Scientific Journal Impact Factor (SJIF): 4.364

International Journal of Modern Trends in Engineering and Research

IJMTER

ISSN (ONLINE) : 2349 - 9745

ISSN (PRINT) : 2393 - 8161

STUDY OF DISEASE PREDICTION SYSTEM METHODS BASED ON PREDEFINED DATA SET

Mr. Pavan Katgaonkar[1] and Prof. Mr. Shrikant P. Akarte[2]

[1] ME (CSE) ,Second Year,Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravai Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

[2] Assistant Professor, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera Amravati.SantGadgebabaAmravatiUniversity, Amravati, Maharashtra, India – 444701.

Abstract-Clinical decision support system, that uses advanced data mining techniques to assist practitioner create correct selections, has received significant attention recently. The benefits of clinical decision support system embody not solely rising identification accuracy however conjointly reducing identification time. Specifically, with massive amounts of clinical knowledge generated every day, naive theorem classification may be utilized to excavate valuable data to boost clinical decision support system. Although clinical decision support system is kind of promising, the flourish of the system still faces several challenges together with data security and privacy issues. In this paper numerous parameters of the disease prediction system are studied on the predefined knowledge sets.

Keywords—Mining, classification, clinical decision support system, privacy issues, data security.

I. INTRODUCTION

Healthcare industry, extensively distributed within the world scope to produce health services for patients, has ne'er faced such a huge amounts of electronic information or intimate with such a pointy rate of information nowadays. As we all know there are a unit varied forms of diseases are there in atmosphere. Some diseases will be simply get cured however some are there that have dangerous impact on body and may result in death [1].

Diseases like dengue, Malaria, cholera, Chicken pox, and diarrhea have similar characteristics at early stage that makes practitioner job troublesome in diagnosing such severe unwellness. For ex. take into account break bone fever. It’s transmitted by the bite of associate degree Aides dipterous insect infected with a dengue virus. The mosquito becomes infected once it bites an individual with dengue virus in their blood. It can’t be unfold directly from one person to a different person. Sometimes, symptoms area unit delicate and might be mistaken for those of the influenza or another virus infection by the victims or patients.

Today's such type of diseases makes us to feel the necessity of implementing the system which might predict the diseases as early as potential based on patient health status. If it's potential to predict the unwellness at early stage based on offered symptoms we will offer several healthful treatment to patient who is suffering.

II. RELATED WORK

Data mining is refer to as mining the knowledge from great amount of datasets. It’s conjointly remarked as knowledge discovery from data (KDD). There’s great amount of knowledge obtainable within the society in numerous fields and it's useful to show such data into helpful information and knowledge. The information gained from this data is used for various functions in several applications. Data processing is a multidisciplinary field that embrace the work areas like database technologies, machine learning, pattern recognition, data retrieval, neural network, computer science. There square measure completely different data mining techniques which may be used for extracting information from great amount of information [2].

Classification is additionally one amongst the technique of data mining wherever a classifier is built to predict categorical labels. It conjointly embrace completely different techniques like Naive Thomas Bayes, k-nearest neighbor, neural network, decision tree. The aim to use data mining technique is typically allows one to gather, store, access, method and ultimately describe and visualize data sets.

III. CLUSTERING AND ITS TECHNIQUES

Clustering is a technique in which objects of similar type from large dataset are grouped into one cluster. It is nothing but partitioning a set of data into a set of meaningful sub-classes called cluster [3].

3.1 Partitioning Based Method

Partitioning algorithmic program may be a non-hierarchical, it construct various partitions so valuate them by some criterion. It construct a partition of an information D of N objects into a collection of K clusters, wherever user ought to predefined the quantity of cluster (K). K-means algorithmic program comes below partitioning primarily based technique. It's one among the largely used cluster algorithmic program.

3.2 Hierarchical Clustering

Hierarchical cluster builds a hierarchic decomposition of the set of data or objects using some criterion. It are often envisioned as a tree like diagram that records the sequences of merges or splits. Any desired range of cluster are often obtained by "cutting" the tree at the correct level.

3.3 Graph Based Clustering

Graph cluster is that the task of grouping the vertices of the graph into clusters taking into thought the sting structure of the graph in such the simplest way that there ought to be several edges among every cluster and comparatively few between the clusters.

IV. CLASSIFICATION & ITS METHODS

Classification is a data mining machine learning technique accustomed predicts group membership for knowledge instances. For instance, you may want to use classification to predict whether or not the weather on a specific day is going to be sunny, rainy or cloudy. Widespread classification techniques include decision trees and neural networks.

4.1 Classification methods

4.1.1. Machine Learning Based Approach

Machine Learning is usually covers automatic computing procedures supported logical or binary operations that learn a task from a series of examples. Here we tend to square measure simply concentrating on classification then attention has focused on decision-tree approaches within which classification results from a sequence of logical steps [4]. These classification results are capable of representing the foremost complex problem given adequate knowledge. Alternative techniques like genetic algorithms and inductive logic procedures (ILP) are presently below active improvement and its principle would enable us to subsume additional general forms of data together with cases wherever the quantity and sort of attributes may vary. Machine Learning approach aims to come up with classifying expressions straightforward enough to be understood simply by the human and should mimic human reasoning sufficiently to produce insight into the choice method [5].

Techniques that's in the main accustomed analyze a given dataset and takes every instance of it and assigns this instance to a selected class such that classification error are going to be least. It's accustomed extract models that accurately outline important data categories inside the given dataset. Classification is a two-step method. throughout beginning the model is formed by applying classification rule on training data set then in second step the extracted model is checked against a predefined test dataset to measure the model trained performance and accuracy. Thus classification is that the method to assign class label from dataset whose class label is unknown [6].

4.1.2 ID3 Algorithm

Id3 calculation starts with the initial set because the root hub. On each cycle of the rule it emphasizes through each unused attribute of the set and figures the entropy (or data acquire IG(A)) of that attribute. At that time chooses the attribute that has the tiniest entropy (or biggest knowledge gain) value. The set is S then split by the chosen attribute (e.g. marks < 50, marks < 100, marks >= 100) to provide subsets of the data. The rule take to recurse on each & every item in set and considering solely things ne'er chosen before.

4.1.3. C4.5 Algorithm

C4.5 is an algorithm accustomed manufacture a decision tree that is an enlargement of previous ID3 calculation. It enhances the ID3 algorithm by managing both continuous and discrete properties, missing values and pruning trees when construction. The choice trees created by C4.5 may be used for grouping and infrequently mentioned as a statistical classifier. C4.5 creats decision trees from a collection of training data same method as Id3 rule [7].

4.1.4. K-Nearest Neighbors Algorithm

The nearest neighbor (NN) rule distinguishes the classification of unknown information on the premise of its nearest neighbor whose class is already well-known. M. cowl and P. E. Hart purpose k nearest neighbor (KNN) during which nearest neighbor is computed on the premise of estimation of k that indicates how many nearest neighbors are to be thought-about to characterize category of a sample data point. It makes utilization of the over one nearest neighbor to work out the category during which the given data point belongs to and consequently it's known as KNN. These data samples are required to be within the memory at the run time and thus they are referred to as memory-based technique [8].

4.1.5. The Naive Bayesian Classifier

Bayes theorem. A Naive Bayesian algorithm is straightforward to create, with no sophisticated repetitious parameter estimation that makes it notably helpful for very massive datasets. Mathematician theorem provides a way of calculating the posterior probability, P(c|D), from P(c), P(D), and P(D|c). Naive bayes classifier assume that the impact of the worth of a predictor (x) on a given class (c) is freelance of the values of alternative predictors [9].

$P(c \mid D)=(P(D \mid c)P(c))/P(D)$

Where,

P(c|D) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(D|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

4.1.6. SVM Algorithm

SVM have attracted an excellent deal of attention within the last decade and actively applied to numerous domains applications. SVMs are usually used for learning classification, regression or ranking operate [10]. SVM are based on statistical learning theory and structural risk reduction principal and have the aim of determinative the location of decision boundaries also known as hyperplane that turn out the optimum separation of classes [11].Maximizing the margin and thereby making the largest possible distance between the separating hyperplane associate degreed the instances on either aspect of it has been established to reduce an edge on the expected generalization error [12]. Efficiency of SVM based mostly classification isn't directly rely upon the dimension of classified entities.

V. CONCLUSION

Classification methods are usually strong in modeling communications. Every of those strategies may be utilized in various situations as required wherever one tends to be useful whereas the opposite may not and vice-versa. These classification algorithms may be implemented on different kinds of data sets like share market data, information of patients, money information,etc. therefore these classification techniques show however a data can be determined and classified once a new set of information is obtainable.

VI. ACKNOWLEDGMENTS

Especially thankful to my guide Mr. Shrikant P. Akarte, Assistant Professor at PRMITR, Badnera who helped me and guided me to complete this paper. Hearty thanks to worthy institute PRMITR, Badnera and all those that we have inadvertently forgotten for their immense support and encouragement.

REFERENCES

[1] "Transforming health care through big data strategies for leveraging big data in the health care industry," http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-bigdata-research-report-download-today/, 2013.
[2] J. Han and M. Kamber, "Data Mining Concepts and Techniques", Elevier, 2011.
[3] M .Aly, "Survey on Multiclass Classification Methods", November (2005).
[4] T.Joachims, "Making large-scale support vector machine learning practical", In Advances in Kernel Methods: Support Vector Machines, (1999).
[5] D. Michie, D.J. Spiegelhalter, C.C. Taylor "Machine Learning, Neural and Statistical Classification", February 17, (1994).
[6] DelveenLuqmanAbdAl.Nabi, ShereenShukri Ahmed, "Survey on Classification Algorithms for Data Mining: (Comparison and Evaluation)" (ISSN 2222-2863)4(8); (2013)
[7] H. Bhavsar, A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231 -2307, 2(4); (2012)
[8] Dasarathy, B. V., Nearest Neighbor (NN) Norms,NN Pattern Classification Techniques. IEEE Computer Society Press, 1990.
[9] Rish, "An empirical study of the naive bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001, pp. 41–46.
[10] V. Vapnik and C. Cortes, "Support Vector Network," Machine Learning, 20; 273-297, (1995).
[11] C. J. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, 2; (1998).
[12] V. Vapnik, "Statistical Learning Theory", Wiley, New York, (1998).