A survey of predictive analytics using big data with data mining

Article *in* International Journal of Bioinformatics Research and Applications · January 2018

DOI: 10.1504/IJBRA.2018.092697

2 authors:

Poornima Selvaraj
SRM Institute of Science and Technology
15 PUBLICATIONS **264** CITATIONS

SEE PROFILE

Pushpalatha Marudappa
SRM Institute of Science and Technology
70 PUBLICATIONS **616** CITATIONS

SEE PROFILE

A survey of predictive analytics using big data with data mining

Article *in* International Journal of Bioinformatics Research and Applications · January 2018

DOI: 10.1504/IJBRA.2018.092697

CITATION
1

READS
307

2 authors:

Poornima Selvaraj
SRM University
8 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Pushpalatha Marudappa
SRM University
40 PUBLICATIONS **128** CITATIONS
SEE PROFILE

A survey of predictive analytics using big data with data mining

S. Poornima* and M. Pushpalatha

Department of CSE,
SRM University,
Chennai, Tamil Nadu, India
Email: poornima.se@ktr.srmuniv.ac.in.
Email: pushpalatha.m@ktr.srmuniv.ac.in
*Corresponding author

Abstract: Today, the world is filled with data like Oxygen. The amount of data being harvested and eaten up is flourishing vigorously in the digital world. The growing exploitation of novel inventions and social media leads to the generation of huge quantities of data called Big data which can bring remarkable information if analysed properly. Organizations may undergo for analysis of big data to having better decisions, thus big data analytics is being paid attention in recent times. For finding the concealed values from big data, society requires new schemes or strategies. Predictive analytics comprises of several statistical and analytical techniques for developing novel strategies for the future possibilities of prediction. Therefore, Predictive analytics becomes vital when an essential quantity of highly sensitive data has to be handled. Based on the perceived events, future probabilities and measures are predicted. With the aid of available data mining techniques, predictive analytics predicts the events in future and can make recommendations called prescriptive analytics. This review paper gives clear idea to apply data mining techniques and predictive analytics on different medical dataset to predict various diseases with accuracy levels, pros and cons, that concludes about the issues of those algorithms and futuristic approaches on big data.

Keywords: big data; classification; data mining; predictive analytics.

Reference to this paper should be made as follows: Poornima, S. and Pushpalatha, M. (2018) 'A survey of predictive analytics using big data with data mining' *Int. J. Bioinformatics Research and Applications*, Vol. 14, No. 3, pp.269–282.

Biographical notes: S. Poornima received her ME (CSE) Degree from Anna University Trichy and currently she is pursuing her doing PhD (CSE) in SRM University. She is working as an Assistant Professor in the Department of Computer Science and Engineering at SRM University. Her research interest includes Big Data Analytics and Data Science.

M. Pushpalatha received her PhD degree from SRM University. Currently working as a Professor in the Department of Computer Science and Engineering, SRM University. Her research interests include Wireless Adhoc Networks, Distributed Systems and Wireless Sensor Networks.

1 Introduction

Big data is a term used for describing the exponential growth along with the structured and unstructured availability of data. As a promising term, it contains the following characteristics:

i Volume: the amount of data generated.

ii Variety: the category to which the big data belongs.

iii Velocity: the speed of generation of data.

iv Variability: the inconsistency which can be shown by the data.

v Veracity: accuracy corresponding to the data is dependent over the truthfulness of the source data which are otherwise the quality of the data.

vi Complexity: data management is becoming very complex when storing large volumes of data from different sources.

Big data analytics is the procedure for the investigation of big data so as to reveal hidden patterns, unknown relations and some other useful information which can be employed to make better decisions.

Today, most of the companies store large volumes of diverse data (i.e. web logs, click streams, sensors and several other sources). The perceptions unknown within this 'Big Data' have significant business value. Several novel schemes have been developed to handle the challenges such as volume, variety, and velocity in big data. It is

i Apache Hadoop software that is a cost-economic, hugely scalable platform for the analysis of big data. It can save and do the processing of petabytes of data, inclusive of every data type which is not suitable for traditional relational database management system (RDBMS).

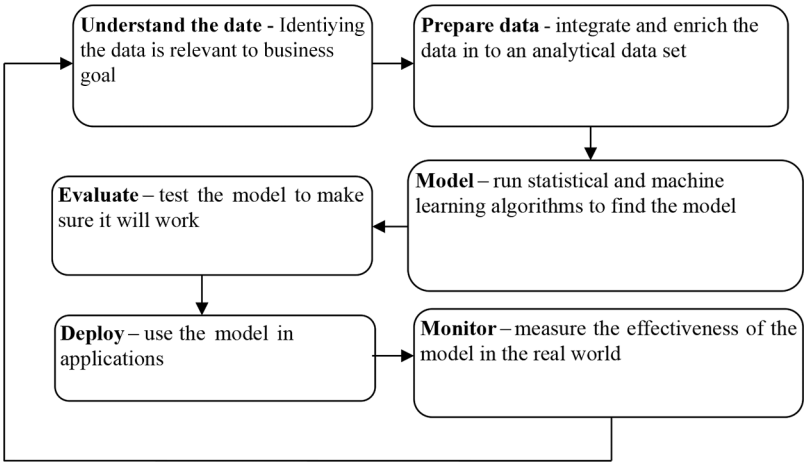
ii Not only structured query language (SQL) database lightens the restraints of the classical RDBMS to be capable of delivering a greater performance along with scalability. SQL databases can then have the abilities of Hadoop clusters extended by yielding low-latency object retrieval or else other data warehouse (DW)-like functionality.

iii Massively parallel-processing (MPP) appliances have the capacity of RDBMS-based data warehouses extended. These systems can save and then process petabytes of structured data.

iv In-memory databases considerably can enhance the performance through the elimination of most data access latencies on the shuttling of data forward and backward between the storage systems and server processors.

In-memory databases can be considered to be an alternative in few of the MPP appliances of today for offering realistic performance for the applications that demand high. Predictive analytics is a type of analytics undergone on big data that deal with extracting information from data and predict the trends and behaviour patterns. Predictive analytics determine the possible future result of an event or even the probability of a condition that can occur. It is one of the branches of data mining related to predict the future possibilities and their trends. Predictive analytics is useful for analysing huge data automatically with multiple variables; it is inclusive of decision trees, clustering, neural nets, market basket analysis, regression modelling, hypothesis testing, decision analytics, genetic algorithms, and text mining etc.[1] It contains different view approaches like integrated reasoning and pattern recognition along with predictive modelling. Many researchers have interest to build an automated reasoning tool for identifying future events and measures. Figure 1 indicates that the process of predictive analytics has to be consistent to guarantee efficiency and accuracy of the data prediction.

Figure 1 Predictive analytics process



Predictive and prescriptive analytics is the future of data mining. The terminologies data mining and data extraction are frequently confused with one another though the difference is significance (Zaman). Data extraction is involved with the receipt of data from one of the data source and having it loaded into a target database. Extraction of data can be done in this manner, from a source system, and it is loaded in a data mart or data warehouse. Data mining also refers extracting inconspicuous or hidden information from data marts or data warehouses. Data mining specifies knowledge discovery as the method used for the search of patterns in repositories of data. For knowledge discovery, data mining employs computational strategies from statistics, machine learning, and pattern recognition. Thus, the characteristics of data mining are described by search for patterns hidden in the data. Various tools are developed using predictive analytical models and strategies of data mining. The first step comprises the extraction of data by having access to huge databases. The data obtained in this way, are then processed with the support of sophisticated algorithms to look for concealed patterns and predictive information. Even though statistics and data mining are related with each other, methods employed in data mining seem to have evolved in domains except statistics.

A predictive model does the analysis for identifying the patterns observed in historical and transactional data so that different risks and potential are determined. The forecasting models acquire the relationships between several factors to permit the evaluation of the risks or else the opportunities that are associated with the certain listing of conditions, thereby directing the making of decision for the candidate transactions. Fundamental strategies for predictive analytics include

i data profiling and transformations

ii sequential pattern analysis

iii time series tracking [1]

The first strategy includes the functions that modify the row and column attributes, combines the fields, evaluates the dependencies, aggregates the records and data formats, and builds rows and columns.[2] Sequential pattern analysis determines that the relationships exist between the rows in database. Sequential pattern analysis is involved with the identification of the sequentially occurring items that are frequently seen across the ordered transactions over time. Time series tracking can be defined as a sequence that is ordered with values at different time intervals spaced with the equal distance.[2] Time series analysis provides the conception of data points that are plotted over time.

2 Literature survey

In past, predictive analytics can be applied in data mining for predicting future events especially in the medical sector, business, education, and crime detection. The health domain contains a bulk of concealed information that is significant in taking effective decisions. Babu and Sastry (2014) concentrated over the predictive abilities of Enterprise Resource Planning (ERP) systems, for the analysis of present data and historical facts so that opportunities and probable risks are identified for the organisations. Analytical decision management and business rules are utilised to make use of a decision in the form of a service.

Bellaachia and Guven (2005) proposed predicting breast cancer lastingness using data mining methods. The authors have examined three data mining methods such as Naïve Bayes, propagated neural networks, and c4.5 decision tree algorithms. Naïve Bayes method is the first method that uses the Bayesian method, because of its simple, clear, and fast predictive nature. The second method is artificial neural networks (ANNs) that uses multilayer network with transmission utilisation. Finally, they used c4.5 decision-tree algorithms. On the whole, the authors' work shows that the preliminary results are challenging prediction problem in medical data sets.

Data mining is the apt technology to predict patterns in the health sector data set. Though it is tedious to make the prediction of few diseases such as heart attack, due to its complexity, such tasks need more skill. Masethe and Masethe (2014) discussed to determine heart disease using classification algorithms. Few data mining algorithms such as j48, Naïve Bayes, REPTREE, and classification and regression trees (CART) are applied to predict heart attacks. The author's research work result shows that prediction accuracy is 99%, and j48, REPTREE, and CART gave a prediction model of 89 cases with a risk factor positive for heart attacks. From these techniques, it was identified that prediction of diagnoses can be done by data-mining algorithms.

A medical data of large size need powerful data analysis tools for processing. Data mining techniques can also be used for the diagnosis and predictive analysis. Ramaraj and Thanamani (2013) proposed predictive analytics methods to identify heart diseases. The authors' aim was to design a predictive method for heart disease detection. Classification accuracy report among various data mining techniques with the difference in error rates is provided in analysis part. The authors' final result shows that CN2Rule performs classification more accurately than the other methods.

Nasridinov et al. (2014) discussed a study on crime pattern prediction using data mining techniques. The authors analysed many data mining techniques with generated test data to determine the best method to perform crime pattern prediction task. Specifically, the authors did an extensive performance analysis of various data mining prediction algorithms such as support vector machine (SVM), decision tree, neural network, k-nearest neighbour, and Naïve Bayes. The authors assumed that wearable sensor devices are attached to the clothes of the user of the proposed method. It captures the inner temperature and heartbeat of a user and sends these data to the server to perform emotion mining. Danger condition was identified when the user developed high heartbeat, inner temperature, and camera surveillance that indicate the danger situation. When a danger condition was detected, the authors employed a test data generation method that cautiously designs test data set which comprises well-known data mining pattern prediction algorithms. This system is useful for law enforcement and emergency agencies to identify, analyse, and predict patterns, trends, and series, and provide useful information to solve, reduce, and prevent various danger situations promptly.

Chandra Shekar et al. (2012) make up a better algorithm for prediction of heart disease using case-based machine learning-based methods technique on non-binary data sets. Mining frequent item-sets in non-binary search space presented fascinating challenges over conventional mining in binary search space. Initially, the non-binary search space needs innovative tactics to calculate support and must be active. As there is a chance of removal of candidate item-set from the non-binary data set due to pruning, applying it at a higher level may become frequent. Support calculation and candidate generation at each level are carried out using separate mechanism. The author's final result was a prototype for generating frequent item sets for non-binary data set that was developed.

Maciejewski et al. (2010) introduced a model to use in spatiotemporal data, because the analysts are looking for the areas of space and time having unpredictably large occurrences of events, developed a predictive visual analytics toolkit which assists the analysts providing them with the linked spatiotemporal and statistical analytic views. Spatiotemporal events are simulated by the system by combining the kernel density estimation for event distribution and seasonal trend decomposition with the support of loss smoothing for the purpose of temporal predictions. Yue et al. (2009) especially addressed the predictive jobs which are related to the prediction of futuristic trends and then introduced RESIN that is an artificial intelligence blackboard-based agent leveraging the interactive visualisation and also the mixed-initiative problem solving so as to facilitate the analysis to look for and preprocess immense quantities of data for performing predictive analytics.

Riensche et al. (2009) explained an approach for supporting the design of games in the context of predictive analytics, developed for collecting the input knowledge, calculating the outcomes of complicated predictive techniques and social models, and examined those outcomes in quite an attractive manner. Huang et al. (2009) used the predictive analytics methods for establishing a decision support system for sophisticated network operation management and also to support the operators in predicting the possible failures in the network and then make the network adapt as a response towards hostile environments. The resulting decision support system facilitates the constant monitoring of the performance of the network and converts huge quantities of data into information that is actionable.

Sanfilippo et al. (2009) gave novel techniques for anticipatory critical thinking which realise a multiperspective approach for performing predictive modelling in aid of naturalistic decision making. In Banjade and Maharjan (2011), this technical work takes the linear regression method into consideration for the analysis of large-scale data set for providing helpful recommendations to aid the e-commerce customers using offline computations of the outcomes of the model.

Kone and Karwan (2011) predicted the expense incurred in delivering bulk (liquefied) gas to new customers making use of a multifactor linear regression model. Development of a single model, i.e. evaluating all the observations one time, leads to poor prediction outcomes. Hence, before regression analysis, a novel supervised learning method is utilised for grouping the customers who have similarity in some or the other perception. Hyperboxes are used to denote classes on customers, and subsequently, a linear regression model is developed within every class. To increase with the combination of data classification and regression, the accuracy of the prediction is indicated.

Bhat et al. (2011) presented a new preprocessing phase along with imputation of missing value for numerical and also categorical data. A hybrid combination consisting of classification and regression trees (CART), genetic algorithms for imputing the missing sequential values and self-organising feature maps (SOFM) for imputing the categorical values are used in the work.

Bhat et al. (2009) introduced an effective imputation technique employing a hybrid combination comprising of genetic algorithm and CART, in the role of a step of preprocessing. The traditional neural network model is used for prediction, over the data set that is preprocessed. Chinchor et al. (2010) address the merging of visual analytics and multimedia analysis to tackle with the information originating from multiple sources, having multiple aims or targets, and comprising multiple media varieties and combinations of those types. The resultant combination results in multimedia analytics. Razi and Athappilly (2005) performed prediction accuracy in three-way comparison that uses nonlinear regression, CART, and NNs models employing a consistent dependent variable along with a set consisting of dichotomous and categorical predictor variables.

Shweta et al. (2012) utilised the Naïve Bayes, decision tree, ANN, and C4.5 algorithms for the prognoses and diagnoses related to breast cancer. The results convey that decision trees offer greater accuracy of 93.62%, ANN yields 86.5%, Naïve Bayes produces 84.5%, and C4.5 produces 86.7%. Chaitrali et al. (2012) made use of Naïve Bayes, decision trees, and neural network algorithms for the analysis of heart disease. The comparison of the result shows that the Naïve Bayes attains about 90.74% of accuracy whereas the decision trees and neural network produces corresponding 99.62% and 100% accuracies, respectively.

Various data mining methodologies were used for predicting the heart disease by Soni et al. (2011). The accuracy of those algorithms are verified, in which the accuracies of Naïve Bayes, ANN and decision tree are said to have accomplished a respective 86.53, 85.53, and 89%. The data mining algorithms such as ANN, decision trees, and C4.5 apply ECG signals to analyse the heart disease. Decision tree algorithm is found to be the best and obtains 97.5% accuracy. The C4.5 algorithm yields an accuracy of 99.20%, whereas Naïve Bayes algorithm produces 89.60% of accuracy (Aneeshkumar and Venkateswaran, 2012). Therefore, these algorithms are employed for estimating the supervision over liver disorder.

C5.0 is a classification algorithm that is applied on huge data sets. It overcomes C4.5 in terms of the memory and speed along with the performance. This technique divides the sample depending on the field which provides the high information gain. Later, the obtained sample subset received earlier will be divided. The action will persist till the sample subset cannot be further divided. At last, the lowest level split in the sample subsets that have less than acceptable level contribution for the model will be eliminated. C5.0 methodology easily deals with the missing attribute and the multivalued attribute from data set (Patil et al., 2012).

Jen et al. (2013) introduced a technique for the analysis of prognostic indicators in the dental implant therapy. In their work, 513 patients are taken for analysis and evaluated 1161 implants from them. Data on 23 items are considered as impact factors over dental implants. These 1161 implants are then evaluated making use of C5.0 method. C5.0 approach generates 25 nodes. This model accomplishes the performance accuracy of 97.67% and specificity of 99.15%.

Jakrarin and Piromsopa (2013) used the K-means clustering making use of Apache Hadoop. Their goal was to efficiently analyse huge data set by reducing time complexity. They also defined whether the accuracy and detection rate are impacted by the number of fields present in log files. Hence, the results from their tests indicate the proper number of clusters and entries in the log files, though the accuracy rate is reduced when there is an increase in the number of entries. It was understood from the results that the accuracy needed to be improved.

Hall et al. (1998) provided an overview of a technique for developing the learning rules of the huge data set. The strategy is having a single decision scheme developed from a big data subset. Meanwhile Patil and Bichkar (2006) followed a hybrid means of pairing between a couple of genetic algorithms and decision tree for generating enhanced decision tree that improves the efficiency and performance of computation. The remarkable increase in the knowledge of big data helps use different methodologies for the analysis of data that are specified in "Data Reduction Techniques for Large Qualitative Data Sets". It defines that the choice for the specific method is dependent on the data set type and the manner in which the pattern has to be evaluated.

Jun et al. (2015) proposed an analytical methodology called divided regression analysis, for big data analysis to reduce the computing burden in regression problem. This approach is a statistical method that focuses much on small sample data; thus, it overcomes the computing burden in big data analysis. This approach should analyse entire data in big data analysis, which is considered as the population in statistics, in which data set is so large.

Tannahill and Jamshidi (2014) attempted to develop a bridge between the SoS and data analytic to evolve trustable models for such kind of systems. One among the recently hopeful data, analytic tool is 'deep learning'. It refers to extensive term applied to the

current evolution and extensions of the neural networks present in the community machine learning that has let for the state-of-the-art outcomes in speech, image, and tasks of natural language processing. Hierarchical learning is a field of research that throws limelight on learning the higher order representations from the low-level data. Learning that is used for object identification from images, recognising words, or syllables from the audio, or the recognition poses along with the movement from video sum to all the good instances of the advanced hierarchical learning research that has a centralised focus over the 'deep learning' effort taken in the community of machine learning and computational statistics. This work provides a further detailed outlook over all the big data analytic tools that are old and new.

Huang et al. collaborated interval regression in addition to the smooth support vector machine (SSVM) for the analysis of big data (Huang et al., 2014). In the recent times, the SSVM is introduced to be an alternate for the standard SVM, which has proved to have more efficiency compared with the conventional SVM in the processing of large-scale data. Moreover, the soft margin technique is introduced for modifying the divergence of separation margin and also to have effectiveness in the gray zone so that the data distribution tends to become difficult to be explained and also the separation margin between the classes.

Ma and Sun (2015) presented a leveraging method for big data regression. One of the emerging statistical methods is called leveraging (Rajan et al., 2015). Leveraging techniques are developed under a subsampling framework, in which a smaller portion of the data (subsample) is basically sampled out from the entire sample, which acts as a surrogate that carries out the necessary computations for the entire sample. The key to the success to leveraging techniques is building nonuniform sampling probabilities such that dominant data points get sampled with big probabilities. Such techniques are supposed to be the very distinct evolution of their kind in big data analytics and permit distributive access to huge quantities of data without moving to cloud computing and high-performance computing.

Quinlan gave a simplified decision tree without affecting the accuracy, which contains information related to thyroid dysfunction (Ma and Sun, 2015). It is downloaded from https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease link. The author needs to find whether a patient has a thyroid that is functioning normally or under-functioning thyroid (hypothyroid) or over-active thyroid (hyperthyroid). Nearly 7200 cases observed totally in the data set, among that 3772 cases ranging from the year 1985 and 3428 from the year 1986. About 92.6% (6666 cases) are under normal group, whereas the hypothyroid class accounts for 5.1% (368 cases) of the observations, and hyperthyroid class represents 2.3% (166 cases) of the data points. This data set is highly unbalanced with notorious difficulty for conventional classification techniques. For every one of the 7200 cases, there exist 21 attributes having 15 binary and six continuous variables that are exploited for determining the patient who belongs to which class under these three categories. Some of the attributes such as age, sex, health condition, and the results of various medical tests represent information on patients (Ma and Sun, 2015) (Table 1).

Table 1 Research gap for existing predictive analytics methods

Author	*Method name*	*Advantage*	*Disadvantage*
Nasridinov et al. (2014)	Decision tree, neural network, SVM, *k*-nearest neighbor, Naïve Bayes	Emergency agencies to identify, analyse, and predict patterns, trends, and series, and provide useful information to solve and reduce the problems	Prevent various danger situations promptly
Razi and Athappilly (2005)	Neural networks (NNs), nonlinear regression, and CART models	NNs and CART models provide better prediction compared with regression models when the predictor variables are binary or categorical and the dependent variable continuous	The regression model yielded a better performance slightly in the construction of model and model verification compared than CART and NN
Banjade and Maharjan (2011)	Linear regression	It allows the speedy model development and flexible integration of various parameters preserving the quality	It is not applicable for real-time data set
Yue et al. (2009)	Blackboard-based approach	It can massively enhance the accuracy of resin having the capability of predicting the future trends with time series analysis	Prediction capacity is less, and it is not focused on a long-range period by finding predictive model
Bhat et al. (2011)	CART and genetic algorithm + SOFM	This approach is simple, easier to be implemented, and is reliable practically	Model construction and model verification are not efficient
Bhat et al. (2009)	Hybrid combination of CART and genetic algorithm, ANN	This approach is simpler, comprehensible. Real-time decision making or else any other groups of applications requiring imputation can follow these steps in an effective manner	The high pace of convergence and in unsteadiness convergence in its training process
Shweta et al. (2012)	Naïve Bayes, ANN, C4.5, and decision tree	Decision tree is found to be the best predictor with 93.62% accuracy	Bayesian network has high efficiency
Chaitrali et al. (2012)	Decision trees, Naïve Bayes, and neural network	Neural networks provide accurate results as compared with decision trees and Naïve Bayes	It is not the suitable large number of data
Jakrarin and Piromsopa (2013)	K-means clustering	It shows that the amount of entry has an effect on the detection rate and the false alarm rate	The adequate iteration and the accuracy of that iteration are not focused
Patil and Bichkar (2006)	Genetic algorithm and decision tree	It is reduced problems such as slower execution, overloading the memory and processor with a huge database. The technique integrates the optimisation along with scalability and comprehensibility	It has a negative impact on predictive accuracy in many cases, and it is not solved
Patil et al. (2012)	C5.0 approach	Statistical analysis accuracy of c5.0 is 99.6 and CART of 94.8	It is unsatisfactory in practical application
		High classifying speed, strong learning ability, and simple construction	

3 Comments on results

In this section, the existing data mining-based predictive analytics algorithms performance is evaluated.

Performance evaluation can be done based on mean absolute error, one of the measurements of prediction accuracy. Also, one can compute the following three error metrics to measure prediction accuracy:

$$\text{Mean absolute percentage error} = \frac{1}{n}\sum\nolimits_{i=1}^{n} |Y_i - \hat{Y}_i / Y_i|$$

$$\text{Mean-squared error} = \frac{1}{n}\sum\nolimits_{i=1}^{n} \left(-\frac{\hat{Y}_i}{Y_i} \right)^2$$

where n is a number of patient's record $\hat{Y}_i$, is the forecasted value of days in bed due to disease (Y). Large prediction error is used in some cases to measure the accuracy in which the forecast error exceeds 100%.

Table 2 shows the graphical representation of accuracy comparison for various prediction methods. C4.5 and C5.0 approaches have achieved high accuracy compared with other methods such as ANN, Naïve Bayes, regression, neural network, decision tree, and CART. But the C5.0 has some disadvantages. The disadvantages are

Table 2 Comparative results of predictive analytics methods

Reference number	*Methods*	*Accuracy*	*Data set*
Shweta (2012)	Naïve Bayes	84.5	Cancer data set
	ANN	86.5	
	C4.5	86.7	
Soni et al. (2011)	Naïve Bayes	86.53	Heart disease data set
	ANN	85.53	
	Decision tree	89	
Razi and Athappilly (2005)	Regression	42.6 (LPE)	Data set of smokers
	Neural network	35.3 (LPE)	
	CART	32.5 (LPE)	
Aneeshkumar and Venkateswaran (2012)	Naïve Bayes	89.6	Liver disorder data set
	C4.5	99.2	
Chiang et al. (2013)	C5.0	97.67	Dental implant therapy data set

LPE, large prediction error

1 The construction of a decision tree is affected badly by irrelevant attributes, e.g. ID numbers.

2 Decision boundaries are rectilinear. Different looking trees may be generated due to small variations in the data.

3 Replication of subtree may occur several times. Considering too many classes for tree generation leads to error prone. Continuous class attributes value prediction is not suitable.

4 Challenges and future directions.

Existing big data analytics offers interesting opportunities, though it is also imposed with several lot of issues. The challenges are given below:

1 Privacy of data: privacy along with possession of data is an important issue. There are several organisations which have a belief that data must be free, and that kind of openness provides them with reasonable benefit.

2 Analysis of the user data: the need to focus and analyse user data is to determine the user's intent. This is the reason for the focus on most predictive analytics in big data.

3 Scaling of user data: possessing more amount of data is generally useful for data-based system; because of the popularisation of social media, massive database collection has been developed, and we are necessitated to put the restrictions by way of scalability of systems. The important challenge that is related with the scaling algorithms is such that the synchronisation and communications overheads increase, and hence, efficiency to the maximum could be lost, specifically in which the computation does not fit correctly into a map reduce model. This requires scalable analytics algorithms to generate results that are right on time (Rajan and Shanthi, 2015).

4 Big data can render the decision support to be easier, rapid, and having more accuracy as the resolutions are dependent on big quantity of data which are more recent and have greater relevance.

5 Analysing the structured data with the current statistical and data mining methods is easy, but for the analysis of unstructured data, improvement is needed in techniques like natural language processing, text analytics, and multimedia analytics.

Many of the recent algorithms have less efficiency on big data analytics that lags in the above issues. Hence, the availability of efficient analytics data mining algorithms becomes a necessity.

5 Conclusion

Predictive analysis is a progressive branch of data engineering that usually does the prediction of any existence or probability of data. Predictive analytics makes use of data-mining methods for making predictions about the events in the future and then yields recommendations by these predictions. The procedure consists of a historic data analysis, and depending on that evaluation, the prediction of the future events is done. Classification and regression hail to be the two chief goals of predictive analytics. It comprises different statistical and analytical methods that are employed for evolving the models which will do the prediction of future occurrence, events, or chances. Predictive analytics is capable of dealing with continuous and discontinuous changes. Classification, prediction and, to the particular extent, affinity analysis comprise the analytical techniques used in predictive analytics. The role taken by these predictive models differs

based on the data which are used by them. In this survey paper, the list of reviews are provided and discussed, how researchers already used predictive analytics for business and medical industry and also mentioned the techniques and algorithms with the issues while applying on big data. Thus, a novel approach or model could be generated to predict making use of predictive analytics modelling methods suitable for big data.

References

Aneeshkumar, A.S. and Venkateswaran, C.J. (2012) 'Estimating the surveillance of liver disorder using classification algorithms', *International Journal of Computer Applications*, Vol. 57, pp.39–42.

Babu, P. and Sastry, S.H. (2014) 'Big data and predictive analytics in ERP systems for automating decision making process', *5th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 27–29 June, Beijing, China, pp.259–262.

Banjade, R and Maharjan, S (2011) 'Product recommendations using linear predictive modeling', *Second Asian Himalayas International Conference on Internet (AH-ICI)*, Kathmandu, Nepal, 46 Nov, pp.1–4.

Bellaachia, A. and Guven, E. (2005) Predicting breast cancer survivability using data mining techniques, Department of Computer Science, the George Washington University, Washington.

Bhat, V.H., Rao, P.G., Krishna, S. and Shenoy, P.D. (2011) *An Efficient Framework for Prediction in Healthcare*, Springer-Verlag, Berlin Heidelberg, p.522–532.

Bhat, V.H., Rao, P.G., Shenoy, D., Venugopal, K.R. and Patnaik, L.M. (2009) 'An efficient prediction model for diabetic database using soft computing techniques', *Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, Springer-Verlag, Berlin Heidelberg, p.328–335.

Chaitrali, S., Sulabha, D. and Apte, S. (2012) 'Improved study of heart disease prediction system using data mining classification techniques', *International Journal of Computer Applications*, Vol. 47, No. 10, pp.44–48.

Chandra Shekar, K., Ravi Kanth, K. and SreeKanth. K. (2012) 'Improved algorithm for prediction of heart disease using case based reasoning technique on non-binary datasets', *International Journal of Research in Computer and Communication Technology*, Vol. 1, No. 7, pp.420–424.

Chiang, H-J., Tseng, C-C. and Torng, C-C. (2013) 'A retrospective analysis of prognostic indicators in dental implant therapy using the C5.0 decision tree algorithm', *Journal of Dental Sciences*, Vol. 8, No. 3, pp.248–255.

Chinchor, N., Thomas, J. and Wong, P. (2010) 'Multimedia Analysis + Visual Analytics = Multimedia Analytics', *IEEE Computer Graphics*, Vol. 30, No. 5, pp.52–60.

Hall, L., Chawla, N. and Bowyer, K. (1998) 'Decision tree learning on very large data sets', *International Conference on Systems, Man and Cybernetics,* San Diego, CA, USA, 14 Oct, pp.2579–2584.

Huang, Z., Wong, P.C., Mackey, P., Chen, Y., Ma, J., Schneider, K. and Greitzer, L. (2009) 'Managing complex network operation with predictive analytics', *Proceedings of the AAAI Spring Symposium on Techno social Predictive Analytics*, California, USA, 23–25 March 2009, pp.59–65.

Huang, C.H., Yang, K.C. and Kao, H.Y. (2014) 'Analyzing big data with the hybrid interval regression methods', *The Scientific World Journal*, Vol. 2014, pp.1–9.

Jakrarin, T. and Piromsopa, K. (2013) 'An analysis of suitable parameters for efficiently applying K-means clustering to large TCP dump data set using Hadoop framework', *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 10th International Conference*, Krabi, Thailand, 15–17 May, pp.1–6.

Jun, S., Lee, S.J. and Ryu, J.B. (2015) 'A divided regression analysis for big data', *International Journal of Software Engineering and its Applications*, Vol. 9, No. 5, pp.21–32.

Kone, E.R. and Karwan, M.H. (2011) 'Combining a new data classification technique and regression analysis to predict the cost-to-serve new customers', *Computers & Industrial Engineering*, Vol. 61, No. 1, pp.184–197.

Ma, P. and Sun, X. (2015) 'Leveraging for big data regression', *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 7, No. 1, pp.70–76.

Maciejewski, R., Hafen, R. and Rudolph, S. (2010) 'Forecasting hotspots—a predictive analytics approach', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 17, No. 4, pp.440–453.

Masethe, H.D. and Masethe. M.A. (2014) 'Prediction of heart disease using classification algorithms', *Proceedings of the World Congress on Engineering and Computer Science*', San Francisco, USA, 22–24 Oct.

Nasridinov, A., Byun, J.Y., Um, N. and Shin, H. (2014) A study on danger pattern prediction using data mining techniques, School of Computer Engineering, Dongguk University at Gyeongju, South Korea.

Patil, D.V. and Bichkar, R.S (2006) 'A Hybrid Evolutionary Approach To Construct Optimal Decision Trees with Large Data Sets', *IEEE International Conference on Industrial Technology*, Mumbai, India, 15–17 Dec, pp.429–433.

Patil, N., Lathi, R. and Vidya, C. (2012) 'Comparison of C5.0 and CART classification algorithms use pruning technique', Vol. 1, No. 4, pp.1–5.

Quinlan, J. (1987) 'Simplifying decision trees', *International Journal of ManMachine Studies*, Vol. 27, pp.221–234.

Rajan, C., Geetha, K. and Sasikala, R. (2015) 'Investigation on Bio-inspired population based metaheuristic algorithms for optimization problems in ad hoc networks', *World Academy of Science, Engineering and Technology*, Vol. 9, No. 3, pp.102–109.

Rajan, C. and Shanthi, N. (2015) 'Genetic based Optimization for multicast routing algorithm for Manet', *Sadhana - Academy Proceedings in Engineering Science*, Vol. 40, No. 7, pp.2341–2352.

Ramaraj, M. and Thanamani, A.S. (2013) Comparative study of CN2 rule and SVM Algorithm and prediction of heart disease datasets using clustering algorithms, Department of Computer Science, NGM College, Pollachi, India.

Razi, M.A. and Athappilly, K. (2005) 'A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models', *Expert Systems with Applications*, Vol. 29, No. 1, pp.65–74.

Riensche, R.M., Paulson, R., Danielson, R., Unwin, D., Scott Butner, R., Miller, M., Franklin, R. and Zuljevic, N. (2009) 'Serious gaming for predictive analytics', *AAAI Spring Symposium on Technosocial Predictive Analytics. Association for the Advancement of Artificial Intelligence (AAAI)*, p.108–113, San Jose, CA.

Sanfilippo, A.P., Cowell, J., Malone, L., Riensche, R., Thomas, J., Unwin, S., Whitney, P. and Wong, P.C. (2009) 'Technosocial predictive analytics in support of naturalistic decision making', *Proceedings of the 9th Bi-annual International Conference on Naturalistic Decision Making*, p.144–151.

Shweta, K. (2012) 'Using data mining techniques for diagnosis and prognosis of cancer disease', *International Journal of Computer Science, Engineering and Information Technology*, Vol. 2, No. 2, pp.55–66.

Soni, J., Ansari, U., Sharma, D. and Soni, S. (2011) 'Predictive data mining for medical diagnosis: an overview of heart disease prediction', *International Journal of Computer Applications*, Vol. 17, No. 8, pp.43–48.

Tannahill, B. and Jamshidi, M. (2014) 'Big data analytic paradigms—from PCA to deep learning', *Proc. AAAI Workshop*, Stanford, CA, USA.

Yue, J., Raja, A., Liu. D., Wang, X. and Ribarsky, W. (2009) 'A blackboardbased approach towards predictive analytics', *Proceedings AAAI Spring Symposium on Techno social Predictive Analytics,* California, USA, 23–25 Mar, pp .154–161.

Zaman, M. (2005) *Predictive Analytics the Future of Business Intelligence*. 08 Nov, Available from: www.mahmoudyoussef.com

Notes

1 http://www.articlesbase.com/strategic-planning-articles/predictiveanalytics-1704860.html

2 http://analyticsweb.com/predictive-analytics