Classification with an improved Decision Tree Algorithm

A. S. Galathiya
Faculty of Technology
D. D. University
Nadiad, India

A. P. Ganatra
Charotar Institute of
Technology, CHARUSAT
Changa, India

C. K. Bhensdadia
Faculty of Technology
D. D. University
Nadiad, India

ABSTARCT
Data mining is for new pattern to discover. Data mining is having major functionalities: classification, clustering, prediction and association. Classification is done from the root node to the leaf node of the decision tree. Decision tree can handle both continuous and categorical data. The classified output through decision tree is more under stable and accurate.

In this research work, Comparison is made between ID3, C4.5 and C5.0 and after that Implementation of system is done. The new system gives more accurate and efficient output with less complexity. The system performs feature selection, cross validation, reduced error pruning and model complexity along with classification.

The implemented system supports high accuracy, good speed and low memory usage. The memory used by the system, is low compare to other classifiers as the rules generated by this system is less.

The major issues concerning data mining in large databases are efficiency and scalability. While in case of high dimensional data, feature selection is the technique for removing irrelevant data. It reduces the attribute space of a feature set.

More reliable estimation of prediction is done by f-fold – cross- validation. The error rate of a classifier produced from all the cases is estimated as the ratio of the total number of errors on the hold-out cases to the total number of cases. By increasing the model complexity, accuracy of the classification is increases.

Overfitting is again major problem of decision tree. The system has also facility to do post pruning that is through reduced error pruning technique. Using this proposed system; Accuracy is gained and classification error rate is reduced compare to the existing system.

Keywords
REP, Decision Tree Algorithm, C5 classifier, C4.5 classifier

1. INTRODUCTION
In data mining, Decision tree structures are a common way to organize classification schemes. Classification using a decision tree is performed by routing from the root node until arriving at a leaf node [7]. The research work is made up from ID3, C4.5 and C5 classifier. [4] Both C4.5 and C5.0 can produce classifiers expressed either as decision trees or rulesets. Here C4.5 embodies new algorithms for classification with improved features.

This research work supports high accuracy, good speed and low memory usage. Memory usage is low compare to other classifier because it generates fewer rules. Accuracy is high as error rate is low on unseen cases. And it is fast due to generating pruned trees.

2. C4.5 CLASSIFIER
The resulting decision tree is generated after classification. The classifier is trained and tested first. Then the resulting decision tree or rule set is used to classify unseen data. C4.5 is the newer version of ID3. C4.5 algorithm has many features like:

- Speed - C4.5 is significantly faster than ID3 (it is faster in several orders of magnitude)
- Memory - C4.5 is more memory efficient than ID3
- Size of decision Trees – C4.5 gets smaller decision trees.
- Ruleset - C4.5 can give ruleset as an output for complex decision tree.
- Missing values – C4.5 algorithm can respond on missing values by '?'
- Overfitting problem - C4.5 solves overfitting problem through ***Reduce error pruning technique***.

2.1 Algorithm C4.5
Input: Example, Target Attribute, Attribute

Output: Classified Instances

In pseudo code the algorithm looks like this [31]:

- Check for the base case
- Construct a DT using training data
- Find the attribute with the highest info gain (A_Best)
- A_Best is assigned with Entropy minimization
- Partition S into S1,S2,S3...
- according to the value of A_Best
- Repeat the steps for S1, S2, S3
- For each ti ϵ D, apply the DT

Base cases are the following:

- All the examples from the training set belong to the same class (a tree leaf labeled with that class is returned).
- The training set is empty (returns a tree leaf called failure).
- The attribute list is empty (returns a leaf labeled with the most frequent class or the disjuction of all the classes).

OUTPUT*: decision tree which classifies the data correctly*

2.2 Comparison – Current Algorithms

2.2.1 Improved features of C4.5 on ID3 algorithm

- Handling both continuous and discrete attributes [5] - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those

whose attribute value is above the threshold and those that are less than or equal to it.

- Handling training data with missing attribute values - C4.5 allows attribute values to be marked as '?' for missing. Missing attribute values are simply not used in gain and entropy calculations.
- C4.5 allows the attributes with different costs.
- Post Pruning - C4.5 creates first decision tree and after creation, it goes back through the tree and attempts to remove branches that do not help by replacing them with leaf nodes.

2.2.2 Problem of decision tree in data mining

- Determining how deeply to grow the decision tree [10]
- Handling continuous attributes
- Focus is not with only relevant attributes
- Missing values of attributes & attributes with different cost are not handled
- Improve computational efficiency

3. RESEARCH WORK

3.1 Proposed Algorithm

This paper reduces the error ratio using cross validation, pruning and class complexity. It is having only focus with the relevant attributes through Feature selection – Genetic search. The following steps are carrying out to classify the decision tree methods [31]:

1. Create a root node for the tree
2. Check for the base case
3. Apply Feature Selection using Genetic Search
4. bestTree = Construct a DT using training data
5. Perform Cross validation
    a. Divide all examples into N disjoint subsets, $E = E_1, E_2, ..., E_N$
    b. For each i = 1, ..., N do
        i. Test set = E_i
        ii. Training set = $E - E_i$
        iii. Compute decision tree using Training set
        iv. Determine performance accuracy P_i using Test set
    c. Compute N-fold cross-validation estimate of performance = $(P_1 + P_2 + ... + P_N)/N$
6. Perform Reduced Error Pruning technique
7. Perform Model complexity
8. Find the attribute with the highest info gain (A_Best)
9. Partition S into S1,S2,S3... according to the value of A_Best
10. Repeat the steps for S1, S2, S3
11. Classification : For each ti ϵ D, apply the DT to determine its class

3.2 Model Evaluation

In this section, a schematic overview is given of feature selection, cross validation, model complexity and reduced error pruning which is used for proposed algorithm.

Feature Selection

Here in the proposed system, feature selection is made with genetic search. 20 populations are generated and among them one subset is selected which is having minimum attributes. These attributes are most relevant and classification accuracy is dependent on these set of attributes. The application where Feature Selection is used, are text classification and web mining. Feature Selection builds the faster model by reducing the number of features, and also helps remove irrelevant, redundant and noisy features.

Reduced Error Pruning

Reduced Error Pruning is a technique which reduces the irrelevant branches from the decision tree. It is post pruning technique, where decision tree is first constructed then from the bottom up approach, one by one each leaf is removed and accuracy is measured. If there is no degradation in the accuracy after removing particular, the branch is removed permanently. By the use of Reduced Error Pruning technique, the problem of overfitting is resolved.

Cross Validation

Cross-Validation is the method of evaluating and comparing learning algorithms. Here numfolds are taken 10. It divides the data into two segments: 1. Train data, 2. Validate data.

Model Complexity

Now a day, some datasets are complex. So, our model should be complex equally. When complexity of the model increases by changing parameters, the accuracy is also increased.

4. EMPIRICAL RESULTS

In this section the properties of datasets are demonstrated. Finally results are presented with the new implementation.

Table 1: Dataset Details

Dataset Name	Attributes	Missing Val	Nominal/ numeric/ string/ Bool	Instances	Classes
Ionosphere	34	No	0,34,0,0	351	2
contact-lenses	4	No	4,0,0,0	24	3
Zoo	18	No	0,2,1,15	101	7
breast_cancer	10	Yes	0,2,3,5	286	2
Annealing	38	Yes	29,9,0,0	798	6
au1_1000	20	No	0,20,0,0	1000	2
weather.nominal	5	No	0,0,3,2	14	2
Diabetes	8	No	8,0,0,0	768	2

Now here it is finding results of the implemented system. Accuracy is improved here with less size generation of decision tree and the features selected too less compare to given in the dataset. Model is being more complex to understand complex dataset.

Table 2: Accuracy with all datasets: Comparison of Existing System & Implemented System

Data Set	Algorithm							
	Proposed System		Naïve bayes		SMO		Multilayer Perceptron	
Results	Correctly classified Instances	Relative Abs Error	Corr. classfi. Inst.	Rel. Abs. Error	Corr. classfi. Inst.	Rel. Abs. Error	Corr. classfi. Inst.	Rel. Abs. Error
Zoo	**98.09%**	**04.45%**	90.40%	56.00%	87.6%	70.30%	90.72%	21.46 %
Ionosphere	**91.16%**	**22.57%**	82.00%	37.00%	88.00%	24.00%	91.10%	20.30%
Contact-lenses	**83.33%**	**45.00%**	70.00%	67.00%	70.00%	83.30%	70.83%	54.80%
Breast Cancer	**73.00%**	**88.00%**	71.00%	78.00%	69.00%	72.00%	64.00%	84.00%
Annealing	**97.99%**	**05.97%**	86.02%	37.50%	97.40%	165.76%	98.11%	01.11%
Au1_1000	**74.90%**	**25.10%**	72.80%	27.20%	74.10%	25.90%	68.6%	31.40%
Weather .nominal	57.14%	70.00%	57.14%	91.00%	64.00%	75.00%	**71.00%**	**60.00%**
Iris	**96.00%**	**07.00%**	96.00%	07.69%	96.00%	51.00%	96.00%	07.35%
Diabetes	75.00%	71.00%	**76.0%**	**62.0%**	77.00%	49.00%	75.00%	65.00%

Below figures is providing comparison of proposed system with existing algorithms in graphical representation with the statistics given in table 2. Assumption is made that PS is proposed system, NB is Naive bayes algorithm and MP is Multilayer perception.

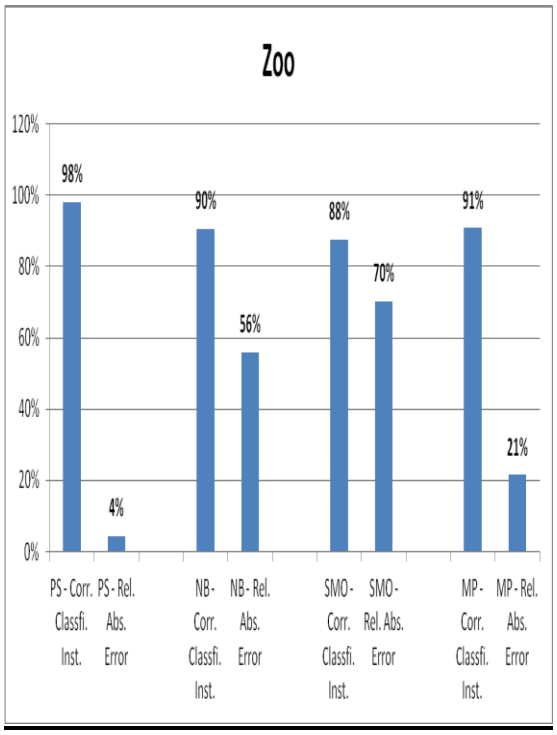

Fig 1: Accuracy with Zoo dataset

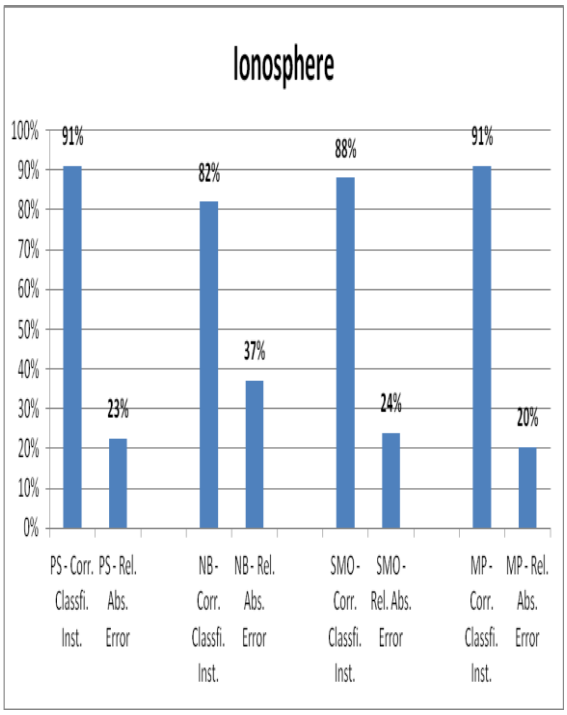

Fig 2: Accuracy with Ionosphere

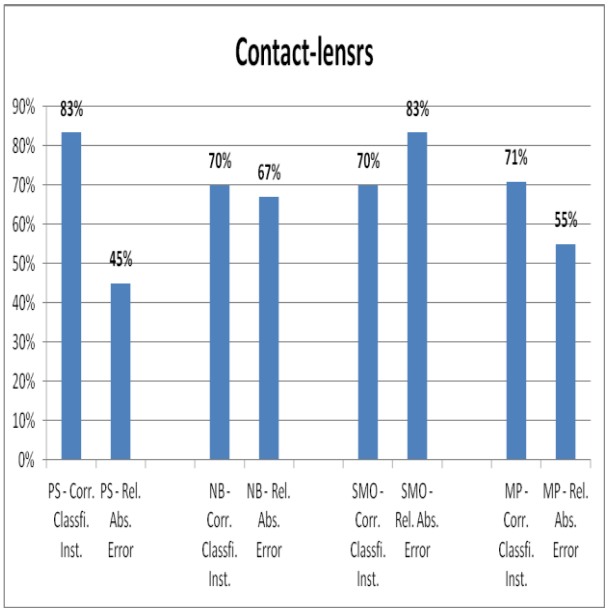

Fig3: Accuracy with Contact-lenses dataset

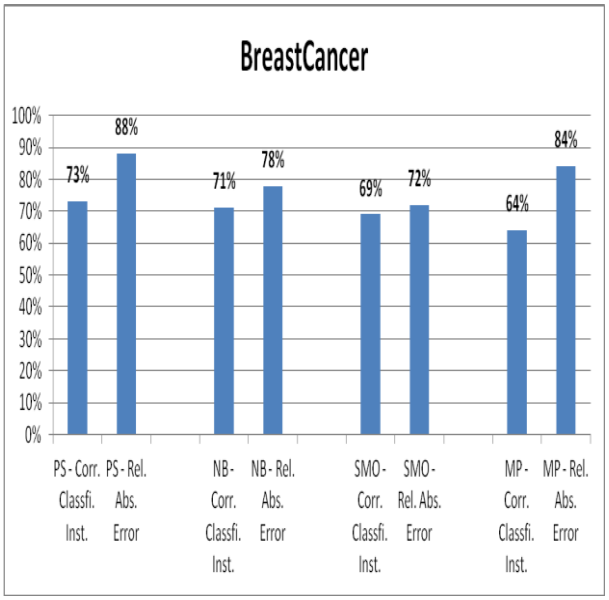

Fig 4: Accuracy with Breast Cancer dataset

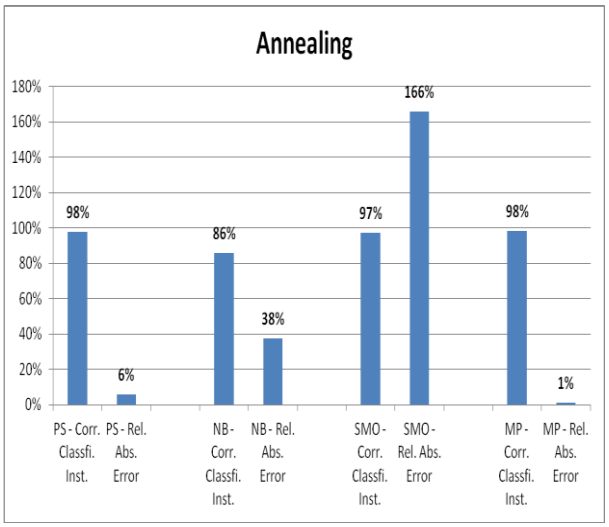

Fig 5: Accuracy with Annealing dataset

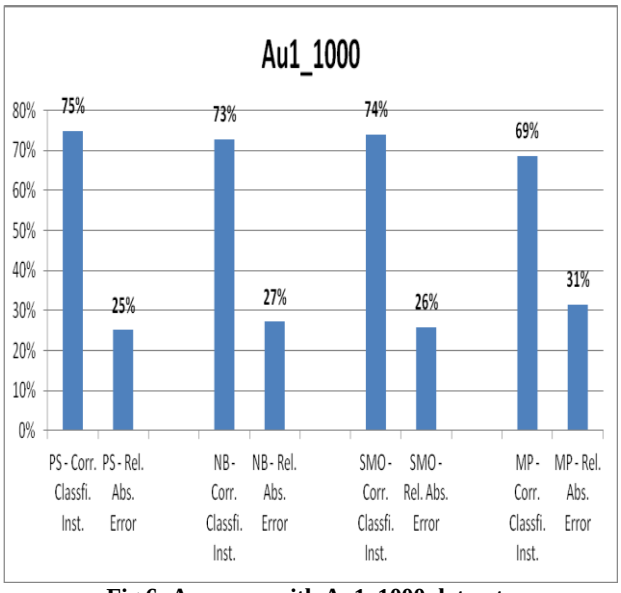

Fig 6: Accuracy with Au1_1000 dataset

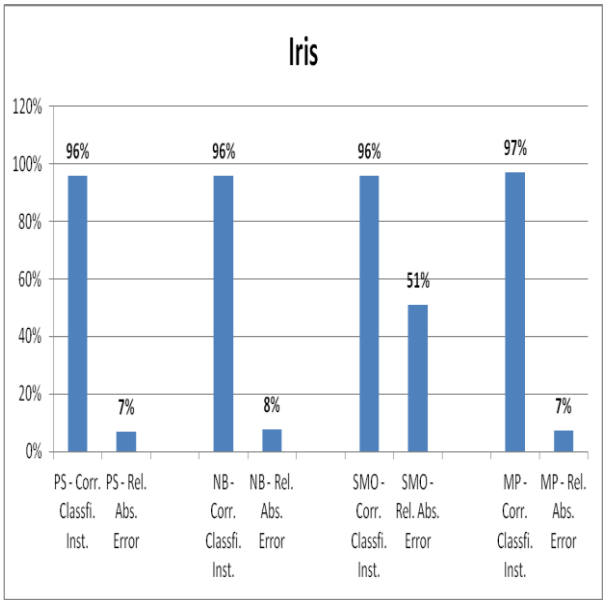

Fig 7: Accuracy with iris dataset

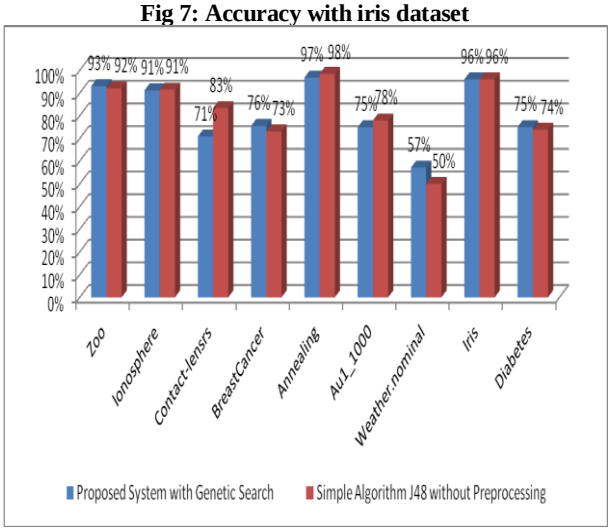

Fig 8: Accuracy with Feature Selection: Comparison of Existing System & Implemented System

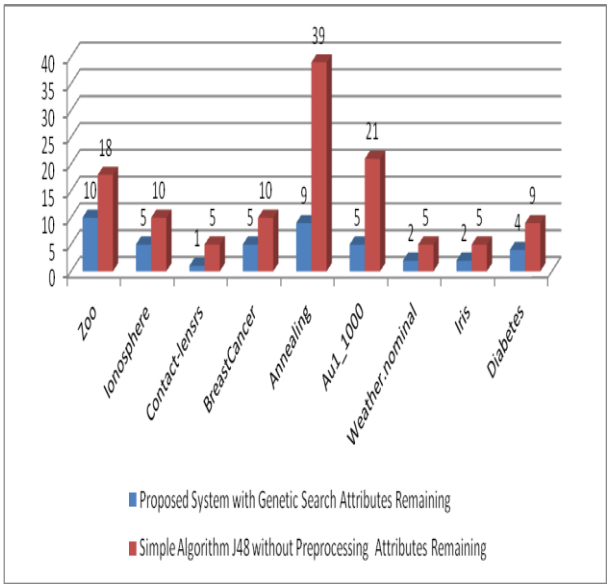

Fig 9: Remaining Attributes with Feature Selection: Comparison of Existing System & Implemented System

4. CONCLUSION & FUTURE WORK

The important task of classification process is to classify new and unseen sample correctly. By the changes done in algorithm, the classification accuracy is improved by implementing the diversities of algorithm using RGUI with weka packages. The gain in accuracy of the solution is acquired by variation of 1-3%, the final output of this modification is more under stable.

All the comparative analysis with the Implemented system is done in the performance study. It can be said for almost all the dataset Implemented system is proved more accurate compare to others. Implemented system is having quite better results. It is giving accuracy with a variation of 1-3% in 7 out of 9 datasets.

As the further scope, the predictive accuracy may still be improved by investigating other kinds of methods. An algorithm based on the input parameter combination can also be investigated for better results.

5. REFERENCES

[1] Sohag Sundar Nanda, Soumya Mishra, Sanghamitra Mohanty, Oriya Language Text Mining Using C5.0 Algorithm, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (1) , 2011

[2] cTomM.Mitchel, McGrawHil, Decision Tree Learning, Lecture slides for textbook Machine Learning, , 197

[3] Zuleyka Díaz Martínez, José Fernández Menéndez, Mª Jesús Segovia Vargas, See5 Algorithm versus Discriminant Analysis, Spain.

[4] XindongWu • Vipin Kumar • J. Ross Quinlan • Joydeep Ghosh • Qiang Yang • Hiroshi Motoda • Geoffrey J. McLachlan • Angus Ng • Bing Liu • Philip S. Yu • Zhi-Hua Zhou • Michael Steinbach • David J. Hand • Dan Steinberg, Top 10 algorithms in data mining, © Springer-Verlag London Limited 2007

[5] J.R, QUINLAN , Induction of Decision Trees, New South Wales Institute of Technology, Sydney 2007, Australia

[6] Rulequest Research, "Data Mining Tools See5 and C5.0, http:/lwww.rulequest.com/see5-info.html, 1997-2004

[7] Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner "Decision Trees— What Are They?"

[8] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong

[9] S. B. Kotsiantis, Department of Computer Science and Technology "Supervised Machine Learning: A Review of Classification Techniques" - , University of Peloponnese, Greece

[10] Classification: basic Concepts, Desision Tree, and model evalution

[11] Terri Oda **,** Data Mining Project, April 14, 2008

[12] Matthew N. Anyanwu manyanwu, Sajjan G. Shiva sshiva, Comparative Analysis of Serial Decision Tree Classification Algorithms

[13] Maria Simi , Decision tree learning

[14] Osmar R. Zaïane, 1999, Introduction to Data Mining, University of Alberta

[15] J. R. B. COCKETT, J. A. HERRERA, Decision tree reduction, University of Tennessee, Knoxville, Tennessee

[16] Hendrik Blockeel , Jan Struyf, Efficient Algorithms for Decision Tree Cross-validation, Department of Computer Science, Katholieke Universiteit Leuven, Belgium

[17] S. Rasoul Safavian and David Landgrebe, A Survey of Decision Tree Classifier Methodology, School of Electrical Engineering Purdue University, West Lafayette

[18] Floriana Esposito, Donato Malerba, and Giovanni Semeraro, A Comparative Analysis of Methods for Pruning Decision Trees, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 19, NO. 5, MAY 1997

[19] Paul E. Utgoff, Neil C. Berkman, Jeffery A. Clouse, Decision Tree Induction Based on Efficient Tree Restructuring, Department of Computer Science, University of Massachusetts, Amherst, MA 01003

[20] Niks, Nikson , Decision Trees, Introduction to machine learning,

[21] Ron Kohavi Ross Quinlan , Decision Tree Discovery, Blue Martini Software 2600 Campus Dr. Suite 175, San Mateo, CA & Samuels Building, G08 University of New South Wales, Sydney 2052 Australia

[22] Paul E. Utgoff , Incremental Induction of Decision Trees, Department of Computer and Information Science, University of Massachusetts, Amherst, MA 01003

[23] Matti K¨a¨ari¨ainen , Tuomo Malinen, Tapio Elomaa, Selective Rademacher Penalization and Reduced Error Pruning of Decision Trees, Department of Computer Science, University of Helsinki, Institute of Software Systems, Tampere University of Technology, Tampere, Finland

[24] Michael Kearns, Yishay Mansour, A Fast, Bottom-Up Decision Tree Pruning Algorithm with Near-Optimal Generalization, AT&T Labs, Tel Aviv University

[25] Zijian Zheng, Constructing new attributes for decision tree learning, Basser Department of Computer Science, the university of Sydney, Australia

[26] Emily Thomas, DATA MINING: DEFINITIONS

AND DECISION TREE EXAMPLES, Director of Planning and Institutional Research, State University of New York
[27] Kurt Hornik, The RWeka Package August 20, 2006
[28] Zhengping Ma, Eli Lilly and Company, Data mining in SAS® with open source software, SAS Global Forum 2011
[29] Simon Urbanek, Package 'rJava', Jan 2, 2012
[30] M. Govindarajan, Text Mining Technique for Data Mining Application, World Academy of Science, Engineering and Technology 35 2007
[31] A S Galathiya, AP Ganatra, CK Bhensdadia, An Improved decision tree induction algorithm with feature selection, cross validation, model complexity & reduced error pruning, IJSCIT march 2012.

6. AUTHORS

Avni S. Galathiya is a student of Master of Technology in computer Engineering at Dharmsinh Desai University, Nadiad, Gujarat, India. She is also an Assistant Professor at R. C. Technical Institute of Computer department, Ahmedabad, Gujarat, India. She has received her B.E. Computer Engineering degree from Dharmsinh Desai University, Nadiad, Gujarat, India in 2007. She has joined M. Tech. at Dharmsinh Desai University, Nadiad, Gujarat, India in 2010. Her current research interest includes Data Mining with classification.

Amit P. Ganatra (B.E.-'00-M.E. '04-Ph.D.* '11) has received his B.Tech. and M.Tech. degrees in 2000 and 2004 respectively from Dept. of Computer Engineering, DDIT-Nadiad from Gujarat University and Dharmsinh Desai University, Gujarat and he is pursuing Ph.D. in Information Fusion Techniques in Data Mining from KSV University, Gandhinagar, Gujarat, India and working closely with Dr.Y.P.Kosta (Guide). He is a member of IEEE and CSI.

His general research includes Data Warehousing, Data Mining and Business Intelligence, Artificial Intelligence and Soft Computing. In these areas, he is having good research record and published and contributed over 70 papers (Author and Co-author) published in referred journals and presented in various international conferences. He has guided more than 90 industry projects at under graduate level and 47 dissertations at Post Graduate level.

He is concurrently holding Associate Professor (Jan 2010 till date), Headship in computer Engineering Department (since 2001 to till date) at CSPIT, CHARUSAT and Deanship in Faculty of Technology-CHARUSAT (since Jan 2011 to till date), Gujarat. He is a member of Board of Studies (BOS), Faculty Board and Academic Council for CHARUSAT and member of BOS for Gujarat Technological University (GTU).

C.K. Bhensdadia is Professor and Head of Department of Computer Engineering at the Dharmsinh Desai University, Nadiad- Gujarat, India. He received a B.E. degree from Dharmsinh Desai University, Nadiad -Gujarat, India in 1990, and M.Tech. degree from IIT Mumbai in 1996. He is currently pursuing Ph.D. His main areas of research are the Genetic Algorithms and Data Mining. He has presented numerous papers on these topics in International journals & Conferences. He is also principal investigator for many research projects, sponsored by Govt. of India.