Vaishali et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 1, March 2018, pp. 157-159

Analysis of Machine Learning Methodologies in Big Data Applications

Vaishali[1] , Sonika[2] ,Manu Narula[3] , Tanvi Dhingra[4]
Department of Information Technology, Northern India Engineering College, New Delhi, India
vaishalimahavar1997@gmail.com[1]
sonikaiskon@gmail.com[2]
manunarula1996oct@gmail.com[3]
tanvi.dhingra@niecdelhi.ac.in[4]

***Abstract* - The field of Big data is a trending research area in computer domain and several of the trades all round the world. It has gained immense affluence in extensive and varied application divisions. This comprises social media, education, stock markets, healthcare, weather forecasting, etc. Various sharp machine learning techniques were proposed and used to impart predictionbased on analytics solutions. So this paper deals with study of commonly used machine learning methods for big data analytics.**

***Keywords* — big data, algorithms, machine learning, predictivemodel, application division.**

I. INTRODUCTION

In this computer oriented age, it is necessary to use advanced analytics strategies on large, diverse big data collection to fabricate useful expertise and information. Big data analytics is an upcoming research area that involves the contribution, storage and study of enormous data sets to backtrack the undisclosed model and other important statistics. Big data analytics helps us to identify the data that are vital parts to the subsequent business decisions. Big data analytics can be mostly found in areas such as banking sector, hospital, education, social media and media industry, biometric applications, geographical applications, agricultural areas etc. It is a difficult task to handle big data using pre-existing data organizing applications. Thus to uncover data patterns, mode and associations, self learning methods can be used. The aim of this research paper is to discuss several machine learning algorithms used by analysts for examiningand modeling big data.The word big data[15] describes extremely huge data sets which are widely used among several researchers all over the world. Conventional relational databases are not competent of dealing with big data. Large quantity of data sets appears from various sources like sensors, contractual applications, internet and social media, etc. The big data occurence can be understood clearly by knowing the different V's related with them- Velocity, Volume, Value and Variety [14].

II.APPLICATIONS

Big data has made it's presence felt in various sectors like:

A. Entertainment and Business Organizations

They tend to study customer statistics besides the behavioral pattern to create a deep customer profile. This outline can be used to promote content for several target onlookers, favour data on demand and calculate content efficiency. For example: "Spotify" is an on request music provider, analyses customer behavior pattern and inspects it using available big data tools to distribute music suggestions. Amazon Prime is a subscription program that provides users an access to live video, music, electronic books, free transportation and several of other Amazon-specific services [13][14]

B. Financial Arenas

The economic conditions of market activities are inspected by the Securities Exchange Commission (SEC) through big data [13]. To avoid illicit trading in the stock markets, SEC is using several language processors and data analytics. Finance sector uses big data for making commitment support analytics, credit risk, sentiment study, prediction analytics etc. Anti-money concealment, fraud reduction and demand operation likelihood management are also considered using big data analytics[14].

C. Health profession

Medical care uses big data for neutral data analysis, disease pattern study, victim care quality and review, medical equipments and pharmaceutics supply chain management, drug uncovering and evolution analysis etc. With the formation of health applications, doctors provide observation based medication. Increase of infectious and transferable diseases are put on a tab using collective health data and social media websites[13].Origin for big data in health sector are genomes, e-health documentation, assessing health care devices, and wellness based mobile phone apps[14].

D. Education

Unites States "Education Department" is using big data to evaluate tutor and learner efficiency. Students 'snap patterns' are studied to measure for how long they retain on a specific topic. Educationist's teaching can be measured against total number of participants, locations and the topic delivered etc. University of Tasmania, an Australian based institute with more than thirty thousand trainees, has established a training and management system to check the professional position of trainer and learner who are involved in web based sessions [12], [13].

III. MACHINE LEARNING TECHNIQUES

A. Connectionist Systems

It can be used for predicting protein customized sites, effect of toxic cells effect in breast cancer, and foretelling electricity generation. Neural network with five hundred concealed neurons and increased conjugate gradient method was used to anticipate protein localization areas [4]. It gives better results as compared to probabilistic classification, decision tree and Bayesian classification. It gives a concise strategy for drug testing, modeling and expectation. Electricity generation foretelling system tells the quantity of power required at a rate approximate to the power utilization[3].This prototype is used to instruct the system for predicting significant power generation based on the composed data [5].

B. Decision tree Method

It was used to track patients at risk and disease patterns, recognition of speech and analyzing users who are present in video conference [12], [1], [6]. "The Medical Centre of Rush University", has formed an automated tool termed as "Guardian" .It uses a decision tree based machine learning steps to recognize at-risk patients and keep a track on disease trends [1].It is applied with large set of words for speech recognition. In evaluating users present in video conference, only the user's data are extricated from photographed images using a binary (white/black) decision tree. This is done so as to decrease data congestion [6].

C. Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

It can be used for segregating real images, initializing the beginning of codebook for image reduction and outlier observation[8].It performs very well on several large datasets and tends to suggest methods that are superior to Clustering Large Applications based on RANdomized Search(CLARANS) in terms of speed and quality for refining real images [7].

D. K means

It can be used to formulate significant connection between dietary habits of children and their tendency of catching cold [9], customer division [10] and image segregation [11]. K means was toextricit important steps from big data. In customer dealing , this algorithm has gained an accuracy measure of 0.95 specifying 95% exact differentiation of the customers. Image segregation was done using modifiable k-means clustering method for three- dimensional and multiple valuedimages[14].

E. Clustering Using Representatives(CURE)

It can be used to differentiate different behaviour. It uses a mixture of random sampling and divisioning that allows it to manage large database systematically[17].

F. Apriori algorithm

It can be used for customized marketing publicity and customer organization,product placing strategies in market stores and inventory management[18], medical data excavation[19],insurance related client analysis [20],packet signature extraction [21], network feed analysis [22] and game strategy analysis [23]. Apriori approach leads to hidden designing in the data.

Those with fascinating facts can be used to identify variables in the data and the similar occurrences of different variables that come along with the greatest frequencies. These patterns can be made to use when launching new products to users based on what they have purchased before or based on which products are bought collectively. It was used to form commonly patterns from available dataset for disease pointing and unfortunate drug reactions.

G. Frequent Pattern(FP) evolution

This algorithm can be used for lottery winning analysis and forecasting [24], web usage mining [21],gene ontology [16] and query solving systems [2]. It was used to review the previous winning numbers and consequently tell the next step of the lottery number in order to improvise the likelihood of winning. It can be used to search the very frequently accessed pattern initiated from the web data. It is applicable for mining either long or short frequent patterns.

IV.CONCLUSIONS

In this paper, we have studied the algorithms on neural networks, association and decision trees that are used for big data applications. Each algorithm is efficient as compared to the other for some applications .We have mentioned the applications of big data. We can extract data from any source and review it by finding suitable machine learning method to find answers that allow cost and time reductions, newer product development and intelligent decision making. In future, we would like to use big data tools for particular application and study their results.

REFERENCES

[1]MasamiAkamine and Jitendra Ajmera, "Decision tree-based acoustic models for speech recognition ",Springer Journal on Audio, Speech, and Music Processing,vol. 1,issue 10,2012.
[2]H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Y. Chang, "Pfp: parallel growth for query recommendation", In Proceedings of the 2008 ACM conference on Recommender Systems, pp.107–114, 2008.
[3] AbdolhosseinQaderi, Neda Dadgar,HamidrezaMansouri, " Modeling and prediction of cytotoxicity of artemisinin for treatment of the breast cancer by using artificial neural networks", SpringerPlus,vol. 2 ,pp.340, 2013.
[4] V Arulmozhi , Rajesh Reghunadhan, "Predicting the protein localization sites using artificial neural networks",Springer. m 8th German Conference on Chemoinformatics: 26 CIC-Workshop Goslar, Germany ,vol.5,issue 1,pp.11-13, 2013.
[5]MohammadNaimur Rahman, Amir Esmailpour, "An Efficient Electricity Generation Forecasting System Using Artificial Neural Network Approach with Big Data",IEEE,First International Conference on Big Data Computing Service and Applications,pp.213-217,2015.
[6]Yunsick Sung and Kyungeun Cho, "Development and evaluation of wireless 3D video conference system using decision tree and behavior network",Springer Journal on Wireless Communications and Networking,vol. 51, 2012.
[7]Zhang.T,Ramakrishnan R and Livny, M, "BIRCH: an efficient data clustering method for very large databases", In Proceedings of the ACM SIGMOD Conference, vol.25,issue 2,pp.103-114,June 1996
[8] Sajana. T, Sheela Rani C. M & Narayana K. V, "A Survey on Clustering Techniques for Big Data Mining",vol. . 9,ssue 3, 2016
[9] BesteEren, EzgiÇÖlgaKarabulut, S. Emre Alptekin, GülfemIúÖklarAlptekin, "A K-Means Algorithm Application on Big Data" Proceedings of the World Congress on Engineering and Computer Science , vol.2 , pp.814-818, October,2015
[10] Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu,," Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services" International Journal of Advanced Research in Artificial Intelligence, vol.4,issue 10, pp .40-44,2015.
[11] Ms. Mamta K. Date, Mr. S.P. Akarte, " Brain Image Segmentation Algorithm using K-Means Clustering" International Journal Of Computer Science And Applications,vol.6,issue 2, pp.285-289,Apr 2013.
[12]www.dezyre.com/article/top-10-machine-learning-algorithms/202
[13] www.informationweek.com/big-data/big-data-analytics/
[14]A. Vinothini; S. BaghavathiPriya, "Survey of machine learning methods for big data applications",2017 International Conference on Computational Intelligence in DataScience(ICCIDS), Pages: 1 – 5,2017
[15] Chun- Wei Tsai, Chin- Feng Lai, Han- Chieh Chao Athanasios V. Vasilakos, "Big data analytics: a survey",Springer, Journal of Big Data,vol.2,issue 21, pp.1-32,2015.
[16] Aleksandra Gruca, "Improvement of FP-Growth Algorithm for Mining Description-Oriented Rules", Man-Machine Interactions Advances in Intelligent Systems and Computing , vol. 242,pp. 183-192,2014.
[17] SudiptoGuha , Rajeev Rastogi , Kyuseok Shim, " CURE: An Efficient Clustering Algorithm for Large Data sets" Proceedings of the ACM SIGMOD Conference,vol. 27,issue 2,pp.73-84,June 1998
[18] Arun,Dr.L.Jabasheela, "Big Data: Review, Classification and Analysis Survey",International Journal of Innovative Research in Information Security,vol.1, Issue 3 ,September 2014.
[19]XiaoyanCui,ShimengYang,Darning Wang, "An algorithm of apriori based on medical big data and cloud computing" IEEE 4th International Conference on Cloud Computing and Intelligence Systems ,pp.361- 365,2016
[20] Jun Liu, Meiling Cong, " The optimization of apriori algorithm based on array and its application in the analysis of insurance clients" IEEE 4th International Conference on e-Technologies and Networks for Development ,vol.4,pp.58-61,2015.
[21]Linhui Tao, Guangjie Liu , Weiwei Liu , "Packet signature mining for application identification using an improved Apriorialgorithm",IEEE International Conference on Progress in Informatics and Computing ,pp.633-637,2015
[22]Preetish Ranjan, Abhishek Vaish, "Apriori Viterbi Model for Prior Detection of Socio-Technical Attacks in a Social Network" ,IEEE,International Conference on Engineering and Telecommunication,pp.97-101,2014.
[23]LiuTianbiao, Hohmann Andreas, "Apriori-based diagnostical analysis of passings in the football game",IEEE International Conference on Big Data Analysis,2016.
[24]Jianlin Zhang, Suozhu Wang, HuiyingLv, Chaoliang Zhou, "Research on Application of FP-growth Algorithm for Lottery Analysis" ,Springer, Proceedings of 3rd International Conference on Logistics, Informatics and Service Science,pp.1227-1231,2015.

