National Conference on Communication and Image Processing (NCCIP- 2017)
3rd National Conference by TJIT, Bangalore

International Journal of Science, Engineering and Technology

An Open Access Journal

Design of the Research Platform for Medical Information Analysis and Data Mining

Sanjana B Pai, Prof. Veena N.

Abstract

A research platform for medical information analysis and data mining is designed. Based on the OpenStack, a cloud computing platform management was built which can realize virtual distribution and management. It also could improve the resource utilization and reduce the requirements for the personal work PC. Moreover, the application of the Hadoop technology provided a better solution for medical data mining and analysis, as a task can be broken into small sub-tasks. This could enable the computing speed of big data. In addition, the platform integrated some classical popular algorithms and it also included a medical database that is useful for the researchers to learn and practice. Modern medicine generates a great deal of information stored in the medical database. Extracting useful knowledge and providing scientific decision-making for the diagnosis and treatment of disease Medical information, data mining, OpenStack,

Keywords: Hadoop, medical image, Neural Networks, Diagnosis

Introduction

Data mining is the key link of knowledge discovery. It is a process of finding effective, novel, useful and ultimately understandable knowledge in a large number of data. Hence, the knowledge that has a special relationship but be hidden in the information usually can be automatically searched from a large amount of data . In this process, we often use the various disease symptoms as the condition attributes and use the diagnosis results as the decision attributes. Then the value of medical diagnosis knowledge can be obtained by data mining for medical decision table. Consequently, a scientific research platform for medical information analysis and mining based on cloud was introduced. In addition, the platform is also can be used for autonomous learning. The system adopted the browser/server architecture. Existing medical system includes hospital management system and decision making system. The focus is on collecting and mining the entire medical Data. Through the virtual machine technology, the researchers can process the big data anywhere just by make simple configuration of the PC machine. And besides, the problems of insufficient resources and long operation time problem also need not to be worried about when processing a super big data.

Heoretical Framework Of Data Mining And Its Development

Basic Process of Data Mining

Data mining generally has the following steps: data collection, collation, mining, mining results evaluation, analysis decision. It takes a cycle of repeated processes to achieve the desired effect. Specifically, in different areas of application, it has its unique nature : first, understanding of the problem in the field of significance, determine the target and success criteria; second, understanding of data; third, understanding of the data, data warehouse -a dynamic process; Fourth, data mining, including data model selection, training and verification process, modelling and model quality evaluation, for the same process can be used by different algorithms, this is only the data of different understanding, each algorithm has a reasonable probability, the actual use of the comparison; fifth, the results of the assessment, the extraction of new knowledge and reasonable interpretation, and need to be understood and have certain application value.

Knowledge Discovery Framework

It presents a framework represented in Fig.1, which generally explains how to lead raw data to discover knowledge. First step is data collection from EHRs on its private and public forms the next process is to create Clinical Datasets (CDSs) from the collected data. Clinical datasets is a warehouse that builds for the purpose of analysis, mining and reporting. Clean CDSs and extract the important data is the next step, Cleansing the raw big data its important stage by using data cleaning algorithms; in other words detect and remove errors or delete duplication or whatever process to improve the quality of data and to avoid processing time consuming. Clinical data warehouse (CDW) is a repository that is collected from multiple clinical software packages and it used to link, analyze, report and search within stored medical information. CDW aimed to provide health investigator by obtaining data from enormous clinical systems. Those Clinical Systems used in health related organizations.

Then this data have to be categorized according to stakeholder's perspectives, this process called information retrieval or data mining level. To get information cleansed big data can be categorized into statistical information or analytical information or tactical information or any different category that aid senior managers and decision maker to do exact decisions. The analysis procedures needs a health plans with a comprehensive view from patients, doctors, physicians, pharmacists, technicians and other intended people.

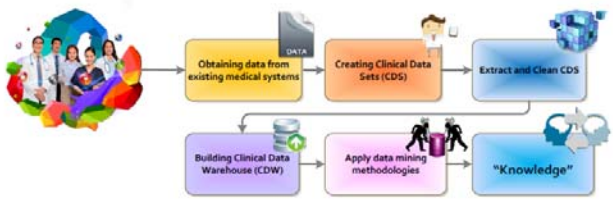
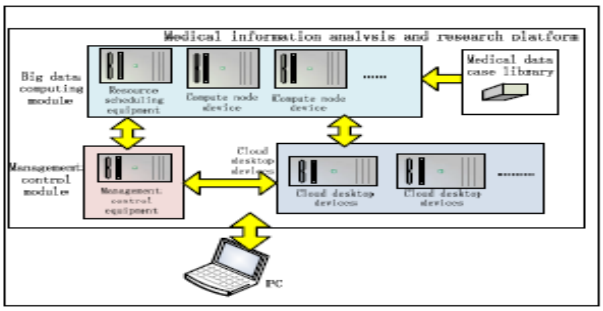


Fig 1- Knowledge discovery through Clinical Big data analysis proposed framework

OVERALL SYSTEM DESIGN

The block diagram of the research platform for medical big data analysis and data mining is composed of four modules: a big data computing module, a management control module, a cloud desktop module, and a medical data library. The big data computing module is constructed in a virtual machine, using the open source ecosystem. So it is very convenient to process and compute big data. The manage control module is a B/S system. It has two main modules. First, it is in charge of the resource scheduling and allocation tasks. Second, it provides a graphical interface and it is also the only entrance to operate the system for the users. The cloud desktop is built based on the OpenStack and some of its components included nova, swift and neutron are used here. Besides, on the basic, a custom management API is also designed. By the cloud desktop devices, we can open a virtual machine to use for the application of algorithms.

Fig 2- Medical Information Analysis



PREPROCESSING OF MEDICAL INFORMATION

In the clinical medical, it involves many medical information. It consists of pure data (such as sign parameters and test results), signals (e.g., EMG, EEG, etc.), medical images (such as the detection results of B ultrasound, CT, and other medical imaging equipment), texts (such as the identity of the patient record and the description of symptom, detection and diagnosis results of textual representation) and other attributes of information.

The medical information mainly has the following four aspects of characteristics:

1. Polymorphism of patterns
2. Incomplete
3. Time
4. Redundancy

Different technologies and measures should be adopted to deal with the medical data that has

different physical properties. By the process, it will make them converge on the attribute or be consistent with the results of the treatment. In addition, the same kind of information sources of different objects may also be different that should also be pre-processed before data mining. The process referred can be called as "the normalization process of medical information".The basic process of the normalization of medical information is shown as in Fig 3.

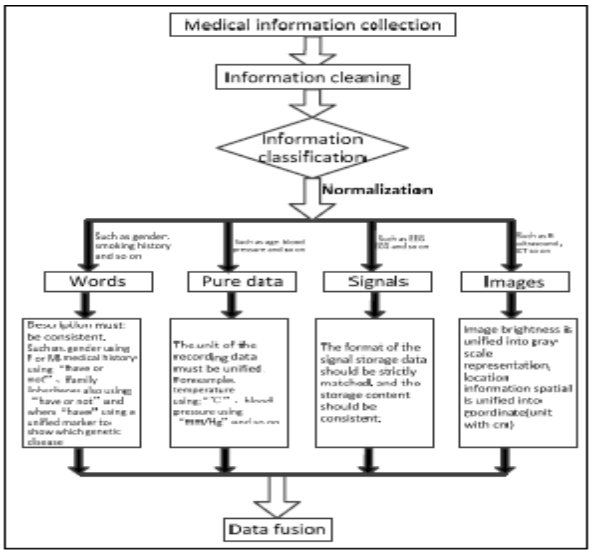


Fig 3- Normalisation of data

ARCHITECTURE OF THE PLATFORM

The platform architecture is shown in Fig 4, both Hadoop cluster and its manager are built in the server based on the OpenStack. By the web server, the Hadoop manager can unify schedule and allocate resources while the Web server sending the information to the OpenStack manager in real time. For the research users, they should first apply for the use of the virtual machines through the UI port and then access the virtual machine by the cloud desktop. The cloud desktop is directly connected to the Hadoop manager. And the Hadoop manager can finish the task of the specific allocation of resources and computing then.

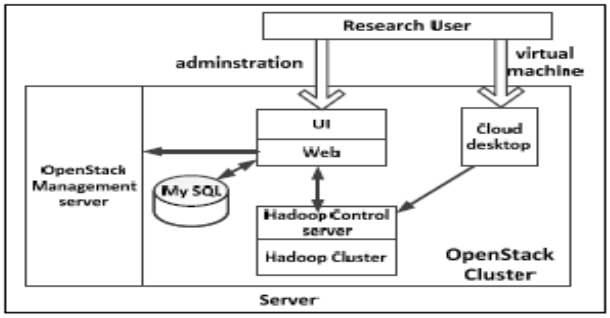


Fig 4- Architecture of the Platform

APPLICATIONS OF DATA MINING IN MEDICAL FIELD

A. Application of data mining in the analysis of physiological parameters of patients

Because of the complexity of the mechanism of medical physiological parameters, it is far higher than the current level of basic theory. It is not enough to quantitatively monitor and describe the relationship between high blood glucose concentration and some complications. The researchers began to establish the physiological data base of diabetes patients, and analyze the physiological parameters such as blood glucose concentration, age, sex, bone density, electrocardiogram, blood pressure, muscle fat content and so on. Data mining theory in the multidimensional data association analysis technology to discover the hidden knowledge in the data. There is a certain regularity between the values of two or more variables, which is called Association. The association can be divided into simple association, temporal association and causal relationship.

B. Healthcare Big Data Applications

1. Provides Advanced Medical Review.
2. Save time for providers in making informed medical and business decisions.
3. Facilitate healthcare analysis.
4. Measuring the clinical and also, healthcare organisations' performance.
5. Assist stakeholders to determine the level of medical decision making.

C. Application of the Platform

Each user has its own set of independent virtual desktop system based on the remote hosted virtual desktop. Each user can share applications with other users, but keep isolate with each other and enhance security. With this mode, the platform improves the utilization of resources by resource sharing and allocation on demand and facilitates learning and scientific research.On the platform, the users can learn and practice big data process using the popular tools. Also, they can process the medical big data by submitting the medical information after pre-processed. Besides, they can do researches on how to optimize the existing algorithms.

EFFICIENCY OF THE PLATFORM

The platform proposed here is convenient for the researchers to fast work and process data. By using the platform, the research institutions are no longer need to repeat construction data platform. The resources can be saved and the cost can be reduced. Besides, the platform also has training functions, and can train the medical personnel to use big data processing tools.

ISSUES AND CHALLENGES

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor's intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

CONCLUSION

A research learning platform for medical information analysis and mining was designed in the report. The platform for medical researchers provided a study platform using data analysis and mining suitable for cloud environment based on the OpenStack and the Hadoop technology. The platform mainly solved two major problems: first, by integrating the most popular analytical mining tools, the researchers can be quickly familiar with a variety of algorithms and then choose an applicable analysis tool by comparing study. Second, as the resource is provided by the cloud, the configuration for the personal working platform can be simplified and the cost for device can be reduced. Healthcare online systems could help to improve the communication between doctors and patients on the other hand to improve the quality of care which may lead to reduce medical errors and costs. In order to get better reviews patients in interest with share, save, manage, and retrieve their medical data, such as their medical history, medications, allergies, x-rays and test results. Accordingly building these online big repositories give them an opportunity to interact with doctors, physicians and pharmacists, but IT experts should take in mind patents' privacy and polices risk. To be concluded that big data is about real time data. Healthcare management follow three main and important stages the first stage is data collection: this data may come from EHRs, hospitals or clinical billing systems, lab and imaging systems or any data attached with patients.

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible. So, with gratitude, I acknowledge all those whose guidance and encouragement crowned my effort with success.

It is my pleasant duty to place on record my deepest sense of gratitude to my guide **Prof. VEENA N,** Assistant Professor, for the constant encouragement, valuable help and assistance in every possible way.

References

[1]Lijun Hao, Shumin Jiang, Boyu Si, Baodan Bai Shanghai University of Medicine and Health Sciences(2016), Design of the Research Platform for Medical Information Analysis and Data Mining.

[2] Deren Li, Wenzhong Shi, Shuliang Wang and Xinzhou Wang, "On spatial data mining and knowledge discovery," Geomatics and Information Science of Wuhan University.

[3]Yijun Song and Ming Zhang, "Exploration of platform construction of unified virtualization teaching environment", Experimental Technology and Management, vol.33,pp:115-118,April 2016.

[4]DaliaAbdul, HadiAbdulAmeer University Of Information Technology and Communications – Iraq, Medical Data Mining: Health Care Knowledge

Author's details

SANJANA B PAI,Student, Department of ISE,BMS Institute of Technology,Yelahanka,Bangalore

Sanjana3295@gmail.com

Under the guidance of:

Mrs. VEENA N, Assistant Professor, Department of ISE, BMS Institute of Technology, Yelahanka, Bangalore.

veena@bmsit.in