International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.241-250
http://dx.doi.org/10.14257/ijdta.2016.9.5.25

An Improved ID3 Decision Tree Algorithm on Imbalance Datasets Using Strategic Oversampling

L. Surya Prasanthi [1], R. Kiran Kumar [2] and Kudipudi Srinivas [3]

1 Research Scholar, Department of Computer Science, Krishna University, Machilipatnam, India
2 Department of Computer Science, Krishna University, Machilipatnam, India
3 Department of Computer Science & Engineering, V.R. Siddartha Engineering College, Vijayawada, India
E-mail:prasanthi.latike@gmail.com, kirankreddi @gmail.com, vrdrks@gmail.com

Abstract

Data mining is the process of extracting useful information from the vast and complex databases. In real time scenario the data sources contain many varied data including imbalance data category. Imbalance data sets contain more percentage of instances from one class and are very less percentage of instances from other class. The traditional decision tree algorithm called Iterative Dichotomiser 3 (ID3) is built for not handling the imbalance datasets. To overcome the drawback of ID3 on imbalance datasets, an improved algorithms are needed. In this paper, propose extension of ID3 algorithm called Over Sampled ID3 (OSID3) for imbalance data learning. The proposed OSID3 approach uses the oversampling technique with unique statistical oversample strategy for removing less privileged instances in the early stage and later on oversampling the high privileged instances for approximate data balance. The experimental observation suggests that the proposed approach improves in terms of Accuracy, Area Under Curve (AUC) and Root Mean Square Error (RMSE) with the benchmark ID3 on 15 imbalance datasets from University of California, Irvine (UCI) repository.

Keywords*: Data Mining, Knowledge Discovery, Classification, Decision Tree, ID3, OSID3*

1. Introduction

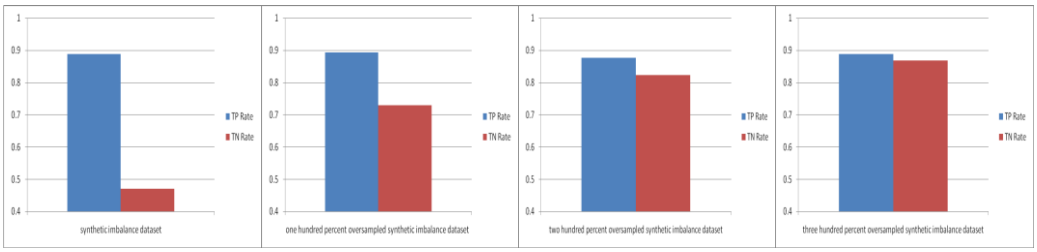
Data mining is one of the most predominate and well versed field for analyzing the varied and complex category of databases. In Data mining, Classification is a simplified and accurate approach for data exploration. In Classification, ID3 is a traditional decision tree approach which uses divide and win strategy for knowledge discovery from the databases. ID3 is a benchmark algorithm which was actually designed for acting on the normal or balanced datasets. One of the shortcomings of the ID3 algorithm is the bottleneck performance for efficient learning of imbalance datasets.

Learning imbalance datasets is a challenging task yet important due to availability in real-time scenario. In the context of imbalance data, most of instances in the dataset belong to class known as majority class and very few instances belong to the other class known as minority class which is usually the more important class. The ID3 decision tree is capable of accurately classifying the majority class which is usually the less important class and the accuracy of minority class drops drastically when compared to majority class. The simulation results in Figure1 presents the three hundred percent oversampled synthetic imbalance dataset demonstrating the above shortcoming. One can observe from the comparative results of ID3 for TP Rate and TN Rate on original data in Figure1(a), 100% oversampled minority subset results in Figure1(b), 200% oversampled minority

ISSN: 2005-4270 IJDTA

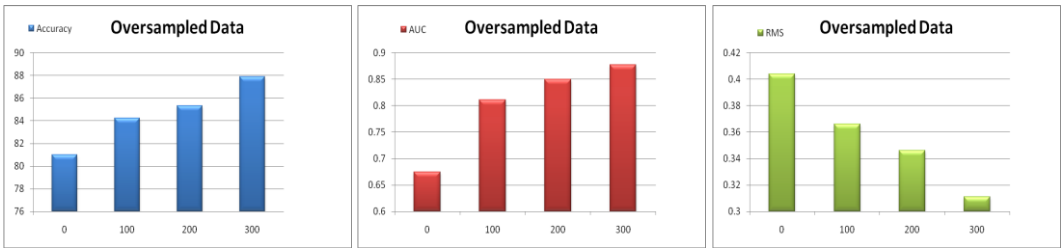
subset results in Figure1(c), 300% oversampled minority subset results in Figure1(d). From the Figure1(a) to Figure1(d) the results of TP Rate are similar but a good improvement can be seen in TN Rate value.

In Figure 2(a): Blue bars represent the accuracy: starting from left: the first bar represents the results of pure synthetic imbalance dataset with an accuracy value of 81.00; the second bar represents the results of 100% minority oversampled dataset with a accuracy value of 84.23; the third bar represents the results of 200% minority oversampled dataset with a accuracy value of 85.32; the fourth bar represents the results of 300% minority oversampled dataset with a accuracy value of 87.90 where the majority and minority instances got balanced; One can make observations from Figure 1(a) that the accuracy value had improved from 81.00 to 87.90.



1(a) 1(b) 1(c) 1(d)

Figure 1. Results Analysis for Synthetic Datasets with Imbalance Nature Problem on ID3 for TP Rate and TN Rate



2(a) 2(b) 2(c)

Figure 2. Results Analysis for Synthetic Datasets with Imbalance Nature Problem on ID3 for Accuracy, AUC and RMS

In Figure 2(b): Red bars represents the AUC: starting from left: the first bar represents the results of pure synthetic imbalance dataset with a AUC value of 0.675; the second bar represents the results of 100% minority oversampled dataset with a accuracy value of 0.811; the third bar represents the results of 200% minority oversampled dataset with a AUC value of 0.850; the fourth bar represents the results of 300% minority oversampled dataset with a AUC value of 0.877 where the majority and minority instances got balanced; One can make observations from Figure 1(b) that the AUC value had improved from 0.675 to 0.877.

In Figure 2(c): Green bars represents the Root Mean Square Error (RMS): starting from left: the first bar represents the results of pure synthetic imbalance dataset with a RMS value of 0.404; the second bar represents the results of 100% minority oversampled dataset with a RMS value of 0.366; the third bar represents the results of 200% minority oversampled dataset with a RMS value of 0.346; the fourth bar represents the results of 300% minority oversampled dataset with a RMS value of 0.311 where the majority and minority instances got balanced; One can make observations from Figure 1(c) that the error value had decreased from 0.404 to 0.311.

This shows that a novel approach can improve the accuracy and AUC of ID3 where as reducing the error rate. In summary, the problem of class imbalance learning is not

considered especially with ID3. In this paper, a novel approach is proposed using minority oversampling for solving above said problems.

The arrangement of paper is follows as. We exhibit in Section 2 the late approaches in learning with decision tree. It will straightforwardly persuade the principle commitment of this work introduced in Section 3, somewhere we propose another structure for OSID3. Assessment criteria's designed for decision tree learning is exhibited in area 4. Test results are accounted for in Section 5. In conclusion, we finish up with Section 6 where we talk about real open issues and upcoming work.

2. Current Approaches in Decision Trees

The decision tree approaches with imbalance data is presented by many of the researchers, one of the contribution is done by Ali Mirza Mahmood [1] as a comprehensive review of current methods for constructing models for learning from class imbalanced data. He also presented a critical review of the nature of the problem. Fahmi Arif *et al*. [2] have proposed the combination of multiple PCA+ID3 algorithm to develop quality prediction model in multi stage manufacturing. Kalpesh Adhatrao *et al*. [3] have developed the system using the predictive performance of ID3 algorithm for evaluating the students for future preparations. Akshaya. *et al*. [4] have proposed a privacy framework for the ID3 decision tree algorithm for achieving better level of accuracy along with improved privacy. Sandeep Kumar *et al*. [5] have proposed an improved ID3 algorithm of decision tree and they used Havrda and Charvat Entropy instead of Shannon Entropy. Ehsan Molaei *et al*. [6] have developed a safe distributed algorithm which is using improved secure sum algorithm and performed on classic ID3. Sagar Manohar *et al*. [7] have proposed the idea of such a classifier which can be built independently and without Bulky Business Intelligence software to effectively forecast future occurrences of any phenomena. Ramanathan L *et al*. [8] have proposed an improved ID3 approach to overcome the short coming of ID3 incling towards attributes with many values. In the proposed approach they used gain ratio and applied weights for each attributes for building decision tree model.

3. The Proposed Approach

In this section, the proposed approach for Improved ID3 is presented.

The different components of our new proposed framework are elaborated in the next subsections.

In the initial stage of our frame work the dataset is divided into minority subset P € pi (i = 1,2,..., pnum) and majority subset N € ni (i = 1,2,..., nnum) respectively. As the ID3 algorithms efficiency drops down on imbalance data to improve the efficiency the dataset’s majority subclass is to the under sampled or minority subclass is to be oversampled. In our proposed approach we initiated the oversampling strategy for the minority sub class. One of the limitations of the existing oversampling algorithms is of not considering for removal of noisy and outlier instances before oversampling. Therefore, in the proposed approach before oversampling phase is started mostly misclassified instances are removed from the dataset. The technique proposed for identifying the mostly misclassified instances is by considering the nearest neighbor instances. If all the nearest neighbor instances of a particular instance are of opposite class then it implies that particular instance comes under the category of a noisy or outlier instance and can be eliminated.

The eliminated instances can boost the performance of the proposed approach in two ways:

First it will reduce the noisy and outlier instances not only from majority but also minority subset and hence improves the quality of the dataset. Second it reduces some of

the outlier and noisy instances from majority subset and so reduces the imbalance nature of the dataset.

In the next phase minority subset is oversampled. The some of the synthetic instances generated are the replica of the existing instances, hybrid instances and pure artificial instances. In the final stage the fine tuned dataset is applied to ID3 algorithm and evaluation metric are generated.

The proposed OSID3 algorithm is summarized as below.

Algorithm: OSID3

Algorithm: New Decision Tree (D, A, GR)
Input: D – Data Partition
A – Attribute List
GR – Gain Ratio
Output : A Decision Tree

Procedure:

Processing Phase:
Step 1. *Take the class imbalance data and divide it into majority and minority sub sets. Let the minority subset be P € pi* (*i* = 1,2,..., *pnum*) *and majority subset be N € ni* (*i* = 1,2,..., *nnum*).

Let us consider
m' = the number of majority nearest neighbors
T = the whole training set
m= the number of nearest neighbors

Step 2. *Find mostly misclassified instances pi*

$pi = m'$; *where* m' $(0 \leq m' \leq m)$

if $m/2 \leq m' < m$ *then pi is a mostly misclassified instance. Then remove the instances m' from the minority set.*

Let us consider
m' = the number of minority nearest neighbors

Step 3. *Find mostly misclassified instances ni*

$ni = m'$; *where* m' $(0 \leq m' \leq m)$

if $m/2 \leq m' < m$ *then pi is a mostly misclassified instance. Then remove the instances m' from the majority set.*

Let us consider
m' = the number of majority nearest neighbors

Step 4. *Find noisy instances pi'*

$pi' = m'$; *where* m' $(0 \leq m' \leq m)$
If $m' = m$, *i.e. all the m nearest neighbors of pi are majority examples, pi' is considered to be noise or outliers or missing values and are to be removed.*

Let us consider
m' = the number of minority nearest neighbors

Step 5. *Find noisy instances ni'*

ni' = m'; *where m'* ($0 \leq m' \leq m$)

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

Step 6. *For every pi'* (i = 1,2,..., *pnum'*) *in the minority class P, we calculate its m nearest neighbors from the whole training set T. The number of majority examples among the m nearest neighbors is denoted by m'* ($0 \leq m' \leq m$).

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

Step 7. *In this step, we generate s × dnum synthetic minority examples from the minority sub set, where s is an integer between 1 and k . One percentage of synthetic examples generated is replica of minority examples and other are the hybrid of minority examples.*

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_w , A_w *)*
8. *label root node N as f(A)*
9. ***for*** *each outcome j of f(A)****do***
10. *subtree j =New Decision Tree(Dj,A)*
11. *connect the root node N to subtree j*
12. ***endfor***
13. ***endif***
14. ***endif***
15. *Return N*

---

4. Investigational Design and Assessment Criteria

We performed the implementation of our new algorithms within the Weka [11] environment on windows 7 with i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM. In order to test the strength of our method, we compared it with existing ID3 algorithm. We evaluated each of the classifiers on the fifteen datasets from UCI data repositories (Table 1).

With a specific end goal to analyze the classifiers, we utilize 10-fold cross acceptance. In 10-fold cross approval, every dataset is broken into 10 disjoint sets such that every set has (generally) the same dissemination. The classifier is found out 10 times such that in every emphasis an alternate set is withheld from the preparation stage, and utilized rather to test the classifier. We then process the accuracy and AUC as the normal of each of these runs.

Datasets used in Decision tree Learning

Table 1 summarizes the benchmark datasets [12] used in the anticipated study.

The details of the datasets are given in Table 1. For each data set, S.no., Dataset, name of the dataset, Instances, number of instances, Attributes, Number of Attributes, IR, Imbalance Ratio are described in the table for all the datasets.

Table 1. UCI Datasets and their Properties

S.no. Dataset	Inst	Attributes	IR
1. Breast-cancer	286	9	2.37
2. Breast-cancer-w	699	9	1.90
3. Horse-colic	368	22	1.71
4. German_credit	1,000	20	2.33
5. Pima diabetes	768	8	1.87
6. Heart-c	303	13	1.19
7. Heart-h	294	13	1.77
8. Heart-statlog	270	14	1.25
9. Hepatitis	155	20	3.85
10. Ionosphere	351	35	1.79
11. Kr-vs-kp	3196	37	1.09
12. Labor	57	17	1.85
13. Mushroom	8124	23	1.08
14. Sick	3772	30	15.32
15. Sonar	208	13	1.15

The most commonly used empirical measure, accuracy distinguish between the numbers of correct labels of different classes. The mathematical notation for calculation of accuracy is give below in eq (i),

$$ACC = \frac{TP + TN}{TP + FN + FP + FN} \text{ --------- (i)}$$

A quantitative representation of a ROC curve is the area under it, which is known as AUC. When only one run is available from a classifier, the AUC can be computed as the arithmetic mean (macro-average) of TP rate and TN rate.

Another important measure used in decision tree is the tree size. The size of the tree is calculated by the depth of the tree and using the number of nodes and leaves.

5. Results

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

Table 2 shows the detailed experimental results of the mean classification accuracy of ID3 [13] and OSID3 on all the 15 data sets. From Table 2 we can see accuracy performance of OSID3 model that it can achieve substantial improvement over ID3 on most data set (14 wins and 1 loss) which suggests that the OSID3 model is potentially a good technique for decision tree learning on imbalance datasets.

Table 2. Accuracy on All the Datasets with Summary of Tenfold Cross Validation Performance

Datasets	ID3	OSID3
Breast-cancer	58.95±9.22	64.70±7.41●
Breast-cancer-w	90.62±3.20	95.41±2.22●
Horse-colic	52.58±8.09	65.57±6.97●
German_credit	8.94±3.03	29.75±3.25●
Pima_diabetes	26.15±4.31	49.35±4.83●
Heart-c	33.62±7.77	44.39±7.79●
Heart-h	27.58±7.75	47.72±7.79●
Heart-statlog	34.67±9.11	61.20±7.61●
Hepatitis	27.75±10.18	47.09±10.69●

Ionosphere	17.32±4.79	44.25±3.89●
Kr-vs-kp	99.60±0.38	99.70±0.33●
Labor	59.33±20.60	71.54±14.55●
Mushroom	100.0±0.0	100.0±0.0
Sick	80.78±1.88	83.75±1.98●
Sonar	0.96±1.93	10.67±5.91●
● Bold dot indicates the win of Proposed approach		

Table 3. AUC on All the Datasets with Summary of Tenfold Cross Validation Performance

Datasets	ID3	OSID3
Breast-cancer	0.593±0.097	0.670±0.075●
Breast-cancer-w	0.953±0.024	0.969±0.019●
Horse-colic	0.716±0.060	0.745±0.050●
German_credit	0.513±0.035	0.536±0.021●
Pima_diabetes	0.539±0.052	0.615±0.037●
Heart-c	0.573±0.088	0.617±0.067●
Heart-h	0.545±0.075	0.607±0.057●
Heart-statlog	0.591±0.084	0.664±0.058●
Hepatitis	0.474±0.043	0.832±0.092●
Ionosphere	0.738±0.064	0.930±0.037●
Kr-vs-kp	0.996±0.004	0.997±0.003●
Labor	0.713±0.193	0.850±0.121●
Mushroom	100.0±0.0	100.0±0.0
Sick	0.871±0.033	0.913±0.015●
Sonar	0.498±0.013	0.597±0.057●
● Bold dot indicates the win of Proposed approach;		

Table 4. RMS Error on All the Datasets with Summary of Tenfold Cross Validation Performance

Datasets	ID3	OSID3
Breast-cancer	0.567±0.072	0.530±0.068●
Breast-cancer-w	0.185±0.070	0.133±0.069●
Horse-colic	0.391±0.105	0.323±0.092●
German_credit	0.595±0.114	0.331±0.067●
Pima_diabetes	0.624±0.059	0.438±0.052●
Heart-c	0.398±0.058	0.344±0.050●
Heart-h	0.379±0.072	0.274±0.059●
Heart-statlog	0.598±0.101	0.398±0.087●
Hepatitis	0.510±0.221	0.257±0.201●
Ionosphere	0.050±0.131	0.018±0.064●
Kr-vs-kp	0.050±0.039	0.042±0.036●
Labor	0.425±0.274	0.336±0.221●
Mushroom	0.0±0.0	0.0±0.0
Sick	0.118±0.025	0.087±0.032●
Sonar	0.130±0.344	0.029±0.137●
● Bold dot indicates the win of Proposed approach;		

Table 3 and 4 shows the detailed experimental results of the AUC and RMS Error of ID3 verses OSID3 on all the data sets. From Table 3 we can see OSID3 model have performed well in terms of AUC (14 wins and 1 loss) and have achieved substantial improvement over ID3. One can observe from Table 4 that RMS Error generated by the proposed OSID3 algorithm is reduced for all the compared datasets.

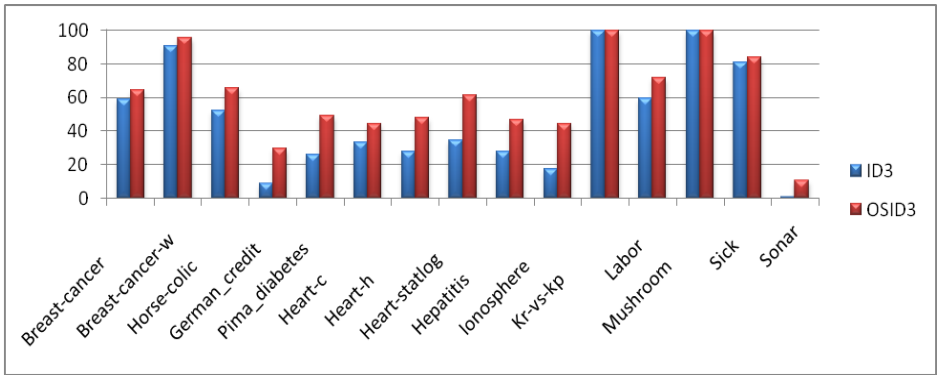


Figure 3. Trends in Accuracy for ID3 versus Proposed Approach on UCI Data Sets

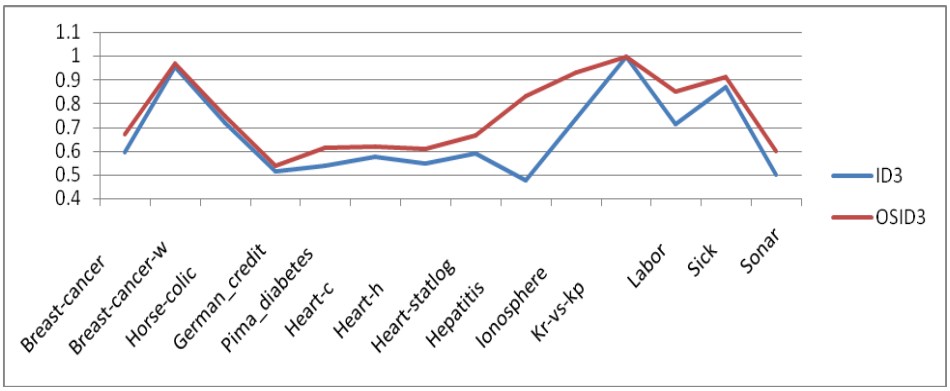


Figure 4. Trends in AUC for ID3 versus Proposed Approach on UCI Data Sets

5. Conclusion

In this paper, we introduced a decision tree approach called Over Sampled ID3 for effective performance on class imbalance datasets. A representative framework for minority subset oversampling is applied for improvising class imbalance learning. Experimental results suggest that the proposed approach performs better than the existing ID3 algorithm on all the evaluation metrics.

In future work, we will like to extend our system for high dimensional and complex datasets.

References

[1] A. M. Mahmood, "Class Imbalance Learning in Data Mining – A Survey", International Journal of Communication Technology for Social Netw Hussin", A Data Mining Approach for Developing Quality Prediction Model in Multi-Stage Manufacturing", International Journal of Computer Applications, vol.69, no. 22, 2013, pp. 0975–8887.

[2] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha and V. Honrao, "Predicting Students' Performance Using Id3 and C4.5 Classification Algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 3, no. 5, **(2013)**.

[3] S. Akshaya, V. L. J. Manchari, A. M. Thoufeeq and K. Kiruthikadevi, "Implementation of Double Layer Privacy on Id3 Decision Tree Algorithm", International Journal of Scientific Engineering and Technology Research, vol. 3, no. 7, **(2014)**, pp. 1194-1200.

[4] S. Kumar and S. Jain, "Intrusion Detection and Classification Using Improved ID3 Algorithm of Data Mining", International Journal of Advanced Research in Computer Engineering & Technology, vol. 1, no. 5, **(2012)**, pp. 352-356.

[5] E. Molaei, H. Vadiatizadeh, A. Mohammadighavam, N. Rajabpour and F. Ziasistani, "Distributed algorithm for privacy preserving data mining based on ID3 and improved secure sum", International Journal of Advanced studies in Computer Science and Engineering IJASCSE, vol. 3, no. 1, **(2014)**, pp. 28-34.
[6] 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.
[7] L. Ramanathan, S. Dhanda and S. D. Kumar, "Predicting Students' Performance using Modified ID3 Algorithm", International Journal of Engineering and Technology (IJET), vol. 5, no. 3, **(2013)**, pp. 2491-2497.
[8] B. R. Patel and K. K. Rana, "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research (IJEDR), vol. 2, no. 1, pp. 1-5.
[9] A. Rajalakshmi and K. Sivaranjani, "A Comparative Study on Student Performance Prediction", International Conference on Engineering Trends and Science & Humanities (ICETSH), www.internationaljournalssrg.org, **(2015)**, pp.53-56.
[10] I. H. Witten and E. Frank, "Data Mining: Practical machine learning tools and techniques", 2nd edition Morgan Kaufmann, San Francisco, **(2005)**.
[11] A. Hamilton and A. 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
[12] J. R. Quinlan, "Induction of Decision Trees", Mach. Learn., vol. 1, no. 1, **(1986)**, pp. 81-106.

