X + 1 \tag{3}$$

where the number of groups will be added to 1.

Example 1. Y = number of rows in the data set, let say, 30,000 rows. X = (30,000/10,000) = 3.3 represents the number of splitting groups that need to go through the data decomposition phase.

Example 2. Y = number of rows in the data set, let say, 37,000 rows. X = (37,000/10,000) = 3.7 which has remainder. Then, one (1) will be added to X. X = 4 represents the number of splitting groups that need to go through the data decomposition phase.

The output of the decomposition process is a group of data set that consists of instance and attribute less than 10,000 and will be labeled as SP(1) until SP(n). Based on literature works and previous experiments, the number of attributes of a data set that used to test the proposed model was less than the number of instances in the data set. Normally, the number of attributes is less than 10,000. Moreover, most of the parameterization tools and algorithms were unable to process the data set if the size is more than 10,000, especially when analyzing the data set using a non-high-performance computer. This task combining an optimistic and owner compute rule approach to split the data to 10,000 instances. It is assumed that the relationship between instances is independent with each other. For condition 2, if the data contains less than 10,000 instances and features, the decomposition process does not need to be executed. This data only needs to go through the attribute reduction process by using a hybrid reduction method.

The proposed hybrid method is comprising of a correlation-based feature selection (CFS) method as an attribute evaluator and best-first search (BFS) as the attribute searching method. This hybrid method will identify the most important attribute to be set as the most optimized attribute set (OAS). The OAS then will go through the attribute reduction process conducted by hybrid CFS and BFS reduction method. The detailed process of the attribute reduction is illustrated in Fig. 2. All the outputs from the hybrid CFS and BFS reduction process for each SP group will be analyzed first before being integrated for the next process. The analysis process is done to identify the number of an optimized attribute for each SP group and to select the highest number of optimized attributes among the SP groups. If more than one SP group has the highest optimized attribute, then the first SP group will be selected. Algorithm 1 presents the process identification of the best SP or the most optimized attribute set (OAS). Given in the algorithm, the input data are from the list of output, SP_1 until SP_n which are presented by R_1 until R_n. As mentioned earlier, SP contains a set of attributes that resulted from the hybrid attribute reduction process (CFSBFS). The output of this process is the most optimized attribute set (OAS).

Algorithm 1: The most optimized attribute set searching algorithm

```
Input: Optimized reduct sets, R1 until Rn
Output: The most optimal reduct set
if Reduct set R has more than one value then
    Select the highest number of attribute values, HR if
    HR does not have the same number with attribute
    value AND HR has more than one value then
        Select the first reduction set, FR of attribute
        values
    else
        Proceed to the next process
else
    Proceed to the next process
```

Definition 3.1. Given $R_1, R_2, \ldots, R_n \in R$ where R is a collection of optimal reduct sets generated by CFSBFS attribute reduction process. Let HR being assigned as the highest number of attribute values where $HR > R_n$ and n is a number of reducts in R and if $HR \neq AV$, attribute values AND $HR > 1$ then $HR = HR_1$.

Example 3. Given the R as a collection of optimized reduct sets $R = \{7, 6, 5, 4\}$ and $HR = 7$ where 7 is the highest number of attribute value.

Example 4. Given the R as a collection of optimized reduct sets $R = \{7, 7, 6, 5\}$ and HR has two values where $HR_1 = 7$ and $HR_2 = 7$. Therefore, select $HR = HR_1$ as the optimized attribute set.

The output of Phase 2 will be used as an input to Phase 3. In this phase, the optimized data set will be going through another parameterization process. This phase focuses on identifying the uncertainty and inconsistency values in the data set by using hybrid mathematical methods, which are soft set and rough set

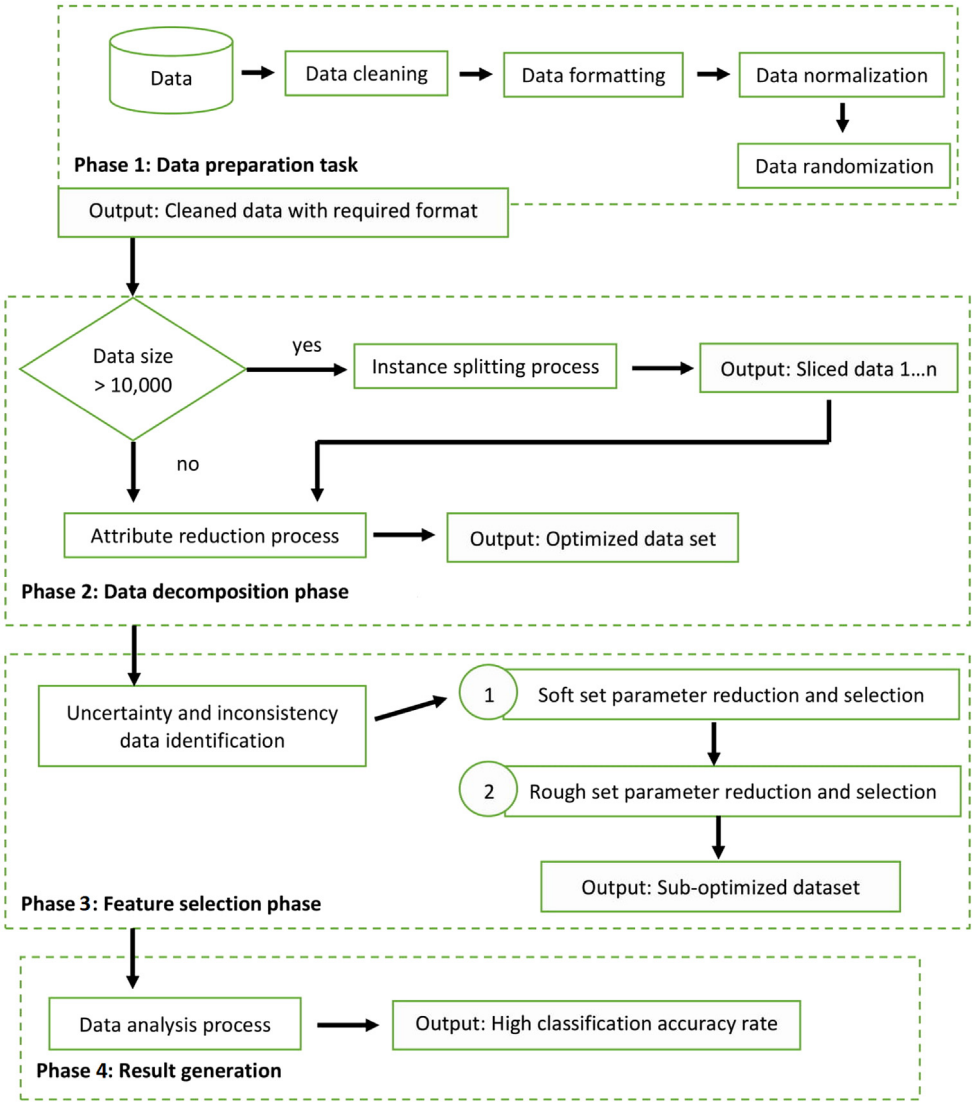

Fig. 1. The architecture of hybrid model.

parameter reduction methods. These vague values then will be eliminated to generate the most optimized data set that cleaned from uncertainty and inconsistency problems. The process of elimination consists of two phases that start with a soft set parameter reduction and selection process and then followed by a rough set parameter reduction and selection process. The hybrid method that used in Phase 3 is labeled as SSRS, which stands for soft-set rough set parameter selection (SSRS) method. The reason for having a double reduction and selection process is because of the credibility of both approaches in generating the optimized and sub-optimized data set.

From the previous experiments and literature review, the output of the soft set parameter selection and selection process unable to produce an optimized data set. This method tends to select all available attributes from the data set as the output of the parameterization process and assume all attributes are important to be analyzed. This issue might cause a problem when dealing with big data, that possibility has a lot of uncertainty values. Therefore, to overcome the weakness of the soft set selection process, a rough set parameter selection method will be implemented as a second selection method. The rough set parameter selection method also will become an examiner to validate the output that has been generated by the soft set parameter selection method. The rough set parameter selection method will re-analyze the data set to identify the uncertainty and inconsistency values and generates the most optimized data set that will be used as an input to the next data analysis process. The processes of Phase 3 are illustrated in Figs. 3 and 4 and defined by Algorithm 2 and Algorithm 3.

The most optimized data that is cleaned from uncertainty and inconsistency value is an output of Phase 3. This output will be an input to Phase 4, which is result generation. In Phase 4, the data will go through a detailed analysis, such as classification, prediction, and regression processes. The performance of the proposed hybrid parameterization model can be evaluated if the obtained results achieved 100% or nearly to 100% of the accuracy rate. The results will indicate the capability of the proposed model, whether it can successfully handle big data, uncertainty, and inconsistency data sets or not. The whole phases of data analysis will be ended with Phase 5, defined as a data recommendation phase that concludes the overall data analysis process by providing a summary of the data set. The decision-maker will be recommended whether the data set is suitable to be employed for data analysis or not. The recommendations are based on the obtained results in terms of accuracy and preciseness of the data analysis method in the analysis process. The decision-maker

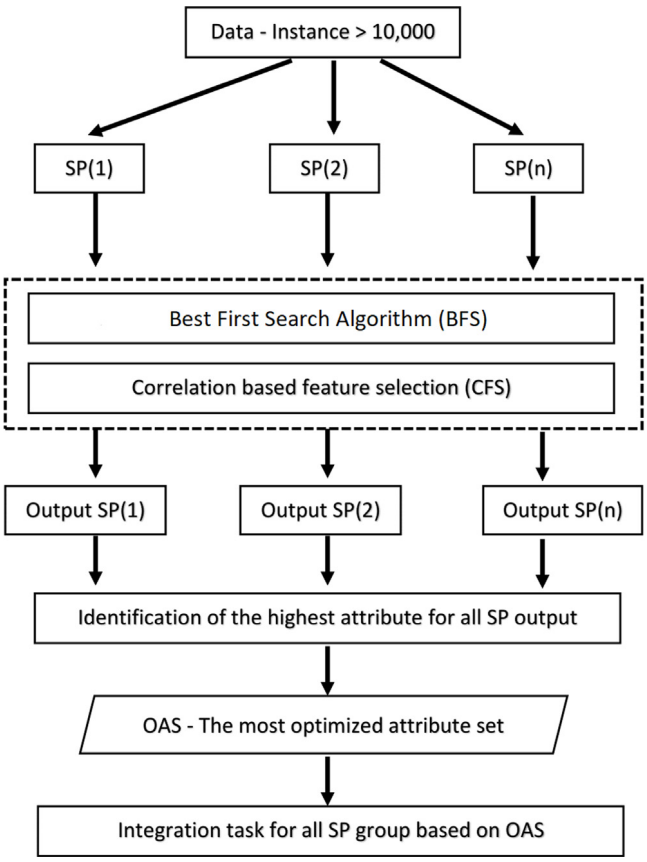

Fig. 2. Data decomposition phase.

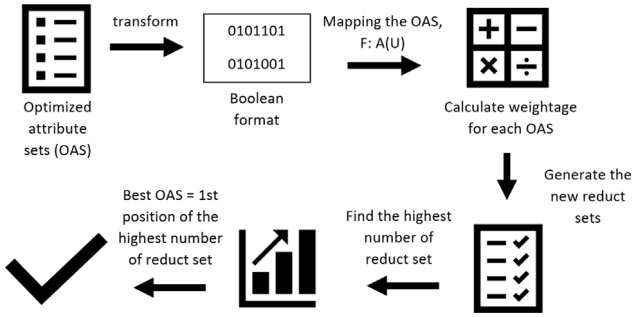

Fig. 3. Feature selection phase (Soft set parameter selection process).

Algorithm 2: Soft set parameter reduction algorithm

1 In tabular representation, let (F, P) represent the soft set. If Q is the reduction of P, the soft set reduction set is defined as (F, Q) of the soft set (F, P) where $P \subset E$
Input: A soft set (F, E), set P
Output: Optimal decision
2 Input the set P of choice parameters.
3 Find all reducts of (F, P).
4 Select one reduct set (F, Q) of (F, P).
5 Find weighted table of soft set (F, Q) according to the decided weights.
6 Find k, for which $c_k = max\ c_i$.
▷ h_k is the optimal choice of value for the selected object. If k has more than one value, any one of the benefits could be chosen.
▷ c_i is the choice of value of an object h_i where $c_i = \sum_j h_{i,j}$ and $h_{i,j}$ is the entries in the table of the reduct soft set.

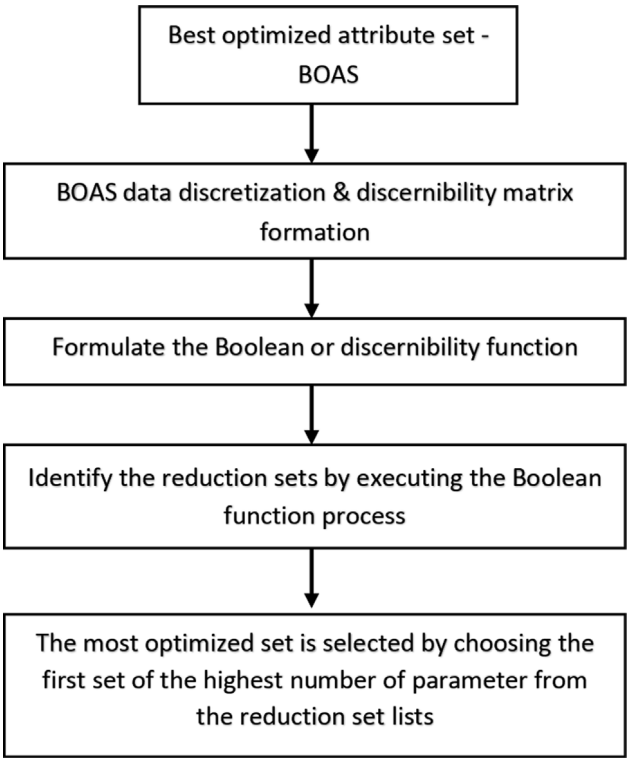

Fig. 4. Feature selection phase (Rough set parameter selection process).

needs to have excellent data in deciding the best solution for any kind of problem. The optimized data set will implicitly assist the analysis method, such as a neural network or support vector machine to generate a sound decision result.

The following mathematical formulation represents the whole architecture of the proposed work.

Definition 3.2. Set S is comprised of elements A, B, C, D. It can be represented as $S = A, B, C, D$ whereas the set S is the construction of union operation for all elements. All elements are executed sequentially one after another. The processes are defined as follows: $S = A \cup B \cup C \cup D$. S can be used in any data analysis process and any kind of data set especially for big data set.

Example 5. Let A represents Phase 1 of the data preparation task, B represents Phase 2 of the data decomposition phase, C represents the feature selection phase, and D represents the results generation phase. The whole proposed model is defined as a set S when these processes were sequentially executed one after another to construct a good data analysis process.

4. Experimental works and results

The aimed of this experimental works is to evaluate the performance of the proposed hybrid model, which is a combination of Phase 2 and Phase 3. The model will be tested whether it is capable of assisting the decision-making task in producing the best-optimized attribute set or not. The proposed model is labeled as BM1, which is a combination of CFSBFS and SSRS parameterization methods. Several experimental works that tested

Algorithm 3: Rough set parameter reduction algorithm

Input: An information system $S = (U, A, V, f)$
▷ U is a finite nonempty set object
▷ A is a finite nonempty set of attributes
▷ V is a nonempty set of values
▷ f is an information function that maps an object in U to exactly one value in V
Output: Simplified reduct sets

1 Input the information table S.
2 Discretization of data.
3 Forming up the $n \times n$ discernibility matrix. The elements of S table is defined as $d(x, y) = a \in A \mid f(x, a) \neq f(y, a)$, $d(x, y)$ is an attribute set distinguishing x and y. For each attribute $a \in A$, if $d(x, y) = a_1, a_2, ..., a_k \neq \varnothing$.
4 Formulate the Boolean function $a_1 \vee a_2 ... \vee a_k$ or discernibility function which represented by $\sum d(x, y)$ as indicated: $F(A) = \prod_{(x,y)\in U\times U)} \sum d(x, y)$.
5 If $d(x, y) = \varnothing$, constant 1 will be assigned to the Boolean function.
6 Execute the attribute reduction process based on the simplified Boolean function.
7 New optimized reduct sets are generated.

a variety of data sets had been conducted in the classification process. WEKA, Matlab, and RSES software were used to run the experimental works. Three classifiers, which are support vector machines (SVM), neural network backpropagation (NNBP), and deep learning (DL) algorithms, were applied in the classification process as these three classifiers were popularly known as a good classifier in analyzing a variety of feature values. To validate the effectiveness of the hybrid model, another two-hybrid method, which is correlation-based with a genetic algorithm (CFSGA) method labeled as BM2 and correlation-based with a greedy stepwise (CFSGS) method labeled as BM3 was implemented in Phase 2.

4.1. Data sets description

Six selected data sets named Arcene, Amazon-commerce-reviews (Amazon), Poker and human activity recognition (HAR), national classification of economic activities (CNAE), and Dota have been used to test the capability of the proposed model. These data sets had been downloaded from https://zenodo.org/record/13748 and UCI Machine Learning Repository website http://archive.ics.uci.edu/ml/index.php. These data sets were selected to test the performance of Phase 2 (Data decomposition phase) and Phase 3 (feature selection phase) in identifying the most optimal attribute to be used in the decision-making process. Three data sets, Poker, Har, and Dota, were used to test the data splitting or decomposition phase; meanwhile, Arcene, Amazon, and CNAE were used to test the feature selection phase (Phase 3). The obtained results will be presented by using the performance measurements such as accuracy rate, precision, recall, F-measure, and Kappa statistic rates. The characteristics of the data sets are shown in Table 1.

4.2. Benchmark models

The performance of the proposed parameterization model which is CFSBFS with SSRS, was validated by using the other two benchmark models; CFSGA with SSRS and CFSGS with SSRS. CFS-GA is a combination of the CFS method with a genetic algorithm meanwhile, CFS-GS is a combination of CFS with the genetic search. The validation process is conducted to identify the most performed model among these three constructed parameterization models. Besides, three well-known classifiers; (i) support vector machine (SVM), (ii) neural network-back propagation (NNBP), and (iii) deep learning (DL) have been employed in the data analysis process. As in previous work, a neural network with backpropagation classifier has shown outstanding performance in data analysis work. Therefore, another two outstanding classifiers have been chosen to validate the performance of neural network backpropagation. These three classifiers were being compared with each other to identify the most outstanding classifier in analyzing the selected data sets in the classification process.

Table 1
Description of data sets.

Data sets	Number of instances	Number of attributes	Attribute Characteristics
Arcene	200	10001	Real
Amazon	1500	10001	Real
Poker	1025009	11	Integer, Real
HAR	10229	562	Real
CNAE	1080	857	Integer
Dota	92650	117	Integer

4.3. Results

The results are analyzed according to the output generated by each phase (Phase 2 and Phase 3). Each of the phases will be evaluated based on the number of optimized attributes or parameters that had been selected. The performance of the decision analysis process relies on the number of attributes to make a good decision. An optimized attribute of a data set might assist the decision analysis method in returning significant and also a good accuracy rate.

4.3.1. Results on parameterization process

The data sets have gone through double parameterization processes. The first parameterization process was conducted in Phase 2 used to reduce the number of the attribute (column) by using either CFSBFS, CFSGA, or CFSGS. It was conducted to identify the correlation between one attribute with other attributes. The second parameterization process (Phase 3) is conducted to eliminate the uncertainty and inconsistency of attribute values in the data set. Table 2 depicted the number of reduced attributes starting from Phase 2 until Phase 3.

The output of Phase 3 is the number of the best-optimized attribute (BOAS) set from the overall parameterization process. The BOAS then will be an input to the classification process. The result of this phase is essential to the next process, whereas it helps to identify the best attribute set that will be used in the decision analysis process. Besides, both parameterization processes help to reduce the processing time and memory space, especially when using a non-high-performance computer.

As depicted in Table 2, the number of BOAS is reduced from the larger size to the smaller size for all models (BM1, BM2, and BM3). It can be seen that after the second parameterization process (SSRS), all attributes were being reduced drastically. It shows that most of the data sets consisted of uncertainty and inconsistency values. The significant of the parameterization process can be proved by looking at the classification process.

Table 2
Results on parameterization process.

Data sets	Attributes	Decomposed	BM1		BM2		BM3	
			CFSBF	SSRS	CFSGA	SSRS	CFSGS	SSRS
Arcene	10001	No	76	4	4298	3	74	3
Amazon	10001	No	41	16	3642	9	41	9
Poker	11	Yes	5	5	5	5	5	5
Har	562	Yes	57	9	265	9	57	9
CNAE	857	No	28	28	309	95	28	5
Dota	117	Yes	20	20	53	53	22	22

Table 3
Results on classification process — Accuracy rate (%).

Data sets	Without PM			BM1			BM2			BM3		
	SVM	NNBP	DL	SVM	NNBP	DL	SVM	NNBP	DL	SVM	NNBP	DL
Arcene	50	NA	78.3	56	68.5	64.5	56.5	69.5	72.5	66	63	63.5
Amazon	36	NA	56.7	16.3	27.5	30.7	9.3	7.8	8.9	16.3	27.5	30.7
Poker	55.8	49.6	48.9	59.1	54.3	48.5	59.1	54.3	48.5	59.1	54.3	48.5
Har	94.9	NA	97.3	83.6	71.4	80.4	61.6	60	57	83.6	71.4	80.4
CNAE	0	NA	NA	78.15	76.6	82	74.2	72.5	85.8	*NA*	*NA*	*NA*
Dota	74.4	72.1	NA	58	58.2	56.6	98.1	98.1	0.93	58.4	58.5	56.8

4.3.2. Results on classification process

In order to validate the efficiency of all hybrid models, the best optimized attribute sets generated from all models were tested in the classification process. Table 3 presented the accuracy rate in percentage value. As can be seen, all hybrid models achieved more than 50% of the classification accuracy rate except for the Amazon data set. All models unsuccessfully helped the classifiers to classify the amazon data set where the accuracy rate only achieved 9.3% to 30.7%. These low results may be caused by the data set itself which has potential duplicates and improper order of data set structure due to the large volume of the attribute set.

Surprisingly, both of the proposed model and BM 3 (CFSGS) had performed quite well in classifying Har data set with SVM and deep learning classifier. Both models returned 83.6% for SVM and 80.4% for deep learning but 71.4% for NNBP. As can be seen, the BM 3 model is failed to help all the classifiers to classify CNAE data set. The results are represented using *NA* and this situation happened when the reduction set generated by BM 3 is not suitable or applicable to be used in the classification task. The results also have shown that only BM 2 (CFSGA) is successfully classified the Dota data set with SVM and NNBP classifiers but not Deep learning classifier. These results indicated that even though the number of best-optimized attribute set is similar among the models, but the value that represents the selected attribute may influence the data analysis process.

Instead of validating the results of the proposed model with other benchmark models, an experiment on all original data sets were also being conducted. The results have shown that most of the data sets were unable to be classified by the classifiers, especially by NNBP. The problem of NNBP was having multiple layers of network that required long processing time and enough memory to execute the analysis process. However, Har data set has been successfully classified by SVM and DL classifiers without applying any parameterization method compared to the other three models. Table 4 indicated the processing time required by the classifiers with the proposed model and without any parameterization model. The difference processing time between the proposed model and without any parameterization model has shown that the parameterization process that consists of data decomposition and parameter selection is required to decrease the processing time and memory usage. The NA or 0 results indicated that whether the classifier was unable to execute the data set or it took a long time to analyze the data set.

Table 4
Processing time in second.

Data sets	Without PM			Proposed work		
	SVM	NNBP	DL	SVM	NNBP	DL
Arcene	0.5	NA	0.03	0	0	0.01
Amazon	16.94	NA	0.4	0.15	0.03	0.1
Poker	1.3	0.02	0.2	2.34	0.01	0.35
Har	22.21	NA	0.39	2.32	0.05	0.37
CNAE	NA	NA	0.08	0	0.02	0.04
Dota	9.94	0.15	NA	18.87	0.1	0.73

4.4. Discussion on the proposed model

Fig. 5 concludes the average classification performance on all data sets (a) and for all models (b). Har data set has returned 72.1%, which is the highest accuracy rate among the other data sets. Meanwhile, the proposed model with deep learning classifier has performed well among the other models, which returned 60.5%. The average results were obtained and showed that the classification results were imbalance among all data sets. Thus, to evaluate the performance of the proposed model, another evaluation measure has been implemented. We conclude that the proposed model is significant to be applied as a parameterization model compared with other bench-marking models. Therefore, precision, recall, and F-measure were applied to analyze the possibility that might occur within the data sets. The F-measure score presents a coherent value between precision and recall. The score that reaches to 1 indicates the better performance [49].

Table 5 indicate the precision (P), recall (R), and F-measure (F-M) of the proposed model for each classifier towards all data sets. The F-measure score indicates two data sets, Amazon and Poker have low rates among the other four data sets. It shows that the proposed model was able to help the classifier to classify the large data sets except for Amazon and Poker data sets when it returned more than 0.5% score on all data sets for all classifiers. Table 5 also indicates that the proposed model can identify all the related instances in all selected data sets except for Poker and Amazon data sets. Moreover, according to the precision value for all data sets, it shows that the proposed model with a combination of NNBP and Deep learning was able to identify the actual data set precisely. Tables 6 and 7 indicate the F-measure score for other bench-marking models (BM2 and BM3). As denoted at both tables, both bench-marking models failed to assist all the classifiers in classifying Amazon and Poker. BM2 also failed with the Dota data set and BM 3 is failed with the CNAE data set.

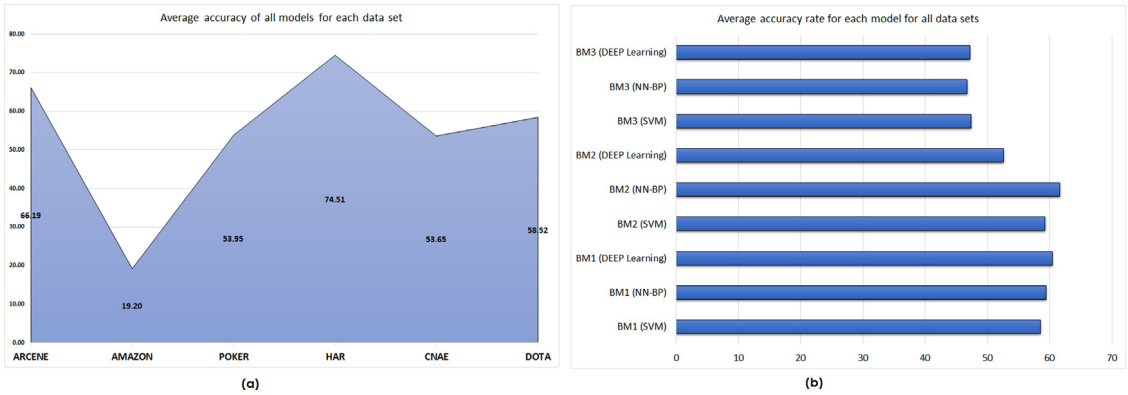

Fig. 5. Average of classification performance on all data sets (a) and for all models (b).

Table 5
Results on precision, recall and F-measure for the proposed model (CFS-BFS).

Data sets	SVM			NNBP			DL		
	P	R	F-M	P	R	F-M	P	R	F-M
Arcene	0	0.56	0	0.68	0.68	0.68	0.67	0.64	0.64
Amazon	0.18	0.16	0.17	0.02	0.26	0.27	0	0.31	0
Poker	0	0.59	0	0	0.54	0	0	0.48	0
Har	0	0.83	0.83	0.73	0.71	0.68	0.80	0.8	0.8
CNAE	0.80	0.78	0.79	0.80	0.76	0.76	0.84	0.82	0.82
Dota	0.58	0.58	0.57	0.58	0.58	0.58	0.56	0.57	0.56

Table 6
F-measure score for BM2 (CFS-GA).

Data sets	SVM	NNBP	DL
Arcene	0.437	0.7	0.73
Amazon	0.09	0.08	0
Poker	0	0	0
Har	0.61	0.6	0.6
CNAE	0.8	0.7	0.9
Dota	0	0	0

Table 7
F-measure score for BM3 (CFS-GS).

Data sets	SVM	NNBP	DL
Arcene	0.61	0.62	0.64
Amazon	0.17	0.26	0
Poker	0	0	0
Har	0.83	0.68	0.8
CNAE	0	0	0
Dota	0.6	0.6	0.6

Table 8
Kappa statistic score for BM1 (CFS-BFS).

Data sets	SVM	NNBP	DL
Arcene	0	0.35	0.31
Amazon	0.15	0.27	0.29
Poker	0.23	0.12	0.15
Har	0.8	0.65	0.76
CNAE	0.8	0.73	0.8
Dota	0.14	0.15	0.12

Table 9
Kappa statistic score for BM2 (CFS-GA).

Data sets	SVM	NNBP	DL
Arcene	0.02	0.39	0.45
Amazon	0.07	0.06	0.07
Poker	0.23	0.12	0.15
Har	0.54	0.52	0.48
CNAE	0.7	0.7	0.8
Dota	0	0	0

4.5. Analysis of used data sets

As discussed in the previous section, Poker and Amazon had returned bad classification results for all parameterization models were only reached less than 60% accuracy rate. However, the accuracy rate obtained for the Poker data set is higher when compared to other research works such as in [50] and most of the research works did not report the result because of a low rate of the classification accuracy [51]. Another evaluation measure which is called Kappa statistics, has been applied to evaluate the correlation coefficient of the data sets used. The value that reaches to 1 shows the strong relationship between the class and the attribute [52,53]. Tables 8–10 present the Kappa statistic for all models on each data sets. As discussed previously, only Har and CNAE data sets have a high correlation between the class and the attribute. However not for model BM3 (CFS-GS) where this model was unsuccessfully identifying the most optimized attribute during the parameterization process which leads to wrong interpretation of data during the classification process.

Overall, three conclusions can be made from the conducted experimental works. Firstly, the number of attributes and instance need to be considered before performing the data analysis process because it will increase the processing time and burned the memory used. Secondly, the characteristics and values of the data set need to be identified earlier because it may result in a low correlation between class and attribute. Thirdly, choose the appropriate method suitable for the data that will be used in the decision-making task because it might generate wrong or inappropriate results.

4.6. Benchmark on related works

Several existing works that applied the same data sets have been used to validate the performance of the proposed work. Table 11 presents the results between the proposed work and the existing works. Work 1 referred to the work proposed by Wang et al. that constructed a randomly partitioned and a Principal Component Analysis (PCA)-partitioned multivariate decision

Table 10
Kappa statistic score for BM3 (CFS-GS).

Data sets	SVM	NNBP	DL
Arcene	0.26	0.23	0.23
Amazon	0.15	0.27	0.29
Poker	0.23	0.12	0.15
Har	0.8	0.65	0.76
CNAE	NA	NA	NA
Dota	0.16	0.16	0.13

Table 11
Accuracy rate on each work.

Data sets	Proposed work	Work 1	Work 2	Work 3
Poker	59.1	54.3	55.1	53.9

tree classifiers for large scale data sets [54]. Meanwhile, Work 2 represents the work proposed by García-Gil et al. that combined principal component analysis (PCA) with Random Discretization (RD) methods [55] for big data sets. Work 3 referred to the work proposed by Maillo et al. which extended the capability of k-Nearest Neighbors with iterative Spark-based architecture in big data sets [56]. As depicted, the proposed work performed well compared to the other three related works. According to the results, it showed that the proposed work was significant to be implemented in analyzing big data sets. In contrast, the obtained result was comparable with the other results, especially with Work 3 that achieves high performance with distributed architecture computers.

5. Conclusion

The decision-making process for big data sets requires a lot of effort, starting from the data collection process until the best results are made. A huge cost is needed to acquire the best hardware, software, and manpower in dealing with big data. Many research works had been done by experts from multi-disciplines and industries to investigate and produce the best approach or method or tools that can be applied for big data processing. Machine learning algorithms and probabilistic theories are some of the preferred methods to be used as a parameterization method. A lot of models and approaches had been proposed to overcome different data issues. Some of the popularly known tools and methods that are beneficial to big data processing are Apache Hadoop, Apache Spark, Apache Storm, Apache Cassandra, MongoDB, R Programming Environment and Neo4j (https://towardsdatascience.com/8-open-source-big-data-tools-to-use-in-2018-e35cab47ca1d).

Inspired by these tools and technologies, this paper presents the analysis of the hybrid parameterization model by constructing several machine learning algorithms in handling big data. This paper also focused on two prominent data characteristics, which are volume and variety. To select the best machine learning method to be applied in the hybrid model, several experimental works have been conducted. As stated in the results and discussion section, the performance of the proposed model entirely is better compared to other bench-marking models. It is proved that the proposed model can be implemented in handling big data with uncertain and inconsistent problems. This model also demonstrated that the large volume of data could be divided into several groups without affecting the relationship between the class and the attribute instance. However, the obtained results were quite low due to imbalance and uncorrelated data sets. Two main factors have been identified that might cause low performance. First, the selections of an attribute during the parameterization process and second, the value or type of the data set itself. Future works need to be implemented by analyzing a balanced data set to avoid a high error rate and low classification accuracy rate.

Acknowledgments

This research has been funded by Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876 and the Fundamental Research Grant Scheme (FRGS) Vot 5F073 supported under Ministry of Education Malaysia. The work is partially supported by the project of Excellence, University of Hradec Kralove, FIM, Czech Republic (ID: 220X–2020). We are also grateful for the support of Ph.D. student Sebastien Mambou in consultations regarding application aspects.

References

[1] D. Kumar, R. Rengasamy, Parameterization reduction using soft set theory for better decision making, in: Pattern Recognition, Informatics and Mobile Engineering, 2013, pp. 3–5.
[2] N. Anitha, G. Keerthika, A framework for medical image classification using soft set, Curr. Trends Eng. Technol. (2014).
[3] M. Mohamad, A. Selamat, Analysis on hybrid dominance-based rough set parameterization using private financial initiative unitary charges data, in: LNAI Asian Conference on Intelligent Information and Database Systems, Springer, Cham, 2018, pp. 318–328.
[4] M. Mohamad, A. Selamat, A two-tier hybrid parameterization framework for effective data classification, in: New Trends in Intelligent Software Methodologies, Tools and Techniques, Vol. 303, IOS Press, 2018, pp. 321–331.
[5] Y. Liu, Y. Zhang, J. Ling, Z. Liu, Secure and fine-grained access control on e-healthcare records in mobile cloud computing, Future Gener. Comput. Syst. 78 (2018) 1020–1026.
[6] S.B.A. Kamaruddin, N.A.M. Ghani, N.M. Ramli, Best forecasting models for private financial initiative unitary charges data of east coast and southern regions in peninsular Malaysia, Int. J. Econ. Stat. 2 (2014) 119–127.
[7] A. Ahmad, M. Khan, A. Paul, S. Din, M.M. Rathore, G. Jeon, G.S. Choi, Toward modeling and optimization of features selection in Big Data based social Internet of Things, Future Gener. Comput. Syst. 82 (2017) 715–726.
[8] P. Sawicki, J. Żak, The application of dominance-based rough sets theory for the evaluation of transportation systems, Proc. Soc. Behav. Sci. 111 (2014) 1238–1248.
[9] M. Cecconello, S. Conroy, D. Marocco, F. Moro, B. Esposito, Neural network implementation for ITER neutron emissivity profile recognition, Fusion Eng. Des. 123 (2016) 637–640.
[10] L. Wang, Y. Wang, Q. Chang, Feature selection methods for big data bioinformatics: A survey from the search perspective, Methods 111 (2016) 21–31.
[11] M.I. Pramanik, R.Y. Lau, H. Demirkan, M.A.K. Azad, Smart health: Big data enabled health paradigm within smart cities, Expert Syst. Appl. 87 (2017) 370–383.
[12] K.Y. Shen, S.K. Hu, G.H. Tzeng, Financial modeling and improvement planning for the life insurance industry by using a rough knowledge based hybrid MCDM model, Inform. Sci. 375 (2017) 296–313.
[13] M. Esposito, A. Minutolo, R. Megna, M. Forastiere, M. Magliulo, G. De Pietro, A smart mobile, self-configuring, context-aware architecture for personal health monitoring, Eng. Appl. Artif. Intell. 67 (2018) 136–156.
[14] X. Ma, Q. Liu, J. Zhan, A survey of decision making methods based on certain hybrid soft set models, Artif. Intell. Rev. 47 (2017) 507–530.
[15] N. Allias, M.N. Megat, N. Megat, M.N. Ismail, A hybrid gini PSO-SVM feature selection based on Taguchi method : An evaluation on email filtering, in: Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, ACM, 2014, pp. 55–59, http://dx.doi.org/10.1145/2557977.2557999.
[16] Z. Masetic, A. Subasi, Congestive heart failure detection using random forest classifier, Comput. Methods Programs Biomed. 130 (2016) 54–64.
[17] B. Ait Hammou, A. Ait Lahcen, S. Mouline, APRA: An approximate parallel recommendation algorithm for Big Data, Knowl.-Based Syst. 157 (2018) 10–19.
[18] M. Mohamad, A. Selamat, A new soft rough set parameter reduction method for an effective decision-making, in: New Trends in Intelligent Software Methodologies, Tools and Techniques, Vol. 297, IOS Press, 2017, pp. 691–704.
[19] A. Hassani, S.A. Gahnouchi, A framework for business process data management based on big data approach, Procedia Comput. Sci. (2017).
[20] Y.-C. Ko, H. Fujita, An evidential analytics for buried information in big data samples: Case study of semiconductor manufacturing, Inform. Sci. 486 (2019) 190–203, http://dx.doi.org/10.1016/j.ins.2019.01.079, http://www.sciencedirect.com/science/article/pii/S002002551930057X.

[21] J. Luo, H. Fujita, Y. Yao, K. Qin, On modeling similarity and three-way decision under incomplete information in rough set theory, Knowl.-Based Syst. (2019) 105251, http://dx.doi.org/10.1016/j.knosys.2019.105251, http://www.sciencedirect.com/science/article/pii/S0950705119305635.
[22] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, Hypotheses analysis and assessment in counter-terrorism activities: a method based on OWA and fuzzy probabilistic rough sets, IEEE Trans. Fuzzy Syst. (2019) 1, http://dx.doi.org/10.1109/TFUZZ.2019.2955047.
[23] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, Improving awareness in early stages of security analysis: A zone partition method based on GrC, Appl. Intell. 49 (2018) 1063–1077.
[24] H. Fujita, A. Gaeta, V. Loia, F. Orciuoli, Resilience analysis of critical infrastructures: A cognitive approach based on granular computing, IEEE Trans. Cybern. 49 (5) (2019) 1835–1848, http://dx.doi.org/10.1109/TCYB.2018.2815178.
[25] J. Akoka, I. Comyn-Wattiau, N. Laoufi, Research on big data – A systematic mapping study, Comput. Stand. Interfaces 54 (2017) 105–115.
[26] L. Koc, T.a. Mazzuchi, S. Sarkani, A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier, Expert Syst. Appl. 39 (18) (2012) 13492–13500.
[27] S. Chebrolua, S.G. Sanjeevi, Attribute reduction in decision-theoretic rough set model using particle swarm optimization with the threshold parameters determined using LMS training rule, Knowl.-Based Syst. 57 (2015) 527–536.
[28] O.S. Soliman, A. Rassem, Correlation based feature selection using quantum bio inspired estimation of distribution algorithm, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNAI, vol. 7694, 2012, pp. 318–329.
[29] N.F. Abubacker, A. Azman, S. Doraisamy, Correlation-based feature selection for association rule mining in semantic annotation of mammographic, Pattern Recognit. Lett. 32 (2011) 482–493.
[30] S. Chormunge, S. Jena, Correlation based feature selection with clustering for high dimensional data, J. Electr. Syst. Inf. Technol. (2018) 4–11.
[31] D. Molodtsov, Soft set theory-first results, Comput. Math. Appl. 37 (4) (1999) 19–31.
[32] J. Chai, E.W.T. Ngai, J.N.K. Liu, Dynamic tolerant skyline operation for decision making, Expert Syst. Appl. 41 (15) (2014) 6890–6903.
[33] Y. Liu, K. Qin, L. Martínez, Improving decision making approaches based on fuzzy soft sets and rough soft sets, Appl. Soft Comput. J. 65 (2018) 320–332.
[34] X. Ma, N. Sulaiman, H. Qin, T. Herawan, J.M. Zain, A new efficient normal parameter reduction algorithm of soft sets, Comput. Math. Appl. 62 (2) (2011) 588–598.
[35] F. Feng, X. Liu, V. Leoreanu-Fotea, Y.B. Jun, Soft sets and soft rough sets, Inform. Sci. 181 (6) (2011) 1125–1137.
[36] M. Irfan Ali, A note on soft sets, rough soft sets and fuzzy soft sets, Appl. Soft Comput. J. 11 (4) (2011) 3329–3332.
[37] M. Mohamad, A. Selamat, Recent study on the application of hybrid rough set and soft set theories in decision analysis process, in: Lecture Notes in Artificial Intelligent, LNAI, 9799, 2016, pp. 713–724.
[38] M. Mohamad, A. Selamat, A new hybrid rough set and soft set parameter reduction method for spam e-mail classification task, in: Lecture Notes in Artificial Intelligent, LNAI, 9806, 2016, pp. 18–30.
[39] Z. Pawlak, Rough set approach to knowledge-based decision support, European J. Oper. Res. 99 (1997) 48–57.
[40] Local rough set: A solution to rough data analysis in big data, Internat. J. Approx. Reason. 97 (2018) 38–63, http://www.sciencedirect.com/science/article/pii/S0888613X17304826.
[41] A. Oussous, F.Z. Benjelloun, A. Ait Lahcen, S. Belfkih, Big data technologies: A survey, J. King Saud Univ. Comput. Inf. Sci. 30 (2018) 431–448.
[42] T.K. Sheeja, A.S. Kuriakose, A novel feature selection method using fuzzy rough sets, Comput. Ind. 97 (2018) 111–121.
[43] J. Liu, Y. Lin, Y. Li, W. Weng, S. Wu, Online multi-label streaming feature selection based on neighborhood rough set, Comput. Ind. 84 (2018) 273–287.
[44] B. Huang, Y.L. Zhuang, H.X. Li, D.K. Wei, A dominance intuitionistic fuzzy-rough set approach and its applications, Appl. Math. Model. 37 (12–13) (2013) 7128–7141.
[45] W.S. Du, B.Q. Hu, Dominance-based rough set approach to incomplete ordered information systems, Inform. Sci. 346–347 (2016) 106–129.
[46] S. Greco, B. Matarazzo, R. Slowi, Algebra and topology for dominance-based rough set approach, in: Z.W. Ras, L.-S. Tsay (Eds.), Advances in Intelligent Information Systems, Springer, 2010, pp. 43–78.
[47] M.I. Ali, B. Davvaz, M. Shabir, Some properties of generalized rough sets, Inform. Sci. 224 (2013) 170–179.
[48] A. Grama, A. Gupta, G. Karypis, V. Kumar, Principles of parallel algorithm design, in: Introduction to Parallel Computing, second ed., Addison Wesley, Harlow, 2003.
[49] H. Li, D. Li, Y. Zhai, S. Wang, J. Zhang, A novel attribute reduction approach for multi-label data based on rough set theory, Inform. Sci. 367–368 (2016) 827–847.
[50] I. Triguero, D. Peralta, J. Bacardit, S. García, F. Herrera, MRPR: A MapReduce solution for prototype reduction in big data classification, Neurocomputing 150 (2015) 331–345.
[51] A. Arnaiz-Gonzalez, J.F. Diez-Pastor, J.J. Rodriguez, C. Garcia-Osorio, Instance selection of linear complexity for big data, Knowl.-Based Syst. 107 (2016) 83–95.
[52] S.K. Pal, S.K. Meher, S. Dutta, Class-dependent rough-fuzzy granular space, dispersion index and classification, Pattern Recognit. 45 (7) (2012) 2690–2707.
[53] G.R. Teixeira de Lima, S. Stephany, A new classification approach for detecting severe weather patterns, Comput. Geosci. 57 (2013) 158–165.
[54] F. Wang, Q. Wang, F. Nie, W. Yu, R. Wang, Efficient tree classifiers for large scale datasets, Neurocomputing 284 (2018) 70–79.
[55] D. García-Gil, S. Ramírez-Gallego, S. García, F. Herrera, Principal components analysis random discretization ensemble for big data, Knowl.-Based Syst. 150 (2018) 166–174.
[56] J. Maillo, R. Sergio, I. Triguero, F. Herrera, kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data, Knowl.-Based Syst. 117 (2017) 3–15.