A Survey Of Machine Learning Algorithms In Health Care

Sweety Bakyarani. E, Dr. Srimathi. H , Dr. M. Bagavandas

Abstract: Data is Knowledge and Knowledge is Power. In this age of information overload, Data Analytics has changed the perspective on how to effectively solve problems in many industries. One of the potential areas where Data Analytics can have a very positive outcome is in the field of health care. Health Care analytics can not only benefit patients but also all the stake holders and key players in the health care industry. It has the potential to prevent disease outbreaks, identify and detect diseases, reduce cost of operation for hospital administrators help government with health care policies and thus improve the overall quality of life. Machine learning is an area of computer science in which we develop algorithms that can effectively self-learn from the data provided. The primary aim is to let the computers learn for themselves without intervention from humans. Data Analytics and Machine learning go hand in hand. In this paper we have reviewed literature on some of the key machine learning techniques employed in the healthcare sector. This systematic review aims at determining the applications and challenges of Machine learning in health care.

Index Terms: Health care, Data analytics, Machine learning, Classification, Regression, Clustering, Deep Learning.

1. Introduction

The term machine learning was coined by Arthur Samuel in 1959, an American Computer scientist and a pioneer in the field of Artificial Intelligence and computer gaming. He defined machine learning as "Field of study that gives computers the ability to learn without being explicitly programmed". It involves developing or using algorithms that can learn and train themselves from experience. Machine learning has garnered a lot of attention in recent times and is finding applications in almost all walks of life. Machine Learning (ML) works well for problems where large volumes of data are available with a lot of underlying patterns that need to be identified and extracted. Based on the relationship that exists between the input data and the expected output, ML algorithms can be grouped as :

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

2. Learning types

2.1 . Supervised Learning

The input data also called as Training data or Training set has clearly distinct labeled values. In supervised learning the algorithm goes through a training phase. The training process is continuous and goes on till correct predictions are made. Supervised Learning algorithms can be further classified as Classification and Regression algorithms.

Classification – It is employed when the desired output variable is a class label

Regression – It is employed when the desired output variable is a continuous value.

2.2. Unsupervised Learning

If the input data is unlabeled then we go for unsupervised learning, here we will not know what sort of output will be got. The general approach followed is training through probabilistic data modeling.

2.3. Semi supervised Learning

Input data is a mixture of few labelled data and lots of unlabeled data. The model derived should learn to organize the data and do predictions. This class of algorithms are gaining huge popularity, as in real life data to be analyzed is usually a mixture of labeled and unlabeled variety.

2.4. Reinforcement Learning

The goal is to build an intelligent agent (RL -agent) which will work with the dynamic problem environment. The RL-agent learns using trial and error method just like human beings. Feedback to model is provided as reward for success and punishments for wrong decisions made in the problem environment. RL-agent will learn by revisiting the past actions it took for which it received rewards.

2.5. Steps Involved in Machine Learning

The process of developing a Machine Learning Algorithm can be depicted as a Six step process

1. Collect Data – Data available from multiple sources can be streamlined, also from the given data select attributes that will have an impact on the study being persued.

2. Preprocess Data – This involves 3 steps

- Formatting – Data must be in an industry standard format such as XML, CSV etc. so that it can be easily worked with.
- Cleaning – It involves removing noise and taking care

- *Sweety Bakyarani. E is currently pursuing Ph.d program in Computer Science in SRM Institute of Science and Technology, India, PH-918754539293. E-mail: sweetye@srmist.edu.in*
- *Dr.H. Srimathi is Professor of Computer Science program at SRM Institute of Science and Technology, India, PH-919940599928.*
- *E-mail: srimathh@ srmist.edu.in*
- *Dr.M. Bagavandas is Head, Centre for Statistics at SRM Institute of Science and Technology, India, PH-919150142330.*

of missing values.

- Sampling – To reduce redundancy data sampling of the data should be carried out at regular intervals.

3. Transform Data – To suit the algorithm used data should be transformed. It can involve decomposing features to extract information or it can also involve aggregation of multiple instance to one feature.

4. Train the Algorithm – Segregate the trainign data and test data from the processed data set.An algorithm learns or extracts knowledge from the training set and its output is stored as a model. Unsupervised learning does not have this Training step.

5. Test the Algorithm – The alogrithm trained in the previous step is given the test data set as input and the accuracy of the outcome is evaluated. If output is not satisfactory the previous step can be repeated.

6. Execute – After testing the algoritm the model generated is validated and can be put to use for realtime predictions, as the algorithm encounters more data, it will continue the learning porcess.

3. Supervised Learning

In supervised learning datasets are labeled. The data set is split into training set and test set. The training set is used by the algorithm to learn under which label the target falls into. It builds a prediction model that can be used in predicting labels for new data. The test set can be used to verify the correctness of the model and if the output is not satisfactory, the learning process continues. The labels of the data set can be categorical, such as ethnicity(O'Brien et al., 2018) where classification is used, or continuous, such as survival rates(Shipp et al., 2002) where regression is used. Feature selection is a subclass of supervised learning. Feature is nothing but information points in the data. Feature selection enables us to identify and select features that are most suited or most identifiable to the labels or remove improper and inaccurate data. In the following section we will discuss some of the common classification algorithms used in Machine learning.

3.1. Bayesian Classification

It is a statistical classification mechanism that helps us to avoid the probability of misclassification of data. Bayesian classification is based on Bayes Theorem. Given the probability of another event that has already occurred, It predicts the probability of an event occurring. Bayes' theorem can be mathematically expressed as follows

$$P(X|Y) = P(Y|X)\, P(X) / P(Y)$$

where X and Y are events.
We try to find the probability of occurrence of event X, given the event Y is true. Event Y is called as evidence.
P(X) is the prior probability of X.
P(X|Y) is a posteriori probability of Y.

3.2. Decision Trees

Decision trees are flow chart like structure where the node in the top is the root node, internal nodes represents a test on an attribute, the branches denote the outcome of the test and the terminal or leaf nodes denote the class labels. Decision trees are built by recursively partitioning the feature set based on information gain, gain ratio, Gini Index etc. Decision tree algorithms follow a Top Down, greedy approach for partitioning the feature space, till a desired maximized criterion is met. C4.5 and CART are commonly used decision tree algorithms

3.3. Random Forest

It is a simple and most popularly used algorithm that gives accurate results most of the time. It can be used for both classification and regression. The algorithm builds an ensemble of decision trees and merges them together.

3.4. Support Vector Machine

Also referred as SVM, it can be used for both classification and regression. In SVM we try to find a hyperplane in N-dimensional space that can be used to distinctly classify data points. (Narathip Reamaroon et al., 2019) Have used SVM to assist doctors in diagnosing patients who are susceptible to ARDS (Acute Respiratory Disorder). eDiag -a privacy preserving online diagnosis framework has been developed using nonlinear kernel support vector machine (SVM) (Hui Zhu et al., 2019).

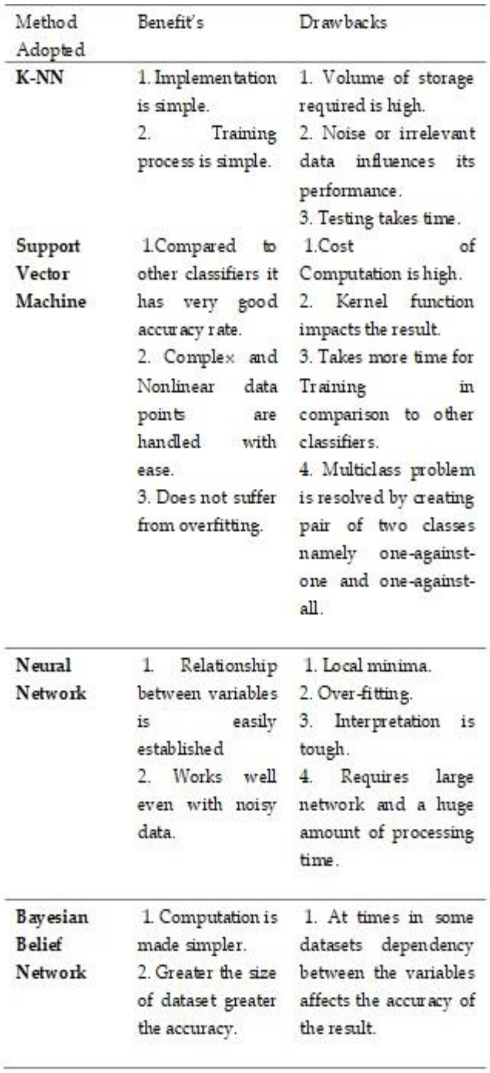
3.5. Artificial Neural Networks

Artificial neural networks or connectionist systems are inspired by the functioning of the central nervous system in human beings. It is a graphical model where the computing units are neurons. Neurons are interconnected and are organized in layers to pass information. The first layer receives raw data and it is called as input layer. The last layer performs the prediction and is called as the output layers. The intermediate layers are hidden layers. Feed forward neural networks (FFNN) have been applied to predict protein site-directed recombination (Bauer et al., 2006). A combination of embedding-based convolutional features and traditional features has been developed to be used with a softmax classifier to extract drug-drug interactions (DDIs) from biomedical literature (Zhao et al., 2016)

3.6. Regression

If the data to be classified is of continuous nature, then we use regression. It is a statistical approach that aims to find the relationship between variables by building equations. The parameters for the equation are obtained from the training set. The most popular regression models are linear regression and regularized linear regression. In regularized regression models the number of coefficients is constrained. Rigid Regression and LASSO are two very popular regularized regression models.

TABLE I *- SUMMARY OF FINDINGS*



Method Adopted	Benefit's	Drawbacks
K-NN	1. Implementation is simple. 2. Training process is simple.	1. Volume of storage required is high. 2. Noise or irrelevant data influences its performance. 3. Testing takes time.
Support Vector Machine	1.Compared to other classifiers it has very good accuracy rate. 2. Complex and Nonlinear data points are handled with ease. 3. Does not suffer from overfitting.	1.Cost of Computation is high. 2. Kernel function impacts the result. 3. Takes more time for Training in comparison to other classifiers. 4. Multiclass problem is resolved by creating pair of two classes namely one-against-one and one-against-all.
Neural Network	1. Relationship between variables is easily established 2. Works well even with noisy data.	1. Local minima. 2. Over-fitting. 3. Interpretation is tough. 4. Requires large network and a huge amount of processing time.
Bayesian Belief Network	1. Computation is made simpler. 2. Greater the size of dataset greater the accuracy.	1. At times in some datasets dependency between the variables affects the accuracy of the result.

4. UNSUPERVISED LEARNING

In unsupervised learning the dataset is unlabeled, and the model generated works with unlabeled data. Here we try to group the data together based on identifying some underlying similarity. Most of the data collected from internet or from automated process are unlabeled. Specially is the fields of life science huge volumes of unlabeled data is generated and extracting meaningful information from them has become a challenge. Here we examine some of the common clustering techniques employed to elicit insight.

4.1. Clustering

Clustering algorithms aim to group together data into some category based on some uniformity shared by the data or some underlying common features exhibited by the data. In exploratory data analysis clustering is used during the exploratory phase. Some of the import clustering techniques are k-mean clustering, hierarchical clustering and mixture models. Clustering takes a quantitative approach thus differentiating it from statistical dimension reduction methods like PCA – Principal Component Analysis and MDS – Multidimensional scaling which groups data quantitatively.

4.2 Hierarchical clustering

Hierarchical clustering creates clusters recursively by dividing the dataset in either a top-down or bottom-up approach. Hierarchical Clustering is of two types:

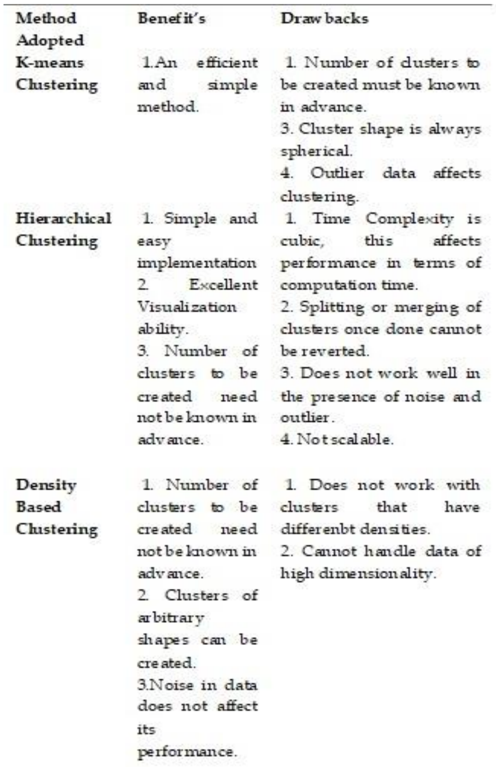
- Agglomerative clustering
- Divisive clustering

In agglomerative hierarchical clustering, each data object is initially treated as an individual cluster and later by applying Ward's method the clusters are merged till, we obtain the required number of clusters. Ward suggests the usage of any objective function that suits the researcher to be used as the criteria for selecting cluster pairs to be merged. In the second method that is in Divisive hierarchical clustering, all the data objects are grouped together as a single large cluster. This single cluster is further divided into smaller clusters till desired number of clusters is reached.

4.3. k-means Clustering

In k-mean clustering we begin by choosing k- random samples and use them as cluster centers also called as centroids. Successively as learning progresses the data will be moved between clusters. Distance between a data points and centroid is measured using techniques like Euclidian distance. Based on the distance a data point is assigned to a nearest cluster center. When every data point is allocated to a cluster center, we recalculate the weight of the cluster center. This process is repeated until a criterion set by the researcher is met. The criteria can be number of times cluster center has been recalculated or distance between the clusters. The success of k-means clustering lies in the value of k – that is initially set because it determines the number of clusters that will be formed. k-means clustering was to cluster individual by ethnicity based on their genomic profile (O'brien et al., 2015).

TABLE II *- SUMMARY OF FINDINGS*



Method Adopted	Benefit's	Draw backs
K-means Clustering	1.An efficient and simple method.	1. Number of dusters to be created must be known in advance. 3. Cluster shape is always spherical. 4. Outlier data affects clustering.
Hierarchical Clustering	1. Simple and easy implementation 2. Excellent Visualization ability. 3. Number of clusters to be created need not be known in advance.	1. Time Complexity is cubic, this affects performance in terms of computation time. 2. Splitting or merging of clusters once done cannot be reverted. 3. Does not work well in the presence of noise and outlier. 4. Not scalable.
Density Based Clustering	1. Number of clusters to be created need not be known in advance. 2. Clusters of arbitrary shapes can be created. 3.Noise in data does not affect its performance.	1. Does not work with clusters that have differenbt densities. 2. Cannot handle data of high dimensionality.

5. Deep Learning

In deep learning we take neural network to the next level through some effective learning algorithms. In a standard neural network (NN) there are many interconnected neurons. Once the input neurons are activated through external stimuli it starts activating the other neurons through already established weighted connection (Schmidhuber, 2015). By keeping many layers of neurons one above the other we can make the model "deep". But simply making a model deep does not guarantee accuracy of results produced. Only an effective training algorithm can make deep learning a success. Deep learning has revolutionized the way image recognition, speech recognition works. It has also hugely impacted biomedical field including genomics and drug discovery. Back Propagation has been successful in identifying complex structures in very large datasets. This algorithm in successful in guiding the NN to change its internal parameters in each layer based on what was learnt in the previous layers. One of the draw backs of deep learning is that it requires voluminous data, but this can be over come by applying transfer learning technique. Deep learning has found applications in biomedical imaging, genomics and signal processing in bioinformatics (Min et al., 2016). Splicing patterns in individual tissues and across tissues in mouse RNA sequence have been analyzed using deep s (Leung et al., 2014). Plis et al., 2014 have shown that using deep learning approach we can models that can learn physiological component analysis (ICA). Buggenthin et al. present a deep neural network that combines a CNN with an RNN architecture to automatically detect local image features and retrieve temporal information about the single-cell trajectories. Buggenthin et al., 2017 have successfully used the above mentioned approach to identify cells that have differently represented lineage-specific genes.

6. Conclusion

In this paper we have surveyed recent papers on machine learning approaches that have been adopted for solving issues in the health care industry. We have tried to give an overview of popular machine learning algorithms that are used to tackle issues in the field of healthcare. During the course of the review, we have identified that the method chosen for analysis is almost always dependent on the data set being analyzed and the outcome that is expected. We have also found that Big Data is ideally suited for machine learning. Traditionally machine learning algorithms learn iteratively and work best when the data is stored locally. But this problem can easily be solved by using distributed computing platforms like Apache Spark and Apache Hadoop. In conclusion machine learning is a field that is rapidly touching new heights specially in the field of health care. Huge breakthroughs and new findings that have the capability to alter the way health care industry functions are imminent.

7. REFERENCES

[1.] Abu-Jamous, B., Fa, R., Nandi, A.K., 2015a. Feature Selection. Integrative Cluster Analysis in Bioinformatics. John Wiley & Sons, Ltd.

[2.] Abu-Jamous, B., Fa, R., Nandi, A.K., 2015b. Mixture Model Clustering. Integrative Cluster Analysis in Bioinformatics. John Wiley & Sons, Ltd.

[3.] Akbani, R., Kwek, S., Japkowicz, N., 2004. Applying support vector machines to imbalanced datasets. In: Proceedings of the Machine Learning, vol. 3201, pp. 39–50, ECML

[4.] Algama, M., Tasker, E., Williams, C., et al., 2017. Genome-wide identification of conserved intronic non-coding sequences using a Bayesian segmentation approach. BMC Genomics 18.

[5.] Amari, S., Wu, S., 1999. Improving support vector machine classifiers by modifying kernel functions. Neural Networks 12, 783–789.

[6.] Bauer, D.C., Boden, M., Thier, R., Gillam, E.M., 2006. STAR: Predicting recombination sites from amino acid sequence. BMC Bioinformatics 7.

[7.] Bauer, D.C., Willadsen, K., Buske, F.A., et al., 2011. Sorting the nuclear proteome. Bioinformatics 27, I7–I14.

[8.] Bauer, D.C., Buske, F.A., Bailey, T.L., Boden, M., 2010. Predicting SUMOylation sites in developmental transcription factors of Drosophila melanogaster. Neurocomputing 73, 2300–2307.

[9.] Beck, D., Foster, J.A., 2014. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLOS ONE 9.

[10.] Blunsom, P., 2004. Hidden markov models. Lecture Notes 15, 18–19. Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152, ACM.

[11.] Boyle, E.A., Li, Y.I., Pritchard, J.K., 2017. An expanded view of complex traits: From polygenic to omnigenic. Cell 169, 1177–1186.

[12.]Bradley, A.P., 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition 30, 1145–1159.

[13.]Leng, N., Li, Y., Mcintosh, B.E., et al., 2015. EBSeq-HMM: A Bayesian approach for identifying gene-expression changes in ordered RNA-seq experiments. Bioinformatics, 31. pp. 2614–2622.

[14.]Leung, M.K.K., Delong, A., Alipanahi, B., Frey, B.J., 2016. Machine learning in genomic medicine: A review of computational problems and data sets. Proceedings of the IEEE 104, 176–197.

[15.] Leung, M.K.K., Xiong, H.Y., Lee, L.J., Frey, B.J., 2014. Deep learning of the tissue-regulated splicing code. Bioinformatics 30, 121–129.

[16.]A. Abbasi and M. Younis. A survey on clustering algorithms

[17.]for wireless sensor networks. Computer communications,

[18.]30(14):2826–2841, 2007.

[19.]Liang, K.C., Wang, X.D., Anastassiou, D., 2007. Bayesian basecalling for DNA sequence analysis using hidden Markov models. IEEE-ACM Transactions on Computational Biology and Bioinformatics 4, 430–440.

[20.] Libbrecht, M.W., Noble, W.S., 2015. Machine learning applications in genetics and genomics. Nature Reviews Genetics 16, 321–332.

[21.] Loewenstein, Y., Portugaly, E., Fromer, M., Linial, M., 2008. Efficient algorithms for accurate hierarchical clustering of huge datasets: Tackling the entire protein space. Bioinformatics 24, I41–I49.

[22.]Loh, W.Y., 2011. Classification and regression trees.

Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery 1, 14–23.

[23.] Lottaz, C., Iseli, C., Jongeneel, C.V., Bucher, P., 2003. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics 19, Ii103–Ii112.

[24.] Maglogiannis, I.G., 2007. Emerging Artificial Intelligence Applications In Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Ios Press.

[25.] A. K. Jain and R. C. Dubes. Algorithms for clustering data. Prentice-Hall, Inc., 1988.