Mining Big Data in Telecommunications Industry: Challenges, Techniques, and Revenue Opportunity

Hoda A. Abdel Hafez

***Abstract*—**Mining big data represents a big challenge nowadays. Many types of research are concerned with mining massive amounts of data and big data streams. Mining big data faces a lot of challenges including scalability, speed, heterogeneity, accuracy, provenance and privacy. In telecommunication industry, mining big data is like a mining for gold; it represents a big opportunity and maximizing the revenue streams in this industry. This paper discusses the characteristics of big data (volume, variety, velocity and veracity), data mining techniques and tools for handling very large data sets, mining big data in telecommunication and the benefits and opportunities gained from them.

***Keywords*—**Mining Big Data, Big Data, Machine learning, Data Streams, Telecommunication.

I. Introduction

TECHNOLOGY revolution enables millions of people to quickly generate massive amounts of data at anytime anywhere from many digital devices such as tablets, smartphones and laptops in addition to continuous streams of digital data from remote sensors. These data of large sizes and complexity have greater value of hidden knowledge and valuable insights. Mining large and complex data sets might extract hidden knowledge and valuable insights. Thus, big data mining has opened many challenges and opportunities. The traditional data mining techniques cause difficulties to extract the hidden knowledge and insights from big data because these techniques cannot handle heterogeneity, volume, speed, accuracy and privacy coming along with big data and big data mining. Thus, many research projects have been initiated in the last couple of years to overcome these challenges [1]. On the other hand, the growing of the telecommunication data traffic according to Cisco annual forecasting will reach 8.6 zettabytes by the end of 2018 up from 3.1 zettabytes per year in 2013 [2]. Therefore, mining the huge amount of data as well as mining real-time data needs to be done by new data mining techniques/approaches. Big data mining can be defined as the capability of extracting valuable information from large datasets or streams of data that due to its volume, variety and velocity it is not possible before to do it [3]. Data mining involves various methods such as decision tree, genetic algorithms, support vector machines and neural network. These methods discover the hidden patterns inside the datasets. The benefit of analysis the pattern is to understand the customer, predict the future, analyze the demands and more [1].

The aim of this paper is to provide detailed study of big data mining its challenges, techniques and related open source tools. It also discusses the implementation of big data mining in telecommunication industry.

This paper has been organized as follows; Section I: Introduction to big data and big data mining. Section II discusses the characteristics of big data and the data mining challenges for handling large datasets. Section III classifies the data mining techniques into machine learning, time series and data streams and examines the applied algorithms in each classification. Section IV describes the main open source tools for mining big data. Section V discusses applying big data mining and its benefits in telecommunications industry. Section VI concludes the discussion**.**

II. Big Data Systems

Big data has a lot of challenges due to its characteristics, and also mining this massive amount of data imposes a critical challenge as the following.

A. Characteristics of Big Data

The characteristics of big data systems include the “4 Vs”. They are Volume, Velocity, Variety, and Veracity as shown in Fig. 1 [4]-[7].

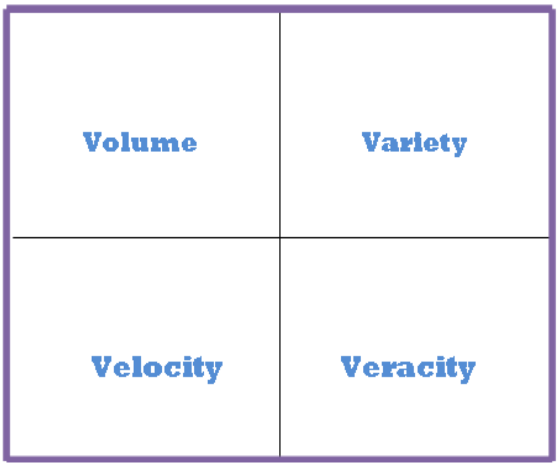

Fig. 1 The 4 V’s of big data

Volume refers to the huge amount of data that is being manipulated and analyzed in order to obtain the desired results. The volume of data increases exponentially and this represents a big challenge. The size of the data that is being processed can be unlimited and the speed of processing operations is constant. The second characteristic of big data is

Hoda A. Abdel Hafez is with the Faculty of Computers & Informatics, Suez Canal University, Egypt (e-mail: Hodaabdelhafez@gmail.com).

velocity or speed. Velocity is the speed at which data generated and the speed at which these data should be processed. Most data sources generate high streamlining data such as phone conversation and web searches. Streaming data in real-time is a big challenge due to high requests that end users have for streamed data over numerous devices. For companies transfer rates of data are limited but requests are unlimited. Variety refers to different types or format of data as text, numerical, sensor data, log files, sequences, images, audio, video, social media and more. It also includes a static data and streaming data. This heterogeneity of unstructured data creates problems for storage, mining and analyzing the data. Moreover, extracting the knowledge from all these types of data needs to be linked together. Veracity has recently added; it came from the idea that the possible consistency of data is good for Big Data. If the data is unreliable and redundant, it can cause an avalanche of errors, incorrect results and the whole company can have a big problem. Finally, I can conclude that the characteristics of big data systems represent key challenges in capturing, storage, analysis, sharing, visualizing, connection and correlation of data, and more.

B. Data Mining Challenges for Handling Big Data

Traditional data mining techniques have been used to discover unknown patterns from structured, homogeneous, and small datasets. Currently, there is a massive volume of data represented by heterogeneous and diverse dimensionalities. Moreover, the autonomous data sources with distributed and decentralized controls as well as the complexity and evolving relationships among data are the characteristics of big data applications. These characteristics cause an extreme challenge for mining big data to discover useful knowledge [8]. The data from different sources possesses many different types and representation forms that lead to the great variety or heterogeneity of big data. Mining from such a massive heterogeneous dataset is a big challenge. Heterogeneity in big data deals with structured, semi-structured, and unstructured data simultaneously. Unstructured data will not fit with database systems. Velocity of big data is really matter because the capability of fast accessing and mining big data is also an obligation especially for data streams. To mining big data with heterogeneity, extreme scale and velocity characteristics, the new mining techniques and algorithms are needed [1]. Other related critical big data mining issues are privacy, accuracy, trust and provenance. The huge volume of big data contains tremendous amount of interconnected personal information. Every piece of information about everybody can be mined out, and when all pieces of the information about a person are dug out and put together, any privacy about that person disappears [1]. For privacy issue, a model needs to be developed where the benefits of data for businesses and researchers are balanced against individual privacy rights [9]. The accuracy and trust of the data sources become more significant because these sources are of many different origins, not all well-known, and not all verifiable, which further propagate to the mining results. To solve this problem, data validation and provenance tracing become a necessary step in the data mining. Unsupervised learning methods have been used to discover the trust measures of suspected data sources using other data sources as testimony [1].

III. Data Mining Techniques for Large Scale Data

This section discusses the data mining techniques that can be used for handling big data. These techniques are classified into machine learning, time series and data streams.

A. Machine Learning

Classification and clustering represent the main machine learning techniques for massive data sets and parallel systems as shown below.

Decision trees are effective algorithms and widely used for classification. For building decision tree, there are many algorithms, which are BOAT (optimistic decision tree construction), ICE (implication counter examples) and VFDT (very fast decision tree) [10]. On the other hand, the hybrid approach combining both decision tree and genetic algorithm are also used to create optimized decision tree in order to improve classification performance [11]. Another classification technique is Artificial Neural network (ANN), which is used for large datasets, described the techniques of SOM (self-organizing feature map) network and LVQ (learning vector quantization) network. SOM reduces the dimensions of data through the use of self-organizing neural networks and it takes input in an unsupervised manner. It is a single layer feed forward network where the output syntaxes are arranged in low dimensional grid. LVQ uses supervised learning to categorize large data set into small sets in order to improve the overall computing time needed to process this large data set [12].

Clustering is a process of grouping objects with similar properties. The main clustering techniques for handling very large datasets are Hierarchical clustering, K-means clustering, and Density Based Clustering [13]. Hierarchical clustering (HC) can be classified into two approaches agglomerative hierarchical clustering and divisive hierarchical clustering. In agglomerative approach, each data points are considered to be a separate cluster, and the clusters are merged based on criteria. In divisive approach all data points are considered as a single cluster and they are split into number of clusters based on certain criteria. Under the hierarchical clustering there are Clustering- Based SVM (support vector machine) algorithm and efficient hierarchical clustering algorithm using P-trees. SVM trains a very large datasets using the hierarchical micro clusters [14]. Efficient hierarchical clustering is suited to multimedia-, stream- and spatial data. It uses P-trees to provide efficient data storage and access. K-means clustering is a method of cluster analysis and its aims to partition *n* observations into *k* clusters. Each observation belongs to the cluster with the nearest mean. Parallel *k/h*-Means algorithm is use for clustering large distributed datasets. It is scalable, thus it enlarges its field of application to clustering tasks. Both advantages of K-Means and HC algorithm are used for high dimensional data set and improving data streams processing.

Finally, Density Based Clustering is another method identifying clusters in large high dimensional data set having different size and shape. It is suitable for handling noise in dataset. DBSCAN and DENCLUE are two examples of density based clustering. DESCRY is a new density based clustering algorithm for a large dataset. It identifies clusters in large high dimensional data set having different size and shape.

B. Time Series

Time series databases consist of sequence of events or values over repeated measurement of time. With the growing deployment of a large number of sensors and online data collection tools, the amount of time series data is increasing rapidly. Time series is a rich and important data source for detecting abnormal states, diagnosing performance issues such as finance, communication, automatic control, and online services, etc. [15].

Time series data mining is a framework of analysis time series data for discovery and use of patterns as well as prediction of future values. Data in time series has a lot of variations, therefore, both clustering and classification methods are used. Classification methods include decision tree and/or SVM (support vector machine) [15].

Clustering high dimensionality and large scale time series datasets become a great challenge because it is required an interactive analysis in real practice. Three clustering algorithms are used for large time series data: CLARANS, DBSCAN and DENCLUE. CLARANS is an improved k-medoids method that is more effective and efficient. DBSCAN, a Density-based algorithm treats clusters as dense regions of objects in the spatial space separated by regions of low density. DENCLUE is a highly efficient Density-based algorithm to deal with high dimensional datasets and it is faster than CLARANS and DBSCAN. Another time series clustering algorithm is YADING for automatically cluster large-scale time series with fast performance and quality results. YADING consists of three steps: data reduction, clustering, and assignment. Data reduction reduces the dimensionality of the input time series instances then clustering is conducted on the sampled dataset, and finally, all the input time series are assigned to the clusters results from the sampled dataset [16].

C. Mining Data Streams

Mining data stream is defined as a process of extracting knowledge structure from continuous, rapid data records. There are two types of data streams: online data stream and offline data stream. Mining online data stream is used in a number of real world applications such as fraud detection and network traffic monitoring. Mining offline data stream is like generating report based on web log streams. Thus, increasing rate from such applications using sensor networks and call detail records need data stream real time analytics to manage data currently generated [17]. In other words, mining data streams face great challenge in storage devices. To extract useful patterns/ knowledge from stream data, algorithms with new or modified techniques are needed [18].

Decision tree algorithms are used for classifying data streams. These algorithms include the Hoeffding Tree, Very Fast Decision Tree method (VFDT) and Concept adaptive Very Fast Decision Tree method (CVFDT). Hoeffding tree provides incremental approach and scale different attributes better than tradition decision tree. It is capable of learning from massive data streams. Both Hoeffding Window Tree and Hoeffding Adaptive tree can cope with concept and distribution drift on data streams [19]. VFDT is based on Hoeffding tree algorithm. It implements basic fundamental of a decision tree learning that is capable of learning from high speed data streams in an incremental anytime fashion. The incremental approach in VFDT means where as a new example come it merged with old data. CVFDT is an extend VFDT to keep trees up to date with time variant data streams. It implements sliding window of various dataset to keep its model consistent. CVFDT continuous monitors the quality of new data and adjusts those that are no longer correct. Every time a new data arrives, the CVFDT incrementing counts for new data and decrementing counts for oldest data in the window [18], [20].

IV. Tools for Mining Big Data

There are many data mining tools for handling large datasets. The main open source tools are scikit-learn, R, WEKA, KNIME, Orange, MOA, RapidMiner and SAMOA as shown below.

Scikit-learn (formerly scikits.learn) is a free package representing machine learning library for the Python programming language. It includes classical learning algorithms, model evaluation and selection tools, as well as preprocessing procedures. Scikit-learn has a well written online documentation for all of the implemented algorithms. The package supports most of the core data mining algorithms, but several groups of data mining algorithms have been omitted currently, including classification rules and association rules. However, the package is strong in function-based methods including many general linear models and various types of SVM implementations. It is also quite fast despite being written in an interpreted language because of optimizing the code in various aspects, such as calling array based NumPy number crunching algorithm and writing wrappers for existing C/C++ implementations in Cython. On the other hand, the use of scikit-learn still requires a skilled programmer in Python because of its command-line interface [21], [22].

R open source programming language is designed for statistical computing and visualization. R is the successor of S, a statistical language originally developed by Bell Labs in 1970s. The source code of R is written in C++, Fortran, and in R itself. Interface to R is command line and use through scripting. It also offers simple GUI for input. From data mining user's perspective, R offers very fast implementations of many machine learning algorithms and statistical data visualizations methods. It has specific data types for handling big data, supporting parallelization, web mining, data streams, graph mining, spatial mining, and many other advanced tasks [21]. R's main problem is its language; it is highly extendable but a difficult one to learn regarding data mining.

WEKA (Waikato Environment for Knowledge Analysis) is a Java-based, open-source data mining platform developed at the University of Waikato, New Zealand. User can access components through JAVA programming or through command line interfaces. WEKA provides graphical user interface through WEKA knowledge flow. It offers four options for data mining: command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. It is much weaker in classical testing than R but stronger in a machine learning. WEKA supports many model evaluation procedures and metrics, but lacks many data survey and visualization methods. In spite of some improvements were made recently with respect to clustering, WEKA is more oriented towards classification and regression problems and less towards descriptive statistics and clustering methods. It provides limited support for big data, text mining, and semi-supervised learning, while deep learning methods are still not considered [21], [23].

TABLE I
CHARACTERISTICS OF BIG DATA MINING TOOLS

Characteristics/ Features	R	Scikit-Learn	WEKA	KNIME	Orange	MOA	Rapid-Miner	SAMOA
Programming Language	G++ Fortran, R	Multiple, Support INRIA & Google summer of code	JAVA	JAVA	Python, C++, QT framew	JAVA	JAVA	JAVA
Main purpose	Computation & statistics	Machine learning package add-on	General data mining	General data mining	General data mining	Machine learning	General Data mining	Mining data streams
User Interface	GUI and command line	Command line	GUI and command line	GUI	GUI and command line	GUI	GUI	API
Data Mining algorithms	Decision tree learner Classification rules Hierarchical cluster Density based clustering (DBSCAN) Partition based clustering including K-means, X-means & fuzzy c-means clustering Naïve Bayes Function based learning Data streams Time series analysis Data visualization	Classification and Regression Trees K-means for clustering Support Vector Machines k-Nearest Neighbor Gaussian Naive Bayes Logistic Regression Visualization	Decision tree learner Classification rules Density based clustering (DBSCAN) Partition based clustering including K-means, X-means Bayesian networks Function based learning Hybrid learning methods Time series analysis Data streams	Decision tree learner Classification rules Density based clustering (DBSCAN) Partition based clustering including K-means, X-means & fuzzy c-means clustering Bayesian Networks Function based learning Hybrid based methods Time series analysis Data Visualization	Decision tree-learner Fuzzy c-means-clustering Naïve Bayes Function based- leaning Data visualization	Massive data streams StreamKM++, CluStream, ClusTree, Den-Stream, D-Stream and CobWeb for clustering boosting, bagging, and Hoeffding Trees, all with and without Naive Bayes classifiers for classification	Decision tree learner Classification rules Density based clustering (DBSCAN) Partition based clustering including K-means, X-means Bayesian networks Function based learning Hybrid based methods Time series analysis Data streams Data visualization	Vertical Hoeffding Tree (VHT) an extended version of streaming decision tree for classification-CluStream for clustering-Decision rule learner for regression

KNIME (Konstanz Information Miner) is a general purpose data mining tool based on the Eclipse platform. It is developed and maintained by the Swiss company KNIME.com AG. KNIME is open-source, though commercial licenses and used by over 3000 companies. The KNIME tool has building blocks called nodes, and more than 1000 nodes are available through the core installation and various extensions. The nodes are organized in a hierarchy and can be searched by name within an intuitive interface. Each node has documentation details and performs a certain function such as filtering, modeling and visualization. It also has input and output ports. Some nodes handle data model as classification tress. One of the greatest strengths of KNIME is the integration with WEKA and R. WEKA integration enables using almost all the functionality available in WEKA as KNIME nodes, while R integration enables running R code as a step in the workflow, opening R views and learning models within R [21], [23], [24].

Orange is a Python-based tool for data mining developed in Bioinformatics laboratory in Ljubljana University. It can be used either through Python scripting as a Python plug-in, or through visual programming. Orange Canva as a visual programming interface offers the following functionalities grouped into nine categories: data operations, visualization, classification, regression, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementations. These functionalities are visually represented by different widgets such as read file and train SVM classifier. The visual programming environment uses graphical widgets to combine methods from the core library and associated modules in order to help users developing custom algorithms. In Orange, the data mining algorithms are organized in hierarchical toolboxes, which make them easy to implement. Although, the number of available widgets seems limited in Orange when compared to other tools (such as KNIME or RapidMiner), the coverage of standard data mining techniques is quite good [21].

MOA (Massive On-Line Analysis) is open source software to perform data mining stream. MOA is related to WEKA, which is a workbench containing implementations of a wide range of batch machine learning methods. It contains collection of offline and online methods for both classification and clustering as well as tools for evaluation. For classification it implements boosting, bagging, and Hoeffding Trees, all with and without Naive Bayes classifiers at the leaves. For clustering, it implements StreamKM++, CluStream, ClusTree, Den-Stream, D-Stream and CobWeb. MOA is written in Java so that the applications can be run on any platform with an appropriate Java virtual machine. MOA streams can be built using generators, reading ARFF files, joining several streams, or filtering streams. They allow for the simulation of a potentially infinite sequence of data. The available generators are Random Tree Generator, STAGGER Concepts Generator, SEA Concepts Generator, LED Generator, Waveform Generator and Function Generator [25].

RapidMiner is Java-based, general data mining tool currently in development by the company RapidMiner, Germany. The tool has become very popular in several recent years and has a large community support. RapidMiner offers an integrating environment with visually appealing and user-friendly GUI. RapidMiner is focused on processes that may contain sub-processes. These processes contain operators in the form of visual components. Operators are implementations of data mining algorithms, data sources, and data sinks. RapidMiner provides data mining and machine learning procedures including: data loading and transformation, ETL (Extract, transform, load), data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. In spite of RapidMiner is quite powerful with its basic set of operators, it is the extensions that make it even more useful. Popular extensions include sets of operators for text mining, web mining, time series analysis and more. Most of the operators from WEKA are also available through extension, which increases the number of implemented data mining methods [21], [26].

SAMOA (Scalable Advanced Massive Online Analysis) is a platform for mining big data streams and it is written in Java. It provides a collection of distributed streaming algorithms for the data mining and machine learning tasks such as classification, clustering, and regression. SAMOA also provides programming abstractions to develop new algorithms. SAMOA has a framework and a library. A framework allows algorithm developers to abstract from the underlying execution engine, and then reuse their code on different engines. It features a pluggable architecture that allows it to run on several distributed stream processing engines such as Storm, S4, and Samza. SAMOA includes a distributed version of CluStream, an algorithm for clustering evolving data streams. CluStream keeps a small summary of the data received so far by computing micro-clusters online. These micro-clusters are further refined to create macro-clusters by a micro-batch process, which is triggered periodically [27].

Table I shows the comparison of data mining tools discussed. The results demonstrate that each tool has features that distinguish it from others, and there is no best tool. The choice among these tools depends on the required features and mining algorithms.

V. Mining Big Data in Telecommunications Industry

This section discusses big data mining in telecommunication companies and the mining techniques/ algorithms that are used to handle very large datasets as well as the benefits and opportunities gain from mining these datasets.

A. Big Data in Telecommunications

In telecommunication industry, data can be classified into three groups: call detail data, network data and customer data [28], [29]. Call detail data is information about the call, which stores as a call detail record (CDR). For every call placed on the network, a call detail record is generated to store the call details. The CDRs include information about originating and terminating phone numbers, duration of calls and period of calls [28], [29]. The increasing use of smart phones also generates a new category of transaction records called extended data records (XDRs). The XDRs capture other transactions, such as mobile payment and the purchase and download of a song or a video clip [30]. Network data such as error generation and status messages are generated in real time and with large volumes. These data are used to support network management functions. Customer data includes information about the customer such as name, address details, subscription type and payment history. To understand key customer attributes, data about millions of customers (stored in database systems and data warehouse) needs to be handled. Therefore, big data in telecommunication companies creates the need to analyze multiple data types including location data, social media (such as facebook and twitter), data from sensors and natural language text in order to provide insights that can enable them to increase revenues and reduce costs [30].

B. Mining Techniques for Handling Big Data in Telecom

Telecommunication faces an increasing churn rate compared with other industries. The types of churn are voluntary and involuntary. If the customer begins the first movement, this is called voluntary churn. There are many reasons for losing the customers such as the phone service changes and the competition. If the company begins to terminate its services to the customer, this is called involuntary churn. In this case customers are churned for reasons like fraud and nonpayment. Increasing churn rate causes a loss of future incomes. Therefore, it is profitable for telecom operators to invest in those customers that already have an experience with the service by renewing their trust, rather than trying to attract new customers. A real world dataset from an Asian mobile operator are used to evaluate the performance of the hybrid model. This hybrid model is built using WEKA including Logistic Regression (LR) in parallel with Voted Perception (VP) for classification, and combined with clustering. LR is used as a major classifier and VP reinforce the prediction. The Data preprocessing step transforms the selected input data in an appropriate format for analysis in WEKA knowledge flow. The evaluation of the model shows that its accuracy is higher than single model. The evaluation of the model shows that its accuracy in explaining the churn behavior is higher than single model. The results illustrate also that four months' period is enough for a churn to show his dissatisfaction before the deactivation [31]. A framework for predicting customer behavior based on hybrid learning approaches for churn prediction in mobile network. It includes decision tree induction C4.5 and genetic algorithm to give a pragmatic churner model. This model is implemented to predict various user defined groupings based on usage time, location and their underlying social network. The experiment has been implemented using 1.2 million subscribers CDRs from large telecom operator in a developing country. The dataset of 0.7 million users is taken for training and the remaining as a test set. The results show that the hybrid model demonstrates higher accuracy than J48 and GP (genetic programming) models [32]. Other proposed model for predicting churn and analyzing customer behavior is a combining decision tree algorithm (C5.0) with misclassification cost factor to predict customer's status (loyal or pre-leave) on different customer groups. A three months' data from Chinese telecommunication operation is used to examine this model. The data contains 1,048,575 records and 25 customer-attributes such as customer ID, costs, traffic, voice package information etc. After excluding abnormal samples, the original datasets remain a total of 1,016,610 records, including 27,925 customer churn records. A random sample of loyal customer and customer churn from the processed dataset is selected and then K-means method is used to cluster the data into three groups (high, medium, low). To conduct customer churn prediction, C5.0 is used with segmentation and misclassification cost and C5.0 without segmentation and misclassification cost. The results show that model accuracy presented by C5.0 with customer segmentation and classification cost are higher than those models without segmentation and misclassification cost [33].

Fraud detection is important for the telecommunications industry because companies and suppliers of telecom services lose a significant proportion of their revenue as a result. Fraud in telecommunication can actually be viewed as fraud scenarios, which are related to the way the access to the network was acquired. An experiment for fraud detection is focused on the evaluation of different user profiles and their effect towards the proper discrimination between legitimate and fraudulent activity. This experiment is based on real data extracted from the CDRs for a period of eight years from an organization's PBX (Private Branch Exchange). Mainly a CDR contains at least data such as the caller ID, the chargeable duration of the call, the date and the time of the call, etc. According to the organization's charging policy, only international and mobile destinations are charged. For each user, three different profile types are constructed and tested. The first profile is build up from the accumulated weekly behavior of the user. The second profile is a detailed daily behavior of a user. The third profile is an accumulated per day behavior. Finally, the last two profiles were also accumulated per week to give overall five different user profile representations. The user daily profiles are used as an input for the two algorithms (K-means and agglomerative clustering algorithm). The k-means algorithm is used to partition the input space in two distinct groups. If the legitimate and the fraudulent behavior cases are sufficiently different from each other, then the k-means algorithm will provide two distinct clusters of data. If this is not the case, then the same input data will be fed into the agglomerative clustering algorithm to check whether there is an output with distinct cases separation [34].

Transactional data streams are a challenging domain in which to apply data mining ideas because of a huge volume of simple records and a dynamic continuous flow of data. For instance, AT&T has approximately 5 million call detail records (CDRs) per day relating to international telecommunications traffic from 12 million accounts. A stream of CDRs associated with AT&T's wireless communications services includes approximately 80 million records per day from 15 million accounts. AT&T has developed signature-based methods for fraud and intrusion detection. Anomaly detection via signatures is used by AT&T for telecommunications fraud detection involving international calling. It uses anomaly detection to measure the unusualness of a new call specific to a particular account (or to a fraudster), and it uses a profile-based approach to characterize this unusualness as fraud [35]. In telecom, time series data generates massive amount of data. For example, long distance data stream in AT&T consists of approximately 300 million records per day from 100 million customers. This time series data has lots of variations. Some data sequences are long like billing data and each data item in a time series is a multidimensional vector. To overcome these difficulties two steps are followed: first, transform the data into equal-length vectors using a model based clustering method. Second, use standard classification method such as decision tree and/or SVM methods [15].

For innovative mining applications, Telefonica is involved in innovative big data applications for real-time marketing and for supporting strategic decisions in retailing. The company has a big base of users and enterprise M2M (machine-to-machine) customers. Telefónica has a Dynamic Insights product portfolio, an analytics engine able to analyze M2M data, network data, as well as other external data. The Dynamic Insights aims to collect and aggregate mobile data to understand the overall behavior of mobile users. The observation of users can be done in near real time. It can show the impact of a marketing campaign, a competitor store opening or even a change in a store's opening hours on society. Smart Steps is a first product of Dynamic Insights. Smart Step is a crowd analytics decision-support tool that can help companies and public sector organizations. It analyzes footfall in any location and that is interesting for retailers [36], [37].

C. Benefits and Opportunities from Mining Big Data in Telecommunications

The increasing use of the platforms, smart phones tablets and laptops allow telecommunication companies to collect huge amount of data mostly unstructured. For example, the data emanating from mobile devices provide information about customers such as past buying patterns and geospatial location. Thus, mining big data in telecom provides insights to better understand customer need and reducing churn rate, which represent an opportunity to improve the relationship with the customers and their satisfaction with the service. Detecting fraud is also important for telecommunication industry in order to reduce lose of their revenue. Other new revenue is through selling analytical data and insights. Analysis mobile data and location data to find patterns of user behavior and then selling them for retail, marketing and advertising purposes represents key initiative and opportunity for this industry.

VI. Conclusion

The enormous data generated from millions of people using digital devices and also the continuous streams of digital data cause many challenges. These challenges include scalability, speed, heterogeneity, accuracy, provenance and privacy. This paper describes and reviews big data mining, its challenges, different techniques and open sources tools for handling large data sets. In addition to that the paper discusses applying big data mining in telecommunication industry. Mining large data sets in telecommunication companies can help them detecting fraudulent, reducing churn rate and improving customer experience as well. Moreover, mining big data in telecommunication and selling the results of analysis and insights to third parties represents a big opportunity to this industry. In this case, a model for privacy issue should be developed to balance the benefits of data for telecom operators against customer privacy rights.

References

[1] D. Che, M. Safran and Z. Peng, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities", DASFAA, workshops, LNCN 7827, Springer, 2013, PP 1-15.

[2] Cisco Analysis, "Cisco global cloud index: forecast and methodology 2013-2018" White Paper, 2014.

[3] Fan W. and Bifet A. (2013), "Mining big data: Current status and forecast to the future" SIGKDD Explarations Vol. 14(2), PP. 1-5.

[4] A. Tole, "Big data challenges" Database Systems Journal, Vol. IV (3), 2013, PP. 31-40.

[5] P. Sharma and C. Navdeti "Securing big data Hadoop: A review of security issues, threats and solution", Vol. 5(2), 2014, PP. 2126-2131.

[6] T. Mitha and V. Kumar, "Application of big data in data mining", International Journal of Emerging Technology and advanced Engineering, Vol. 3(7), 2013, PP. 390-393.

[7] M. TRIFU and M. IVAN, "Big Data: present and future", Database Systems Journal, vol. 5(1), 2014, PP. 32-41.

[8] X. Wu and X. Zhu, "Data Mining with Big Data", IEEE transactions on Knowledge and Data Engineering, Vol. 26 (1), 2014, PP. 97-107.

[9] O. Tene, J. Polonetsky, "Privacy in the Age of big data: A Time for Big Decisions", Stanford Law Review Online, vol. 64, 2012, pp. 63-69.

[10] A. Franco-Arcega, J. Carrasco-Ochoa, G. Sánchez-Díaz, and J. Martínez-Trinidad, "Decision Tree based Classifiers for Large Datasets", Computacióny Sistemas Vol. 17(1), 2013, pp. 95-102.

[11] C. Yada, S. Wang and M. Kumar, "Algorithm and approaches to handle large data survey", IJCSN International Journal of Computer Science and Network, Vol 2(3), ISSN (Online): 2277-5420, 2013.

[12] Y. Lu and C. Fahn, "Hierarchical Artificial Neural Networks for recognizing high similar large data sets", Proceeding of the sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 2007, PP. 1930-1935.

[13] M. Vijayalakshmi and M. Renuka devi, "A Survey of Different Issue of Different Clustering Algorithms Used in Large Data sets", IJCSN International Journal of Computer Science and software Engineering, Vol 2(3), 2012, PP. 305-307.

[14] H. Yu, J. Yang, and J. Han, "Classifying Large Data Sets Using SVMs with Hierarchical Clusters", SIGKDD'03 Washington, DC, USA, 2003.

[15] R. Mahajan, A. Thangavelu and M. Shahakar, "Data Mining Techniques for Identifying Temporal Patterns of Time Series Data", Journal of Engineering Research and Applications (IJERA) Vol. 2(6), 2012, pp.185-187 185.

[16] R. Ding, Q. Wang, Y. Dang, Q. Fu, H. Zhang, D. Zhang, "YADING: Fast Clustering of Large-Scale Time Series Data", Proceedings of the VLDB Endowment, Vol. 8(5), 2015, PP. 473-484.

[17] D. Parikh and P. Tirkha, "Data mining and data stream mining – open source tools", International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), Vol. 2(10), 2013, PP. 5234-5239.

[18] T. Trambadiya and P. Bhanodia, "A comparative study of stream data mining and innovative technology", International Journal of Engineering and innovative technology (IJEIT), Vol. 2(3), 2012, PP. 149-154.

[19] A. Bifet and R. Gavalda "Adaptive parameter-free learning from evolving data streams", Proceeding of 8th International Symposium on Intelligent Data Analysis, IDA in the series lecture notes in computer science, Vol. 5772, 2009, PP. 249-260.

[20] G. Hulten, L. Spencer and P. Domingos, "Mining time-changing data streams", proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining, 2001, PP 97-106.

[21] A. Jovic, K. Brkic and N. Bogunovic, "An overview of free software tools for general data mining", 37th International Conventiion on Information & Communication Technology Electronics & Microelectronics., 2014, PP. 1112-1117.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", JMLR, Vol. 12, 2011, PP.2825-2830.

[23] A. Bifet, "Mining big data in real time", infomatica, Vol. 37, 2013, PP. 15-20.

[24] M. Berthold, N. Cebron, F. Dill, T. Gabriel, T. Kotter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel, "KNIME – The Konstanz Information Miner: Version 2.0 and Beyond", ACM SIGKDD Explorations, Vol. 11(1), 2009, PP. 26-31.

[25] A. Bifet, G. Holmes, R. Kirkby. and B. Pfahringer, "MOA: Massive online analysis", Journal of Machine Learning Researches, Vol. 11, 2010, PP. 1601-1604.

[26] RapidMiner Review, 2015, http://butleranalytics.com/rapidminer-review/

[27] G. Morales and A. Bifet, "SAMOA: Scalable Advanced Massive Online Analysis", Journal of Machine Learning Research, Vol. 16, 2015, PP.149-153.

[28] G. Weiss, "Data mining in Telecommunication", in O. Maimon and L. Rokach (Eds) Data mining & Knowledge discovery handbook: A complete guide for practitioners and research, Kluwer Academic publisher, 2005.

[29] M. Joseph, "Data mining and business intelligence applications in telecommunication Industry" International Journal of Engineering and advanced Technology (IJEAT), 2013.

[30] R. Dam, "Big data a sure thing for telecommunications: telecom's future in big data", International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 2013

[31] G. Olle and S. Cai, “A hybrid churn prediction model in Mobile telecommunication industry”, International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 4(1), 2014, PP. 55-62.

[32] V. Yeshwanth, V. Raj and M. Saravanan, “Evolutionary Churn Prediction in Mobile Networks Using Hybrid Learning” Proceeding of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 2011.

[33] Y. Liu, and Y. Zhuang, “Research Model of Churn Prediction Based on Customer Segmentation and Misclassification Cost in the Context of Big Data”, Journal of Computer and Communications, Vol. 3, 2015, PP. 87-93

[34] C. Hilas, P. Mastorocostas and I. Rekanos, “Clustering of Telecommunications User Profiles for Fraud Detection and Security Enhancement in Large Corporate Networks: A case Study”, An International Journal Applied Mathematics & Information Sciences, Vol. 9(4), PP. 2015, 1709-1718.

[35] C. Cortes and D. Pregubon, “Signature-Based Methods for Data Streams”, Data Mining and Knowledge Discovery, Kluwer Academic Publishers. Vol. 5, 2001, PP. 167–182.

[36] M. Musolesi “Big Mobile Data Mining: Good or Evil?” IEEE Internet Computing, 2014, PP. 2-5.

[37] S. Nakajima and J. Gaudemer, “Big data for Telcos: How big data can get new revenue and reduce costs”, IDATE Research, December 2013.

Hoda Abdelhafez, hold PhD in Information Technology from Alexandria University, Egypt in 2002. She is a member of editorial board in International Journal of Data Science (Interscience) in 2014 and reviewer of the International Institute of Informatics and Systems since 2005. She has over 15 publications in international journals and conferences related to decision support systems, data warehouse, data mining, e-learning, big data and e-government. She has also two book chapters published in IGI Global in 2014.