A NOVEL METHOD FOR DATA MINING TECHNIQUES IN ALZHEIMER'S DISEASE

[1]Saravanan N, [2]Kesavan K
[1]Assistant Professor, [2]II MCA Student
[1 2]Department of Computer Applications,
[1 2]Priyadarshini Engineering College, Vaniyambadi, Vellore, Tamilnadu, India

Abstract: Finding a cure for Alzheimer's disease has been facing challenges due to the lack of reliable biomarkers for detection of risk. Fluid based biomarkers provide some criteria for identification of the disease's current stage in patients. But these markers are not reliable predictors for disease progression or response to treatment also most of these markers are tested in cerebrospinal fluid which reduces the applicability of the method, significantly. The main purpose of this paper is to describe research surveys in effect of blood-based biomarkers and diagnostic imaging in AD, using data mining techniques.

***IndexTerms- Data mining, SVM, feature vector, Alzheimer's* disease**

I. INTRODUCTION

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. This could be a win/win overall [1]. But due to the complexity of healthcare and a slower rate of technology adoption, our industry lags behind these others in implementing effective data mining and analytic strategies.

The treatments for Alzheimer's disease has been a promising and disappointing endeavor over the past two decades, yielding a greater understanding of the disease yet still failing to generate successful new drugs [2]. To identify similar disease types, using a multilayer clustering algorithm to sort through dozens of data points from two large studies of the Alzheimer's Disease Neuroimaging Initiative. Study data included cognition tests, brain scans and spinal fluid biomarkers from cognitive impairment.

Alzheimer's Disease (AD), with the goal being the analysis of risk factors and identifying tests that can help diagnose AD [3]. The existing multiple studies that analyze the factors that can help diagnose or predict AD, the study of non-image data, while using a multitude of techniques from machine learning and data mining [4]. The applied methods include classification tree analysis, cluster analysis, data visualization, and classification analysis. All the analysis, except classification analysis, resulted in insights that eventually lead to the construction of a risk table for AD. The study contributes with new insights, but also by demonstrating a framework for analysis of such data. The insights obtained in this study can be used by individuals and health professionals to assess possible risks, and take preventive measures [5].

II. DATA MINING TECHNIQUES

A. Classification Analysis

This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes [6]. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster [7]. So, in classification analysis you would apply algorithms to decide how new data should be classified. A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.

B. Association Rule Learning

It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases [8]. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset [9]. Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning [10].

C. Anomaly Or Outlier Detection

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of

data [11]. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention. This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy [12].

D. Clustering Analysis

The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters [13]. Clustering analysis is the process of discovering groups and clusters in the data in such a way that the degree of association between two objects is highest if they belong to the same group and lowest otherwise. A result of this analysis can be used to create customer profiling.

E. Regression Analysis

In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting [7, 8].

III. DATA MINING IN HEALTHCARE

Data mining possesses great potential for the healthcare industry, but it also comes with a few privacy concerns. Massive amounts of patient data being shared during the data mining process may leave some patients worried that their personal information could fall into the wrong hands. Some intelligent data mining techniques to guess the most accurate illness that could be associated with patient's symptoms [14]. If the system is not able to provide suitable results, it informs the user about the type of disease or disorder it feels user's symptoms are associated with. If symptoms do not exactly match any disease in our database, are shows the diseases user could probably have judging by his/her symptoms.

IV. ALZHEIMER'S DISEASE DATASETS

The AD data set used in this study was obtained from Ray *et al* (2007). The following section describes them in detail:

The AD data set has a total of 259 samples with 120 known signaling proteins in MS-Excel format. The plasma samples associated with these data points were collected from several academic centers specializing in neurological or neurodegenerative diseases (Ray *et al*, 2007). The data set is divided into a number of subsets as shown in Table1.

Clinical diagnosis	Number
Alzheimer disease (AD)	85
Non-demented control (NDC)	79
	Training set
	AD: 43
	NDC : 40
	Test set
	AD: 42
	NDC: 39
Other dementia (OD)	11
Mild cognitive Impairment (MCI)	47
	MCI -> AD: 22
	MCI -> OD: 8
	MCI -> MCI: 17
Other neurological disease (OND)	21
Rheumatoid arthritis (RA)	16

Table1. Description of subsets of Alzheimer Diseases

All data sets had the same format as samples were arranged in columns and proteins were arranged in rows. Although the data sets did not require any data pre- processing, the format of the data needs to be modified in order to work with the different programs SAM, PAM and W EKA as they use different input formats. The data format used for PAM was the same format as SAM's format, but the PAM program allows the class labels to be in both numeric and alphabetic format, therefore there was no

need to change the original class labels (AD and NDC) to numeric values. The original data supplied with rows for proteins (features) and columns for samples. This format works fine with SAM and PAM, but it is not suitable for WEKA because WEKA requires the data in the format of rows for samples and columns for features, and the class labels must be i n the last column . Therefore the data need to be converted to the WEKA required format.

V. BENEFITS OF DATA MINING IN HEALTHCARE INDUSTRY

The solutions might favor healthcare providers or insurance companies, data mining benefits everyone concerned, from healthcare organizations to insurers to patients [15, 16].Patients receive more affordable and better healthcare services. This happens when healthcare officials use data mining programs to identify and observe high-risk patients and chronic diseases and design the right interventions needed. These programs also reduce the number of claims and hospital admissions, further streamlining the process.

Healthcare providers use data mining and data analysis to find best practices and the most effective treatments. These tools compare symptoms, causes, treatments and negative effects and then proceed to analyze which action will prove most effective for a group of patients. This is also a way for providers to develop the best standards of care and clinical best practices [12, 13]. Insurers are now able to better detect medical insurance abuse and fraud because of data mining. Unusual claims patterns are easier to spot with this tool and it can identify inappropriate referrals and fraudulent medical and insurance claims. When insurers reduce their losses due to fraud, the cost of health care also decreases.

Healthcare facilities and groups use data mining tools to reach better patient-related decisions. Patient satisfaction is improved because data mining provides information that will help staff with patient interactions by recognizing usage patterns, current and future needs, and patient preferences. Electronic health records (EHR) are quickly becoming more common among healthcare facilities [10, 15]. With increased access to a large amount of patient data, healthcare providers can now optimize the efficiency and quality of their organizations using data mining.

VI. CONCLUSION

In healthcare, data mining has proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare and measuring the effectiveness of certain treatments. The purpose of data mining, whether it's being used in healthcare or business, is to identify useful and understandable patterns by analysing large sets of data. These data patterns help predict industry or information trends, and then determine what to do about them. In the healthcare industry specifically, data mining can be used to decrease costs by increasing efficiencies, improve patient quality of life, and perhaps most importantly, save the lives of more patients.

REFERENCES

[1] Bellazzi, R., Zupan, B.: Predictive data mining in clinical medicine: Current issues and guidelines. International Journal of Medical Informatics 77, 81–97 (2008).

[2] Canlas Jr., R.D.: Data Mining in Healthcare: Current Applications and Issues (2009).

[3] Chen, H., Fuller, S., Friedman, C., Hersh, W.: Medical Informatics. Knowledge Management and Data Mining in Biomedicine. Springer Science (2005).

[4] Cios, K.J., Moore, G.W., Uniqueness of medical data mining. Artificial Intelligence in Medicine 26, 1–24 (2002).

[5] Kolce, E., Frasheri, N., A Literature Review of Data Mining Techniques Used in Healthcare Databases (2012).

[6] Healthcare Information and Management Systems Society. Electronic Health Records. A Global Perspective. White paper. HIMSS Enterprise Systems Steering Committee and the Global Enterprise Task Force (2010).

[7] Janarthanan Y, Balajee J.M, and Srinivasa Raghava S. "Content based video retrieval and analysis using image processing: A review." International Journal of Pharmacy and Technology 8, no.4 (2016): 5042-5048.

[8] Jeyakumar, Balajee, MA Saleem Durai, and Daphne Lopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.

[9] Kamalakannan, S. "G., Balajee, J., Srinivasa Raghavan.,"Superior content-based video retrieval system according to query image"." International Journal of Applied Engineering Research 10, no. 3 (2015): 7951-7957.

[10] Kambatla, K., Kollias, G., Kumar, V. and Grama, A., Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), pp.2561-2573, 2014.

[11] Ranjith, D., J. Balajee, and C. Kumar. "In premises of cloud computing and models." International Journal of Pharmacy and Technology 8, no. 3 (2016): 4685-4695.

[12] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* 61 (2015): 85-117.

[13] Ushapreethi P, Balajee Jeyakumar and BalaKrishnan P, Action Recongnition in Video Survillance Using Hipi and Map Reducing Model,International Journal of Mechanical Engineering and Technology 8(11), 2017,pp. 368–375.

[14] Ushapreethi, P. and Lakshmipriya, G.G."Survey on Video Big Data: Analysis Methods and Applications." International Journal of Applied Engineering Research, 12(10), pp.2221-2231.2017.

[15] Sethumadahavi R Balajee J "Big Data Deep Learning in Healthcare for Electronic Health Records," International Scientific Research Organization Journal, vol. 2, Issue 2, pp. 31–35, Jul. 2017.

[16] Saravanan.N, Pavithra.K, Nandhini.C, "Iris Based E-Voting System Using Aadhar Database",International Journal of Scientific & Engineering Research, Volume 8, Issue 4, pp. 62-64,Apr. 2017.