Journal of King Saud University – Computer and Information Sciences 33 (2021) 787–797

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Improving the classification performance on imbalanced data sets via new hybrid parameterisation model

Masurah Mohamad [b,f], Ali Selamat [a,b,c,e,*], Imam Much Subroto [d], Ondrej Krejcar [e]

[a] Media and Games Center of Excellence (MagicX), Universiti Teknologi Malaysia, 81310 Skudai, Johor Bahru, Johor, Malaysia
[b] School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), 81310 Skudai, Johor Bahru, Johor, Malaysia
[c] Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia
[d] Universiti Islam Sultan Agung, Semarang, Indonesia
[e] University of Hradec Kralove, Rokitanskeho 62, 500 03 Hradec Kralove, Czech Republic
[f] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Tapah Road, 35400 Perak, Malaysia

ARTICLE INFO

Article history:
Received 18 November 2018
Revised 17 March 2019
Accepted 13 April 2019
Available online 24 April 2019

Keywords:
Soft set theory
Rough set theory
Parameter selection
Neural network
Hybrid method
Imbalanced data

ABSTRACT

The aim of this work is to analyse the performance of the new proposed hybrid parameterisation model in handling problematic data. Three types of problematic data will be highlighted in this paper: i) big data set, ii) uncertain and inconsistent data set and iii) imbalanced data set. The proposed hybrid model is an integration of three main phases which consist of the data decomposition, parameter reduction and parameter selection phases. Three main methods, which are soft set and rough set theories, were implemented to reduce and to select the optimised parameter set, while a neural network was used to classify the optimised data set. This proposed model can process a data set that might contain uncertain, inconsistent and imbalanced data. Therefore, one additional phase, data decomposition, was introduced and executed after the pre-processing task was completed in order to manage the big data issue. Imbalanced data sets were used to evaluate the capability of the proposed hybrid model in handling problematic data. The experimental results demonstrate that the proposed hybrid model has the potential to be implemented with any type of data set in a classification task, especially with complex data sets.

1. Introduction

In our previous work, a new hybrid parameterisation model that integrated two mathematical methods, soft set and rough set theories, was proposed. Soft set and rough set theories were selected in constructing this model because of their capability in dealing with uncertain and inconsistent data. These two models were integrated because each of them will compensate for lapses of the other. The proposed parameterisation model consists of several important phases that were used to process data before the decision-making process was executed. Some of the phases involved in the proposed work were data pre-processing, parameter reduction and parameter selection. The results show that the proposed approach can assist the selected classifier in identifying the optimal reduction set that will be used in the classification process. This paper presents an extension of the previous study by enhancing the capability of the hybrid model in identifying the most important attributes of large imbalanced data sets for the classification process. Two issues were considered and analysed in this study: the size and the type or category of the data set. Based on previous studies (Mohamad et al., 2017; Mohamad et al., 2017), the number of data sets is one of the factors that influences the experimental results. It might affect the implementation of the whole framework of the classification task.

For example, some of the parameterisation methods are unable to manage or analyse a large volume of data at one time. Not only was the parameterisation method unable to handle the data, but even the processors and the parameterisation tools faced this kind of difficulty. This issue worsened when it dealt with big data where researchers need to consider many issues, such as processing time and storage availability, which might correlate with the implementation of big data in the decision-making process (Arnaiz-González

* Corresponding author at: Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia.
E-mail address: aselamat@utm.my (A. Selamat).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2019.04.009
1319-1578/

et al., 2016). Data pre-processing and parameter reduction are among the important processes that are usually highlighted in data mining, especially in handling big data. Inconsistent and uncertain decisions might be generated if these two processes are misconducted (Houari et al., 2016).

The category of data set such as imbalance data may also lead to inconsistent and inaccurate results. The imbalanced data consisted of an uneven distribution of classes or difference in numbers (Wang et al., 2016). These data sets were comprised of different categories such as real, integer and nominal that will be explained in detail in Section 4 (Experimental work and results discussion). Biased results might be obtained when imbalanced data sets were ignored and not properly analysed during the data collection phase. Moreover, most research works suggest a preference for implementing the standard data sets which were already cleaned and balanced before or during the time the decision-making process was executed (Zhou et al., 2017). These processes might also generate biased results leading to an incorrect results analysis (Derrac et al., 2013).

Therefore, in order to execute the large imbalanced data sets, an alternative parameterisation model is proposed which is an integration of mathematical methods: soft set theory and rough set theory as a parameterisation method, and neural network as a classifier. This proposed model can accommodate uncertain and inconsistent data issues, and is also able to manage a large volume of data sets. Soft set and rough set theories were selected due to their ability in handling inconsistent and uncertain data sets. These theories have been implemented and proved in various works of research and different application fields (Luo et al., 2016; Meng et al., 2016). Most research has applied these theories to dealing with approximation problems, feature reduction, feature selection approaches or even as a classification method (Azar et al., 2016; Raza et al., 2016; Mohamad et al., 2016). This paper also presents alternative steps, that is, a SRS identifier to help the proposed model in identifying the best-optimised parameter set. The detailed explanation of these proposed steps is explained in the Methodology section.

The best-optimised parameter set will be used as an input to the decision analysis process. A classification process is a data analysis task that is normally used to test the performance of any machine learning method. The classification process is conducted using a neural network due to its capability in analysing complex data. The neural network is a well-known method that can replace human activity in dealing with complex variables and complex relationships, and has been successfully proved by many researchers such as in Lam et al. (2014) and Weng et al. (2016). The neural network is also known as an artificial neural network (ANN) that can be applied in different application areas and to problems such as forecasting, classification, optimisation and regression (Paradarami et al., 2017; Kim et al., 2017). It is hoped that the proposed work can be an alternative parameterisation model in the big data pre-processing task.

This paper consists of five sections. Section 1 introduces the highlighted issues, while Section 2 presents the background knowledge of the important topics. Section 3 explains the methodology of the proposed work and Section 4 presents the experimental results with the discussion. Finally, Section 5 concludes the proposed work based on the results analysis from the previous sections.

2. Background knowledge

This section discusses several basic concepts of related topics to provide an understanding of soft set and rough set parameter reduction approaches and the neural network technique as a good classifier. Both soft set and rough set parameter reduction approaches are approaches that deal with uncertain and unclear data. Both approaches apply mathematical concepts in identifying the important attributes or parameters, and are commonly used by the decision-maker to solve many complicated problems.

2.1. Soft set parameter reduction approach

Soft set parameter reduction is an approach that implements soft set theory as a parameterisation tool in dealing with uncertainties problems. It was initiated by Molodtsov in 1999, in order to improvise the fuzzy concept which was also used to deal with uncertainties and fuzzy problems. Molodtsov had claimed that the soft set theory was easier to understand and implement compared to fuzzy sets theory. Soft set theory also implements the theory of approximation which enables the non-mathematical expert to understand the whole structure of the theory. It is used to solve the desired problem easily by not focusing only on the mathematical part. Soft set theory does not apply any restriction in the parameterisation process because it applies an approximation approach to initialise each object. Therefore, as highlighted by Molodtsov, any type of parameterisation approach can be implemented with the assistance of numbers, functions, mapping, words and sentences (Molodtsov, 1999).

Due to the capability of soft set theory in solving the uncertainty problems, many researchers enhanced this theory by integrating it with other mathematical theories in order to solve desired problems. Some of the hybrid theories are soft rough fuzzy sets and soft fuzzy rough sets proposed by Meng et al. (2011), soft rough sets proposed by Feng et al. (2011) and multi-fuzzy soft set proposed by Yang et al. (2013). Most hybrid theories were proposed in order to generalise the functionality of the selected theories. Therefore, many researchers tended to test the ability of the soft set theory itself or its hybrid theories in different application fields especially as a parameter selection method.

The following definitions are the basic formulation of how the soft set theory works in the parameterisation process. The basic formulation was taken from Molodtsov's paper published in 1999 (Molodtsov, 1999).

Definition 2.1a. Let U represent the set of universe and E represents a set of parameters. A pair of (F, E) is defined as a set of soft set over set of universe U, when F is a mapping of the set E in all of the subsets of the set U. For $\varepsilon \epsilon A, F(\varepsilon)$ maybe considered as a set or an approximate set of the soft set (F, E). Therefore, a soft set is not a crisp set. The approximate set is comprised of different types of values such as missing values or uncertain values.

The theory then was applied and enhanced by Herawan in 2010. In Herawan et al. (2010), Herawan implemented the soft set theory by using maximal supported objects to analyse patients who were suspected of influenza. He also proposed a multi-soft sets theory of approximation for a multi-valued information system (Herawan et al., 2010). The soft set of universe can also be considered as a binary-valued information system. Therefore, the decision making process can be made by using binary-valued representation format. Definition 2.1b indicates the formulation on how the parameter in the soft set is reduced in the set of universe.

Definition 2.1b. For soft set (F, E) over the universe U and $u \epsilon U$. An object u is an optimal decision if u is maximally supported by E. This formulation is a derivation from the Definition 2.1c and Definition 2.1d.

Definition 2.1c. Definition 2.1c: Let (F, E) be the soft set over the universe U and $A \subset E$. A is defined as indispensable if $U/A = U/E$. Otherwise A is set to be dispensable. This definition was used in the parameter reduction process without modifying the set of optimal and sub-optimal decisions.

Definition 2.1d. Let soft set (F, E) be the set of universe U and $A \subseteq E$. A is a reduction set of E if and only if A is a indispensable and supported by all sets of E.

The above definitions have been used as a guideline in the parameter reduction process and also in selecting the optimal and sub-optimal parameter sets. This algorithm was modified by adding Step 4 in order to select the most-optimised at-tribute set. The implementation of these definitions and soft set parameter reduction steps are discussed in the Methodology section. The steps of the soft set parameter reduction process are as follows:

- Prepare the data set and transform into binary representation 0 and 1 format.
- Identify the reduction sets based on the attribute value.
- Calculate the weighted for each reduction set.
- Select the most optimised reduction set by choosing the highest number of attribute set.

2.2. Rough set parameter reduction approach

Rough set theory is a well-known theory which is able to manage uncertain or incomplete data effectively. Rough set was proposed by Pawlak and has similar functionality to other theories such as fuzzy sets, Bayesian inference and evidence theory (Pawlak et al., 1998). The main idea of rough set is the approximation concept that was represented by a boundary region between the upper and lower approximations. Rough set theory has been extended and generalised towards many application areas such as pattern recognition, decision analysis, image processing, inductive reasoning and machine learning (Feng et al., 2011). The following paragraph defines the basic notions of rough set theory as proposed by Pawlak.

The set of elements is defined as rough set or imprecise when the set cannot be defined or identified in the set of universe. The data has the possibility of not being a member or its companion in the set of the data.

For an information system $S = (U, A), X \subseteq U and B \subseteq A$. For every $X \subseteq U$,
B is defined as upper approximation of $X, B^*(X)$ when, $\bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \varnothing\}$and
B is defined as lower approximation of $X, B_*(X)$ when, $\bigcup_{x \in U} \{B(x) : B(x) \subseteq X\}$.

The rough set is derived from the minus operation of upper approximation and lower approximation operations.

Rough set is also known as a boundary set which is part of the membership set for any data. Thus, it is suitable for use with many situations and to solve different types of problems such as data reduction, parallel processing and identifying patterns from hidden data (Pawlak, 1997). The major concern in this work is to evaluate the ability of rough set theory in assisting soft set theory to identify the optimal set of data, especially for the large data set problem. The following steps describe the processes that are conducted during the rough set parameter reduction process. All the processes include the mathematical formula which was defined earlier. This rough set parameter reduction algorithm has been modified by adding Step 6 in order to select the most-optimised reduction set.

- Prepare the data set.
- Data discretisation.
- Forming up the *mxn* discernibility matrix.
- Calculate the discernibility function.
- Identify the reduction sets.
- Select the most-optimised reduction set by choosing the highest number of attribute set in the reduction set.

Several works of research which reported the integration between these two theories can be found in Meng et al. (2011), Montazer et al. (2015) and Ma et al. (2016). Some papers that implemented the enhancement of rough set theory as a parameter reduction and selection method can be found in Raza et al. (2016), Chen et al. (2016), Chen et al. (2016). All proposed methods showed that rough set theory could increase the performance of the decision-making process and was able to reduce the processing time in any task, especially in parameter reduction and selection processes.

2.3. Parameterisation process

The parameterisation process is normally conducted after the data pre-processing task and before data analysis is executed. The parameterisation process is comprised of two processes: parameter reduction and parameter selection. The output of the parameterisation process is an optimised attribute set which is generated by the parameterisation method. Many parameterisation methods have been proposed that are based on different approaches such as Filters, Wrappers and Embedded (Chormunge et al., 2018). Each of these approaches has its advantages and disadvantages. A proper selection of which method needs to be applied for the parameterisation process in solving a given problem should be considered.

The parameter reduction process is also known as attribute reduction or feature extraction. It is used to reduce the number of attributes of the data set according to certain criteria and specified characteristics. Normally, the uncertainty and inconsistent data set will be eliminated to avoid misinterpretation of the data analysis process. Different algorithms and methods have been proposed such as the incremental algorithm with variable precision rough sets (Chen et al., 2016) and dimension reduction using Copulas and LU-Decomposition (Houari et al., 2016).

Parameter selection is a process of selecting the most important attribute among the available attributes of the data set. Some methods are able to conduct both reduction and selection processes at one time. Parameter selection is one of the important processes before the data set is used in determining the best solution. Among the well-known methods and algorithms that are usually used for parameter selection are the SVM, random forest, decision tree, ReliefF and Fisher Score (Zhou et al., 2017; Masetic et al., 2016).

2.4. Neural network

Neural network is also known as artificial neural network (ANN), and is a machine learning technique that can accomplish various decision-making tasks. Neural network imitates the human brain in its processing task. Neural network has two features which are artificial neuron or node and node's connectivity. The artificial neuron represents an information processing unit which is the main component of the neural network process. The artificial neuron has three basic components: i) a set of connecting links from different inputs x_i called synapses that are characterised by use of a weight or strength w_{ki}, where $i = 1, 2, \ldots, n$and n is a number of input data, ii) one integrator to add the input signals X_i weighted with the synaptic strengths $W_k i$, and iii) an activation function f for limiting the amplitude of the output y_k of the neuron (Weng et al., 2016).

The most common type of neural network layer is divided into three layers: input, intermediate and output. The intermediate

layer, also known as the hidden layer, can consist of several layers and hidden nodes. Neural network can be applied to identify complex relationships between features and independent parameters, to determine the interactions of high polynomial parameters, classification, prediction and optimisation (Paradarami et al., 2017). Neural network has been widely used because of its simplicity. It is easily applied in different application areas and can return good results in solving problems (Massimiani et al., 2017). Back propagation and feed forward are examples of neural network algorithms that were frequently used in solving problems (Kim et al., 2017). Neural network was implemented in previous work and was proven capable of handling a complex data set (Mohamad et al., 2017; Mohamad et al., 2018).

2.5. Related existing works

Recently, researchers have tended to integrate more than one method to create a hybrid model. This is due to the capability of each method in handling different kinds of data problems, especially big data. Big data is comprised of different types of data sets, most of which are vague and imbalanced (Ahmad et al., 2017). These kinds of data sets really need an effective and efficient data analysis model in order to generate effective decisions. The following paragraphs describe several existing parameterisation models that deal with problematic data sets, especially big data and imbalanced data sets.

Nowadays, the most popular way to manage big data is by using the MapReduce data processing model. MapReduce is used to process and produce large data sets by implementing parallel processing in an efficient way (Triguero et al., 2015). It provides many benefits in handling big data in terms of reducing processing time and use of memory space. Some research, which implemented MapReduce, were the hierarchical attribute reduction algorithm (Qian et al., 2015), secured smart health care monitoring and alerting system (Manogaran et al., 2017) and the four layers of the architectural model for feature selection in Big Data IoT (Ahmad et al., 2017). Most research has claimed that the implementation of MapReduce in their proposed frameworks and models resulted in an improvement in the data processing performance.

Instead of implementing MapReduce in handling the big data issue, some researchers proposed hybrid models in order to solve multiple categories of data issues. The hybrid model can be defined as more than one of many models that were integrated. The hybrid models were probably proposed to overcome the weakness of, and to improve the performance of, the existing single model in handling any specified problems (Mohamad et al., 2016; Paradarami et al., 2017). To be more focused, this paper will list several hybrid models related to the highlighted data issues on imbalanced data, inconsistent and uncertain data problems. The combination of more than two models might generate complicated models which are difficult to understand and execute. Some existing hybrid models that have been proposed to solve the highlighted issues are presented in Table 1.

Table 1
Existing works on imbalanced, inconsistent and uncertain data issues.

Data issues	Existing hybrid models
Imbalanced	A systematic online banking fraud detection approach by using three algorithms: contrast pattern mining, neural network and decision forest (Wei et al., 2013), Multi-criteria optimisation classifier using fuzzification, kernel and penalty factors for predicting protein interaction hot spots (Zhang et al., 2014) and a new approach based on fuzzy rough sets and evolutionary algorithms to improve the performance of one neural network classifier (Derrac et al., 2013).
Complicated (inconsistent & uncertain)	An implementation of dominance-based neighbourhood rough sets (DNRS) to reduce the attribute of big data set using parallel processing (Chen et al., 2016), a new attribute reduction approach for multi-label data that consisted of complementary decision reduction, a discernibility matrix-based method and a heuristic algorithm (Li et al., 2016) and improved dominance-based rough set approach (IDRSA) which was proposed to handle complex and uncertain nominal attributes (Azar et al., 2016).

3. Methodology

The proposed methodology consists of several phases and sub-processes. The size of data set must be determined in the beginning of the decision analysis process. It is important to specify the size of data as not all parameter reduction methods are able to process large data at one time. All data must go through an evaluation process in order to identify the size of data set. If the size of data is more than 10,000, the data then should be decomposed, or else the data will be processed by using the hybrid soft set and rough set parameterisation model. Fig. 1 presents the framework that was applied in this study and the following sub-sections explained each specified process.

3.1. Phase 1: data sets collection

The data sets collection phase is a process of acquiring the desired input data from different resources. The following issues were considered during the data collection process: i. size of data (large data set), ii. characteristics of data (uncertain and inconsistent values), and iii. imbalanced data (in terms of data division).

3.2. Phase 2: data pre-processing

The collected data will go through several processes in preparation for the classification task. The processes are data formatting, data normalisation and data randomisation. The raw data is formatted into a required scheme according to the methods or software used during the classification task. Basically, the data is presented by using $m \times n$ matrix including the decision class at the end of the data column. The formatted data is then normalised in order to standardise the value of each column, to increase the computer processing performance, and to decrease the memory usage. In addition, the normalised data will be randomised to avoid any bias issues and to increase the accuracy rate of the classification task.

3.3. Phase 3: data decomposition

This phase is applied after the data has been through the pre-processing task and when the size of the data or instances is more than 10,000. It has been proposed as an alternative approach for processing a large size of data instead of using big data analytic tools. Processing time and the cost of operation are the factors behind implementing the slicing technique in this study. Most data processing methods require a long time to analyse a big data set and require expensive high-performance tools to process the data. The data will be fragmented into a number of groups by dividing the total number of instances by 10,000. If the calculation contains a remainder, to the number of groups will be added 1. Let G be defined as the number of groups and D as a number of data.

$$G = (D/10000) \tag{1}$$

If G contains remainder, then

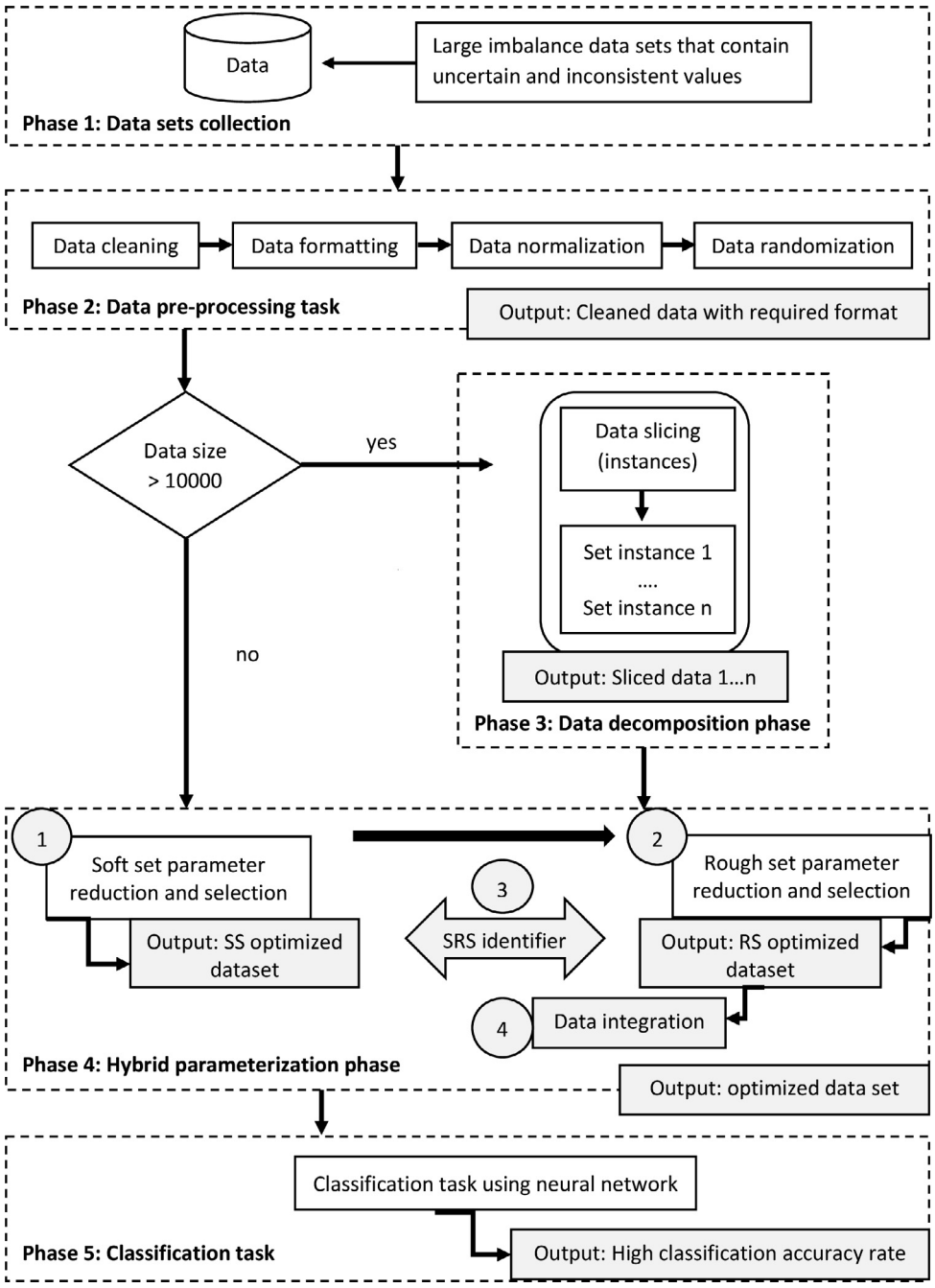

Fig. 1. Proposed framework of the hybrid model.

$$G = G + 1 \tag{2}$$

For example, if the total data is 12,000, the data needs to be divided by 10,000. The answer is 1 and it contains 2,000 as a remainder. Therefore, 1 must be added to the answer, that is, 1 + 1 = 2. In this case, the data should be divided or sliced into 2 groups where the number of instances of each group will be distributed equally.

The slicing technique will be executed only when the number of instances is more than 10,000. 10,000 is a constant value. The reason for defining the constant value as 10,000 is that most parameter reduction methods which are executed by the normal processor (not a super computer) can only manage this value or less. If more than 10,000 instances are being analysed, the processor either takes a longer time to process or is unable to be execute the process at all. If the number of instances is equal to or less than 10,000, the next process which is phase 4 (as shown in Fig. 1) will be executed. This slicing technique has been tested several times in previous experimental works (Mohamad et al., 2017; Mohamad et al., 2016).

3.4. Phase 4: soft rough set parameter reduction process

Phase 4 is a hybrid parameterisation process. It is divided into four parts where part 1 will execute the soft set parameterisation process, part 2 will execute the rough set parameterisation process, part 3 will execute the optimised parameter selection process and part 4 will execute the data integration process. Part 1 and 2 were executed after phase 3 had been accomplished. Meanwhile part 3

was executed after the both parts had been conducted. Both part 1 and part 2 produced their own optimal reduction sets. These outputs were labelled as soft set (SS) optimised data set and rough set (RS) optimised data set. These output data sets were then integrated into one set as an input to the classification process.

Both part 1 and part 2 were sequentially executed one after another in order to ensure an optimal reduction set would be generated. Then, a soft rough set (SRS) identifier was applied in order to select the best-optimised input for the classification task. Basically, the SRS identifier chose the highest number of attribute set as the set to be processed in the next phase. The algorithm for selecting the optimised reduction set which was implemented by the SRS identifier can be identified in Mohamad et al. (2017). Part 4 was only executed when the size of data was more than 10,000 as discussed in phase 3. Phase 4 was repeated a number of times according to the number of groups produced during the data decomposition phase.

3.5. Phase 5: classification task

Classification task is the last phase that needs to be executed. Cleaned and simplified data set was ready to be used as an input to the task. Any classifier can be applied to execute the classification task. The results obtained from this phase will be evaluated using several standard evaluation measures.

3.6. Proposed hybrid model

To the best of our knowledge, soft set theory and rough set theory are among the most efficient theories in handling uncertain and inconsistent data (Ma et al., 2017; Du et al., 2016). The main objective in proposing the hybrid model is to have a good parameterisation method that is capable of handling uncertain and inconsistent data. The fundamental concept of the soft set parameterisation method itself has its own weakness in generating the optimised reduction set. The algorithm proposed by Maji et al. (2002) is one which is unable to generate optimised and sub-optimised reduction sets. Meanwhile the algorithm proposed by Kong et al. has its own limitations (Ma et al., 2017). This was proven when many researchers proposed several enhancements of the soft set parameterisation method by improving the fundamental theory or by hybridising it with other theories (Mohamad et al., 2017).

This work took the initiative to overcome the weakness of the soft set theory by integrating it with rough set theory. The rough set parameter reduction method was selected to be integrated with the soft set parameter reduction algorithm. Rough set has the capability to handle uncertain and inconsistent data successfully, using the theory of approximation. The fundamental concept of rough set theory has been enhanced by various works of research such as the improved dominance-based rough set approach (Azar et al., 2016) and dominance-based neighbourhood rough sets (Chen et al., 2016).

The basic concept of how the hybrid parameterisation process was executed is illustrated using the following definitions.

> In a data classification process, $D:X$ is defined as a soft set parameter reduction process and Y is defined as a rough set parameter reduction process. Both X and Y produced an optimised reduction set which are defined as $S \leftarrow X$ and $T \leftarrow Y$.
> H is defined as a hybrid parameterisation process of D, when $S \cup T$ which produced a result Z, which is an optimised reduction set from both sets that needs to be executed in sequence. Therefore, $Z \leftarrow S \cup T$ and Z can be used in D in order to increase the classification performance.

The input of the selection process was a list of optimised reduction sets which had been generated from both the soft set and rough set parameter reduction processes. Each of the sets was evaluated based on the number of attributes that had been selected as an optimised reduction set. If the produced reduction set was more than 1 value, the SRS identifier would select the highest number of the produced attribute value among the available sets. If the produced reduction set was equal to 1, the produced reduction set would be directly used in the next process. Then, the selected reduction set which was the highest number of attributes value would be evaluated for the second evaluation process.

In the second evaluation process, two evaluation questions will be considered:

1. Does the highest attribute value is equal to the number of attributes of the original data set?
2. Does the highest attribute value has more than one reduction set?

If both conditions were met, the SRS identifier would choose the first reduction set as an optimised reduction. The optimised reduction set would then be used as an input for the next process.

4. Experimental work and results discussion

Various experiments were conducted in order to evaluate the performance of the proposed approach. Two important software packages, Matlab R2014a and rough set exploration system version 2.2 (RSES), were used to ensure the experimental work was successfully executed. Almost all the data processing processes were executed using Matlab: the data pre-processing phase, the soft set parameter reduction phase and the classification phase. Meanwhile RSES was used only for the rough set parameter reduction process. 19 imbalanced data sets were properly selected. The data sets were downloaded from the www.keel.es web-site, known as a Knowledge Extraction based on Evolutionary Learning (KEEL) data repository.

4.1. Data description

Instead of implementing the imbalanced data sets into the proposed approach, the proposed work also considered the large data set to be analysed. Most previous work had not included a large data set to be processed and tested. Many data sets of more than 1000 instances were ignored and not tested. Therefore, the performance of the proposed method in handling large data was not really tested and validated.

Imbalanced data is the data set that is unevenly distributed among the given classes. Most of the data were grouped as a negative class and the least group was a positive class. Some of the data sets consisted of a multiple class imbalanced problem. These data sets were grouped into three categories, i) imbalanced ratio between 1.5 and 9, ii) imbalanced ratio higher than 9 and iii) multiple class imbalanced problem. These imbalanced data sets also contained uncertain and inconsistent data problems. The data was presented by listing the data set's name, number of instances, number of attributes, data type, missing values and the ratio of the instances. The details of each data set are presented in Table 2, Table 3 and Table 4.

4.2. Evaluation measures

This study aimed to analyse the experimental results according to several criteria such as the number of instances that were executed during the parameter reduction process and classification

Table 2
Imbalanced ratio between 1.5 and 9.

Data sets	Number of instances	Number of attributes	Data type	Missing values	Instance imbalanced ratio (%)
segment0	2308	19	Real	No	Positive = 14.25 Negative = 85.75
vehicle0	846	18	Integer	No	Positive = 23.53 Negative = 76.47
page-block0	5472	10	Integer Real	No	Positive = 10.21 Negative = 89.79
glass0	214	9	Real	No	Positive = 32.68 Negative = 67.32
haberman	306	3	Integer	No	Positive = 26.46 Negative = 73.54

Table 3
Imbalanced ratio higher than 9.

Data sets	Number of instances	Number of attributes	Data type	Missing values	Instance imbalanced ratio (%)
vowel0	988	13	Integer Real	No	Positive = 9.11 Negative = 90.98
shuttle-c0-vsc4	1829	9	Integer	No	Positive = 6.72 Negative = 93.28
abalone19	4174	8	Nominal Real	No	Positive = 0.77 Negative = 99.23
kddcup	2225	41	Nominal Real	No	Positive = 0.99 Negative = 9.01
lymphography-normal-fibrosis	148	18	Integer Nominal	No	Positive = 4.05 Negative = 95.95
shuttle-2_vs_5	3316	9	Integer	No	Positive = 1.48 Negative = 98.52

Table 4
Multiple class imbalanced problem.

Data sets	Number of instances	Number of attributes	Data type	Missing values	Instance imbalanced ratio (%)
Penbased	1100	16	Real	No	Positive = 33.9 Negative = 66.1
Contraceptive	1473	9	Real Nominal	No	Positive = 34.6 Negative = 65.4
Dermatology	366	34	Integer	Yes	Positive = 15.27 Negative = 84.73
Autos	159	15	Real	No	Positive = 5.88 Negative = 94.12
Shuttle	2175	9	Real	No	Positive = 0.12 Negative = 99.88
Thyroid	720	21	Real Nominal	No	Positive = 2.64 Negative = 97.36
Ecoli	336	7	Real	No	Positive = 1.38 Negative = 98.62
Wine	178	13	Real	No	Positive = 40 Negative = 60

process, the number of attributes that were eliminated after the reduction process and the time taken for each data set during the classification process. The obtained results were analysed based on standard evaluation measures in order to identify the performance of the proposed approach in classifying the selected data. The effectiveness and the efficiency of the proposed approach was evaluated using accuracy rate (ACC), specificity rate (SPEC), sensitivity rate (SENS), positive predictive value (PPV), negative predictive value (NPV) and F-measure value. The equations of the six evaluation measures are indicated below using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) formulations (Son et al., 2012; Hu et al., 2010).

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (3)$$

$$SPEC = \frac{TN}{TN + FP} \times 100 \quad (4)$$

$$SENS/Recall = \frac{TP}{TP + FN} \times 100 \quad (5)$$

$$PPV/Precision = \frac{TP}{TP + FP} \times 100 \quad (6)$$

$$NPV = \frac{TN}{TN + FN} \times 100 \quad (7)$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

4.3. Results discussion

After several experimental works were conducted for all selected data sets, the results show that the proposed model is effective and efficient, and can be implemented for any types of data sets in any classification problems. Table 5, Table 6 and Table 7 provide the details of the experimental results in terms of classification accuracy rate, number of attributes without using any parameterisation model and number of attributes when implementing the proposed model (hybrid parameterisation model). The performance of the proposed model is also compared with three other hybrid models, namely GA-CFS,WRAP-RS and PCA-Ranker. GA-CFS is a combination of correlation-based feature selection and genetic algorithm, WRAP-RS is a combination of wrapper subset evaluation and random search methods, while PCA-Ranker is a combination of principle component approach and ranker methods. These hybrid parameterisation models were

Table 5
Accuracy rate (%) for data sets that contain imbalanced ratio between 1.5 and 9.

Data sets	No. of attributes	No. of attributes after PR	Without PR	With SRS	GA-CFS	WRAP-RS	PCA-Ranker
segment0	19	5	94.5	92.2	98.3	99.52	97.83
vehicle0	18	6	90.1	91.3	76.4	65.96	64.18
page-block0	10	5	94.5	92.2	94.6	95.38	96.36
glass0	9	5	78.1	84.4	90.6	73.36	77.10
haberman	3	3	80.4	80.4	84.8	74.84	71.57

Table 6
Accuracy rate (%) for data sets that contain imbalanced ratio higher than 9.

Data sets	No. of attributes	No. of attributes after PR	Without PR	With SRS	GA-CFS	WRAP-RS	PCA-Ranker
vowel0	13	5	100	98.6	100	95.34	98.28
shuttle-c0-vsc4	9	3	100	98.5	100	93.44	99.89
abalone19	8	8	99.8	99.8	98.9	99.23	99.23
kddcup	41	4	100	100	100	100	99.73
lymphography	18	7	100	100	100	97.97	97.97
shuttle-2_vs_5	9	2	100	100	100	100	99.91

Table 7
Accuracy rate (%) for data sets that contain multiple class imbalanced problem.

Data sets	No. of attributes	No. of attributes after PR	Without PR	With SRS	GA-CFS	WRAP-RS	PCA-Ranker
Penbased	16	6	98.8	98.1	97.6	34.45	59.36
Contraceptive	9	9	71.9	71.9	59.7	46.78	52.82
Dermatology	34	5	99.4	87.9	100	49.73	91.53
Autos	15	8	95.9	94.4	83.3	53.46	70.44
Shuttle	9	4	99.4	99.5	100	93.84	94.3
Thyroid	21	19	96.9	98.1	96.3	94.38	96.69
Ecoli	7	5	88.9	89.1	54	62.5	77.68
Wine	13	4	97.5	100	100	63.48	97.19

selected as benchmark methods because of their ability in predicting the best attributes to be used in the decision-making process. They are effective in identifying and eliminating the unimportant and redundant attributes (Koc et al., 2012; Bouhana et al., 2013).

As presented, the proposed model represented by the SRS label performed well for all categories of data sets. The classifier returned an accuracy rate of more than 80%, with the use of the optimised attribute set generated by the proposed model. Unfortunately, the contraceptive data set was unsuccessfully classified by the classifier and returned an accuracy rate of only 71.9%. The results show that not only was the proposed model unable to assist the classifier, but also the other hybrid model was unable to generate the optimised attribute set.

4.3.1. Number of reduced attributes

As depicted in Table 5, Table 6 and Table 7 the attributes of all data sets for all three categories was reduced except for haberman and abalone19. The attributes were reduced by more than 50% from the original attribute number. The proposed approach shows that the classification performance for most data sets which contain an imbalanced ratio between 1.5 and 9 gained quite a high accuracy rate when compared to the classification results that do not implement any parameterisation model or other hybrid models. Meanwhile, Table 6 shows that the proposed approach performed quite well in classifying the imbalanced ratio higher than 9 data sets, where the results were similar to the benchmark approach except for vowel0 and shuttle-c0-vcs4 data sets. Table 7 lists the data sets that contained multiple class imbalanced problem data sets, which presented an improvement of the classification results when implementing the proposed approach except for pen-based, dermatology and autos data sets. The best performance of the proposed approach in this problem set was with the wine data set where the obtained result was 100% and the difference from the benchmark approach is about 2.5%. It can be concluded that a good classification result can be achieved with a small number of attributes.

4.3.2. Time taken to process the classification task

Processing time is another factor that must be considered in evaluating the performance of any method in the decision-making process. Instead of reducing the number of attributes, reducing the time taken in processing the data also might influence the generated results. The larger the size of data, the longer time taken in the analysis process. Fig. 2 presents the obtained results of time taken in processing imbalanced data sets. The processing time is presented by terms TIME WPR and TIME PR, where TIME WPR denotes the time taken during the classification task without applying the parameter reduction process. TIME PR represents the time taken for executing the classification task when applying the parameter reduction process. The processing time was recorded during the execution of the classification task and was measured in seconds.

As denoted in Fig. 2, four data sets, segment0, pageblock0, kddcup and dermatology, show an improvement in processing time especially on kddcup and dermatology data sets. The processing time for kddcup and dermatology took more than 50 s to accomplish without using any parameterisation model, whereas kddcup took 302.4 s and dermatology took 51 s to finish. This was because both data sets contained a large number of instances and attributes. However, the proposed model only took less than a second in assisting the classifier to finish analysing the data sets. Unfortunately, the proposed model did not really help the classifier in classifying the shuttle-c0-vsc4 and pen-based data sets in the classification task. The processing time for both data sets increased from 0 to 0.01 s. Overall, this proves that the parameterisation model is required when dealing with a large volume of data set during the data analysis process in order to reduce the processing time. Besides, the processing time can also be reduced by considering use of a high-performance processor. It is beneficial to have software that can measure the time in smaller units so that the processing time can be measured precisely.

4.3.3. Discussion on overall performance

The overall performance of the proposed model was measured not only by looking at the accuracy rate but also by considering the precision, recall and F-measure values. Overall, all the data sets were successfully classified without using any parameterisation model. However, processing time and available space are two main issues that might be faced by decision-makers. These issues can be eliminated by implementing the data parameterisation method. The significance of implementing the parameterisation method had been proven by the obtained results. Even though the obtained results were not exactly the same as the results obtained by not implementing any parameterisation method, they are still

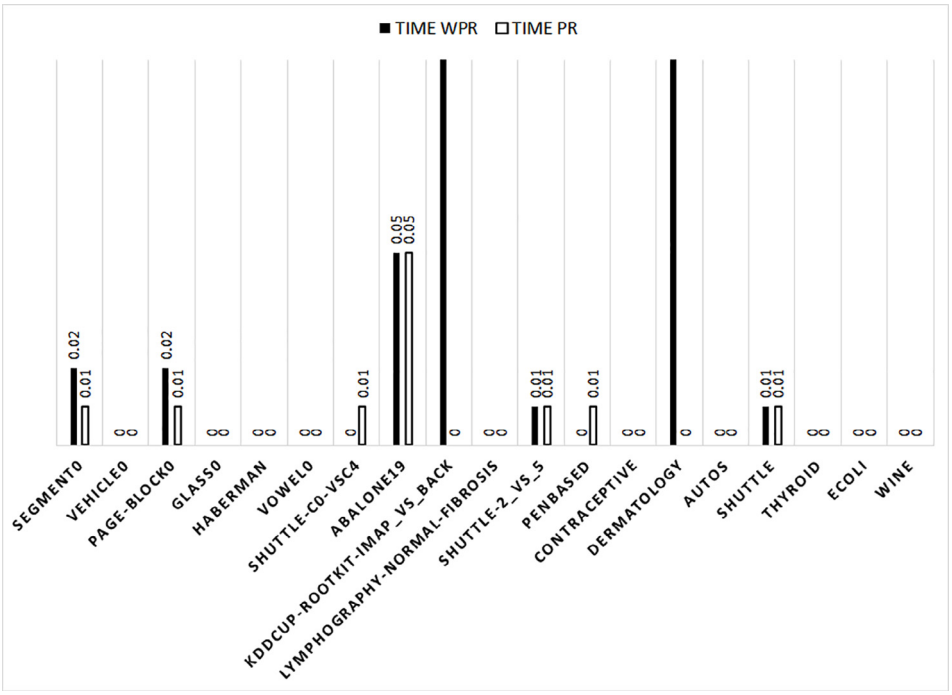

Fig. 2. Performance evaluation on processing time.

satisfying for some data sets which achieved accuracy rate of more than 70%. The results also show that the proposed model effectively and precisely assisted the classifier in handling imbalanced data sets. This is proven by the value of F-measure for all data sets that exceed more than 50% except for thyroid and ecoli data sets.

The precision, recall and F-measure values of the proposed model can be seen in Tables 8–10. The results show that all the parameterisation models successfully assisted the classifier when using the data sets categorised as an imbalanced ratio higher than 9 where the value of precision, recall and F-measure were nearly achieved or achieved to 1. The exceptions were the abalone9 and lymphography data sets. This is because the data division was not properly allocated according to the balance ratio. Thus, the classifier may easily classify the data for the class that was dominant in the data set. Therefore, in order to test the capability of any parameterisation model, the data division must be considered as one of the main criteria. A good division of data which has a balance ratio for each class might help in improving the performance of the decision-making model. Besides, the decision-making process will also fail to return good results without having a balanced data set, even though the best parameterisation model is used or the best classifier is implemented.

4.3.4. Discussion on the performance of the proposed model

The proposed model has been evaluated based on several evaluation measures as discussed in previous sections. The performance of the proposed model has also been validated with several benchmark hybrid models that are regarded as reliable and efficient parameterisation methods. Fig. 3 indicates the average performance of all parameterisation models upon all categories

Table 8
Overall results for imbalanced ratio between 1.5 and 1.9.

Data sets	Precision		Recall		F-measure	
	Without PR	With PR	Without PR	With PR	Without PR	With PR
segment0	0.98	0.70	0.98	0.91	0.98	0.75
vehicle0	0.81	0.83	0.79	0.83	0.79	0.82
page-block0	0.81	0.60	0.86	0.90	0.83	0.64
glass0	0.78	0.76	0.74	0.84	0.75	0.78
haberman	0.55	0.55	0.9	0.9	0.54	0.54

Table 9
Overall results imbalanced ratio higher than 9.

Data sets	Precision		Recall		F-measure	
	Without PR	With PR	Without PR	With PR	Without PR	With PR
vowel0	1	0.93	1	0.99	1	0.96
shuttle-c0-vsc4	1	0.95	1	0.95	1	0.95
abalone19	0.5	0.5	0.499	0.499	0.50	0.50
kddcup	1	1	1	1	1	1
lymphography	0.5	0.5	0.5	0.5	0.5	0.5
shuttle-2–5	1	1	1	1	1	1

Table 10
Overall results for multiple class imbalanced problem.

Data sets	Precision		Recall		F-measure	
	Without PR	With PR	Without PR	With PR	Without PR	With PR
Penbased	0.94	0.91	0.92	0.90	0.94	0.90
Contraceptive	0.56	0.56	0.58	0.58	0.57	0.5
Dermatology	0.98	0.52	0.97	0.56	0.97	0.53
Autos	0.61	0.64	0.56	0.6	0.59	0.61
Shuttle	0.59	0.59	0.59	0.58	0.59	0.59
Thyroid	0.72	0.33	0.98	0.32	0.75	0.33
Ecoli	0.56	0.44	0.49	0.37	0.50	0.38
Wine	0.97	1	0.97	1	0.97	1

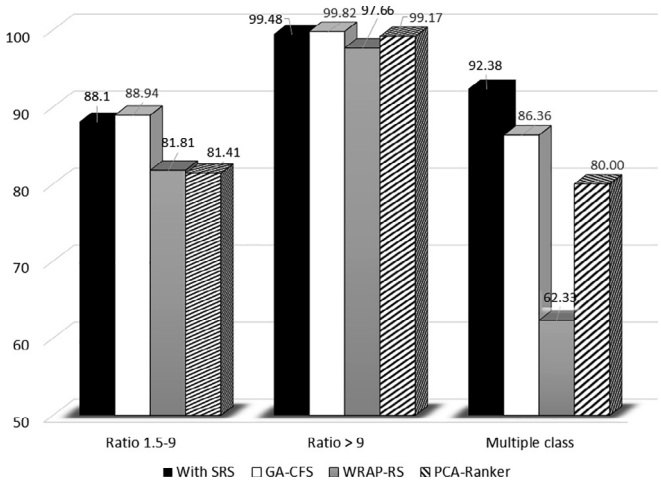

Fig. 3. Average performance of all parameterisation models upon each categories of data set during the classification process.

of data set. The proposed model, which is depicted in black, performed quite well in assisting the classifier in the classification process where it returned 88.1% for imbalanced ratio 1.5–9, 99.48% for imbalanced ratio $>$ 9 and 92.38% for multiple class imbalanced data set. The proposed model has assisted the classifier in achieving the highest average accuracy rate for a multiple class imbalanced data set.

Besides, the proposed model also helped the classifier to return the highest average accuracy rate for all data sets in the classification process. The proposed model returned 93.49%, followed by GA-CFS, which returned 91.29%. PCA-Ranker achieved 86.42% and WRAP-RS only achieved 78.61%. Out of 19 data sets, only one data set, contraceptives, returned quite a low percentage accuracy rate of only 71.9%. The contraceptive result was compared with the results achieved in Arnaiz-González et al. (2016). It shows that the contraceptive data set is not a reliable data set to be used in the testing process where the results of the proposed method returned an accuracy rate of less than 43%. Other data sets such as pageblock0, kddcup, thyroid and wine were also tested in the benchmark paper and were compared with the data sets used in this paper. All three data sets returned similar results except for the wine data set, where the proposed model achieved a higher accuracy rate of 100%, whereas the benchmark models only returned 64.85% and 62.04%. Both benchmark models used a linear complexity algorithm with a combination of nearest neighbour and J48 classifiers.

The results show that the proposed hybrid parameterisation model has achieved its goals, as it is able to deal with a large imbalanced data set and is able to manage an uncertain and inconsistent data set well. The results also prove that a large volume of data set needs to be reduced and selected properly before undertaking the decision-analysis process. Moreover, the results also prove that the combination of soft set and rough set parameterisation methods is beneficial and efficient in reducing and selecting the optimised attribute for any category of data set, and especially for a complex data set.

5. Conclusion

This work has proposed an alternative parameterisation model in analysing complex data, especially for the uncertainty and inconsistency issue. The model integrates three main phases that are comprised of data decomposition, parameter reduction using soft set theory and parameter reduction using rough set theory. The aim of this hybrid model is to reduce the number of parameters of large data sets and to eliminate uncertainty and inconsistency data problems. Imbalanced data sets were chosen to be the testing data sets as the data sets are not preferable for use in experimental works because this might generate insignificant classification results. The experimental works show that the proposed model had achieved the goal in that it successfully reduced the number of attributes, is able to manage an imbalanced data set and increased the classification performance. However, the proposed work needs to be executed phase by phase in sequence. Moreover, the data analysis phase only records the processing time in seconds but not milliseconds. In conclusion, it is hoped that the proposed hybrid model can be executed by using a single application system. This approach might reduce the number of parameterisation tools used in the decision-making process. Moreover, this single tool will also be able to record the data processing time in a more detailed format. Besides, it is recommended to implement the statistical analysis of the selected data before any analysis process is executed, in order to determine the pattern of the data; that is, whether it is evenly distributed or not.

Acknowledgement

The authors wish to thank 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 for the completion of the research. The work is partially supported by the SPEV project, University of Hradec Kralove, FIM, Czech Republic (ID: 2102-2019). We are also grateful for the support of Ph.D. student Sebastien Mambou in consultations regarding application aspects.

References

Ahmad, A., Khan, M., Paul, A., Din, S., Rathore, M.M., Jeon, G., Choi, G.S., 2017. Toward modeling and optimization of features selection in Big Data based social Internet of Things. Future Gener. Comput. Syst. 82, 715–726.

Arnaiz-González, A., Díez-Pastor, J.F., Rodríguez, J.J., García-Osorio, C., 2016. Instance selection of linear complexity for big data. Knowl.-Based Syst. 107, 83–95.
Azar, A.T., Inbarani, H.H., Renuga Devi, K., 2016. Improved dominance rough set-based classification system. Neural Comput. Appl., 1–16
Bouhana, A., Fekih, A., Abed, M., Chabchoub, H., 2013. An integrated case-based reasoning approach for personalized itinerary search in multimodal transportation systems. Transp. Res. Part C: Emerging Technol. 31, 30–50.
Chen, D., Yang, Y., Dong, Z., 2016. An incremental algorithm for attribute reduction with variable precision rough sets. Appl. Soft Comput.
Chen, H., Li, T., Cai, Y., Luo, C., Fujita, H., 2016. Parallel attribute reduction in dominance-based neighborhood rough set. Inf. Sci. 373, 351–368.
Chormunge, S., Jena, S., 2018. Correlation based feature selection with clustering for high dimensional data. J. Electr. Syst. Inf. Technol., 4–11
Derrac, J., Verbiest, N., García, S., Cornelis, C., Herrera, F., 2013. On the use of evolutionary feature selection for improving fuzzy rough set based prototype selection. Soft. Comput. 17, 223–238.
Du, W.S., Hu, B.Q., 2016. Dominance-based rough set approach to incomplete ordered information systems. Inf. Sci. 346–347, 106–129.
Feng, F., Liu, X., Leoreanu-Fotea, V., Jun, Y.B., 2011. Soft sets and soft rough sets. Inf. Sci. 181 (6), 1125–1137.
Herawan, T., Deris, M.M., 2010. Soft decision making for patients suspected influenza. LNCS 6018 – Computational Science and Its Applications... ICCSA 2010, 405–418.
Herawan, T., Deris, M.M., Abawajy, J.H., 2010. Matrices representation of multi soft-sets and its application, LNCS 6018 – Computational Science and Its Applications. ICCSA 2010, 201–214.
Houari, R., Bounceur, A., Kechadi, M.T., Tari, R., Kamel Euler, A., 2016. Dimensionality reduction in data mining: a Copula approach. Expert Syst. Appl. 64, 247–260.
Hu, Y., Guo, C., Ngai, E.W.T., Liu, M., Chen, S., 2010. A scalable intelligent non-content-based spam-filtering framework. Expert Syst. Appl. 37 (12), 8557–8565.
Kim, S.H., Vu, T.M., Pyeon, C.H., 2017. A preliminary study on applicability of artificial neural network for optimized reflector designs. Energy Procedia 131, 77–85.
Koc, L., Mazzuchi, T.A., Sarkani, S., 2012. A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. Expert Syst. Appl. 39 (18), 13492–13500.
Lam, H.K., Ekong, U., Liu, H., Xiao, B., Araujo, H., Ling, S.H., Chan, K.Y., 2014. A study of neural-network-based classifiers for material classification. Neurocomputing 144, 367–377.
Li, H., Li, D., Zhai, Y., Wang, S., Zhang, J., 2016. A novel attribute reduction approach for multi-label data based on rough set theory. Inf. Sci. 367–368, 827–847.
Luo, C., Li, T., Yi, Z., Fujita, H., 2016. Matrix approach to decision-theoretic rough sets for evolving data. Knowl.-Based Syst. 99, 123–134.
Ma, L.Q., Xueling, Zhan, J., 2016. A survey of decision making methods based on certain hybrid soft set models. Artif. Intell. Rev., 1–24
Ma, L.Q., Xueling, Zhan, J., 2017. A survey of decision making methods based on certain hybrid soft set models. Artif. Intell. Rev. 47, 507–530.
Maji, P.K., Roy, A.R., Biswas, R., 2002. An application of soft sets in a decision making problem. Comput. Math. Appl. 44, 1077–1083.
Manogaran, G., Varatharajan, R., Lopez, D., Kumar, P.M., Sundarasekar, R., Thota, C., 2017. A new architecture of Internet of Things and big data ecosystem for secured smart healthcare monitoring and alerting system. Future Gener. Comput. Syst. 82, 375–387.
Masetic, Z., Subasi, A., 2016. Congestive heart failure detection using random forest classifier. Comput. Methods Programs Biomed. 130, 54–64.
Massimiani, A., Palagi, L., Sciubba, E., Tocci, L., 2017. Neural networks for small scale ORC optimization. Energy Procedia 129, 34–41.
Meng, Z., Shi, Z., 2016. On quick attribute reduction in decision-theoretic rough set models. Inf. Sci. 330, 226–244.
Meng, D., Zhang, X., Qin, K., 2011. Soft rough fuzzy sets and soft fuzzy rough sets. Comput. Math. Appl. 62 (12), 4635–4645.
Mohamad, M., Selamat, A., 2016. A new hybrid rough set and soft set parameter reduction method for spam e-mail classification task. Lecture Notes in Artificial Intelligent, LNAI 9806 (9806), 18–30.
Mohamad, M., Selamat, A., 2017. An analysis of rough set-based application tools in the decision-making process, recent trends in information and communication technology. IRICT 2017. Lecture Notes on Data Engineering and Communications Technologies 5, 467–474.
Mohamad, M., Selamat, A., 2017. A new soft rough set parameter reduction method for an effective decision-making. New Trends in Intelligent Software Methodologies, Tools and Techniques 297, 691–704.
Mohamad, M., Selamat, A., 2018. A two-tier hybrid parameterization framework for effective data classification. New Trends in Intelligent Software Methodologies, Tools and Techniques 303, 321–331.
Molodtsov, D., 1999. Soft set theory-first results. Comput. Math. Appl. 37 (4), 19–31.
Montazer, G.A., ArabYarmohammadi, S., 2015. Detection of phishing attacks in Iranian E-banking using a fuzzy-rough hybrid system. Appl. Soft Comput. 35.
Paradarami, N.D., Tulasi, K., Bastian, Wightman, J.L., 2017. A hybrid recommender system using artificial neural networks. Expert Syst. Appl. 83, 300–313.
Pawlak, Z., 1997. Rough set approach to knowledge-based decision support. Eur. J. Oper. Res. 99, 48–57. https://doi.org/10.1016/S0377-2217(96)00382-7.
Pawlak, Z., 1998. Rough set theory and its applications. J. Telecommun. Inf. Technol. 29, 7–10.
Qian, J., Lv, P., Yue, X., Liu, C., Jing, Z., 2015. Hierarchical attribute reduction algorithms for big data using MapReduce. Knowl.-Based Syst. 73, 18–31.
Raza, M.S., Qamar, U., 2016. An incremental dependency calculation technique for feature selection using rough sets. Inf. Sci. 343–344, 41–65.
Son, C.-S., Kim, Y.-N., Kim, H.-S., Park, H.-S., Kim, M.-S., 2012. Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches. J. Biomed. Inform. 45 (5), 999–1008.
Triguero, I., Peralta, D., Bacardit, J., García, S., Herrera, F., 2015. MRPR: a MapReduce solution for prototype reduction in big data classification. Neurocomputing 150, 331–345.
Wang, L., Wang, Y., Chang, Q., 2016. Feature selection methods for big data bioinformatics: a survey from the search perspective. Methods 11, 21–31.
Wei, W., Li, J., Cao, L., Ou, Y., Chen, J., 2013. Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web 16, 449–475.
Weng, C.-H., Huang, T.C.-K., Han, R.-P., 2016. Disease prediction with different types of neural network classifiers. Telematics Inform. 33, 277–292.
Yang, Y., Tan, X., Meng, C., 2013. The multi-fuzzy soft set and its application in decision making. Appl. Math. Model. 37 (7), 4915–4923.
Zhang, Z., Gao, G., Yue, J., Duan, Y., Shi, Y., 2014. Multi-criteria optimization classifier using fuzzification, kernel and penalty factors for predicting protein interaction hot spots. Appl. Soft Comput. J. 18, 115–125.
Zhou, P., Hu, X., Li, P., Wu, X., 2017. Online feature selection for high-dimensional class-imbalanced data. Knowl.-Based Syst. 136, 187–199.