ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol.2 , No.5, Pages : 31-34 (2013)
Special Issue of ICACET 2013 - Held during October 14-15, 2013, Kuala Lumpur, Malaysia

A Novel and Cyclic Approach to solve a Classification Problem

Dr.R.Siva Rama Prasad[1], D.Bujji Babu[2] ,Vijaya Sreenivas.K[3]

[1]*Coordinator- International Business Studies, Acharya Nagarjuna University,Guntur,A.P.,India.*raminenisivaram@yahoo.co.in
[2]*Associate Professor, Prakasam Engineering college, Kandukur, A.P., India. bujji_bict@yahoo.com*
[3]*Assistant Professor, Prakasam Engineering college, Kandukur, A.P., India. Vijaysrinivas@gmail.com*

Abstract: In this Paper, we propose a Novel solution to solve the classification problems in the area of Data Mining. We are more concentrated on classification problem because the current test data set will become as training data set after a period of time, hence classification model must face the post-mortem operation to know the facts, how good the model is induced from the training data set. To induce a good and an efficient classification model, the model must undergo the post-mortem and the test data set must undergo the pre-processing and the process must be a cyclic process.

Keywords: Data mining, Classification, Pre-processing, Post-mortem. Classification Model.

INTRODUCTION

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 can not 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 . The known data set is known as a supervised data. This is called the classification problem[1][2]. During the classification process, there are so many problems like noise data, irrelevant data, erroneous data ..etc[4][5]. To overcome the 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.

EXISTING METHOD

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. The following figure illustrates an approach for building a classification problem.

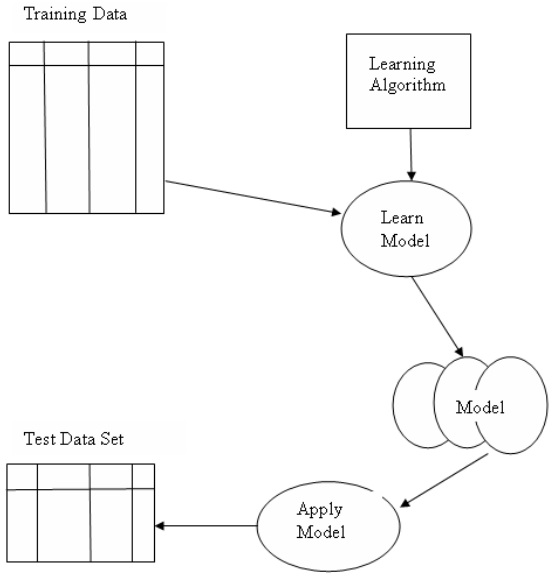

Fig:1. Approach for building a classification problem

Even the traditional process is providing a solution for the classification problem ,this approach is suffering with some problems[7][9][10] 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 becomes as historical data for tomorrow . i.e, the test data set will becomes as training data set ,it seems to be a cyclic process , which this approach is not suppose.

PROPOSED METHOD

The following is the proposed Novel and Cyclic Algorithm.

Step 1: Initially select some sample data set as training data set to build a learning model.
Step 2: Induce a classification learning model with applying classification Algorithms, and form various learning models.
Step 3: Apply models on the test data set for model Evaluation.
Step 4: Perform Post Mortem process on learning model with finding the classification accuracy and the classification error rate as

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol.2 , No.5, Pages : 31-34 (2013)
Special Issue of ICACET 2013 - Held during October 14-15, 2013, Kuala Lumpur, Malaysia

Accuracy = $\frac{\text{Number of correct Predictions}}{\text{Total no of predictions}}$

Error Rate = $\frac{\text{Number of Wrong Predictions}}{\text{Total no of predictions}}$

If classification error rate is negligible then goto step 7
Step 5: Pre process Test Data Set and Transform as training data set to induce a learning model.
Step 6: goto Step 2
Step 7: Stop.

The proposed novel and cyclic approach is as shown below.

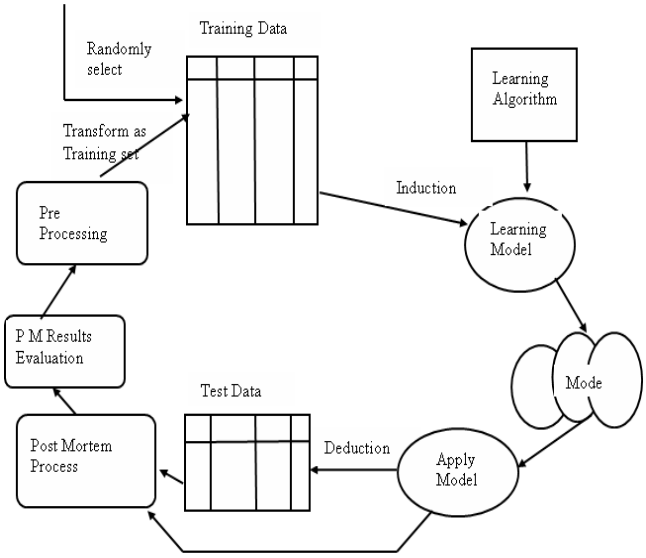

Fig 2:.Proposed Novel and Cyclic Approach for building a classification problem.

We conduct the Post mortem operation on the test data set only if its error rate satisfies 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 improves to uplift best practices. The general post-mortem process has the following five fundamental steps (adapted from [11]):

1. 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 summary of 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 improvement activity we perform the pre processing to the test data set before transforming as training data . The pre processing 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. The mean value is calculated as follows. Consider n no.of samples as $X_1, X_2, X_3, \ldots\ldots X_n$. $\bar{x}$ is the mean value.

Mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

Weighted arithmetic mean $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

$W_1, W_2, W_3, \ldots\ldots W_n$ .are different weights .

$$median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$$

Mode is the value that occurred frequently.

The relation between the mean, mode and median is

$$mean - mode = 3 \times (mean - median)$$

ISSN 2278 - 3091

International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE), Vol.2 , No.5, Pages : 31-34 (2013)
Special Issue of ICACET 2013 - Held during October 14-15, 2013, Kuala Lumpur, Malaysia

The relationship is symmetric and skewed as follows

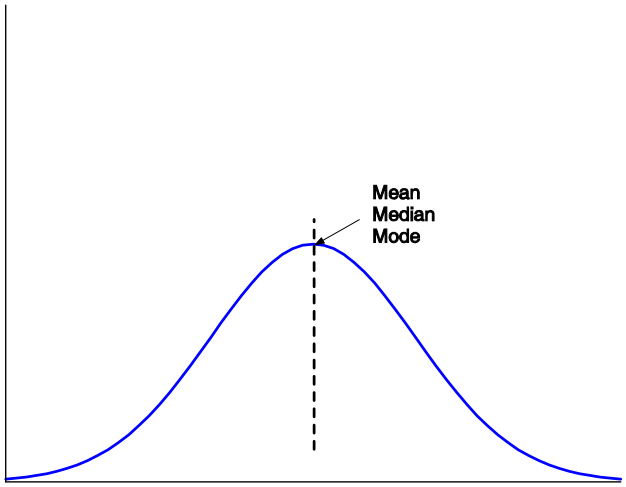

Fig 3:.Symmetric Relation

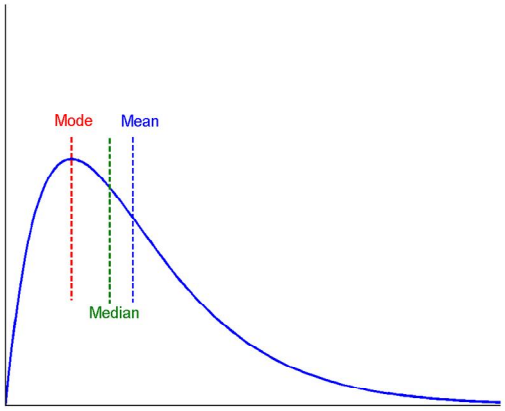

Fig 4: .Positively Skewed

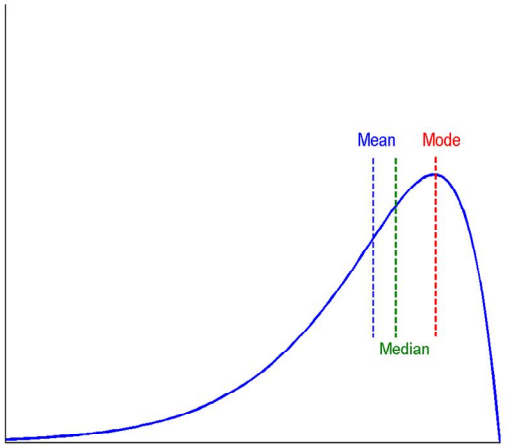

Fig 5:.Negatively Skewed

The remaining techniques also useful to make the data clean and qualitative dataset. Like this the process is cycled.

CONCLUSION

In this paper we presented a novel and a cyclic approach for solving a classification problems which can deal with variety of training data sets. In this cyclic approach the test data will be transformed as training data after data post mortem and pre-processing. This model can induce more accurate learning models. We are going to perform some experimental work on different types of data using this proposed novel and cyclic approach.

ACKNOWLEDGEMENTS

We are very grateful to the Secretary and correspondent of Prakasam Engineering College, kandukur Sri. Dr.Kancharla Ramaiah garu for his marvellous encouragement and support to do the research with providing the research environment.

REFERENCES

[1] Chan, Lois Mai. *Cataloging and Classification: An Introduction*, second ed. New York: McGraw-Hill, 1994. ISBN 978-0-07-010506-5, ISBN 978-0-07-113253-4.

[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 Jorma Rissanen

[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 informative patterns and data cleaning", In*: Fayyad UM, Piatetsky-Shapiro G, Smyth P and Uthurusamy R. (ed) Advances in knowledge discovery and data mining*, AAAI/MIT 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 Uthurusamy R. (Eds.) Advances in knowledge discovery and data mining*, AAAI/MIT Press, California, 1996, pp. 353-373

[7] B. Pfahringer, "Supervised and unsupervised discretization of continuous features", *Proc. 12th Int. Conf. Machine Learning*, 1995, pp. 456-463.

[8] J. 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 for classi_cation learning. In *Intl. Joint Conf. on Arti_cial Intelligence(IJCAI)*, pages 1022. 1029, 1993.

[10] C.W. Hsu, C.C. Chang and C.J. Lin, "A practical guide to support vector classification", http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide .pdf, 2003.

[11] Bonnie Collier: Project Review Process Web Site - Postmortem Toolkit; http://www.projectreview.net/ prtoolkit.asp [Retrieved March 28, 2004], 1996

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

BIOGRAPHIES

Dr.R.Sivarama Prasad is currently working as a coordinator for International Business Studies at Acharya Nagarjuna University,Guntur,A.P.India. He is a interdisciplinary researcher in computer science and management Sciences.He Published a dozens of research papers. His interested area of research is software Engineering, Data Warehouse and Data Mining, e-commerce,Business Intelligence, Systems Architectures. Customer Relationship Management. He Authored seven books.

D.Bujji Babu is an Associate Professor in the department of Computer Science and Engineering at Prakasam Engineering College, Kandukur, A.P. India. He published many research papers in different referred journals. His interested area of research is software Engineering, Data Warehouse and Data Mining, Business Intelligence, Systems Architectures.

Vijaya Sreenivas.K is working as an assistant professor in the department of Computer Science and Engineering at Prakasam Engineering College, Kandukur, A.P. India. His area of interest is Networks,Data Mining,Programming Languages , e-commerce..etc.