DIABETIC DIAGNOSIS WITH DATA ANALYTICS IN NEURAL FRAMWORK WITH MACHINE LEARNING

S Prabhavathi[1],
Assistant Professor, Department of IT
SNS College of Engineering,
Coimbatore,India.
prabhaatoffice@gmail.com

Dr. P Sumathi[2],
Head of Department, Department of IT
SNS College of Engineering,
Coimbatore,India.
psumathi.it@gmail.com

K Narayana Prakash[3],
UG Scholar, Department of IT
SNS College of Engineering,
Coimbatore,India.
narayanaprakash3734@gmail.com

G karthikeyan[4],
UG Scholar, Department of IT
SNS College of Engineering,
Coimbatore,India.
Karthikeyan.g201999@gmail.com

M JAGADEESH[5],
UG Scholar, Department of IT
SNS College of Engineering,
Coimbatore,India.
jagadeeshmulla7777@gmail.com

ABSTRACT

Personal health record has emerged as a patient-centric model of health information exchange. To ensure patient-centric control over their own diabetic symptoms analyze statistics are framed, which is essential to have fine-grained data access control mechanisms that scrutinize the performance of diabetic system. Different type of algorithms is framed to create the cluster of record based on the medical history of the diabetic patients. The data set is mined to analyze the symptoms and the un-necessary data are pruned to create a valid set. A disease, if predicted wrongly, there may not be any chance of curing it. With the advancement of computing facility provided by information technology, it is now possible to predict many conditions of ailments more accurately. The first big advantage of information technology is that a huge data storage of past patient's records are maintained and monitored by hospitals continuously.

The stored medical data can be helpful to doctors to examine patterns in the data set. The patterns found in data sets may be used for clinical diagnosis. Clustering and Genetic Algorithms are quite suitable for pattern analyses. The principal objective of clustering computation was to find different groups of diabetes patients with similar symptoms within a group but different symptoms of other groups. The classification was performed based on selected training parameters

Key words - PHR - Personal health record, GA- Genetic Algorithm

I.INTRODUCTION

The Personal health record data set are used by researchers to classify various medical conditions by using neural network based tools such as Perception with adaptive learning routine algorithms. In other research studies, the PHR database set performance were analyzed which had shown similar results to that of the adaptive algorithm with few rules and faster training rate. A pruning algorithm further reduces the number of rules significantly. It also increased accuracy levels. All these techniques are complex, difficult to implement and consumes large computing resources and long training times for converging on expected results. Genetic algorithm models have been used for similar analysis of modeling of worker and manufacturing problems. The model basically focused on classification problem. This work presents analyses made while comparing the other cluster algorithm, H-means clustering and genetic algorithm for clustering analyses in a very simple way. This work is organized by data collection methods and introduces the selection criterion of chosen data set for scrutinize experiments.

K-means algorithm performs clustering based on partitioning the total population as k-number of clusters. Here, the convergence speed depends on the number of iterations chosen for achieving stated clusters. The k-means+ algorithm may contain empty clusters. Lloyd's h-means+ algorithm improve the concept of k-mean+ algorithm. It is more appropriate for finding a local optimal solution and is faster compared to k-means. The h-means+ algorithms improve the cluster classification by performing iterations in order to reduce error function. The algorithm includes repeat until loop which is executed repeated until optimal clustering is obtained with non-empty clusters. The two algorithms can be used together referred as two-phase hk-means algorithm.

II. Literature Survey

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy n company strength. Once these things r satisfied, ten next steps is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

A. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models: Computational Statistic and Data Analysis, Vol. 41, pp. 561–575, 2003.

Simple methods to choose sensible starting values for the EM algorithm to get maximum likelihood parameter estimation in mixture models are compared. They are based on random initialization, using a classification EM algorithm (CEM), a Stochastic EM algorithm (SEM) or previous short runs of EM itself. Those initializations are included in a search/run/select strategy which can be compounded by repeating the three steps. They are compared in the context of multivariate Gaussian mixtures on the basis of numerical experiments on both simulated and real data sets in a target number of iterations. The main conclusions of those numerical experiments are the following. The simple random initialization which is probably the most employed way of initiating EM is often outperformed by strategies using CEM, SEM or shorts runs of EM before running EM. Also, it appears that compounding is generally profitable since using a single run of EM can often lead to suboptimal solutions.

B. Unsupervised Learning of Finite Mixture models:Pattern Analysis and Machine Intelligence, IEEE Transactions on pattern analysis. (Volume:24 , Issue: 3)

This paper proposes an unsupervised algorithm for learning a finite mixture model from multivariate data. The adjective "unsupervised" is justified by two properties of the algorithm: 1) it is capable of selecting the number of components and 2) unlike the standard expectation-maximization (EM) algorithm, it does not require careful initialization. The proposed method also avoids another drawback of EM for mixture fitting: the possibility of convergence toward a singular estimate at the boundary of the parameter space. The novelty of our approach is that we do not use a model selection criterion to choose one among a set of pre estimated candidate models; instead, we seamlessly integrate estimation and model selection in a single algorithm. Our technique can be applied to any type of parametric mixture model for which it is possible to write an EM algorithm .These experiments testify for the good performance of our approach.

C. “Mean shift-based clustering”, Pattern Recognition, Vol. 40, pp. 3035-3052, 2007

The mean shift clustering algorithm is a useful tool for clustering numeric data. The shift clustering algorithm for circular data that are directional data on a plane. In this work, we extend the mean shift clustering for directional data on a hyper sphere. The three types of mean shift procedures are considered. With the proposed mean shift clustering for the data on a hyper sphere it is not necessary to give the number of clusters since it can automatically find a final cluster number with good clustering centers. Several numerical examples are used to demonstrate its effectiveness and superiority of the proposed method..

Today's world of digitalization, sensitive and confidential medical records are stored in data centers by healthcare providers across the world. The computation on the medical records can be carried out by third-party service providers. This increases concerns about security and privacy of the sensitive information. The sensitive patient's record must be kept confidential in the healthcare sector. The privacy of the data can be guaranteed, if it is encrypted by the owner of the record before uploading in any clinical sites or in data centers. Only the owner of the record can able to decrypt the data using the private key. This ensures privacy and the security of the data. But, according to the encryption schemes the computation cannot be carried out on the encrypted data, it requires the private key to decrypt the data . Providing private keys to the third-party service provider again results in privacy and security concern. In all the standard encryption schemes, arbitrary computations such as addition or multiplication cannot be carried out on the encrypted data.

` III.EXISTING SYSTEM

Existing System a novel framework of evaluating of personal health records based on classical methods. The Framework addresses the unique challenges brought by multiple PHR Owners and Users. In other research studies, the PHR database fuzzy ARTMAP test set performance were analyzed which had shown similar results to that of the adaptive algorithm with few rules and faster training rate. An ARTMAP pruning algorithm further reduces the number of rules significantly. It also increased accuracy levels. All these techniques are complex, difficult to implement and consumes large computing resources and long training times for converging on expected results.

Disadvantage:

□In a PHR system, there are multiple records who may modify according to their own ways, possibly using different sets of mining methods

□The existing works do not differentiate between the personal and public domains, which have different attribute definitions, key management requirements and scalability issues

□Less accuracy to find out diabetics

IV. PROPOSED SYSTEM

The proposed system uses a novel framework of evaluating of personal health records in cloud dataset. The framework addresses the unique challenges brought by multiple PHR owners and users, in that we greatly reduce the complexity of key management while enhance the privacy guarantees compared with previous works. This work is organized as follows: First it describes data collection methods and introduces the selection criterion of chosen data set for simulation experiments. The evaluation is presented with three types of clustering techniques of EM algorithm, H-means clustering techniques and GA.

Advantage:

- ✓ We focus on the multiple data owner scenario, and divide the users in the PHR system
- ✓ enables dynamic modification of access policies or file attributes, supports efficient on-demand user/attribute revocation
- ✓ Data accuracy is high
- ✓ Quality of evaluation is high
- ✓ Diabetics can be easily found by existing methodology.

V.IMPLEMENTATION

The work is divided in six modules :

1)Data collection

2)Symptoms analysis

3)K-means clustering

4)Genetic Algorithm

5)Classification and Pruning

6)Reports

1.Data collection:

The diabetic's patient's details are collected with help of the data collection through the data entry in the data set. Huge collection of data set is created to evaluate the symptoms. Different type of enquires are made based on the diabetics symptoms.

2. Symptoms analysis

The diabetes diagnosis requires more than just one abnormal blood sugar result. The blood sample results and other basic samples are collected for the symptom evaluation. Use of Symptom Checker to help determine possible causes and treatments.

3.K-means clustering

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells.

4.Genetic Algorithm

In a genetic algorithm, a population of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible

5.Classification and Pruning

The principal objective of clustering computation was to find different groups of diabetes patients with similar symptoms within a group but different symptoms of other groups. The classification was performed based on selected training parameter.

6.Reports

The report on diabetes detection methodology using artificial intelligent tools of data mining techniques. The hypothesis was tested based on test carried out on two types of data sets using algorithms tool used for computation.

VI. CONCLUSION

In the work the diabetic dataset is taken as input and different type of parameters are applied and the data is clustered into groups. The diabetic data is analysed with all the parameters and genetic algorithms is applied and the final diabetic identification in implemented. The results are analysed with the performance and accuracy levels are identified.

REFERENCE

[1] B. Settles, "Active learning literature survey," Dept. Comput. Sci.,Univ. Wisconsin–Madison, Madison, WI, USA, Tech. Rep. 1648, 2010.

[2] D. Pelleg and A. Moore, "Active learning for anomaly and rare-category detection," in Proc. Adv. Neural Inf. Process. Syst., Cambridge, MA, USA, Dec. 2004, pp. 1073–1080.

[3] Z. Qiu, D. J. Miller, B. Stieber, and T. Fair, "Actively learning to distinguish suspicious from innocuous anomalies in a batch of vehicle tracks," Proc. SPIE, vol. 9079, pp. 90790G-1–90790G-11, Jun. 2014.

[4] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in Proc. IEEE Symp. Secur. Privacy, Oakland, CA, USA, May 2010, pp. 305–316.

[5] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," Neural Comput., vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[6] L. M. Manevitz and M. Yousef, "One-class SVMs for document classification," J. Mach. Learn. Res., vol. 2, pp. 139–154, Dec. 2001.

[7] S. Dasgupta and D. Hsu, “Hierarchical sampling for active learning,” in Proc. 25th Int. Conf. Mach. Learn., Helsinki, Finland, Jul. 2008, pp. 208–215.

[8] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[9] K. Das, J. Schneider, and D. B. Neill, “Anomaly pattern detection in categorical datasets,” in Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Las Vegas, NV, USA, Aug. 2008, pp. 169–176.

[10] Z. Qiu, D. J. Miller, and G. Kesidis, “Detecting clusters of anomalies on low-dimensional feature subsets with application to network traffic flow data,” in Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process., Boston, MA, USA, Sep. 2015, pp. 1–6.