Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings XX (2016) XXX–XXX

www.materialstoday.com/proceedings

PMME 2016

An Efficient Decision Tree for Imbalance data learning using Confiscate and Substitute Technique

Salina Adinarayana[a,*] , E.Ilavarasan[b]

[a]*Department of IT, Shri Vishnu Engineering College for Women, Bhimavaram, Andhra Pradesh, India*

[b]*Professor, Department. of CSE, Pondicherry Engineering College, Pondicherry, India.*

Abstract

Data mining and knowledge discovery is the process of discovering knowledge from the real world datasets. One of the limitations of the real world datasets is the existence of contamination in the dataset. The existing algorithms performance will degrade due to the contamination in the real world datasets in the form of noisy and missing values. In this paper, we propose a novel algorithm dubbed as Confiscate and Substitute Imbalance Data Learning (CSIDL) for better knowledge discovery from real world datasets. The process of confiscate is implemented in the majority subset for the removal of noisy, border line and missing instances and substitute of missing instances is done in the minority subset for improving the strength of the dataset. Experimental comparisons are done on six real world dataset with bench mark traditional algorithms. The results suggest that the proposed CSIDL algorithm performed better than the compared algorithms in terms of Accuracy, AUC, Precision and F-measure.

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

Keywords: Knowledge Discovery, Decision Trees, Confiscate and Substitute Imbalance Data Learning (CSIDL)

1. Introduction

In Data mining the two major approaches for knowledge discovery are Classification and Clustering. Classification is the process of classifying the labelled instances with the help of model built by the training data. Clustering is the process of grouping data by investigating the intrinsic properties of the instances. In Classification, decision trees are one of the traditional ad simple approaches for knowledge discovery. Decision tree follow the simple strategy of splitting data into different braches for building the model. In Decision Tree, one of the traditional and benchmark model is C4.5 [8]. In Data mining, the problem of decision trees had received a good amount of attention in recent years. The some of the advances in this field are as follows.In [1] author proposed the Chi-FRBCS-Big Data algorithm, a linguistic fuzzy rule-based classification system that uses the MapReduce framework

Selection and Peer-review under responsibility of International Conference on Processing of Materials, Minerals and Energy (July 29th – 30th) 2016, Ongole, Andhra Pradesh, India.

to learn and fuse rule bases. In [2] author presented the performance of ID3 classification and cascaded model with RBF network. In [3] author proposed a windowed regression over-sampling (WRO) method for oversampling of instances in the minority subset to change the class distribution through adding virtual samples. WRO not only reflects the additive effects but also reflects the multiplicative effect between samples.

In [4] author presented a review of existing solutions to the class-imbalance problem both at the data and algorithmic levels. In [5] author summarized a comprehensive study of different feature selection schemes in machine learning for the problem of mood classification in weblogs. A novel use of a feature set based on the affective norms for English words (ANEW) lexicon studied in psychology is also proposed. In [6] author presented a neural network-based finite impulse response extreme learning machine (FIR-ELM) for studying of medical datasets. In [7] author proposed a secure k-NN classifier over encrypted data in the cloud. The algorithm is used for solving the classification problem over encrypted data by protecting the confidentiality of data, privacy of user’s input query and hides the data access patterns. Obviously, there are many other algorithms which are not included in this literature. A profound comparison of the above algorithms and many others can be gathered from the references list.

In real time scenario, the preparation of dataset is a complex and tedious job. In this process there are many challenges to prepare ideal dataset which doesn’t have any misleading information such as noise, missing values. The misleading information will enter into the dataset due to many reasons. One of the challenges to the data mining research community is to efficiently use datasets with noisy information for proper knowledge discovery. This paper proposes an algorithm of one of its kind to address issues relating to efficient knowledge discovery from the real world datasets.

The paper is organized as follows. In Sect. 2, we propose a new framework for imbalance data learning known as CSIDL. Experimental design and evaluation criteria’s for decision tree learning is presented in section 3. Experimental results and discussions are presented in section 4. Finally, we conclude with Sect. 5 with an indication towards our future work.

2. The Proposed Method

In this section, the Confiscate and Substitute Imbalance Data Learning (CSIDL) approach is presented.
The CSIDL approach follows a unique pattern of solving the problem of class imbalance learning. In any dataset, we find noisy, missing and borderline instances. The definitions of these instances are as follows:

Noisy instances are those which may belong to any of the specified class but the intrinsic characteristics of those instances are very far from the intrinsic properties of the other vast number of instances in the same class. In most of the cases the noisy instances are entered in the dataset due to improper data collection, improper pre-processing and rare conditions in the real time scenario. The benefit provided by these instances is in fact less than the damage they made to the data collection. The traditional approaches will lose their path for building the proper model for classification when they encounter with the noisy instances. The identification of the noisy instances can be easily done by the mathematical analysis. In this research work we used the well-established distributive and probability theory for proper identification of the noisy instances for their removal.

Missing value instances are those where the values for one or more attributes are missing. The instances with zero or null values are not considered as missing values. The reason for the missing values is due to the non-availability of the data for data preparation. These instances may not give a proper picture for building the model for classification. The removal of missing value instances from the majority subset is a welcome sign for solving the problem of class imbalance learning. The missing values are represented in the instances as ‘?’. The identification and removal of the missing values are done in our CSIDL approach.

Borderline instances are those instances which are in the border region of the two or more classes. These instances are very dangerous for the efficient performance of the built model for classification. The removal of these instances will improve the performance of the model built.

In the majority subset the number of instances is to be eliminated for reducing the problem of class imbalance.

The noisy, missing and borderline instances are removed to form the improved majority subset.

In the minority subset the instances are to the increased so as to reduce the problem relating to class imbalance nature. The missing value instances present in the minority subset can be populated with the appropriate values. The novel technique used in this research for generating the appropriate values is by computing the mean of the existing values of the instances. The new mean value computed is replaced with the missing values in the instances of the minority subset. This technique will improve the quality of the minority model built there by giving scope for improved efficiency.

The improved majority and minority subsets are combined to form a strong and less imbalance dataset. In the next stage of the framework the C4.5 [8] algorithm is used as the base algorithm and the evaluation metrics are computed.

Suppose that the whole training set is *T*, the minority class is *P* and the majority class is *N*, and

$P = \{p1, p2 ,..., ppnum\}$, $N = \{n1,n2 ,...,nnnum\}$

Where *pnum* and *nnum* are the number of minority and majority examples. The detailed procedure of CSIDL is as follows.

Algorithm: CSIDL

Algorithm: New Decision Tree (D, A, GR)
Input: D – Data Partition
A – Attribute List
GR – Gain Ratio
D: A set of minor class examples *P*, a set of major class examples *N*, *jPj<jNj*, and *Fj*,the feature set, j > 0.

Output : A Decision Tree with Average Measure { AUC, Precision, F-Measure, TP Rate, TN Rate }

Procedure:

External selection Phase

Step 1: For every ni(i= 1,2,..., nnum) in the majority class N, we calculate its m nearest neighbours from the whole training set T. The number of majority examples among the m nearest neighbours is denoted by m' (0 ≤ m'≤ m) .

Step 2: If m/ 2 ≤ m'<m , namely the number of ni ’s minority nearest neighbours is larger than the number of its majority ones, ni is considered to be easily misclassified and put into a set MISCLASS.

MISSCLASS = m'

Remove the instances m' from the majority set.

Step 3: For every ni’ (i= 1,2,..., pnum’) in the minority class N, we calculate its m nearest neighbours from the whole training set T. The number of majority examples among the m nearest neighbours is denoted by m' (0 ≤ m'≤ m).

If m'= m, i.e. all the m nearest neighbours of ni are majority examples, ni’ is considered to be noise or outliers or missing values and are to be removed.

Step 4: The examples in minority set are the prominent examples of the minority class P, and we can see that PR⊆P . We set

PR= {p'1 ,p'2 ,..., p'dnum}, 0 ≤ dnum≤ pnum

Step 5: Substitute missing values from minority subset P with the mean value.

Building Decision Tree:

```
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;
6. else
7.         apply Gain Ratio(D,A)
8.         label root node N as f(A)
```

```
for each outcome j of f(A)do
subtree j =New Decision Tree(Dj,A)
        connect the root node N to subtree j
endfor
endif
endif
Return N
```

3. Experimental Design and Evaluation Criteria

The experimental simulation is done on the open source tool Weka [9]. The six dataset from UCI repository [10] are used for evaluating the CSIDL approach. The comparative study of the approach is done on the standard bench mark algorithms C4.5 [8], Reduced Error Pruning Tree (REP) [12], Classification and Regression Trees (CART) [13] and Naïve Bayes Tree (NB Tree) [14] in our experiments. The details of the dataset are provided in the table 1 below,

Table 1 The UCI datasets and their properties

S.no.	Dataset	Instances	Missing values	Attributes	Classes	Majority/Minority	IR
1.	Breast-cancer	286	Yes	9	2	201/85	2.36
2.	Crx	690	Yes	15	2	383/307	1.24
3.	Hepatitis	155	Yes	19	2	123/32	
4.	Horse-colic	368	Yes	22	2	232/136	
5.	House votes	435	Yes	16	2	267/168	
6.	Post Operative	90	Yes	8	2	66/24	

The experimental methodology used for experimental simulation is 10 fold cross validation. In 10 fold cross validation the data source is divided into 10 equal partitions. In each run, one of the folds is used for testing and remaining folds are used for training the model. The mean of 10 runs are used for computing of evaluation metrics such as accuracy, AUC, TP rate, TN rate etc… The framework for 10 fold cross validation is shown in the figure 1 below.

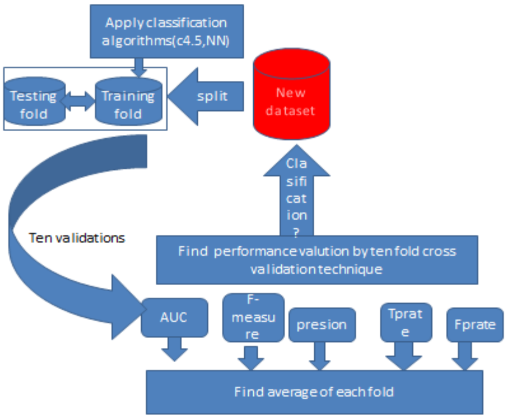
Apply classification algorithms(c4.5,NN)
Testing fold
Training fold
split
New dataset
Clasification?
Ten validations
Find performance valution by ten fold cross validation technique
AUC
F-measure
presion
Tprate
Fprate
Find average of each fold

Figure 1: Frame work for 10 Fold Cross Validation

The 10 fold cross validation technique splits the data into 10 folds and in each run it uses 9 folds for training and 10th fold for testing. The process is repeated 10 times and in each run the testing data is replaced with untested fold. In this paper, the researcher use AUC, Precision, F-measure, TP Rate and TN Rate as performance evaluation measures. Let us define a few well known and widely used measures:

Receiver Operating Characteristic (ROC) curve is the recent evaluation metric used for supervised learning dealing with imbalanced data study. This ROC curve can be used for projecting results depending upon the user perspective with different combinations of basic components such as true positives, false positives, true negatives and false negatives. The summary of the ROC curve can be given as the area under it, which is known as Area Under Curve (AUC). AUC can be computed simple as the micro average of TP rate and TN rate when only single run is available from the classification algorithm.

The Area under Curve (AUC) [15] measure is computed by,

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \qquad (1)$$

Or

$$AUC = \frac{TP_{RATE} + TN_{RATE}}{2} \qquad (2)$$

The Precision [15] measure is computed by,

$$\Pr ecision = \frac{TP}{(TP)+(FP)} \qquad (3)$$

The F-measure [15] Value is computed by,

$$F - measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \qquad (4)$$

The True Positive Rate [15] measure is computed by,

$$TruePositiveRate = \frac{TP}{(TP)+(FN)} \qquad (5)$$

The True Negative Rate [15] measure is computed by,

$$TrueNegativeRate = \frac{TN}{(TN)+(FP)} \qquad (6)$$

4. Results and Discussion

In this section, the results of the CSIDL approach are compared and discussed. The results are summarized as follows.

Table 2 shows the detailed experimental results of the mean classification accuracy of C4.5, REP, CART, NB Tree on all the data sets. From Table 2 we can see accuracy performance of CSIDL model that it can achieve substantial improvement over C4.5 on most data set (3 wins 1 tie and 1 loss) which suggests that the CSIDL model is potentially a good technique for decision trees. The CSIDL method can also gain significantly improvement over REP (4 wins 1 tie and 1 loss) and is comparable to two state-of-the-art technique for decision trees, CART (4 wins 1 tie and 1 losses) and NB Tree with all the wins (6 wins).

Table 2 Summary of tenfold cross validation performance for Accuracy on all the datasets

Datasets	**C4.5**	**REP**	**CART**	**NB Tree**	**CSIDL**
Breast-cancer	74.46±5.40○	69.03±6.29●	70.13±4.83●	71.47±6.63●	72.77±5.39
Crx	85.01±3.91●	84.28±4.19●	85.19±4.10●	85.36±4.48●	86.07±4.03
Hepatitis	79.22±9.57●	78.75±6.96●	77.10±7.12●	79.11±9.78●	79.96±9.54
Horse-colic	85.13±5.89●	85.02±5.75●	85.40±5.37●	82.33±6.40●	96.69±3.87
House votes	96.57±2.56	95.33±3.10	95.79±2.67	95.05±3.53●	96.81±2.94
Post Operative	72.33±6.22○	73.00±6.17○	73.22±5.49○	69.11±8.87●	70.92±7.01

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

Table 3 Summary of tenfold cross validation performance for AUC on all the datasets

Datasets	C4.5	REP	CART	NB Tree	CSIDL
Breast-cancer	0.610±0.100○	0.597±0.118○	0.590±0.100○	0.682±0.108○	0.589±0.094
Crx	0.881±0.047●	0.878±0.041●	0.877±0.043●	0.914±0.037○	0.891±0.047
Hepatitis	0.668±0.184●	0.620±0.151●	0.564±0.126●	0.766±0.142○	0.746±0.164
Horse-colic	0.843±0.070●	0.846±0.065●	0.849±0.069●	0.863±0.069●	0.867±0.219
House votes	0.979±0.025	0.975±0.024	0.973±0.027	0.987±0.017●	0.965±0.033
Post Operative	0.489±0.032○	0.490±0.062○	0.499±0.007○	0.388±0.017●	0.488±0.034

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

Table 4 Summary of tenfold cross validation performance for Precision all the datasets

Datasets	C4.5	REP	CART	NB Tree	CSIDL
Breast-cancer	0.753±0.042○	0.722±0.038●	0.728±0.038●	0.763±0.056○	0.741±0.039
Crx	0.835±0.063●	0.800±0.058●	0.810±0.062●	0.858±0.066○	0.854±0.058
Hepatitis	0.510±0.371●	0.298±0.392●	0.232±0.334●	0.514±0.343●	0.718±0.202
Horse-colic	0.851±0.055○	0.857±0.057○	0.853±0.052○	0.848±0.059○	0.590±0.473
Housevotes	0.960±0.042●	0.932±0.049●	0.942±0.052●	0.934±0.058●	0.967±0.042
Post-operative	0.731±0.054○	0.732±0.055○	0.733±0.055○	0.726±0.063○	0.720±0.055

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

Table 5 Summary of tenfold cross validation performance for F-measure on all the datasets

Datasets	C4.5	REP	CART	NB Tree	CSIDL
Breast-cancer	0.838±0.040○	0.808±0.039●	0.813±0.038●	0.805±0.057●	0.828±0.035
Crx	0.832±0.044●	0.831±0.047●	0.841±0.044●	0.831±0.053●	0.849±0.044
Hepatitis	0.409±0.272●	0.208±0.255●	0.179±0.235●	0.438±0.264●	0.681±0.163
Horse-colic	0.888±0.044○	0.886±0.044○	0.890±0.040○	0.862±0.050○	0.603±0.473
Housevotes	0.955±0.033●	0.940±0.041●	0.946±0.033●	0.936±0.045●	0.973±0.025
Post-operative	0.838±0.043○	0.842±0.043○	0.844±0.037○	0.812±0.065●	0.828±0.051

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

Table 6 Summary of tenfold cross validation performance for TP Rate on all the datasets

Datasets	C4.5	REP	CART	NB Tree	CSIDL
Breast-cancer	0.954±0.041○	0.908±0.096●	0.919±0.077●	0.855±0.072●	0.939±0.049
Crx	0.833±0.062●	0.870±0.077○	0.881±0.073○	0.810±0.070●	0.849±0.064
Hepatitis	0.374±0.256●	0.183±0.235●	0.169±0.236●	0.442±0.305●	0.717±0.239
Horse-colic	0.930±0.054○	0.920±0.062○	0.933±0.049○	0.881±0.069○	0.630±0.485
House votes	0.953±0.045●	0.951±0.056●	0.953±0.046●	0.942±0.056●	0.980±0.033
Post Operative	0.987±0.054○	0.996±0.043○	0.999±0.014○	0.930±0.111●	0.977±0.061

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

Table 7 Summary of tenfold cross validation performance for FP Rate on all the datasets

Datasets	**C4.5**	**REP**	**CART**	**NB Tree**	**CSIDL**
Breast-cancer	0.750±0.153○	0.823±0.168●	0.815±0.181●	0.618±0.170○	0.759±0.130
Crx	0.136±0.060●	0.179±0.065●	0.172±0.072●	0.111±0.058○	0.129±0.061
Hepatitis	0.100±0.097○	0.055±0.093○	0.072±0.094○	0.120±0.109○	0.161±0.127
Horse-colic	0.283±0.119●	0.269±0.121●	0.280±0.112●	0.275±0.126●	0.008±0.024
House votes	0.026±0.029○	0.045±0.035○	0.039±0.037○	0.044±0.041○	0.049±0.062
Post Operative	1.000±0.000	1.000±0.000	1.000±0.000	0.963±0.129●	1.000±0.000

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

Table 8 Summary of tenfold cross validation performance for TN Rate on all the datasets

Datasets	**C4.5**	**REP**	**CART**	**NB Tree**	**CSIDL**
Breast-cancer	0.250±0.153○	0.177±0.168●	0.185±0.181●	0.382±0.170○	0.241±0.130
Crx	0.864±0.060●	0.821±0.065●	0.828±0.072●	0.889±0.058○	0.871±0.061
Hepatitis	0.900±0.097○	0.945±0.093○	0.928±0.094○	0.880±0.109○	0.839±0.127
Horse-colic	0.717±0.119●	0.731±0.121●	0.720±0.112●	0.725±0.126●	0.992±0.024
House votes	0.974±0.029○	0.955±0.035○	0.961±0.037○	0.956±0.041○	0.951±0.062
Post Operative	0.000±0.000	0.000±0.000	0.000±0.000	0.037±0.129●	0.000±0.000

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

Table 9 Summary of tenfold cross validation performance for FN Rate on all the datasets

Datasets	**C4.5**	**REP**	**CART**	**NB Tree**	**CSIDL**
Breast-cancer	0.046±0.041○	0.092±0.096●	0.081±0.077●	0.145±0.072●	0.061±0.049
Crx	0.167±0.062●	0.130±0.077○	0.119±0.073○	0.190±0.070●	0.151±0.064
Hepatitis	0.626±0.256●	0.817±0.235●	0.831±0.236●	0.558±0.305●	0.283±0.239
Horse-colic	0.070±0.054○	0.080±0.062○	0.067±0.049○	0.119±0.069○	0.370±0.485
House votes	0.047±0.045●	0.049±0.056●	0.047±0.046●	0.058±0.056●	0.020±0.033
Post Operative	0.013±0.054○	0.004±0.043○	0.001±0.014○	0.070±0.111●	0.023±0.061

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

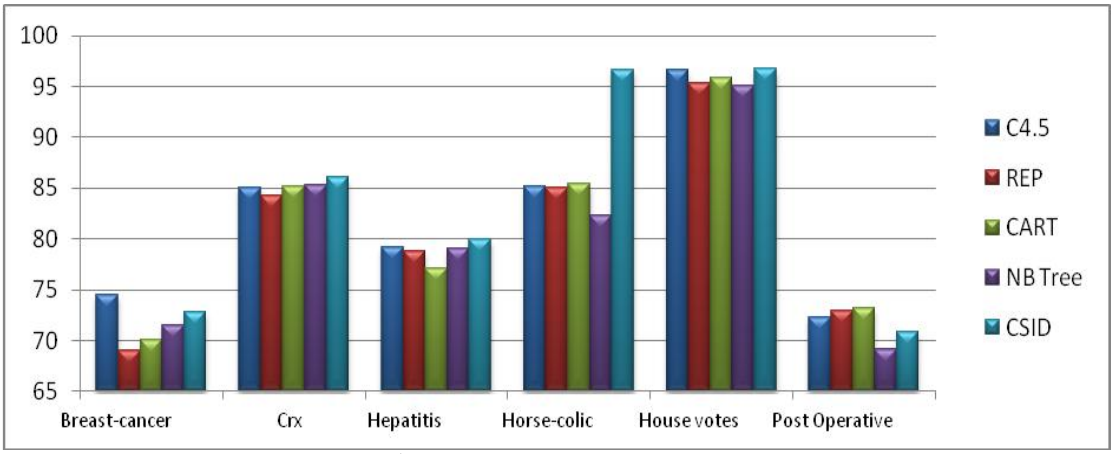

Figure 2: Trends in accuracy for C4.5, REP, CART and NB Tree versus CSIDL on UCI datasets

Figure 2 presets the bar chart representation of accuracy results compared with benchmark algorithms with CSIDL. Table 3 shows the detailed experimental results of the AUC of C4.5, REP, CART, NB Tree on all the data sets. From Table 3 we can see CSIDL model have performed well in terms of AUC and have achieve substantial improvement over C4.5 (3 wins 1 tie and 2 loss) and moderate improvement over REP (3 wins 1 tie and 2 loss), CART (3 wins 1 tie and 2 loss) and NB Tree (3 wins and 3 loss). Table 3 – 9 presents the results of Precision, F-measure, TP Rate, FP rate, TN Rate and FN Rate respectively. Table 10 presents the summary of all the performance metrics on comparative algorithms. The unique properties of the datasets such as the ratio of the missing values, imbalance ratio etc are some of the reason for unexpected results.

Table 10 Summary of tenfold cross validation performance

Metric/System	C4.5	REP	CART	NB Tree
Accuracy	3/1/2	4/1/1	4/1/1	6/0/0
AUC	3/1/2	3/1/2	3/1/2	3/0/3
Precision	3/0/3	4/0/2	4/0/2	2/0/4
F-measure	3/0/3	4/0/2	4/0/2	5/0/1
TP Rate	3/0/3	3/0/3	3/0/3	5/0/1
FP Rate	2/1/3	3/1/2	3/1/2	2/0/4
TN Rate	2/1/3	3/1/2	3/1/2	2/0/4
FN Rate	3/0/3	3/0/3	3/0/3	5/0/1

(W/T/L) – Win /Tie /Loss

5. Conclusion

In this paper, we propose a novel algorithm dubbed as Confiscate and Substitute Imbalance Data Learning (CSIDL) for better knowledge discovery from real world datasets. The experimental results indicate that the proposed approach is a competitive one for wide range of datasets. In future work, we want to implement our algorithmic approach on multi class imbalance learning complex data sources.

REFERENCES

[1] Sara del R, Victoria L´opez , Jos´e Manuel Ben´ıtez , Francisco HeCSera, “A MapReduce Approach to Address Big Data Classification Problems Based on the Fusion of Linguistic Fuzzy Rules”,International Journal of Computational Intelligence Systems, Vol. 8, No. 3 (2015) 422-437.

[2] Dharm Singh, Naveen Choudhary&JullySamota,” Analysis of Data Mining Classification with Decision treeTechnique:”, Global Journal of Computer Science and TechnologySoftware& Data Engineering, Volume 13 Issue 13 Version 1.0 Year 2013.

[3] Yong Hu, DongfaGuo, Zengwei Fan, Chen Dong, Qiuhong Huang, ShengkaiXie,Guifang Liu, Jing Tan, Boping Li, QiweiXie.” An Improved Algorithm for ImbalancedData and Small Sample Size Classification”, Journal of Data Analysis and Information Processing, 2015, 3, 27-33.

[4] VaishaliGanganwar, “An overview of classification algorithms for imbalancedDatasets”, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 4, April 2012).

[5] Thin Nguyen, DinhPhung, Brett Adams, Truyen Tran, and SvethaVenkatesh,” Classification and Pattern Discovery of Mood inWeblogs, M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 283–290, 2010.

[6] Kevin Lee,ZhihongMan,Dianhui Wang, Zhenwei Cao,” Classification of bioinformatics dataset using finite impulse response extreme learning machine for cancer diagnosis”, Neural Comput&Applic, DOI 10.1007/s00521-012-0847-z.

[7] Bharath K. Samanthula, Member, IEEE, YousefElmehdwi, and Wei Jiang, Member, IEEE,”k-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data,IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015.

[8] J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA:Morgan Kaufmann, 1993.

[9] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

[10] HamiltonA. Asuncion D. Newman. (2007). *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

[11] Keel

[12] J. Quinlan. Induction of decision trees, Machine Learning, vol. 1, pp. 81C106, 1986.

[13] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Belmont, CA: Wadsworth, 1984.

[14] Ron Kohavi: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: Second International Conference on Knoledge Discovery and Data Mining, 202-207, 1996.

[15] Maimon, O., Rokach, L.: Data Mining And Knowledge Discovery Handbook. Springer, Berlin (2010)