REVIEW OF RESEARCH
ISSN: 2249-894X
IMPACT FACTOR : 5.7631(UIF)
VOLUME - 10 | ISSUE - 12 | SEPTEMBER - 2021

PREDICTION OF DISEASES USING DATA MINING TECHNIQUE

Prashant Devidas Jadhav
Research Scholar

ABSTRACT

The transformation of information technology has created many applications in healthcare information. Healthcare Informatics is generating large amounts of data. These data can be processed to predict diseases using hybrid data mining techniques. Data mining is the process of analyzing data, extracting data, and presenting data as knowledge that builds relationships between available data. Data mining techniques include association, clustering, sorting and others. These techniques are applied to analyze the data and to achieve maximum accuracy on the data set obtained. Significant advances in information technology have led to the proliferation of data in healthcare information. Healthcare informatics data includes hospital details, patient details, disease details and treatment costs. This huge data is created from a variety of sources and formats. It may contain irrelevant attributes and missing data. Data mining involves various methods to extract knowledge from huge disease data sets. Data mining techniques such as classification, bunch and rules will be accustomed to analyze mining information and extract non-vegetarian information.*

KEYWORDS: *transformation of information technology, extracting data, process of analyzing data.*

INTRODUCTION

In today's fast-paced world, the risk of disease is much higher, whether in developed or developing countries. Due to the tremendous advances in technology, diode medical information systems in hospitals and medical establishments are getting bigger and bigger and the method of extracting useful information is becoming extra and difficult and time consuming. Currently the establishment area unit's one-day extra range has become expensive and inefficient to calculate on manual information analysis, from keeping the patient's primary based records to computerized patient records. We currently need advances in technology that is not only affordable but also very easy to use so we need laptops based primarily on analysis that can effectively diagnose patient defects.

Data mining and information retrieval is a method of extracting information from a vast body of knowledge and is not limited to completely different classification functions as well as decision making, fault finding, pattern identification, prediction. And image processing. The purpose of extracting information from information is to create models from information to predict long-term behavior. Prior to data mining, CPR was known as electronic patient records and was simply a centralized data archive that provided extremely limited possibilities for analyzing and processing data and more data was only able to diagnose simple cases. Data mining is the key to evaluating, interpreting, and processing large amounts of data and processing queries with great accuracy and increasing CRM value.

Data mining uses algorithms and tools to transform the life cycle of knowledge and uses formalities to extract patterns, information and knowledge extracted from data stored in CPR. Thus, we can say that Data Mining is useful for converting transaction data in CPR from horse intelligence to more useful and efficient explicit knowledge.

Data Mining is not only knowledge creation but also a set of techniques for data analysis; It is the key to extracting information from huge data sets. Without data processing, knowledge as cardiac resuscitation is not necessary as it does not make a difference in recognition. Data processing is the aggregation of heterogeneous tools and techniques to perform completely different tasks on the data creation method. They use every descriptive and prophetic model. The descriptive model facilitates specific similar patterns in the knowledge analyzed by the victim classification, association rules, and visual image on the opposite hand prophetic model, using classification, regression, and statistical analysis to show the effect of treatment on the patient. We can summarize how past model distinguishes data mining techniques using building and clustering techniques.

Due to the lack of medical services and specialist physicians in most developing countries, the death rate is quietly high. Late recognition of the disease with inappropriate treatment support, and deficiency of number and quality of medical specialists in developing countries, contribute to high death rate. The current approach of the healthcare model is to provide preventive measures instead of waiting for treatment once the disease is detected. Medical experts use their own hypotheses to rule out the possibility of any disease. Their assumptions depend on factors such as daily life, medical history or demographics. With the help of data mining and machine learning, it is possible to predict the disease with more accurate results. But recent advances in advanced algorithmic techniques have led to the development of robust models for disease detection. Current medical practice in developing countries like India involves steps such as visiting patients, waiting and consulting a doctor for further assistance in treatment and diagnosis.

Many doctors may not have much experience in dealing with high risk diseases like cancer. However, the waiting period for treatment is usually a few days, weeks or even months. In most cases, it is a waste of time to make an appointment, then consult, and then further investigation. During this time many high-risk diseases have already spread and there is no chance of survival. Therefore, late detection of the disease leads to unexpected death. Since most high-risk diseases can be cured only at an early stage, patients have to endure a lifetime. The modern approach to healthcare is to take preventive measures instead of seeking treatment after diagnosis. A computer program or software developed by imitating human intelligence can be used to help doctors make decisions without consulting a specialist directly. This software was not intended to replace a specialist or a doctor, however it was developed to help general practitioners and specialists diagnose and predict a patient's condition based on certain rules or “experience”. Patients with high-risk factors or symptoms, or who are prone to be severely affected by a particular disease or illness, may be shortlisted to see a specialist for further treatment.

Diseases Prediction System:

Support Vector Machine and Multilinear Regression with Algorithm:

The evolution of modern technologies such as data science and machine learning has paved the way for early detection of diseases in the healthcare community and medical institutions and has helped to provide better services to patients. When we do not have complete medical data, the accuracy of detecting potential diseases decreases. Furthermore, some diseases are area-based, which can weaken the prognosis. When something goes wrong in your body, your body shows symptoms, sometimes it may be just a minor problem but sometimes you may get serious illness and if you do not take care of these symptoms at an early stage, recovery may be delayed. Disease. A diagnostic system that can diagnose potential diseases based on symptoms so that it can be cured at an early stage. This saves time for a complete diagnosis of the patient and we can diagnose the diseases required by the patient based on the instructions given by the system.

It is a system built using machine learning algorithms to predict potential diseases based on patient symptoms. The development of technology has so far improved our lives. It provides many tools that can save millions of lives, and machine learning is one of them. Machine learning is used to develop systems that predict many diseases based on symptoms. This may indicate to the doctor, the possibility of possible diseases. And the diagnosis can be made according to the advice, so the cost can be reduced.

We are living in the age of technology and nowadays human beings can say that almost anything is possible with the help of technology. Today we have so many tools and methods to get information from any region of the world and at this age information is so important that we cannot live without information. We have tools that can give us relevant information at our fingertips, and the Internet is one of them. Today, billions of search queries are made every day and sometimes the results given are relevant and sometimes they are not. In those search queries, thousands of searches are related to medical advice. People usually want to know if they have any serious disease based on their signs and symptoms. But there are no tools available to give them proper information. This research seeks to provide them with tools to provide potential disease predictive information to the end user.

Figure 4.1 Flowchart Methodology

Symptom Dataset Collection
Symptom Dataset Collection
Symptom Dataset Collection
Symptom Dataset Collection
Symptom Dataset Collection

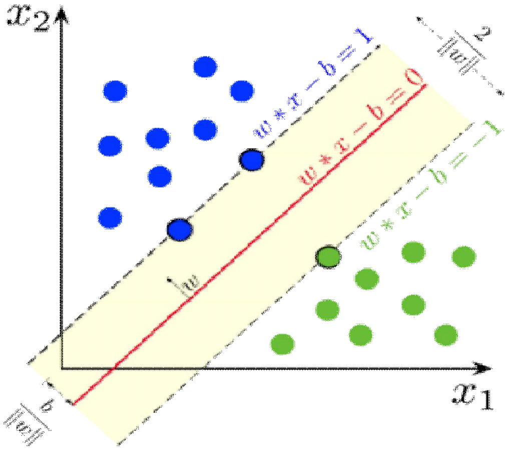
Our proposed method involves the following steps:

Step1: First collect a dataset of symptoms and functional problems in their body.
Step2: Then collect the information related to the possible disease related symptoms so that the information related to the related disease will be collected.
Step3: Then get the symptoms as input from the patient and process it by multilinear regression.
Step4: Multilinear regression then predicts potential diseases for those acquired symptoms.
Step5: The system will then show the diagnosis in the form of the most probable disease and the least probable disease.

Algorithm:

The Support Vector Machine (SVM) classifies N-dimensional hyperplanes by creating well-separated data into two categories. SVM is a supervised learning method used for classification and regression. SVM belongs to the family of generalized linear classification. An important feature of SVM is that the geometric average is maximized in this classification and the empirical classification error rate is reduced. Hence it is known as maximum margin classifiers. Structural Risk Minimization (SRM) is another concept on which SVM is based. The SVM map input vector is a high dimensional space where the maximum separating hyperplane is built. Two hyperplanes are created to classify data into two separate classes. Separate hyperplanes are hyperplanes that increase the distance between two parallel hyperplanes. It is assumed that the generalization error of the classification will be corrected by increasing the distance or margin between the separated hyperplanes.

Figure 4.2 Classifier of SVM



SVM represents the instance data as points in space, then the data is mapped to the space of that feature so that the instance data of different categories are spaced as wide as possible. Consider an example, given a set of points corresponding to one of two classes, SVM finds a hyperplane containing the largest number of points of the same square. This separation hyperplane is called optimal separating hyperplane (OSH) which increases the distance between two parallel hyperplanes and can reduce the risk of incorrect classification of examples of test datasets. If some training data is given as 'D', the set form of n points.

$$D = \left\{(x_i, y_i) | x_1 \in \mathbb{R}^p, y_1, \in \{-1,1\}\right\}_{i=1}^{n}$$

Equation – 1

The value of y_i is 1 or -1 which indicates the point to which x_i belongs. Basically x_i is the actual vector of dimension 'p'. We are interested in hyperplanes that increase margins and separation points with the value $y_i = -1$, rather than $y_i = 1$.

The definition of a hyperplane with a set of $'X'$ points is: "$W.X - b = 0$". Where 'w' is the common vector of the hyperplane? Offset of hyperplane with respect to common vectors 'w' is $\frac{b}{\|W\|}$.

In the case of linear separation training data, we can select two hyperplanes in such a way that they can separate data with no dots between them and then try to increase their distance. The area bordering them is called the "margin". These hyperplanes can be described by equations.

Equation - 2

$$W.X - b = 1$$

Equation – 3

$$W.X - b = -1$$

A description of the distance between two hyper planes from the two equations given above done by $\frac{2}{\|W\|}$, So the main objective is to reduce $\|W\|$ So as to prevent data points from entering the margin. Thus, we add the following limit for each 'i'.

Equation – 4

$$W.X_i - b \geq 1$$

- SVM works relatively well when there is a clear gap between the classes.
- SVM is more effective in high dimensional spaces.
- SVM is effective in cases where the number of dimensions exceeds the number of samples.
- SVM is relatively memory efficient
- SVM is a very useful method if we have little idea about data. It can be used for data like image, text, audio etc. It can be used for data that is not regularly distributed and has an unknown distribution.
- SVM provides a very useful technique known as kernel and we can solve any complex problem using related kernel function.
- The kernel allows you to select a function that is not necessarily linear and may have different variants in relation to the different data it operates and thus is a non-parametric function.
- In classification problems, the data contains linearly separable patterns but with the introduction of the kernel, there is a strong belief that input data can be converted into high dimensional data without the need for this assumption.
- SVMs generally do not tolerate overfitting conditions and perform well when there are clear signs of segmentation. SVM can be used when the total number of samples is less than the number of dimensions and performs well in terms of memory.
- SVM works well out of sample data and generalizes. This proves that SVM proves itself to be fast on sample data outside of normalization because it is certain that in SVM for a sample classification, the kernel function is evaluated and done for each support vector.
- Another important advantage of the SVM algorithm is that it is capable of handling even high dimensional data and is a great help considering its use and application in the field of machine learning.
- The support vector machine is useful for detecting isolated hyperplanes, it can be useful to locate hyperplanes to properly classify data between different groups.
- SVM has the form of convex optimization which is very useful as we are assured of optimum in results so the answer will be global minimum instead of local minimum.
- In SVM, we can create it with big differences, we can fit in more data and classify it perfectly.

CONCLUSION:

The purpose of the SVM algorithm is to detect hyperplanes in n-dimensional space that clearly classify data points. The magnitude of the hyperplane depends on the number of features. If the number of input features is two, the hyperplane is only one line. If the number of input features is three, the hyperplane becomes a 2-D plane. It's hard to imagine when the number of features is more than three.

REFERENCES:

1. Arun Pushpan, Ali Akbar N. (2017), 'Data Mining Applications in Healthcare', IOSR Journal of Computer Engineering, ISSN 2278-0661, pp. 1-4.
2. Behal Sunny and Krishan Kumar (2016), 'Trends in Validation of DDoS Research', International Conference on Computational Modeling and Security, Procedia Computer Science, 85 (2016), 7-15.
3. Deepti Mishra and Devpriya Soni (2018), 'Outliers in Data Mining: Approaches and Detection', International Journal of Engineering & Technology, International Journal of Engineering & Technology, Vol-7, pp. 189-198.

4. Durairaj M. and Ranjani V. (2013), 'Data Mining Applications In Healthcare Sector: A Study', International Journal Of Scientific & Technology Research, Vol-2, Issue-10, pp. 29-35.
5. Kavita, Mahani Priyanka and Ruhil Neelam (2016), 'Application of Data Mining in Health Care', International Journal of Advacne Technology in Engineering and Science, Vol-4, Issue-2, pp. 65-70.
6. Niya Werts and Monica Adya(2000), 'Data Mining in Healthcare: Issues and a Research Agenda', Association for Information Systems, Americas Conference on Information Systems, pp. 94-97.
7. Ray Rakhi (2018), 'Advances in Data Mining: Healthcare Applications', International Research Journal of Engineering and Technology, Vol-05, Issue-03, pp. 3738-3742.
8. Ramesh G.S., Rajinikanth T.V. and Vasumathi D. (2017), 'Explorative Data Visualization Using Business Intelligence and Data Mining Techniques', International Journal of Applied Engineering Research ISSN 0973-4562 Volume-12, Issue-24, pp. 14008-14013.