International Journal of Control and Automation
Vol. 8, No. 7 (2015), pp. 325-334
http://dx.doi.org/10.14257/ijca.2015.8.7.32

# An Efficient Approach for Knowledge Discovery in Decision Trees using Inter Quartile Range Transform

Bhanu Prakash Battula, KVSS Rama Krishna and Tai-hoon Kim*

*Vignan College, Andhra Pradesh, India*
*Vignan College, Andhra Pradesh, India*
*Department of Convergence Security, Sungshin Women's University,*
*249-1, Dongseon-dong 3-ga, Seoul, 136-742, Korea*
*battulaphd@gmail.com, ksai.mb@gmail.com, *taihoonn@daum.net*

***Abstract***

*Data mining and knowledge discovery is used for discovery of hidden knowledge from large data sources. Decision trees are one of the most famous classification techniques with simple and efficient generalization technique. This paper presents a new decision tree algorithm IQ Tree for classification problem. The IQ Tree assumes using an inter quartile range conversion of attributes with C4.5 as the base algorithm for performing induction can improve all the measures such as accuracy, tree size.*

***Keywords:*** *Data Mining, Classification, Decision Tree, inter quartile range*

## 1. Introduction

In Machine Learning community, and in data mining works, classification has its own importance. Classification is an important research application field in the data mining [1].

A decision tree gets its name because it is shaped like a tree and can be used to make decisions. ―Technically, a tree is a set of nodes and branches and each branch descends from a node to another node. The nodes represent the attributes considered in the decision process and the branches represent the different attribute values. To reach a decision using the tree for a given case, we take the attribute values of the case and traverse the tree from the root node down to the leaf node that contains the decision. "A critical issue in artificial intelligence (AI) research is to overcome the so-called ―knowledge-acquisition bottleneck [2]" in the construction of knowledge-based systems. Decision tree can be used to solve this problem. Decision trees can acquire knowledge from concrete examples rather than from experts [3]. In addition, for knowledge-based systems, decision trees have the advantage of being comprehensible by human experts and of being directly convertible into production rules [4].

One of the main problems in the effective operation of decision trees is its complexity. The complexity of the decision trees should be minimized to have better generalization. In this paper, the statistical procedure concerned with elucidating the covariance structure of a set of variables and outlier detection techniques are introduced to provide improved performance. The rest of this paper was organized as follows: The related work is given in Section 2. The proposed algorithm is discussed in Section 3. In Section 4, the details of experimental framework are presented. Simulation results are listed in Section 5 and conclusion is presented in final section.

* Corresponding Author

ISSN: 2005-4297 IJCA

## 2. Literature Review

In Data mining, the problem of decision trees has also become an active area of research. In the literature survey of decision trees we may have many proposals on algorithmic, data-level and hybrid approaches. The recent advances in decision tree learning have been summarized as follows:

A parallel decision tree learning algorithm expressed in MapReduce programming model that runs on Apache Hadoop platform is proposed by [5].A new adaptive network intrusion detection learning algorithm using naive Bayesian classifier is proposed by [6]. A new hybrid classification model which is established based on a combination of clustering, feature selection, decision trees, and genetic algorithm techniques is proposed by [7]. A novel roughest based multivariate decision trees (RSMDT) method in which, the positive region degree of condition attributes with respect to decision attributes in rough set theory is used for selecting attributes in multivariate tests is proposed by [8].

A novel splitting criteria which chooses the split with maximum similarity and the decision tree is calledmstreeis proposed by [9]. An improvedID3 algorithm and a novel class attribute selection method based on Maclaurin-Priority Value First method is proposed by [10]. A modified decision tree algorithm for mobile user classification, which introduced genetic algorithm to optimize the results of the decision tree algorithm, is proposed by [11].A new parallelized decision tree algorithm on a CUDA (compute unified device architecture), which is a GPGPU solution provided by NVIDIA is proposed by [12]. A Stochastic Gradient Boosted Decision Trees based method is proposed by [13]. A modified Fuzzy Decision Tree for the fuzzy rules extraction is proposed by [14].

A decision tree approach is applied for finding solutions to diagnose the disease by analyzing the patterns [15]. C4.5 approach is used for selection of attributes for prediction of credit default [16].Decision tree model is used for pre-processing, feature extraction and classification of ECG signals for medical applicability [17].

In [18] author proposes a classifier which can be built independently and without Bulky Business Intelligence software to effectively forecast future occurrences of any phenomena. In [19] author proposes framework and process models that provide user results, graphs and trees that help user in fraud clustering and classification. After analyzing the existing recent literature, we found that new classification algorithm for varied data source is the need of the hour.

## 3. The Proposed Approach

In this Section, we investigate to propose a new decision tree induction algorithm known as Inter Quartile (IQ) Range Decision Tree. Our IQ Decision Tree induction method depends on inter quartile ranges which was described in the above section. We assume that the subset of the training data is small, *i.e.*, it is computationally cheap to act on such a set in a reasonable time.

We focus on a set of improved attribute range filters using attribute transformations. Next, we try to adapt and deploy them as IQ Tree components. Since the IQ Tree scheme is based on a restricted list of candidates, this list could be represented by features that seems to be relevant or those that might provide incremental usefulness to the selected feature subset. For the IQ Tree construction stage we opt for selection scheme capable of generating attribute ranking. Hence, the weights associated to features will serve as one of the selection criterion in the new heuristic function for inducing decision trees. The next stage of IQ Tree tries to consider both entropy and weights for splitting of attributes. The quality of solution fine-tuning, mainly, depends on the nature of the filter involved and the parameters of attribute transformation. The following algorithm, detail different design alternatives for both attributes transform and filter procedure search for IQ Tree components. The algorithm for IQ decision tree is shown below:

 

**Algorithm IQ: New Decision Tree (D, A, RGR)**
**Input:**D – Data Partition
A – Attribute List
GR – Gain Ratio
**Output:** Decision Tree Measures – Accuracy, Tree Size.
**Procedure:**
1. Create a node N
2. If samples in N are of same class, C then
3. return N as a leaf node and mark class C;
4. If A is empty then
5. return N as a leaf node and mark with majority class;
else
6. ( , ) = apply Inter Quartile Range (D, A)
7. apply Gain Ratio( , )
8. label root node N as f(A)
9. for each outcome j of f(A)do
10. subtree j =New Decision Tree( j, )
11. connect the root node N to subtree j
12. endfor
13. endif
14. endif
15. Return N

**The Algorithm: IQ Tree can be Explained as Follows:**

The inputs to the algorithm are data partition "D", attribute set "A" and splitting criteria gain ratio "GR". The output of the algorithm will be the average measures such as accuracy and tree size produced by the IQ Tree method. The algorithm begins with the create node for same class. In the next stage, attribute rages are applied to the inter quartile for transformation. In the later on stage, the transformed dataset is applied for the splitting criteria gain ratio for decision tree induction. The induced decision tree is applied for the tree pruning process for generalization of the tree. In the final the measures for decision tree validation *i.e.*, accuracy ad tree size are generated.

## 4. Experimental Setup and Algorithms Compared

In the study, we have considered 14 data-sets which have been collected from the UCI [20] machine learning repository web sites. The complete details regarding all the datasets can be obtained from UCI Machine Learning Repository.

We have obtained the accuracy and tree size metric estimates by means of a 10-fold cross-validation. That is, the data-set was split into ten folds, each one containing 10% of the patterns of the dataset. For each fold, the algorithm is trained with the examples contained in the remaining folds and then tested with the current fold. Table 1 summarizes the properties of the selected datasets.

**Table 1. The 14 UCI Datasets and Their Properties**

| S. No. | Dataset | Instances | Missing values | Numeric attributes | Nominal attributes | Classes |
|---|---|---|---|---|---|---|
| 1. | Balance-scale | 625 | No | 4 | 0 | 3 |
| 2. | Breast-cancer | 286 | Yes | 0 | 9 | 2 |
| 3. | Credit-g | 1,000 | No | 7 | 13 | 2 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 4. | Pima diabetes | 768 | No | 8 | 0 | 2 |
| 5. | Glass | 214 | No | 9 | 0 | 6 |
| 6. | Heart-statlog | 270 | No | 13 | 0 | 2 |
| 7. | Ionosphere | 351 | No | 34 | 0 | 2 |
| 8. | Iris | 150 | No | 4 | 0 | 3 |
| 9. | Lymphography | 148 | No | 3 | 15 | 4 |
| 10. | Sonar | 208 | No | 60 | 0 | 2 |
| 11. | Vehicle | 846 | No | 18 | 0 | 4 |
| 12. | Vowel | 990 | No | 10 | 3 | 11 |
| 13. | Waveform | 5,000 | No | 41 | 0 | 3 |
| 14. | Zoo | 101 | No | 1 | 16 | 7 |

The algorithms used in the experimental study and their parameter settings, which are obtained from the WEKA [21] software tool. Several decision tree methods have been selected and compared to determine whether the proposal is competitive in different domains with the other approaches. Algorithms are compared on equal terms and without specific settings for each data problem. The parameters used for the experimental study in all decision tree methods are the optimal values from the tenfold cross-validation, and they are now detailed in Table 2.

**Table 2. Experimental Settings for Standard Decision Tree Algorithms**

| Algorithm | Parameter | Value |
|---|---|---|
| C4.5 | confidence factor | 0.25 |
| | min number of objects | 2.0 |
| REP | maximum depth | no restriction |
| | min number of objects | 2.0 |
| | min variance proportion | 0.001 |
| CART | number of folds pruning | 5 |
| | min number of objects | 2.0 |
| NB Tree | technique used at leaves | naive bayes |

## 5. Results and Discussion

We experimented with 10 standard datasets from the UCI repository, these datasets are standard benchmark imbalanced datasets used in the context of supervised learning. The goal is to examine whether the proposed IQ Tree achieve better predictive performance than a number of existing standard learning algorithms. We compared the above method with the C4.5, BPN, REP, CART and NB Tree state-of-the-art metric learning algorithms. In all the experiments we estimate accuracy using 10-fold cross-validation and control for the statistical significance of observed differences using t-test (sig. level of 0.05).

In Table 3 and 4, we present the results of the comparison between C4.5, BPN, REP, CART, NB Tree and IQ Tree. From these results we can make several observations. The developed IQ Tree compared with C4.5, REP, CART, NB Tree generally given

 

competitive results; the advantage of our methods is most visible in the balance, diabetes, glass, ionosphereand sonar datasets. Finally, the method that most often win is IQ Tree.

**Table 3. Summary of Tenfold Cross Validation Performance for Accuracy on All the Datasets**

| S. No. | Datasets | C4.5 | REP | CART | NB Tree | IQ Tree |
|---|---|---|---|---|---|---|
| 1. | Balance-scale | 77.82● | 78.54● | 78.73● | 75.96● | 81.53 |
| 2. | German_credit | 71.25● | 72.02● | 73.43● | 74.64● | 79.15 |
| 3. | Pima_diabetes | 74.49● | 74.46● | 74.56● | 74.96● | 77.21 |
| 4. | Glass | 67.63● | 65.54● | 71.26● | 69.84● | 73.43 |
| 5. | Heart-statlog | 78.15● | 76.15● | 78.07● | 80.93● | 91.61 |
| 6. | Ionosphere | 89.74● | 89.46● | 88.87● | 90.03● | 91.23 |
| 7. | Iris | 94.73● | 93.87● | 94.20● | 93.47● | 95.52 |
| 8. | Lymphography | 75.84● | 75.33● | 77.21● | 81.90● | 82.23 |
| 9. | Sonar | 73.61● | 72.69● | 70.72● | 77.11● | 78.16 |
| 10. | Vehicle | 72.28● | 70.18● | 69.91● | 70.98● | 72.36 |
| 11. | Vowel | 80.20● | 66.67● | 79.61● | 92.35● | 93.19 |
| 12. | Waveform | 75.25● | 76.57● | 76.65● | 79.84● | 80.32 |
| 13. | Zoo | 92.61● | 40.61● | 40.61● | 94.73 ○ | 92.65 |
| **Win/Tie/Loss** | | **(13/0/0)** | **(13/0/0)** | **(13/0/0)** | **(13/0/1)** | |

● Bold dot indicates the win of proposed method; ○ Empty dot indicates the loss of proposed method.

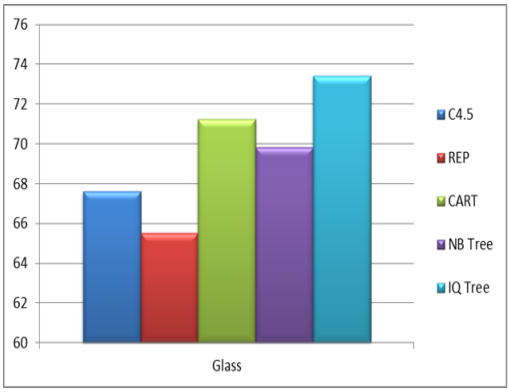


Figure 1. (a)

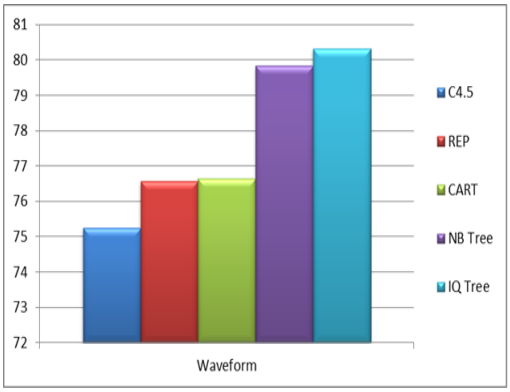


Figure 1. (b)

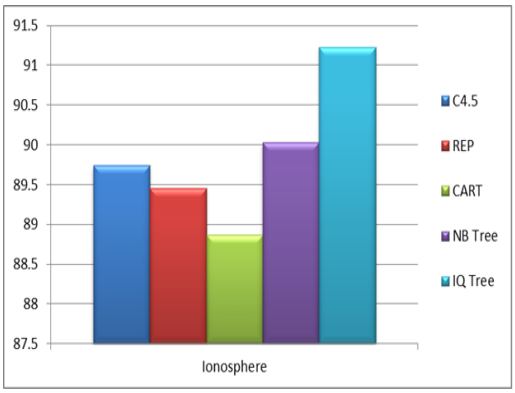


Figure 1. (c)

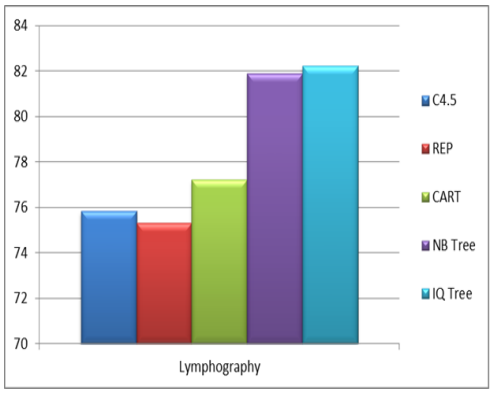


Figure 1. (d)

**Figure 1. Test Results on Accuracy between the C4.5, REP, CART, NB Tree and IQ Tree on 1(a) glass 1(b) Waveform 1(c)Ionosphere and 1(d) Lymphography Dataset**

**Table 4. Summary of Tenfold Cross Validation Performance for Tree Size on All the Datasets**

| S. No. | Datasets | C4.5 | REP | CART | NB Tree | IQ Tree |
|---|---|---|---|---|---|---|
| 1. | Balance-scale | 82.20● | 42.36○ | 55.28● | 17.38○ | 52.43 |
| 2. | German_credit | 126.85● | 76.81● | 24.46○ | 12.07○ | 61.76 |
| 3. | Pima_diabetes | 43.40○ | 30.98○ | 17.36○ | 5.18○ | 57.82 |
| 4. | Glass | 46.16● | 19.70○ | 21.16○ | 10.0○ | 42.35 |
| 5. | Heart-statlog | 34.64● | 14.78● | 15.36● | 9.62○ | 11.66 |
| 6. | Ionosphere | 26.74● | 8.76○ | 8.42○ | 16.20○ | 24.52 |
| 7. | Iris | 8.28● | 5.84○ | 7.40○● | 4.38○ | 6.41 |
| 8. | Lymphography | 28.00● | 11.46○ | 13.92○ | 10.24○ | 17.65 |
| 9. | Sonar | 27.90● | 10.20○ | 10.50○ | 13.74○ | 19.85 |
| 10. | Vehicle | 138.0● | 58.52○ | 92.54○ | 57.70○ | 93.76 |
| 11. | Vowel | 209.81○ | 254.36● | 171.74○ | 70.10○ | 217.58 |
| 12. | Waveform | 591.94● | 167.24○ | 98.32○ | 94.48○ | 211.62 |
| 13. | Zoo | 15.70● | 1.00○ | 1.00○ | 8.34○ | 7.41 |
| **Win/Tie/Loss** | | **(11/0/2)** | **(3/0/10)** | **(3/0/10)** | **(0/0/13)** | |

● Bold dot indicates the win of proposed method; ○ Empty dot indicates the loss of proposed method.

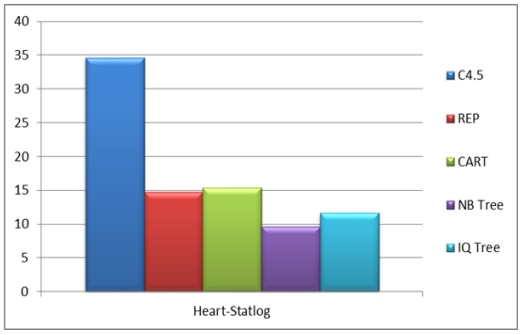


Figure 2. (a)

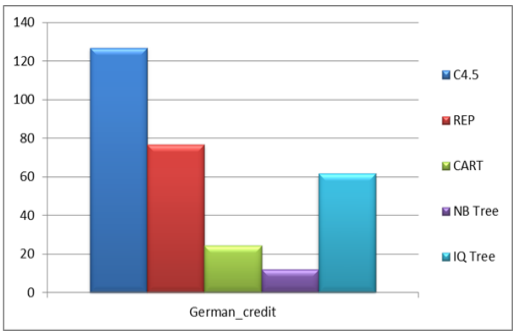


Figure 2. (b)

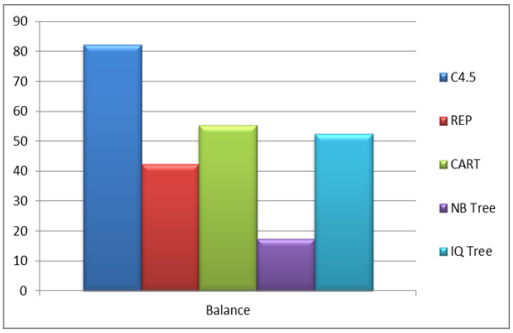


Figure 2. (c)

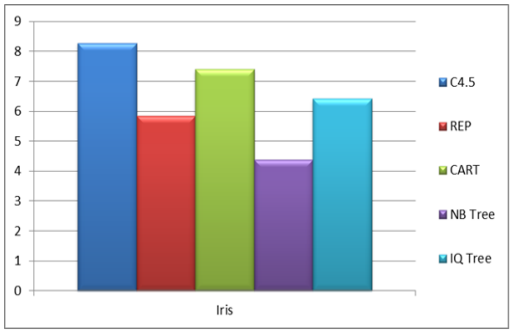


Figure 2. (d)

**Figure 2. Test Results on Tree Size between the C4.5, REP, CART, NB Tree and IQ Tree on 2(a) Heart-statlog2 (b) Grema Credit 2 (c) Balance and 2 (d) Iris Dataset**

Table 3 and 4 presents the comparative results of proposed algorithm IQ Tree against C4.5, REP, CART and NB Tree. The value in the table; example: "11/0/2" specifies that the proposed algorithm has registered 11 wins, 0 ties and 2 losses against compared algorithm for that specified measure. One can observe from the Table 3 and 4 that our proposed algorithms have registered good number of wins against the compared algorithms on all the datasets.

These results suggest that in the majority of the high dimensional datasets, the feature interactions are not important, and hence the methods that do not account for feature interactions have in general better performances. Alternatively, it might suggest that stronger regularization is needed. Moreover, it is interesting to note that the cases for which the good performance are difficult classification problems from the UCI datasets. This hints that there might be a bias of method development towards methods that perform well on UCI datasets; however, one can argue that they are really representative of the real world.

These results are remarkable since IQ Tree, which is based on a simple idea, performs equally well as the more elaborate standard learning algorithm that has been reported to consistently outperform other metric learning techniques over a number of non-trivial learning problems. Finally, we mention that the surprisingly poor performance of IQ Tree on sonar, vehicle and vowel datasets in Tables 4, might be explained by the fact that its conversion function is not convex and hence it is sensitive to the unique properties of the datasets.

In overall, from all the tables and figures we can conclude that our proposed IQ Tree have given good results when compared to benchmark algorithms. The unique properties of datasets such as size of the dataset and the number of attributes will also effect on the results of our proposed IQ Tree. The above given results are enough to project the validity of our approach and more deep analysis should be done for further analysis.

## 6. Conclusion

This paper presents a new decision tree algorithm IQ Tree for class classification problem. The IQ Tree assumes using an inter quartile range conversion of attributes with C4.5 as the base algorithm for performing induction can improve all the measures such as accuracy, tree size.

The experiments conducted with IQ Tree specify that improved performance can be achieved. We have conducted experiments on 10 datasets from UCI which suggest that IQ Tree can quickly remove redundant, irrelevant and weak attributes as long as the properties of the dataset are normal. Excellent improvement in measures on some natural domain datasets shows the compatibility of IQ Tree approach on real-time applications. One of the shortcomings seen in IQ Tree is when used for datasets with unique properties; Because IQ Tree will not consider unique properties of datasets for removing instances from data source. Finally, we can conclude that IQ Tree can be a good contribution as a decision tree induction method for efficient learning of the varied datasets.

## References

[1] J. Hu, J. Deng and M. Sui, "A New Approach for Decision Tree Based on Principal Component Analysis", International Conference on Computational Intelligence and Software Engineering, **(2009)**.
[2] Bergsma and Shane, "Large-scale semi-supervised learning for natural language processing", University of Alberta, **(2010)**.
[3] Durkin, Jack and J. Durkin, "Expert systems: design and development", Prentice Hall PTR, **(1998)**.
[4] Quinlan and J. Ross, "C4. 5: programs for machine learning", Elsevier, **(2014)**.
[5] V. Purdila and S.-G. Pentiuc, "MR-Tree - A Scalable Map Reduce Algorithm for Building Decision Trees", Journal of Applied Computer Science & Mathematics, vol. 16, no. 8, **(2014)**, pp. 16-19.
[6] Farid, Md. Dewan, N. Harbi and M. Z. Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection", International Journal of Network Security & Its Applications, vol. 2, no. 2, **(2010)**, pp. 12-25.
[7] M. Khanbabaei and M. Alborzi, "The use of Genetic Algorithm, Clustering and Feature Selection Techniques in Construction of Decision Tree Models for Credit Scoring", International Journal of Managing Information Technology (IJMIT), vol. 5, no. 4, **(2013)**, pp. 13-31.
[8] W. Dianhong, X. Liu, L. Jiang, X. Zhang and Y. Zhao, "Rough set approach to multivariate decision trees inducing", Journal of Computers, vol. 7, no. 4, **(2012)**, pp. 870-879.
[9] Z. Xinmeng and S. Jiang, "A Splitting Criteria Based on Similarity in Decision Tree Learning", Journal of software, vol. 7, no. 8, **(2012)**, pp. 1775-1782.
[10] W. Ying, X. Peng and J. Bian, "Computer Crime Forensics Based on Improved Decision Tree Algorithm", Journal of Networks, vol. 9, no. 4, **(2014)**, pp. 1005-1011.
[11] D.-s. Liu and S.-j. Fan, "A Modified Decision Tree Algorithm Based on Genetic Algorithm for Mobile User Classification Problem", The Scientific World Journal, **(2014)**, pp. 11.
[12] Lo, Win-Tsung, Y.-S. Chang, R.-K. Sheu, C.-C. Chiu and S.-M. Yuan, "CUDT: a CUDA based decision tree algorithm", The Scientific World Journal, **(2014)**.
[13] C. Tarun and J. Vajpai, "Fault Diagnosis in Benchmark Process Control System using Stochastic Gradient Boosted Decision Trees", International Journal of Soft Computing and Engineering, vol. 1, **(2011)**, pp. 98-101.
[14] S. V. S. G. Devi, "Fuzzy Rule Extraction for Fruit Data Classification", COMPUSOFT, An international journal of advanced computer technology, vol. 2, no. 12, **(2013)**, pp. 400-403.
[15] I. Aiswarya, S. Jeyalatha and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques", International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 1, **(2015)**, pp. 1-14.
[16] M. Bach, J. Zoroja and V. Šimicevic, "Attribute Selection for Predicting Credit Default with Decision Trees", World Academy of Science, Engineering and Technology, International Science Index, Economics and Management Engineering, vol. 2, no. 6, **(2015)**, pp. 1297.
[17] P. Mondal and K. Mali, "Cardiac Arrhythmias Classification using Decision Tree", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, **(2015)**, pp. 540-542.
[18] S. Manohar, A. Mittal, S. Naik and A. Ambre, "A Dynamic Classifier using Decision Tree Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, no. 1, **(2015)**, pp. 628-631.
[19] A. Muhammad, K. A. Alam and M. Hussain. "Crime Mining: A Comprehensive Survey", International Journal of u-ande-Service, Science and Technology, vol. 8, no. 2, **(2015)**, pp. 357-364.
[20] A. Asuncion and D. Newman, "UCI Repository of Machine Learning Database (School of Information and Computer Science)", Irvine, CA: Univ. of California [Online]. Available:http://www.ics.uci.edu/~mlearn/MLRepository.html.

 

[21] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd edition Morgan Kaufmann, San Francisco.
[22] Quinlan and J. Ross, "Induction of decision trees", Machine learning, vol. 1, no. 1, **(1986)**, pp. 81-106.
[23] L. Breiman, J. Friedman, R. A. Olshen and C. J. Stone, "Classification and Regression Trees (Wadsworth, Belmont, CA, 1984)", Proceedings of the Thirteenth International Conference, Bari, Italy, **(1996)**.
[24] Chawla, V. Nitesh, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority over-samplingtechnique", Journal of artificial intelligence research, vol. 16, no. 1, **(2002)**, pp. 321-357.