International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-4, April 2019

Disease Prediction using Big Data Analytics and SVM

V. Sahaya Sakila, S. Sri Gayathri

Abstract: ***The biomedical communities have been facing a rapid big data growth. Medical data can be effectively used in disease detection, treatment and cure. There is a need for proper and effective methods in order to analyse the data accurately. The frequency and the common kinds of diseases may vary from region to region but the characteristics of the diseases exhibited by different regions can be different, thus, making the prediction of disease outbreaks difficult. Here, machine learning algorithms are streamlined for effective prediction of diseases. The algorithm to be used is Support Vector Machine, also known as SVM. Support vector machines are models that come under supervised machine learning, and generally consist of algorithms that are used to analyse data for regression analysis after classification. That is, taking a set of training examples, each of it is segregated based on the category it belongs to among the two and a model is built by an SVM training algorithm that would assign examples to either of the two categories.***

Index Terms: Data mining, Support Vector Machine, Medical communities, datasets.

I. INTRODUCTION

Every year, all over the world, people are dying in millions because of either ignorance of their ailment or because of leaving a chronic disease which has been treated unsupervised. This is because most deadly diseases have only subtle chronic symptoms which can be easily discarded as minor inconvenience. To tackle this problem scientists and doctors have come up with wearable monitoring systems, which collect various information about the person who wears them around the clock and stores it. This data is immensely huge, and finding coherent variables to link and deduce any particular disease is a herculean task. Extraction of patterns or information manually from data has been occurring for ages. Some of the methods made use of in the early stages include Bayes' theorem and regression analysis. It is evident that the increasing power of technology has increased the ability to collect, store, and manipulate data dramatically. Since datasets keep growing in size and complexity, direct analysis of data has been replaced with the indirect method of automated processing, supported by other discoveries in computer science. To mention a few, cluster analysis, neural networks, genetic algorithms. Data mining is the process of applying these methods in order to uncover patterns in large data sets. It acts as a bridge to fill the gap between AI and applied statistics to database management by making full use of the way the data is stored to execute the actual discovery of algorithms with more efficiency. Data mining refers to the practice where large datasets are examined to generate new information by discovering patterns. It involves methods at the intersection of statistics and machine learning. The main goal of data mining is to extract information from datasets for further use. The "knowledge discovery in databases" is a process that consists of phases such as selection, pre-processing, transforming, mining and interpretation. CRISP-DM refers to The Cross Industry Standard Process for Data Mining. It consists of six stages which are data understanding, business understanding, data preparation, modelling, interpretation and deployment.

There currently exists an opportunity for healthcare systems for defining what "predictive analytics" means to them and how its use can be effective in order to make improvements. Machine learning is a discipline which has a long history of success in many fields. The usage of machine learning concept and algorithms in the healthcare sector will only prove to be more beneficial. Predictive analytics and machine learning are two disciplines that are rapidly becoming some of the most-discussed topics in the field of healthcare analytics.

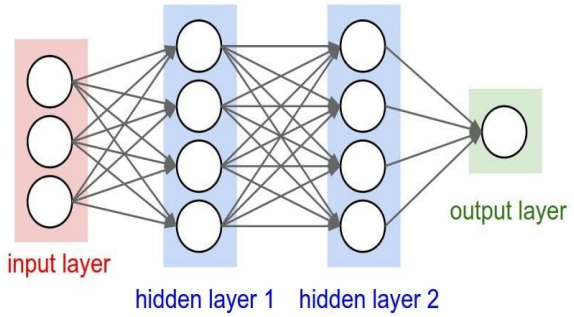
II. EXISTING SOLUTION

The pre-existing solution for disease prediction using data mining applies a different machine learning algorithm for the predictive analysis. The used algorithm is neural network.

A. Neural Networks

An artificial neural network is basically a group of interconnected nodes, that is similar to the network of neurons in a brain. A neural network is a framework for various machine learning algorithms in order for them to work together to process data inputs that are complex in nature. The individual nodes known as artificial neurons model the neurons of a brain and the connections, that act like the synapses of a brain are capable of transmitting signals from one node to another. There is an input layer and an output layer in a neural network model. Other than these two, there's a certain number of hidden layers that perform individual functions. For instance, when it comes to image recognition, one layer can identify the edges of the image, another might recognize texture, etc.

Manuscript published on 30 April 2019.
* Correspondence Author (s)
Ms. V. Sahaya Sakila*, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.
S. Sri Gayathri, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India.



If x_1, x_2… and y_1 , y_2… and w_1 , w_2… are vectors in R^m, R^n, R^e respectively, first the weight w_i is computed as:

$$(x_i, y_i, w_{i-1})$$

The output gives a new function:

$$x \mapsto f_N(w_p, x)$$

There are certain major disadvantages of the existing system. Neural networks require a large dataset. This can require more training time. Another disadvantage is that when the features are heterogenous, the weights and updates will not be on the same scale. Thus, the inputs need to be standardized in some way. Neural networks are expensive for productional uses.

Due to these reasons, a different algorithm is preferred in order for the system to be more effective and yield better and more accurate results.

III. LITERATURE SURVEY

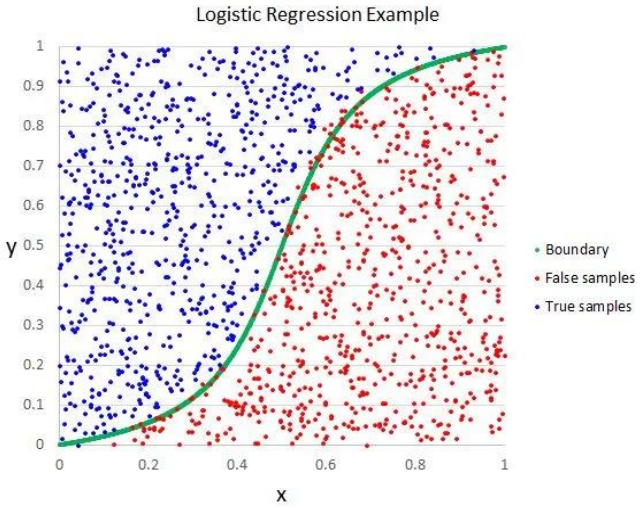
Since data mining is a study that's proving to be necessary in almost every field, several attempts have been made in the healthcare community to use the datasets from the databases effectively for the purpose of improvement in healthcare. Some methods and ideas pre-existing in this field and discipline are as follows:

A. Decision Tree

Decision tree is a model used in predictive analysis. It is a tool that has a tree structure similar to a flowchart. Each one of the internal nodes denotes a test on a particular attribute, a branch constitutes the outcome of one particular test, and the terminal nodes or the leaf nodes hold class labels. Classification of an instance is done by starting at the tree's root node, then the attribute specified by that node is tested, moving along the branches downwards corresponding to the attribute's value. This process is then carried out again for the subtree that contains the new node as the root node. Decision trees are capable of classifying without requiring much computation. But the weaknesses of decision trees include less appropriation when it comes to estimation tasks in which the main goal would be to predict the value of an attribute. Decision trees are expensive to train. Splitting field of each candidate at each node has to be sorted before the best split for it can be identified. When it comes to classification problems that have many classes and relatively small number of training examples, these trees are prone to errors. Since many sub-trees need to be established and compared, pruning algorithms are also expensive. Due to these disadvantages of decision trees, the need for a better algorithm arises.

B. Logistic Regression

It is a method under predictive analysis. Logistic regression is a statistical model for binary dependent variables. As in a binary logistic model, a variable that is said to be dependent can have two values '0' or '1', in this model, a variable with a value of '1' is said to be a linear combination of independent variables. An independent variable might be binary or continuous. If a variable can take on two particular real values such that it can also include all of the real values between them, the variable is said to be continuous in that particular interval. Logistic regression is known to measure the relationship between the dependent variable and independent variables. This is done by estimating probabilities. A logistic function is used for that purpose. Thus, it treats the particular set of problems as probit regression, which uses a cumulative normal distribution curve. In these two methods, when it comes to the latent variable interpretations, the first method assumes a logistic distribution of errors whereas the second assumes a normal distribution of errors. The limitations of logistic regression include limited outcome variables, the requirement of independent observations, overfitting of the models. Also, logistic regression mainly produces probabilistic values and does not make an absolute prediction. It doesn't rely completely on the data to give a final decision. This may be good for estimation purposes but there isn't enough confidence in the data.



Considering a model that has two predictors x_1 and x_2, where these may be continuous variables or indicator functions of binary variables, the representation of log odds is of the form:

$$l = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

Where, the log odds are represented using 'l'.
The logit of the probability, which is the logarithm of the probability is defined as:

$$\operatorname{logit} p = \ln \frac{p}{1-p} \quad \text{for } 0 < p < 1.$$

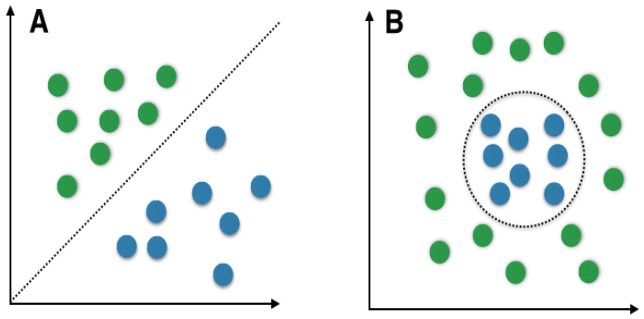
The logit function in a generalized model is given as:

$$\operatorname{logit} \mathrm{E}(Y) = \alpha + \beta x$$

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

C. Naive Bayes Classifier

Naive Bayes is a technique to build classifiers or models that allocates labels to instances of the problem, represented using vectors of the feature values. The labels are obtained from some finite set. This method also requires a family of algorithms for training the classifiers. A Naive Bayes Classifier expects each feature to contribute independently to the probability of what it is, without consideration for any correlations that may be possible among the features. It makes strong assumptions on the shape of the data distribution, which means that the features are not dependent given the output class. Another problem with Naive Bayes classifier is data scarcity. It requires an estimation of likelihood by frequentist approach for any possible value of a particular feature. Another disadvantage is that when it comes to continuous features, a lot of necessary information might be thrown out by the binning process as the binning method commonly used is to make them discrete. Due to reasons such as these the Naive Bayes classifiers are not always reliable in all cases.



The conditional probability, according to the Bayes' theorem is expressed as:

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

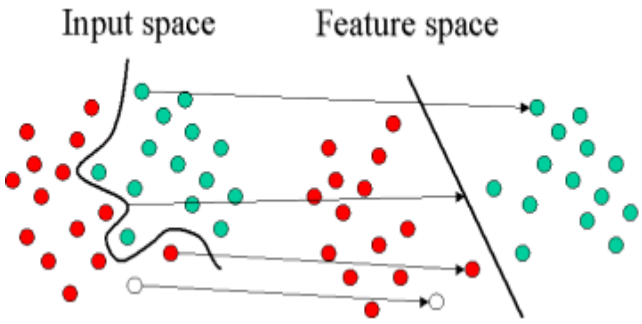
A Bayes classifier is a function which allocates a label $\hat{y} = C_k$ as follows:

$$\hat{y} = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

Due to the number of disadvantages in the above-mentioned methods, a better algorithm is required for the purpose of disease prediction which hasn't been used. One of the most reliable machine learning algorithms is what is expected to be trained with for this model.

IV. PROPOSED SOLUTION

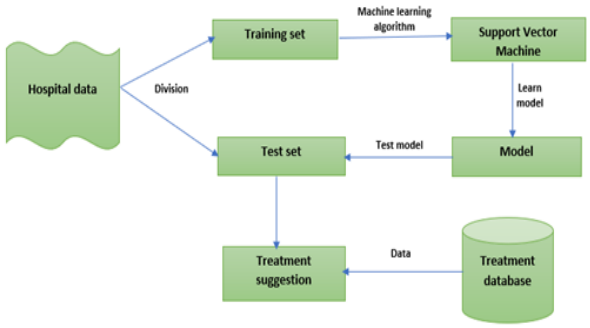
The major reason to develop this system is to make the analysis of data collected from hospitals more accurate than it already is. The proposed system is better than all the existing ones since the algorithm preferred here is SVM or Support Vector Machine. Support Vector Machine is considered to be one of the highly efficient algorithms because it gives high accuracy. It is an extended version of linear regression. SVM has many applications in computer vision.



The system proposed predicts diseases based on the data sets obtained from hospital medical records. Some of the data are the electronic health records of the patients, image data and gene data. Disease prediction is mainly analysing data and checking if the patient is amongst the high-risk population of some or the other disease based on their medical history. The Support Vector Machine is responsible for classifying the disease into one of the two classes, based on the analysis if it's malign or benign. A way to make the system more effective is to give a suggestion as to what has to be the solution for the particular disease. This is done with the help of a dataset containing the treatments for all possible diseases. This database would provide aid for the patient by deciding the proper treatment for the solution. Among the three types of machine learning, the most common and appropriate one for big data analytics is supervised machine learning. Supervised machine learning is responsible for classification and regression, which are the key methods in data analytics. Therefore, the same type has been applied here as well. Supervised machine learning is a type that depends entirely on the dataset provided. The model is trained based on a machine learning algorithm for it to do the analysis and hence the prediction.

V. SYSTEM ARCHITECTURE

The system architecture depicts the flow of control. There isn't a requirement for special hardware devices for this system since it is regarding analysis of the data. The system architecture is as follows:



The data from the hospitals in a particular locality has been collected for analysis. This data is divided randomly into the training set and the test set. The training set is used for the purpose of training the model that would perform the analysis on the data. The test set is where the data is actually analysed.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

Once the testing model has been developed and trained, machine learning algorithms are applied for analysis. In this case, Support Vector Machine is used. So once the testing has been done and the identification of common and frequent diseases has been done, the next step is performed which is called the treatment suggestion. There exists a database consisting of the appropriate treatments or solutions for all sorts of diseases. This database would be used to give a suggestion for the diseases that are common or frequent in that specific region. Thus, effective disease prediction and cure can be done.

VI. MODULE IDENTIFICATION

The working of this approach is carried out with the help of certain modules which are as follows:

A. Data Set

The methods under supervised learning attempt to study the relationship among the input attributes and the target attributes. The input attributes can be referred to as independent attributes and the target attributes are known as the dependent attributes. Thus, the input attributes or the dataset is highly essential when it comes to supervised machine learning. Here, the data set would be the information collected from hospitals in a specific area. The information would mostly include the number of patients and their individual information from the patients' medical records such as their age, the symptoms etc. This form of data are of two categories. One is the structured data which includes all forms of definite information regarding the patient and the other type is the unstructured data which is mainly the details that may not be accurate or true. For instance, the structured data might be the patient's personal details like his age, blood group, etc., whereas the unstructured data would include the patients' narration of the problem faced by them. Thus, the data set can be considered as the fundamental need in supervised type of machine learning. It is applicable here since machine learning algorithms play a crucial role in data analytics.

Data category	Item	Description
Structured data	Demographics of the patient	Patient's gender, age, height, weight, etc.
	Living habits	Whether the patient smokes, has a genetic history, etc.
	Examination items and results	Includes 682 items, such as blood, etc.
	Diseases	Patient's disease, such as cerebral infarction, etc.
Unstructured text data	Patient's readme illness	Patient's readme illness and medical history
	Doctor's records	Doctor's interrogation records

B. Coherence of Data

Coherence in general refers to the quality of being whole, that is the quality of being consistent and logical. It is essential in data mining since suitable patterns are to be obtained from the data in order to analyse it. For instance, for predicting a patient's disease, all of the required factors are to be tested for so that the accuracy of the prediction is high. The qualities of different diseases may overlap but the differentiation has to be done based on the extent or the intensity of the individual factors. The prediction system cannot give the wrong outcome for a particular case. In order to avoid such faults, coherence of data is essential. When it comes to dealing with personal probability assessments, coherence is a property of self-consistency in the whole set of such assessments. It is required so that consistent decisions can be obtained from all the probabilities.

C. Model Training

This is a step that's as crucial as obtaining the data set. Data analytics is all about studying data and making predictions without human effort. Thus, in order for this to take place without human interference, a model has to be trained so that it could use the dataset to generate the required information accordingly. This is where machine learning comes into play. Machine learning algorithms are required in big data analytics to train the system according to the need. The algorithm used here is Support Vector Machine. Support Vector Machines are models that come under supervised machine learning. These are used for classification and regression analysis. Since these two are the most commonly used methods in data analysis, this algorithm is highly applicable in this case. Also, Support Vector Machine has a lot of advantages compared to the other machine learning algorithms. It is one of the most reliable ones and gives an output with high accuracy. This algorithm builds a model that categories each piece of data into one of the two categories. Thus, it is a non-probabilistic binary classifier. This model represents all the examples in space as points, so that they can be sorted out into the suitable categories. SVM is capable of performing not only linear classification but non-linear as well, by using what is known as a kernel trick, in which mapping of the inputs into high dimensional feature spaces takes place.

D. Disease Detection

Disease detection is done with the help of the information obtained or derived. This step checks if the disease predicted is malign or benign based on the values of the factors responsible for the prediction, that is, by checking if it's high or low or moderate. This is essential since one disease should not be mistaken for another due to errors or inaccuracies. The coherence of the data obtained plays an important role here. That is because the pattern obtained should match with the pattern of a particular disease among all of them. The outcome would prove to be ineffective and sometimes lethal if there are faults while making the prediction because there is also a module that suggests the appropriate treatment for the specific disease and the whole procedure might go wrong due to inaccuracies. Therefore, disease detection is quite crucial.

E. Cure Suggestion

Treatment suggestion is a step where, as the name suggests, the appropriate treatments for the diseases are suggested. There is a database that contains all the techniques for cure, and when the disease is fed to it, it chooses the solution and provides it to the patient. This step is quite simple compared to the rest and doesn't involve much complications since it just has to give the information from the database to the system.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)

VII. CONCLUSION

A data mining model for the purpose of disease prediction and treatment suggestion is provided. The machine learning algorithm known as Support Vector Machine has been used for training the model. It can produce accurate and reliable results since Support Vector Machine model is a very popular and important algorithm in the discipline of machine learning. Here, treatment is also being suggested with the help of a database consisting of all the possible treatments for all kinds of diseases. The system proposed here can be further developed by applying more efficient algorithms. The healthcare community still has room for development and improvement which can be provided by researching on data mining concepts. Since data mining is being introduced in almost every field, more developments in this field would make lives easier.

ACKNOWLEDGMENT

This project was supported by Prof. Ms. V. Sahaya Sakila, thus I would like to express special thanks of gratitude to her, for providing me with the opportunity to work on this project. During the research on this project, many new terms came to be known for which I am thankful.

REFERENCES

1. Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L., 2017. Disease prediction by machine learning over big data from healthcare communities. Ieee Access, 5, pp.8869-8879.
2. Srinivas, K., Rani, B.K. and Govrdhan, A., 2010. Applications of data mining techniques in healthcare and prediction of heart attacks. International Journal on Computer Science and Engineering (IJCSE), 2(02), pp.250-255.
3. Delen, D., Walker, G. and Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2), pp.113-127.
4. Durairaj, M. and Ranjani, V., 2013. Data mining applications in healthcare sector: a study. International journal of scientific & technology research, 2(10), pp.29-35.
5. Srinivas, K., Rao, G.R. and Govardhan, A., 2010, August. Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In Computer Science and Education (ICCSE), 2010 5th International Conference on (pp. 1344-1349). IEEE.
6. Xing, Y., Wang, J. and Zhao, Z., 2007, November. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on(pp. 868-872). IEEE.
7. Cheng, T.H., Wei, C.P. and Tseng, V.S., 2006, June. Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches. In Computer-Based Medical Systems, 2006. CBMS 2006. 19th IEEE International Symposium on (pp. 165-170). IEEE.
8. 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, 17(8), pp.43-48.
9. Burke, H.B., Goodman, P.H., Rosen, D.B., Henson, D.E., Weinstein, J.N., Harrell Jr, F.E., Marks, J.R., Winchester, D.P. and Bostwick, D.G., 1997. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer, 79(4), pp.857-862.
10. Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi. "The survey of data mining applications and feature scope." arXiv preprint arXiv:1211.5723 (2012).

AUTHORS PROFILE

Ms. V. Sahaya Sakila, M.E., Assistant Professor (O.G) at Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai

S. Sri Gayathri, Student, 3rd year, B. Tech, Department of Computer Science at SRM Institute of Science and Technology, Chennai. Previously published research papers in the domain of machine learning.