sciendo

Int. J. Appl. Math. Comput. Sci., 2021, Vol. 31, No. 4, 659–683
DOI: 10.34768/amcs-2021-0046

amcs

APPLICATIONS OF ROUGH SETS IN BIG DATA ANALYSIS: AN OVERVIEW

PIOTR PIĘTA [a,*], TOMASZ SZMUC [a]

[a] Department of Applied Computer Science
AGH University of Science and Technology
Mickiewicza 30, 30-059 Kraków, Poland
e-mail: pipieta@agh.edu.pl

Big data, artificial intelligence and the Internet of things (IoT) are still very popular areas in current research and industrial applications. Processing massive amounts of data generated by the IoT and stored in distributed space is not a straightforward task and may cause many problems. During the last few decades, scientists have proposed many interesting approaches to extract information and discover knowledge from data collected in database systems or other sources. We observe a permanent development of machine learning algorithms that support each phase of the data mining process, ensuring achievement of better results than before. Rough set theory (RST) delivers a formal insight into information, knowledge, data reduction, uncertainty, and missing values. This formalism, formulated in the 1980s and developed by several researches, can serve as a theoretical basis and practical background for dealing with ambiguities, data reduction, building ontologies, etc. Moreover, as a mature theory, it has evolved into numerous extensions and has been transformed through various incarnations, which have enriched expressiveness and applicability of the related tools. The main aim of this article is to present an overview of selected applications of RST in big data analysis and processing. Thousands of publications on rough sets have been contributed; therefore, we focus on papers published in the last few years. The applications of RST are considered from two main perspectives: direct use of the RST concepts and tools, and jointly with other approaches, i.e., fuzzy sets, probabilistic concepts, and deep learning. The latter hybrid idea seems to be very promising for developing new methods and related tools as well as extensions of the application area.

Keywords: rough sets theory, big data analysis, deep learning, data mining, tools.

1. Introduction

In 1991, Professor Zdzisław Pawlak ignited a new approach to data analysis (Pawlak, 1991). Rough set theory (RST), invented in the early 1980s, has undoubtedly become a well-known framework for processing uncertain knowledge and often compared to fuzzy set theory (Zadeh, 1965) or used jointly as the fuzzy-rough approach (Dubois and Prade, 1990; 1992). During the last decades, many generalisations and extensions of the classical RST were proposed by researchers that come not only from Europe but also from the USA, the UK, Japan, China and other countries. This theory has been successfully applied in several domains: machine learning, pattern recognition, data mining, decision analysis and support, rule generation, data reduction, granular computing and other areas.

In order to extract useful knowledge from large amounts of data from databases, generated by sensors or other systems, a great need for the proper processing of such sources of information appears inexorably. In recent years research interesting data analysis methods, have been developed especially when big data and the Internet of things are of primary concern. Most of them refer to the stimulating works about hybrid models originated by combining the RST concepts placed within the proposed architecture as one of the crucial elements. For example, Chen *et al.* (2020a) consider and develop the RST connected with support vector machine big data fusion technology for feature extraction and information mining carried out in the process of intelligent prediction of economic trend indexes.

In this article, we briefly present recent studies on RST in the context of big data analysis. Section 2 gives a fast and straightforward insight into the big data technology. The next section concentrates on the basic notions of RST and the related tools. Section 4

[*] Corresponding author

points out several difficulties in large-scale data set processing. Section 5 summarizes the foundations of the local rough set (LRS) approach based on the classical RS theory. An overview of selected RS applications for big data analysis is presented in the next two sections. Section 6 concentrates on direct applications of RS methods. The next section focuses on applications of RST combined with other methods. The hybrid approaches are grouped into two categories: RS combined with similar (fuzzy, probabilistic) approaches and merged with neural networks. Finally, Section 8 provides a summary and suggestions for future research.

2. Big data

Nowadays, the volume of information is increasing at an uncommon rate. We are witnessing the rapid development of information technologies and their impact on our personal life and environment. Extraction of information from a traditional database is usually a simple task. More commonly, we have to retrieve information from multiple, heterogeneous, autonomous, and distributed data sources (e.g., IoT) with complex and evolving relationships and increasing volumes. Properly defining big data may cause troubles due to its ambiguous meaning, placed in different contexts (Stefanowski *et al*., 2017).

According to Isitor and Stanier (2016) the term big data describes a data environment in which scalable architectures support the requirements of analytical and other applications which process, with high speed, high volume data which may have a variety of data formats and which may include high velocity data acquisition. Big data has become a recent area of strategic investment for businesses by providing extremely powerful business intelligence when data are properly analyzed, synthesized, and visualized (Venkatraman and Venkatraman, 2019). This technology and services involve a variety of hardware or software resources, tools and techniques such as in-memory databases (IMDBs), NoSQL databases, massive parallel processing (MPP), Hadoop, Phoenix, Spark or MapReduce file systems, virtualization, cloud platforms and related software as well as analytics solutions (Sedkaoui, 2018; Chao, 2018).

Occasionally, the data also perish at the equivalent high speed as they are produced. Infrastructural technology is considered as a basis of the big data ecosystem for the storage, analytics and visualisation of data (Venkatraman and Venkatraman, 2019).

2.1. 3V model. The 3V model was described in 2001 under the META Group report. It tries to characterize and define the phenomenon that is big data in the current time of such rapid technological development around the world. It uses a 3V perspective: volume–variety–velocity (GARTNER, 2001).

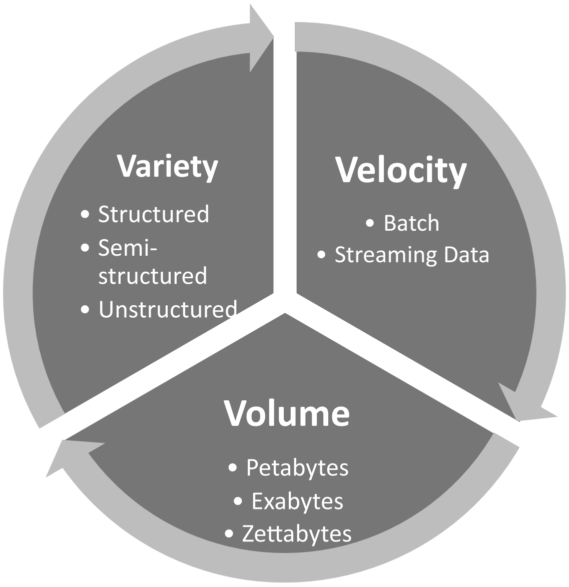

Fig. 1. Big data—the 3V model (after Ivanov *et al.*, 2013).

The Gartner company extended the 3V model in 2011 with two further dimensions: variability and complexity (Isitor and Stanier, 2016; Bulger *et al*., 2014).

2.2. Big data analysis. Data analytics is a process of inspecting, cleaning, transforming and modelling data to discover useful information, suggest conclusions, and support decision making. It focuses on knowledge discovery for predictive and descriptive purposes to find new ideas or to confirm existing ones (Sedkaoui, 2018). There are two main profiles of such analytics in big data: descriptive and predictive. According to Delen and Demirkan (2013), big data add the ability to perform a third type of analytics, known as perspective analytics. Descriptive analytics techniques depict what is contained in a data set or database (past data), and they use simple statistics for them. The predictive approach uses more advanced statistical methods and supports building models that identify trends, future events and relationships not readily observed in the descriptive analysis. The perspective analytics provides methodologies that allow optimal use of allocable resources. For instance, linear programming models may be used for optimal allocation of budget to various advertising media (Sedkaoui, 2018).

With the rise of popularity of big data in industrial environments, the need to store massive volumes of data and their proper processing still increases. Machine learning (ML), as a significant part of artificial intelligence (AI), is a learning system aimed at detecting unknown regularities in big databases, inducting rules, creating analogies and modifying data. In a pervasive sense, the term big data refers to processing a large amount of data in a specified and automated way. ML is often compared to data mining as the same approach to data analysis. Still, they differ, i.e., ML algorithms are used as a tool for solving data mining problems. Over the few last years, many algorithms and improved approaches for data analysis were proposed and tested in order to provide the best results. Some of them were utilized in combination with rough sets.

The available ML methods may be divided into the two categories: supervised and unsupervised. Supervised algorithms are trained on annotated (or labeled) input data to build predictive models (Cichosz, 2015). Unsupervised algorithms try to automatically discover complex patterns in unlabeled data sets. Regression techniques take a finite set of relations between dependent variables and independent variables and create a continuous function to generalize these relations. Regression methods predict the continuous or real number value based on previous observations from a training set (Watt *et al*., 2016).

The linear model is an example of the simplest form of the regression. It is based on the assumption that there is a linear relationship between the input (observations) and the output (predictions),

$$\widehat{Y}_i = \beta_0 + \beta_1 x_i,$$

where x is an observation vector, and $\widehat{Y}$ is approximated from the real observations. The algorithm tries to find a line parameterized by β_0 and β_1, which fills the training data better. Classification refers to a predictive modelling problem where a target class is predicted for a given example of input data.

K-nearest neighbors (kNN) is a well-known algorithm that may be used both for classification and regression. The method takes k data points closest to the studied point in order to predict its label. The kNN algorithm classifies a given data set in one of the categories by calculating the distance between the category and each point of the training set. The technique takes the nearest k elements, and it chooses the dominant label among the k elements representing the category of the data set element.

The decision tree (Quinlan, 1983) is a classification method based on using a tree structure to define the final decision. Connections between the tree nodes are labelled by conditions. The model is built using an ML method, e.g., CART (Gordon *et al*., 1984), C4.5 (Quinlan, 1993), or LMT which are classification trees with logistic regression functions at the leaves (Landwehr *et al*., 2005). The algorithm starts with a set of predefined classes and searches interactively for the most different variables in the classified entities. Once the variable is identified, and the decision rules are determined, the data set is segmented into several groups according to the rules. Data analysis is performed recursively on each subset until all key classification rules are identified.

A random forest (Breiman, 2001) is a collection of decision trees. The algorithm provides better predictive results and requires almost no data preparation and modelling (Sedkaoui, 2018).

Logistic regression (Cessie and Houwelingen, 1992) is a statistical methods for performing binary classifications. It takes qualitative or ordinal predictors as input and measures the probability of the output value using the sigmoid function. In comparison with this model, the support vector machine (also named the large margins classifier) chooses the clearest separation possible between the two classes. Naive Bayes (John and Langley, 1995) is an approach to both binary and multiclass classification problems. The method relies on the Bayes theorem:

$$P\left(A|B\right) = \frac{P\left(B|A\right)P\left(A\right)}{P\left(B\right)}. \tag{1}$$

The idea is based on treating each feature independently. The algorithm evaluates the probability of each feature, regardless of any correlation, and makes the prediction based on the Bayes theorem. The advantages of the using the method include its simplicity and ease of understandings. In addition, it performs well for data sets with irrelevant features, since the probabilities contributing to the output are low. The naive Bayes method usually results in good performance in terms of consumed resources, since it only needs to calculate the probabilities of the features and classes.

Artificial neural networks (NNs) are models inspired by the human brain. They allow finding complex patterns in datasets. The model consists of neurons (nodes) placed in several layers of the network. Input data (environment) trigger neurons in the first layer while the other neurons are triggered through weighted links from previously active neurons. The output layer calculates classification/decision results. The models require a lot of learning data and are not suitable for all problems, especially if the number of input parameters is too low (Sedkaoui, 2018). Deep neural networks (DNNs) (Cios, 2018), a branch of neural networks, have been rapidly developed for the last decade. The networks are built from a cascade of layers corresponding to a hierarchy of abstraction concepts (multidimensional learning). Advanced DNNs could process abstract features and are commonly used in image recognition (convolutional neural networks, CNNs), natural language and time variant signals processing (recurrent neural networks, RNNs), and processing inputs in the tree order (recursive neural networks). Currently, this research field constitutes the main trend in machine learning and data mining.

Unsupervised algorithms help to discover complex patterns in untagged datasets (Sedkaoui, 2018). There is no labeling of input data. Clustering is a kind of unsupervised approach—the algorithms group similar objects into clusters (or separate groups). Cluster analysis became a branch of statistical multivariate analysis (Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990). The most popular metrics for quantifying the similarity between two elements are the Euclidean distance, the Manhattan distance or the Hamming distance.

K-means is a commonly used clustering algorithm. There are various extensions of the classical k-means proposed in the literature (Alhawarat and Hegazi, 2018; Meng *et al.*, 2018; Lv *et al.*, 2019; Sinaga and Yang, 2020). This method divides a set of data entities into separate groups, where the parameter k is the number of created clusters. The approach allows an assignment of entities to different clusters by iterative calculations of the average midpoint or centroid for each cluster. The created centroids become the focal points of the iterations, which refine their locations and reassign the data entities to fit the new locations. The steps of the algorithm are repeated until the groupings are optimized and the centroids do not move anymore.

The same calculations are applied in another clustering algorithm named K-Medians, using the median vector instead of the mean one. In comparison with the first approach, this method is much slower for larger data (due to sorting in each iteration) and less sensitive to outlier instances.

The mean shift clustering is sliding-window-based method that attempts to find dense areas of data points. The density-based spatial clustering of applications with the Noise (DBSCAN) algorithm (Ester *et al.*, 1996) does not require a pre-set number of clusters. In comparison with similar approaches, it can find arbitrarily sized and arbitrarily shaped clusters. Additionally, the algorithm identifies outliers as noises, unlike mean-shift clustering which simply throws them into a cluster (even if the data point is very different).

Expectation-maximization (EM) clustering using Gaussian mixture models (GMM) is another clustering approach. In this case, we assume that data points have Gaussian distribution. An optimization algorithm named Expectation-Maximization is used to find Gaussian parameters (the mean and standard deviation) for each cluster. Agglomerative hierarchical clustering algorithms (Murtagh, 1983; Murtagh and Contreras, 2012) can be divided into top-down and bottom-up categories. Bottom-up methods treat each data point as a single cluster at the beginning. Next, they successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster containing all data points.

Big data analysis may be supported by many available software tools (e.g., Datawrapper, Lumify, Apache Storm, Talend, Knime) or distributed libraries/modules (e.g., NLTK, TensorFlow, Keras, Pandas, Scikit-learn). For instance, R and Python are the most popular programming languages used in academic and industrial environments for computational analysis of experiments. They provide techniques and packages for overall data analysis starting from pre-processing to advanced visualization of achieved results (e.g., generating reports). RapidMiner is a cross-platform which offers an integrated environment for data science, machine learning and predictive analytics. Apache SAMOA is an open-source platform for big data stream mining and machine learning. The free tools allow users to create distributed streaming ML algorithms and run them on multiple distributed stream processing engines.

2.3. Reproducibility in data science.

Computational reproducibility in data science means that an experiment must be able to be reproduced to validate achieved results. This approach should facilitate reuse of a current experiment by other scientists or to allow them extending it in a simple way. Ensuring reproducibility of work is a labour-intensive task due to the requirements for strict documentation of each stage of an experiment. There is no existing platform providing reproducibility verification (Ivie and Thain, 2018) of experiment outcomes (for different methods and big data sets) and standard statistical evaluation of differences between them.

Some tools have been proposed as an attempt for solving this problem, e.g., ReproZip (Chirigati *et al.*, 2016) or WholeTale (Brinckman *et al.*, 2019). ReproZip consists of packing and unpacking phases. In the first phase, the ReproZip creates an .rpz file containing all the necessary information and components for the experiment done in the original environment. In the second phase, the work can be reproduced from the file, even if the current environment runs under a different operating system from the original one. The Whole Tale is an open source, web-based and multi-user platform for reproducible research enabling the creation, publication, and execution of tales—executable research objects that capture data, code, and the complete software environment used to produce research findings.

Mondelli *et al.* (2019) developed a conceptual model of the reproducibility of an experiment and a generic framework for findable, accessible, interoperable and reusable (FAIR) computational experiments. These four principles contains a set of requirements on how data, metadata, and infrastructure must be managed, allowing machines to retrieve them automatically, or at least with minimal human intervention (Wilkinson *et al.*, 2016). The model is built using an entity-relationship diagram. Each aspect involved in the process of computational experiment reproducibility is mapped into an entity in this model. The main idea behind the proposed conceptual model and the relationships between its entities (selected entity names were included in the brackets) is as follows: a user specifies and runs the experiment under the operating system (OS) using a specific hardware. This OS has some packages installed (OS package), and among them, we can consider the scripting language package that is used to specify the experiment (such as R or Python). The experiment is defined through a script, from an existing package in the OS. It can contain calls to user-defined functions or functions (function) that are part of a specific language package or a module (script package). These modules need to be installed in the OS. The functions comprise the activities of the experiment that consume input and produce output. Parameters can also be used as input to functions, and constitute an attribute of the consume relation (Mondelli *et al.*, 2019).

In the same paper, the authors proposed another framework for the verification of computational experiments. It is based on the previously mentioned conceptual model for reproducible experiments and adaptation of the FAIR principles. The first step, named packaging (or encapsulation), consists of importing the experiment and the related packages or libraries into a virtual machine (VM). Validation is the second step. The main idea is to access the VM and re-execute the experiment in the new environment. This step requires user interaction, who is responsible for validating the generated results. All modifications must be recorded to update the default VM specifications. Thus, a user intending to reproduce the experiment will be able to rebuild the virtual environment in which it has been validated. The reproduction does not need to install the required applications manually. Publishing is the last step and can be done in a repository that allows sharing research results. The authors demonstrated the framework implementation through ENM case study (Mondelli *et al.*, 2019).

The above mentioned models and tools do not deliver a common solution for verification of the reproducibility. It is currently a crucial and open problem in data science.

3. Rough set theory

Rough sets theory is a popular mathematical framework proposed by Z. Pawlak for dealing with uncertainty, imprecision and vagueness in the context of knowledge acquisition and reasoning from such data. The fundamentals of the classical theory were presented in the referenced book (Pawlak, 1991). The key issue of the methodology is how to define approximately an inexact set based only on known sets (crisp sets) contained in the universe of discourse. The RST based methods can be perceived as a useful tool in decision and non-decision problems covering a wide spectrum of domains. It can be basically used for finding hidden patterns in data sets or to enhance the proper measure of significance of the attributes, pointing out dependencies between attributes and in data reduction by calculation of the core and the reducts as the subset of the essential attributes.

A book in memory of Professor Pawlak was published in 2013 (Skowron and Suraj, 2013a; 2013b). The two volumes contain more than 50 chapters written by scientists who cooperated with Professor Pawlak during his life. The monograph is a kind of summary (up to that date) of RST and its applications.

3.1. Notions. RST uses the notion of an information system (IS) for a representation of gathered data. During the last few years, researchers proposed several generalizations of the IS to handle complex and large-scale data, e.g., a multi-scale information system (Wu and Leung, 2011), a multi-source information system (Lin *et al.*, 2015) and a multi-modality information system (Hu *et al.*, 2018). In particular, the IS is defined using decision tables where conditional and decision attributes are determined. This way of the IS description (interpretation) directly corresponds to a data model. Many tools and algorithms based on the RS concept exist. Data in the form of decision tables are used as an input to the related systems.

Let U be a universe of discourse, $R \subseteq U \times U$ is an equivalence relation on U, called classification R. The pair $K = (U, R)$ is a relational system called a knowledge base (or an approximation space). The equivalence relation R divides the universe U into several disjoint sets, each having the same property. The partition is denoted by U/R.

Any equivalence class consists of objects having the same property (for example, objects with the same specific colour). All of the equivalence classes are considered as the current knowledge of the agent (system). For a given set $X \subseteq U$, we can use the agent's actual knowledge to describe the set. The relation R determines the exactness of the description, which means that in some cases we are not able to decide which objects belong to X. It may happen when an equivalence class groups objects on the border of X. In such cases, the agent's knowledge is insufficient to give a precise representation of X (or to gather all the objects using ambiguous description). To define an inexact set (a concept) in the universe, Pawlak proposed two exact sets called lower and upper approximations of X. The approximation sets are described in terms of subsets of their attributes.

In addition, a granule of information has been introduced as describing a clump of objects drawn together by indiscernibility, similarity, connectivity, and the proximity of functionality (Zadeh, 1997; Pedrycz *et al.*, 2015b). The notions form a basis for granular computations and similar approaches (Skowron and Stepaniuk, 2001; Pal and Meher, 2013; Lin *et al.*, 2015; Pedrycz *et al.*, 2015a; Skowron *et al.*, 2016; Xu and

Yu, 2017).

There is no need to have any transcendental knowledge (a probability distribution or a level of membership) regarding the data when utilizing the RS model to handle data (Pawlak, 1991). The related reasoning is carried out based on knowledge directly gathered from data; however, combinations with other approaches (fuzzy, Bayesian, etc.) are also applied. The hybrid approaches will be discussed later in the paper.

3.2. Methods. RST allows reducing dimensions of given datasets by removing the superfluous data (attributes and some of their values can be omitted) using the key concepts: the reduct and the core (the discernibility matrix and the Boolean function). It provides an extensive methodology to extract useful knowledge, i.e., hidden relationships between objects and their classes in the form of rules induced from data (Pawlak, 1991).

3.3. Applications. The RS methodology in data analysis is a scientific field that cannot remain in strong theoretical frames without practical applications. Many researches have continued the work of Z. Pawlak (*selected publications*) with regard to, e.g., Boolean reasoning: Skowron and Nguyen (1999), Pawlak and Skowron (2007); approximate reasoning: Skowron and Stepaniuk (1996), Skowron (2001); granular computing: Skowron and Stepaniuk (2001), Skowron *et al.* (2016); conflict analysis: Skowron *et al.* (2006); interactive granular computing: Skowron *et al.* (2009), Skowron and Wasilewski (2011a; 2011b), Skowron and Dutta (2017); rough mereology: Polkowski nad Skowron (1996; 2000), Polkowski (2011; 2020); neurocomputing: Polkowski (2005); spatial reasoning: Polkowski and Osmialowski (2010); multi-source decision system: Lin *et al.* (2016); methods devised for acceleration of big data computations in database engines: Chądzyńska-Krasowska *et al.* (2017); object recognition and content-based image retrieval: Przyborowski *et al.* (2018); approach to learning forecasting models over large multi-sensor data sets: Ślęzak *et al.* (2018); LERS system for data mining: Grzymała-Busse (1997); probabilistic approximations: Clark and Grzymała-Busse (2011), Clark *et al.* (2019; 2020); approximation spaces: Skowron and Stepaniuk (1996); methods for big data set processing: Czolombitko and Stepaniuk (2017), Kopczyński *et al.* (2016; 2017); rough web caching: Sulaiman *et al.* (2009); dynamic reducts and statistical inference: Bazan (1996; 1998); inductive reasoning: Bazan *et al.* (2005); hierarchical classifiers: Bazan (2008), Bazan *et al.* (2020); Petri net with RSs: Peters *et al.* (2000); decomposition methods: Pancerz and Suraj (2013); fuzzy forward reasoning methodology for rule-based systems using the functional representation of rules: Suraj *et al.* (2015); fuzzy rough granular neural networks in classification: Ganivada and Pal (2011); granular social network: Pal and Kundu (2017); double bounded RSs: Kundu and Pal (2018); granulated deep learning: Pal *et al.* (2019), Pal (2021); dynamic dominance rough sets: Huang *et al.* (2020); attribute reduction in fuzzy-rough sets: Yuan *et al.* (2021); inhibitory rules: Delimata *et al.* (2008), Delimata *et al.* (2009); multiple classifiers: Delimata and Suraj (2013); adaptive fuzzy rough approximate time controller: Peters *et al.* (1998); near sets: Peters and Naimpally (2012), Peters (2013); software defect classification with the rough-fuzzy-neural hybrid approach: Bhatt *et al.* (2009); topology: Peters (2020); fuzzy modeling: Pedrycz and Gomide (1994), Hirota and Pedrycz (1999), Izakian *et al.* (2015); granular computing: Pedrycz and Bargiela (2002), Pedrycz and Vukovich (2001); data fusion: Pedrycz *et al.* (2021); multi-modality information system: Hu *et al.* (2017a); neighbourhood rough sets: Hu *et al.* (2013); decision systems: Wakulicz-Deja *et al.* (1998), Simiński and Wakulicz-Deja (2003), Ilczuk and Wakulicz-Deja (2005), Nowak-Brzezińska and Wakulicz-Deja (2019); incomplete knowledge: Wakulicz-Deja *et al.* (2011); conflict analysis: Wakulicz-Deja *et al.* (2013), Wakulicz-Deja and Przybyła-Kasperek (2016); rough representations of graded ill-known sets: Inuiguchi (2013); algebraic structures of different kinds of information systems: Khan and Banerjee (2013); information fusion: Wei and Liang (2019); neutrosophic fusion of RST: Zhang *et al.* (2020); discretization methods: Nguyen (1997; 1998); approximate Boolean reasoning: Nguyen (2006), Cornelis *et al.* (2015); D-stripped quotient set and dependencies in degree k between attributes: Nguyen *et al.* (2017); topology: El-Bably and Kozae (2014); measure for the induction of fuzzy rough classification trees (FRCTs): Bhatt and Gopal (2006); RS based on Galois connections: Madrid *et al.* (2020).

From our perspective, applications of RSs in big data analysis yielded new significant methods and tools. For example, the research group supervised by Nguyen proposed a scalable method for classification problem in the client-server environment (Kwiatkowski *et al.*, 2010). The new approach named FDP was a modified version of the frequent-pattern discovery algorithm (FP) from the transaction data set. In the first step, the data structure called the frequent decision pattern tree is created (note that the algorithm is applicable for decision tables). Next, the set of frequent decision rules is generated from the structure built in the previous step. In order to get a set of irreducible decision rules, the last phase of the FDP consists in inserting the obtained rules into a data structure—a minimal rule tree. The implemented algorithm was tested on benchmark data sets from the UCI Machine Learning Repository (with different sizes of training data). The accuracy and computational time of the approach were compared with the nearest-neighbour classifier and the naive Bayes classifier available in WEKA. The experiment results confirm a linear dependence of the computation time on the size of the training set.

The big data analytics in combination with RS methodology is still an important research field. Many open issues should be concerned in future works, e.g., reducing computational time-complexity, scalability, merging the interactive granular computing approach with big data analysis (Skowron and Dutta, 2018), non-determinism (Sakai *et al*., 2020) and others.

It is worth noting that there are about 117 000 publications (including books, monographs and tools) that refer to RST (as of March 2021). According to Google Scholar (2021), 8 040 of them have RST in the title.

3.4. Tools. Researchers proposed many RST modifications during the last decades and developed several software tools based on RSs. One of the most significant systems is Learning from Examples based on Rough Sets (LERS). It was developed by Grzymała-Busse (1992) for machine learning and data mining. The main functionality of LERS covers rule induction from raw data (especially when inconsistencies and missing values are of primary concern) and classification of the new examples using a set of generated rules (Grzymała-Busse, 1997). A comparative overview of the tools is presented by Pięta *et al*. (2019). A short description of the tools is presented below.

The Rough Set Exploration System (RSES) (RSES, 2005; Bazan *et al*., 2002) employed the RSES-lib 2 library for computations. The library and GUI were designed and implemented at the Group of Logic, Institute of Mathematics, Warsaw University, and the Group of Computer Science, Institute of Mathematics, University of Rzeszów, Poland. The tool allows the user to perform complex data mining experiments on decision tables while providing a simple GUI interface. It is worthwhile to mention that a new and entirely redesigned version 3 of the library was released in 2019 (RSlib, 2019).

The Rough Set Data Explorer (ROSE2) (ROSE, 1998; Prędki and Wilk, 1999; Prędki *et al*., 1998) has been created at the Laboratory of Intelligent Decision Support Systems of the Institute of Computing Science in Poznań, Poland. It provides both basic and advanced data analysis methods based on classical RSs and variable–precision rough set theory (Ziarko, 1993). The tool has more methods implemented than the above-mentioned RSES.

The Rough Set Toolkit for Analysis of Data (ROSETTA) (ROSETTA, 1994; Ohrn and Komorowski, 1997; Ohrn, 2000) has been developed at the University of Uppsala. Apart from the basic functionality as the import and export of data, it supports also the ODBC interface for extracting data from data base. ROSETTA uses the RSES library for elementary computation and adds its implementations of well-known algorithms based on the classical RST and its extensions: variable-precision rough set approximation and based on tolerance relations.

The Waikato Environment for Knowledge Analysis (WEKA) (WEKA, 2009; Jensen, 2014) is an integrated tool for data analysis and predictive modelling. The supported methods are implemented in the form of filters to perform experiments. WEKA is based not only on RST but also on other mathematical formalisms (in the form of attached packages), such as fuzzy-rough sets. Thanks to a GUI consisting of Experimenter and Explorer windows, WEKA helps a user through a series of stages in data mining: preprocessing, feature selection, instance selection and classification, proposing many practical functionalities outside the classical RSs.

The Rough Set Data Analysis Framework (jMAF) (jMAF, 2021; Błaszczyński *et al*., 2012) supports the methods based on dominance-based rough sets (Greco *et al*., 1999) and the variable consistency-based rough set approach (Greco *et al*., 2001). It is suitable for analysing data gathered in the decision table with predefined profiles based on background knowledge about ordinal evaluations of objects from a given universe and about monotonic relationships between these evaluations.

We can also refer to Rseslib 3 (RSlib, 2019), an interesting tool not mentioned in the above publication. The open-source library written in Java delivers extensible, modifiable algorithms and computational models from RST and machine learning. It bases on modular component-based architecture (six modules: Discretization, Logic, Discernibility, Reducts, Rules and Rough Set Classifier) that enables us to implement unconventional combinations of data mining methods. Rseslib is available in WEKA as an official WEKA package.

RoughSets is an R package for data modelling and data analysis. R is one of the most popular programming languages mainly used for statistical computations and data science. The library is described widely by Riza *et al.* (2014). The tool integrates several algorithms based on classical RSs and the fuzzy-rough approach (FRST) in a single software library. There are more than 40 functions included in this package. The strength of the R lies in many modules delivered from the CRAN repository that help researches on each step of the data analysis (e.g., preprocessing, discretization, visualization of the achieved results using embedded methods or `ggplot2`/other packages). With the easy-to-use integrated development environments (like RStudio), we can conduct non-trivial experiments on data and import only required packages.

The most recent version of the FRS-based algorithms is implemented in Python as fuzzy-rough-learn 0.1 library (Lenz *et al.*, 2020). The following algorithms can be found in this tool: fuzzy-rough feature selection (Cornelis *et al.*, 2010), variants of fuzzy-rough nearest neighbours (Jensen and Cornelis, 2008), fuzzy-rough rule induction (Jensen *et al.*, 2009), fuzzy-rough prototype selection (Verbiest *et al.*, 2013), fuzzy-rough OVO combination multiclass classification (Vluymans *et al.*, 2018b) and fuzzy-rough neighbourhood consensus multilabel classification (Vluymans *et al.*, 2018a).

4. Constraints in application of rough sets

Processing large-scale data sets consisting of thousands of rows and columns and covering a broad spectrum of domains cause many problems with storage, understanding and management of such data. It may especially appear when we want to acquire knowledge from distributed databases where data are represented in different ways. This section points out some difficulties that may arise within the data analysis process. The properties described in the following subsections formed motivations for introduction of local rough set theory (Skowron *et al.*, 2018).

4.1. Properties of big data. The basic rough set model requires a representation of the information system. Decision tables are widely used for the agent's knowledge about the outer realm. The algorithms based on classical RSs process labelled data with the last attribute named the decision attribute. In big data, we cannot be confident that all data are labelled. Labelling a large amount of data is an expensive and laborious task, and sometimes even infeasible and many machine learning algorithms require complete data sets. In such cases, the classical RS approach may be computationally expensive and practically useful for relatively small data sets. Local rough sets (LRSs) (Skowron *et al.*, 2018) provide theoretical basics that may overcome the inefficiency mentioned above.

4.2. Time-consuming computation. The majority of existing conventional RS algorithms require all data to be pre-loaded into memory, and then the data are processed by the algorithms (Hu and Wang, 2008). The computing of approximations or attribute reductions may be time-consuming; therefore, scalability and decomposition tools are needed to improve efficiency (Skowron *et al.*, 2018).

4.3. Over-fitting in attribute reduction. The next issue refers to an over-fitting problem in the data reduction process. There is a high possibility that an attribute reduct obtained from noisy data might have more attributes in some cases. The research efforts to ensure the monotonicity of an attribute reduction process is the primary concern and motivation to overcome this issue (Skowron *et al.*, 2018).

5. Local rough set theory

To overcome the disadvantages (expressed in the previous section) of the classical (global) rough set model in the processing of the large data, Skowron *et al.* (2018) proposed a new framework called local rough sets (LRSs). The concept is based on combining of the classical RS approach with the decision-theoretic RS framework (Yao, 2007) formed into one representation and developing a series of approximation concepts and attribute reduction algorithms of linear time complexity. The algorithms may efficiently work in limited labelled big data domains.

Theoretical analysis and experimental results presented by the authors show that each algorithm based on local rough sets significantly outperforms its original counterpart in classical RST (Skowron *et al.*, 2018). A brief description is given below.

Let (U, R) be an approximation space with R being an equivalence relation on U and $U/R = \{[x]_R : x \in U\}$ the set of all equivalence classes generated by the equivalence relation R. Then, for any subset $X \in U$, the α-lower and β-upper approximations of the set X are defined by (Skowron *et al.*, 2018)

$$\begin{aligned} R_*(\alpha,)(X) &= \{x : P(X|[x]_R) \geq \alpha, x \in U\}, \\ R^*(\beta)(X) &= \{x : P(X|[x]_R) > \beta, x \in U\} \end{aligned}$$

where $P(\cdot)$ is a conditional function often depicted as an inclusion degree with the constraints, and α, β are two parameters from the decision-theoretic rough set.

The pair $\langle R_*(\alpha)(X), R^*(\beta)(X)\rangle$ is called a local rough set. The authors proposed definition of the local boundary region: $BN_R = R_*(\beta)(X) - R^*(\alpha)(X)$. In the case when the $\alpha = 1$ and $\beta = 0$ the LRS can be reduced to the classical rough set.

Other concepts of the classical RST (accuracy measure, accuracy of approximation) were also transformed into the related notions in LRS theory and used for investigating monotonicity of heuristic attribute reduction algorithms in the LRS.

In general, the LRS computation model of lower or upper approximations does not obtain information granules of all objects in advance as in the RS model but only calculates the objects within a target concept. Compared with the classical RS model, these algorithms can take unlabelled data, which may be considered a significant advantage of the approach.

Computing approximation and attribute reduction of a target concept and a target decision are the new methods based on LRS investigated in the cited paper. The following algorithms have been proposed: LLAC for computing a local lower approximation of the target concept, LARC for searching for a local attribute reduct of the target concept, LLAD for calculating a local lower approximation of a target decision, and LARD for finding a local attribute reduct of the target decision. The linear time complexity characterizes the proposed methods so that they are well suited for big data analysis (Skowron *et al.*, 2018).

The LLAC algorithm takes as an input the information system IS, a target concept X, and a parameter value α. To compute a local lower approximation LA of X with respect to attributes, in the first step, of generation equivalence classes to approximate the target concept is carried out. The initial values for LA and i are set in the next step. The comparison between the equivalence classes $|X|$ with the target concept X for obtaining its local lower approximation is made in the next phase of LLAC as the most important part of the proposed method.

The second algorithm, called LLAD, computes the local lower approximation of a target decision with some extensive stages. Similarly, it takes as an input the parameter α and a decision table (with the last attribute called the decision or the class), but in comparison with LLAC, generation of equivalence classes to approximate every target concept are carried out as the second step of the method.

According to the authors, there is no difficulty to extend originated algorithms by computation of a local upper approximation of a target concept or a local upper approximation of a target decision based only on the existing methods. Still, they did not continue in this way in the work of Skowron *et al.* (2018).

LARC is a forward greedy local attribute reduction algorithm for a target concept. It takes several arguments as an input of the proposed method: an information system IS, a parameter value α and $X \subseteq U$. Similarly, the designed algorithm for finding a local attribute reduct with respect to a target decision (named LARD) in local rough sets takes only a decision table and a parameter α. LARC utilizes the inner and the outer importance of α with respect to X. The former measure determines the significance of every attribute. The latter value is used in the forward feature selection process.

LARD starts with the attribute with the maximal inner importance. In the next phase, it takes an attribute with the maximal outer significance into the attribute subset. This step is done in a loop, until the attribute subset satisfies a stopping criterion. In a similar way, the LARD algorithm computes a local positive region with respect to the decision. At the end of these two methods, a local attribute reduct is obtained (Skowron *et al.*, 2018).

6. Direct applications of rough sets in big data analysis

In this section, we describe interesting applications of the rough set methods in big data environments. Most of the works mentioned in this chapter do not concentrate on building a new theory based on rough set concepts, but rather on combining the well-known methods and algorithms within the concrete phases of the data mining process in such a manner (e.g., Kang *et al.*, 2011). The main section is divided into subsections, each of which describes applications in a related data mining stage or usage of RSs jointly with other approaches. Each subsection delivers a simple view on some remarkable results achieved by researchers and connected with the main topic.

To conclude the presentation, Tables 1 and 2 provide an overview (references and brief descriptions). Table 1 regards the data mining phases and gives several works in this manner. Table 2 focuses on presenting general applications of RST in combination with other approaches.

6.1. Discretization of continuous values on attributes. Rough set based algorithms are mainly used in combination with other methodologies to provide better efficiency and classification results and improve the effectiveness of knowledge acquisition in a specific manner. One of the most critical phases of data analysis is preprocessing within the discretization has an important role. Several techniques have been originated to enhance transformations of continuous values into the discrete ones.

Li and Shen (2020) proposed a discretization algorithm for incomplete economic information in RS-based processing on big data. In the first step, they used a deep neural network for filling-in incomplete economic information. After the supplement, the algorithm for discretization in the RS is used to implement the discretization based on supplementary economic information theory.

As mentioned in the referenced paper, when the number of breakpoints increases, it still has a higher computational efficiency and can effectively improve the integrity of incomplete economic information. Finally, the application performance is superior (Li and Shen, 2020).

Chen *et al.* (2021) present a hybrid metric method of feature discretization for classification of high-resolution remote sensing images in coastal areas. In the proposed methodology, as one of the essential phases of the technique (after calculating of the stability of pixel categories in discrete intervals), they borrowed the degree of dependence among knowledge from the classical RST as the evaluation criterion of the discretization scheme. Then, each band was scanned in turn with the strategy

Table 1. Rough sets in big data (*selected works*).

Task	Subtask/approach	References
Preprocessing	Discretization	Li and Shen (2020), Qiong *et al*. (2021)
	Data reduction/Feature selection	Skowron *et al*. (2018), Kong *et al*. (2020), Jingjing *et al*. (2019), Hamidinekoo *et al*. (2018), Venkatraman and Venkatraman (2019), Qiong *et al*. (2021), Sun *et al*. (2021), Thuy and Wongthanavasu (2021)
	Approximation concept	Skowron *et al*. (2018), Yun (2014), Dagdia *et al*. (2017), Dagdia *et al*. (2018), Kong *et al*. (2020), Cui and Huang (2015), Liu and Zhang (2019)
	Missing values	Shan *et al*. (2016)
KDD support	Prediction/regression model	Chen *et al*. (2020a)
	Incremental learning	Yang *et al*. (2017), Huang *et al*. (2017), Luo *et al*. (2016; 2018), Wang *et al*. (2016a), Li *et al*. (2015)
	Decision fusion	Shan *et al*. (2016)
	Decision support	Zhang *et al*. (2012), Sun *et al*. (2019), Sachin and Shubhangi (2015), Jing *et al*. (2014), Banerjee and Badr (2018), Hong-Wei and Xindi (2016), Narayanan *et al*. (2017), Chowdhury *et al*. (2016), Pal (2020), Li *et al*. (2019), Vluymans *et al*. (2015), Wang *et al*. (2016b), Zhao *et al*. (2020)
	Cloud computing	Zhang *et al*. (2012), Sun *et al*. (2019), Kune (2014), Qu *et al*. (2019), Li *et al*. (2015), Wang *et al*. (2016b), Grzegorowski *et al*. (2017)
	Rule induction	Zhou and Lin (2018), Wang *et al*. (2016b)
	Data clustering	Wan and Li (2019), Cui and Gao (2019), Xie (2018), Grzegorowski *et al*. (2017), Li *et al*. (2021), Janusz and Ślęzak (2014)
	RS-based approximate SQL	Naouali and Missaoui (2005), Ślęzak *et al*. (2012; 2018)
Hybridizations	Local rough sets	Skowron *et al*. (2018), Yang *et al*. (2017), Qiana *et al*. (2017)
	Granular computing (GrC)	Hu and Wang (2008), Tang *et al*. (2019), Chen (2017), Xia *et al*. (2020), Li *et al*. (2015), Pal (2020), Zhao *et al*. (2020)
	Multigranulation	Qiana *et al*. (2017)
	NN and deep learning	Chu and Zhang (2020), Pal (2020), Vluymans *et al*. (2015), Xiaoguang *et al*. (2018), Li *et al*. (2016), Hassan (2017)
Computational platforms	MapReduce	Zhang *et al*. (2012), Sachin and Shubhangi (2015), Cui and Huang (2015), Jing *et al*. (2014), Chowdhury *et al*. (2016), Pandu (2020)
	Apache Spark	Dagdia *et al*. (2017), Vluymans *et al*. (2015)

of splitting and merging to obtain an optimal discrete feature set. Remote sensing image features were the input of the algorithm, and discretized features were the output. After the initialization of needed parameters and discretization thresholds, getting discrete intervals by the hybrid metric method, the evaluation of the information system compatibility was done using the well-known equation from RST ($card(\cdot)$ means the cardinality of a given set):

$$\gamma_C(D) = \frac{card(POS_C(D)}{card(U)}$$

The obtained value (in the range of $[0,1]$) from the above equation indicates the degree of dependence between knowledge marked here as C and D. Experiments conducted by authors delivered interesting results; in comparison with other discretization-based techniques, they got fewer discrete feature intervals and data errors and achieved better classification results on an SVM and a neural network classifier (Chen *et al*., 2021).

Qiong *et al*. (2021) proposed a discretization algorithm based on a fuzzy-rough (FR) model to analyze and process high-resolution remote sensing big data. They determined the membership degree of each pixel in training samples through linear decomposition, and established the individual fitness function based on the FR model. An adaptive genetic algorithm was applied for the selection of discrete breakpoints, and the MapReduce framework calculated the individual fitness of the population in parallel to obtain an optimal discretization scheme in the minimum time.

6.2. Approximation concept. The concept of approximation stands a key issue in the classical RS-based methodology. Due to the lower and upper approximations of a given inexact set X of the object's instances in considered domain, we can describe roughly (or only approximately) X based on the current agent's knowledge. During the last decades, several generalizations of the concept approximation have been proposed in terms of a binary relation.

As mentioned before, the LRS model's strength lies in some inconspicuous details; for example, it needs to calculate information granules of objects within a given target concept, and also only compares them with the target concept for determining its lower/upper approximation (Skowron *et al.*, 2018). For example, suppose we want to approximate a concept in the form of a subset X included in the universe of discourse U using the classical RS methodology. In that case, the time complexity of the algorithm for computing all equivalence classes in such universe is $O(|U|^2)$, but the same action done in the method proposed by Skowron as the first step of the LLAC costs only $O(|X||U|)$. As the authors added in this publication, they achieved the linear time complexity for LLAC $O(|X||U| + |X|^2)$ in terms of $|U|$.

In a similar way, due to computing only r local lower approximations as a next step in the LLAD algorithm, the time complexity is $(\sum_{j=1}^{r} |X^j||U|) = |\bigcup_{j=1}^{r} X^j||U|$ (Skowron *et al.*, 2018). From this point of view, we can conclude that the LRS method can be treated as more suitable than the classical RS counterpart approach for large-scale data analysis.

On the other hand, in some cases, such complexity remains still unfeasible for big data processing. To overcome that issue, researchers investigated recently several methods based on the improved form of the local RS called double-local rough sets (Wang *et al.*, 2021). The new framework defines a local deletion matrix, an upper addition matrix and an upper deletion matrix for computing the approximations. In comparison with LRSs, the approach is effective and efficient in dealing with completely or partially labelled large-scale data sets. The double-local RS model avoids the repeated computation of equivalence classes and sets comparisons.

6.3. Data reduction. Data reduction or feature selection is one of most important stages of the data analysis. Methods based on RST can effectively support computation of the reducts and core, and as consequence of the approach, irrelevant attributes can be omitted from a given data set.

Traditional reduct computation techniques fail in big data processing. They can be computationally intensive or yield poor performance in terms of the size of the resulting reducts. To overcome such issues, Janusz and Ślęzak (2014) developed two algorithms for the computation of multiple decision reducts. The methods are based on a greedy heuristic approach and attribute clustering results to obtain a set of diverse and short reducts. The authors evaluated the proposed techniques and showed that, by applying clustering results, it is possible to speed up the search for decision reducts significantly. In addition, the obtained reducts tended to be smaller than those reached without clustering (Janusz and Ślęzak, 2014).

As mentioned in the previous section, Skowron *et al.* (2018) proposed the LARC algorithm for searching a local attribute reduct of a target concept based on local rough sets and the LARD algorithm for finding a local attribute reduct of a target decision with linear time complexity of each algorithm.

Dagdia *et al.* (2017) deal with a novel efficient distributed algorithm based on RST for large-scale data pre-processing in the Spark framework. To reduce the computational effort of the RS computations, the authors split a given data set into partitions with smaller numbers of features that are then processed in parallel. Next, they demonstrated the effectiveness of the approach using the Amazon Commerce reviews data set from the UCI Machine Learning Repository (UCI, 2021). The data set was characterized with 10000 features and 1500 data items equally spread over 50 classes.

An extension of neighbourhood rough sets (NRSs) that replace the membership function with the neighbourhood concept is presented by Venkatraman and Venkatraman (2019). The extension allows an NRS to handle scenarios where no prior knowledge is available. A novel rough set based method named GBNRS was proposed. It is the first parameter-free RS algorithm for processing continuous data. It does not require any membership functions or the optimization of any mid-computation parameters for processing continuous data. It demonstrably out-performs the current state-of-the-art NRS algorithm with its time complexity of $O(N)$. The adaptive method of selecting the neighbourhood radius improved the quality of attribute reduction. Examination of GBNRS using popular feature selection benchmark data sets led to higher classification accuracy than both classical NR and the current best NRS algorithm, FARNeMF. It showed that efficiency was improved by more than 90% on a relatively large benchmark data set. The granular computing ball was also improved, and MGBNRS was proposed to achieve an even higher efficiency than GBNRS (Venkatraman and Venkatraman, 2019). A distributed fuzzy-rough set (DFRS) algorithm for feature selection was discussed by Kong *et al.* (2020).

The role of feature space granulation was presented by Grzegorowski *et al.* (2017). They evaluated lastly introduced feature space granulation approaches and discussed the meaning of similarity, proximity and functionality in the context of the physical existence of granules or potentially derivable attributes.

Li *et al*. (2016) used classical RS techniques for data reduction (using core and reduct concepts) in order to simplify the decision table.

In many real applications, especially when we consider big data, a number of labels of training samples are randomly missed, and multilabel classification can have great complexity and ambiguity. To address this issue, Sun *et al.* (2021) proposed a feature selection algorithm based on multilabel NRSs combined with fuzzy neighbourhood RSs was proposed. Promising results have been obtained, but the experiments pointed out that better classification performance cannot be achieved when the percentage of missing labels is very high. This implies an open question, and indicates direction for future research, which should concentrate on improving classification performance and decreasing computational costs of the proposed model for multilabel data with missing labels, and also finding more efficient optimal search strategies and uncertainty measures in this approach (Sun *et al.*, 2021).

7. Rough sets combined with other methods

The section presents selected worth considering results in combining the theory with the other methodologies that cover various aspects of big data processing. The applications are distributed among two subsections related to merging with similar methods (fuzzy, probabilistic) and combining with neural network approaches.

7.1. Rough sets merged with similar approaches. Chen *et al.* (2020a) proposed an intelligent prediction model of economic trends index based on an RS support vector machine. Shan *et al*. (2016) recommended RST for decision fusion of incomplete information systems and proposed a new approach to evaluate the impact of missing data. The authors introduced an improved metric called the α-classification quality of approximation in order to measure the quality of decision fusion with various identical degrees (IDs). Yang *et al*. (2017) proposed a unified framework of dynamic three-way probabilistic rough sets for incrementally updating three-way probabilistic regions (positive, boundary, and negative). D'eer *et al*. (2016) discussed a semantically sound approach to Pawlak's model using a descriptive language. Luo *et al*. (2016) focused on efficiently updating probabilistic approximations with incremental objects in a dynamic information table. Wang *et al.* (2016a) deal with an efficient algorithm for updating rough approximations with a multi-dimensional variation of ordered data based on the dominance-based rough sets approach and the incremental learning strategy. Selected papers (Ishizu *et al*., 2007; Krishnamurthy and Janardanan, 2018; Pierzchała, 2014; Yap and Kim, 2013) show the connection between ontologies and RST. For example, Ishizu *et al*. (2007) generalize the concept of an ontology by using RST and define a rough ontology. They built a rough ontology upon the information system (Pawlak, 1991) and formulated its properties.

Hu *et al.* (2017b) proposed four algorithms for updating rough approximations based on fuzzy probabilistic rough sets over two universes. Luo *et al*. (2018) dealt with a formalization of the dynamic characteristics of knowledge granules with the cut refinement and coarsening through attribute value taxonomies in the hierarchical multi-criteria decision systems. Several uncertainty measures of neighbourhood granules were proposed by Chen *et al.* (2017): neighbourhood accuracy, information quantity, neighbourhood entropy and information granularity in the neighbourhood systems. Some rudiments about the processing of large-amount of data were considered by Hu and Wang (2008).

An interesting paper by Zhang *et al*. (2012) presented an algorithm proposed for knowledge acquisition using the MapReduce framework from big data in combination with parallel rough sets (implemented on the Hadoop platform). In the first step of the proposed method, they divided decision table S into m decision sub-tables. Next, based on each sub-decision table, they computed independently: numbers of elements in equivalence classes, decision classes and union classes. Simultaneously, the classes of different sub-decision tables were combined if their information sets were the same. Hence, it can be transformed into a MapReduce problem, and they designed three parallel methods based on RST for knowledge acquisition.

Hu and Wang (2008) provided a principle and a method for processing massive data sets based on RST and granular computing. Sun *et al.* (2019) proposed a decision-making method based on fuzzy theory and Bayesian-rough sets. In the first step in the above-mentioned approach, a decision-making model was established. The comprehensive operational efficiency, cost, cycle and risk were chosen as decision-making factors. Evidence theory and cloud model theory were used to optimize the comprehensive operational efficiency factor. In the next step, a Bayesian RS was introduced in order to solve the redundancy problem in decision-making factors. Yun (2014) employed classical RST to enhance the information processing capacity, and an electromagnetism-like algorithm can also avoid a local optimum search in the context of big data analysis system based on a three-tier structure. Zhao *et al*. (2020) dealt mainly with application of RST and granular computing theory for intelligent evaluation. Firstly, through the discretization, evacuation situational elements and behavioral evolution characteristics of the crowd were expressed in the form of knowledge, and the redundant knowledge was eliminated by reduction. Then, through the rigorous inverse discretization, the rough set's meta-rules were reduced to generalized rules with practical significance. Cui and Gao (2019) carried out experiments on RS processing outliers in cluster analysis. Li *et al.* (2015) introduced a solution called PICKT on big data analysis based on the theories of granular computing and RST.

A knowledge acquisition method based on variable granularity (Zhao *et al.*, 2020) was designed to simplify the complex data and further improve the reliability of the obtained rules. Finally, a knowledge base for the crowd's evacuation and stability was constructed, the decision rules were exported, and the interpretation was generalized.

A discussion about the meaning of similarity, proximity and functionality while considering the granules of physically existing or potentially derivable attributes was presented by Grzegorowski *et al.* (2017). The authors showed several examples of utilization of the granulation structures defined over the feature spaces in the feature selection algorithms. As a case study they considered algorithms developed within RST, aimed at finding irreducible subsets of attributes that are sufficient to distinguish between the cases belonging to different target decision classes.

Accelerating standard big data computing is an essential issue in scientific research. The RS paradigm may be used for this purpose. The group of researchers supervised by Ślęzak investigated a new database engine that acquired and utilized granulated data summaries for the purposes of fast approximate execution of analytical SQL statements (Chądzyńska-Krasowska *et al.*, 2017). Wnuk *et al.* (2020) presented an approach to data and information granulation known from the Infobright Community Edition (ICE) (Ślęzak *et al.*, 2008; 2010; Ślęzak and Eastwood, 2009)—an analytical database engine developed in order to minimize the need for accessing and decompressing the data while resolving SQL queries and showed how to re-implement them (ICE's granulated tables) into other libraries: Apache Parquet and ROOT. The RS-based approximate SQL extensions may be used to support knowledge discovery in databases and business intelligence operations on big data.

The issues related to the process of global-decision making on the basis of information stored in several local knowledge bases was described widely by Przybyła-Kacperek and Wakulicz-Deja (2013; 2014; 2016a; 2016b; 2017).

7.2. Rough sets and neural network approaches.

During recent years researchers developed some significant improvements in the classical RS model. An interesting and very prospective direction combines RST with neural networks (NNs) in solving many real-life problems—the so-called rough-neural computing paradigm (Pal *et al.*, 2004). Hassan (2017) investigated a new model, where a hierarchy of actors and their behaviours in social networks deeply learn from an individual decision table level. The objective of this work is to propose a model that uses more decision tables and approximates these tables to a classification system. The framework introduced by deep rough sets can be regarded as the first attempt that uses deep learning combined with rough set methods. In a similar way to classical RST, the proposed model may be considered as the sextuple (U, C, D, J, f, W), where U is the universe of discourse defined as $U = U_1 \cup \cdots \cup U_n$, and each U_i is a set of objects in the decision table, C is a non-empty finite set of attributes that can be divided into subsets of conditional attributes and D is a set of decision attributes $(C_i \cap D = \emptyset)$, J a set of deep relations, f is an information function, W is a matrix of real-valued non-zero weights such that $\sum_{i,j} c_i w_{ij} = 1, \forall j \in U/C, c_i \in C_j$, and a total weight w_i^* for decision table can be calculated as

$$w_i^* = 2\tilde{c_i} + \sum W_{ij} c_j$$

where $\tilde{c_i}$ is the mean attribute value, for the decision table.

The author's key idea is to replace the indiscernibility relation I used in conventional RST with a deep relation J, which is defined for each pair of objects x and y as (Hassan, 2017)

$$xJy = f(x, c \in C) = f(y, c \in C).$$

Both single and multi-decision tables can be handled by this model (Hassan, 2017). The proposed architecture has three types of layers called the decision table layer, the deep decision system layer, and deep rough neural layer (Hassan, 2017). To demonstrate the power of the proposed framework in a practical area, some experiments were carried out on social networks (represented as a graph), a Facebook dataset, a Twitter dataset and a Wikipedia dataset. They showed that with the deep rough set approach, we can obtain better results measured on classification rate (the accuracy above 90%).

A similar, but not the same, combination of RS methods with deep learning has been presented by Li *et al.* (2016). The authors developed a DLRSA model (feature extraction based on deep learning and situation assessment based on rough set analysis named SARSA), which can be treated as an extension of the cyberspace situational awareness (CSA) model. However, in comparison with the results of, e.g., Li and Shen (2020), Chen *et al.* (2020a) or Hassan (2017), RST can be used in the classical way (Pawlak, 1991) andserve as a technique for realizing a cyberspace situation assessment

Table 2. Selected applications with references.

Reference	Topic of the selected papers
Yang *et al.* (2017)	A unified framework of dynamic three-way probabilistic rough sets ○ a novel matrix approach is investigated
D'eer *et al.* (2016)	Rough sets and covering-based rough sets
Qiana *et al.* (2017)	Incremental rough set approach for hierarchical multicriteria classification
Hu and Wang (2008)	Algorithms for computing positive region and attribute core based on divide and conquer method
Wan and Li (2019)	Clustering data stream with rough sets ○ introduces upper and lower approximations in rough sets to describe the uncertainty of the data stream ○ lets a time decay model to describe the evolution of data flow
Lu *et al.* (2019)	Method for big data sets of high-resolution earth observation images
Kune (2014)	Genetic algorithm based data-aware group scheduling for big data clouds
Sachin and Shubhangi (2015)	Parallel RS and MapReduce from big data
Tang *et al.* (2019)	Granular computing based online public opinion
Bello and Falcon (2017), Pal (2020)	Addresses the significance of granular computing in several mining applications ○ an elaborated study on the context of big data ○ review of machine learning and rough sets
Li *et al.* (2019)	A new privacy protection algorithm in attribute-related data ○ proposes an information entropy differential privacy solution for correlation data privacy issues based on rough set theory.
Vluymans *et al.* (2015)	Distributed fuzzy rough prototype selection for big data regression ○ the work is aimed at learning a regression model ○ builds over large high-dimensional data sets
Xiaoguang *et al.* (2018)	Neural networks in combination with rough sets ○ problem of classification of LBS service facilities
Li *et al.* (2016)	Situation assessment based on rough set analysis (SARSA) ○ classical rough set techniques are used
Hassan (2017)	Deep rough sets architecture based on multi-decision tables ○ integration of different decision tables into deep architecture
Jian *et al.* (2019)	Apply the Apriori-based framework to analyzing data set ○ rule-based model ○ estimation of missing values in the data by using the obtained rules
Qiong *et al.* (2021)	Discretization method for high-resolution remote sensing big data
Thuy and Wongthanavasu (2021)	Attribute selection method for high-dimensional mixed decision tables ○ introduces a new concept of stripped neighbourhood covers to reduce unnecessary tolerance classes from the original cover

to enhance the process of transforming information into knowledge (Li *et al.*, 2016). It is worth mentioning that the concepts from the classical rough set approach were adopted by the authors into specific needs, like the set of monitoring data in DLRSA which can be perceived as a conditional attribute set, and the set of situation values would be regarded as a resulting attribute set (Li *et al.*, 2016). In addition, all the historical data were adopted into a decision table form. Simplification of the decision table and a minimal decision rule induction based on RS techniques were applied to select situational factors from monitoring indices. In the process of SARSA, the participation of experts is necessary for pattern recognition (they analyze the cyberspace situation patterns by judging and scoring) (Li *et al.*, 2016).

In the paper of Xiaoguang *et al.* (2018) RS methods were used to enhance proper classification of location based services (LBS)—service facilities. Algorithms based on classical RTS were adopted for preprocessing data in combination with neural networks. Experiments carried out by authors led to some interesting results: they got the same classification results at the end, but in the case of preprocessed data, the training time was shorter, fewer training steps were performed, and a higher precision was achieved.

In the work of Li *et al.* (2021) the missing value was regarded as a decision feature, and then the prediction was generated for the objects that contained at least one missing value. The algorithms called jointly fuzzy c-means, vaguely quantified rough sets based nearest neighbour imputation (JFCM-VQNNI) and jointly fuzzy c-means and fitted vaguely rough sets based nearest neighbour imputation (JFCM-FVQNNI) have been proposed. The first method clustered the complete object set into several groups using a fuzzy c-means algorithm, and implemented fuzzy similarity relations to judge the relevance degree of the missing object with its similar records. The second algorithm can be treated as the improved JFCM-VQNNI and added the analysis of the fuzzy membership of dependent features for instances with the corresponding clusters. The authors carried out experiments on two complete and three incomplete data sets provided by the UCI Repository (UCI, 2021).

A novel efficient semi-supervised algorithm (SSFRCNN) for image classification was proposed by Riaz *et al*. (2019). The presented approach fused the fuzzy-rough c-means clustering algorithm (FRCM) with neural networks in the overall architecture. This perspective has not been discussed before, but several attempts have been made in this way (Deng *et al*., 2017; Yeganejou and Dick, 2018; Rajesh and Malar, 2013). In comparison with other methods on the representation learning task, the combination approach (Riaz *et al*., 2019) dealt with the labelled and unlabelled data. The later remains the most strength among the similar methodologies (Wu and Prasad, 2018; Shi *et al*., 2015; Zhou *et al*., 2014), especially in the context of big data (Sedkaoui, 2018; Chao, 2018). The authors' framework (Riaz *et al*., 2019) applied FRCM algorithm to learn k centroids from an unlabelled data set U, which consists of four important parts: unsupervised, supervised, and semi-supervised learning modes, and a task-driven classification layer. The experiments were carried out on four large-scale benchmark data sets (ImageNet, MNIST, CIFAR-10 and Scene-15m) related to the image classification problem and obtained promising results.

Pal *et al*. (2019) proposed a granulated deep learning system for motion detection and object recognition with linguistic description. In the first step of the methodology, they performed a granulation on an input image frame f_i. Then, the object O_b and the background models B_g are computed on the granulated input. In the next step, the computed O_b and B_g feed a deep neural network (DNN) for recognition of static and moving objects. For this task, the authors applied a convolutional neural network (CNN) due to the known abilities of this neural architecture for typical image recognition tasks. In general, the CNN layer takes as an input a raw-pixel frame, but in this approach the input is in granulated form. Experiments were done on frames granulated with various techniques (uniform-sized rectangular granules with spatial similarities, unequal-sized rectangular granules with grey level and spatial similarities, and natural arbitrary-sized/shaped (neighbourhood) granules with spatio-colour similarity). The comparison of results was carried out in terms of time and accuracy between the proposed method, deep learning without granulation, and other state-of-the-art algorithms.

In the context of artificial intelligence, deep learning should be considered as a promising, fastest developing field that creates new interesting directions in the current research. Many algorithms, approaches, frameworks and publications have been presented in this area in recent years, and the general tendency is still growing due to the obtained new results in different areas. It is worth noticing that deep neural networks have gained particular importance as a strong tool for fighting the COVID-19 pandemic among other machine learning methodologies as show by Islam *et al*. (2020).

In addition, RST has also been spotted by researchers for addressing the specific needs using neural networks. Hassan (2017) presented a deep learning architecture based on rough sets from a theoretical perspective. Some interesting results have been achieved by a combination of the convolutional neural networks (CNNs) with the fuzzy-rough set approach (Chen *et al.*, 2020b). The constructed framework called the Type 2 fuzzy rough convolutional neural network was used as a model for facial expression recognition problem in terms of the fuzzy classification task (it varied from a traditional picture classification).

Finally, a recent book on granular computing using rough sets and deep learning should be mentioned (Chakraborti and Pal, 2021). Hybrid methods are applied for tracking objects from video sequences. Several new algorithms are proposed and compared with existing solutions. The content is interesting for video processing and seems to be also inspiring for other application areas.

8. Summary

The aim of this article is to present recent rough set theory applications for big data analysis. The presentation is not restricted to classical RST and the related tools but also describes applications of novel RST concepts and hybrid approaches. The diversity of the analyzed applications caused a division of the application section into three general thematic fields:

1. Applications of the classical RST,
2. RS usage jointly with the other approaches,
3. RS combined with neural networks.

The first field covers the concepts based on classical RST in data mining. RST is applied according to its primary aim, i.e., reducing uncertainty in terms of indiscernibility. The classical approach can be considered separable as a concrete phase/phases data analysis process. In this way, we can improve classification results and use RS for data preprocessing (e.g. discretization of continuous attribute values), data reduction (reduct-core computations during conducted experiments), approximation of concepts, rule induction, etc.

Hybrid rough set approaches are significantly more common due to numerous extensions and available modifications of the classical RST model, e.g., fuzzy-rough sets, probabilistic rough sets, dominance-based rough sets, neighbourhood rough sets. The majority of such papers provide details of the modified methodology, proved by various experiments. RS combined with the well-known frameworks dealing with processing a large amount of data, like MapReduce or Spark, is a good example of the approach.

Other techniques, like local rough sets (LRSs) investigated and developed by Skowron *et al.* (2018), can be regarded as a significant contribution to the overall rough set approach in providing effective, sustainable and scalable methods for large-scale data analysis. It seems that LRSs are a very prospective tool for further applications in big data analysis. To enable LRSs to more efficiently handle both completely labelled data and partially labeled data, an enhanced local rough set framework, called double-local rough sets was proposed in 2021. It is worth noting that several software tools and libraries implementing algorithms based on RS, or RS extensions, were designed and implemented in academic environments.

The last application field deals with RST extensions combined with neural networks. This section has been extracted from hybrid approaches to point out its potential applicability for big data analysis. These methods achieve the most promising results in many areas now. Observing growing big data sets and still increasing popularity of deep learning among scientists and industry, we claim that this research field will be increased constantly and provide promising results in the future.

Acknowledgment

The authors wish to thank the reviewers for their valuable remarks which helped to improve the paper.

The research project has been supported by the program *Excellence Initiative—Research University* of the AGH University of Science and Technology.

References

Alhawarat, M. and Hegazi, M. (2018). Revisiting k-means and topic modeling: A comparison study to cluster Arabic documents, *IEEE Access* **6**: 42740–42749.

Banerjee, S. and Badr, Y. (2018). Evaluating decision analytics from mobile big data using rough set based ant colony, *in* G. Mastorakis *et al.* (Eds), *Mobile Big Data*, Springer, Cham, pp. 217–231.

Bazan, J. (1996). Dynamic reducts and statistical inference, *Proceedings of the 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Granada, Spain,* Vol. 3, pp. 1147–1152.

Bazan, J., Drygaś, P., Zaręba, L. and Molenda, P. (2020). A new method of building a more effective ensemble classifiers, *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Glasgow, UK*, pp. 1–6.

Bazan, J.G. (1998). A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables, *Rough Sets in Knowledge Discovery* **1**: 321–365.

Bazan, J.G. (2008). Hierarchical classifiers for complex spatio-temporal concepts, *in* R.H. Peters *et al.* (Eds), *Transactions on Rough Sets IX*, Lecture Notes in Computer Science, Vol. 5390, Springer, Berlin/Heidelberg, pp. 474–750.

Bazan, J.G., Szczuka, M. and Wroblewski, J. (2002). A new version of rough set exploration system, *in* J.F. Peters *et al.* (Eds), *Rough Sets and Current Trends in Computing*, Springer, Berlin/Heidelberg, pp. 397–404.

Bazan, J., Peters, J. and Skowron, A. (2005). Behavioral pattern identification through rough set modelling, *10th International Conference Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Regina, Canada,* pp. 688–697.

Bello, R. and Falcon, R. (2017). Rough sets in machine learning: A review, *in* A. Skowron *et al.* (Eds), *Thriving Rough Sets*, Springer, Cham, pp. 87–118.

Bhatt, R.B. and Gopal, M. (2006). On the extension of functional dependency degree from crisp to fuzzy partitions, *Pattern Recognition Letters* **27**(5): 487–491.

Bhatt, R., Ramanna, S. and Peters, J.F. (2009). Software defect classification: A comparative study of rough-neuro-fuzzy hybrid approaches with linear and non-linear SVMs, *in* A. Abraham *et al.* (Eds.), *Rough Set Theory: A True Landmark in Data Analysis*, Studies in Computational Intelligence, Vol 174, Springer, Berlin/Heidelberg, pp. 213–231.

Błaszczyński, J., Greco, S., Matarazzo, B., Słowiński, R. and Szeląg, M. (2012). jMAF—Dominance-based rough set data analysis framework, *in* A. Skowron and Z. Suraj (Eds.), *Rough Sets and Intelligent Systems—Professor Zdzislaw Pawlak in Memoriam*, Intelligent Systems Reference Library, Vol. 42, Springer, pp. 185–209.

Breiman, L. (2001). Random forests, *Machine Learning* **45**(1): 5–32.

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M.B., Kowalik, K., Kulasekaran, S., Ludäscher, B., Mecum, B.D., Nabrzyski, J., Stodden, V., Taylor, I.J., Turk, M.J. and Turner, K. (2019). Computing environments for reproducibility: Capturing the whole tale, *Future Generation Computer Systems* **94**: 854–867.

Bulger, M., Taylor, G. and Schroeder, R. (2014). *Engaging Complexity: Challenges and Opportunities of Big Data*, NEMDOE, London.

Cessie, S. and Houwelingen, J.C. (1992). Ridge estimators in logistic regression, *Applied Statistics* **41**(1): 191–201.

Chakraborti, D.B. and Pal, S.K. (2021). *Granular Video Computing with Rough Sets, Deep Learning and IoT*, World Scientific, Singapore.

Chao, W. (2018). *High Performance Computing for Big Data: Methodologies and Applications*, CRC Press, Portland.

Chądzyńska-Krasowska, A., Stawicki, S. and 'Ślęzak, D. (2017). A metadata diagnostic framework for a new approximate query engine working with granulated data summaries, *in* L. Polkowski *et al.* (Eds), *Rough Sets*, Lecture Notes in Computer Science, Vol. 10313, Springer, Cham, pp. 623–643.

Chen, L., Li, Z., Lv, M. and Xiong, M. (2020a). Intelligent prediction algorithm of economic trend index based on rough set support vector machine, *Journal of Intelligent and Fuzzy Systems* **38**(1): 147–153.

Chen, X., Li, D., Wang, P. and Yang, X. (2020b). A deep convolutional neural network with fuzzy rough sets for FER, *IEEE Access* **8**: 2772–2779.

Chen, Q., Huang, M. and Wang, H. (2021). A feature discretization method for classification of high-resolution remote sensing images in coastal areas, *IEEE Transactions on Geoscience and Remote Sensing* **59**(10): 8584–8598.

Chen, Z. (2017). Exploring dynamic granules for time-varying big data, *IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, Orlando, USA*, pp. 1092–1097.

Chirigati, F., Rampin, R., Shasha, D. and Freire, J. (2016). ReproZip: Computational reproducibility with ease, *Proceedings of the 2016 International Conference on Management of Data, New York, USA,* pp. 2085–2088.

Chowdhury, T., Chakraborty, S. and Setua, S.K. (2016). Knowledge extraction from big data using MapReduce-based Parallel-Reduct algorithm, *5th International Conference on Computer Science and Network Technology (ICCSNT), Changchun, China,* pp. 240–246.

Chu, Z. and Zhang, Y. (2020). An accurate financially-challenged college student identification model based on rough set-bp neural networks and its application, *International Conference on Computer Information and Big Data Applications (CIBDA), Guiyang, China*, pp. 144–150.

Chen, Y., Xue, Y., Ma Y. and Xu, F. (2017). Measures of uncertainty for neighborhood rough sets, *Knowledge-Based Systems* **120**: 226–235.

Cichosz, P. (2015). *Data Mining Algorithms: Explained Using R*, Wiley, Chichester.

Cios, K. (2018). Deep neural networks—A brief history, *in* A. Gawęda *et al.* (Eds), *Advances in Data Analysis with Computational Methods: Dedicated to Professor Jacek Zurada*, Studies in Computational Intelligence, Vol. 738, Springer, Cham, pp. 183–200.

Clark, P.G. and Grzymała-Busse, J.W. (2011). Experiments on probabilistic approximations, *2011 IEEE International Conference on Granular Computing, GrC-2011, Kaohsiung, Taiwan*, pp. 144–149.

Clark, P.G., Grzymała-Busse, J.W., Hippe, Z.S., Mroczek, T. and Niemiec, R. (2020). Complexity of rule sets mined from incomplete data using probabilistic approximations based on generalized maximal consistent blocks, *in* M. Cristani *et al.* (Eds), *Knowledge-Based and Intelligent Information & Engineering Systems*, Procedia Computer Science, Vol. 176, Elsevier, Amsterdam, pp. 1803–1812.

Clark, P.G., Grzymała-Busse, J.W., Mroczek, T. and Niemiec, R. (2019). Rule set complexity in mining incomplete data using global and saturated probabilistic approximations, *in* R. Damasevicius and G. Vasiljeviene (Eds), *Information and Software Technologies*, Communications in Computer and Information Science, Vol. 1078, Springer, Cham, pp. 451–462.

Cornelis, C., Nguyen, H.S., Pal, S., Skowron, A. and Wu, W.-Z. (2015). Rough sets and fuzzy sets preface, *Fundamenta Informaticae* **142**(1–4): 5–8.

Cornelis, C., Verbiest, N. and Jensen, R. (2010). Ordered weighted average based fuzzy rough sets, *in* J. Yu *et al.* (Eds), *Rough Set and Knowledge Technology*, Springer, Berlin, pp. 78–85.

Cui, G. and Gao, H. (2019). Rough set processing outliers in cluster analysis, *4th International IEEE Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China*, pp. 111–115.

Cui, W. and Huang, L. (2015). Knowledge reduction method based on information entropy for port big data using MapReduce, *International Conference on Logistics, Informatics and Service Sciences (LISS), Barcelona, Spain*, pp. 1–6.

Czolombitko, M. and Stepaniuk, J. (2017). Scalable maximal discernibility discretization for big data, *in* L. Polkowski *et al.* (Eds), *Rough Sets*, Lecture Notes in Computer Science, Vol. 10313, Springer, Cham, pp. 644–654.

Dagdia, Z.C., Zarges, C., Beck, G., Azzag, H. and Lebbah, M. (2017). A distributed rough set theory based algorithm for an efficient big data preprocessing under the Spark framework, *IEEE International Conference on Big Data (Big Data), Boston, USA,* pp. 911–916.

Dagdia, Z.C., Zarges, C., Beck, G., Azzag, H. and Lebbah, M. (2018). A distributed rough set theory algorithm based on locality sensitive hashing for an efficient big data preprocessing, *IEEE International Conference on Big Data, Seattle, USA*, pp. 2597–2606.

D'eer, L., Cornelis, C. and Yao, Y. (2016). A semantically sound approach to Pawlak rough sets and covering-based rough sets, *International Journal of Approximate Reasoning* **78**: 62–72.

Delen, D. and Demirkan, H. (2013). Data, information and analytics as services, *Decision Support Systems* **55**(1): 359–363.

Delimata, P., Moshkov, M., Skowron, A. and Suraj, Z. (2008). Comparison of lazy classification algorithms based on deterministic and inhibitory decision rules, *3rd International Conference on Rough Sets and Knowledge Technology, Chengdu, China*, pp. 55–62.

Delimata, P., Moshkov, M., Skowron, A. and Suraj, Z. (2009). *Inhibitory Rules in Data Analysis: A Rough Set Approach*, Springer, Berlin/Heidelberg.

Delimata, P. and Suraj, Z. (2013). Hybrid methods in data classification and reduction, *in* A. Skowron and Z. Suraj (Eds), *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam*, Intelligent Systems Reference Library, Vol. 43, Springer, Berlin/Heidelberg, pp. 263–291.

Deng, Y., Ren, Z., Kong, Y., Bao, F. and Dai, Q. (2017). A hierarchical fused fuzzy deep neural network for data classification, *IEEE Transaction on Fuzzy Systems* **25**(4): 1006–1012.

Dubois, D. and Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets, *International Journal of General Systems* **17**(2–3): 191–209.

Dubois, D. and Prade, H. (1992). Putting rough sets and fuzzy sets together, *in* R. Słowiński (Ed.), *Intelligent Decision Support*, Theory and Decision Library, Vol. 11, Springer, Dordrecht, pp. 203–232.

El-Bably, M. and Kozae, A. (2014). New generalized definitions of rough membership relations and functions from topological point of view, *Journal of Advances in Mathematics* **8**(3): 1635–1652.

Ester, M., Kriegel, H., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise, *2nd International Conference on Knowledge Discovery and Data Mining, Portland, USA,* pp. 226–231.

Ganivada, A. and Pal, S.K. (2011). A novel fuzzy rough granular neural network for classification, *International Journal of Computational Intelligence Systems* **4**(5): 1042–1051.

GARTNER (2001). Big data, *Gartner Glossary*, Gartner, Inc., Stamford, `https://www.gartner.com/en/information-technology/glossary/big-data`.

Google Scholar (2021). Rough sets, *Search,* `https://scholar.google.com/scholar?hl=pl&as_sdt=2007&q=allintitle%3A+%22Rough+sets%22&btnG=`.

Gordon, A.D., Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). Classification and regression trees, *Biometrics* **40**(3): 874.

Greco, S., Matarazzo, B., Słowiński, R. and Stefanowski, J. (2001). Variable consistency model of dominance-based rough sets approach, *in* W. Ziarko and Y. Yao (Eds.), *Rough Sets and Current Trends in Computing*, Lecture Notes in Computer Science, Vol. 2005, Springer, Berlin/Heidelberg, pp. 170–181.

Greco, S., Matarazzo, B. and Słowinski, R. (1999). Rough approximation of a preference relation by dominance relations, *European International Operational Research* **117**(1): 63–83.

Grzegorowski, M., Janusz, A., Ślęzak, D. and Szczuka, M. (2017). On the role of feature space granulation in feature selection processes, *IEEE International Conference on Big Data (Big Data), Boston, USA,* pp. 1806–1815.

Grzymała-Busse, J.W. (1992). LERS—A system for learning from examples based on rough sets, *in* R. Słowiński (Ed.), *Intelligent Decision Support*, Theory and Decision Library, Vol. 11, Springer, Dordrecht, pp. 3–18.

Grzymała-Busse, J.W. (1997). A new version of the rule induction system LERS, *Fundamenta Informaticae* **31**(1): 27–39.

Hamidinekoo, A., Dagdia, Z.C., Suhail, Z. and Zwiggelaar, R. (2018). Distributed rough set based feature selection approach to analyse deep and hand-crafted features for mammography mass classification, *IEEE International Conference on Big Data, Seattle, USA*, Vol. 1, pp. 2423–2432.

Hassan, Y.F. (2017). Deep learning architecture using rough sets and rough neural networks, *Kybernetes* **46**(4): 693–705.

Hirota, K. and Pedrycz, W. (1999). Fuzzy computing for data mining, *Proceedings of the IEEE* **87**(9): 1575–1600.

Hong-Wei, Y. and Xindi, T. (2016). Based on rough sets and L1 regularization of the fault diagnosis of linear regression model, *International Conference on Intelligent Transportation, Changsha, China,* pp. 490–492.

Hu, F. and Wang, G. (2008). Huge data mining based on rough set theory and granular computing, *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia*, pp. 655–658.

Hu, Q., Li, L. and Zhu, P. (2013). Exploring neighborhood structures with neighborhood rough sets in classification learning, *in* A. Skowron and Z. Suraj (Eds), *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam*, Springer, Berlin/Heidelberg, pp. 277–307.

Hu, Q., Zhang, L., Zhou, Y. and Pedrycz, W. (2017a). Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets, *IEEE Transactions on Fuzzy Systems* **26**(1): 226–238.

Hu, Q., Zhang, L., Zhou, Y. and Pedrycz, W. (2018). Large-scale multimodality attribute reduction with multi-kernel fuzzy rough sets, *IEEE Transactions on Fuzzy Systems* **26**(1): 226–238.

Hu, J., Li, T., Luo, C. and Li, S. (2017b). Incremental fuzzy probabilistic rough sets over two universes, *International Journal of Approximate Reasoning* **81**: 28–48.

Huang, Q., Li, T., Huang, Y., Yang, X. and Fujita, H. (2020). Dynamic dominance rough set approach for processing composite ordered data, *Knowledge Based Systems* **187**: 104829.

Huang, Y., Li, T., Luo, C., Fujita, H. and jinn Horng, S. (2017). Dynamic variable precision rough set approach for probabilistic set-valued information systems, *Knowledge-Based Systems* **122**(5): 131–147.

Ilczuk, G. and Wakulicz-Deja, A. (2005). Rough sets approach to medical diagnosis system, *in* P.S. Szczepaniak *et al.* (Eds), *Advances in Web Intelligence*, Lecture Notes in Computer Science, Vol. 3528, Springer, Berlin/Heidelberg, pp. 204–210.

Inuiguchi, M. (2013). Rough representations of ill-known sets and their manipulations in low dimensional space, *in* A. Skowron and Z. Suraj (Eds), *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam*, Springer, Berlin/Heidelberg, pp. 309–331.

Ishizu, S., Gehrmann, A. and Nagaw, Y.Y. (2007). Rough ontology: Extension of ontologies by rough sets, *in* G. Smith and M.J. Salvendy (Eds), *Human Interface and the Management of Information: Methods, Techniques and Tools in Information Design*, Springer, Berlin/Heidelberg, pp. 456–462.

Isitor, E. and Stanier, C. (2016). Defining big data, *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies, New York, USA*, pp. 1–6.

Islam, M., Inan, T., Rafi, S., Akter, S., Sarker, I.H. and Islam, A. (2020). A systematic review on the use of AI and ML for fighting the COVID-19 pandemic, *IEEE Transactions on Artificial Intelligence* **1**(3): 258–270.

Ivanov, T., Korfiatis, N., Zicari, R. (2013). On the inequality of the 3V's of big data architectural paradigms: A case for heterogeneity, *arXiv* abs/1311.0805.

Ivie, P. and Thain, D. (2018). Reproducibility in scientific computing, *ACM Computing Surveys* **51**(3): 1–36.

Izakian, H., Pedrycz, W. and Jamal, I. (2015). Fuzzy clustering of time series data using dynamic time warping distance, *Engineering Applications of Artificial Intelligence* **39**: 235–244.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*, Prentice-Hall, Upper Saddle River.

Janusz, A. and Ślęzak, D. (2014). Rough set methods for attribute clustering and selection, *Applied Artificial Intelligence* **28**(3): 220–242.

Jensen, R. (2014). Fuzzy-rough data mining with WEKA, *Tutorial*, `http://users.aber.ac.uk/rkj/Weka.pdf`.

Jensen, R. and Cornelis, C. (2008). A new approach to fuzzy-rough nearest neighbour classification, *in* C.C. Chan *et al.* (Eds), *Rough Sets and Current Trends in Computing*, Lecture Notes in Computer Science, Vol. 5306, Springer, Berlin/Heidelberg, pp. 310–319.

Jensen, R., Cornelis, C. and Shen, Q. (2009). Hybrid fuzzy-rough rule induction and feature selection, *2009 IEEE International Conference on Fuzzy Systems, Jeju, South Korea*, pp. 1151–1156.

Jian, Z., Sakai, H., Watada, J., Roy, A., Hilmi, M. and Hassan, B. (2019). An Apriori-based data analysis on suspicious network event recognition, *IEEE International Conference on Big Data (Big Data): Suspicious Network Event Recognition, Los Angeles, USA*, pp. 5888–5896.

Jing, S., Yang, J. and She, K. (2014). A parallel method for rough entropy computation using MapReduce, *10th International Conference on Computational Intelligence and Security, Kunming, China*, pp. 707–710.

Jingjing, J., Hongzhe, X. and Zhuangzhuang, S. (2019). Application of attribute reduction algorithm of rough set based on mix_fp tree in computer teaching, *2nd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China*, pp. 78–81.

jMAF (2021). *jMAF—Rough Set Data Analysis Framework*, `http://www.cs.put.poznan.pl/jblaszczynski/Site/jRS.html`.

John, G.H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers, *11th Conference on Uncertainty in Artificial Intelligence, Montreal, Canada*, pp. 338–345.

Kang, X., Liu, X. and Zhai, M. (2011). Instances selection for NN with fuzzy rough technique, *International Conference on Machine Learning and Cybernetics, Guilin, China*, pp. 1097–1100.

Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Khan, M.A. and Banerjee, M. (2013). Algebras for information systems, *in* A. Skowron and Z. Suraj (Eds), *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam*, Intelligent Systems Reference Library, Vol. 42, Springer, Berlin/Heidelberg, pp. 381–407.

Kong, L., Qu, W., Yu, J., Zuo, H., Chen, G., Xiong, F., Pan, S. and Siyu Lin, M.Q. (2020). Distributed feature selection for big data using fuzzy rough sets, *IEEE Transactions on Fuzzy Systems* **28**(5): 846–857.

Kopczyński, M., Grześ, T. and Stepaniuk, J. (2016). Core for large datasets: Rough sets on FPGA, *Fundamenta Informaticae* **147**(2–3): 241–259.

Kopczyński, M., Grześ, T. and Stepaniuk, J. (2017). Hardware supported rule-based classification on big datasets, *in* L. Polkowski *et al.* (Eds), *Rough Sets*, Lecture Notes in Computer Science, Vol. 10313, Springer, Cham, pp. 655–668.

Krishnamurthy, S. and Janardanan, A. (2018). Rough set based ontology matching, *International Journal of Rough Sets and Data Analysis* **5**(2): 46–68.

Kundu, S. and Pal, S.K. (2018). Double bounded rough set, tension measure, and social link prediction, *IEEE Transactions on Computational Social Systems* **5**(3): 841–853.

Kune, R. (2014). Genetic algorithm based data-aware group scheduling for big data clouds, *BDC'14: Proceedings of the 2014 IEEE/ACM International Symposium on Big Data Computing, London, UK*, pp. 96–104.

Kwiatkowski, P., Hoa, N. and Nguyen, H.S. (2010). On scalability of rough set methods, *in* E. Hüllermeier *et al.* (Eds), *Information Processing and Management of Uncertainty in Knowledge-Based Systems: Theory and Methods*, Communications in Computer and Information Science, Vol. 80, Springer, Berlin/Heidelberg, pp. 288–297.

Landwehr, N., Hall, M. and Frank, E. (2005). Logistic model trees, *Machine Learning* **59**: 161–205.

Lenz, O.U., Peralta, D. and Cornelis, C. (2020). Fuzzy-rough-learn 0.1: A Python library for machine learning with fuzzy rough sets, *in* R. Bello *et al.* (Eds), *Rough Sets*, Lecture Notes in Computer Science, Vol. 12179, Springer, Cham, pp. 491–499.

Li, D., Zhang, H., Li, T., Bouras, A. and Wang, X.Y.T. (2021). Hybrid missing value imputation algorithms using fuzzy c-means and vaguely quantified rough set, *IEEE Transactions on Fuzzy Systems*, DOI: 10.1109/TFUZZ.2021.3058643, (early access).

Li, T., Luo, C., Chen, H. and Zhang, J. (2015). PICKT: A solution for big data analysis, *in* D. Ciucci *et al.* (Eds), *Rough Sets and Knowledge Technology*, Lecture Notes in Computer Science, Vol. 9436, Springer, Cham, pp. 15–25.

Li, X., Li, X. and Zhao, Z. (2016). Combining deep learning with rough set analysis: A model of cyberspace situational awareness, *6th International Conference on Electronics Information and Emergency Communication (ICEIEC), Beijing, China*, pp. 182–185.

Li, X., Luo, C., Liu, P. and Wang, L. (2019). Information entropy differential privacy: a differential privacy protection data method based on rough set theory, *IEEE International Conference on Dependable, Autonomic and Secure Computing/Pervasive Intelligence and Computing/Cloud and Big Data Computing/Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan*, pp. 918–923.

Li, X. and Shen, Y. (2020). Discretization algorithm for incomplete economic information in rough set based on big data, *Symmetry* **12**(8): 1245.

Lin, G., Liang, J. and Qian, Y. (2015). An information fusion approach by combining multigranulation rough sets and evidence theory, *Information Sciences* **314**(1): 184–199.

Lin, G., Liang, J., Qian, Y. and Li, J. (2016). A fuzzy multigranulation decision-theoretic approach to multi-source fuzzy information systems, *Knowledge-Based Systems* **91**: 102–113.

Liu, P. and Zhang, G. (2019). Research on key attributes of learning behavior based on rough set, *14th International Conference on Computer Science & Education (ICCSE), Toronto, Canada*, pp. 1030–1034.

Lu, Z., Liu, K., Liu, Z., Wang, C., Shen, M. and Xu, T. (2019). An efficient annotation method for big data sets of high-resolution earth observation images, *ICBDT 2019: Proceedings of the 2nd International Conference on Big Data Technologies, Jinan, China*, pp. 240–243.

Luo, C., Li, T., Chen, H., Fujita, H. and Yi, Z. (2016). Efficient updating of probabilistic approximations with incremental objects, *Knowledge-Based Systems* **109**(C): 71–83.

Luo, C., Li, T., Chen, H., Fujita, H. and Yi, Z. (2018). Incremental rough set approach for hierarchical multicriteria classification, *Information Sciences* **429**: 72–87.

Lv, Z., Liu, T., Shi, C., Benediktsson, J.A. and Du, H. (2019). A novel land cover change detection method based on k-means clustering and adaptive majority voting using bitemporal remote sensing images, *IEEE Access* **7**: 34425–34437.

Madrid, N., Medina, J. and Ramírez-Poussa, E. (2020). Rough sets based on Galois connections, *International Journal of Applied Mathematics and Computer Science* **30**(2): 299–313, DOI: 10.34768/amcs-2020-0023.

Meng, Y., Liang, J., Cao, F. and He, Y. (2018). A new distance with derivative information for functional k-means clustering algorithm, *Information Sciences* **463**: 166–185.

Mondelli, M.L., Peterson, A. and Gadelha, L. (2019). Exploring reproducibility and FAIR principles in data science using ecological niche modeling as a case study, *in* G. Guizzardi *et al.* (Eds), *Advances in Conceptual Modeling*, Lecture Notes in Computer Science, Vol. 11787, Springer, Cham, pp. 23–33.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms, *Computer Journal* **26**(4): 354–359.

Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: An overview, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**(1): 86–97.

Narayanan, U., Paul, V. and Joseph, S. (2017). Different analytical techniques for big data analysis: A review, *International Conference on Energy, Chennai, India,* pp. 372–382.

Nguyen, H. (1997). *Discretization of Real Value Attributes, Boolean Reasoning Approach*, PhD thesis, University of Warsaw, Warsaw.

Nguyen, H.S. (1998). Discretization methods in data mining, *in* L. Polkowski (Ed.), *Rough Sets in Knowledge Discovery*, Physica-Verlag, Heidelberg, pp. 451–482.

Nguyen, H.S. (2006). Approximate Boolean reasoning: Foundations and applications in data mining, *in* J.F. Peters and A. Skowron (Eds), *Transactions on Rough Sets*, Lecture Notes in Computer Science, Vol. 4100, Springer, Berlin/Heidelberg, pp. 334–506.

Nguyen, H.S., Nguyen, N.T., Nguyen, H.S. and Nguyen, L. (2017). Some observations on representation of dependency degree k, *9th International Conference on Knowledge and Systems Engineering (KSE), Hue, Vietnam,* pp. 13–17.

Naouali, S. and Missaoui, R. (2005). Flexible query answering in data cubes, *in* A.M. Tjoa and J. Trujillo (Eds), *Data Warehousing and Knowledge Discovery*, Lecture Notes in Computer Science, Vol 3589, Springer, Berlin/Heidelberg, pp. 221–232.

Nowak-Brzezińska, A. and Wakulicz-Deja, A. (2019). Exploration of rule-based knowledge bases: A knowledge engineer's support, *Information Sciences* **485**(2): 301–318.

Ohrn, A. (2000). ROSETTA: Technical reference manual, *Technical report*, Norwegian University of Science and Technology, Trondheim.

Ohrn, A. and Komorowski, J. (1997). ROSETTA: A rough set toolkit for analysis of data, *Workshop on Rough Sets and Soft Computing (RSSC'97), Durham, USA,* Vol. 3, pp. 403–407.

Pal, S. (2021). Rough set and deep learning: Some concepts, *Academia Letters* (1849): 1–6, DOI: 10.20935/AL1849.

Pal, S.K. (2020). granular mining and big data analytics: Rough models and challenges, *Proceedings of the National Academy of Sciences, India, A: Physical Sciences* **90**: 193–208.

Pal, S.K., Bhoumik, D. and Chakraborty, D. (2019). Granulated deep learning and Z-numbers in motion detection and object recognition, *Neural Computing and Applications* **32**(4): 1–16.

Pal, S.K. and Kundu, S. (2017). Granular social network: Model and applications, *in* A.Y. Zomaya and S. Sakr (Eds), *Handbook of Big Data Technologies*, Springer, Cham, pp. 617–651.

Pal, S.K. and Meher, S.K. (2013). Natural computing: A problem solving paradigm with granular information processing, *Applied Soft Computing* **13**(9): 3944–3955.

Pal, S.K., Polkowski, L. and Skowron, A. (2004). *Rough-Neural Computing-Techniques for Computing with Words*, Springer, Berlin.

Pancerz, K. and Suraj, Z. (2013). A rough set approach to information systems decomposition, *Fundamenta Informaticae* **127**(1–4): 257–272.

Pandu, S. (2020). MapReduce based improved quick reduct algorithm with granular refinement using vertical partitioning scheme, *Knowledge-Based Systems* **189**: 1872–7409.

Pawlak, Z. (1991). *RoughSets and Theoretical Aspects of Reasoning About Data*, Kluwer, Dordrecht.

Pawlak, Z. and Skowron, A. (2007). Rough sets and Boolean reasoning, *Information Sciences* **177**(1): 41–73.

Pedrycz, W. and Bargiela, A. (2002). Granular clustering: A granular signature of data, *IEEE Transactions on Systems, Man, and Cybernetics B: Cybernetics* **32**(2): 212–224.

Pedrycz, W., Gacek, A. and Wang, X. (2015a). Clustering in augmented space of granular constraints: A study in knowledge-based clustering, *Pattern Recognition Letters* **67**: 122–129.

Pedrycz, W. and Gomide, F. (1994). A generalized fuzzy Petri net model, *IEEE Transactions on Fuzzy Systems* **2**(4): 295–301.

Pedrycz, W., Succi, G., Sillitti, A. and Iljazi, J. (2015b). Data description: A general framework of information granules, *Knowledge-Based Systems* **80**: 98–108.

Pedrycz, W. and Vukovich, G. (2001). Granular neural networks, *Neurocomputing* **36**(1): 205–224.

Pedrycz, W., Zhao, J., Jing, X. and Yan, Z. (2021). Network traffic classification for data fusion: A survey, *Information Fusion* **72**: 22–47.

Peters, J. (2013). How near are Zdzisław Pawlak's paintings? Study of merotopic distances between digital picture regions-of-interest, *in* A. Skowron and Z. Suraj (Eds), *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam*, Intelligent Systems Reference Library, Vol 42, Springer, Berlin/Heidelberg, pp. 545–568.

Peters, J. (2020). Computational geometry, topology and physics of visual scenes, *in* J.F. Peters (Ed.), *Computational Geometry, Topology and Physics of Digital Images with Applications*, Springer, Cham, pp. 1–85.

Peters, J.F., Skowron, A. and Suraj, Z. (2000). An application of rough set methods in control design, *Fundamenta Informaticae* **43**(1–4): 269–290.

Peters, J.F., Ziaei, K., Ramanna, S. and Ehikioya, S.A. (1998). Adaptive fuzzy rough approximate time controller design methodology: Concepts, Petri net model and application, *IEEE International Conference on Systems, Man, and Cybernetics, San Diego, USA*, Vol. 3, pp. 2101–2106.

Peters, J. and Naimpally, S. (2012). Applications of near sets, *Notices of the American Mathematical Society* **59**: 536–542.

Pierzchała, D. (2014). Application of ontology and rough set theory to information sharing in multi-resolution combat, *in* J. Sobecki *et al.* (Eds), *Advanced Approaches to Intelligent Information and Database Systems*, Springer, Cham, pp. 193–203.

Pięta, P., Szmuc, T. and Kluza, K. (2019). Comparative overview of rough set toolkit systems for data analysis, *3rd International Conference of Computational Methods in Engineering Science (CMES18), Kazimierz Dolny, Poland,* pp. 1–7.

Polkowski, L. (2005). Rough-fuzzy-neurocomputing based on rough mereological calculus of granules, *International Journal of Hybrid Intelligent Systems* **2**(2): 91–108.

Polkowski, L. (2011). *Approximate Reasoning by Parts: An Introduction to Rough Mereology*, Springer, Berlin/Heidelberg.

Polkowski, L. (2020). On the compactness property of mereological spaces, *Fundamenta Informaticae* **172**(1): 73–95.

Polkowski, L. and Osmialowski, P. (2010). Navigation for mobile autonomous robots and their formations: An application of spatial reasoning induced from rough mereological geometry, *in* A. Barrera (Ed.), *Mobile Robots Navigation*, IntechOpen, London, DOI: 10.5772/8987, `https://www.intechopen.com/chapters/10248`.

Polkowski, L. and Skowron, A. (1996). Rough mereological approach to knowledge-based distributed AI, *3rd World Congress on Expert Systems, Seoul, Korea,* pp. 774–781.

Polkowski, L. and Skowron, A. (2000). Rough mereology in information systems with applications to qualitative spatial reasoning, *Fundamenta Informaticae* **43**(1–4): 291–320.

Prędki, B., Słowiński, R., Stefanowski, J., Susmaga, R. and Wilk, S. (1998). ROSE—Software implementation of the rough set theory, *in* A.S. Polkowski (Ed.), *Rough Sets and Current Trends in Computing*, Springer, Berlin, pp. 605–608.

Prędki, B. and Wilk, S. (1999). Rough set based data exploration using ROSE system, *in* A.S.Z.W. Ras (Ed.), *Foundations of Intelligent Systems*, Springer, Berlin, pp. 172–180.

Przyborowski, M., Tajmajer, T., Grad, L., Janusz, A., Biczyk, P. and Ślęzak, D. (2018). Toward machine learning on granulated data: A case of compact autoencoder-based representations of satellite images, *2018 IEEE International Conference on Big Data, Seattle, USA*, pp. 2657–2662.

Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2013). Application of reduction of the set of conditional attributes in the process of global decision-making, *Fundamenta Informaticae* **122**(4): 327–355.

Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2014). A dispersed decision-making system—The use of negotiations during the dynamic generation of a systems structure, *Information Sciences* **288**(1): 194–219.

Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2016a). Global decision-making in multi-agent decision-making system with dynamically generated disjoint clusters, *Applied Software Computing* **40**: 603–615.

Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2016b). The strength of coalition in a dispersed decision support system with negotiations, *European Journal of Operational Research* **252**(3): 947–968.

Przybyła-Kasperek, M. and Wakulicz-Deja, A. (2017). Comparison of fusion methods from the abstract level and the rank level in a dispersed decision-making system, *International Journal of General Systems* **46**(4): 1–28.

Qiana, Y., Lianga, X., Lin, G., Guo, Q. and Liang, J. (2017). Local multigranulation decision-theoretic rough sets, *International Journal of Approximate Reasoning* **82**(6): 119–137.

Qiong, C., Huang, M., Wang, H. and Xu, G. (2021). A feature discretization method based on fuzzy rough sets for high-resolution remote sensing big data under linear spectral model, *IEEE Transactions on Fuzzy Systems*: 1–1, (early access).

Qu, W., Kong, L., Wu, K., Tang, F. and Chen, G. (2019). Distributed fuzzy rough set for big data analysis in cloud computing, *IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), Tianjin, China*, pp. 109–116.

Quinlan, J.R. (1983). Learning efficient classification procedures and their application to chess end games, *in* R.S. Michalski *et al.* (Eds), *Machine Learning: Symbolic Computation*, Springer, Heidelberg, pp. 463–482.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Francisco.

Rajesh, T. and Malar, R.S.M. (2013). Rough set theory and feed forward neural network based brain tumor detection in magnetic resonance images, *International Conference on Advanced Nanomaterials & Emerging Engineering Technologies, Chennai, India*, pp. 240–244.

Riaz, S., Arshad, A. and Jiao, L. (2019). A semi-supervised CNN with fuzzy rough c-mean for image classification, *IEEE Access* **7**: 49641–49652.

ROSE (1998). *ROSE—Rough Set Data Explorer,* `https://idss.cs.put.poznan.pl/site/rose.html`.

ROSETTA (1994). *ROSETTA—A Rough Set Toolkit for Analysis of Data,* `http://bioinf.icm.uu.se/rosetta/downloads.php`.

RSES (2005). *RSES—Rough Set Exploration System,* `https://www.mimuw.edu.pl/~szczuka/rses`.

RSlib (2019). *Rseslib 3—Rough Set and Machine Learning Open Source in Java,* `http://rsproject.mimuw.edu.pl/help.html`.

Riza, L.S., Janusz, A., Bergmeir, C., Cornelis, C., Herrera, F., Ślęzak, D. and Benítez, J.M. (2014). Implementing algorithms of rough set theory and fuzzy rough set theory in the R package RoughSets, *Information Sciences* **287**: 68–89.

Sachin, J. and Shubhangi, S. (2015). Knowledge acquisition using parallel rough set and MapReduce from big data, *International Conference on Information Processing (ICIP), Pune, India,* pp. 16–20.

Sakai, H., Nakata, M. and Watada, J. (2020). NIS—Apriori-based rule generation with three-way decisions and its application system in SQL, *Information Sciences* **507**: 755–771.

Sedkaoui, S. (2018). *Data Analytics and Big Data*, Wiley, Hoboken.

Shan, H., Xiaoning, J. and Jianxun, L. (2016). An assessment method for the impact of missing data in the rough set-based decision fusion, *Intelligent Data Analysis* **20**(6): 1267–1284.

Shi, W., Gong, Y., Ding, C., Ma, Z., Tao, X. and Zheng, N. (2015). Transductive semi-supervised deep learning using min-max features, *Proceedings of the European Conference on Computer Vision (ECCV), Cham, Switzerland*, pp. 299–315.

Simiński, R. and Wakulicz-Deja, A. (2003). Decision units as a tool for rule base modeling and verification, *in* M.A. Klopotek *et al.* (Eds), *Intelligent Information Processing and Web Mining*, Springer, Berlin/Heidelberg, pp. 553–556.

Sinaga, K.P. and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm, *IEEE Access* **8**: 80716–80727.

Skowron, A., Bazan, J. and Wojnarski, M. (2009). Interactive rough-granular computing in pattern recognition, *in* S. Chaudhury *et al.* (Eds.), *Pattern Recognition and Machine Intelligence*, Lecture Notes in Computer Science, Vol. 5909, Springer, Berlin/Heidelberg, pp. 92–97.

Skowron, A. and Dutta, S. (2017). From information systems to interactive information systems, *in* G. Wang *et al.* (Eds), *Thriving Rough Sets*, Studies in Computational Intelligence, Vol. 708, Springer, Cham, pp. 207–223.

Skowron, A. and Dutta, S. (2018). Rough sets: Past, present, and future, *Natural Computing* **17**: 855–876.

Skowron, A., Jankowski, A. and Dutta, S. (2016). Toward problem solving support based on big data and domain knowledge: Interactive granular computing and adaptive judgement, *in* N. Japkowicz and J. Stefanowski (Eds), *Big Data Analysis: New Algorithms for a New Society,* Studies in Big Data, Vol. 16, Springer, Cham, pp. 49–90.

Skowron, A. and Nguyen, H.S. (1999). Boolean reasoning scheme with some applications in data mining, *in* J.M. Żytkow and J. Rauch (Eds), *Principles of Data Mining and Knowledge Discovery*, Springer, Berlin, pp. 107–115.

Skowron, A., Ramanna, S. and Peters, J. (2006). Conflict analysis and information systems: A rough set approach, *in* G.Y. Wang *et al.* (Eds), *Rough Sets and Knowledge Technology*, Lecture Notes in Computer Science, Vol. 4062, Springer, Berlin/Heidelberg, pp. 233–240.

Skowron, A. and Stepaniuk, J. (1996). Tolerance approximation spaces, *Fundamenta Informaticae* **27**(2–3): 245–253.

Skowron, A. and Stepaniuk, J. (2001). Information granules: Towards foundations of granular computing, *International Journal of Intelligent Systems* **16**(1): 57–85.

Skowron, A. and Suraj, Z. (Eds) (2013a). *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam, Volume 1*, Springer, Berlin/Heidelberg.

Skowron, A. and Suraj, Z. (Eds) (2013b). *Rough Sets and Intelligent Systems—Professor Zdzisław Pawlak in Memoriam, Volume 2*, Springer, Berlin/Heidelberg.

Skowron, A. and Wasilewski, P. (2011a). Information systems in modeling interactive computations on granules, *Theoretical Computer Science* **412**(42): 5939–5959.

Skowron, A. and Wasilewski, P. (2011b). Toward interactive rough-granular computing, *Control and Cybernetics* **40**(2): 213–235.

Skowron, A., Yuhua, Q., Xinyan, L., Qi, W., Liang, J., Bing, L., Yiyu, Y., Jianmin, M. and Dang, C. (2018). Local rough set: A solution to rough data analysis in big data, *International Journal of Approximate Reasoning* **97**: 38–63.

Skowron, A. (2001). Approximate reasoning by agents in distributed environments, *Proceedings of the 2nd Asia-Pacific Conference on Intelligent Agent Technology, Maebashi, Japan*, pp. 28–39.

Ślęzak, D., Glick, R., Betliński, P. and Synak, P. (2018). A new approximate query engine based on intelligent capture and fast transformations of granulated data summaries, *Journal of Intelligent Information Systems* **50**: 385–414.

Ślęzak, D., Synak, P., Toppin, G., Wróblewski, J. and Borkowski, J. (2012). Rough SQL—Semantics and execution, *in* S. Greco *et al.* (Eds), *Advances in Computational Intelligence*, Communications in Computer and Information Science, Vol. 298, Springer, Berlin/Heidelberg, pp. 570–579.

Ślęzak, D. and Eastwood, V. (2009). Data warehouse technology by Infobright, *Proceedings of SIGMOD, Providence, USA*, pp. 841–846.

Ślęzak, D., Grzegorowski, M., Janusz, A., Kozielski, M., Nguyen, S.H., Sikora, M., Stawicki, S. and Wróbel, Ł. (2018). A framework for learning and embedding multi-sensor forecasting models into a decision support system: A case study of methane concentration in coal mines, *Information Sciences* **451–452**: 112–133.

Ślęzak, D., Synak, P., Wróblewski, J. and Toppin, G. (2010). Infobright analytic database engine using rough sets and granular computing, *IEEE International Conference on Granular Computing, San Jose, USA,* pp. 432–437.

Ślęzak, D., Wróblewski, J., Eastwood, V. and Synak, P. (2008). Brighthouse: An analytic data warehouse for ad-hoc queries, *Proceedings of VLDB Endowment* **1**(2): 1337–1345.

Stefanowski, J., Krawiec, K. and Wrembel, R. (2017). Exploring complex and big data, *International Journal of Applied Mathematics and Computer Science* **27**(4): 669–679, DOI: 10.1515/amcs-2017-0046.

Sulaiman, S., Shamsuddin, S.M. and Abraham, A. (2009). Rough web caching, *in* A. Abraham *et al.* (Eds), *Rough Set Theory: A True Landmark in Data Analysis*, Studies in Computational Intelligence, Vol. 174, Springer, Berlin/Heidelberg, pp. 187–211.

Sun, Y.Q., Wu, L.Y., Zeng, Y. (2019). A decision-making method for weapon demonstration based on fuzzy theory and Bayesian rough sets, *ICMSS 2019: Proceedings of the 3rd International Conference on Management Engineering, Wuhan, China*, p. 74–77.

Sun, L., Yin, T., Ding, W., Qian, Y. and Xu, J. (2021). feature selection with missing labels using multilabel fuzzy neighborhood rough sets and maximum relevance minimum redundancy, *IEEE Transactions on Fuzzy Systems*: 1–1, (early access).

Suraj, Z., Grochowalski, P. and Bandyopadhyay, S. (2015). Optimization of backward fuzzy reasoning based on rule knowledge, *Proceedings of the International Workshop on Concurrency, Specification and Programming, Rzeszów, Poland*, pp. 177–186.

Tang, J., Wang, J. and Wu, C. (2019). Research progress on network public opinion based on rough sets from the big data perspective, *IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, Chongqing, China*, pp. 1074–1077.

Thuy, N.N. and Wongthanavasu, S. (2021). A novel feature selection method for high-dimensional mixed decision tables, *IEEE Transactions on Neural Networks and Learning Systems*: 1–14, (early access).

UCI (2021). *UCI Machine Learning Repository,* `https://archive.ics.uci.edu/ml/index.php`.

Venkatraman, R. and Venkatraman, S. (2019). Big data infrastructure, *Proceedings of the 3rd International Conference on Big Data and Internet of Things, BDIOT 2019, New York, USA,* pp. 13–17.

Verbiest, N., Cornelis, C. and Herrera, F. (2013). OWA-FRPS: A prototype selection method based on ordered weighted average fuzzy rough set theory, *in* D. Ciucci *et al.* (Eds), *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Lecture Notes in Computer Science, Vol 8170, Springer, Berlin/Heidelberg. pp. 180–190.

Vluymans, S., Asfoor, H., Saeys, Y., Cornelis, Y., Tolentino, M., Teredesai, A. and De Cock, M. (2015). Distributed fuzzy rough prototype selection for big data regression, *Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS)/5th World Conference on Soft Computing (WConSC), Redmond, USA*, pp. 1–6.

Vluymans, S., Cornelis, C., Herrera, F. and Saeys, Y. (2018a). Multi-label classification using a fuzzy rough neighborhood consensus, *Information Sciences* **433–434**(9): 96–114.

Vluymans, S., Fernández, A., Saeys, Y., Cornelis, C. and Herrera, F. (2018b). Dynamic affinity-based classification of multi-class imbalanced data with one-versus-one decomposition: A fuzzy rough set approach, *Knowledge and Information Systems* **56**(1): 55–84.

Wakulicz-Deja, A., Boryczka, M. and Paszek, P. (1998). Discretization of continuous attributes on decision system in mitochondrial encephalomyopathies, *1st International Conference, RSCTC'98, Warsaw, Poland*, pp. 483–490.

Wakulicz-Deja, A., Nowak-Brzezińska, A. and Jach, T. (2011). Inference processes in decision support systems with incomplete knowledge, *in* J. Yao *et al.* (Eds), *Rough Sets and Knowledge Technology,* Springer, Berlin/Heidelberg, pp. 616–625.

Wakulicz-Deja, A., Nowak-Brzezińska, A. and Przybyła-Kasperek, M. (2013). Complex decision systems and conflicts analysis problem, *Fundamenta Informaticae* **127**(1–4): 341–356.

Wakulicz-Deja, A. and Przybyła-Kasperek, M. (2016). Pawlak's conflict model: Directions of development, *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016, Gdańsk, Poland*, pp. 191–197.

Wan, R. and Li, Y. (2019). Clustering data stream with rough set, *ICCPR'19: Proceedings of the 8th International Conference on Computing and Pattern Recognition, Beijing, China*, pp. 52–56.

Wang, G., Li, T., Zhang, P., Huang, Q. and Chen, H. (2021). Double-local rough sets for efficient data mining, *Information Sciences* **571**(2–3): 475–498.

Wang, S., Li, T., Luo, C. and Fujita, H. (2016a). Efficient updating rough approximations with multidimensional variation of ordered data, *Information Sciences* **372**(7): 690–708.

Wang, X., Wang, L., Li, Y., Wang, B., Hei, X. and Cao, Z. (2016b). A quick algorithm for rule acquisition based on distributed domputing, *IEEE International Conference on Smart Cloud, New York, USA,* pp. 278–281.

Watt, J., Borhani, R. and Katsaggelos, A.K. (2016). *Machine Learning Refined: Foundations, Algorithms, and Applications*, Cambridge University Press, Cambridge.

Wei, W. and Liang, J. (2019). Information fusion in rough set theory: An overview, *Information Fusion* **48**: 107–118.

WEKA (2009). *WEKA—Waikato Environment for Knowledge Analysis*, http://users.aber.ac.uk/rkj/book/wekafull.jar.

Wilkinson, M., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Bonino da Silva Santos, L.O., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R. and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship, *Scientific Data* **3**: 1–9, Article 160018.

Wnuk, M., Stawicki, S. and Ślęzak, D. (2020). Reinventing infobright's concept of rough calculations on granulated tables for the purpose of accelerating modern data processing frameworks, *IEEE International Conference on Big Data (Big Data), Atlanta, USA*, pp. 5405–5412.

Wu, W.-Z. and Leung, Y. (2011). Theory and applications of granular labelled partitions in multi-scale decision tables, *Information Sciences* **181**(18): 3878–3897.

Wu, H.S.P. and Prasad S. (2018). Semi-supervised deep learning using pseudo labels for hyperspectral image classification, *IEEE Transactions on Image Processing* **27**(3): 1259–1270.

Xia, S., Zhang, Z., Li, W., Wang, G., Giem, E. and Chen, Z. (2020). GBNRS: A novel rough set algorithm for fast adaptive attribute reduction in classification, *IEEE Transactions on Knowledge and Data Engineering*: 1–1, (early access).

Xiaoguang, Y., Qisong, Z. and Guojun, S. (2018). Research on classification of LBS service facilities based on rough sets neural network, *Chinese Control and Decision Conference (CCDC), Shenyang, China,* pp. 2843–2848.

Xie H., Hu, X., Peng, Z., Yao, X. and Chen, Y. (2018). A method of electricity consumption behavior analysis based on rough set fuzzy clustering, *2nd IEEE Conference on Energy Internet and Energy System Integration (EI2), Beijing, China*, pp. 1–5.

Xu, W. and Yu, J. (2017). A novel approach to information fusion in multi-source datasets: A granular computing viewpoint, *Information Sciences* **378**(C): 410–423.

Yang, X., Li, T., Dun, D.L., Chen, H. and Luo, C. (2017). A unified framework of dynamic three-way probabilistic rough sets, *Information Sciences* **420**(C): 126–147.

Yao, Y. (2007). Decision-theoretic rough set models, *in* J. Yao *et al.* (Eds), *Rough Sets and Knowledge Technology,* Lecture Notes in Computer Science, Vol. 4481, Springer, Berlin/Heidelberg, pp. 1–12.

Yap, C.E. and Kim, M.H. (2013). Instance-based ontology matching with rough set features selection, *International Conference on IT Convergence and Security (ICITCS), Macao, China*, pp. 1–4.

Yeganejou, M. and Dick, S. (2018). Classification via deep fuzzy c-means clustering, *IEEE International Conference on Fuzzy Systems, FUZZ-IEEE, Rio de Janeiro, Brazil*, pp. 1–6.

Yuan, Z., Chen, H., Xie, P., Zhang, P., Liu, J. and Li, T. (2021). Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions, *Applied Soft Computing* **107**(2): 107353.

Yun, S. (2014). Research of big data analysis on rough set and electromagnetism-like mechanism algorithm, *IEEE International Conference on Computer and Information Technology, Xi'an, China*, pp. 923–926.

Zadeh, L. (1965). Fuzzy sets information and control, *Information and Control* **8**(3): 338–353.

Zadeh, L.A. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, *Fuzzy Sets and Systems* **90**(2): 111–127.

Zhang, C., Li, D., Kang, X., Song, D., Kumar, A. and Said, B. (2020). Neutrosophic fusion of rough set theory: An overview, *Computers in Industry* **115**: 103117.

Zhang, J., Li, T. and Pan, Y. (2012). Parallel rough set based knowledge acquisition using MapReduce from big data, *BigMine'12: Proceedings of the 1st International Workshop on Big Data, Beijing, China,* pp 20–27.

Zhao, R., Wang, Y., Liu, Q., Dong, D. and Li, C. (2020). Knowledge acquisition model for stability situation judgement used in crowd evacuation, *5th International Conference on Computer and Communication Systems (ICCCS), Shanghai, China,* pp. 510–514.

Zhou, C. and Lin, Z. (2018). Study on fraud detection of telecom industry based on rough set, *IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, USA*, pp. 15–19.

Zhou, S., Chen, Q. and Wang, X. (2014). Fuzzy deep belief networks for semi-supervised sentiment classification, *Neurocomputing* **131**: 312–322.

Ziarko, W. (1993). Variable precision rough set model, *Journal of Computer and System Sciences* **46**(1): 39–59.

Piotr Pięta is a PhD student at the AGH University of Science and Technology in Cracow, Poland. He received his MSc degree in computer science in 2018. His research activity over the last years has focused on various aspects of rough set theory in big data analysis and data mining, machine learning, collective intelligence and artificial intelligence.

Tomasz Szmuc holds an MSc (1972) in electrical and control engineering from the AGH University of Science and Technology (AGH). Since the beginning of his professional career he has been employed at the AGH University of Science and Technology, receiving there his PhD (1979) and DSc (1989) degrees (both in computer science), and the professorial title (1999). His research may be assigned to two main areas: formal methods supporting software development and hybrid approaches in big data analysis. The former field focuses on applications of Petri nets, process algebras and temporal logics in the development. The combined use of fuzzy-rough sets and neural networks forms the basis of the latter research field. Professor Szmuc is a member of the Computer Science Committee of the Polish Academy of Sciences, and of two scientific committees of the Polish Academy of Arts and Sciences.

Received: 26 April 2021
Revised: 29 July 2021
Accepted: 25 August 2021