The Evolution of Data Mining Techniques to Big Data Analytics: An Extensive Study with Application to Renewable Energy Data Analytics

Dina Fawzy[1*], Sherin Moussa and Nagwa Badr

Department of Information Systems
Faculty of Computer and Information Sciences
Ain Shams University, Cairo, Egypt

[*]*Corresponding author's email: dina.fawzy [AT] cis.asu.edu.eg*

ABSTRACT--- ***Recently big data have become a buzzword, which forced the researchers to expand the existing data mining techniques to cope with the evolved nature of data and to develop new analytic techniques. Big data analytic techniques are serving many domains. In this paper, we provide a detailed comprehensive analysis and discussion of the data mining techniques, studying the changes that have been introduced to some of them that have been successfully developed into big data analytic techniques. The analysis also investigates the reasons behind the rest of data mining techniques that could not be evolved to big data analytics. A detailed study is also presented to discuss the application of big data analytics in the field of renewable energy studies.***

Keywords— Big Data Analytics, Data Mining, Renewable Energy, Wind Energy, Wind Farms

1. INTRODUCTION

Data mining is the process of analyzing data sets in novel ways to find unsuspected relationships. The relationships and summaries derived through data mining are often referred to as models or patterns that extract implicit, unknown and potential useful information from data. This is required in order to predict future trends and behaviors, to make proactive decisions, and to answer business questions that consume too much time to answer [48]. Different data mining techniques have been studied to process and to analyze several types of data patterns, where the most popular data mining tasks are classification, summarization, association rules mining, and clustering [49].

Big data term is a fast-emerging notation referring to the collections of the huge data sets that cannot be processed using traditional database management systems and existing techniques. Big data introduce new approaches for data storage, processing models, analysis and visualization of such enormous data size within an accepted time duration that can be achieved with typical computational systems [52]. This is due to the mainly characterized 4Vs; (i) Volume, which indicates dealing with huge amount of data in terms of petabytes scale collections. (ii) Variety, where the categorization of big data belongs to structured, semi-structured, or unstructured data. (iii) Velocity, which refers to the speed of data generation or how fast the data are required for processing to meet the demand. (iv) Veracity, which refers to the inconsistency and the low quality of data that can be detected in massive data sets, affecting the processing of data [2], [7], [25], [27].

Big data are generated from many domains; business intelligence, spatio-temporal, healthcare and medical records, online newspapers, social networks, and renewable energy [15], [16], [21]. Renewable energy systems generate massive amounts of data from different sources like biomass, solar, wind and nuclear energy. Data are generated as different series of time, and from different sensors and devices that needs to be analyzed in real time in most of the applications. According to the unique nature of big data, it has caused an explosion in the use of newly-evolved data mining techniques, where many challenges have been put under study regarding their algorithmic design, taking into consideration big data volumes, sparse, fast in generation, heterogeneous, uncertain, incomplete, and multi-source data [9], [22], [25], [26], [54]. These new features of big data have forced the researchers to dig more into finding new ways that fit the processing and for better understanding of such data. Hence, big data analytics field has arisen as the approach of analyzing huge amounts of data that have low quality and differ in their structures to extract useful information. It allows the scalability of processing to analyze these data in real-time in order to make time-sensitive decisions. Some data mining techniques have been used to analyze renewable energy data. However, the usage of big data analytics can

help renewable energy enterprises to utilize energy efficiently and to reduce energy costs. In addition, analyzing these data streams enables such enterprises to keep track of production and consumption, while maintaining balance between them during continuous demands [22].

In this paper, we present a detailed study for the data mining techniques and the enhancements that have been introduced to some of them in order to be able to handle big data revolution. We also analyze the reasons of the non-scalability of other data mining techniques to big data nature. In addition, we discuss the data mining and big data analytic techniques that have been applied to the renewable energy domain in specific. The paper is structured as follows: Section 2 presents various data mining tasks with their corresponding techniques. Section 3 studies and analyzes how some data mining techniques have evolved to cope as big data analytic techniques. Section 4 discusses the other data mining techniques that could not be developed to adapt with the big data nature. Section 5 focuses on the data mining techniques applied to the renewable energy domain, indicating which of these techniques were developed to big data analytics. Finally, the conclusion summarizes our study and analysis.

2. DATA MINING TECHNIQUES

In order to ensure meaningful data mining results, it is necessary to understand the data being processed. Data mining approaches are usually affected by several factors, such as noisy data that include null values and untypical values (i.e. outliers). According to the changing nature of the data to be mined, extensions have been introduced to data mining; spatial data mining, for mining spatial data; web usage mining and web content mining, for mining users' behaviors and specific topics over the web respectively; graph mining, for mining data in networks; and recently big data mining, which is an evolved branch of big data analytics to fit different types of data [20], [43], [47], [51]. To achieve a successful data mining model, there are basic phases to follow as shown in the summary presented in figure 1.

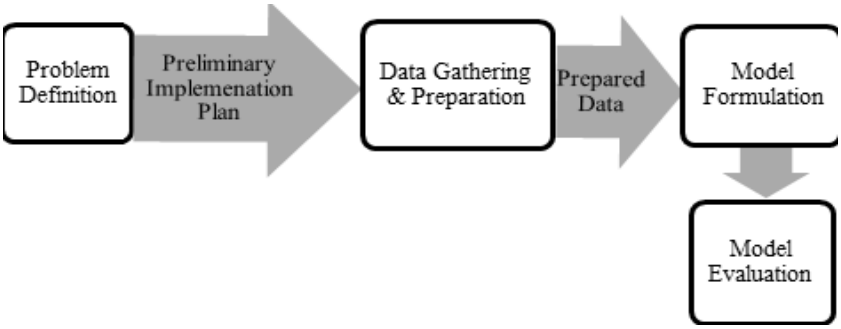

Figure 1: Phases of building a data mining model

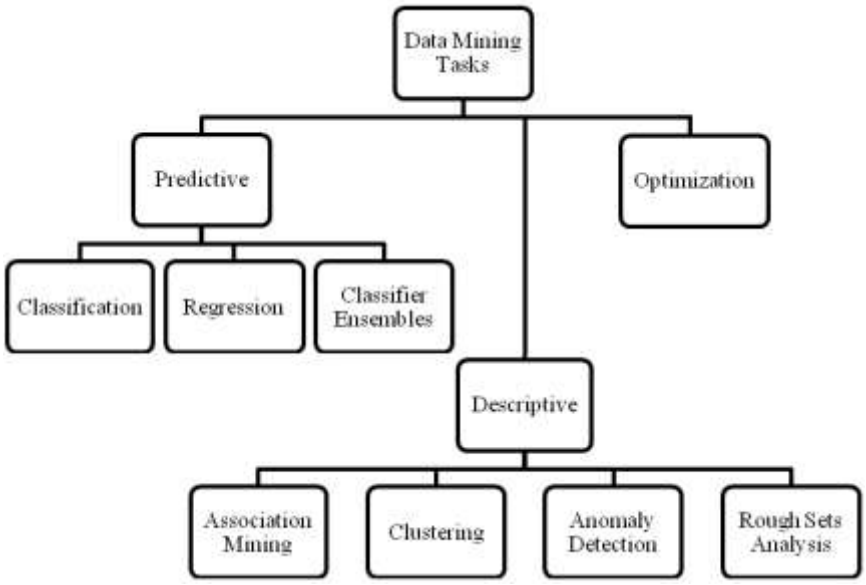

Figure 2: Main data mining tasks

First, the problem definition of the data mining project should be determined, which focuses on understanding the objectives and requirements of the project. Once they are specified from a business point of view, data can be structured in the form of a data mining task and an initial implementation plan is developed. Second, the data gathering and

preparation phase starts, where data sources and data format must be specified, then once data are gathered, data will prepared progressively through multiple rounds. Next, a variety of modeling techniques can be applied in the model formulation step. Thus, many approaches have been investigated for the same category of data mining process, taking into consideration the variety of data forms. Finally, comes the model building and evaluation. In this phase, various modeling techniques are applied to evaluate how well the model satisfies the objectives [1], [32]. Data mining techniques are based on three main tasks as presented in figure 2.

2.1. Predictive Data Mining

The predictive task uses specific variables or values in the data set to predict unknown or future values of other variables of interest [33]. Several approaches have been proposed for prediction as follows:

2.1.1.Classification

The data mining task identifies the class to which a new observation belongs. Given a training data set that has several attributes, where a model is identified as a function of the other attributes' values. This requires a training set of correctly identified observations. The classification is applied to automatically assign records to pre-defined classes, ex: to classify credit card transactions as legitimate or fraudulent, or to classify news stories as finance, entertainment, sports, etc. Many techniques have emerged for classification. However, the most common approaches that have been used in solving real world problems are decision tree-based methods [10], neural networks [11], and support vector machines (SVM), naive bayes classifier, and k-nearest neighbor (KNN) [11]. Decision tree-based methods deduce meaningful rules for predictive information in order to be used for data classification. One of the most popular algorithms is CART (Classification and Regression Tree), ID3 (Iterative Dichotomiser 3), and C4.5 [10].

Neural networks, which are also used in classification because of their ability to extract meaningful information from complex data, they are applied to detect patterns that are considered to be too complicated to be performed by humans. Neural networks consist of networks of "neurons", having a similar neural structure as in the brain. On the other hand, SVMs outline decision boundaries depending on the decision plans concept, which separates between objects belonging to different classes. Whereas naive Bayes classifier is a straight-forward probabilistic classifier that applies Bayes' theorem and assumes strong independent relationships among the features [11]. K-nearest neighbor is another popular classification technique, which uses the common election of the neighbors to assign a data item to the class having the least distance function. In addition, comes the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) technique for classifying objects using a collection of "if . . . then . . . " rules. This technique generates a detection model composed of resource rules that are built to detect future examples of malicious executables [5], [23], [35].

2.1.2.Regression

The other side of predictive data mining is regression, which is a supervised mining function for predicting a numerical target [57]. In the training process of the regression model, it evaluates the target value in terms of a function of each data item's predictors. The relationship between the target value and the predictors are then formulated in a model that can be applied to various data sets with unknown target values. Generalized Linear Model (GLM) is one of the main techniques that apply regression, which performs linear regression for continuous target values [28], in which the dependent variable is continuous, whereas the independent variable(s) can be continuous or discrete, having the nature of regression line is linear [61]. Whilst it applies logistic regression for binary target values classification.

2.1.3. Classifier Ensembles

Classifier Ensembles present the concept of aggregating multiple classifiers as a novel approach to improve the performance of classifiers that work individually [58]. These classifiers can be based on a variety of classification methodologies, achieving different rates of correctly classified individuals. Bagging is an example for classifier ensembles for bootstrap aggregating. It is a method for generating an ensemble of models constructed from bootstrap replicates samples [14]. Random forest is another classifier ensemble consisting of many decision trees, and outputs the node of the class by individual trees. For many data sets, it produces a highly accurate classifier and it can run efficiently on large databases [11]. Rotation forest, on the other hand, uses feature extraction in order to build classifier ensembles. For a base classifier, the training data is created through separating the feature set into k subsets, and then applying the Principal Component Analysis (PCA) on each subset. The principal components are usually reserved to maintain the information variability. Therefore, k axis rounds are performed in order to formulate the new features for the base classifier [24].

2.2. Descriptive Data Mining

Descriptive models analyze past events in the data for insight on how to approach future events. These models can understand past performance by mining historical data to look for the reasons behind past success or failure. This can be used to quantify relationships in data in a way to classify, for example, customers into assemblies. Thus, it differs

from the other predictive models that concentrate on evaluating the behavior of a single customer [28], [34]. Several approaches have been deduced from descriptive models as follows:

2.2.1. Association Rules Mining

It is an approach for exploring the relationships of interest between variables in huge databases [13]. Considering groups of transactions, it discovers rules that forecast the existence of an item depending on the existences of other items in the transaction. It is applied to guide positioning products inside stores in such a way to increase sales, to investigate web server logs in order to deduce information about visitors to websites, or to study biological data to discover new correlations. Examples for association rules mining techniques are: Frequent Pattern (FP) Growth and Apriori. Apriori explores rules satisfying support and confidence values that are greater than a predefined minimum threshold value [34].

2.2.2. Clustering

Cluster Analysis is one of the unsupervised learning techniques, which collects similar objects together that are far different from the rest of objects in other groups [56]. Examples include grouping of related documents in emails, or proteins and genes having similar functionalities. Many types of clustering techniques have been introduced like the non-exclusive clustering, where the data may belong to multiple clusters. Whereas fuzzy clustering considers a data item to be a member to all clusters with different weights ranging from 0 to 1. Hierarchical (agglomerative) clustering, on the other hand, creates a group of nested clusters that are arranged in the form of a hierarchical tree. K-means is the most famous clustering algorithm, where it uses a partitioned approach to separate the data items into a pre-determined number of clusters having a centroid; data items that are in one cluster are closer to its centroid. K-medoids algorithm is a clustering algorithm related to K-means algorithm, which chooses data points as centers [36].

2.2.3. Anomaly Detection

This technique is responsible for detecting outliers, that is, the set of data points that are considerably different from the rest of data. For example, anomaly detection is used for credit card fraud detection, telecommunication fraud detection, network intrusion detection, and fault detection. It builds a pattern or summary statistics of the "normal" behavior for the overall population to detect anomalies. There are several types of anomaly detection, including the graphical-based, where its main functionality is to spot anomalous network entities (e.g., nodes, edges, subgraphs) given the entire graph structure, in addition to the statistical-based, the distance-based, where data is represented as a vector of features and it computes the distance between every pair of data points, and the model-based, which's assumes a parametric model describing the distribution of the data and focusing on finding outliers from data based on this model [44].

2.2.4. Rough Sets Analysis

Rough sets analysis is mainly concerned with the analysis of uncertain and incomplete information [30], [31]. Rough sets represent a major infrastructure for knowledge discovery, where mathematical computations are provided to explore hidden patterns in data. It is used for data reduction, feature selection and extraction, and generation of decision rules.

2.3. Optimization Data Mining

Optimization is the process of finding the most cost effective or highest achievable performance alternatives under some given constraints by maximizing the desired factors and minimizing the undesired ones. Genetic algorithms are of the most well-known algorithms for optimization and search problems, where a method of "breeding" computer solutions of simulated evolution is used. A population of randomly created individuals initiates the evolution. For every generation, the optimization technique evaluates the fitness of every individual in the population to be selected in the next iteration of the algorithm. The algorithm stops when either a threshold maximum number of generations has been created, or an acceptable fitness level has been achieved for the population [29], [50]. Thus, data mining techniques are used in data preprocessing, where data can be cleaned from outliers by the usage of clustering techniques, and then can be smoothed from noisy values by applying regression techniques. Sampling techniques are one kind of the statistics approaches that are needed in data preprocessing before applying most of the data mining techniques. Sampling is usually used with data mining because processing the entire data set of interest is too expensive and time-consuming [59].

3. EVOLUTION TO BIG DATA ANALYTIC TECHNIQUES

Analyzing huge amounts of data allows analysts, researchers, and business users to make better and faster decisions using data that were previously not obvious before, inaccessible, or unusable. However, the dramatically increase of data amounts have made the well-known data mining algorithms unsuitable for such data sizes. Therefore, many studies have currently been directed towards the enhancements that can be introduced to data mining techniques in order to cope with big data, where big data analytics field has emerged. Big data analytic techniques are concerned with several data mining functions, where the most important functions are: association rules mining and classification tree analysis. In this

section, we analyze the main data mining tasks that have been adopted to big data analytic techniques, clarifying the enhancements that have been introduced to achieve such adoption, in addition to the "V" dimension of big data that has been handled by such modifications. Table 1 represents our comprehensive summary of the analysis done for the evolution of data mining tasks to big data analytics [18], [37], [41], [42]. Techniques are grouped according to their data mining task. The table presents the status of each technique whether it has been developed to big data analytics and the dimension of big data that is handled by this developed technique. The following sub-sections describe the enhancements that have been introduced to the different data mining techniques to handle the dimensions of big data in order to evolve to big data analytic techniques.

Table 1: Evolution of data mining techniques to big data analytics

Data Mining Task	Technique	Developed to Big Data Analytics?	Dimension Handled
Classification	K-Nearest Neighbor	Y	Volume & Veracity
	Decision Trees	Y	Volume & Velocity & Variety
	Support Vector Machines	N	-
	Naive Bayes classifier	N	-
	RIPPER	N	-
	Neural Networks	Y	Volume
Association Mining	Apriori	Y	Volume & Velocity
	FP Growth	Y	Velocity
Clustering	K-Means Clustering	Y	Volume
	K-Medoids	N	-
	Agglomerative	N	-
Optimization	Genetic Algorithms	N	-
	Sampling Techniques	N	-
Classifier Ensembles	Bagging	N	-
	Random Forests	N	-
	Rotation Forests	N	-
Regression		N	-
Anomaly Detection		N	-
Rough Set Analysis		N	-

3.1. Dealing with Big Data Volume

Volume refers to the huge amount of data under study, which represents a main challenge in terms of indexing, storage, retrieval, and analysis. Many analytical approaches have handled big data volume. We have investigated these different techniques to determine the specific modifications applied to handle volume as shown in the analysis presented in figure 3. As for the classification, a new version of K-Nearest Neighbor (KNN) algorithm has been introduced for big data, which is the Near Filter Classifier (NFC). The main difference between KNN and NFC is adding the step of dimensionality reduction. In this step, NFC computes the class distribution in each attribute of the original data set and sorts attributes by the classification contribution. Dimensionality reduction reduces data complexity as much as possible, and hence, the computation of model will be simpler.

Another popular way to handle volume in most of the big data analytic techniques is parallel processing, as in the case of the distributed decision tree learning for mining big data streams [10] and the decision trees on the cloud [17]. A machine-learning library is used to implement decision trees in order to carry out analysis on very large data sets, hosting the system on a cloud server. This methodology is responsible for solving the volume dimension only of big data, ignoring the remaining issues of big data [17]. In addition, another technique called "Scalable Advanced Massive Online Analysis (SAMOA)" used distributed streaming machine learning framework with decision trees to address three big data dimensions (volume, velocity and variety), where it focuses on the classification task [55]. The parallel processing was also used in multi-view K-means clustering on big data, where K-means clustering is applied on multiple views of features, then combines the results [56]. Random forests technique uses parallel processing as well on several decision trees and combines their result to handle large volume data sets [57].

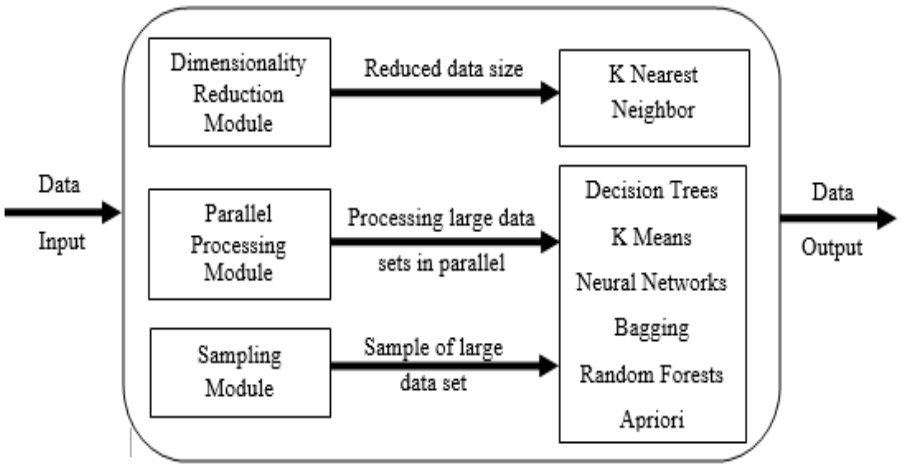

Figure 3: Methods for handling big data volume

Moreover, bagging technique handles big data volumes using parallel and distributed models of classifiers ensembles [58]. Other techniques used sampling that helps in reducing data instead of working on large data sets, like in the randomized algorithm for approximate association rule mining [13]. This approach extracts approximations of frequent items by mining several small samples in parallel. After getting samples, it uses a mining algorithm such as Apriori to get the frequent items. Some other techniques solved the issue of large data volumes using a hierarchical unsupervised growing neural network for clustering large data sets into similar smaller sets, which improved the time needed for processing [53].

3.2. Dealing with Big Data Velocity

The fast data generated had been a challenge for data management. Some data mining techniques have dealt with the velocity issue. We investigated these techniques to deduce the proposed modifications to handle velocity as shown in the analysis presented in figure 4, through using a stream processing engine in order to perform parallel big data streams mining. An example is the SAMOA technique [10], which achieved scalability using stream processing engines like Storm [55]. PARMA is another technique that used sampling with parallel processing for the MapReduce framework [13]. In this approach, all samples were processed in parallel to allow the model to achieve faster results and decrease the running time.

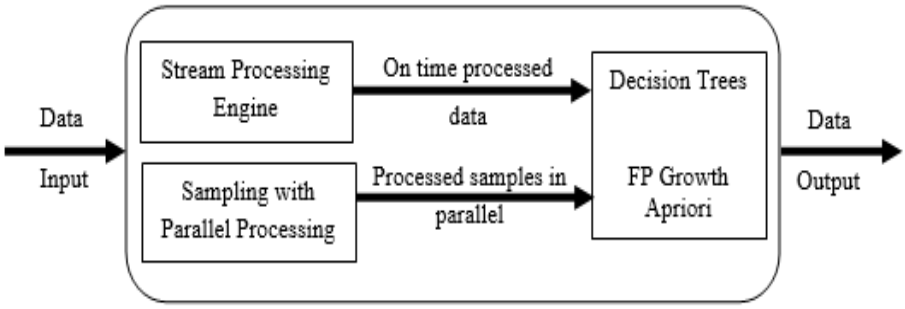

Figure 4: Methods for handling big data velocity

Another big data streams mining technique named "Online Association Rule Mining over Fast Data" [19] was introduced to extract the frequent items from data sets of different sources like Complex Event Processing (CEP) engines at real time using Apriori algorithm or FP Growth. It was implemented in a parallel processing manner to handle the fast velocity of the generated data [19].

3.3. Dealing with Big Data Veracity

Ensuring the quality associated with huge amounts of data is considered to be a key element for reliable analytical outcomes. Some techniques like NFC, the variant of the K-Nearest Neighbor algorithm [9], handle big data veracity. In the proposed analysis, we studied these approaches in order to define the enhancements suggested to handle veracity as shown in the analysis presented in figure 5. Such model has to update itself regularly to reach a certain accuracy degree. Thus, it satisfies the challenge of veracity.

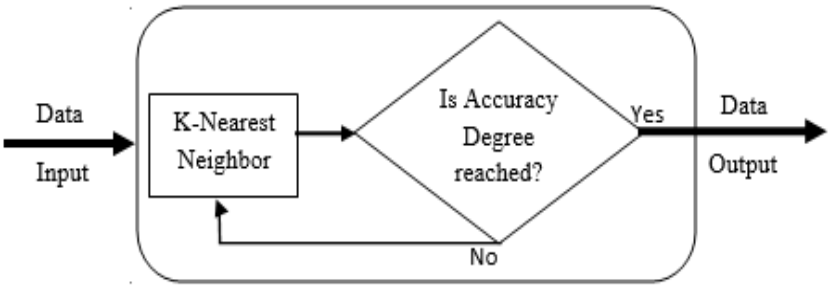

Figure 5: Methods for handling big data veracity

3.4. Dealing with Big Data Variety

Some machine learning frameworks that handled different data types, such as WEKA, have addressed big data variety, as in the case of the SAMOA technique [55]. Another 3-tier framework structure was introduced in [25], [26] to scale up to the exceptionally large volume of varied data. Tier I is responsible for data accessing and computing, which handles the different storage locations of continuously growing data volumes. Tier II, on the other hand, is responsible for data privacy and domain knowledge that handles the different information sharing mechanisms between data producers and data consumers for different domain applications. Finally, Tier III that manipulated the big data mining algorithms. This tier is responsible for mining sparse, uncertain, and incomplete data. Handling big data variety means the ability to process different kinds of data that draw the nature of big data like text, images, audio, and video [27].

4. NON-EVOLVED DATA MINING TECHNIQUES

Big data analytic techniques investigate many challenges and issues, including data understanding and the quality of data, where decision making processes consider the quality of data to be very critical in terms of their accuracy and timing. These are identified as vital metrics for any data analysis. However, the huge volume of information residing in big data makes the case even more acute. This applies not only on volume, but also on data consistency, the variable structure of data, handling data mistakes, the storage requirements for these data, the processing model that will be applied, handling data speed and the hosting location of this processing. Some data mining techniques could not be adopted to cope with this emerging nature of data as presented in table I [38], [39], [40], [45]. In this section, we analyze these techniques and investigate the reasons behind their un-scalability to fit the processing of big data.

4.1. Non-evolved Classification Techniques

Agglomerative lacks scalability as it consumes high complexity for processing stable data. In addition, it faces serious problems during the learning process of raw data that should be very fast when dealing with streams of data. Thus, it needs a stream processing engine. In case of dealing with high dimensionality of data streams; it needs a big data volume reducer module. Naive bayes, SVM, and k-medoids techniques are based on mathematical equations to be performed. With large data sets and high dimensional data, they consume dramatically processing time. Thus, a data volume reducer is highly required as a pre-processing step. In case of RIPPER technique, the huge amount of data makes the process of extracting if...then rules harder. Thus, it needs to reduce data size as well.

4.2. Non-evolved Clustering Techniques

There is no doubt about the low quality of data in dealing with big data, data are inconsistent, out of range, and maybe missing values in these data, so it becomes a challenge to cluster these data using clustering technique such as the k-medoid, which has some limitations including the identification of the number of cluster parameter k and its sensitivity to the order of input dataset and data outliers. Thus, a handling data veracity module is essential.

4.3. Non-evolved Optimization Techniques

Considering the nature of big data, optimization techniques are facing many problems and challenges, which prevent them from developing into big data analytic techniques. This includes the high dimensional data in volume in case of the sampling techniques, in addition to the high complexity of mathematical models used in fitness function in optimization techniques. During big data processing, it is very difficult to collect all data set to get a sample. Whereas with the fast generation of big data, the optimization techniques cannot guarantee finding the optimal solution in a limited time. Moreover, complex mathematical operations are needed for processing, which consume more time.

4.4. Non-evolved Classifier Ensembles

The multiple predictors in Rotation Forest have many weakness points; the sizes of each classifier itself and the ensemble highly affect the total storage (i.e. the number of classifiers in an ensemble). Not only this, but also the required computation that dramatically increases to classify input data, where the processing of all component classifiers

together is considered instead of that of a single classifier. Thus, it requires more execution time. All these concerns are exaggerated with the big data volume, ensuring that a data volume reducer is a necessity.

4.5. Non-evolved Regression Techniques

The ability to apply regression techniques on large datasets with complex relationships between data is still unsatisfied, because of the complex mathematical equations used, and the low quality of data. Regression is sensitive to data outliers, which can have huge effects on regression results. According to big data nature, outliers can really play a scary limitation while applying regression techniques. A data volume reducer is then needed to decrease data size, as well as a data veracity handling module to improve data quality.

4.6. Non-evolved Anomaly Detection Techniques

With so much data volumes and high velocity of generated data as in big data, it could become impossible to completely identify items that do not comply with an expected pattern or other items in a data set in order to remove or to process in a different manner. This is due to the large volume of the outliers themselves that could barely be characterized as "outliers".

4.7. Non-evolved Rough Sets Analysis Techniques

Rough set analysis techniques are dependent on the complete information, and this make these techniques harder to be applied on big data due to the errors, and missing data. The mathematical equations used in rough sets analysis techniques, with the large data sets that reach petabytes; it becomes difficult to process them with complex mathematical equations.

5. RENEWABLE ENERGY DATA ANALYTICS

Renewable energy represents the generated energy from natural renewable resources like solar, wind, biomass, geothermal, hydropower, compressed natural gas and nuclear power [11]. Energy organizations need to evaluate and track energy consumption and production in order to enhance their perception about their energy needs, allowing better real-time decision making for energy usages, while maintaining enough energy resources for the forecasted demands [4], [5], [6], [8], [12], [22], [23], [46]. The relation between big data analytics and renewable energy arises from the fact that huge data streams are increasingly needed to be observed and studied in real time in order to achieve the main target of energy saving [3]. Studying the related works in this domain, one framework for big data and renewable energy analysis was presented in [60]. Hadoop [7] was used in the big energy data loading and analysis by integrating a statistical data mining technique for predictive studies. Based on our research, no other approaches have been introduced to apply big data analytics to renewable energy data.

Wind energy generation, in specific, refers to the process of using wind to generate electricity or mechanical power. Wind turbines are used in order to convert the energy in wind into mechanical power, where they are usually assembled together into a single specifically-designed wind power plant, named wind farm, to generate bulk electrical power. Electricity generated from such turbines is then fed into an electrical grid and dispersed to customers. Designing wind farms is a challenge, as it definitely influences the quantity of generated power. It is considered to be an optimization problem of many factors, including cost issues, by balancing site preparation and wind turbines installation costs. Environmental concerns, like land topology and wind specification (i.e. speed, direction, etc.) are also included in such affecting factors, in addition to supply and transport factors, since good wind sites are usually residing in remote destinations away from any sources of electric power. In this section, we present the different research approached that have been directed towards the data analytics of renewable energy. In addition, we analyze the applied techniques, investigating both data mining and big data analytics approaches.

Some data mining techniques have been used for renewable energy analytics. These techniques are bagging, random forests, rotation forests, RIPPER, k-nearest neighbor, and neural networks [11], [14]. Based on our study, other data mining techniques have never been used on renewable energy data sets. A comparison between the data mining techniques applied for renewable energy data analytics is presented in Table 2. As shown in the comparison, the main advantages of the bagging technique are its low-error degree and bias reduction, whereas neural networks can model complex relations between data more than other techniques can do. However, both techniques lack interpretation. Moreover, neural networks may even over-fit with large data sets as in the case of random forest as well.

Table 2: Comparison between renewable energy analytics techniques

Technique	Advantages	Disadvantages
Bagging	Less error degree and reduce bias	Lack of interpretation
RIPPER	Highly expressive as decision trees	Needs large memory capacity
Rotation Forest	High accuracy degree	Needs large memory capacity
Random Forest	High accuracy degree	Over-fitting for large data sets

On the other hand, both random forest and rotation forest can achieve high accuracy degrees, but as a trade-off, rotation forest needs large memory capacity in order to perform efficiently similar to the RIPPER technique. K-nearest neighbor technique has been applied to renewable energy data as it is a fast and simple approach, but it achieves low accuracy rates when data have noise, since it is sensitive to noisy data. Thus, by the huge data generated from wind turbines' sensors at very high speeds, more studies should be addressed towards big data analytics for renewable energy, especially wind energy. This is strongly needed to analyze these amounts of data in order to optimize the design of the wind farms to be able to efficiently predict the power generated from them [5], [6], [23].

6. CONCLUSION

Big Data is concerned with the huge amount of data that are continuously growing, besides their unprecedented speed that need to be dealt with in a timely manner. The presence of big data has produced a unique moment in the history of data analysis. With the incremental demand to analyze huge amounts of data, resulting from variant sources and generated at very high rates, researchers at different domains have studied the expansion of the existing data mining techniques to cope with the evolved nature of data and to develop new analytic techniques. In this paper, we provide a detailed comprehensive study of the data mining techniques, analyzing the new developments that have been introduced to some of them that have been successfully developed into big data analytic techniques. In addition, by discussing the reasons behind the rest of data mining techniques that could not be evolved to big data analytics.

Finally, we investigate the data analytic approaches that have been applied in the field of renewable energy studies, as huge amounts of energy data are required to be analyzed to efficiently produce power on demand. Limited efforts have been investigated to apply big data analytics to renewable energy data, especially wind energy. Thus, more studies should be addressed towards big data analytics for wind energy for the optimization of the wind farms design to efficiently predict the power generated.

REFERENCES

[1] Xindong Wu,Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J, McLachlan, Angus Ng, Bing Liu, Philip S, Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg: Top 10 algorithms in data mining", Springer-Verlag, 4 December 2007.

[2] Hsinchun Chen, Roger H. L. Chiang, Veda C. Storey, "Business Intelligence And Analytics: From Big Data To Big Impact, Big Data Analytics An Oracle White Paper", MIS Quarterly vol. 36 no. 4, pp. 1165-1188/December 2012.

[3] S. San M. Negnevitsky, N. Hatziargyriou, "Applications of Data Mining and Analysis Techniques in Wind Power Systems", 42440178X/06/$20.00 ©2006 IEEE.

[4] Anushree A. Wasu, HarshadaM .Kariya, Shreyas S. Tote, "Evaluating renewable energy using data mining techniques in developing India", Journal of IJSER, IJSER (International Journal of Scientific & Engineering Research), vol. 4, Issue 12, December 2013.

[5] Lionel Fugon, J´er´emieJuban and George Kariniotakis, "Data mining for wind power forecasting", European Wind Energy Conference - Brussels, Belgium, April 2008.

[6] Muhammad Shaheen, Muhammad Shahbaz, Khalid Afsar Khan Jadoon, "Data Mining For Wind Energy Site Selection", Proceedings of the World Congress on Engineering and Computer Science 2012 vol I WCECS 2012, October 24-26, 2012, San Francisco, USA.

[7] Youssef, M., Gamal Attiya, and El-Sayed Ayman. "New Framework For Improving Big Data Analysis Using Mobile Agent."

[8] Krioukov, Andrew, "Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities." IEEE Data Eng. Bull. 34.1 (2011): 3-11.

[9] Niu, Kun, Fang Zhao, and Shubo Zhang. "A fast classification algorithm for big data based on KNN." Journal of Applied Sciences 13, no. 12, pp.2208.

[10] Arinto Murdopo, "Distributed Decision Tree Learning for Mining Big Data Streams", July 2013.

[11] A Min Tjoa, Iman Paryudi, Ahmad Ashari, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", Journal of IJACSA, IJACSA (International Journal of Advanced Computer Science and Applications), vol. 4, no. 11, 2013.

[12] Abu-Taha, Rimal. "Multi-criteria applications in renewable energy analysis: A literature review." Proceedings of PICMET (Technology Management in the Energy Smart World), 11: IEEE, 2011.

[13] Riondato, Matteo, and Eli Upfal. "Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees." Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2012. 25-41.

[14] Machová, Kristína, Frantisek Barcak, and Peter Bednár. "A bagging method using decision trees in the role of base classifiers." Acta Polytechnica Hungarica3.2 (2006): 121-132.

[15] "Big data & green energy opportunities", Copyright IBM Corporation 2010, Copyright IBM Corporation 2010.

[16] Shahrokni, van der Heijde, Lazarevic, Brandt, "Big Data GIS Analytics Towards Efficient Waste Management in Stockholm", 2nd International Conference on ICT for Sustainability (ICT4S 2014).
[17] Chinmay Bhawe, "Big Data Classification Using Decision Trees On The Cloud", Master's Projects. Paper 317.
[18] Chanchal Yadav, Shuliang Wang, Manoj Kumar, "Algorithm and approaches to handle large Data-A Survey", Journal of IJCSN, IJCSN (International Journal of Computer Science and Network), vol. 2, no. 3, 2013.
[19] Erdi Ölmezoğulları, Ismail Ari, Online Association Rule Mining over Fast Data, 2013 IEEE International Congress on Big Data.
[20] Suthaharan, Shan, "Big data classification: problems and challenges in network intrusion prediction with machine learning." ACM SIGMETRICS Performance Evaluation Review 41.4 (2014): 70-73.
[21] Evans, Michael R., "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities." CyberGIS: Fostering a New Wave of Geospatial Innovation and Discovery, Springer Book, 2013.
[22] Mark J. Embrechts, "bigDAARE: Big Data Analytics for Renewable Energy", CFES 2012-2013 Annual Conference January 25, 2013.
[23] Anoop Verma, Andrew Kusiak, "Prediction of Status Patterns of Wind Turbines: A Data-Mining Approach", Journal of JSEE, JSEE (Journal of Solar Energy Engineering), February 2011.
[24] Kuncheva, Ludmila I., and Juan J. Rodríguez. "An experimental study on rotation forest ensembles." Multiple Classifier Systems. Springer Berlin Heidelberg, 2007. 459-468.
[25] Kale Suvarna Vilas, "Big Data Mining", Journal of CSMR, CSMR (International Journal of Computer Science and Management Research eETECME), October 2013.
[26] Mrs. Deepali KishorJadhav, "The New Challenges in Data Mining", Journal of IJIRCST, IJIRCST (International Journal of Innovative Research in Computer Science & Technology), September 2013.
[27] Rong Liu, Qicheng Li, Feng Li, Lijun Mei, Juhnyoung Lee, Big Data Architecture for IT Incident Management, 2014 IEEE.
[28] Han, Jiawei, Micheline Kamber, and Jian Pei, " Data mining: concepts and techniques: concepts and techniques." Elsevier, 2011.
[29] Minaei-Bidgoli, Behrouz, and William F. Punch. "Using genetic algorithms for data mining optimization in an educational web-based system." Genetic and Evolutionary Computation—GECCO 2003. Springer Berlin Heidelberg, 2003.
[30] Slimani, Thabet. "Application of rough set theory in data mining." arXiv preprint arXiv: 1311.4121 (2013).
[31] Zdzisław Pawlak, Roughsets And Data Mining, Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Baltycka 5, 44 100 Gliwice, Poland.
[32] Hegland, Markus. "Data mining techniques." Acta Numerica 2001 10 (2001): 313-355.
[33] Mohammed J. Zaki, Limsoon Wong, Data Mining Techniques, August 9, 2003 WSPC/Lecture Notes.
[34] Freitas, Alex A, "A survey of evolutionary algorithms for data mining and knowledge discovery." Advances in evolutionary computing. Springer Berlin Heidelberg, 2003. 819-845.
[35] Ozer, Patrick, "Data Mining Algorithms for Classification." Radboud University Nijmegen, January 2008.
[36] Berkhin, Pavel, "A survey of clustering data mining techniques." Grouping multidimensional data. Springer Berlin Heidelberg, 2006. 25-71.
[37] Aloisioa, G., "Scientific big data analytics challenges at large scale" Proceedings of Big Data and Extreme-scale Computing (BDEC) (2013).
[38] Ularu, Elena Geanina, "Perspectives on Big Data and Big Data Analytics", Journal of DBSJ, DBSJ (Database Systems Journal) pp.3-14.
[39] Labrinidis, Alexandros, and H. V. Jagadish. "Challenges and opportunities with big data." Proceedings of the VLDB Endowment 5.12 (2012): 2032-2033.
[40] Ms. Ashwini Mandale, and Prof. Shriniwas Gadage, "Big Data Analytics: Challenges, Tools", Journal of IJIRCST, IJIRCST (nternational Journal of Innovative Research in Computer Science & Technology), vol.3, no.3, May 2015.
[41] Yadav, Chanchal, Shuliang Wang, and Manoj Kumar, "Algorithm and approaches to handle large Data-A Survey." arXiv preprint arXiv: 1307.5437(2013).
[42] Wu, Xindong, "Data mining with big data." Knowledge and Data Engineering, IEEE Transactions on 26.1 (2014): 97-107.
[43] Li, Deren, and Shuliang Wang. "Concepts, principles and applications of spatial data mining and knowledge discovery." Proceedings of the International Symposium on Spatio-Temporal Modeling, (STM'05), Beijing, China. 2005.
[44] Gupta, Richa, "Journey from Data Mining to Web Mining to Big Data." arXiv preprint arXiv: 1404.4140 (2014).
[45] Fan, Wei, and Albert Bifet, "Mining big data: current status, and forecast to the future." ACM SIGKDD Explorations Newsletter 14.2 (2013): 1-5.
[46] Davenport, Thomas H., and Jill Dyché, "Big data in big companies." May 2013(2013).
[47] Zaki, Mohammed J., and Wagner Meira Jr, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014.

[48] Shunxiang, Xu, and Chen Dezhi. "2013 Third International Conference on Intelligent System Design and Engineering Applications ISDEA 2013."
[49] Han, Jiawei, Micheline Kamber, and Jian Pei, "Data mining, southeast Asia edition: Concepts and techniques", 2006.
[50] Sastry, Kumara, David Goldberg, and Graham Kendall. "Genetic algorithms." Search methodologies. Springer US, 2005. 97-125.
[51] Washio, Takashi, and Hiroshi Motoda, "State of the art of graph-based data mining." ACM SIGKDD Explorations Newsletter 5.1 (2003): 59-68.
[52] Tamhane, Deepak S., and Sultana N. Sayyad, "Big Data Analysis Using Hace Theorem", Journal of IJARCET, IJARCET (International Journal of Advanced Research in Computer Engineering & Technology), vol.4, 2015.
[53] Shafaque, Uzma, and Parag D. Thakare, "Algorithm and Approaches to Handle Big Data." IJCA Proceedings on National Level Technical Conference X-PLORE 2014.no. 1. Foundation of Computer Science (FCS), 2014.
[54] Ularu, Elena Geanina, "Perspectives on Big Data and Big Data Analytics." Journal of DBSJ, DBSJ (Database Systems Journal) 2012.
[55] De Francisci Morales, Gianmarco, "SAMOA: A platform for mining big data streams. "Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.
[56] Cai, Xiao, FeipingNie, and Heng Huang, "Multi-view k-means clustering on big data." Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, 2013.
[57] Lim, A., L. Breiman, and A. Cutler, "BIGRF: Big Random Forests: Classification and Regression Forests for Large Data Sets, 2014.
[58] Kleiner, Ariel, "The big data bootstrap."
[59] Hand, David J., "Statistics and data mining: intersecting disciplines." ACM SIGKDD Explorations Newsletter 1.1 (1999): 16-19.
[60] Ceci, Michelangelo, "Big Data Techniques For Renewable Energy Market.
[61] Buck, Samuel F, "A method of estimation of missing values in multivariate data suitable for use with an electronic computer." Journal of the Royal Statistical Society, 1960.