International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 4, April - 2014

Survey on Predictive Medical Data Analysis

Using pattern recognition algorithm

Dr. T. Senthil Kumar
Professor Dept. of Computer Science,
Amrita School of Engg.,
Coimbatore, Tamil Nadu, India

Adarsh Suresh Mangalath
Student Amrita School of Engg.,
Coimbatore, Tamil Nadu, India

Kirti Kiron Pai
Student
Amrita School of Engg.,
Coimbatore, Tamil Nadu, India

Priyanka Chinnaswamy
Student
Amrita School of Engg.,
Coimbatore, Tamil Nadu, India

***Abstract*—The medical field is one of the most important industries that can benefit vastly from the many advantages of Cloud Computing and data mining tools. The system we propose combines both domains by enabling doctors, patients, pharmacists, pharmaceutical companies to find hidden trends in medical data (EMRs) stored on a cloud and thus, predict trends for the future. The system is to be supported by a knowledge base that will store recorded data and enhance predictions. The dataset consists of the attributes namely: age, gender, region, climate, time period, diseases and respective diagnosis. A user of the system can find patterns of diseases under one of the categories: age, gender, region and time period. The proposed system explores the application of the k-means algorithm to cluster the data and tests for any modification of the algorithm that maybe required for producing more efficient and accurate results..**

***Keywords*—cloud computing, clustering, data mining, prediction**

I. INTRODUCTION

Even in the twenty-first century, infectious diseases continue to emerge at a rapid pace, in turn causing economic and social disruptions. Methods for enhanced detection, verification, and response capabilities places uniquely three industries—life sciences, food and agriculture, and health care to mitigate the impact of emerging diseases on society. Being able to find patterns in the occurrences of diseases with respect to certain factors like gender, age, change of climate, etc. will help people to be better equipped to take preventive and precautionary measures. This paper undertakes that purpose and tries to propose a system for the same. The aim of the system is to mine interesting and hidden trends of diseases and disorders that dwell in the community by clustering medical records of patients based on a conditional search. A knowledge base consisting of medical data, that is, names of frequent diseases and the causes and preventive measures for them are built. A graphical representation of the mined pattern is also created and the diseases based on the criteria of time period, gender, region and age is enlisted to the user. The stakeholders of this system are pharmaceutical companies, doctors and users or patients in general. To implement our proposed model and evaluate its performance, we use a dataset from a single tertiary hospital. The following sections talk about the significance of the component/technique/technology involved.

A. CLOUD COMPUTING

With the advent of cloud computing, many applications make use of its various benefits such as mobility of data, ease for storage and heavy computing etc. Cloud Computing is germinating its benefits to many industrial sectors in medical scenarios. In Cloud Computing, IT-related capabilities and resources are provided as services, via the distributed computing on demand [2].

Cloud is necessary for this project because large no of EMRs are being used to cluster the various diseases that occur in different ages, genders etc. This increases the computational load, thus, calling in for a system that is scalable and reliable for heavy computing. This makes cloud a necessity as it's easier to store and retrieve large amounts of data in cloud whilst performing intense operations on them. Mining of large data is a tedious job and it includes a lot of computational work which is easier and faster in cloud compared to other methods.

Cloud is a specialized form of distributed computing and we implement this using the Hadoop framework.

Hadoop is an open source cluster-based framework implementing the MapReduce algorithm and is used for running distributed application that processes large amounts of unstructured and structured data. MapReduce algorithm breaks up both the query and the data set into constituent parts-this forms the mapping. The mapped components of the query can be processed simultaneously-or reduced- to rapidly return results[9]. The entire dataset is divided into key-value pairs. The Map function is specified by the user, processes the key-value pairs generating intermediate key-value pairs. After processing, the Reduce function will merge all the intermediate results. The final result is then given to the user.*B. Knowledge Base*

A knowledge base is used in our system to store all necessary information such as known diagnosis for the disease, symptoms, etc. The information is consulted with during the phase of mining and extracting patterns. The knowledge base also stores the extracted patterns.

As elaborated in the work of Jae-Kwon Kimet.al. [1], a knowledge base will contain domain knowledge and important supportive information along with medical guidelines, all which are verified by medical experts. In the above work, a decision-tree rule induction technique creates mining-based rules that are subjected to validation by medical experts. As the rules may not be medically suitable, the experts add rules that have been verified and delete inappropriate rules.

C. Role of Data mining

Mining technique is necessary for our project is because:

- Extracting information on how and what diseases affect different patients from different with the help of EMRs.
- Analyzing EMR's through the several methods bounded along with mining such as Automated discovery of previously unknown pattern.
- On basis of performance, mining is better in gathering and analyzing data faster through parallel processing systems.
- As applying several methods Association rule learning, we are able to identify the pattern within the EMR's and extract interrelationship among the dataset. From that we can predict the future pattern.
- Data mining techniques have recently been attracting attention as a means of enhancing data processing capacity and solving complex problems using computers [1]. Expert systems, which use data mining techniques, help to deal with complex and specialized decision-making issues.

D. Pattern Recognition Algorithm suggested

We employ clustering techniques in our proposed system so as to form clusters of the diseases based on the input conditions. Clustering techniques are unsupervised learning methods for grouping similar data, hence, popularly used for various data mining and pattern recognition purposes. Essentially, clustering is the process of finding nearby points in an n-dimensional space, where each vector represents a point in this space , and each element of a vector represents a dimension in this space.

Fuzzy K-means (also called Fuzzy C-Means or Soft clustering) is an extension of K-means (or Hard Clustering) is the algorithm chosen for the proposed system. K-means algorithm discovers clusters wherein a point belongs to only one cluster whereas Fuzzy K-means discovers soft clusters, that is, a particular point can belong to more than one cluster with certain probability. This is more apt for finding patterns and trends. We explore the use of both the clustering algorithms. K-means algorithm is initially used with clustering various conditions under a single disease, eg: different types of cancer. Then, we progress to explore the use of fuzzy k-means to incorporate all diseases.

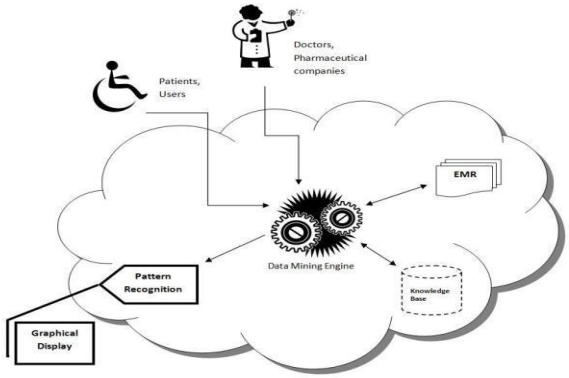
II. RELATED WORK

The literature chosen to support our work is basically derived from two research papers:

We aim to imitate the use of the knowledge base as the knowledge-based Clinical Decision Support System built in the work by Jae-Kwon Kimet.al.[1] executes. This paper proposes the Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model (FbACHD_PSM), which gives content recommendation to coronary heart disease patients. The proposed model uses a mining technique validated by medical experts to provide recommendations.

The aim of the work by N. Karthikeyan .is to provide palm vein pattern recognition based on a medical record retrieval system, using cloud computing. This system was made in consideration for mentally affected, differently abled and unconscious patients who can't communicate about their
medical history to the medical practitioners during an emergency. The paper reveals an efficient means to view, edit or transfer the DICOM images instantly which was a challenging task for medical practitioners, thus making use of the efficiency of cloud.

III. ***System Design***



IV. PLATFORMS AND FRAMEWORKS USED

Using the Apache Mahout framework, we make use of the Machine Learning Library, employing the K-means and Fuzzy K-means algorithm as MapReduce jobs onto a Hadoop framework. Storing our data in HDFS(Hadoop Distributed File System), we run the clustering algorithms on the medical records.

V. DISCUSSION

K-means algorithm is an unsupervised learning algorithm for performing hard clustering, i.e. a vector must belong to one and only one of the clusters. We aim to use k-means on records belonging to a single disease with various conditions, e.g. different types of cancer.

Giving input: set of conditions to cluster, for e.g. age, gender and region, we apply the clustering algorithm.

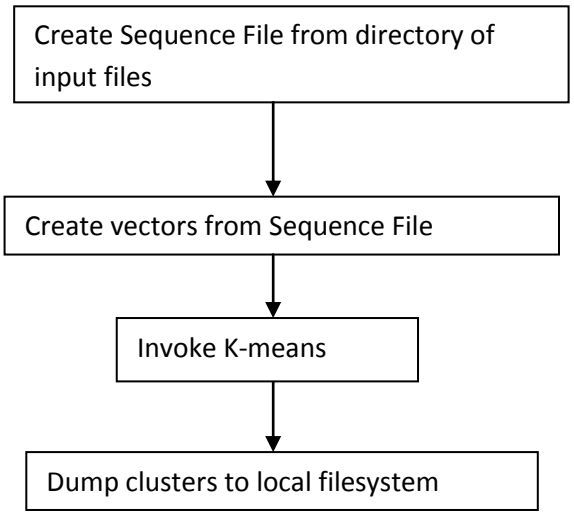
The expected output is clusters of various medical condition types with records matching the chosen criteria of age, gender and region.

We run the algorithm repeatedly giving lesser criteria expecting lesser number of clusters. Thus, we need to find under what conditions does a record probably fall in another cluster, with what probability does it fall into another cluster with a change in the criteria. Here, we must bring in a predictive model that will help decide whether a vector will fall into which cluster with more probability.

. Fuzzy k-means is used in further advancements of the system to include various diseases with various conditions.

It is important to choose the right vector format for the clustering algorithm. We use the SequentialAccessSparseVector for K-means. The output of the clustering algorithm can be read using the Mahout cluster dumper subcommand. We can measure the quality of clusters by measuring the intercluster and intracluster distances.

The general procedure for running k-means algorithm is shown below:



The value of k (number of clusters) is provided by the caller based on the knowledge of the data. We can eliminate this guesswork by using the Canopy clusterer with appropriate distance thresholds to indicate the size of the clusters.

Thus, the system's performance is evaluated and compared with a non-distributed system. This forms the core purpose of the system, i.e, how cloud can be used to improve the computation.

REFERENCES

[1] Jae-Kwon Kim · Jong-Sik Lee · Dong-Kyun Park ·Yong-Soo Lim · Young-Ho Lee · Eun-Young Jung , "Adaptive mining prediction model for content recommendation to coronary heart disease patients" ,August 2013.

[2] N. Karthikeyan and R. Sukanesh, "Cloud Based Emergency Health Care Information Service in India", 27 July 2012 / Published online: 3 August 2012.

[3] Chen Xu,"Biomimetic pattern recognition,Topological pattern recognition",July 20.

[4] Ji-Wen Chio and Shu-Yuan Chen ,"Legend Extraction From E-Learning Video Streams;pattern recognition", August 2011.

[5] Jackson K.R. , Ramakrishnan, L . Muriki , Canon S , Cholia.S., Shalf.J. ,Wasserman and Harvey J., "Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud", 2010.

[6] Gaizhen Yang ,Zemin Zhu and Fen Zhu, "The Application of SaaS-Based Cloud Computing in the University Research and Teaching Platform" ,August 2011

[7] UmaMaheswari.P and Rajaram, M , " A novel approach for mining Association rules on sports data using principle component analysis", 2009.

[8] Saraee M.H , Isfahan Ehghaghi. Z , Meamarzadeh, H., "Applying Data Mining in medical data with focus on mortality related to accident in children", 2008.

[9] Chuck Lam , "Hadoop in Action".

[10] sujitpal.blogspot.in/2012/09/learning-mahout-clustering.html

VI. LITERATURE SURVEY

Author	Method/ Algorithm	Merit	De-merit
Jae-Kwon Kim; Jong-Sik Lee ; Dong-Kyun Park ; et al.	Proposes the Fuzzy Rule-based Adaptive Coronary Heart Disease Prediction Support Model (FbACHD_PSM), which gives content recommendation to coronary heart disease patients. The proposed model uses a mining technique validated by medical experts to provide recommendations.	To increase its prediction accuracy, knowledge and mining-based rules were generated, and the decision treetechnique was used. treetechnique was used.	Not scalable to a large dataset.
N.Karthikeyan and R.Sukanesh	With software as a service (SaaS) by means of Cloud computing, it aims to bring emergency health care sector in an umbrella with physical secured patient records. Also giving a ubiquitous access to appropriate records using Palm vein pattern recognition	It reveals an efficient means to view, edit or transfer the DICOM images. Merits of cloud computing well supported.	
Huang Jun Ni2, Yuanyuan Dan3, Sen Xu4	This paper presents a skewed gene selection algorithm that introduces a weighted metric into the gene selection procedure.	Classification performance is improved and over fitting is avoided by combining multiple feature gene pairs (decision rules) into an ensemble learning framework.	Mining of the decision rules is quite time consuming
Rajaram and UmaMaheswari	Association using principal component Analysis	The efficiency of mining algorithm is improved provided that Principal Component Analysis generates frequent patterns.	Experiment outcomes may not be fully accurate

Table 1. Related Work in Cloud Computing, Data Mining and Pattern Recognition

About Author (s):

T.Senthil kumar completed his B.Tech(Computer Science and Engineering) from Sethu Institute of Technology, Madurai in 1999. He then completed his M.Tech(Distributed Computing Systems) from Pondicherry Engineering college, Pondicherry . He completed his Ph.D in Information and Communication Engineering from Anna University, Coimbatore. He has around 12 years of teaching experience and 2 year of industry Experience. His area of interest include cloud computing, software Engineering, Video processing, Wireless Sensor Networks , Dot Net Programming , JIST simulator, Data Mining. He is currently working as a Assistant Professor(Selection Grade) in computer science and Engineering Department at Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore. He has publication in 10 National Conferences and 6 International Conferences and six International Journals.

Adarsh Suresh is pursuing B.Tech final year in Amrita University. His areas of interest are Cloud computing, Data Mining, Algorithms

Kirti Kiron Pai is pursuing B.Tech final year in Amrita University. Her areas of interest are Cloud computing, Data bases.

Priyanka Chinnaswamy is pursuing B.Tech final year in Amrita University. Her areas of interest are Cloud computing, DBMS, Data Mining.