A SURVEY ON DISEASE DIAGNOSIS ALGORITHMS

Aanchal Oswal, Vachana Shetty, Mustafa Badshah, Rohit Pitre, Manali Vashi

***Abstract*— Lots of data related to healthcare is available but knowledge is not mined from this huge data. There is lack of effective analysis tools to discover hidden relationships and trends in data. Data mining methods are used to analyse the medical data information resources. The knowledge gathered from the data of patients collected in databases facilitate in the diagnosis process. Even if we have records it is necessary to tap the hidden information which might turn out to be helpful in critical situations and this can be achieved by various data mining algorithms, which we will be discussing in this survey.**

***Index Terms*—C4.5, K-NN, SVM, RVM**

I. INTRODUCTION

Medical data mining has been a great potential for exploring hidden patterns in data sets of medical domain. In healthcare, though data mining is not widely used but it has become increasingly popular. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data. There are two primary goals of data mining-prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. On the other hand Description focuses on finding patterns describing the data that can be interpreted by humans.

With the rapid increase in population the number of diseases occurring is also increasing. Many diseases have closely related symptoms which make it difficult for the doctors to predict the exact disease in one go. This is where data mining comes into assistance; it helps in predicting the disease which is nearly accurate. Even though the prediction is not very accurate it at least gives the doctor a brief idea what the disease could be. Thus, in no way disease prediction using data mining is a substitute to doctors whereas; it's a compliment to the doctors.[1][8]

Manuscript received Nov, 2014.

***First Author Aanchal Oswal**, Computer Science, KJ College of Engineering and Management Research, Pune, India, (email: aanchaloswal@gmail.com).*

***Second Author Vachana Shetty**, Computer Science, KJ College of Engineering and Management Research, Pune, India, (email: vachanakshetty23@gmail.com).*

***Third Author Mustafa Badshah**, Computer Science, KJ College of Engineering and Management Research, Pune, India, (email: mustafab325@gmail.com).*

***Fourth Author Rohit Pitre**, Computer Science, KJ College of Engineering and Management Research, Pune, India, (email: rohitpitre3@gmail.com).*

***Fifth Author Manali Vashi**, Computer Science, KJ College of Engineering and Management Research, Pune, India, (email: manali.vashi@kjsedu.com).*

II. EXISTING SYSTEMS

We have done a survey on disease diagnosis research, there are two types of systems are available for this purpose-

- Domain specific which is for doing diagnosis of particular diseases such as system involved with diagnosis of heart diseases only it won't identify any other diseases.
- Second types of system which concentrate on multiple diseases diagnosis.

Further we can divide medical diagnosis system into two category such as system which uses symptoms to identify diseases and not considering other variables such as family medical history, age, etc. and other kind of such system which considers symptoms along with variables such as medical history, age etc.

In literature various techniques are available in data mining domain. Ontology's are also created as a previous work for disease diagnosis.

III. DATA MINING FOR DISEASE DIAGNOSIS

Domain specific algorithms for diagnosis

There are many individual disease diagnosis algorithms are available, we have done survey on heart disease, breast cancer and diabetes.

Diabetes

In [2], there is prediction of diabetes using amalgam K-NN with K-means. The inconsistent data from Pima Indian Diabetes Database (PIDD) is identified and corrected using K-means clustering algorithm thereafter K-NN classification algorithm is used to classify the data.

Methodology:

- The first step is the pre-processing where the inconsistent values are removed.
- The K-means clustering algorithm is used to identify and correct these inconsistencies.
- The missing values are replaced by means and medians.
- After the inconsistencies are removed the K-NN classification algorithm is used to classify the data.
- This model is then tried for different values of *k.*

Diabetes is a disease in which body does not produce insulin or use it properly. There are two types of diabetes; Type-1 diabetes- also called insulin dependent and type-2 diabetes which is with relative insulin deficiency. [3] Paper deals about the classification/prediction of Type II diabetes.

Methodology:

This study consists of two stages, data pre-processing and decision tree construction.

- **Data Pre-processing:** [3] uses attribute identification and selection, handling missing values, and numerical discretization techniques for pre-processing.
- **Decision Tree:** After the dataset has been prepared, Weka software was used to construct the decision tree. The dataset was then classified by choosing the J48 algorithm which is a decision tree learner and is the implementation of Quinlan C4.5 in Weka software.

Algorithm/methods	Limitations/disadvantages	Various measures used
Amalgam of K-means and K-NN[2]	It is lazy learner. There is no thumb rule to determine value of parameter k (Number of nearest neighbours).	Accuracy ,sensitivity and specificity
J48 Decision Tree Algorithm & C4.5[3]	There are other risk factors that the data collection does not consider. Dataset contains data of only female patients.	Accuracy – 78.1768%

Table1: Summary of disease diagnosis algorithm for diabetes [2][3]

Heart Disease

Diagnosis of heart disease is a significant and tedious task in medicine.

[4] Gives us a simple technique to predict risk of heart attacks. The data classification is based on MAFIA algorithms which result in accuracy. C4.5 algorithm is used as the training algorithm to show rank of heart attack with the decision tree. Finally, the heart disease database is clustered using the K-means clustering algorithm. The results showed that the system is capable of predicting the heart attack successfully.

Methodology:

- The database is pre-processed successfully by eliminating identical records and providing missing values.
- The polished data set is then collected by K-means algorithm with the K value of 2.
- The forms are mined efficiently from the collection applicable to heart disease, using the MAFIA algorithm.
- The model consortiums of heart attack parameters for ordinary and risk level along with their values and levels are detailed.
- In that, ID lesser than of (#1) of weight contains the normal level of prediction and higher ID other than #1 comprise the higher risk levels.
- A subsequent decision tree is created to show the heart attack level using C4.5 algorithm by information gain.

[5] Proposes a classification algorithm which combines KNN and genetic algorithm, to predict heart disease of a patient. The author uses genetic search as a measure to check redundant and irrelevant attributes, and to rank the attributes which contribute towards classification. Least ranked attributes are removed, and classification algorithm is built based on evaluated attributes. The classifier classifies heart disease data set as either healthy or sick.

Methodology:

- Apply genetic search on the data set
- Attributes are ranked based on their value
- Select the subset of higher ranked attributes
- Apply (KNN+GA) on the subset of attributes that maximizes classification accuracy
- Calculate accuracy of the classifier

Algorithm/methods	Limitations/disadvantages	Results
K-Mean based MAFIA with ID3 and C4.5 [4]	Prediction of heart attack using patient prescription is not included.	Accuracy – 92% Precision – 0.82
KNN and genetic algorithm[5]	(KNN+GA) was not successful for breast cancer and primary tumour. As the k value goes on increasing accuracy of data sets is decreasing.	Accuracy – 95.73%

Table2: Summary of disease diagnosis algorithm for heart diseases [4][5]

Cancer

In paper [6] the author has used J48 classification algorithm for detecting breast cancer using a two level diagnosis. In the first level the diagnosis is done on the basis of Wisconsin Breast Cancer Dataset (WBCD); the result obtained from the WBCD is classified into malignant and benign classes. At the second level diagnosis is based on the pathological and physiological parameters of malignant breast cancer dataset and then classified into various types.

Methodology:

- At the first level of diagnosis the input is taken from the WBCD.
- The result obtained from the WBCD is then classified into malignant and benign classes.
- At the second level of diagnosis, based on pathological reports, if the tumour is detected as malignant, then the type of breast cancer is classified into five types- Ductal Carcinoma in Situ (DCIS), Lobular Carcinoma in Situ (LCIS), Invasive Ductal Carcinoma (IDC), Invasive Lobular Carcinoma (ILC) and Mucinous Carcinoma (MC).

In this paper [7], we focus on mass detection using SVM classifier and texture analysis. Mammography is the dominant method for detecting breast cancer assisted by computers. Segmentation of mammogram plays a major role in isolating

areas which can be subject to tumors. In this we segment the mammogram in three areas: the pectoral muscle, fibro glandular tissue and adipose tissue, then we analyze the dense tissues and we classify them in normal tissue and pathological ones. The Identification of these involves three stages: pectoral muscle segmentation, hard density zone detection and texture analysis of regions of interest.

Methodology

1. **Texture analysis of mammograms:**
   - Several works were devoted to the segmentation of the pectoral muscle. The co-occurrence matrix is the most common method to analyse mammographic image textures.
   - This characteristic makes it possible to differentiate normal tissue from pathological tissue.
   - To test our approach we used 95mammographic images: calcification (20 cases); circumscribed masses (22cases); speculated masses (19 cases); ill-defined (15 cases) and 19 normal mammograms. The size of all the images is 1024 pixels x 1024 pixels.
   - SVM classification based on Haralick vector**:** one has to choose a window of suitable size. The size of this window must be as small as possible to reduce risk of mixing of textures and as large as possible to extract robust and significant statistics. Ideal size being 7x7 and 9x9.

2. **Pectoral muscle detection:**

   - To detect the pectoral muscle, active contours is one of the better performing methods but the disadvantage being that we have to initialize close to the required result, for this auto initialization is used.
   - This is done by S.M.Kwok and R. Chandrasekhar where first pectoral muscle is estimated by a straight line which is validated for correctness of location and orientation. Then accuracy is redefined by "Cliff detection".

Algorithm/methods	**Limitations/disadvantages**	**Result**
J48 classification algorithm [6]	J48 is not feasible when larger dataset is used, as small change in dataset reflects in larger modification in decision tree.	Sensitivity 94%-100%
SVM classification based on Haralick vector Algorithm developed by S.M. Kwok and R. Chandrasekhar. [7]	The displacement and the orientation used for the calculation of co-occurrence matrix significantly affect the results. The mixing of two approaches(co-occurrence and contours) just gave satisfactory results	95% of classification rate can be achieved by using pre-segmented mammograms by maxima thresholding.

Table3: Summary of disease diagnosis algorithm for cancer [6][7]

Multiple Diseases

In [8]MediQuery medical system is proposed for multiple disease diagnosis based on symptoms, recent medical trends and misdiagnosis factor.

Method

- Workflow of proposed algorithm includes three steps, which are as follows: First step is pattern matching. Second step is to apply recent medical history and third step is differential diagnosis.
- In pattern matching Symptoms probability is computed with group of disease instead of individual disease. Bayes classification is used to classify disease into sub groups. Support and confidence is calculated for each sub group of disease. Confidence and support is calculated for symptoms entered by patient, which is compared with disease subgroup and appropriate disease sub group is found. After that disease is searched in that sub group only.
- In step-2 recent medical records of 3 months are used for diagnosis purpose.
- In step-3 Differential diagnosis is performed using Hebbian learning and weights are adjusted using back propagation.[9]
  In paper [8] Authors have proposed system for multiple disease diagnosis,
  Method
- It uses different algorithm on each steps of paper [8].In step-1 it uses K-NN classification instead of Baysian classification used in[9].In Step-2 it uses recent trends. In step-3 it uses Hopfield network with weight assigning using LAMSTAR network where as in [8]Hebbian learning and back propagation is used.[8]

IV. RESULTS

Result shows that K-NN,Hopfield and LAMSTAR used in [8] is faster than Baysian, Habbian and Back propagation used in [9].

In paper [10] meta classifier are used for diagnosis purpose. The problem is that it is not known from beginning which is best classifier, so in paper [10] authors have developed meta classifier combination scheme which combines various

classifier for medical diagnosis. Authors have tested meta classifier based proposed algorithm for diagnosis of breast cancer, diabetes, new thyroid problem, hepatic lesions. Meta classifier is composed of three levels (Lower, Middle, and Upper). Lower level includes simple classification algorithms such as neural network, Support Vector Machine, C4.5 Descion Tree. There are combination methods (Bagging, Boosting) in middle level that combines classifier level below. The proposed meta classifer combines various methods of middle level in upper level. The four different combination scheme at upper level are crisp voting(class labels without weights),weighted voting(class labels with weights),Average class probability(class probability without averaging),class probabilities weighted sum(class probabilities with weighted sum).Result shows that proposed meta classifier performs better than Lower level & Middle level classification algorithms.[10]

Algorithm / Diseases	SVM	K-NN	Decision Tree	C4.5	RVM
Heart Disease		Yes		Yes	
Diabetes		Yes	Yes		
Cancer	Yes		Yes	Yes	Yes
Multiple Diseases		Yes			

Table4: Summary of disease diagnosis algorithms

V. DATASET CONSIDERED

- **PIDD[2][3] - Pima Indian Diabetes Database**
  Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

- **UCI Repository[5]**
  The performance of proposed approach has been tested with 6 medical data sets and 1 non medical data set. Out of 7 data sets, 6 data sets were chosen from UCI Repository and heart disease A.P was taken from various corporate hospitals in Andhra Pradesh, and attributes are selected based on opinion from expert doctor's advice.

- **WBCD[6] - Wisconsin Breast Cancer Dataset**
  Samples arrive periodically as Dr. Wolberg reports his clinical cases. The database therefore reflects this chronological grouping of the data. This grouping information appears immediately below, having been removed from the data itself.

- **MIAS[7] - Mammographic Image Analysis Society**
  Most of the mammographic databases are not publicly available. The most easily accessed databases and therefore the most commonly used databases is the Mammographic Image Analysis Society (MIAS) database.

VI. OPEN RESEARCH PROBLEMS

- We cannot predict the specific disease through the patient's prescription data collection.
- Specific disease prediction algorithms do not give the same accuracy in predicting different diseases by using the same algorithm and similar attributes.
- All the risk factors are not included in the attributes for the process of disease prediction.
- Pre-processing cannot be standardized as every disease requires processing on different parameters.
- Large and complex datasets decreases the accuracy of prediction and increases the steps of processing and calculating the accurate prediction or diagnosis.

VII. CONCLUSION

We have done survey on various data mining algorithms for disease diagnosis. Still No method is 100% even doctors diagnosis can't be 100%.Such system are useful for doctors to provide assistance. Various algorithms have been proposed for diagnosis. For multiple disease diagnosis system must provide faster diagnosis result, system must be self learning. Such system must consider various parameters such as recent medical trends, seasonal effect etc.

REFERENCES

[1] Ming Li, Member, IEEE, Shucheng Yu, Member, IEEE, Yao Zheng, Student Member, IEEE, Kui Ren, Senior Member, IEEE, and Wenjing Lou, Senior Member, IEEE "Scalable and Secure Sharing of Personal Health Records in Cloud Computing Using Attribute-Based Encryption", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 1, JANUARY 2013
[2] NirmalaDevi.M, Appavu alias Balamurugan.S,
Swathi U.V, An amalgam KNN to predict Diabetes Mellitus,
[3] Asma A. AlJarullah, Decision Tree Discovery for the Diagnosis of Type II Diabetes
[4] Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods
[5] Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm
[6] Rajkumar Gaur Grewal, Babita Pandey "Two level Diagnosis of Breast Cancer using data mining."
[7] Fatima Eddaoudi, Fakhita Regragui, Abdelhak Mahmoudi and Najib Lamouri. Masses Detection Using SVM Classifier Based on Textures Analysis
[8] Rahul Isola, Student Member, IEEE Rebeck Carvalho, Student Member, IEEE, Amiya Kumar Tripathy, Member, IEEE Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and *k*-NN.
[9]Rebeck Carvalho,Rahul Isola,Amiya KumarTripathy "MediQuery-An Automated Decision Support System",IEEE,ISSN :1063-7125,Page No-1-6
[10]George L. Tsirogiannis, Dimitrios Frossyniotis, Konstantina S. Nikita, and Andreas Stafylopatis "A Meta-classifier Approach for Medical Diagnosis",G.A. Vouros and T. Panayiotopoulos (Eds.): SETN 2004, LNAI 3025, pp. 154–163, 2004

First Author Aanchal Oswal is BE student of K.J. College of Engineering & Management Research, Pune. Her current research interests are data mining and android development.

Second Author Vachana Shetty is BE student of K.J. College of Engineering & Management Research, Pune. Her current research interests are data mining and cloud computing.

Third Author Mustafa Badshah is BE student of K.J. College of Engineering & Management Research, Pune. His current research interests are data mining.

Fourth Author Rohit Pitre is BE student of K.J. College of Engineering & Management Research, Pune. His current research interests are data mining and web development.

Fifth Author Manali Vashi is Assistant Professor at K.J. College of Engineering & management Research, Pune. Her current research interests are data mining, image processing, and parallel processing