Data Analysis using Hierarchical Computing

Aliasger Kagalwala, Ankush Khurana, Shashwat Kaul
[1,2,3]UG Student, Dept. of Computer Engineering, M.E.S.
College of Engineerig, Pune

Govind Pole
Assistant Professor, Dept. of Computer Engineering, M.E.S.
College of Engineerig, Pune

Abstract:Supervised learning algorithm can be used to mine datasets on the internet. Stock market, Medical organizations, education institutes all store a huge amount of data .For the purpose of analyzing this data classification algorithms can be used. The processing of these algorithms can be done using eithera single machine either sequentially or parallel or on multiple machine either using 1)Parallel approach 2)Cloud approach 3)Hierarchical approach.

Keywords*: machine learning; supervised learning; parallel computing; cloud computing;*

*****

I. Introduction

In today's age of technology a huge amount of data is being produced daily. To analyse this data lot of computation power is needed. This can be costly. Small companies or organizations may not be able to afford it. Also to choose the best machine learning algorithm can take a lot of resources. With our proposed system one can use the available resources in an organization and convert them into a master-slave configuration and then divide the data on this network, the divided data set then goes through three different classification algorithms which then produces the result.

II. Related work

Data mining is used for examining pre-existing databases in order to generate new information, to analyse behaviour, pattern , to do certain surveysand much more. Data mining is achievable by using supervised machine learning algorithms like in [1] author Rui Rui and Changchun Bao has applied supervised machine learning to classify musical instruments by creating a musical retrieval system. It can also be applied on signal processing for this system[2][3],Supervised learning algorithms can also be applied in sleep monitoring for posture reorganization[4],and in predicting bankruptcy[5].

Processing of this data requires a lot of computational power. For this purpose MLaaS(machine learning as a service)[6] can be used. The service can then be used by either a single user or multiple users. This platform can run on single pc based architecture or multicomputer based architecture.

Hierarchical architecture has many advantages over a single system implementing supervised learning algorithms. By distributing the data on mobile devices for independent remote execution [7] the overall performance of the algorithm can significantly be improved. In [8] authors Pengcheng Shen, Xin Du, Chunguang Li used distributed computing for metric learning on data under pervasive constraints.

Cloud Computing is another platform on which data mining algorithms can be implemented .Encrypting algorithm can be used to securely divide datasets based on attributes . A good example is proposed by Deepti Mittal, Damandeep Kaur, Ashish Aggarwal in [9].

III.Proposed architecture

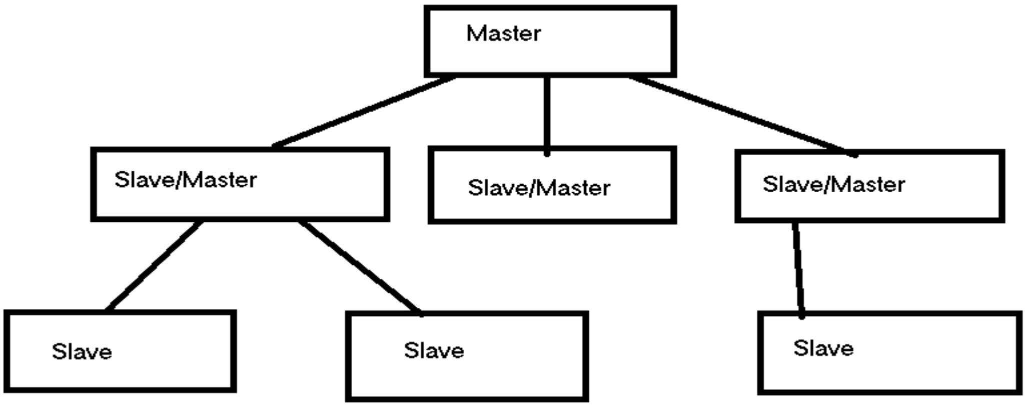


Fig. Master-Slave Hierarchical Architecture

The Architecture is a Master-Slave Hierarchical architecture in which Master is responsible for distributing the workload between the slaves. In our architecture the slaves can further act as masters and distribute workload among their slaves. The architecture uses 3 supervised learning algorithms namely ID3, Naïve Bayes and FP Growth. Supervised learning provides the system with training sets and trains the system to correctly classify the test data set.

Our architecture provides a cost effective solution for data analytics for small organizations and hospitals. The main idea is to utilize the resources and computational power of idle computers that is usually wasted and perform analysis on data instead of buying expensive hardware or outsourcing it to a third party.

The detailed working of the system is as follows:

1. Workload is submitted to master.
2. Master checks the available slaves and distributes workload among them evenly.
3. Slaves fetches the data and checks for sub-slaves. If present the slave distributes work among sub-slaves.
4. Slaves and Sub-slaves perform analysis on data and the result is accumulated at master.

IV. ALGORITHMS USED

A. ID3

The ID3 algorithm works by creating a decision tree. For this purpose it calculates the information gain to find the splitting attribute. The attribute with highest information gain is selected as splitting attribute for set S. This is iterated for each unvisited attribute. At each level the untouched attribute is explored. The non-terminal attributes are the attributes which are under consideration for decision tree and terminal nodes denote class labels

B. NAÏVE BAYES

Naive Bayes is classifier constructing technique based on probability prediction that helps to assign class labels to unclassified data sets. The important consideration for Naïve Bayes classifier is that it assumes each feature to be independent. For example a fruit is considered as mango if it is yellow, elliptical and around 10cm in size. These features contribute to the uniqueness of the fruit independently. Naïve Bayes is a Supervised Learning algorithm in which the classifier is trained using training set and tested against test set.

C. FP Growth

The FP- Growth is a faster alternative to apriori algorithm for finding frequent itemsets. It does this without generating candidate sets. It uses divide and conquer strategy. It uses a data structure named FP-Tree for holding the association information. FP-Growth works in two phases. In First phase it generates FP-tree using the compressed database and in second phase it mine frequent itemsets iterating through the patterns.

V. ADVANTAGES

- **Faster Execution**

As the architecture uses master-slave approach the execution becomes faster by utilizing the computational power of every machine in the organization.

- **Better Resource Consumption**

The architecture uses the resources already available in the organization. This ensures that no machine is idle and its computational power is used to its full potential. This helps in achieving maximum throughput.

- **Cuts Processing Costs**

No need for the organization to purchase heavy processing units for analysing the data or outsourcing the data to other organization. All the analysis can be done within the organization using the existing resources.

VI. DISADVANTAGES

- **Socket Stream Corruption**

In the case where the socket is loaded with data more than its capacity the stream can get corrupted.
This can halt the execution of the analysis or a_ect the accuracy of the results.

- **Abnormal Slave/s Shutdown:**

If the Slave/s shutdown abnormally due to power loss or system failure, it can halt the execution and also affect the load balancing.

VII. APPLICATIONS

- **Banking System:**

Data mining is essential part of banking system using machine learning algorithm a lot of useful information can be produced. This information can be used to find customers transaction patterns, different policy analysis, to find patterns in rise and drop of stocks and much more. By using suitable machine learning algorithms bankruptcy can be predicted. If a investor is investing in stocks of a company then it is important for them to know if the current stock rate will lead to company going bankrupt or not. Mining of previous record might generate a prediction model for determine the probability of if a company will go bankrupt or not.

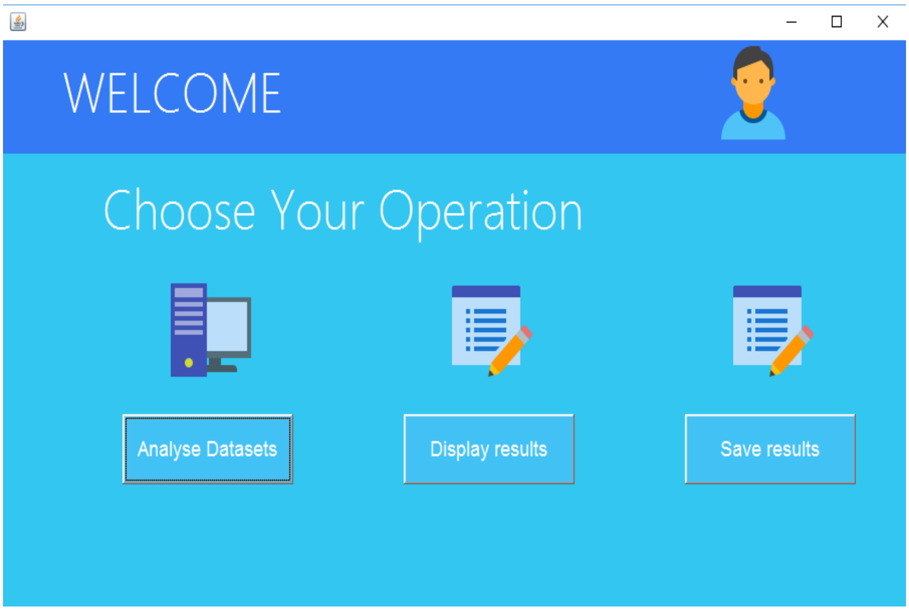
- **Health:**

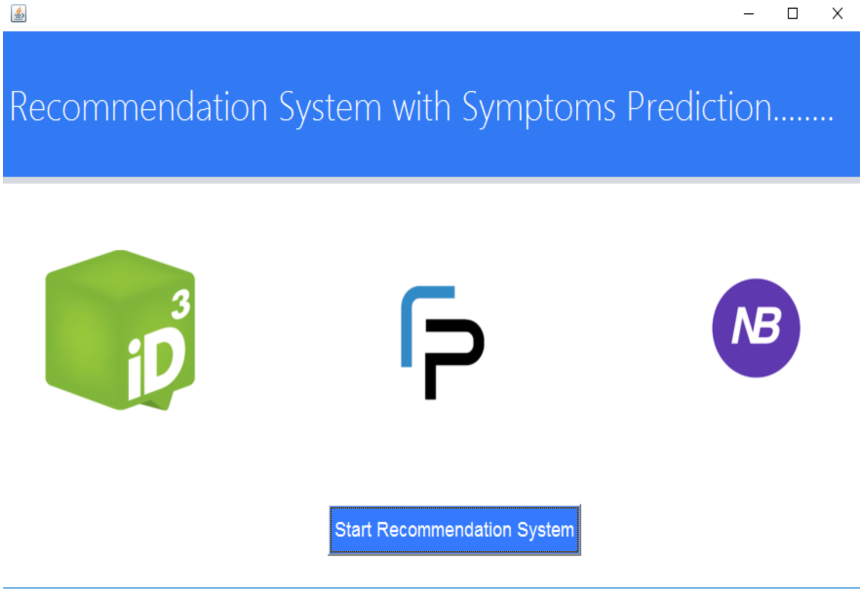
Machine learning algorithms are used in many health-based application for .e.g. fitness bands uses them to monitor users body condition, diet, to monitor its exercising pattern and generate pictorial result for analysis. In sleep monitoring for posture reorganization supervised learning can be efficiently applied. Pressure sensors can be deployed in mattresses under the person’s body to acquire body pressure distribution .The data collected is sent to the main computer for processing and the result are communicated for monitoring and diagnostics.

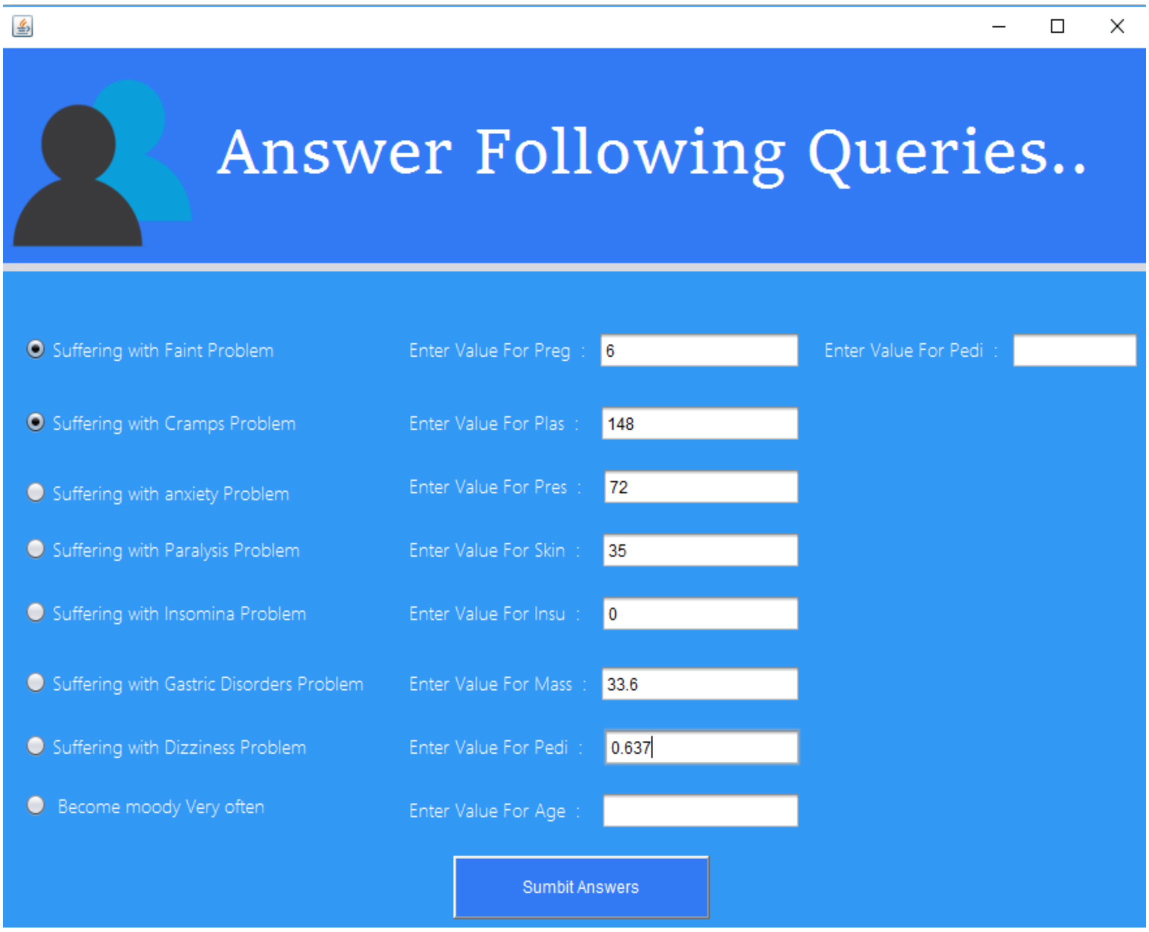
- **IOT:**

Data collected from sensors is huge. To extract meaningful information from these sensors machine learning is used. Sensor continuously produces data for e.g. a temperature detecting sensor will continuously send temperature to the data sets this data can be used to _nd whether of a particular region. It can also be used to measure rainfall, and sea level. All this information can be then used to analyse rain patterns. And may even be able to inform about danger in advance. Machine Learning when applied with IOT can have endless possibilities and can be used to make our day to day life better.

VIII. SYSTEM IMPLEMENTATION




Recommendation System with Symptoms Prediction........
Start Recommendation System

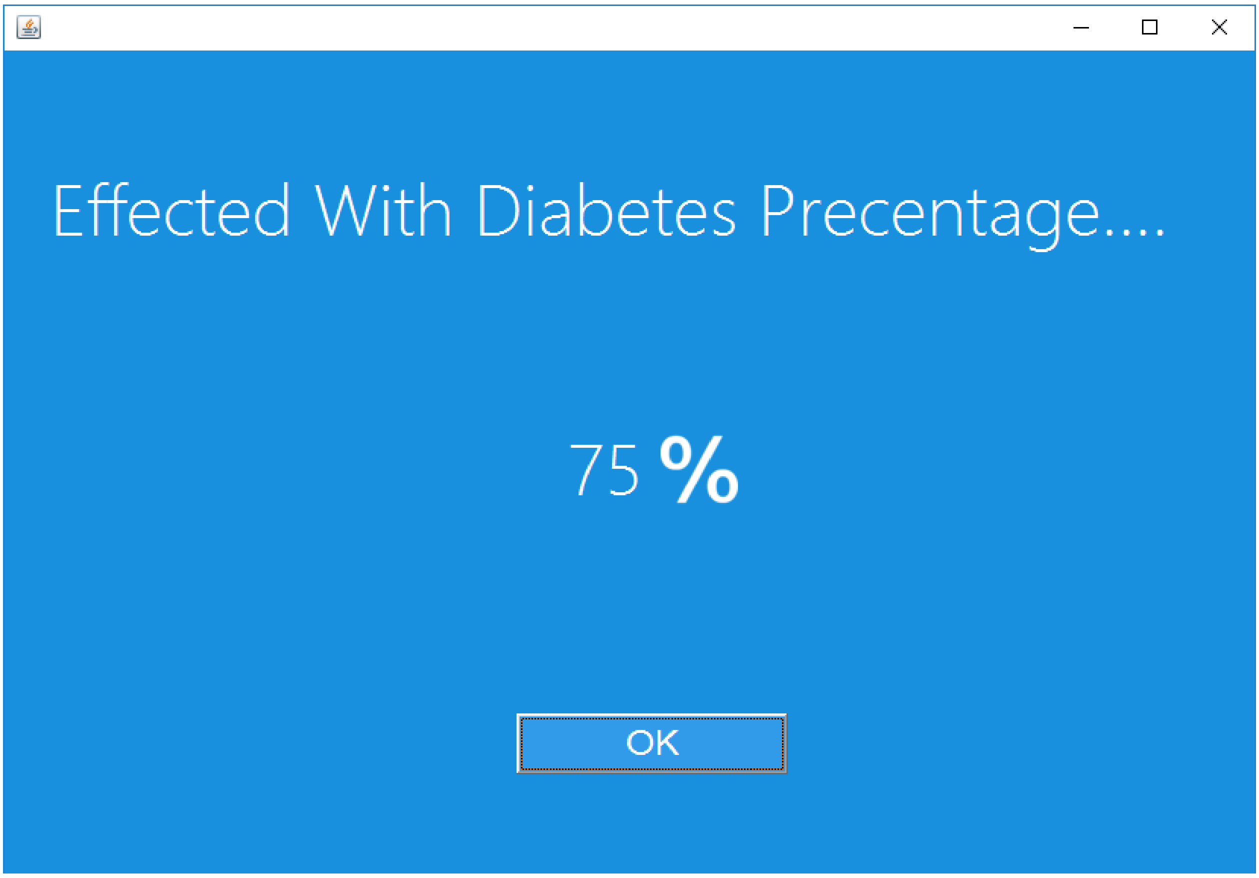


Answer Following Queries..

● Suffering with Faint Problem
● Suffering with Cramps Problem
○ Suffering with anxiety Problem
○ Suffering with Paralysis Problem
○ Suffering with Insomina Problem
○ Suffering with Gastric Disorders Problem
○ Suffering with Dizziness Problem
○ Become moody Very often

Enter Value For Preg : 6
Enter Value For Plas : 148
Enter Value For Pres : 72
Enter Value For Skin : 35
Enter Value For Insu : 0
Enter Value For Mass : 33.6
Enter Value For Pedi : 0.637
Enter Value For Age :

Enter Value For Pedi :



Effected With Diabetes Precentage....

75 %

IX. CONCLUSION

The System can be used by Hospitals and small scale organizations where data analysis is secondary. This system will help such organization in setting up a cost effective solution for their data analyzing needs. This will help them to focus their budget on requirements which are really necessary for the organization and use only the resources that are already available with them for analysis.

REFERENCES

[1] Rui Rui and Changchun Bao, "A Novel Supervised Learning Algorithm for Musical Instrument Classification," 978-1-4673-1118-2/12/$31.00 ©2012 IEEE. (*references*)

[2] S. Essid, G.Richard and B.David, "Musical instrument recognition by pairwise classification strategies," IEEE Transactions on Audio, Speech, and Language Processing, vol. 14, no. 4, pp.1401-1412, 2006.

[3] M. R. Every, "Discriminating between pitched sources in music audio," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 2, pp.267-277, 2008.

[4] Georges Matar1, Jean-Marc Lina1,2, Julie Carrier2, Anna Riley3, Georges Kaddoum1"Internet of Things in Sleep Monitoring: An Application for Posture Recognition Using Supervised Learning"2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom).

[5] Stamatis Karlos,Nikos Fazakis,Sotiris Kotsiantis ,Kyrgiakos Sgarbas "Effectiveness of semi-supervised learning in bankruptcy prediction" IISA.2016.7785435.

[6] Mauro Ribeiro, Katarina Grolinger, Miriam A.M. Capretz "MLaaS: Machine Learning as a Service" 2015 IEEE 14th International Conference on Machine Learning and Applications.

[7] Jianlin Xu, Yifan Yu, Zhen Chen, Bin Cao, Wenyu Dong, Yu Guo, and Junwei Cao, "Cloud Computing Based Forensic Analysis for Massive Mobile Applications Using Data Mining" TSINGHUA SCIENCE AND TECHNOLOGY ISSN 110070214 10/10 pp418-427 Volume 18, Number 4, August 2013.

[8] Pengcheng Shen, Xin Du, Chunguang L Distributed Semi-Supervised Metric Learning

[9] Deepti Mittal, Damandeep Kaur, Ashish Aggarwal"Secure Data Mining in Cloud using Homomorphic Encryption"

[10] Sagar S. Nikam, "A Comparative Study Of Classification Techniques in Data Mining Algorithms" Oriental Journal Of Computer Science and Technology ISSN:0974-6471 VOL 8, No 1, Pgs 13-19, April 2015

[11] Bin Liu, Shu-Gui Cao, Qing-Chun Li, Qi Li, "A Hierchical Distributed Data Mining Architecture" Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, Guilin, 10-13 July, 2011.

[12] Seema Sharma, Jitendra Agrawal, Shikha Agarwal,Sanjeev Sharma, "Machine Learning Techniques for Data Mining: A Survey"

[13] Ruxandra-tefania PETRE, "Data Mining in Cloud Computing" Database System Journal vol. III, no.3/2012

[14] Mr.A.Srinivas,M.Kalyan Srinivas, A.V.R.K.Harsha Vardhan Varma, "A Study On Cloud Computing Data Mining" International Journal Of Innovative Research In Computer and Communication Engineering.Vol.1, Issue 5, July 2013

[15] Mohammed Abdul Khaleel, Satish Kumar Pradham, G.N. Dash, "A Survey Of Data Mining Techniques on Medical Data for Finding LocallyFrequent Diseases" International Journal Of Advanced Research in Computer Science and Software Engineering" Volume 3, Issue 8, August 2013.