Empirical Rapid and Accurate Prediction Model for Data Mining Tasks in Cloud Computing Environments

Samaher Al-Janabi
Department of Information Networks,
Faculty of Information Technology (IT),
University of Babylon
Babylon 00964, Iraq
samaher@itnet.uobabylon.edu.iq

Ahmed Patel
School of SOFTAM, Faculty of
Information Science and Technology
(FTSM), Universiti Kebangsaan Malaysia,
43600 UKM , Malaysia
whinchat2010@gmail.com

Hayder Fatlawi
Information Technology Center,
University of Kufa
Najaf 00964, Iraq
hayder_alnaji2@uokufa.edu.iq

Ibrahim AlShourbaji
Computer Network Department,
Computer Science and Information System College,
Jazan University
Jazan 82822-6649, Saudi Arabia
i_shurbaji@yahoo.com

Kenan Kalajdzic
School of SOFTAM, Faculty of Information
Science and Technology (FTSM),
Universiti Kebangsaan Malaysia,
43600 UKM , Malaysia
kalajdzic@gmail.com

Abstract— **With the arrival of big data and cloud computing as a computing concept, it is becoming ever more critical to efficiently choose the most optimum machine on which to execute a program, for example in the healthcare environment. This process of choice is also complicated by the fact that numerous machines are available as virtual machines. Hence, predicting the most optimum choice of machine based on a target application is a challenge. Prediction techniques consume large amount of computing resources when operating with multi-dimensional data that can cause long delays compounded by cross validation process in evaluating and choosing the most optimum prediction model. We propose a model of prediction techniques to predict and classify some of the health datasets to retrieve useful knowledge to illustrate how a data miner can choose a suitable machine especially in cloud environment with good accuracy in a timely manner. Our results show that the execution time has an inverse relation with the use of resources of a machine and the accuracy of prediction could be different from one machine to another using the same predicting technique and dataset.**

Keywords--**Computer architectures, Data Miner, Predicting techniques, Cloud computing, and Healthcare Datasets**.

I. Introduction

Data mining is the process of extracting knowledge from large amounts of data. Over time, huge data are stored and we can apply some intelligence techniques on these data to support the decision making process. Generally, data mining tasks are classified into two categories: ***Prediction tasks*** and ***Description tasks*** The first one includes classification, regression and time-series analysis which are used to find an unknown value of the target feature by using known values of other features, the second one includes clustering, summarization, mining associations, and sequence discovery which are utilized to find understandable patterns of data [1].

Cloud computing is migration of computing from physical hardware and local platforms to virtualized services that host in the cloud [2]. Cloud systems are divided according to their accessibility: to ***public, private and hybrid*** (see Figure 1).

Public cloud systems do not specify to specific organizations, it can be accessible from anywhere over the internet such as the Amazon Web Services (AWS), Microsoft Azure, Salesforce.com, and Google AppEngine [2][3]. The main issue of care with public cloud is the security of information and data.

Private cloud can overcome the problem which was explained in public cloud by building the infrastructure of the cloud system especially for specific originations that have sensitive information. Many of the software were developed to manage private cloud environments such us Open stack and Eucalyptus.

Hybrid cloud has a combination between public and private systems which provide services for both sides [3].

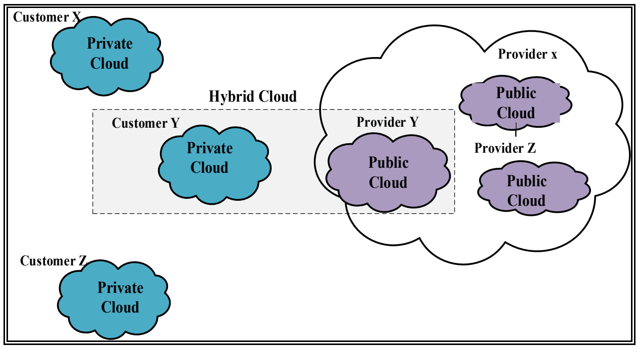

Figure 1. Public Cloud, Private Cloud And Hybrid Cloud

978-1-4799-8021-5/14/$31.00 ©2014 IEEE

According to the kind of service, cloud systems are classified into ***three major categories:*** (1) Infrastructure as a Service (IaaS) when demands of the customer is to have specific resources like the speed of CPU or size of memory, or it can be physical or virtual resources. (2) Platform as a Service (PaaS) which provides the required platform especially as an environment of programming or to execute any solution. (3) Software as a Service (SaaS) that supports the most popular software for customers such as Google Apps and Sales force [2] [3].

In this paper, we focus on prediction techniques to predict and classify some of the health datasets to retrieve useful knowledge. In many cases, prediction algorithms consume large amount of computing resources with high dimensionality data. This can be delayed to hours with some prediction techniques like Random Forest Regression and classification (RFRC). Usually, Cross validation process is used to evaluate the prediction model with all possible distribution of data; such a process can take a long time. The time issue is critical alongside with accuracy. This work attempts to answer many questions that can be asked about the benefits of using cloud computing to solve data mining problems

Why should we use cloud computing? Is this possible? What are the main advantages and disadvantages of this approach? How could we develop data mining algorithms to work under cloud? What are the challenges that still act as open problems in this field? Table 1 attempts to answer the above questions.

Table 1: Advantages, Disadvantages and Challenges of using Data Mining user Cloud

Advantages	Disadvantages	Challenges
Specific cost: "pay as you go".	Security risks for sensitive data	Security problems
No need to have fixed resources	Not suitable for small data mining tasks	Specify infrastructure correctly with (IaaS)
No need to maintain hardware	Time for Cloud management could delay the task	Reduce cloud management time
Availability of fast computational resource in cloud computing	Wrong infrastructure	Saving accuracy of task with less time
Less time of processing with huge/big data	May be constrained by internet transfer speeds if the data is hosted at different sites	Need to verify its accuracy

One of the most important challenges in cloud environment is to ***decide which resources we need in cloud computing*** (in other words, *what is the suitable architecture of a virtual machine to execute the specific task of prediction?).* In specific words; choosing the architecture' perfect from the following points: Central Processing Unit (CPU), Random Access Memory (RAM) and type of operating systems (OSs). All of them could affect the performance of the prediction model. Another challenge, is to choose a prediction technique with the best performance that affects the time, accuracy and cost of prediction task.

In this paper, we attempt to find which machine is more suitable for prediction by applying five prediction techniques on three different datasets in five different architectures. Also, we propose a new abstraction for prediction task on cloud computing that aims to choose the best prediction technique and to distribute computing of cross validation over multi cloud computing nodes.

The rest of the paper is structured as follows. Section 2 presents the related works Section 3 presents the main tools used while in Section 4, the suggested architecture that contains the two parts are explained. Section 5 shows the experiments. Finally, the discussion and conclusion of the paper is presented in Section 6.

II. Related Works

Both Data mining and cloud computing have received significant interest in recent years. Many works are presented trying to improve data mining techniques using the abilities of cloud computing.

In 2008 Christopher et al., [4] started the first steps in this field. They tried to scale up the classifiers for Cloud Computing computers/machines by making a comparison amongst the three classification techniques (decision trees, k-nearest neighbors and support vector machines) and to evaluate their performance with distributed data. They worked with six different data sets (Protein, KDDCup, Alpha, Beta, Syn-SM, and Syn-LG). On the other hand our work extends and converges with their research by using decision trees techniques and proposing an implementation of an abstraction classification method; whilst our work diverges from theirs by using alternative classification techniques and different datasets.

In 2009 Yuzhanget al., [5] improved the decision tree construction (SPRINT algorithm) with service-oriented architecture based on the principles of Cloud Computing by using Distributed Computational Service Cloud (DCSC). They provided a SPRINT model to handle the user-defined dataset. Our work goes further by building a cloud computing model which distributes cross-validation tasks and we use different datasets.

In 2010 Jianzong Wang, et al., [6] worked with Data Mining of Mass Storage based on Cloud Computing by implementing a combination model between three techniques (Global Effect, K-NN and Restricted Boltzmann Machines) for Netflix Prize dataset mining, They observed that Global Effect and K-NN worked very well but RBM did not perform well. We are similar to their work with regards to data mining based on cloud computing but we are different by implementing a new model for classification of tasks.

In 2011 Gopalakrishnan and K. Lakshmi [7] proposed a new Hierarchical Virtual K-Means Approach (HVKM) with two models of cloud computing system PaaS and SaaS for the user who desired to provide Business Analysis as a Service. They used the Sample insurance data to test their approach. We converges with their work on implementing the tow model of cloud computing system for data mining purposes but we differ from it by developing a model for classification and prediction rather than simple clustering.

In 2012 Tong et al., [8] they built a web application for their data mining analysis in the forecasting service based on cloud computing. They called it Forecasting as a Service (FaaS), which provides forecasting services for users. They evaluated the performance of six data mining techniques (Logistic Regression, Time Series, ANN, Random Forest, SVM, MARS) on the SaaS model based on using R, PHP and MySQL tools to analyze the manufacturing data of the industrial index information forecasting in Taiwan. From a technical point of view, our work is similar with theirs on building a model for prediction which uses multiple prediction techniques but we are differ from them by using health care datasets for medical diagnosis purposes as opposed to an industrial index dataset.

In 2012 N. R. Sheth and J. S. Shah [9] implemented one of the most popular Association Rule algorithms called Apriori which improved on MapReduce programming model to work on the Hadoop platform. They built an interface between Hadoop and the Sector file System (Sector/Sphere Cloud system) which give all Hadoop application the ability to work on Sector data and they observed a decline in performance in the Sector file system due to I/O and JNI overhead. We converge with their work on the data mining side of cloud computing while our work differs from theirs by using prediction techniques rather than yet Association Rule algorithms.

In 2012 Juan and Pallavi [10] developed a sequential association rule algorithm (Apriori) by redesigning it from the original concept and applied it to work on MapReduce on the Amazon EC2 cloud model which provided a parallel computing platform. They used four different datasets (chess, mushroom, connect and T10I4D100K).. Our work is similar to this research by developing a cloud model but we use different datasets for health care and medical diagnosis purposes.

In 2012 Jing and Shanlin [11] developed a new solution for classification rules techniques by using a Genetic algorithm on Master/PU Model cloud system model. They used UCI Machine Learning Repository breast cancer symptom dataset to test this model. That model had better accuracy compared with traditional genetic classification (TGC) and required less execution time compared with the decision tree model DTC when the data is increased. We are broadly similar with their research approach in implementing the cloud model for classification and prediction tasks using the breast cancer dataset, but we differ by using multiple environments to test the performance of the different prediction techniques using different datasets.

In 2012 Mahendiran, et. al., [12] implemented the K-means algorithm in Cloud environment (Google App Engine, Cloud SQL). They used three datasets (Iris Dataset, Blood Transfusion and Service Center Data Set). The experimental results showed that it worked well in a Cloud Computing environment. Our work takes a similar approach with theirs by implement a model for cloud computing environment but it is applied on the classification tasks rather than clustering, as well as we use different medical datasets.

In 2012 Nandini and Saurabh [13] proposed high performance cloud data mining algorithm by improving the Apriori algorithm using Genetic algorithm approach to work on sector/sphere cloud framework, and they used multi transaction datasets to validate their model. Our approach is similar to theirs by developing the cloud model for data mining but we differ from it by using alternative datasets in order to test classification techniques in multiple computing environments.

In 2013 Kawuu and Yu-Chin [14] presented four efficient algorithms (Association rule Equal Working Set (EWS), Request On Demand (ROD), Small Size Working Set (SSWS) and Progressive Size Working Set (PSWS)) to utilize cloud nodes in cloud computing environment with IBM's Quest synthetic data generator. They observed that the four algorithms are more scalable than TPFP-tree and BTP-tree schemes. PSWS required only 12.2% and 18% of the execution time used respectively by TPFP-tree and BTP-tree. Our work is similar to theirs by utilizing the resources of the cloud nodes to distribute the computation of the data mining tasks but our model is better developed for classification and prediction tasks rather than just the Association rule.

III. The Tools of Research

In this paper, we will use a number of tools including ***computer architecture, prediction techniques and evaluation measures***, all of them can support our goal to find which environment is suitable and which technique is the best for prediction task.

A. *Techniques of prediction*

Some of the data analysis tasks have to find continuous values or ordered values for target variable and this model is called predictor and the tasks are called numerical prediction. Regression is the main statistical method used with numerical prediction. Classification and numeric prediction are the two major types of prediction problems which need careful usage. [1]. There are many machine learning methods which can be used with predication techniques of which five of the most optimum ones are described below.

1. Decision Tree (DT)

This technique is used to classify data with easier and more understandable way. To classify the problem, the value of the target variable should be amended by using some interested variable. It makes recursively a split procedure from top to down trying to building a tree, each branch represents a question about the value of one of the variables and to answer which direction of child nodes to follow, right or lift, If there are no more questions to ask about a specific direction we will reach the terminal node. There are many implementations of DT such us ID3, J48 and CART.

2. Random Forest Regression and Classification (RFRC)

In a dataset, we have many types of variables representing possible permutations for nested split operations and there is a chance to lose the optimal one. Random Selection is made by Random Forest for the subset from a variable to make a tree, and then choose another subset randomly (again) to make another tree. Finally, we have a forest from trees resulting from the training set and then testing the records to classify them according to the class based on the highest voting by Random Forest technique.

RFRC has accuracy as good as Adaboost and sometimes better. It's relatively robust to outliers and noise and faster than bagging or boosting. It gives useful internal estimates of error, strength, correlation and variable importance. It's simple and easily parallelized [15].

When using Random Forest importance, its training time required from hours to even days of computation, especially for larger sets [16]. RFRC depends on randomness that makes it suffer from problems of randomness.

3. *Bayesian Neural Network Classifier (BNNC)*

Bayesian networks classifier is one of the classification methods that used both probability theory and graph theory to solve the problem of non-deterministic relation between variables set and target class. Graph is used as a qualitative part with the node representing the variables and arcs representing dependency between them. Arcs must be directed and without any cycle DAG. A set of parameters is used as quantitative part to represent conditional probability distributions [17] [18] [19].

Despite of consuming time in constructing the network which requires a large amount of effort, it is more flexible than the Naive Bayesian classifiers by allowing some dependency between variables. BNNC can deal with incomplete data. Because the data is combined probabilistically with prior knowledge, the method is quite robust to model over-fitting [17].

4. *Naïve Bayesian*

Naïve Bayesian Classifier (NBC) is one of classification methods that used probability theory to solve the problem of non-deterministic relation between variables set and target class. It works to find conditional property of target class with hard assumption that there is conditional independency between variables. With continuous values, Naïve found conditional probability by using either discretization method or Gaussian distribution [17].

Despite the low performance with presence of correlated variables, it can deal with noise data and missing values. It is not affected by irrelevant variables [17]. NBC has easy construction without complicated parameter estimation schemes which make it fast with huge datasets [20].

5. *PART Rule*

It's a combination between Decision Tree algorithm C4.5 and RIPPER algorithm for generating rules by repeatedly building partial decision trees. The rule set of this method has the same size and accuracy of the rules which is generated by C4.5 and better than RIPPER, Also, it is faster because it does not need post processing. The main advantage of PART is generating good rule sets without global optimization [21].

B. *Computer Architecture*

Mostly, data mining tasks require considerable amount of computational resources which would affect the performance of the task specially the execution time. In this part, we will discuss four factors of computer architecture and the relation of them with the performance of the data mining tasks.

1. *Kind of operating system*

Windows and Linux are the most popular operating systems and all applications and computer users work on one of them. Windows have the highest number of users and applications on the personal computer (PC). On the other hand, Linux is the main operating system for servers which provide services for multi users like the web server. Linux is used now with cloud computing environment as virtual machines. In this paper, we make a comparison between both windows and Linux on data mining tasks.

2. *Type of machine*

Physical machine is an operating system which runs directly on the physical hardware of the computer while a virtual machine is an application which simulates the work of an operating system and runs indirectly on the physical hardware of the computer. The main concept of cloud computing is to use virtualization which means create a virtual machine for each customer who is provided by cloud services. In this paper, we compare the performance of data mining tasks between a physical machine and virtual machine.

3. *Memory size*

The data mining tasks and specially prediction and classification techniques require a considerable amount of memory resources to hold data to be predicted or classified. Some of these techniques like the Decision Tree need all the data to be on the memory for the building tree process. In this paper, we compare the performance of five prediction and classification techniques in five different machines with different memory sizes.

4. *CPU speed*

The core of computational system is CPU and the execution time of all processes depends mainly on the speed of CPU. Multi cores CPUs have more throughputs which can support the performance of data mining tasks. In this paper, we use five machines with different CPUs to execute five prediction and classification tasks.

5. *Addressing Buses*

Addressing buses are computer buses used to read or write on/from memory location. They affect the time of processes which need to reach the memory. Data mining tasks usually have wide usage of data that is located on the memory especially with classification and prediction methods. In this paper, we use both 32 bit and 64 bit addressing buses.

C. *Evaluation of Prediction Techniques*

The main purpose of using evaluation of predictor or classifier result is to obtain a more reliable model. In this paper, we use two metrics: ***accuracy and execution time***.

1. The accuracy of the model represents the percentage of instances which are predicted correctly. Binary

classification means the class is labeled with two values only and we can consider one value as positive class and the other as negative class. True positive represents the instances which belong to the positive class and predicted correctly. True negative represents the instances which belong to the negative class and predicted correctly [1]. The following equation calculates the accuracy of the model:

$$\boldsymbol{accuarcy} = \left(\frac{\boldsymbol{true_{pos}}}{\boldsymbol{pos}} \times \frac{\boldsymbol{pos}}{\boldsymbol{pos+neg}}\right) + \left(\frac{\boldsymbol{true_{neg}}}{\boldsymbol{neg}} \times \frac{\boldsymbol{neg}}{\boldsymbol{pos+neg}}\right) \quad (1)$$

2. The execution time of the model is in seconds. It has a major effect on the decision about the best technique. Because a best technique of predication or classification must give the results in short time especially with the execution of critical tasks.

IV. Proposed Methodology

As we explained in the previous sections, choosing a suitable machine architecture still acts as a challenge especially in cloud computing environment. We propose two models to support the decision about specific machine. The first model is developed for choosing the best computer architecture for prediction and classification tasks. The second model is developed to utilize the cloud computing resources for prediction and classification task and we can use the first model for virtual machines architecture.

A. Proposed Model for Finding the Best Computer Architecture for Prediction Task

The first model we proposed in this paper can support the data miner to select suitable computer resources and machines to execute classification or prediction tasks. In the first stage, a number of popular datasets is used from the bank of datasets. We use ***three popular datasets; Breast cancer, Leukemia and Pima Diabetes***.

The second stage consists of five different machines with different computational resources; each one of these machines will use the three datasets from the first stage and apply five prediction and classification techniques on each dataset. The ***five techniques*** used in the second stage are ***Decision Trees (DTs), RFRC, BNNC, NBC and PART***.

The third stage of the model is an evaluation stage for predication and classification technique's performance according to their time and accuracy. Finally, we display the best prediction for the best computer architecture. We describe these stages in Figure 2

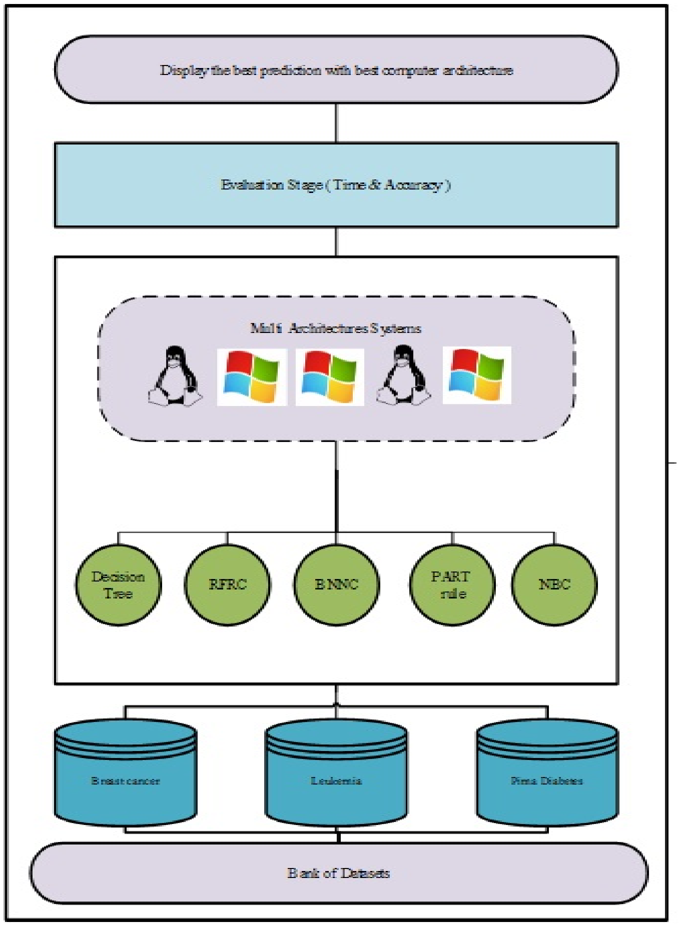
Display the best prediction with best computer architecture
Evaluation Stage (Time & Accuracy)
Multi Architectures Systems
Decision Tree
RFRC
BNNC
PART rule
NBC
Breast cancer
Leukemia
Pima Diabetes
Bank of Datasets

Figure 2. Proposed Model for Finding Best Computer Architecture for Predicting Task

B. Proposed model for predicting task based on cloud computing

The second model that we proposed in this paper utilizes cloud computing features to improve prediction and classification performance. There is (m) of prediction and classification techniques and we aim to find the best one. Also there are ((m*n)+2) nodes or virtual machine in cloud environment.

The first stage has one node for making (m) copy of the dataset. The second stage has (m) nodes for (m) techniques, each node has **many functions** (P) for partitioning of dataset to (n) subset for cross-validation process, (PR Train) for training process, (PR Test) for testing process, (R) for collecting result and (Ev) for evaluating (n) results from cross-validation.

The third stage consists of one node for final evaluation between results of (m) techniques and chooses the best according to time and accuracy. We describe these stages in Figure 3.

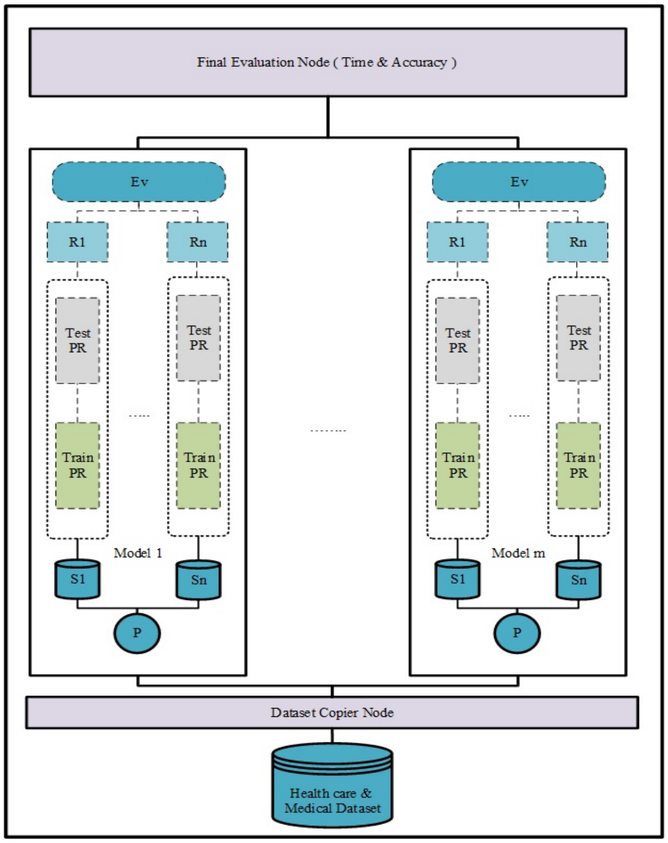
Final Evaluation Node (Time & Accuracy)
Ev
R1
Rn
Test PR
Train PR
Model 1
Model m
S1
Sn
P
Dataset Copier Node
Health care & Medical Dataset

Figure 3. Proposed Model for Predicting Task Based On Cloud Computing

V. 5. Experimental Results

The datasets that we used in this paper are considered very popular for classification and prediction tasks. The main information of that datasets is explained in Table 2. Our results focus on applying the first proposed model on these datasets. The first dataset is the Breast cancer dataset which contain (9) attributes, the class attribute and (286) instances, the task of this dataset is binary classification.

TABLE 2: Datasets Information

Databases	Number of attributes	Number of instances	Task
Breast cancer [22]	9	286	binary classification
Leukemia[23]	7129	38	binary classification
Pima [24]	8	768	binary classification

The second dataset is the Leukemia dataset which consists of (7129) genes, the class attribute and (38) instances, the task of this dataset are binary classification. The third dataset is Pima Diabetes data Set which has (8) attributes, the class attribute and (768) instances, the task of this dataset are binary classification.

The environment of applying the classification and prediction techniques is in five different machines. We describe the details of these machines in Table 3.

TABLE 3: Information Of Machines

Machine	OS type	Add.B uses	CPU	RA M	Hard Driv e Size	Machin e Type
Machine 1	Windows 7 Ultimate	32 bit	1 CPU	2 G	16 G	Virtual Machine
Machine 2	Windows 7 Ultimate	64 bit	Corei5 CPU	4 G	1 T	Physical Machine
Machine 3	Ubuntu 10.04 Desktop	32 bit	Dual Core 2.10 CPU	2 G	120 G	Physical Machine
Machine 4	Windows 7 Ultimate	32 bit	Dual Core 2.10 CPU	2 G	120 G	Physical Machine
Machine 5	Ubuntu 10.04 Desktop	32 bit	1 CPU	2 G	16 G	Virtual Machine

From Table 4, we observe that the Decision Tree has the best accuracy for the Breast Cancer dataset and NBC and PART have less time with this dataset. Also, NBC has the best accuracy with the Leukemia dataset and RFRC has less time for it. For Pima Diabetes, NBC has the best accuracy and less time.

TABLE 4: Performance Of Prediction Techniques On Machine 1

	Breast Cancer		Leukemia		Pima Diabetes	
	Time	Accuracy	Time	Accuracy	Time	Accuracy
Decision Tree	0.02	75.52%	0.14	84.21%	0.03	73.82%
RFRC	0.03	69.93%	0.09	78.94%	0.14	73.43%
BNNC	0.04	72.02%	Out of Memory		0.02	74.34%
NBC	0.01	71.67%	0.12	94.73%	0.01	76.30%
PART	0.01	71.32%	0.15	84.21%	0.03	75.26%

From Table 5, we observe that the Decision Tree has the best accuracy for the Breast Cancer dataset and NBC has less time with this dataset. Also, NBC has the best accuracy with the Leukemia dataset and RFRC has less time for it. For Pima Diabetes, NBC has the best accuracy and Decision Tree and RFRC has less time.

TABLE 5: Performance Of Prediction Techniques On Machine 2

	Breast cancer		Leukemia		Pima Diabetes	
	Time	Accuracy	Time	Accuracy	Time	Accuracy
Decision Tree	0.06	75.52%	0.12	84.21%	0.01	73.82%
RFRC	0.02	66.78%	0.11	78.94%	0.01	68.09%
BNNC	0.03	71.32%	Out of Memory		0.06	74.34%
NBC	0	71.67%	0.07	94.73%	0.05	76.30%
PART	0.02	72.02%	0.1	84.21%	0.03	75.26%

From Table 6, we observe that the Decision Tree has the best accuracy for the Breast Cancer dataset and NBC has the less time with this dataset. Also, NBC has the best accuracy with the Leukemia dataset and RFRC has less time for it. For Pima Diabetes, NBC has the best accuracy and less time.

TABLE 6: Performance Of Prediction Techniques On Machine 3

	Breast cancer		Leukemia		Pima Diabetes	
	Time	Accuracy	Time	Accuracy	Time	Accuracy
Decision Tree	0.04	75.52%	0.12	84.21%	0.09	73.82%
RFRC	0.23	66.53%	1.15	86.84%	0.35	73.82%
BNNC	0.02	72.02%	Out of Memory		0.07	74.34%
NBC	0.01	71.32%	0.35	94.73%	0.03	75.86%
PART	0.07	71.32%	0.28	84.21%	0.03	75.86%

From Table 7, we observe that the Decision Tree has the best accuracy for the Breast Cancer dataset and BNNC has less time with this dataset. Also, NBC and PART have the best accuracy with the Leukemia dataset and Decision Tree has less time for it. For Pima Diabetes, NBC has the best accuracy and less time.

TABLE 7: Performance Of Prediction Techniques On Machine 4

	Breast cancer		Leukemia		Pima Diabetes	
	Time	Accuracy	Time	Accuracy	Time	Accuracy
Decision Tree	0.07	75.52%	0.15	84.21%	0.12	73.82%
RFRC	0.04	69.93%	1.06	86.84%	0.1	73.43%
BNNC	0.02	72.02%	Out of Memory		0.03	74.34%
NBC	0.03	71.67%	0.19	84.21%	0.01	76.30%
PART	0.03	71.32%	1.06	86.84%	0.05	75.26%

From Table 8, we observe that the Decision Tree has the best accuracy for the Breast Cancer dataset and NBC has less time with this dataset. Also, NBC has the best accuracy and less time with Leukemia. For Pima Diabetes, NBC has the best accuracy and less time.

TABLE 8: Performance Of Prediction Techniques On Machine 5

	Breast cancer		Leukemia		Pima Diabetes	
	Time	Accuracy	Time	Accuracy	Time	Accuracy
Decision Tree	0.03	75.52%	0.14	84.21%	0.26	73.82%
RFRC	0.07	68.53%	1.06	86.84%	0.09	73.82%
BNNC	0.03	72.02%	Out of Memory		0.04	74.34%
NBC	0.01	71.67%	0.13	94.73%	0.02	76.30%
PART	0.03	71.32%	0.19	84.21%	0.04	75.25%

VI. Discussion

Evaluation of prediction techniques tasks performance in a multi architectural system has many difficulties. There is a lack on the knowledge analysis system which could work under different operating systems like windows and Linux but Weka can do that efficiently, also it covers the most popular techniques of classification and predication.

There is another deficiency on the Cloud environment builder which could enable us to apply the second proposed model for utilizing cloud resources to find the best prediction techniques. There are some open source applications for building private cloud but they are so difficult to use for testing issues, so the experimental results in this paper focused on the first proposed model.

Form the experimental results in Section 5 of this paper we discover many interesting points. Initially, we discussed the change of execution time depending on the prediction technique and machine type. The best execution time for Breast Cancer dataset was 0 second and 0.07 second for Leukemia dataset by NBC technique on Machine2. This was the physical machine and had high amount of computational resources with Windows 7 operating and 64 bit addressing buses, the time was 0 second. Even it is optimal to have zero second time but we considered it the best execution time because it was so short to calculate. The worst execution time for Breast Cancer dataset was 0.23 and 1.15 Leukemia dataset second for second by RFRC technique on machine3 which was the physical machine and had Ubuntu Linux operating system and 32 addressing buses. For the Pima Diabetes dataset, the best execution time was 0.01 by the NBC technique on machine1 and machine2 and by the Decision Tree and RFRC techniques on machine2. The worst execution time for Pima Diabetes dataset was 0.35 by RFRC on machine3. Overall, we observed that RFRC had the longest time for all datasets especially on machine3 and NBC had shorter time for all datasets especially on machine2.

As expected, there was an inverse relationship between increasing computational sources and execution time of prediction technique. But the most interesting issue in the experimental results was the changing of accuracy of prediction tasks on different machines. The best accuracy for Breast Cancer dataset was 75.52% using the Decision Tree and it still same on all machines, the worst accuracy for this dataset was by RFRC but it changed from 69.93% on machine1 and machine4, and to 66.53% on machine3. The best accuracy for Leukemia dataset was 94.73% by all machines except machine4 and the worst accuracy was 78.94% on machine2 only. For Pima Diabetes dataset, the best accuracy was by NBC with accuracy 76.30% for all machines except machine3, the worst accuracy for this dataset was by RFRC on all machines with different values especially on machine2 which has accuracy 68.09%.

Another interesting result was that the Decision Tree was the most stable prediction technique on all machines on both evolutions metrics i.e. time and accuracy. BNNC technique had memory out error on all machines with Leukemia dataset. In general, we think that this occurrence is because the high dimensionality (i.e., higher number of attributes) of the Leukemia dataset.

VII. Conclusion

We proposed two models in this paper. The first model was to select the best computer architecture for predicting a task which could help the data miner to choose a suitable machine especially in cloud environment. The second model was developed for utilizing cloud computing resources on predicting tasks by distribute cross-validation process on multiple cloud nodes then to choose the best predicting technique based on time and accuracy. We observed that the execution time had an inverse relation with resources of machine and the most important observation was that the accuracy of prediction could be different from one machine to another with the same predicting technique and dataset. Also, we noticed that exceeding memory space as an error occurred when predicting Leukemia dataset when using the BNNC technique. For our future work we hope to apply the second model on real cloud environment and use the results of the first model to choose suitable machine architecture on cloud environment of the second model.

References

[1] Jiawei Han and Micheline Kamber, (2006), book, " Data Mining: Concepts and Techniques ", Second Edition, Elsevier & Morgan Kaufmann Publishers, United States of America.

[2] L. Wang, R. Ranjan, J. Chen and B. Benatallah, (2011), book, "Cloud Computing: Methodology, Systems, and Applications, (2011). " CRC Press, Taylor and Francis Group..

[3] Zaigham Mahmood, Richard Hill,(2012). book,"Cloud Computing for Enterprise Architectures", Springer-Verlag London Limited.

[4] C. Moretti, K. Steinhaeuser, D. Thain and N. Chawla, (2008), "Scaling Up Classifiers to Cloud Computers, Eighth IEEE International Conference on Data Mining, ISSN: 1550-4786, pp (472 – 481.

[5] Y.Han, P. Brezany and I. Janciak, (2009). "Cloud-Enabled Scalable Decision Tree Construction", Fifth International Conference on Semantics, Knowledge and Grid", ISBN: 978-0-7695-3810-5, pp.128–135.

[6] J. Wang, J. Wan, Z. Liu and P.Wang,(2010). "Data Mining of Mass Storage based on Cloud Computing", Ninth International Conference on Grid and Cloud Computing. ISBN: 978-1-4244-9334-0 , pp. 426–431.

[7] T. G. Nair and K. L. Madhuri,(2011). "Data Mining using Hierarchical Virtual K-means Approach Integrating Data Fragments in Cloud Computing Environment" , IEEE CCIS, ISBN: 978-1-61284-203-5, pp. 230–234.

[8] T. Yang, B. Shia, J. Wei and K. Fang, (2012). "Mass Data Analysis and Forecasting Based on Cloud Computing", Journal of Software, vol. 7, no. 10, October.

[9] N.i R. Sheth and J. S. Shah, (2012). "Implementing Parallel Data Mining Algorithm on High Performance Data Cloud ", International Journal of Advanced Research in Computer Science and Electronics Engineering Volume 1.

[10] J. Li, P. Roy, S. Khan, L. Wang and Y. Bai, (2012). "Data Mining Using Clouds: An Experimental Implementation of Apriori over MapReduce", 12th International Conference on Scalable Computing and Communications (ScalCom), Changzhou, China, December.

[11] J. Ding and S. Yang, (2012). "Classification Rules Mining Model with Genetic Algorithm in Cloud Computing", International Journal of Computer Applications Volume 48, No.18.

[12] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam, (2012). "Implementation of K-Means Clustering in Cloud Computing Environment", Research Journal of Applied Sciences, Engineering and Technology.

[13] N. Mishra, S. Sharma and A. Pandey, (, 2013). "High performance Cloud data mining algorithm and Data mining in Clouds", IOSR Journal of Computer Engineering (IOSRJCE) Volume 8, Issue 4.

[14] K. W. Lin and Yu-Chin Lo, (2013). "Efficient algorithms for frequent pattern mining in many-task computing environments", Journal of Knowledge-Based Systems.

[15] L. Breiman, (2001). "Random forest", Machine Learning, ISSN: 0885-6125, Volume 45, Issue 1, pp.5-32.

[16] M. B. Kursa, (2014). "Robustness of the Random Forest-based gene selection methods", BMC Bioinformatics, ISSN 1471-2105, Volume 15.

[17] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, (2006). book, "Introduction to Data Mining", Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

[18] N. Friedman, D. Geiger and M. Goldszmidt, (1997)"Bayesian Network Classifiers", Machine Learning, 29, Kluwer Academic Publishers. Manufactured in The Netherlands.

[19] S. H. Ali, (2007). "Evolving Optimal Bayesian Neural Network Using Breeder Genetic Programming to Acoustic Radar Pattern Identification".

[20] X. Wu , V. Kumar, (2007). "Top 10 algorithms in data mining", Knowledge and Information Systems, Volume 14, Issue 1, pp.1-37.

[21] E. Frank and I. H. Witten, (1998). "Generating Accurate Rule sets Without Global Optimization", Department of Computer Science, University of Waikato,.

[22] M. Zwitter & M. Soklic, Breast cancer data, Institute of Oncology, University Medical Center, Ljubljana, Yugoslavia, repository: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer.

[23] Leukemia data, Science,(1999). Vol 286, pp. 531-537..http://www.upo.es/eps/bigs/datasets.html

[24] Pima Indians Diabetes Database, National Institute of Diabetes and Digestive and Kidney Diseases, repository: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes