A Novel Cloud based and Cyclic Approach for Supervised Learning in Data Mining

R. Vinodharasi, PG Student[1]
Department Of Computer Science and Engineering,
[1]PG Student, Parisutham Institute of Technology and Science, Thanjavur, Tamilnadu, India.

Abstract: **The Clouds provide more services to the users. In the past era the software users have to purchase the software with license. But nowadays clouds provide these softwares on pay and use basis to use it. This pay and use service brought more comfort to the users. Through this paper, we propose 'A Novel Cloud based and Cyclic Approach for Supervised Learning', in the area of Data Mining. Because the data warehouses are utilizing the virtualized Iaas cloud service. i.e., the data warehouses are stored in the clouds and utilizes services of cloud. We concentrated more on classification problem in the area of data mining, because the global business scenario is entirely changed. According to these changes the need of classification become essential in all areas. Particularly to the data, which is stored in the virtual data warehouses. In the cloud hosted data warehouses the current test data set will become as training data set after some period of time. In our proposed approach we introduced the post- mortem technique on the classification model to know the facts, how good the model is induced to classify the data set from the training data set. To provoke a high-quality and an efficient classification model, the model must go through the post-mortem operation to know the reality of the classification. The test data set must go through the pre-processing operation to make the data set pure and clean. This process must be done in routine.**

Key Words : Data warehouse, Cloud, Data mining, Supervised Learning, Classification, Pre-processing, Post-mortem.

I. INTRODUCTION

In this section we introduce the abstraction of knowledge Discovery Process and the abstraction of the cloud architecture which gives a brief narration about data mining and the cloud architecture

A. DataMining

The entire business scenario depends on the business analytics in the current age. Data mining is a technique of extracting previously un-known, hidden and treasured information from large volume of data sources like data bases, Relational Data Bases, Data Warehouses, Data Marts, World Wide Web and many more. So, the data mining can be called as Knowledge Discovery process because the data mining process provides the analytical knowledge regarding to the domain. The abstraction of knowledge discovery process. The Figure. 1 is extracted from kamber[22].

With reference to the figure there are seven iterative steps, among them four steps are different forms of data preprocessing technique. The forms of preprocessing techniques are Data Selection, Data Cleaning, Data Integration and Data Transformation. It is very clear that to take right decisions to improve the business, right knowledge is required. We can extract different types of knowledge with various data mining techniques[22].

The increase in demand for low power and low voltage VLSI circuits can be investigated different levels of design, such as the architectural, circuit, layout and the process technology. At the device level, reduction in supply voltage and reduction in the threshold voltage to reduce the power consumption, where as in layout use of short channel transistors, poly and diffusion areas, shorter metal lines for connecting two various devices. It reduces capacitances in circuit and device level.

The data mining techniques are used to extract various kinds of knowledge. First, the Class/Concept Description technique. In this technique the data set is formed as classes or categories

based on their behavior or characteristics. The well known Periodic Table groups are examples for classes. Behind a class there must be some characterization or behavior. The Second, association rule mining helps to know the frequent item from the transactional database. Third, classification and prediction technique helps for forecasting purpose. Fourth, clustering technique is very convenient to organize the unsupervised objects. In our research we concentrated on the classification. For the classification we must take the known dataset. Known dataset mean the class properties are very clear. Rest of the techniques covers Outliers and some other.

B. CloudArchitecture

Now a days, the software Service Oriented Programs, which encompasses small computer program, mission critical programs and business application software those are widely utilizing the cloud services on rental basis or pay and use basis. The following is the abstract and overview of cloud computing[13][14], [1 5], 161.

The application users use the cloud applications by using some hand hold devices or PCs networks from the Cloud. Cloud consumers did not have any control over the cloud infrastructure, and often employs multi-tenancy system architecture. The term multi-tenancy means different cloud consumers' applications are organized in a single logical environment. The SaaS cloud achieve financial feasibility and optimization in terms of security, speed, availability, disaster recovery and maintenance. Examples of SaaS are schoolDude.com, www.rightscale.com. SalesForce.com. Google Mail, Google Docs ...etc.,

The Platform as a Service allows the customers to develop the Development Lifecycle directly. i.e., this Service acts as platform cloud. Hence, the difference between SaaS and PaaS is that SaaS only hosts completed cloud applications whereas PaaS offers a development platform that hosts both developed and in under development cloud applications. i.e., this service requires some additional development infrastructure to facilitate the services both for completed applications and in-progress applications. Google AppEngine, Oracle, Force.com ,Bungee Connect are some of the examples for PaaS.

The Cloud clients/customers can directly use cloud infrastructures like the storage, network, processing and some other resources facilitated in the Iaas Cloud. In other words it provides the infrastructure for computation, storage..etc. The IaaS mainly provides the virtual machines (VM) to fulfill the virtualization concept. Virtualization is extensively used in IaaS cloud in order to integrate/ decompose physical resources in an ad-hoc manner to meet growing or shrinking resource demand from cloud consumers. This aims to transform the application software architecture. So that multiple instances (from multiple cloud consumers) can run on a single application (i.e. the same logic machine). Ex:Amazon's Ec2.

The Data as a Service can be seen as a special service of IaaS. Because this service is originated from the data delivery demand from the virtualized storage. The virtual storage concept is used in Iaas. Due to this service the costs of software purchase (licensing cost), installation of dedicated servers and their maintenance costs are drastically reduced. Some of Examples of DaaS are DaaS include Amazon S3, Google BigTable, and Apache HBase, etc.

In this paper, we propose 'A Novel Cloud based and Cyclic Approach for Supervised Learning' , in the area of Data Mining. Because the data warehouses are utilizing the virtualized IaaS cloud. i.e., the data warehouses are accommodating in the clouds and utilizing the services of cloud. We concentrated more on classification problem in the area of data mining, for the virtual data warehouses. In the cloud hosted data warehouses the current test data set will become as training data set after some period of time. Hence we introduced the post- mortem operation on the classification model to know the facts, how good the model is induced to classify the data set from the training data set. To induce a good and an efficient classification model, the model must undergo the post-mortem operation to know the reality of the classification and the test data set must undergo the pre-processing operation to make the data set pure and clean. The paper is organized as follows, In the first section we introduced knowledge discovery process and the cloud abstraction. i.e., brief description of the cloud services are explained. Second section describes the motivation of classification problem, which is one of the data mining techniques and also states the problem definition. Third section illustrates the related work to our research and the literature survey and the review findings. Fourth section describes the traditional approach for classification. Fifth section describes 'A Novel Cloud based and Cyclic Approach for Supervised Learning' method and its activities are explained along with its architecture. Sixth section describes results. Finally, the conclusions, acknowledgements and references are mentioned.

II. MOTIVATION AND PROBLEM DEFINITION

The motivation problem for the classification is identification of birds in the real world. In Nature several birds are living. Among them, only some birds can be identified / classified by the human beings but some of them cannot be identified!classified. The reason is some birds properties are known and some are unknown. It is very clear that we can classify a bird to a certain family based on its properties which are well known . In the same way the known data set is known as a supervised data. This is called the classification problem [1][2]. The classification is one of the learning process and it uses the supervised dataset. In supervised learning, the algorithm works with a set of examples whose labels are known. The labels can be nominal values in the case of the classification task, or numerical values in the case of the regression task. During the classification process, we are facing so many problems like noise data, irrelevant data,

erroneous data .. etc[4][5]. To overcome these difficulties we propose a novel approach to solve the classification problem. Definition:-Classification is a process of learning a function f(x), which can classify the x into any pre defined category.

III. RELATED WORK

A. Literature Survey

Mitchell (1997) [18] Presented a very good and an excellent coverage on several Classification techniques like decision tree based classification, Bayeis classsifier,Classification by back propagation.. etc.,. Duda et al.(2001) [19],Webb (2002) [20], Fukunaga (1990)[21], Han and Kamber (2001)[22] gave an extraordinary illustrations and excellent description on different classification algorithms like decision tree based classification, Bayeis classification, Backpropagation ..etc., algorithms with suitable examples. The C4.5 rules algorithm for extracting classification rules from decision trees was proposed by Quinlan(1993)[23]. An indirect method for extracting classification rules from artificial neural networks was proposed by Andrews et al.(1995)[24], P.Langley, W. Iba and K. Thompson (1992)[25] Investigated the NaIve Bayes Classifiers, the method has worked well for applications such as text classification. Pang-Ning Tan (2007)[26] introduced an approach for building a classification problem. Armbrust, M et al.,(2009) [15] described the cloud services like Software as a Service, Platform as a Service, Infrastructure as a service and Network as a Service

B. Review Findings

From the above review we found that the clouds are providing different services like Paas, Iaas, Saas, Dass. Various authors described different algorithms for classification problem with a generalized approach to solve a classification problem. But this generalized approach is not supported the cloud environment, At the same time the general approach for classification problem is not inducing high quality classification rules. Therefore we introduced a novel and cyclic approach for classification problem[13]. It's extensive work is proposed in this paper , we proposed as 'A Novel Cloud based and Cyclic Approach to solve a Classification Problem', which induces high quality rules

IV. EXISTING METHOD FOR CLASSIFICATION

The current classification learning approach first takes a training data set and then induces a model from that with following some learning algorithms . After model induction[6][7][8] the learned model is applied on some other new test data set then evaluate the accuracy of the learning model. Figure.2 illustrates an approach for building a classification problem. This architecture is referenced from Tan[26], gives the abstraction of classification problem.

Though the traditional process is providing a solution for the classification problem, this approach is suffering with some problems[7][9][1O] due to the dynamic and incremental data sets. In this approach there is no complete result analysis. So that, we can not find the root cause of misclassification. It is very clear that today's current data will become as historical data for tomorrow. That is the test data set will become as training data set, it is like a cyclic process, which this approach is not supposed. This approach is not providing the solution for the virtual data warehouses stored in cloud environment.

All the classification algorithms classifies the training data set by using the above approach. The following is the brief narration of some of the classification techniques. Rule based classifiers are in the fonn if-then structure. The rules are framed with two main parts. First antecedent and next rule consequent. The rule antecedent, is the if part, specifies a set of conditions referring to predictor attribute values, and the rule consequent, the then part, specifies the class predicted by the rule for any example that satisfies the conditions in the rule antecedent. These rules can be generated using different classification algorithms, the most well known being the decision tree induction algorithms and sequential covering rule induction algorithms [27]. A Bayesian network (BN) consists of a directed, acyclic graph and a probability distribution for each node in that graph given its immediate predecessors [28]. A Decision Tree Classifier consists of a decision tree generated on the basis of instances. At the root node and at each internal node, a test is applied. The outcome of the test determines the branch traversed, and the next node visited. The class for the instance is the class of the final leaf node [29]

V. PROPOSED APPROACH FOR CLASSIFICATION

We Proposed a new architecture for cloud based data classification. The following is the algorithm for the proposed 'A Novel Cloud based and Cyclic Approach to solve a Classification Problem'. Which is the extensive work of A Novel and Cyclic Approach to Solve a Classification problem[13]. The following is the algorithmic representation of the proposed framework for classification in the cloud environment

Step 1: Login to the Cloud System

Step 2: Initially select some sample data set as training data set to build a learning model.

Step 3: Induce a classification learning model with applying c1assi fication Algorithms.

Step4: Each classification algorithm forms a variety of learning model.

Step 5: Apply models on the test data set for model Evaluation.

Step 6: Perform Post Mortem process on learning model with fmding the classification accuracy and the classification error rate as

Step 7: Accuracy = Number of correct Predictions Total no of pre dictions

Step 8: Error Rate = Number of Wrong Predictions Total no of predictions

If classification error rate is beyond the given threshold value go to step 9 otherwise go to step 1 1 .

Step 8: Error Rate = Number of Wrong Predictions Total no of predictions
If classification error rate is beyond the given threshold value go to step 9 otherwise go to step 1 1 .

Step 9: Pre process Test Data Set and Transform as training data set to induce a learning model.

Step 10: go to Step 3

Step 11: Logout the cloud and Stop.

The proposed architecture is as shown in Fig 3. In this method we introduced the pre processing and the post-mortem processes. The Proposed architecture is utilizing the cloud Storage feature and providing features like Platform as a service for analytics and the Software as a service. To obtain more quality in the classification process, we conduct the post mortem operation on the test data set, only if its error rate is greater than the threshold value. Post mortem is a process that determines whether the classification process is successful or not. This process reduces the future risks and helps to improve and to uplift best practices.

The general post-mortem process has the following five fundamental steps (adapted from [1 1]):

l. A project review is planned to identifY the most suitable methods and tools used in the other steps. The post-mortem reviews, the reasons for the review, the focus and the participants are defined;
2. Both objective and subjective data are collected from all the project participants via pre-defined metrics, surveys, debriefings, etc. to identifY the useful information for the" following step" (workshop/review) ;
3. A"project history day" is the most important step, and it is held to combine reflective analysis of project events with the actual project data after a project's major milestone (post-iteration), or after a project has finished (post-mortem). In the case of large projects , only a few key people participate in this session;
4. The findings are analyzed, prioritized and synthesized as lessons learned. This is often started during the project history day after identifYing and prioritizing the positive events and problems;

5. The summaryof the findings is published and presented in a way that enables future projects to know what processes or tools are important to continue, and to turn problems into improvement activities.
As a part of quality improvement in the test data set, we perform the pre processing to the test data set before transforming as training data. The preprocessing ensures the data quality in multi dimensional views like accuracy, completeness,

consistency, timeliness, accessibility ... etc. Pre processing encompasses Data Cleaning [12], Data Integration, Data Reduction, Data Transformation and Data Discretization activities.
The descriptive data summarization increases the understandability of the data . The measures mean, median and mode are the related measures of Central Tendency[26]. The

Mean value is calculated as follows. Consider 'n' no.of samples

asXl,X2,X3, ...Xn.

$x' = 1/n(\sum x_i ; 1 \le i \le n)$ $\mu = \sum x/N$ ---------->(1)

weighted arithmetic mean

$x' = (\sum w_i x_i / \sum w_i ; 1 \le i \le n)$ ---------->(2)

L fa is the sum of the frequencies or number of scores up to the interval containing the median. fw is the frequency or number of scores within the interval containing the median.
I is the size or range of the interval. Mode is the value that occurred frequently.

Wl,W2,W3,..... Wn are different weight

$Mn = (LIM + (N/2) - (\sum f_0)/ f_w)I$ ---------->(3)

where:

Mn is the median. LIM is the lower limit.
N is the scores total number.

The relation between the mean, mode and median

is mean - mode = 3 x (mean - median)

The relationship is symmetric and skewed as follow

A shows the symmetric relationship of mode, mean and median. A relation is said to be symmetric, if and only if (b,a) is in the relation whenever (a,b) is in relation. Figure.4.B. shows the positively skewed relationship of mode, mean and median. FigureA.C shows the negatively skewed relationship of mode, mean and median. Fig 4.A, Fig 4.B and Fig 4.C are useful to understand the relationship of mode, mean and median measures. Equations (1), (2) and (3) are useful to compute mean, weighted mean and median values for discrete values respectively.

The rest of the pre-processing techniques are useful to make the data clean, qualitative and noise free dataset. The quality input gives quality out put, like a computer characteristic GIGO (Garbage In Garbage Out). At final we present the flow chart for the proposed method.

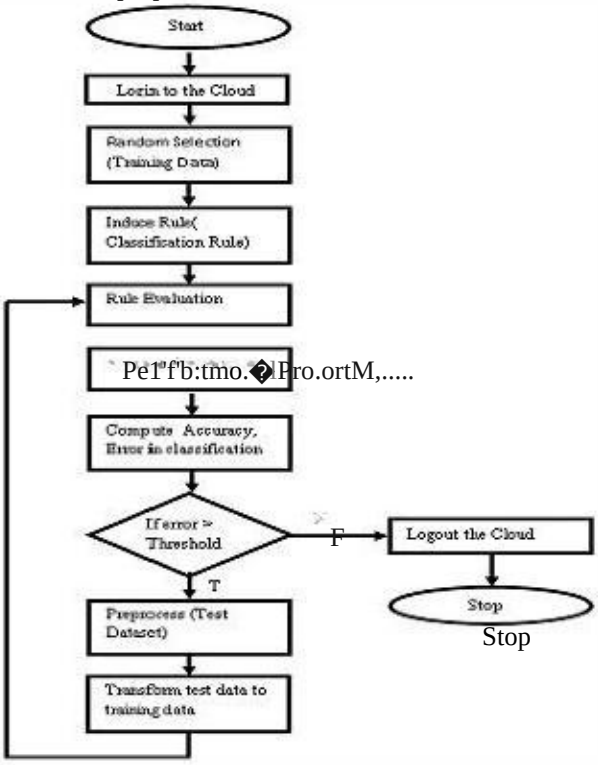


Fig: 5 Flow chart for the proposed method

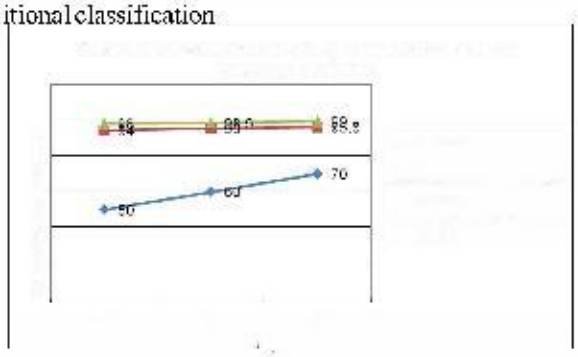
VI. RESULTS ANALYSIS

We have taken supervised sample data of different customers of a bank to induce a classification rule. First we induced the classification rule with the help of traditional classification approach. Later we calculated the classification accuracy of the rule with the help of new test dataset. After that, we performed the post-mortem operation with applying various filters, different analytical views and some other techniques to fmd the root cause of misclassification based on the classification error. Then we removed the noise in the test data set by using various pre processing techniques like smoothing, aggregation, missing

Table.I.Training and Test Data set attribute

S.No.	Own House	Marital Status	Income per Annum.	Defaulted Borrower
I	yes	married	600000	No

itional classification



Flg.6.Results

values concept... etc., Again we induced the rule, by using the preprocessed testing data set (test data set will change as training data set), Finally, we observed that more quality rule is induced after conducting the post-mortem of classification rule

Figure.6. shows the test results of3 different sized data sets using the Weka tool.

VII. CONCLUSION

In this paper we presented A Novel Cloud based and Cyclic Approach for Supervised Learning, which can deal with variety of training data sets in the cloud environments. In this cyclic approach the test data will be transformed as training data after data post mortem and pre-process. The post mortem and the pre-processing steps play a very important role in getting pure and qualitative data. This model can induce more accurate learning models for classification in virtual data warehouses those are accommodated in cloud environment. We are going to perform some experimental work using this approach on different types of data on various configured machines and in different types of environments.

VIII. ACKNOWLEDGEMENTS

We are so grateful to Sri. Dr.Kancharla Ramaiah garu the Secretary and correspondent of Prakasam Engineering College, kandukur for extending his marvelous encouragement and support to do the research with providing the research environment. Last but not least, we are very much thankful to all the authors and co-authors of the reference papers for providing us knowledge about clouds, cloud environment and the data mining techniques particularly about classification.

REFERENCES

[I] Chan, Lois Mai. Cataloging and Classification: An Introduction, second ed. New York: McGraw-Hill, 1994. IS B N 978-0-07-010506-5,ISBN978-0-07-113253-4.

[2] G. Dong, X. Zhang, L. Wong, and J. Li. Classi_cation by aggregating emerging patterns. In Discovery Science, Dec. 1999.

[3] SLIQ: A fast scalable Classifier for Data Mining; Manish Mehta, Rakesh Agarwal and JormaRissanen

[4] D. Pyle, Data preparation for data mining, 1st Vol., Morgan Kaufmann publisher, San Francisco, 1999

[5] I. Guyon, N. Matic and V. Vapnik, "Discovering i n f o r mat i v e patterns and data cleaning", In: Fayyad UM, Piatetsky Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in know I edge discovery and data mining, AAAIIMIT Press, California, 1996. pp. 181- 203

[6] E. Simoudis, B. Livezey B and R. Kerber R , "Integrating inductive and deductive reasoning for data mining", In: Fayyad UM, Piatetsky-Shapiro G, Smyth P and R. (Eds.) Advances in knowledge discovery and data mining, AAAIIMIT Press, California, 1996, pp. 353-373

[7] B. Pfahringer, "Supervised and unsupervised discretization 0 f continuous features", Proc. I2thInt. Con£ Machine Learn i ng, 1995, pp. 456-463.

[8] l Catlett, "On changing continuous attributes into ordered discrete attributes", In Y. Kodratoff (ed), Machine Learning-EWSL-91, Springer-Verlag, New York,1991, pp 164-178

[9] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes forclassi_ cation learning. In I n t I . Joint Con£ on Arti_cial Intelligence(IJCAI), pages 1022. I 029, 1993.

[10] C. W. Hsu, c.c. Chang and C.l Lin, "A practical guide to support v e c t o r c l a s s i f i c a t i o n " , http://www.csie.ntu.edu.tw/-cjlin/papers/guide/guide .pdf, 2003.

[II] Bonnie Collier: Project Review Process Web Site- Postmortem Toolkit; http://www.projectreview.netip r t o o I k i t . a s p [Retrieved March 28, 2004], 1996

[12] E. Rahm and H. H. Do. Data Cleaning: Problems and Current Approaches. IEEE Bulletin ofthe Technical Committee on Data Engineering. Vo1.23, No.4.

[13] Dr.R.Sivarama Prasad,D.Bujji Babu,Vijaya Srinivas,K:"A Novel and Cycl ic Approach to Solve a Classification prob lem" ProceedingsofICACET-2013 Kuala Lumpur, Malaysia ISBN 978-93-5137-024-6. Page.No:31-34.

[14] Cloud Computing for E-Governance. A white paper, lIlT Hyderabad, India.

[15] Armbrust, M et aI., "Above the Clouds: A Berkeley View of Cloud Computing, Technical Report" No. UCB/EECS-2009

[16] "Swamp Computing" a.k.a. Cloud Computing". Web Security Journal. 2009-12-28. Retrieved 20I0-01-25.

[17] "Thunderclouds: Managing SOA-Cloud Risk", Philip Wik". Service Technology Magazine. 2011-10. Retrieved 2011-21-21.

[18] T. Mitchell." Machine learning ," Mc Graw Hill, Boston,M.A,1997.

[19] R.O.Duda, P.E.Hart, and D.G.stock. "Pattern Clasification," John Wiley & Sons, Inc., New York, 2nd edition,200I.

[20] A.R. Webb. "Statistical Pattern Recognition," John Wiley & Sons, 2nd edition, 2002.

[21] K.Fukunaga "Introduction to statistical Pattern Recognition," Academic Press, New York, 1990.

[22] lHanand m.Kamber. "Data Mining: Concepts and Techniques," Morgan Kaufinann Publishers, San Francisco, 200I.

[23] lR.Quinlan. "C4.5:ProgramsformachineLearning," Morgan-Kaufmann Publishers, San Mateo, CA, 1993.

[24] R.Andrews, lDiedrich, and A. Tickle. "A survey and critique of techniques for extracting rules from Trained Artificial Neural Networks. Knowledge Based Sysyems," 8(6):373-389,1995.

[25] P.Langley, W. IbaandK.Thompson. "An analysis of B ayesi an Classifi cati on," In Proc. Of the 10th National Con£ onArtificial Intelligence, Pages 223-228, 1992.

[26] Pang-ning-Tan,Vipin Kumar,Michael Steinbach."lntroduction to Data Mining" Pearson 2007. ISBN978-81-317-1472-0.

[27] G.L. Pappa and A.A. Freitas, Automating the Design of DataMining Algorithms. An Evolutionary Computation Approach, Natural Computing Series, Springer, 20 I 0

[28] A. Darwiche, Modeling and Reasoning with Bayesian Networks, Cambridge University Press, 2009.

[29] M. Garofalakis, D. Hyun, R. Rastogi and K. Shim, B u i I d i n g Deci sion Trees with Constrai nts", Data Mi ning and Know I edge Discovery, vol.7, no. 2,2013, pp.187-214.