International Journal of Scientific Research in Computer Science, Engineering and Information Technology

© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952154

Generalized Disease Prediction based on Symptoms

Ramandeep Singh Sethi*, Aniket Thumar, Vaibhav Jain, Sachin Chavan

Department of Computer Science and Engineering, NMIMS, Shirpur, Maharashtra, India

ABSTRACT

We are right now facing a daily reality where mobile utilization is developing exponentially. Mobile technology is omnipresent. It offers services that is customized to us – the 21st century user. Innovation has empowered us incredibly, we look for data anyplace and anytime. Digital health is acquainting new methodologies with the administration of health conditions. Research has exhibited noteworthy development in the effect that digital health is having on patients and overall healthcare. The selection of digital health tools, such as mobile healthcare apps, holds incredible guarantee with proof of these tools playing a positive role in both patient results and the expenses. Portable applications can enable patients to be effectively associated with each phase of their healthcare venture. This fundamentally enhances patient commitment and the patient experience, and urges purchasers to be responsible for their own health. Portable apps can tailor health content as indicated by the patients, or healthcare providers, mobile history and current conduct. These customized mobile experiences help convey highly pertinent information at the right time, based on user priority.

Keywords : Data Mining, Classification, Clustering

I. INTRODUCTION

The disease prediction systems available in the market currently have decent accuracy but they are not available to everyone. The systems which are available publicly don't provide personalized treatment and remedies. Also, these systems don't take BMI (Body Mass Index) and drugs that patient is taking currently into consideration during the prediction process. This affects their accuracy significantly. Data mining is a pattern discovery technique that is used to find the concealed qualities from huge measure of information. As the patient's populace and medications increases, the restorative databases also grows day by day. The examination of these therapeutic data is intricate without the PC-based analysis architecture. The PC-based analysis architecture provides the robotized medical determination system. This robotized determination system supports the medical expert to make systematic decision in therapy and ailment forecast. Data mining is the quickly developing area for the doctors to deal with huge amount of patient's data sets from multiple point of view such as understanding of complex symptomatic tests, interpreting past outcomes and accumulating the different information together. Customarily hospital's conclusion is molded by the medical expert's inspection and predicting the result rather than the inference obtained from the huge data. This robotized determination system leads to increase the service's standards and reduces the medical cost.

II. RELATED WORK

Darcy A. Davis Used ICD9-CM to predict future disease risks. They used clustering to predict the disease based on similar patient's medical history [1].

T.F. Michael Raj and S. Prasanna proposed the model that trains the machine and it proves the probabilistic models are stable and reliable to identify the disease [2]. K.Rajalakshmi, Dr.S.S.Dhenakaran, N.Roobini pre-processed data collected from different sources was given as input to the different clustering methods. When K-Means algorithm was applied to preprocessed data, it showed low accuracy [3]. But when it was used along with different classifiers, it showed decent accuracy [4].Shalet K.S, V.Sabarinathan, V.Sugumaran, V.J. Sarath Kumar proposed a model where REPTree was used for the process of feature selection. This selection helped us get structural information which can be analyzed easily. Feature classification is done using SVM (support vector machine) [5]. K.M. Al-Aidaroos, A.A. Bakar and Z. Othman have conducted the research for the best medical diagnosis mining technique. These authors compared Naïve Bayes with five other classifiers i.e. Logistic Regression (LR), K-Star (K*), Decision Tree (DT), Neural Network (NN) and a simple rule-based algorithm (ZeroR) [6]. For this, 15 real-world medical problems from the UCI machine learning repository (Asuncion and Newman, 2007) were selected for evaluating the performance of all algorithms. In the experiment it was found that NB outperforms the other algorithms in 8 out of 15 data sets [7]. So, it was concluded that the predictive accuracy results in Naïve Bayes is better than other techniques [8].

III. PROPOSED SOLUTION

Our proposed system aims at predicting the disease of a person based on his/her symptoms. The system will ask simple and relevant questions and understand what's wrong with the patient's health. In every assessment, our system will take all of a patient's information into consideration, including past medical history, symptoms, risk factors and more. The system will act as a prescreen consultation before being handed off to a real doctor for further advice. This will save significant time during any follow up consultation. Our system will be available to everyone in the form of a mobile application so that everyone can access our healthcare service. Number of questions that we will ask a patient will be less than what is asked by present systems. Our system will take a patient's past medical history and drugs that he is currently taking into consideration and then provide a personalized treatment/remedy for him/her. Based on the disease detected, our system will also recommend nearby hospitals where this disease can be treated.

IV. DATA MINING

Data mining is a process in which different data sources are accumulated and some hidden patterns are discovered from that data. The following diagram depicts different phases involved in process of data mining. The data mining procedure uses various methods. Clustering and classification techniques used in data mining are discussed in this paper.

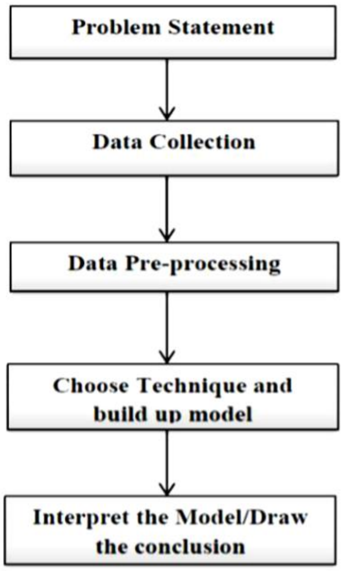
Problem Statement
Data Collection
Data Pre-processing
Choose Technique and build up model
Interpret the Model/Draw the conclusion

Fig.1 Data mining process

V. CLUSTERING

Clustering is a machine learning method. It is an unsupervised algorithm which has no appropriate predefined groups. Clustering is the method of grouping set of data points into relevant sub-groups called clusters. The phases of clustering procedure can be depicted as:

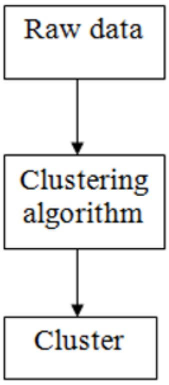

Fig.2 Stages of clustering

VI. CLASSIFICATION

Classification is a supervised learning method by which we find to what category a certain data lie. In classification, training sample is provided. This method classifies the data into one of the predefined classes by using class labels. In order to find the correct category for a given data, machine learning technology does the following:

- Applies a classification algorithm to recognize features of different classes.
- Compares those features to the data we are seeking to classify.
- Uses that knowledge to evaluate how likely it is that data fits to a specific class.

VII. K-MEANS ALGORITHM

K-means clustering is a clustering technique based on partition. It is an unsupervised learning algorithm that is used when we have unlabeled data. In K-means clustering algorithm, data points are split into fixed number of clusters denoted by variable K. This algorithm consists of two steps which are as follows:

1. Select the number of clusters i.e. K value and choose K centroid values in arbitary way.
2. Allocate data points to nearest centroid

i) Method:

The procedure of K-Means algorithm is as follows:

Input: Number of desired clusters K and a database D= {p1, p2… pn} containing ‘n’ data points.
Output: A set of K clusters.

ii) Steps:

1. Fix the centroid of K clusters
2. Evaluate the distance of each data point from each centroid.
3. Set the new centroid position of each cluster to mean of all data points lying in that cluster
4. Repeat step no. 2 and 3 until no data points change the clusters

By use of iterative method, K-means algorithm reduces the sum of distances from each data point to its cluster’s centroid. These algorithm works until the sum of distance from respective cluster cannot be minimized beyond some specific value.

VII. NAÏVE BAYES CLASSIFIER

It supposes that each feature depends only on the class. So, this is supposed to mean that each feature has only the class as parent. In some real-world problems, there should be a hindrance in which the independency hypotheses of features with respect to class are disrupted. It is simple, rapid to implement with the easy structure, and efficient. Probability Each feature is approximated individually such that it is useful for high spatial data. Let C denote the class of an observation X. To predict the class of the observation X by using the Bayes rule, the highest posterior probability of

$$P(C|\mathbf{X}) = \frac{P(C)P(\mathbf{X}|C)}{P(\mathbf{X})}$$

should be found.

In the NB classifier, using the assumption that features X1, X2,...,Xn are conditionally independent of each other given the class, we
get

$$P(C|\mathbf{X}) = \frac{P(C)\prod_{i=1}^{n} P(X_i|C)}{P(\mathbf{X})}. \qquad \text{IX.}$$

CONCLUSION

The disease prediction systems available in the market currently have decent accuracy but they are not available to everyone. The systems which are available publicly don't provide personalized treatment and remedies. Also, these systems don't take BMI (Body Mass Index) and drugs that patient is taking currently into consideration during the prediction process. This affects their accuracy significantly. Our proposed system will take a patient's past medical history and drugs that he is currently taking into consideration and then provide a personalized treatment/remedy for him/her. Based on the disease detected, our system will also recommend nearby hospitals where this disease can be treated.

VIII. REFERENCES

[1]. Davis, D., V. Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008) "Predicting Individual Disease Risk Based on Medical History"

[2]. T.F. Michael Raj and S. Prasanna "Implementation of ML Using Naīve Bayes Algorithm for Identifying Disease-Treatment Relation in Bio-Science Text"

[3]. K.Rajalakshmi, Dr.S.S.Dhenakaran and N.Roobini "Comparative Analysis of K-Means Algorithm in Disease Prediction"

[4]. Purvashi Mahajan, Abhishek Sharma "Role of K-Means Algorithm in Disease Prediction"

[5]. Shalet K.S, V.Sabarinathan, V.Sugumaran, V.J. Sarath Kumar "Diagnosis of Heart Disease Using Decision Tree and SVM Classifier"

[6]. K.M. Al-Aidaroos, A. B. (n.d.). 2012. "Medical Data Classification with Naive Bayes Approach"

[7]. Kazmierska J, Malicki J. "Application of the Naïve Bayesian Classifier to optimize treatment decisions"

[8]. Adam, S., & Parveen, A. (2012) "Prediction System for Heart Disease Using Naive Bayes"

Cite this article as :

Ramandeep Singh Sethi, Aniket Thumar, Vaibhav Jain, Sachin Chavan, "Generalized Disease Prediction based on Symptoms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 543-546, March-April 2019. Available at doi : https://doi.org/10.32628/CSEIT1952154
Journal URL : http://ijsrcseit.com/CSEIT1952154