BEST: International Journal of Management, Information Technology and Engineering (BEST: IJMITE)
ISSN(Print):2348-0513; ISSN(Online): 2454-471X;
Vol. 3, Issue 9, Sep 2015, 61-68
© BEST Journals

PREDICTIVE ANALYTICS USING BIG DATA: A SURVEY

G. KUMARESAN[1] & P. RAJAKUMAR[2]

[1]Research Scholar, Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India

[2]Assistant Professor, Computer Applications, Jeyaram College of Engineering, Tiruchirappalli, Tamil Nadu, India

ABSTRACT

Now-a-days, Information Technology is in the new era of Big Data, which provides more volume of data to researchers and analysts. We have large-enough data in our hand, this available data have precious insight values that make the administrators, policymakers and business analyst to take correct decision in right time. In order to find out the hidden values from the existing data, we need some methods or techniques. Hence, Predictive analytics is an essential technique while dealing with vital amount of potentially sensitive data. Here, based on perceived events, to predict the future probabilities, trends and measures. With the help of existing data mining methods, in order to make predictions about future events and recommendations. Predictive analytics is composed of various statistical and analytical methods used to develop a new techniques to predict future possibilities. In the end, this research work discussed predictive analytics various issues and challenges, available tools, applications and modeling techniques in big data.

KEYWORDS: Big Data, Predictive Analytics, Predictive Modeling, Learning Analytics, Educational Data Mining

INTRODUCTION

Big data is a term that is used to describe exponential growth and availability of data, both structured and unstructured. As a promising term it contains the following characteristics: (i) Volume: The quantity 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 of the data is based on the veracity of the source data that is quality of the data. (vi) Complexity: Data management is becomes very complex when storing large volumes of data from different sources. Big data analytics is the process of investigating big data to uncover hidden patterns, unknown relations and some other useful information that can be used to make better decisions.

Today, most of the companies are store large volumes of diverse data (i.e. web logs, click streams, sensors, and many other sources). The perceptions unknown within this "big data" hold tremendous business value. To handle the big data challenges such as Volume, Variety, and Velocity a number of new technologies has been evolved. They are (i) Apache Hadoop software is a cost-effective, massively-scalable platform for analysing big data. It can store and process petabytes of data, including all the data types that don't fit into traditional RDBMS. (ii) Not only SQL (NoSQL) databases relax the constraints of a traditional RDBMS to deliver higher performance and scalability. NoSQL databases can extend the capabilities of Hadoop clusters by providing low-latency object retrieval or other DW-like functionality. (iii) Massively parallel-processing (MPP) appliances extend the capabilities of RDBMS-based data warehouses. These systems can store and process petabytes of structured data. (iv) In-Memory databases dramatically improve performance by eliminating most of the data access latencies associated with shuttling data back and forth between storage systems and server processors.

In-Memory databases are available as an option on some of today's MPP appliances to provide real-time performance for the most demanding applications. Predictive analytics is a branch of big data that deals with extracting information from data and using it to predict trends and behaviour patterns. Predictive analytics applies innovative methods such as, real-time regression analysis and machine learning analysis to predict future measures. It contains different view approach like integrated reasoning, pattern recognition and predictive modeling. Many researchers have interest to building an automated reasoning tool for identifying future events and measures.

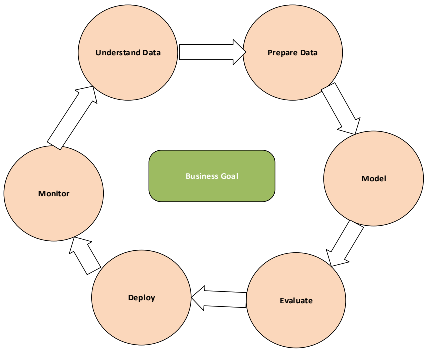

Figure 1: Predictive Analytics Process

Figure 1 shows that the predictive analytics process must be continuous to ensure effectiveness and accuracy of data prediction.

Related Work

Past years, predictive analytics can be applied in data mining for predicting future events especially in medical sector, business, education, and crime detection. The health sector today contains bulk of hidden information that can be significant in making powerful decisions. Abdelghani Bellaachia et al. proposed [3] predicting breast cancer lastingness using data mining methods. The authors have examined three data mining methods such as Naive Bayes, propagated neural networks and c4.5 decision tree algorithms. The first method is Naive Bayes method following Bayesian method, it is a simple, clear and fast predictive method. Artificial neural networks is the second method, it is a multi-layer networks with transmission is utilized. Finally, they uses c4.5 decision-tree algorithms. Overall, the authors study shows that the preliminary results are promising prediction problem in medical datasets.

Rajeswari at al. proposed three datasets of different diseases such as Heart, Breast cancer and Diabetes disease and tried to apply Feature Selection algorithm [5]. Feature selection is an important step in classification and also for dimensionality reduction. Feature Selection is a preprocessing method is used to identify the significant attributes, which play a dominant role in the task of classification. They analysed two different approaches for feature selection is done especially for medical datasets. They also shows a novel approach for feature selection using correlation and by generating association rules. The authors conclude that feature selection really helpful for dimensionality reduction and building cost effective model for disease prediction in medical datasets.

Data mining is the correct technology to predict patterns in the health sector dataset. Though, it is very difficult to predict some diseases like heart attack, as it is a complex job that needs more skill. Hlaudi Daniel Masethe et al. discussed to predict heart disease using classification algorithms [4]. They applied some algorithms to predict heart attacks such as j48, Naive Bayes, REPTREE and CART. The author’s research work result shows that prediction accuracy is 99% and j48, REPTREE and CART given a prediction model of 89 cases with a risk factor positive for heart attacks. This techniques strongly suggested that data mining algorithms are able to predict a diagnoses.

Huge amount of medical data that need powerful data analysis tools for processing. Data mining techniques can also be used for diagnosis analysis which is very important but difficult task that would be performed accurately and efficiently. Ramaraj. M et al. proposed [6] a predictive analytics techniques to identified heart diseases. The authors aim was to design a predictive method for heart disease detection. Analysis part provides the report for the classification accuracy among various data mining techniques with difference in error rates. The authors final result shows that CN2 Rule perform classification more accurately than the other methods.

IBM authors applied predictive analytics to higher education that can help institutions to accurately predict student behaviours in the areas of learning results, recruitment, and retention [2]. By analysing past data, predictive analytics can suggests an institution as to which students are most likely to enroll and which are likely to continue and graduate. Eduventures undertook research that included interviews with experts from a diversity of institutions such as public universities, private universities (minimum four years' experience) and community college. Adding predictive analytics to the institutional management toolbox allows for a continuous learning loop in which analysis informs decisions. These decisions lead to outcomes that are then assessed and combined with updated data to make better-informed decisions. By aggressively monitoring the risk factors, the institution knows who is most at risk – and with whom to intervene. The intervention steps often include "intrusive advising," in which students at risk must hold with an advisor in order to even register for the next term.

Predictive analytics also used in business for predicting business needs and improving strategies. Rick Nicholson et al. proposed [7] business contests facing oil and gas trade relation to asset optimization and present stage of predictive analytics for quality management. In this oil and gas industry, predictive analytics builds on prior investments in enterprise asset management systems, combines real-time data from sensors and other acquisition techniques with historical data to predict potential asset failures and permits the move from reactive to proactive maintenance. Predictive analytics can be used to analyse the real-time data from the sensors in the context of historical data and asset information held in this system to predict future conditions such as faults or failures and produce alarms or schedule maintenance or replacement. In the end, the author's analysis shows that, with the help of predictive analytics, to improve reduce non-productive time, recovery rates and reduce maintenance costs.

Aziz Nasridinov et al. discussed [8] a study on crime pattern prediction using data mining techniques. The authors analysed a variety of data mining techniques using the generated test data to determine which was potentially best for performing crime pattern prediction task. Specifically, authors analysed extensive performance evaluation on several data mining prediction methods, including Decision Tree, Neural Network, SVM, k-nearest neighbor, Naive Bayes, etc. The authors assumed that a user of the proposed method has wearable sensor devices attached to his/her clothes. It senses the heartbeat and inner temperature of a user, and sends these data to the server to perform emotion mining. Danger situation was detected when user produced high heartbeat, inner temperature, and camera surveillance indicates the danger situation. Once a danger situation was identified, the authors use a test data generation method, which carefully designs test dataset so that it comprises with well-known data mining pattern prediction algorithms. The proposed system can be 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 in a timely manner.

K. Chandra Shekar et al. proposed a new improved algorithm for prediction of heart disease using case based technique on non-binary datasets [9]. Mining Frequent Item-sets over non-binary search space presented the interesting new challenges over traditional mining in binary search space. In First, the non-binary search space needs new tactics to compute support and it has to be active. In Second, pruning cannot be applied to non-binary dataset as it may eliminate candidate item-set which at higher level may become frequent. The authors used separate mechanism for support calculation and candidate generation at each level. The author's final result was a prototype for generating frequent item-sets for non-binary dataset was developed.

Education Data Mining is a promising discipline which has an imperative impact on predicting students' academic performance. Suchita Borkar et al. evaluated student's performance using association rule mining algorithm [10]. Experiment was conducted using Weka and real time data set available in the college premises. The authors presented the potential use of education data mining using association rule mining algorithm in enhancing the quality and predicting students' performances in university result. The analysis revealed that student's university performance is dependent on Unit test, Assignment, Attendance and graduation percentage. The results reveal that the student's performance level can be improved in university result by identifying students who are poor unit Test, Attendance, Assignment and graduation and giving them more guidance to improve the university result.

B R Prakash et al. proposed [15] several applications for educational data mining and learning analytics which are in turn employed predictive analytics tools and techniques. The application areas include, User knowledge modeling, User behaviour modeling, User experience modeling, User profiling, Domain modeling, Learning components analysis and instructional principle analysis, Trend analysis, Application area, and Adaptation and Personalization. The authors proposed application will help both teachers and students to gain insights into student performance.

Jia Yue et al. applied predictive analytics to Global Terrorism Database (GTD) to resolve the missing information about a particular terrorism event and for predicting the possibility to such similar events in the near future [1]. By using machine learning techniques, black-board perspective reasoning, time series analysis and the GTD visualization tool they tried to make predictions based on available historical data. They have identified abstract representations of the tasks that could largely improve the accuracy of automated system.

Aditi Jain et al. summarized [16] the available tools and approaches for Higher Education Learning Analytics. The authors compared four Learning Analytics tools such as SNAPP, C4S, AWE, and PASS. The SNAPP tool is used to generate visual representations of user interactions, activities and patterns of behaviour. The visual representation shows the user's level of involvement in learning. It easily identifies users at low level of participation. It generates reports based on user interactions. C4S is an early warning tool and it automatically flag users who are likely to require additional support to complete their studies. The AWE is an alert engine and evidence based system that helps to identify the learners who are at high-risk may be struggling or experiencing disengagement in their courses. PASS is also an early alert tool it enhances learners retention and engagement in an online learning environment. It generates performance levels, visual signals, recommend contents and activities.

Issues and Challenges

Predictive analytics focuses on extracting hidden information or knowledge from large volume data and it building methods that can predict future events. Based on this events, security and accuracy are the major concerns in big data and some of the issues and challenges listed below.

Privacy of Data

Privacy and ownership of data is an important issue. There is always difference between producer and consumer of data, there are many organizations that believe that data should be open and that frankness provide them with a reasonable benefit.

Analysis of User Data

The need to focus and analyse user data is to determine the user’s intent. This is the reason for the focus of most predictive analytics in big data.

Scaling of User Data

Having more data is always useful for data based system, due to popularization of social media huge database repository has been created , we are necessitated to put the limits in terms of scalability of systems. The major problem associated with scaling algorithms is that communications and synchronization overheads rise and so most of efficiency can be lost, particularly where the computation does not fit correctly into a MapReduce model.

PREDICTIVE ANALYTICS TECHNIQUES

The available methods to examine predictive analytics can be categorized into regression and machine learning techniques. It has become easy to store, process huge amounts of data both structured and unstructured with Central Processing Unit, low-priced memory and new technologies like Hadoop, MapReduce, Text analytics and Memory database. This helps to discover unknown patterns and it provide new insights.

Regression Techniques

The statistical process for estimating the relationship between variables. Linear regression model finds the association between a dependent variable y and more than one independent variable X. Logic regression method converts data about the binary dependent variables into a boundless continuous variables. Discrete choice models explains, and guess choices between two or more discrete alternatives. Probit regression model permits the dependent variable can take two values for example male or female. Logit versus Probit both are sigmoid methods with a range between 0 and 1 and also they are inverse of the Cumulative Distribution Function (CDF).

Machine Learning Techniques

Neural Network (NN) provides the systems of interrelated neurons that can calculate values from inputs by feeding input through the network. Multilayer Perceptron (MLP) is Artificial Neural Network model this maps sets of input data into a set of appropriate outputs. Radial basis function is a real valued function their values only depend on the distance from the origin. A support vector machine is learning models and contains related learning algorithms that analyse and identify patterns in data. Naive Bayes are classifier based on applying Baye’s theorem. Classification and Regression are k-nearest and non-parametric functions. This model predicts objects class or value based on the k-closet objects in the feature space. Geospatial predictive model limits the distribution where the occurrences of events being modelled. Occurrences of events are not in uniform or random in distribution.

TOOLS FOR PREDICTIVE ANALYTICS

Predictive modeling is the task of developing, testing and validating a model to predict the probability of the future event. It can be done using a number of modeling techniques from machine learning, artificial intelligence, and statistics as we saw in earlier chapter. Since the improvement in technology, predictive analytics tools are no longer controlled to advanced users. The available tools come with minimized mathematical complexity and maximize graphical user interfaces. Recent predictive analytic tools also provide simple charts, graphs, and scores that help the business peoples to recognize the likelihood of promising outcomes. Some of the open source tools for predictive analytics are:

R Language

R is a platform for statistical calculations and graphics that runs on a wide variety of Windows, UNIX, and Mac OS Platforms. R provides an extensive range of statistical functionalities such as linear, non-linear modeling, statistical tests, classification and clustering models. It is extremely extensible and provides capabilities for data handling, calculation, and graphical display, calculations on array, and tools for data analysis, programming language that includes loops, conditionals and many other features.

Orange

Orange is a data visualization and data analysis tool. Data mining achieved through Python scripting or through visual programming. Orange remembers the choices and suggests most frequently used combinations.

Rapid Miner

RapidMiner is developed in java programming language. And it is a tool for data analysis. It contains data mining and machine learning algorithms such as data loading and transformation, data preprocessing, visualization, modeling, evaluation, and deployment. It utilizes learning schemes and attributes evaluators from Weka and statistical modeling schemes from R projects. It used for text mining, data stream mining, distributed data mining and for development of ensemble methods.

Weka

Weka is a java written group of machine learning algorithms for data mining process. This algorithm may be applied directly or indirectly to a dataset with the help of java code. Weka comprises tools for clustering, association, regression and data preprocessing. In this environment for developing a new machine learning algorithms.

Graph Lab Create

It is a machine learning tool and mainly built for developers and data scientists who have functional programming skills and data science. It permits them to simply prototype and scales their thoughts. Developers and data scientists are allowed to rapidly deploy and easily incorporate with other applications.

APPLICATIONS OF PREDICTIVE ANALYTICS

Predictive analytics can be placed to use in many applications, few examples where predictive analytics has exposed optimistic force in current years are as follows.

Analytical Customer Relationship Management (CRM)

Analytical CRM is a common commercial application of Predictive Analytics. Techniques of predictive analytics are applied to purchaser data to hunt CRM objectives in the company. CRM utilizes predictive analytics in applications for marketing, sales, and customer services. These tools are required for a company to position and focus their endeavours efficiently across the size of their customers. The company must analyse and recognize the products in demand or the products which are will be in most demand in near future, predict customers buying practices in order to promote applicable products at numerous finger tips, and practically recognize and moderate problems that have the potential to drop customers or skill to expand new customers.

Clinical Decision Support System

Experts use predictive analytics in health care mostly to determine about patients who are at risk and developing certain conditions for them like diabetes, heart disease, and other lifetime illnesses. And also for complicated clinical decision support systems that incorporates predictive analytics to conclude view of care.

Collection Analytics

Several organizations have a set of crook clients who do not make their payments on exact time particularly the financial institution has to commence collection activities on these clients to recover the amounts due. Sometimes impossible to collect due amount from client, such situations predictive analytics can help by identifying correct collection agencies, thus considerably increasing recovery as well as reducing time.

CONCLUSIONS

This paper provides the necessary details about predictive analytics. The area of predictive analytics can be applied and briefly discussed. Here, we introduced the number of tools and techniques in predictive analytics. Finally, list of reviews are provided and we discussed, how researchers already used predictive analytics for business and medical industry and also mentioned the techniques and algorithms. This predictive analytics has wide area for research concept. It provides more tools and techniques that can be applied in any environment where prediction is required.

ACKNOWLEDGEMENTS

The authors thank our beloved parents, family members and friends for their constant help and encouragement throughout this research work.

REFERENCES

1. Jia Yue, Anita Raja, Dingxiang Liu, Xiaoyu Wang, William Ribarsky, "A Blackboard-based Approach towards Predictive Analytics", Department of Software and Information Systems, Department of Computer Science, The University of North Carolina.
2. IBM "Predictive Analytics in Higher Education (2013), "Data-Driven Decision-Making for the Student Life Cycle, Eduventures.
3. Abdelghani Bellaachia, Erhan Guven "Predicting Breast Cancer Survivability Using Data Mining Techniques" *D*epartment of Computer Science, the George Washington University, Washington.
4. Hlaudi Daniel Masethe, Mosima Anna Masethe (2014), "Prediction of Heart Disease using Classification Algorithms" Proceedings of the World Congress on Engineering and Computer Science.

5. K. Rajeswari, Dr.V.Vaithiyanathan and Shailaja V.Pede, “Feature Selection for Classification in Medical Data Mining” Pune University, Pune, India. ISSN: 2278-6856.

6. Ramaraj.M, Dr.Antony Selvadoss, Thanamani, “A 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. ISSN: 2225-0603.

7. Rick Nicholson, Jill Feblowitz, Catherine Madden, Roberta Bigliani (2010), “The Role of Predictive Analytics in Asset Optimization for the Oil and Gas Industry”.

8. Aziz Nasridinov, Jeong-Yong Byun, Namkyoung Um, HyunSoon Shin, “A Study on Danger Pattern Prediction Using Data Mining Techniques” School of Computer Engineering, Dongguk University at Gyeongju, South Korea.

9. K. Chandra Shekar, K. Ravi Kanth, K. Sree Kanth (2012),” Improved Algorithm for Prediction of Heart Disease Using Case based Reasoning Technique on Non-Binary Datasets”, Department of CSE Department of IT, Hyderabad, India. International Journal of Research in Computer and Communication technology, ISSN: 2278-5841, Vol.:1, Issue. 7.

10. Suchita Borkar, K. Rajeswari (2013), “Predicting Students Academic Performance Using Education Data Mining”, MCA, Pune University, PCCOE, India, International Journal of Computer Science and Mobile Computing, ISSN: 2320–088X, Vol.:2, Issue. 7, pp. 273 – 279.

11. http://www.predictiveanalyticstoday.com/predictive-analytics-tools/

12. http://www.predictiveanalyticstoday.com/top-predictive-analytics-freeware-software/

13. Moty Fania, Parviz Peiravi, Ajay Chandramouly, Chandhu Yalla, “Predictive Analytics and Interactive Queries on Big Data”, Whitepaper, Big Data/Advanced Analytics, Intel IT.

14. Nishchol Mishra, Dr.Sanjay Silakari (2012), “Predictive Analytics: A Survey, Trends, Applications, Opportunities & Challenges”, International Journal of Computer Science and Information Technologies, ISSN: 0975-9646, Vol.:3 (3), pp. 4434-4438.

15. B R Prakash, Dr.M. Hanumanthappa, Vasantha Kavitha (2014), “Big data in Educational Data Mining and Learning Analytics”, International Journal of Innovative Research in Computer and Communication Engineering, ISSN: 2320 – 9798, Vol.:2, Issue. 12.

16. Amara Atif, Deborah Richards, Ayse Bilgin, Mauricio Marrone (2014), “Learning Analytics in Higher Education A Summary of Tools and Approaches”, 30th Ascilite conferences Proceedings.