A Review On Different Prediction based model for cloud-based data mining

[1]Er. Spinder Kaur, [2]Dr. Sandeep Kautish
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda,Punjab India.
Department of Computer Engineering
Guru Kashi University
Talwandi Sabo Bathinda ,Punjab India.

Abstract- Cloud-based healthcare data mining framework. Under such framework, various cloud based healthcare data mining services can be developed, deployed, and provisioned to the general healthcare industry for knowledge discovery and decision-making support. In our proposed framework, 1) population-level healthcare data scattered across disparate local data sources are integrated, which provides abundant data for the data mining process; 2) computational infrastructure and resources can be delivered by cloud computing platforms in a reliable, scalable, and cost-effective manner, which satisfies the computational and financial requirement for building healthcare data mining services; 3) the service development process is modularized, which makes the service development, update, and maintenance easier and faster; 4) the healthcare data mining services are deployed and provisioned to the healthcare practitioners as either cloud applications or web services, which ensures high service accessibility.

Keywords- FLOS, Cloud, Random Forest, Gradient Boosting.

I. INTRODUCTION

In recent time various types of services are emerging in the society. These services are related to the different fields of the society. Out of those fields major field is medical. India is a large country where large population resides. Various types of organized and unorganized medical facilities are available in the country. But due to the population explosion each facility remain scarce. To overcome and catalyst the growth in this part of the applications. Various researchers are involved which are growing with different researches so that the problem of scarcity of the resources can be catered without increasing the much cost.

In this researches cloud is one of the major thrust area. Which can solve the problem of this medical mismanagement. According to this research paper there will be a cloud of the medical data. Which is consisting of integration of various small city level data. Any company will provides processing ability which can process this large integration of the data. So that any patient data can be provided at the required place. His case history or we can say medical history can be recorded at each step.

1.1 Introduction to Data mining

Data Mining is a process to analyzing the data from large databases. As it is also clear from its name Data Mining **:**"searching for valuable information in a large database". Data mining is also known as knowledge discovery.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase

Importance of Data Mining

We can simply define data mining as a process that involves searching, collecting, filtering and analyzing the data. It is important to understand that this is not the standard or accepted definition. But the above definition caters for the whole process. Large amount of data can be retrieved from various websites and databases. It can be retrieved in form of data relationships, co-relations and patterns. With the advent of computers, internet and large databases it is possible collect large amounts of data. The data collected may be analyzed steadily and help identify relationships and find solutions to the existing problems. Governments, private companies, large organizations and all businesses are after large volume of data collection for the purposes of business and research development. The data collected can be stored for future use. Storage of information is quite important whenever it is required. It is important to note that it may take long time for finding and searching for information from websites, databases and other internet sources.

1.2 How does Data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

1.3 Data mining Techniques

a) Classification

Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a

training set consisting of pre-classified examples. The task is to build a model that can be applied to unclassified data in order to classify it. Examples of classification tasks include:
•Classification of credit applicants as low, medium or high risk
•Classification of mushrooms as edible or poisonous
• Determination of which home telephone lines are used for internet access
is the task of segmenting a diverse group into a number of similar subgroups or clusters. What distinguishes from classification is that does not rely on predefined classes. In , there are no predefined classes. The records are grouped together on the basis of self similarity. is often done as a prelude to some other form of data mining or modeling. For example, might be the first step in a market segmentation effort, instead of trying to come up with a one-size-fits-all rule for determining what kind of promotion works best for each cluster.

b) Association Rules

An association rule is a rule which implies certain association relationships among a set of objects (such as "occur together" or "one implies the other") in a database. Given a set of transactions, where each transaction is a set of literals (called items), an association rule is an expression of the form XY, where X and Y are sets of items. The intuitive meaning of such a rule is that transactions of the database which contain X tend to contain Y. An example of an association rule is: 30% of farmers that grow wheat also grow pulses; 2% of all farmers grow both of these items". Here 30% is called the confidence of the rule, and 2% the support of the rule. The problem is to fund all association rule that satisfy user-specified minimum support and minimum confidence constraints.

c) Regression

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques. The simplest form of regression, linear regression, uses the formula of a straight line (y = mx + b) and determines the appropriate values for m and b to predict the value of y based upon a given value of x.The regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.

II. LITERATURE SURVEY

[1] Peng Zhang(2016) et al: in this paper the research has been undertaken for creating cloud of the healthcare data. So that it can be integrated from different place. With the help of processing ability and cloud services different types of patient data be processed to generate prediction based scenario.

[2] Nima Jafari Navimipour(2016) et al: in current research Replica selection requires information about the capabilities and performance characteristics of a storage system. It is based on the user demand and failure occurs during response time. In data cloud, the selection of replica is an important issue for users and to access a data file. There research is mainly focused on replica selection mechanism in order to achieve the best performance. This research proposes new replica selection base on ant colony optimization to improve average access time.

[8] E.K. Burke et al.discussed automatic timetable generation with the use of traditional methods such as graph coloring and advanced methods such as the genetic algorithms. This paper presents the examination timetabling. This paper discussed Genetic algorithm is very useful general purpose optimization tools that may be applied to wide range of very difficult problems

[9] KhaledMahar proposed genetic algorithm has a simple representation that handles all the university timetables at once and easily modified to creation of a accurate timetable which satisfies constraints that must not be broken. This paper presents that algorithm is applied to create timetables for the college of Arab Academy for Science and Technology in Egypt and the results are very satisfactory and there is no hard constraint violation encountered. The program tested with different population sizes, a crossover and mutation rates. This paper also provides an overview of different techniques for automatic generation of university time tables like tabu search, simulated annealing, genetic algorithms, graph coloring heuristics, constraint programming, network flow models, and constraint programming .This papers proves that as long as population size increases the cost changes faster and large size takes too much running time and memory consumption.

[10] DiptiSrinivasan et al. stated an evolutionary algorithm based approach to solving a large constrained university timetabling problem. Other techniques also used for obtaining feasible timetables in a appropriate time that are Heuristics and context-based reasoning. The complete course timetabling system presented in this paper has been accurate, tested and discussed using data from a university. The results have shown that implementing the intelligent adaptive mutation operator has led a more than 10 times of improvement in the performance of evolutionary algorithm.

III. COMPARATIVE TABLE

Author,Year,Title	Technique	Constraints
D. Sumathi, P. Poongodi,2015, An Improved Scheduling Strategy in Cloud Using Trust Based Mechanism	Trust based scheduling to improve cloud security by modifying Heterogeneous Earliest Finish Time (HEFT) algorithm	This paper has not considered any prediction based system for cloud data scheduling.
Seny Kamara,2010, Cryptographic Cloud Storage	This paper has focused on the cloud data security. So that no illegal access can be taken place while data is being shared by multiple persons.	there is no data prediction for security threats based on previous attacks performed on to the cloud data.
John A. Doucette, Atif Khan,2011, A Framework for AI-Based Clinical Decision Support that is Patient-Centric and Evidence-Based.	This paper is focused on prediction based architecture where clinical data is used for patient centric prediction.	This technique only uses evidence based data. Have not considered any hypothesis.

Dimitris IAKOVIDIS,2012, A Semantic Model for Multimodal Data Mining in Healthcare Information Systems	this paper has used the multi modal based technique for data mining in Healthcare information systems. So that data can be used for patients for prediction purposes.	This paper requires the further improvement in the prediction for information centric architecture for data processing and prediction.
Francisco Bermudo Guitarte, 2017 ,Integrating Electronic Systems for Requesting Clinical Laboratory Test into Digital Clinical Records: Design and Implementation	This paper is based on the technique for building a system where electronics records are kept for laboratory data. At any time user can request for the data access.	This requires further security of data to share the private data amongst two requesting persons.

IV. CONCLUSION

The study of various research papers has been performed while doing the research work for prediction based system. Various techniques are being done with different levels of success rates. Random forest and the Gradient Boosting based techniques are the best technique with better accuracy. So that the prediction in various strategic fields can be done with higher levels of accuracy. This type of procedure uses the large data streams. With the increase of the data stream size the accuracy of the both techniques is increased. In random forest the tree based on splitting the dataset values. This process of splitting is based for identifying the left sub tree members and then the right sub tree members. In Gradient Boosting technique the loss factor value will be reduced at each iteration. This technique is based on optimization of prediction based on iteratively increasing the dataset size.

V. FUTURE WORK

From the current review of various research papers it is clear that the prediction based system is required most in health care system. It will be helpful in reduction of the cost and also used for better treatment to the patients. Gradient Boosting Machine Based technique can be used for prediction of Future length of stay for the patient in the hospital.

REFERENCES

[1] Nima Jafari Navimipour, Bahareh Alami Milani "Replica selection in the cloud environments using an ant colony algorithm " 2016 IEEE

[2] Peng Zhang, Shang Hu, Jing He, Yanchun Zhang, Guangyan Huang, Jiekui Zhang ," building cloud-based healthcare data mining Services", 2016 IEEE International Conference on Services Computing.

[3]Baden, R., A. Bender, N. Spring, B. Bhattacharjee andD. Starin, 2009. Persona: An online social networkwith user-defined privacy. Proceedings of the ACMSIGCOMM 2009

Conference on DataCommunication, Aug. 16-21, ACM Press, NewYork, pp: 135-146. DOI: 0.1145/1592568.1592585

[4]Fong, P.K. and J.H. Weber-Jahnke, 2012.Privacypreserving decision tree learning using unrealizeddata sets. IEEE Trans. Knowl. Data Eng., 24: 353-364. DOI: 10.1109/TKDE.2010.226

[5]Isdal, T., M. Piatek, A. Krishnamurthy and T. Anderson,2010.Privacy-preserving P2P data sharing withOneSwarm. Proceedings of the ACM SIGCOMM2010 Conference, Aug. 30-Sep. 03, ACM Press, NewYork, pp: 111-122. DOI: 10.1145/1851182.1851198

[6] Hayrinen K., Saranto K., & Nykanen P. (2008). Definition, structure, content, use and impacts of electronic health records: a review of the

research literature. International journal of medical informatics, 77(5), 291-304.

[7] Khan A., Doucette J., Jin C., Fu L., & Cohen R. (2011, April). An ontological approach to data mining for emergency medicine. In 2011 Northeast Decision Sciences Institute Conference Proceedings 40th Annual Meeting. Montreal, Quebec, Canada (pp. 578-594).

[8] Iakovidis D., & Smailis C. (2012). A semantic model for multimodal data mining in healthcare information systems. Stud Health Technol Inform, 180, 574-578.

[9] Kamara S., & Lauter K. (2010). Cryptographic cloud storage. In Financial Cryptography and Data Security (pp. 136-149). Springer Berlin Heidelberg.

[10] Hamlen K., Kantarcioglu M., Khan L., & Thuraisingham B. (2010). Security issues for cloud computing. International Journal of Information Security and Privacy (IJISP), 4(2), 36-48.

[11]Kagal, L. and J. Pato, 2010. Preserving Privacy basedon semantic policy tools. Security Privacy IEEE,8: 25-30. DOI: 10.1109/MSP.2010.89Lin, K.P., 2011. On the design and analysis of the

[12]privacy-preserving SVM classifier. Proceedings ofthe IEEE Transactions on Knowledge and DataEngineering, (TKDE' 11), IEEE Xplore Press, pp:1704-1717. DOI: 10.1109/TKDE.2010.193Lu, R., X. Lin and X. Shen, 2012. SPOC:

[13]A secure andprivacy-preserving opportunistic computingframework for mobile-healthcare emergency.IEEETrans. Parallel Distrib. Syst. DOI:10.1109/TPDS.2012.146

[14] Alberto Colorni, Marco Dorigo, Vittorio Maniezzo, "A Genetic Algorithm to Solve theTimetable Problem", Centre for Emergent Computing, Napier University, Edinburgh EH105DT, UK,2000.

[15] E.K.Burke, D.G.Elliman, R.F.Weare, "The Automation of the Timetabling Process in Higher Education".