A Survey of Big Data Analytics in Healthcare

Muhammad Umer Sarwar*, Muhammad Kashif Hanif†, Ramzan Talib‡, Awais Mobeen§, and Muhammad Aslam¶
Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—Debate on big data analytics has earned a remarkable interest in industry as well as academia due to knowledge, information and wisdom extraction from big data. Big data and cloud computing are two most important trends that are defining the new emerging analytical tools. Big data has various applications in different fields like traffic control, weather forecasting, fraud detection, security, education enhancement and health care. Extraction of knowledge from large amount of data has become a challenging task. Similarly, big data analysis can be used for effective decision making in healthcare by some modification in existing machine learning algorithms. In this paper, drawbacks of existing machine learning algorithms are summarized for big data analysis in healthcare.**

Keywords*—Big data; Analytics; Healtcare; Analytical tools; Machine learning

I. Introduction

According to Sutherland and Shan, big data is based on three main properties, i.e., volume, velocity, and variety [1]. A large volume of data is being produced by various sources like astronomy, environmental data, transportation data, stock market transactions, census, airline traffic, internet images etc. The rate at which data is being generated from different resources is referred to velocity. Variety is different types of data including text, audio, images, video etc. In statistical perspective, big data is not big just in volume only but it is also big in terms of dimensions. Dimensions are also termed as features. Existing traditional methods of data mining can hardly give useful information. There is need for modification in existing machine learning methods for better data extraction and decision making [2].

Recently, the trend in healthcare is now diverting from cure to prevention due to rapid increase in data. Scientist have focused to make improvement in reliability and efficiency of healthcare systems to minimize the curing cost in healthcare and also to deliver better medication to patients. Hospitals and healthcare system are good warehouse of big data like patient history, test reports, medical image [3]. It is challenging task to cope with large amount of unstructured data and missing values. Improvements in existing data mining and machine learning approaches can help to develop personalized medicines which can cure and prevent diseases [4]. Existing traditional machine learning algorithms works on centralized databases which requires a lot of time to train and analyze large amount of data. Similarly, it is not feasible to store and process big data on single machine. Therefore, there is need to parallelize existing approaches and modify these approaches with hybrid approaches that have enough capabilities to overcome the challenge of storing and processing large data set in distributed environment [1].

A lots of research has been carried out in healthcare domain to train the system and predict the expected result for the patient. Microarray data has become a major interest area in healthcare domain in recent few years. There are thousands of dimensions in microarrays data that require huge amount of processing in terms of cost and time [5]. Dimension reduction techniques are applied to extract relevant information from large dimensions.

II. Big Data

Data is considered as much important as oil. However, oil in unprocessed form is hardly of use. Similarly, data in unprocessed form in not useful. Important information can be extracted from data by using different analytical methods. According to Hermon [3], big data analytics can bring the revolution in healthcare industry. This data in healthcare provide opportunity to perform predictive analysis. Big data analytics has a great potential to process large amount of data in parallel and solution of hidden problems can also be find. This analytical approach can be implemented to minimize cost of processing time on large amount of data. For example, any disease that has occurred earlier in any parts of the world, prediction of that disease can be done efficiently. Although, clinics and hospitals can reprocess the data to analyze and calculate the patient preferences. In predictive analysis, we implement different statistical methods, data mining, and machine learning approaches to analyze, process, and predict the conclusion for undiscovered data. Healthcare domain has a lot of possibilities to provide better cure for disease using different analytical tools.

Processing of big data can be organized into four layers (Fig. 1). To process large amount of data collected from different sources which can be in different formats is a challenging task. Due to unstructured data, traditional database management system cannot be use for knowledge extraction form data. Big data may include structured, semi-structured, and unstructured data [6]. First, we collect data that is generated from different origins and then collect and store it into one common platform. Most commonly, we use Apache Hadoop that is open source framework to provide Hadoop Distributed File System (HDFS) for distributed storage and fault tolerance [7]. MapReduce is the programming model of Hadoop which can be used to process huge amount of data as quick as possible [8]. The dataset is partitioned into training and testing subsets [9]. Machine learning algorithm can be implemented to perform intelligent analysis on input data and produce information that can be used to produce reports in processing layer.

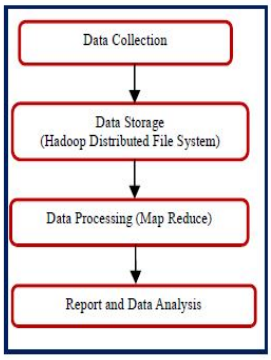

Fig. 1. Big data processing

III. Analysis and Prediction using Machine Learning Approaches

In industry, how the market is growing can be predicted using data analysis. Prediction is done using machine learning algorithms. Different factors are investigated in making prediction. There should be a prior knowledge about the class about which the prediction model is going to work. Two types of learning approaches used in healthcare are supervised and unsupervised learning. To deal with large amount of features, we apply dimensionality reduction approaches to obtain relevant features. Dimensionality reduction eliminates the unnecessary features to speed up computation and prediction for accurate decision [1]. In the following section of paper, different data mining approaches are summarized.

A. Feature Selection and Evaluation

Initially, data pre-processing is performed to reduce the noise and redundant data to speedup computation. As the dataset is divided into training and testing subsets, the training subset is used for the feature extraction and selection.In this manner, subset from given features are selected.These features are derived from pixel intensity,colors and geometric features such as contours, edges and shapes [9]. These features are further used for elimination of redundant and noisy features. This step will help for model interpretation. Different approaches can be used to obtain optimal feature subset.

- **Complete search**
  This search guarantees the best solution. It can be applied for finding optimal solution of big data problems. Heuristic approaches like branch and bound helps to minimize the searching whole feature space.

- **Sequential search**
  In this type of search, heuristic approach is applied [10]. This approach search either from whole feature set or null feature subset to obtain optimal solution by adding or removing features, respectively. Added features cannot be removed and removed features cannot be added. This approach does not guarantee for optimal solution. However, solution of this approach become acceptable due to less processing time.

- **Random Search**
  Random features are selected to start this search. Local optimal solution can be achieved using randomness. Performance of random search can be increased by integrating with sequential search for generation of random subset like simulated annealing and random start hill climbing algorithm.

Feature evaluation step involves values that are assigned to feature subset depending on specific criteria. Similarity is determined using class labels for classification. Finding irrelevant features is a challenging task in clustering. The concept behind feature evaluation claims: “Good features subsets contains features highly correlated with class, yet uncorrelated with each other” [11]. Commonly used for feature evaluation are wrapper, filter, and hybrid appraoches.

Wrapper approach has ability to provide optimal solution which can be tuned for classifier learning. Algorithm performs searching to select particular feature subset based on criterion function (Fig. 2) [12]. Computational cost increases since algorithm is run at each iteration. This technique involves high computational cost and not suitable for solving big data problems.

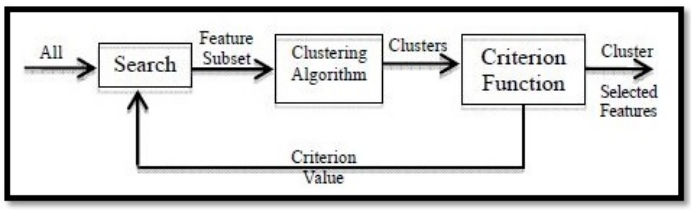

Fig. 2. Wrapper approach

Filter approach can be used to determine relevance between selected features (Fig. 3). Subset production is not used for classifier learning. Therefore, generic result is produced instead of specific algorithm tuning [12]. Its suitability becomes high for solving big data problems. This approach performs batter than other approaches like Relief attributes estimator [13].

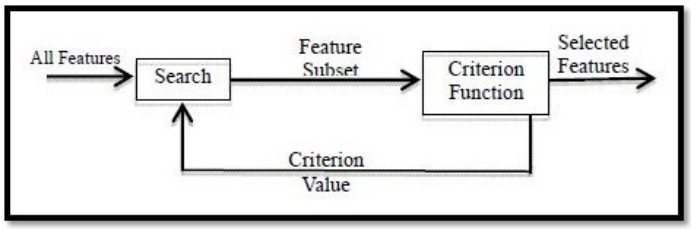

Fig. 3. Filter approach

Hybrid approach performs evaluation on features and make a feature subset by choosing the best among them in further iterations. Comparison between different types of features subsets is performed for optimal learning of algorithm [14]. Hybrid approaches have gain more importance as compared with other approaches. It involves trade-off between time and efficiency. From healthcare aspect, it is more suitable approach than other approaches. There is scope of sub-optimal solution [15].

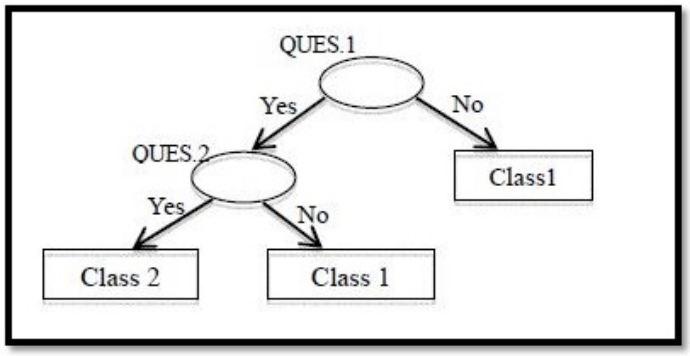

Fig. 4. Classification by decision tree

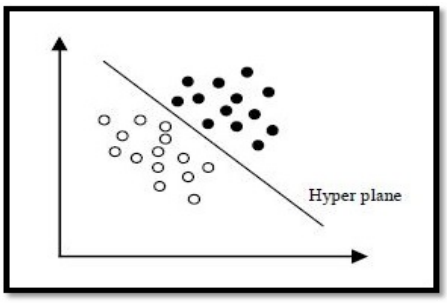

Fig. 5. Classification by support vector machine

B. Classification

Classification model classifies input data and target class is used by classifier for training and testing purpose [16]. Input is given to machine learning algorithm; target class data is also given for performing correct decision by classifier that involve features of input data based on classifier model for target class. After training of classifier, next phase is testing in which input data is given to perform prediction about target class. Due to increase in volume of big data, existing classification approaches have some limitations and involve high processing cost [1]. Commonly used classification techniques in healthcare domain are decision tree, support vector machine, neural network, k-nearest neighbour, and Bayesian approach [17].

A decision tree is a tree like structure used for classification [18]. Decision tree is applied for gaining accurate and fast results due to its simple structure. It difficult to construct decision tree with huge amount of data because a lot of time is required to construct it. Decision tree is most commonly classification approach in healthcare domain for problem solving by assigning class label to patient. Fig. 4 represents a sample decision tree.

Support Vector Machine (SVM) is a statistical model used for classification. SVM is capable for making decisions on large data set.SVM is very beneficial specially multidomain applications in big data environment. SVM is mathematically complex and computational expensive [19]. Multi-level or binomial classification can be performed using SVM [20]. SVM is most popular approach among existing machine learning techniques. The performance of SVM degrades on larger data set that consist of noisier data. This method performs prediction very fast after training [21]. Prediction is done based on hyper plane and support vector that performs separation in higher dimensional space. To overcome problem of noisy data, SVM is combined with other machine learning techniques for obtaining better results [22]. Fig. 5 shows SVM classification is represented by hyper plane for decision making.

In traditional machine learning techniques, Neural Network (NN) depicted lot of variation. In NN, weights of connecting links are adjusted between neurons until reaching optimal value. NN is most commonly used for problem solving is Multilayer Perceptron (MLP) [23]. It works similar to human brain. First, NN is trained to perform classification. After completion of training, testing is performed on input data. Major drawback of using NN is computation time for large data set. Memory requirement also increases as size of data set increases. NN is suitable for gaining optimal results but it requires more time for larger data set. There is need to adopt hybrid approach to minimize computational time by combining NN with other approaches to overcome issue of computational time for larger data set [24].

Convolutional Neural Network (CNN) is the advance form of NN. CNN is a multilayer neural network that takes input in vector form. However in case of medical images pixels or voxels are information source.In standard multi-layer neural network, Convolutional layers interlard with sub-sampling layer followed by fully connected layer is designed to batter use of spatial information by taking 2D or 3D images as input [25].

K-Nearest Neighbour (KNN) is a simple classification model that works according to nearest neighbor of existing class label [26]. In KNN, value of k is computed to find nearest neighbor where k represents number of nearest neighbours. To obtain accurate results, optimization algorithms are applied to minimize computational cost. KNN provides better results when compared with Bayesian method in various applications [26].

Bayesian method works according to Bayes theorem.Bayes theorem provide mathematically grounded tools to find out the uncertainty of a model [27]. For larger data set, Bayesian classifier can perform better in classification [28]. Naive Bayesian model provide high accuracy only when attributes values are independent [17]. It is a statistical model and provides high accuracy. This appraoch assumes all attributes are independent according to each other. This classifier can perform well in healthcare either by pre-processing or without pre-processing.

C. Clustering

Clustering groups similar data together to make clusters. Target class label is not provided initially in clustering. There exists higher similarity within same cluster. Different clusters have lower similarity between data points. Traditional approaches that are being used for similarity measures are Jaccard measure, Pearson correlation, euclidean distance and Cosine [15], [29]. There is no need for previous information about data is required to work upon clustering. It is suitable for applying on microarray data in which very little information is required about genes. Tapia et al. implemented genetic algorithm on expression data to analyze it [30]. It has capability to

represent information in compact form without losing much information by producing optimal clusters with reference to big data [31]. There exists different clustering algorithms.

Partitioned clustering involves predefined number of clusters. In this technique, data set is classified into predefined number of partitions. There is none of empty partition and data belongs to exactly one cluster. This approach can be further classified based on cluster centroid and similarity measures, i.e., K-Medoids and K-Means. In K-Medoids, medoids are used instead of centroid. Center point of a cluster is medoid that exist in data set. Belciug et al. applied clustering for detection of breast cancer to obtain better accuracy [32].

In hierarchical clustering, it is not necessary to pre-define number of clusters [31]. There are two main categories of hierarchical clustering with respect of working. Agglomerative technique is a bottom up technique in which every data point is assumed as cluster while two different clusters are merged on the basis of few similarity measures [33]. Desired cluster can be obtained after some iterations. Major drawback of this appraoch is integration cannot be rollback. Divisive technique is top-down technique in which all data points are treated as a cluster. After some iterations, cluster is further classified into two classes based on some measures [33]. Required number of clusters can be achieved by repeatedly running this algorithm. This approach also has major drawback that once cluster is divide into sub-clusters, these cannot be grouped together to make original cluster. In hierarchical clustering algorithms, iterations can be stop by gaining desired number of clusters.

IV. Fusion of Different Classifier

All existing traditional data mining techniques have not capability to perform better classification for big data in healthcare domain. Therefore, there is need to merge the distinct techniques together to perform better classification [34]. One popular approach for requiring relationship between data is Association. It is necessary to obtain relationship between diseases for finding similar treatment. Apriori algorithm is applied in association to find relationship between items and also to perform separation between similar as well as different items. PSO optimization approach is mostly combined with SVM to perform optimization first on feature set and then classifier is applied to separate data [33]. KNN is fused with fuzzy logic to decrease computation time [35]. It is better to combine different techniques in healthcare domain for obtaining better classification.

V. Related Work

Different researchers are investigating to improve performance of existing approaches by combining with other techniques. From few recent years, modification in traditional machine learning approaches has earned remarkable interest in healthcare domain. To increase classifier accuracy, different fusion techniques have been proposed. Few traditional machine learning related works in healthcare domain is discussed. Chest disease is classified using artificial neural network [36]. Chronic illness classification has been carried out [34]. For targeting high-dimensional data, conventional method of filter has been modified [37]. Distributed K-Nearest Neighbour has been applied for learning [38]. For detection of Parkinsons disease, fuzzy approaches have been fused with K-Nearest Neighbour to better classification [35]. To find irregular relationship in health data set, Association has been implemented [39]. Different big data processing platforms have been applied for classification. To reduce drawbacks of traditional approach, Map Reduce application of decision tree method has been proposed [40].

VI. Conclusion

Due to rapid enhancement in big data prediction and analysis, healthcare domain has got a valuable attention from recent few years. Existing traditional machine learning approaches take much time in computation when data set volume increase. Similarly, in healt care domain huge data is collected from different sources and researchers always try to make problem simpler to patient. In big data, we can find remarkable and hidden information that can help in understanding the nature of problem more deeply.

References

[1] S. Suthaharan, "Big data classification: Problems and challenges in network intrusion prediction with machine learning," *SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 4, pp. 70–73, 2014.

[2] B. P. RAO, "of the notes: Brief notes on big data: A cursory look," 2015.

[3] R. Hermon and P. A. Williams, "Big data in healthcare: What is it used for?" 2014.

[4] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and Systems*, vol. 2, no. 1, p. 1, 2014.

[5] O. H. Fang, N. Mustapha, and M. N. Sulaiman, "Integrating biological information for feature selection in microarray data classification," in *Second International Conference on Computer Engineering and Applications (ICCEA)*, vol. 2. IEEE, 2010, pp. 330–334.

[6] P. Groves, B. Kayyali, D. Knott, and S. Van Kuiken, "The big data revolution in healthcare," *McKinsey Quarterly*, vol. 2, 2013.

[7] S. Chandra and D. Motwani, "An approach to enhance the performance of hadoop mapreduce framework for big data," in *Micro-Electronics and Telecommunication Engineering (ICMETE), 2016 International Conference on*. IEEE, 2016, pp. 178–182.

[8] K. R. Satish and N. Kavya, "Big data processing with harnessing hadoop-mapreduce for optimizing analytical workloads," in *International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2014, pp. 49–54.

[9] A. S. Panayides, C. S. Pattichis, and M. S. Pattichis, "The promise of big data technologies and challenges for image and video analytics in healthcare," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 1278–1282.

[10] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[11] H. Liu and H. Motoda, *Computational methods of feature selection*. CRC Press, 2007.

[12] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[13] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[14] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on knowledge and data engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[15] R. Dharavath and A. K. Singh, "Entity resolution-based jaccard similarity coefficient for heterogeneous distributed databases," in *Proceedings of the Second International Conference on Computer and Communication Technologies*. Springer, 2016, pp. 497–507.

[16] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," in *Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*. IEEE, 2013, pp. 1–7.

[17] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare," *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[18] D. Wang, X. Liu, and M. Wang, "A dt-svm strategy for stock futures prediction with big data," in *IEEE 16th International Conference on Computational Science and Engineering (CSE)*. IEEE, 2013, pp. 1005–1012.

[19] S. Suthaharan, "Support vector machine," in *Machine Learning Models and Algorithms for Big Data Classification*. Springer, 2016, pp. 207–235.

[20] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[21] G. Cavallaro, M. Riedel, M. Richerzhagen, J. A. Benediktsson, and A. Plaza, "On understanding big data impacts in remotely sensed image classification using support vector machine methods," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 8, no. 10, pp. 4634–4646, 2015.

[22] Y. Tang and J. Zhou, "The performance of pso-svm in inflation forecasting," in *2015 12th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE, 2015, pp. 1–4.

[23] B. Chandra and R. K. Sharma, "Fast learning for big data applications using parameterized multilayer perceptron," in *IEEE International Conference on Big Data*. IEEE, 2014, pp. 17–22.

[24] V. Bijalwan, V. Kumar, P. Kumari, and J. Pascual, "Knn based machine learning approach for text and document mining," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 61–70, 2014.

[25] D. S. S. Kevin Zhou, Hayit Greenspan, *Deep Learning for Medical Image Analysis*, 1st ed. joe Hayton, 1 2017.

[26] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*. Chapman & Hall/CRC Boca Raton, FL, USA, 2014, vol. 2.

[27] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[28] G. J. Torres, R. B. Basnet, A. H. Sung, S. Mukkamala, and B. M. Ribeiro, "A similarity measure for clustering and its applications," *Int. J. Electr. Comput. Syst. Eng.*, vol. 3, no. 3, pp. 164–170, 2009.

[29] J. J. Tapia, E. Morett, and E. E. Vallejo, "A clustering genetic algorithm for genomic data mining," in *Foundations of Computational Intelligence Volume 4*. Springer, 2009, pp. 249–275.

[30] A. Fahad, N. Alshatri, Z. Tari, A. Alamri, I. Khalil, A. Y. Zomaya, S. Foufou, and A. Bouras, "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," *IEEE transactions on emerging topics in computing*, vol. 2, no. 3, pp. 267–279, 2014.

[31] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[32] K. Sasirekha and P. Baby, "Agglomerative hierarchical clustering algorithm–a review," *International Journal of Scientific and Research Publications*, vol. 3, no. 3, p. 1, 2013.

[33] C. Xiang, Y. Xiao, P. Qu, and X. Qu, "Network intrusion detection based on pso-svm," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12, no. 2, pp. 1502–1508, 2014.

[34] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early-warning system for chronic illnesses," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8852–8858, 2012.

[35] W.-L. Zuo, Z.-Y. Wang, T. Liu, and H.-L. Chen, "Effective detection of parkinson's disease using an adaptive fuzzy k-nearest neighbor approach," *Biomedical Signal Processing and Control*, vol. 8, no. 4, pp. 364–373, 2013.

[36] O. Er, N. Yumusak, and F. Temurtas, "Chest diseases diagnosis using artificial neural networks," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7648–7655, 2010.

[37] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, vol. 3, 2003, pp. 856–863.

[38] R. Mall, V. Jumutc, R. Langone, and J. A. Suykens, "Representative subsets for big data learning using k-nn graphs," in *IEEE International Conference on Big Data*. IEEE, 2014, pp. 37–42.

[39] Y. Ji, H. Ying, J. Tran, P. Dews, A. Mansour, and R. M. Massanari, "Mining infrequent causal associations in electronic health databases," in *IEEE 11th International Conference on Data Mining Workshops*. IEEE, 2011, pp. 421–428.

[40] W. Dai and W. Ji, "A mapreduce implementation of c4. 5 decision tree algorithm," *International Journal of Database Theory and Application*, vol. 7, no. 1, pp. 49–60, 2014.