ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Modified Cross Validation for Improving the Accuracy Based on Distinct Classifiers

D.Udhayakumarapandian[1], RM. Chandrasekaran[2], A.Kumaravel[3]

Research Scholar, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India[1]

Professor, Department of Computer Science and Engineering, Annamalai University, Chidambaram, India [2]

Professor and Dean, Department of Computer Science and Engineering, Bharath University, Selaiyur, Chennai, India[3]

ABSTRACT: The conventional cross validation for train/test phase of any data mining task is usually based on selecting unique classifier at a time. This approach is commonly tackled for getting better accuracies either by increasing the number of folds or by selecting appropriate classifier. In this paper we establish the different orientation namely for each iterations we select a different classifier and get the average accuracy at the exit of the iterations. We show better results by this new approach comparing to the conventional cross validation in the context of diabetes algorithm.

KEYWORDS: Data mining, Classification, Diabetes data set, Search Methods, Tree, Meta boost, Bayes.

I. INTRODUCTION

In knowledge discovery or data mining, a typical task is to get a learning model from available data. Such a model may be represented by decision trees, rules, bayes and meta-learner. The inherent problem with evaluating such a model is that it may demonstrate adequate prediction capability on the training data, but might fail to predict future unseen data. cross-validation is a procedure for estimating the generalization performance in this context. In 1930s [1] the idea for cross-validation was initiated. The authors Mosteller and Turkey [2], and similar researchers further carried out this idea. Well defined statement of cross-validation, (same as current version of k-fold cross-validation), at the beginning coined in [3]. The two authors Stone and Geisser [4,5] applied cross-validation in 1970s as means for tuning the better model parameters, as against cross-validation only for estimating model performance. Currently, cross-validation is widely accepted in data mining and machine learning community, and serves as a standard procedure for performance estimation and model selection. The main two possible goals in cross-validation are firstly to estimate performance of the learned model from available data using one algorithm. The emphasis is to measure the generalizability of an algorithm. Secondly it is to compare the performance of two or more different algorithms and find out the best algorithm for the available data, or alternatively to compare the performance of two or more types of a parameterized model.

II. DATA PREPARATION

In this section, we dwell the collection of data and format in which the data has to be presented for mining experiments following the iterative steps in Fig 1.We use java based implementation namely Weka tool from University of Waikato, Newzealand.

A. DATASET

The datasets for these experiments are from [18]. The original data format has been slightly modified and extended in order to get relational format.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

i. Dataset Description

The database of diabetes describes a set of eight attributes11 as shown in the below list 2.2. The class attribute has binary values 'tested negative' and 'tested positive'. The number of instances in this database is 768.

B. LIST OF DESCRIPTION OF ATTRIBUTES

For each attribute (all numeric-valued), the description and the units are shown:

1. Number of times pregnant
2. Plasma glucose concentration at 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1) ' tested negative' or 'tested positive'

C. BRIEF STATISTICAL ANALYSIS

Attribute number	Mean	Standard Deviation
1.	3.8	3.4
2.	120	32.0
3.	69.1	19.4
4.	20.5	16.0
5.	79.8	115.2
6.	32.0	7.9
7.	0.5	0.3
8.	33.2	11.8

D. RELATED WORK IN DIABETES DATASET

For the long time the research in diabetes prediction have been conducted. The main objectives are to predict what variables are the causes, at high risk, for diabetes and to provide a preventive action toward individual at increased risk for the disease. Several variables have been reported in literature as important indicators for diabetes prediction. However obtaining the accuracy for recommendation for assisting the physician is a paramount issue. Increased awareness and treatment of diabetes should begin with prevention. Much of the focus has been on the impact and importance of preventive measures on disease occurrence and especially cost savings resulted from such measures. A risk score model is constructed by Lindstrom and Tuomilehto (2003) which includes Age, BMI, waist circumference, history of antihypertensive drug treatment, high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables as categorical variables. A sequential neural network model is obtained by Park and Edington (2001) for indicating risk factors, in the final model, as well as cholesterol, back pain, blood pressure, fatty food, weight index or alcohol index. Concaro et al, (2009) present the application of a data mining technique to a sample of diabetic patients. They consider the clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk in the model.

III. METHODS DESCRIPTION

Here we select a standard set of methods for predicting from the data set described above. We consider three types of classifiers for our study, such as tree based, Bayes approach based, and Meta level based classifiers. The following sections describe briefly the methods for classifier and results of such methods are tabulated further. Then final results are interpreted

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

A. TREE CLASSIFIERS

Supervised Learning is performed conducted using tree classifiers .We select four types of tree classifiers as shown below.

i. Decision Stump

One of the tree classifier is a decision stump, is a machine learning model consisting of a one-level decision tree as described in [3] . That is, it is a decision tree with one internal node (the root) which is immediately connected to the terminal nodes. A decision stump makes a prediction based on the value of just a single input feature

ii. J48

This method description is given from the tool descriptor found in The first number is the total number of instances (weight of instances) reaching the leaf. The second number is the number (weight) of those instances that are misclassified. If your data has missing attribute values then you will end up with fractional instances at the leafs. When splitting on an attribute where some of the training instances have missing values, J48 will divide a training instance with a missing value for the split attribute up into fractional parts proportional to the frequencies of the observed non-missing values. This is discussed in the Witten & Frank Data Mining book as well as Ross Quinlan's original publications on C4.5.

iii. ADTree

Class for generating an alternating decision tree. This version currently only supports two-class problems. The number of boosting iterations needs to be manually tuned to suit the dataset and the desired complexity/accuracy tradeoff. Induction of the trees has been optimized, and heuristic search methods have been introduced to speed learning.

B. BAYES CLASSIFIERS

These types of classifiers includes probability measure for the class values and comes under supervised learning.

i. Naïve Bayes

This belongs to the class implemented in a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an Updateable Classifier you need the Updateable Classifier functionality, use the Naïve Bayes Updateable classifier. The Naïve Bayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build Classifier is called with zero training instances.

ii. Bayes Net

Bayes Network learning using various search algorithms and quality measures. Base class for a Bayes Network classifier. Provides data structures and facilities common to Bayes Network learning algorithms like K2 and B.

C. META CLASSIFIERS

Most of the time, the aggregation of more than one classifier has better performance. Such combinational methods are shown below.

i. Adaboost

Class for boosting a nominal class classifier using the Adaboost M1 method. Only nominal class problems can be tackled. Often dramatically improves performance, but sometimes over fits.

ii. Bagging

Class for bagging a classifier to reduce variance. Can do classification and regression depending on the base learner. Generate B bootstrap samples of the training data: random sampling with replacement. Train a classifier or a regression function using each bootstrap sample For classification: majority vote on the classification results. For regression: average on the predicted values. Reduces variation. Improves performance for unstable classifiers which vary

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

significantly with small changes in the data set, e.g., CART. Found to improve CART a lot, but not the nearest neighbor classifier.

iii. Logit Boost

This classifier is for performing additive logistic regression. This class performs classificationusing a regression scheme as the base learner, and can handle multi-class problems. This method belongs to the type of meta classifiers.

iv. Multi Boost AB

Class for boosting a classifier using the Multi Boosting method. Multi Boosting is an extension to the highly successful AdaBoost technique for forming decision committees. Multi Boosting can be viewed as combining AdaBoost with wagging. It is able to harness both Ada Boost's high bias and variance reduction with wagging's superior variance reduction. Using C4.5 as the base learning algorithm, Multi-boosting is demonstrated to produce decision committees with lower error than either AdaBoost or wagging significantly more often than the reverse over a large representative cross-section of UCI data sets. It offers the further advantage over AdaBoost of suiting parallel execution.

IV. METHOD FOR CROSS VALIDATION

The conventional K-fold cross validation is in the following main algorithm. The 'partition' in the below indicates the ratio of the sizes of training set and testing set at each step of the conventional as <{2,……,k},{1}>to <{1,....k-1},{k}>

A. DEFAULT CV METHOD

Input D= Training set

K=No folds (assumed k=10 for our experiment), C=Selected Classifier

Default CV Method

1. Divide D in to K folds
2. Get the model based on C using K-1 folds
3. Test the model based on C obtained in the step2 using Kth fold.
4. Repeat the testing step 3 for every fold.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

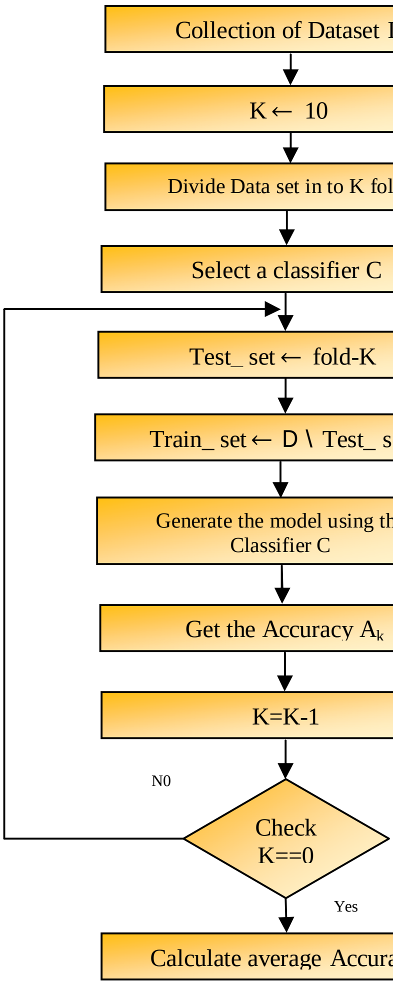
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Output

Average accuracy A



4.1 Flow chart for Default CV method

B.PROPOSED CV METHOD

The experiment for validating our approach is depicted in the following flow chart

Input

D= Training set; K=No folds (assumed to be K=10 for our experiment);
C= {C 1, C 2,…. C k}

Proposed CV Method

Divide D in to K folds
Get the model based on Ck using K-1 folds
Test the model based on Ck obtained in the step2 using Kth folds
Get the accuracy Ak.
Decrement k
Using the results of the models, calculate the average accuracy A
Check 'k==0'if yes then stop else go to step 2.

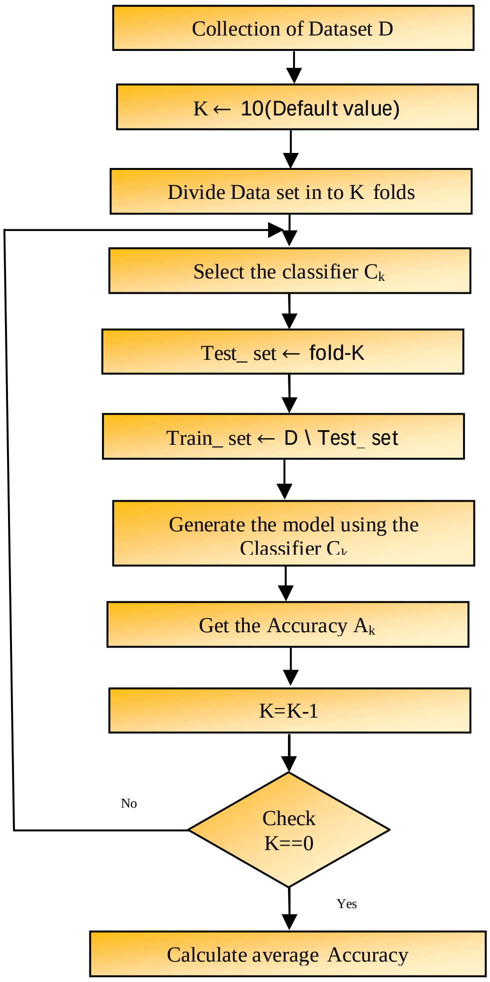
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Output: Average accuracy A= $(\sum_{i=1}^{K} A_i)/K$



4.2 FLOWCHART FOR PROPOSED CV METHOD

V. EXPERIMENTAL RESULTS

In the following table the partition TiSi represents with Ti, test set 10% and Si, train data 90%.

S.No	Classifiers	T1S1(Accuracy)
1.	Bayes Net	78.9474
2.	Naïve bayes	67.1053
3.	Ada boost	65.5475
4.	Bagging	68.65792
5.	Logit boost	67.1053
6.	Multi Boost	60.5263
7.	J-Rip	65.7895
8.	ADTree	67.1053
9.	Decision Stump	60.5263
10.	J48	68.4211
		66.97319

S.No	Classifiers	T2S2(Accuracy)
1.	Bayes Net	78.9474
2.	Naïve bayes	82.8947
3.	Ada boost	76.3158
4.	Bagging	76.3158
5.	Logit boost	82.8947
6.	Multi Boost	75
7.	J-Rip	78.9474
8.	ADTree	78.9474
9.	Decision Stump	72.3684
10.	J48	80.2632
		78.28948

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

Table of T1S1classifiers and < Train, Test > Partition

S.No	Classifiers	T3S3(Accuracy)
1.	Bayes Net	64.4737
2.	Naïve bayes	72.3684
3.	Ada boost	69.7368
4.	Bagging	81.5789
5.	Logit boost	77.6316
6.	Multi Boost	68.4211
7.	J-Rip	69.7368
8.	ADTree	71.0526
9.	Decision Stump	68.4211
10.	J48	71.0526
		71.44736

Table of T3S3classifiers and < Train, Test > Partition

S.No	Classifiers	T5S5(Accuracy)
1.	Bayes Net	73.6842
2.	Naïve bayes	75
3.	Ada boost	72.3684
4.	Bagging	78.9474
5.	Logit boost	73.6842
6.	Multi Boost	73.6842
7.	J-Rip	75
8.	ADTree	77.6316
9.	Decision Stump	71.0526
10.	J48	77.6316
		74.86842

Table of T5S5classifiers and < Train, Test > Partition

S.No	Classifiers	T7S7(Accuracy)
1.	Bayes Net	78.9474
2.	Naïve bayes	80.2632
3.	Ada boost	78.9474
4.	Bagging	84.2105
5.	Logit boost	78.9474

S.No	Classifiers	T9S9(Accuracy)
1.	Bayes Net	73.6842
2.	Naïve bayes	73.6842
3.	Ada boost	73.6842
4.	Bagging	84.2105
5.	Logit boost	73.6842
6.	Multi Boost	78.9474
7.	J-Rip	71.0526
8.	ADTree	76.3158
9.	Decision Stump	69.7368
10.	J48	78.9474
		75.39473

Table of T9S9classifiers and < Train, Test > Partition

6.	Multi Boost	72.3684
7.	J-Rip	76.3158
8.	ADTree	78.9474
9.	Decision Stump	67.1053
10.	J48	81.5789
		77.76317

Table of T7S7classifiers and < Train, Test > Partition

Table of T2S2classifiers and < Train, Test > Partition

S.No	Classifiers	T4S4(Accuracy)
1.	Bayes Net	61.8421
2.	Naïve bayes	68.4211
3.	Ada boost	65.7895
4.	Bagging	63.1579
5.	Logit boost	64.4737
6.	Multi Boost	65.7895
7.	J-Rip	61.8421
8.	ADTree	59.2105
9.	Decision Stump	65.7895
10.	J48	59.2105
		63.55264

Table of T4S4classifiers and < Train, Test > Partition

S.No	Classifiers	T6S6(Accuracy)
1.	Bayes Net	76.3158
2.	Naïve bayes	75
3.	Ada boost	78.9474
4.	Bagging	85.5263
5.	Logit boost	80.2632
6.	Multi Boost	75
7.	J-Rip	80.2632
8.	ADTree	76.3158
9.	Decision Stump	75
10.	J48	85.5263
		78.8158

Table of T6S6 classifiers and < Train, Test > Partition

S.No	Classifiers	T8S8(Accuracy)
1.	Bayes Net	84.2105
2.	Naïve bayes	82.8947
3.	Ada boost	81.5789
4.	Bagging	94.7368
5.	Logit boost	86.8421
6.	Multi Boost	80.2632
7.	J-Rip	85.5263
8.	ADTree	81.5789
9.	Decision Stump	72.3684
10.	J48	97.3684
		84.73682

Table of T8S8classifiers and < Train, Test > Partition

S.No	Classifiers	T10S10(Accuracy)
1.	Bayes Net	74.1176
2.	Naïve bayes	75.2941
3.	Ada boost	80
4.	Bagging	81.1765
5.	Logit boost	81.1313
6.	Multi Boost	78.8235
7.	J-Rip	75.2941
8.	ADTree	81.1765
9.	Decision Stump	77.6471
10.	J48	78.8235
		78.34842

Table of T10S10classifiers and < Train, Test > Partition

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

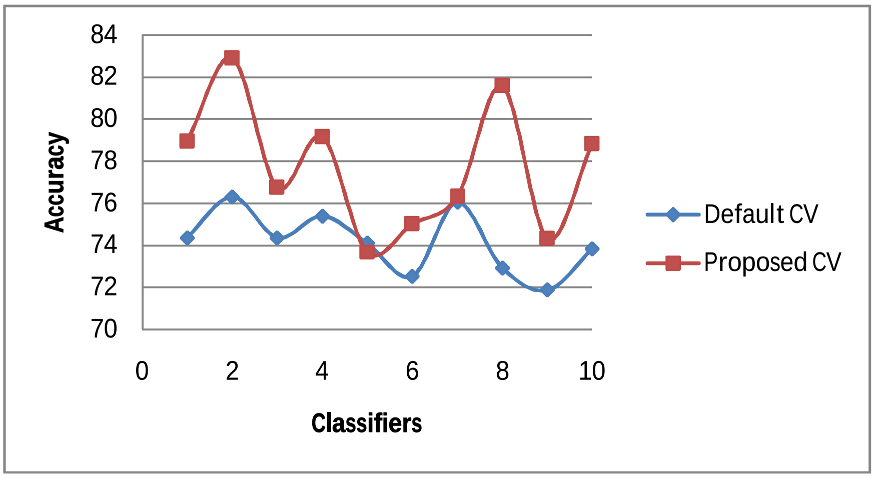


Figure 4.3 Comparison of original reduced dataset Vs for accuracy

VI. CONCLUSION AND FUTURE WORK

We establish the power of varying the classifiers instead of applying single classifier on each part of the training and testing parts. The outputs of our experiments as shown in the Figure 4.3 answer our query of better performance. Specifically even in the small range of data sizes and collection of classifiers we achieve increment 0 to 10%

Future remarks: The approach proposed in this paper can be further modified with the randomizing the indices of the train/test partitions. Since this involves extra iterations for this randomizing process the overall complexity will be increased. But this can be tried with huge datasets in a parallel environment.

REFERENCES

1. Larson S. The shrinkage of the coefficient of multiple correlation. J. Educat. Psychol., 22:45–55,1931.
2. Mosteller F. and Wallace D.L. Inference in an authorship problem. J. Am. Stat. Assoc., 58:275–309, 1963.
3. Mosteller F. and Turkey J.W. Data analysis, including statistics. In Handbook of Social Psychology. Addison-Wesley, Reading, MA, 1968.
4. Stone M. Cross-validatory choice and assessment of statistical predictions. J. Royal Stat. Soc., 36(2):111–147,1974.
5. Geisser S. The predictive sample reuse method with applications. J. Am. Stat. Assoc., 70(350):320–328,1975.
6. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of International Joint Conference on AI. 1995, pp. 1137–1145, URL http:// citeseer.ist.psu.edu/kohavi95study.html.
7. Liu H. and Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. Knowl. Data Eng., 17(4):491–502,2005, doi:http://dx.doi.org/10.1109/ TKDE.2005.66.
8. Efron, B. and Morris, C. (1973). Combining possibly related estimation problems (with discussion). J. R. Statist. Soc. B, 35:379.
9. Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. J. Amer. Statist. Assoc., 92(438):548–560.
10. Refaeilzadeh P., Tang L., and Liu H. On comaprison of feature selection algorithms. In AAAI-07 Worshop on Evaluation Methods in Machine Learing II. 2007.
11. Salzberg S. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min. Knowl. Disc., 1(3):317–328, 1997, URL http://citeseer.ist.psu.edu/salzberg97comparing.html.
12. V. Vapnik. Estimation of Dependences Based on Empirical Data [in Russian]. Nauka, Moscow, 1979.(English translation: Springer Verlag, New York, 1982).
13. D.Udhayakumarapandian.,RM.Chandrasekaran., andA.Kumaravel "A Novel Subset Selection For Classification Of Diabetes Dataset By Iterative Methods" Int J Pharm Bio Sci ,5 (3) : (B) 1 – 8, July(2014)
14. A.Kumaravel., Udhayakumarapandian.D.,Consruction Of Meta Classifiers For Apple Scab Infections , Int J Pharm Bio Sci, 4(4): (B) 1207 – 1213, Oct(2013)
15. A.Kumaravel., Pradeepa.R., Efficient molecule reduction for drug design by intelligent search methods.Int J Pharm Bio Sci, 4(2): (B) 1023 – 1029,Apr (2013)
16. https://www.waset.org/journals/waset/v68/v68-21.pdf world academy of science, engineering and technology, 2012.
17. H.Dunham, Data Mining, Introductory and Advanced Topics, Prentice Hall, 2002
18. Source about wekahttp://www.cs.waikato.ac.nz/ml/weka/ downloaded on 3rd august 2014
19. L. Breiman, " RandomForests,"inMachine Learning, vol. 45, pp. 5-32, 2001.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 9, September 2016

20. Steve R. Gunn., University Of Southampton,Support Vector Machines for Classification and Regression.
21.Dietterich, T. G., Jain, A., Lathrop, R., Lozano-Perez, T. (1994). A comparison of dynamic reposing and tangent distance for drug activity prediction.Advances in Neural Information Processing Systems, 6. San Mateo, CA: Morgan Kaufmann. 216--223.
22.A.Stensvand, T. Amundsen, L. Semb, D.M. Gadoury, and R.C. Seem. 1997. Ascospore release and infection of apple leaves by conidia and ascospores of Venturia inaequalis at low temperatures. Phytopathology 87:1046-1053.
23. Website for attribute description http://archive.ics.uci.edu/ml/machine-learning databases/pima-indians-diabetes., accessed on 3rd august 2014
24.Bal,Hp.2005.Bioinformatics-principles and applications.Tata McGraw-Hill Publishing company Ltd New Delhi.
25.Bo.Th and Jonassen,I-2002 New feature subset selection procedures for classification of expression profiles.Genome Biology 3:research 00170.-0017.11
26. Khalid AA Abakar & Chongwen Yua., Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity, Indian Journal of Fibre & Textile Research, Vol. 39: (B) 55-59, March (2014).
27. Steve R. Gunn., Support Vector Machines for Classification and Regression Technical Report., Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science .,10 May 1998
28. F. Girosi., An equivalence between sparse approximation and Support Vector Machines.A.I. Memo 1606, MIT Artificial Intelligence Laboratory, 1997.
29. N. Heckman., The theory and application of penalized least squares methods or reproducing kernel hilbert spaces made easy, 1997.
30. G. Wahba. Spline Models for Observational Data. Series in Applied Mathematics,Vol. 59, SIAM, Philadelphia, 1990.

BIOGRAPHY

First Author D.Udhayakumarapandian received the MTech in Computer Science and Engineering and pursuing Phd Degree in Computer Science and Engineering from Annamalai University, TamilNadu. He is currently working as an Assistant Professor at the Department of Computer Science and Engineering, Bharath University, Tamil Nadu, India. He has presented and published more than 8 papers in technical conferences and reputed Journals. His areas of research include Data Mining and its applications, Algorithms and Computer networks.

Second Author Dr. R. M. Chandrasekaran received the B.E Degree in Electrical and Electronics Engineering from Maduari Kamaraj University in 1982 and the MBA (Systems) in 1995 from Annamalai University, M.E in Computer Science and Engineering from Anna University and PhD Degree in Computer Science and Engineering from Annamalai University, Tamil Nadu, India in 1995,1998 and 2006 respectively. He is currently working as a Professor as well the Controller of Examinations at the Department of Computer Science and Engineering, Annamalai University, Annamalai Nagar, Tamil Nadu, India. From 1999 to 2001 he worked as a software consultant in Etiam, Inc, California, USA. He has conducted Workshops and Conferences in the Areas of Multimedia, Business Intelligence and Analysis of algorithms, Data Mining. He has presented and published more than 32 papers in conferences and journals and is the author of the book Numerical Methods with C++ Program (PHI, 2005). His Research interests include Data Mining, Algorithms, Networks, Software Engineering, Network Security, Text Mining. He is Life member of the Computer Society of India, Indian Society for Technical Education, Institute of Engineers, Indian Science Congress Association.