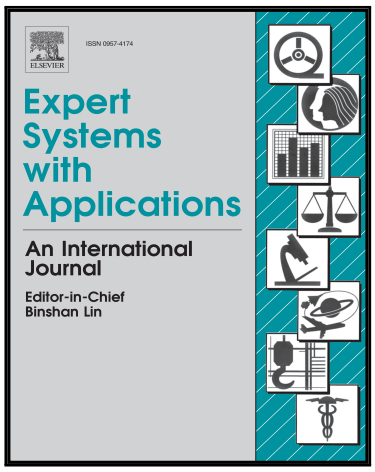


# Dispersion Ratio Based Decision Tree Model for Classification

Smita Roy, Samrat Mondal, Asif Ekbal, Maunendra Sankar Desarkar

PII: S0957-4174(18)30543-8
DOI: https://doi.org/10.1016/j.eswa.2018.08.039
Reference: ESWA 12170

To appear in: *Expert Systems With Applications*

Received date: 6 April 2018
Revised date: 10 August 2018
Accepted date: 20 August 2018

Please cite this article as: Smita Roy, Samrat Mondal, Asif Ekbal, Maunendra Sankar Desarkar, Dispersion Ratio Based Decision Tree Model for Classification, *Expert Systems With Applications* (2018), doi: https://doi.org/10.1016/j.eswa.2018.08.039

**Highlights**

- Here Dispersion Ratio is proposed to identify splitting attribute in Decision Trees.
- The proposed method works for any type of variable: categorical, nominal, continuous.
- Efficient discretization method is proposed for continuous features.
- Has no bias towards features with more distinct values like many other methods.
- Extensive evaluation and analysis is performed on a large number of datasets.

# Dispersion Ratio Based Decision Tree Model for Classification

Smita Roy[a,*], Samrat Mondal[b,d], Asif Ekbal[b], Maunendra Sankar Desarkar[c]

[a]*Department of Computer Science, Central University of South Bihar, Patna, India*

[b]*Department of Computer Science and Engineering, Indian Institute of Technology Patna, India*

[c]*Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India*

[d]*Daniel Felix Ritchie School of Engineering & Computer Science, University of Denver, CO, USA*

**Abstract**

In predictive tasks like classification, Information Gain (IG) based Decision Tree is very popularly used. However, IG method has some inherent problems like its preference towards choosing attributes with higher number of distinct values as the splitting attribute in case of nominal attributes and another problem is associated with imbalanced datasets. Most of the real-world datasets have many nominal attributes, and those nominal attributes may have many number of distinct values. In this paper, we have tried to point out these characteristics of the datasets while discussing the performance of our proposed approach. Our approach is a variant of the traditional Decision Tree model and uses a new technique called Dispersion_Ratio, a modification of existing Correlation Ratio (CR) method. The whole approach is divided into two phases - firstly, the dataset is discretised using a discretization module and secondly, the preprocessed dataset is used to build a Dispersion Ratio

*Smita Roy, Central University of South Bihar,BIT Campus, P.O.- B. V. College,Patna - 800 014, Bihar, India, Phone-(+91)8294976444

*Email addresses:* smitaroy@cub.ac.in (Smita Roy), samrat@iitp.ac.in (Samrat Mondal), asif@iitp.ac.in (Asif Ekbal), maunendra@iith.ac.in (Maunendra Sankar Desarkar)

based Decision Tree model. The proposed method does not prefer the attributes with many unique values and indifferent about class distribution. It performs better than previously proposed CR based Decision Tree (CRDT) Model since an efficient discretization module has been added with it. We have evaluated the performance of our approach on some benchmark datasets from various domains to demonstrate the effectiveness of the proposed technique and also compared our model with Information Gain, Gain Ratio and Gini Index based models. Result shows that the proposed model outperforms other models in majority of the cases that we have considered in our experiment.

*Keywords:* Data Mining, Decision Tree, Information Gain, Correlation Ratio, Dispersion_Ratio, Classification.

---

## 1. Introduction

Increasing use of affordable Internet and mobile technology have made the growth of data exceeding the capacity of traditional computing. Recently, it is reported (Brueckner, 2018) that Human- and machine-generated data is experiencing an overall 10x faster growth rate than traditional business data, and machine data is increasing even more rapidly at 50x the growth rate. Thus, we have abundant data and those data need to be analyzed for extracting meaningful information. This has led the research community to explore the capability of different existing data mining techniques.

One of the important tasks of data mining is classification. For classification purpose, the commonly used techniques are - Decision Tree algorithm (Quinlan, 1986)(Han et al., 2006), Naive Bayes Classifier (John and Langley, 1995)(Domingos and Pazzani, 1997)(Han et al., 2006), Neural Network model (Murata et al., 1994)(Han et al., 2006), k-Nearest Neighbour algorithm(Han et al., 2006), Support Vector Machine (Vapnik, 1995)(Han et al., 2006), etc. Feature selection is an important subtask that needs to be performed before applying this classification task. Normally real-world datasets contain many features and there may be some irrelevant or redundant features which may be responsible for degrading the performance of the learning models. To illustrate this point, let's consider the following example .

**Example 1.1.** *Suppose in a patient dataset, we have kept a patient-ID attribute by mistake and using the dataset to learn a classifier for disease diagnosis. The classifier will become exactly consistent with the training dataset due to some spurious relationship but its predictive power for unseen examples will get reduced due to over-fitting.*

Thus, selection of important features is a major concern. Feature selection techniques are divided into three categories: *filter*, *wrapper* and *embedded* methods (Wang et al., 2014). In *filter* method, the best subset of features is selected based on some inherent property of the features and after that the learning algorithm is applied. *Wrapper* method evaluates the selected feature subset on the basis of the performance of the classifier. In *embedded* approach, feature selection is an integral part of the learning model. Decision Tree classification algorithm uses embedded approach for feature selection.

Decision Tree model normally employs Information Gain (IG) as the criterion to compute the significance for splitting on an attribute. At each level of the Decision Tree, an attribute having the highest IG is chosen for splitting the associated dataset. But IG has a bias towards the attributes that have many distinct values. Let's clarify this with following example.

**Example 1.2.**

Table 1: Student Dataset

| *Roll-No* | *Behaviour* | *Regular in class* | *Academic Performance* | *Parents Edu* | *Class* |
|---|---|---|---|---|---|
| *1* | Good | Yes | Good | Literate | Pass |
| *2* | Good | No | Good | Illiterate | Fail |
| *3* | Bad | Yes | Good | Illiterate | Fail |

So if a dataset contains a key attribute, like Roll-No as shown in the above Table 1, then it will be chosen as the splitting attribute as it can discriminate each record belonging to the dataset and this would produce a considerable number of partitions (equal to the number of different values), where each one will have only one record. Since for each partition, records in each of them have same class labels, so the information required to classify any tuple in dataset $D$ depending on Information Gain (Han et al., 2006) principle

would amount to 0. Hence, the reduction in entropy by segmenting on this attribute will be maximum. Therefore, such a partitioning is actually of no use for classification (Han et al., 2006).

So, in this case, the generalization of the model may not be possible. Thus, the existing IG based approach does not suit every types of datasets. For example, if a dataset contains some attributes with different numbers of distinct values, it favours most distinct valued attribute to be chosen as splitting attribute though, in fact, some other attributes with relatively few distinct values may be more relevant for classification. Owing to the above mentioned reason, IG based approach gives less accuracy for some datasets. Also the other attribute selection methods like *Gain Ratio*, *Average Gain*, *Gini Index* have their respective disadvantages. For example, *Gini Index* has similar problem like *Information Gain*. The *Average Gain* measure performs better than *Gain Ratio* in terms of tree size and run time but no significant improvement in terms of accuracy can be observed. More such details have been discussed in (Han et al., 2006) and (Dianhong and Liangxiao, 2007).

The aim of our work is to propose a better alternative for Information Gain method which will be used for significant feature selection in Decision Tree model and which will also overcome the difficulties associated with the other existing approaches to some extent. So, we propose a technique called Dispersion_Ratio (DR) which uses the concept of Correlation Ratio (Battiti et al., 2009) or CR as the splitting measure with an added discretization module. In this approach, the attribute selection is based on the attribute which is important enough to recognize at least one outcome class. This paper is an extended version of the paper in (Roy et al., 2016). We have extended this version in various aspects such as - we modified and improved the algorithm, shown the suitability of proposed method with many more examples, experimented on more number of datasets, analysed the performance of the proposed technique on different datasets and finally compared the performances of the proposed technique with more number of existing approaches. In this paper we have added a proper discretization method, unlike equal binning approach in the earlier paper (Roy et al., 2016). Here, discretization has been done in a more systematic way with the help of K-means clustering. We also perform cluster analysis through assessing the cluster performance with varying K (number of clusters). This is indeed useful, as it can be seen from the experimental results that this method does better than the earlier method with equal binning. In order to show the efficacy of the proposed

approach, we evaluate it with many other datasets, and provide necessary comparisons with proper analysis. Essentially, changes made in this paper are in the discretization method (which leads to better performance) and experiments.

In a nutshell, the contributions of the presented work can be summarized as follows:

- Proposing a Dispersion_Ratio based feature selection technique which is suitable for nominal as well as categorical attributes.
- Discretization of numerical attributes using K-means clustering in the preprocessing phase.
- Building a model for classification using Decision Tree algorithm which incorporates our proposed Dispersion_Ratio based feature selection method.
- Comparison of our proposed method with Information Gain, Gain Ratio and Gini Impurity based Decision Tree on multiple datasets.

The organization of the paper is as follows: Section 2 mentions about various research work done in this area. Section 3 gives the problem definition and elaborately describes our proposed approach. The result and analysis of the proposed approach are shown in Section 4. Section 5 contains conclusion and gives future direction of the work.

## 2. Related Work

Many splitting criteria have come up over the time for Decision Tree based model. Some of the well known approaches are - Information Gain (Han et al., 2006), Gain Ratio (Han et al., 2006), Average Gain (Dianhong and Liangxiao, 2007), Gini Index (Han et al., 2006), etc. The basis of most of these approaches is entropy. Besides the well-known drawback of Information Gain approach i.e. its inclination towards many different-valued attributes, there is one more drawback of IG method as pointed out by Liu et. al. in (Liu et al., 2010). The paper discusses the problem faced in using IG on imbalanced data and also proves that Information Gain of an attribute for imbalanced dataset will be biased towards majority class. To overcome this problem the authors propose a measure called Class Confidence proportion (CCP). The idea of CCP is to focus on a class and consider the attribute

as important for that class which has the most significant association with that class. The CCP of a rule $X \implies y$ is computed by the expression: $\frac{CC(X \implies y)}{CC(X \implies y)+CC(X \implies \neg y)}$. Here $CC(X \implies y)$ means the confidence of the rule $X \implies y$ where $X$ is an attribute and $y$ is the class. However, the major disadvantage of this method is that CCP needs to be computed for each rule associated with each value of an attribute and each class label.

The Gain Ratio method is basically same as normalized Information Gain where the ratio between Information obtained after splitting the dataset $D$ on attribute $A$ and the Split Information (related to the information resulted after splitting dataset $D$ on attribute $A$ into $v$ subsets where $v$ is the number of distinct values of attribute $A$) is obtained. Next, another measure, namely, Average Gain measure has come up in (Dianhong and Liangxiao, 2007) to subjugate the problem of computation of Gain Ratio that when the denominator i.e. the Split Information of Gain Ratio becomes zero, then Gain Ratio becomes undetermined (Han et al., 2006). Also GR has a tendency to prefer the attribute with low split info (Han et al., 2006). The Average Gain measure is expressed to be the ratio of Information Gain and the number of different values the attribute can take. But the problem of Average Gain is lying with the numeric attributes. The Average Gain measure cannot be applied to numeric attributes (Dianhong and Liangxiao, 2007).

Another important splitting criteria is Gini Index (Han et al., 2006). In CART (Classification and Regression Tree)(Breiman et al., 1984), Gini Index is used as the splitting criteria. In this measure the binary split of each attribute is considered. While considering a binary split, the weighted sum of impurity of each resulting partition is calculated. All possible binary splits are considered for each attribute. For any attribute, the subset resulting in minimum Gini Index for that attribute is considered as the splitting subset. The attribute which minimizes the Gini Index is the splitting attribute. Computation of GI for attribute having many different values takes considerable amount of time as it considers different binary partitions of that attribute and then chooses the best among them. The Gini Index also has a preference towards multivalued attributes and encounters problem when number of classes is more (Han et al., 2006).

Another alternative splitting criterion used in CART is Twoing criterion. For two classes, both Gini and Twoing measures are equivalent. Generally, at any node, Gini measure splits the dataset into two parts- a small and pure

partition and a large and impure partition. On the contrary, Twoing method splits a dataset into two balanced and impure partitions. So, Gini measure is superior than the Twoing criterion in terms of yielding pure children nodes (Breiman et al., 1984). The twoing method works slower than Gini Index method when the number of classes is more. For example, if there are K (a considerably larger number) classes, then we will be having 2K-1 different sets of groupings of the classes into two sets (Kantardzic, 2002).

In (Chandra et al., 2010), authors have come up with a new splitting technique called Distinct Class based Splitting Measure (DCSM) for Decision Tree construction where the number of different classes in a partition has been taken into account. The measure is computed by multiplying two terms - the first term considers the number of unique classes in each partition and if the partition is pure then this term decreases. The second term takes into account the number of examples from different classes and it decreases with the increase in the number of examples belonging to a particular class. The idea is to select the attribute for splitting having minimum DCSM measure. The major disadvantage of this measure is that it prefers pure partitions and increases exponentially with the increase in number of distinct classes in a partition.

In case of most of the data mining tasks, Feature Subset Selection plays an important role in the preprocessing step. But in case of embedded approach the feature selection is already there in the particular data mining technique used. Sun et al. have evolved an idea of a new Feature Subset selection method in (Sun et al., 2013) which is referred as dynamic-weighting based feature selection algorithm. In this approach, the significance of an attribute is obtained by computing $J(f)$ value based on Correlation-Ratio. It is a kind of association of the feature with the class attribute and indicates the weight of the corresponding feature. Subsequently for the remaining features belonging to the feature set the weights are updated dynamically one-by-one based on Correlation-Ratio and existing weights of the respective features. In comparison to other methods this method improves the performance significantly. But the time complexity of this approach is not linear.

In (Kozak and Boryczka, 2016), the significant role that the pheromone plays in Ant Colony Decision Tree (ACDT) is examined. It is a dynamic approach for discretization applied during construction of the Decision Tree model. Pheromone maps are constructed and the direction of movement of

the pheromone trail during the traversal of ACDT was investigated.

In (Gama and Rocha, 2003) the authors propose a classification approach using decision tree on stream data. The proposed model VFDTc is an extension of the VFDT (Very Fast Decision Tree Learner) system (Domingos and Hulten, 2000) in two ways mainly it is able to handle continuous data and it uses an efficient classification technique at tree leaves. The proposed system, VFDTc, can classify new information online. The most interesting property of the system is that it is able to achieve a similar performance as a standard decision tree model even for medium size datasets. This uses information gain as the splitting criteria in the decision tree. VFDTc chooses a possible split point if and only if the number of instances in each of the subsets is greater than a certain percentage of the overall number of instances seen in the node. The time needed to calculate the best split point is $O(nlogn)$ which is not linear.

The above study not only reflects the usefulness and wide variety of applications of different variants of Decision Tree models (obtained by using different splitting strategies) in different domains but also points out some of the disadvantages of the existing approaches. So, we have worked on building a classification model which addresses some of the drawbacks of exiting techniques.

## 3. Problem Formulation and Proposed Approach

### *3.1. Problem Formulation*

In this paper, we attempt to develop a Decision Tree based model for classification. Given a set of training examples, our goal is to develop a Decision Tree that can be used to further classify unseen test example into one of the predefined classes.

However, unlike other decision tree model, our Decision Tree model is targeted to use an attribute selection criterion that is free from any kind of bias and independent of the class distribution. Also, the model to be constructed here should be integrated with a proper discretization module.

### *3.2. Proposed Approach*

Here, we provide the details of our proposed Dispersion_Ratio (DR) based Decision Tree construction approach which also uses an efficient discretization technique. Before that we present a brief description of the existing Correlation Ratio (CR) method since our proposed DR method is based on the concept of CR.

#### *3.2.1. Idea of existing Correlation Ratio*

The measure of association or correlation between the class attribute and other attributes plays an important role in the part of prediction of the outcome by the classification model. The Correlation Coefficient method suits those applications with quantitative outcome. Sometimes, categorical outcomes like "yes/no" may be desired from the learning algorithm. Thus, whenever categorical outcome is desired, Correlation Coefficient cannot be applied there. Also, this method is able to capture the relationship between the attributes only if it is linear. But attributes may have non-linear relationships between them. To address this problem, Correlation Ratio (CR)(Weisstein and Eric, 1951) (Crathorne, 1922) (Roy et al., 2016) has come into existence. Correlation Ratio is applied to find the association between the two attributes where one of the attribute is numeric (or quantitative) and the other is nominal. As an example, "Age" can take any numeric value within a range say "1-100" and it is an example of quantitative attribute. "Eye Colour" is an example of categorical attribute ( values of such attribute act as labels) whose values can be generally any one of the followings- "black", "brown" and "green". So, the association in this case is non-linear and CR is able to retrieve the non-linear dependencies.

#### *3.2.2. Proposed Dispersion_Ratio (DR) method*

To overcome the aforementioned limitation, we propose an approach named Dispersion_Ratio (DR) which is based on the concept of CR and it can be applied to find association between any pair of nominal or categorical attributes. So at first, we define Dispersion_Ratio (DR).

**Definition 3.1.** ***Dispersion_Ratio (DR)****: DR of an attribute is defined to be the square-root of the ratio of the two components : the numerator is the dispersion (summation of the squared deviation) in the relative importances of*

*that attribute among individual classes, and the denominator is the dispersion in the importance of that attribute across the whole population.*

The following Equation 1 gives the expression for computing DR for attribute 'i':

$$DR_i = \sqrt{\frac{\sum_{y \in Y} n_y (\overline{m}_y^{(i)} - \overline{m}^{(i)})^2}{\sum_{y \in Y} \sum_{j=1}^{y} (v_{jy}^{(i)} - \overline{m}^{(i)})^2}} \quad (1)$$

where $y$ is class label, $Y$ is the set of classes and $y \in Y$, $n_y$ is the number of tuples with class label $y$, $m$ stands for relative importance, $\overline{m}_y^{(i)}$ is the relative importance of the $i^{th}$ attribute with respect to class $y$, $\overline{m}^{(i)}$ is the overall importance of the $i^{th}$ attribute irrespective of classes, $v_{jy}^{(i)}$ is the relative importance of $j^{th}$ value of the $i^{th}$ attribute in class $y$.

The computation of proposed Dispersion_Ratio (DR) of an attribute is illustrated using the following example.

**Example 3.1.** *In Table 2, we consider Income as the first attribute of the feature-set in a dataset and it has three possible values - Low, Medium and High. For the class label attribute, we have considered two possible classes with labels as - Buy and Don't Buy. The frequencies of different values of the attribute with respect to each class is given in each cell of the table. The maximum frequency value for the attribute Income in a particular class is used to calculate the relative importance of the attribute in that class. In the above example* $\bar{m}_{Buy}^{(1)}$ *and* $\bar{m}_{Don'tBuy}^{(1)}$ *are the relative importances indicating relative weights of the attribute in the classes 'Buy' and 'Don't Buy' respectively. The overall importance (weight)* $m^{(1)}$ *of the attribute is the ratio of the summation of the maximum frequencies of the two classes and the total number of instances in the two classes. The calculation of DR based significance of the attribute Income for predicting class Y,* $DR_{Income}$ *is shown in the example and Table 3.*

Table 2: Example dataset

| | **Income** | | |
|---|---|---|---|
| **Class** | *Low* | *Medium* | *High* |
| *Buy* | 2 | 3 | 5 |
| *Don't Buy* | 4 | 2 | 1 |

Table 3: Computed parameter values

| $\bar{m}^{(1)}_{Buy}$ | $\bar{m}^{(1)}_{Don'tBuy}$ | $\bar{m}^{(1)}$ | $DR_{Income}$ |
|---|---|---|---|
| 0.5 | 0.571 | 0.529 | 0.209 |

$\bar{m}^{(1)}_{Buy} = \frac{5}{10} = 0.5$ $\bar{m}^{(1)}_{Don'tBuy} = \frac{4}{7} = 0.571$ $\bar{m}^{(1)} = \frac{9}{17} = 0.529$

$$DR^2{}_{Income} = \frac{num}{den}$$

*where* $num = 10 * (\frac{5}{10} - \frac{9}{17})^2 + 7 * (\frac{4}{7} - \frac{9}{17})^2$ *and* $den = (\frac{2}{17} - \frac{9}{17})^2 + (\frac{3}{17} - \frac{9}{17})^2 + (\frac{5}{17} - \frac{9}{17})^2 + (\frac{4}{17} - \frac{9}{17})^2 + (\frac{2}{17} - \frac{9}{17})^2 + (\frac{1}{17} - \frac{9}{17})^2$

$$DR^2_{Income} = 0.026$$
$$DR_{Income} = 0.161$$

*3.2.3. Proposed Algorithms*

**Algorithm 1** Constructing DR based Decision Tree

**Input:** D,$N_\ell$
**Output:** A decision tree

1: Create initially the root node associating the whole dataset.
2: Choose the best attribute based on Dispersion_Ratio using Algorithm 2
3: Split the dataset based on the attribute chosen in previous step.
4: **for** each subset obtained after splitting **do**
5: a) if all instances are of same class, create leaf node with that class label
6: b) if the subset is empty then assign majority class of the parent node in the associated leaf node;
7: c) if instances belong to different class label, then go to step 2.
8: **end for**

Algorithm 1 show the steps of constructing a Decision Tree using proposed DR technique as the attribute selection criterion for splitting. Initially the root node is created associating the whole dataset. The dataset is further divided based on DR. During Decision Tree construction, at each level, the

DR value between each attribute and the class attribute is computed and the attribute whose DR value with the Class attribute is the highest is chosen as the partitioning attribute for the dataset. The root node gets the label of the corresponding splitting attribute. The branches corresponding to the subtrees of the root node are labeled with different distinct values of the selected root attribute and child nodes are constructed from the root node for each splitted sub-datasets respectively. If any partition has the same class label for all the records, then the associated leaf node has label as same as the corresponding class label. Incidentally, if there is no record in some partition, then the majority class label in its parent's partition is used to label the corresponding leaf node. The same process is iterated until all the data points in each partition have the same class labels.

Our proposed DR based method selects the attribute for splitting the dataset which has the maximum value of DR for a particular class in comparison to other classes. We have considered the relative importance of the $i^{th}$ attribute within each outcome class as the ratio of the two components - the highest frequency value of occurrence for a distinct value of $i^{th}$ attribute with respect to the class and the number of total records in that class. Subsequently, the overall importance of the $i^{th}$ attribute is obtained which is equal to the ratio of the summation of individual class' maximum frequencies and the total number of records in the dataset. Then the ratio of the differences or dispersion in the relative importances of the attribute '$i$' among individual classes and the dispersion across the whole population for the $i^{th}$ attribute is calculated which is equal to the square of the DR for the $i^{th}$ attribute from which the square root is calculated to get the actual DR value as given in Equation 1. Algorithm 2 shows the necessary steps to compute DR using the proposed approach.

The time complexity to select an attribute for splitting using proposed DR based approach is similar to IG based approach. For each node, DR based significance is computed for each of the remaining attributes, which amounts to $O(N)$, where $N$ is the number of attributes.

The proposed DR based method can be applied on datasets having numeric attributes also. In that case, the numeric attributes needs to be discretized first in the preprocessing phase. Any discretization method can be used for that purpose. In (Roy et al., 2016), no heuristic method for discretization has been used and all the numeric attributes were blindly discretized using

**Algorithm 2** Compute DR

**Input:** $A_i$: Attribute $i$ ,$Y$: Class attribute
**Output:** $DR$

1: Find the relative importance of the attribute within each outcome class as the ratio of maximum frequency of a distinct value of that attribute and the number of records in that class.
2: Find the overall importance of the attribute as ratio of summation of maximum frequencies of different classes and the total number of records in the dataset.
3: Compute the ratio between within class dispersion of relative importance and overall dispersion of importance.
4: Get the square root of the ratio computed in previous step and that gives the Dispersion_Ratio.

a fixed number of intervals. But we intend to use an efficient discretization method which has the following properties: 1) Requires a very less number of parameters, 2) Does not require any prior assumption in data distribution. Equal Interval or Equal Width discretization and Equal Frequency discretization (Han et al., 2006) ( Both ) are the popular unsupervised discretization techniques as they are simple to use (Boulle, 2005) and possess the two properties mentioned before. But Equal Interval discretization may result into unbalanced distribution of values into bins (Dash et al., 2011). In case of Equal Frequency discretization, same values of an attribute may not remain in the same group (Dash et al., 2011). So, here we use K-means clustering approach to discretize the numeric features as given in Algorithm 3 as K-means clustering algorithm has all the properties mentioned previously. A clustering algorithm can effectively be used for the purpose of discretization of a numerical attribute, A, by segregating the values of A into clusters or groups. In Clustering, the distribution of the values of A is taken into account. Also the distances among the data points are considered to produce high-quality discretization results (Han et al., 2006). Among unsupervised discretization methods, K-means discretization has shown good performance when used for classification. It is not affected by the selection of classification method as well as indifferent about class labels of instances. Same values will always remain in the same group unlike Equal Frequency discretization method. K-means clustering also have a benefit of handling boundary cases, whereas in case of equal frequency method, two occurrences of the same value

may fall on different sides of the boundary thereby forcing to take arbitrary decision about the inclusion of the value in a particular interval (Maslove et al., 2013). We have used K-means discretization as a pre-processing step. So, our discretization module is suitable to be applied wherever required before performing the main data mining task.

**Algorithm 3** Discretization

**Input:** Dataset $D$ with continuous features, featureset $F$, value of $K$
**Output:** Dataset $D'$

1: **for** each feature in $F$ **do**
2: Find value of $K$ (less or equal to user specified value) which returns best set of clusters
3: Apply K-means clustering approach for discretization
4: Add the discretized feature with values in dataset $D'$
5: **end for**
6: Return new dataset $D'$ with discretized features

The continuous attributes present in different datasets taken from UCI machine learning repository (Dheeru and Karra Taniskidou, 2017) were converted to discrete type using Algorithm 3. User needs to specify some value of $K$ and the algorithm finds the best value of $K$ which returns the best set of clusters. Then it uses this value of $K$ for discretization of the attribute using K-means clustering technique. In each level during growing the Decision Tree, we have splitted the dataset based on the attribute which has the highest value of DR with the class attribute.

Our proposed method has no inclination towards attributes having many distinct values. The following example illustrates this fact.

**Example 3.2.** *Table 4 shows a weather dataset. The DR for different attributes of the dataset in Table 4 have been computed using Algorithm 2 as well as the IG, GR and GI values have been found out for the same dataset and obtained the following results as shown in Table 5:*

Table 4: Weather Dataset

| **Outlook** | **Temp** | **Humidity** | **Windy** | **Play** |
|---|---|---|---|---|
| rainy | cool | low | T | N |
| rainy | mild | low | T | N |
| rainy | hot | high | F | N |
| rainy | mild | high | F | N |
| overcast | mild | high | F | Y |
| overcast | cool | normal | T | Y |
| rainy | mild | normal | F | Y |
| rainy | mild | low | F | Y |
| rainy | mild | normal | T | Y |

Table 5: DR,IG,GR and GI values for weather dataset. The values in bold shows that the corresponding attribute will be selected as the splitting attribute

| | *Outlook* | *Temp* | *Humidity* | *Windy* |
|---|---|---|---|---|
| *IG* | 0.225 | 0.157 | **0.379** | 0.0008 |
| *GR* | 0.295 | 0.129 | **0.24** | 0.000807 |
| *GI* | 0.381 | 0.416 | **0.296** | 0.489 |
| *DR* | **0.540** | 0.351 | 0.152 | 0.239 |

According to the proposed DR method, the most important attribute is *Outlook* and it has two distinct values whereas IG and GR have chosen *Humidity* as the most significant one and *Humidity* has three different values. GI also selects *Humidity*, as it chooses the attribute which has the lowest value of GI.

Another disadvantage of IG method as discussed in (Liu et al., 2010) is with imbalanced dataset since the value of Information Gain decreases in case of imbalanced dataset, given the same true positive rate and false positive rate (Liu et al., 2010). Information Gain uses entropy and the equation for entropy for binary class is given in equation 2 (Han et al., 2006) (Liu et al., 2010):

$$Entropy(t) = -\sum_{j=1}^{2} p(\frac{j}{t})\log_2 p(\frac{j}{t}) \tag{2}$$

In equation 2, $p(\frac{j}{t})$ is the probability of class *'j'* at node *'t'*. So, it is evident that Information Gain is dependent on the class distribution (Liu et al., 2010). On the other hand, the value of DR is not directly dependent on the class distribution (i.e. proportions of classes) since it does not involve any term like $p(\frac{j}{t})$ (see equation 1). Rather it is dependent on the term $n_y$ (number of instances in a class $y$) along with other terms like relative importance of the attribute with respect to class $y$, overall importance of the attribute irrespective of the classes etc. which in turn depend on the frequency distribution of different values of the attribute. There might be one class having much smaller number of examples than another class, but if the within-class dispersion of an attribute for the smaller class is higher compared to the larger class, then its contribution in the computation of DR will be higher. Thus it is not affected by the imbalance class distribution. This claim is supported by the following set of examples where in each case [ Table 6-Table 8 ] we have considered two class labels $C1$ and $C2$ and three possible values of an attribute as: $V1$, $V2$ and $V3$. The values of DR for Table 6 (class distribution - 7:11), Table 7 (class distribution - 9:9) and Table 8 (class distribution - 9:9) are respectively 0.26, 0.21 and 0.37. For balanced class distribution as shown in Table 7 and Table 8, DR values are different. In case of imbalance class distribution (Table 6), the value of DR is more compared to the value of DR for balance class distribution in Table 7. These reflect the fact that the value of DR is not dependent on class distribution and only depends on the distribution of frequencies of the attribute-values with respect to individual classes.

Table 6: Three different values $V1$,$V2$,$V3$ of an attribute with the corresponding frequencies listed in the table (Considering imbalance class distribution)

| | *V1* | *V2* | *V3* |
|---|---|---|---|
| *C1* | 3 | 4 | 0 |
| *C2* | 8 | 2 | 1 |

Table 7: Three different values $V1$,$V2$,$V3$ of an attribute with the corresponding frequencies listed in the table (Considering balance class distribution)

| | $V1$ | $V2$ | $V3$ |
|---|---|---|---|
| $C1$ | 3 | 5 | 1 |
| $C2$ | 6 | 1 | 2 |

Table 8: Three different values $V1$,$V2$,$V3$ of an attribute with the corresponding frequencies listed in the table (Considering balance class distribution)

| | $V1$ | $V2$ | $V3$ |
|---|---|---|---|
| $C1$ | 4 | 5 | 0 |
| $C2$ | 7 | 1 | 1 |

## 4. Experimental Results and Analysis

We have experimented our DR based approach with various datasets. Among them many datasets are from medical domain. One of the general characteristics of medical datasets is their variation. The variation not only reflects in number of classes but also in types of attributes. So, we have used most of the datasets from medical domain so that we can demonstrate the performance of our method on datasets of varied nature. However, to check the applicability of our proposed method, we have also used some datasets from other domain too.

Overall we have used datasets such as Pima Indian Diabetes, Mammography Masses, Spect-heart, Statlog(heart), Diabetic Retinopathy Debrecen, Ecoli, Thyroid (Allbp), Thyroid (Allhyper), Thyroid (Allhypo), Thyroid (Allrep), Hayes Roth, Mushroom, Bank Marketing, Credit Approval, Bankruptcy, Congressional and Balance Scale. All these datasets are taken from UCI machine learning repository (Dheeru and Karra Taniskidou, 2017).

### *4.1. About the datasets*

Table 9 shows the characteristics of the datasets considered here. Traditional CR is suitable for datasets having all numeric predictor attributes. In case the dataset has only nominal attributes or mixture of nominal and numeric attributes (numeric attributes need to be discretized) then DR method is suitable. The last column indicates the method(s) applicable to a particular dataset.

Table 9: Nature of the Dataset

| **Dataset** | **Domain** | **Number of instances** | **No. of attributes** | **Number of numeric attributes** | **No. of classes (No. of examples in each class)** | **Presence of Missing Values** | **Method(s) applicable** |
|---|---|---|---|---|---|---|---|
| Diabetic Retinopathy | Life | 1151 | 20 | Sixteen real attributes | 2 (540:611) | No | CR and DR |
| Ecoli | Life | 327 | 6 | All attributes are numeric | 5 (143:77:35:20:52) | No | CR and DR |
| Hayes Roth | Life | 132 | 4 | All nominal attributes | 3 (51:51:30) | No | DR |
| Mammography | Life | 961 | 6 | Only one integer attribute and discretized into 2 distinct values | 2 (516:445) | yes | DR |
| Mushroom | Life | 8124 | 22 | All nominal attributes | 2 (4208:3916) | Yes | DR |
| Pima Indian Diabetes | Life | 768 | 9 | All numeric values and all discretized into two different distinct values | 2 (500:268) | No | CR and DR |
| Spect-heart | Life | 267 | 23 | All are binary attributes | 2 (55:212) | No | DR |
| Statlog (heart) | Life | 270 | 13 | Six real attributes | 2 (151:119) | No | DR |
| Thyroid (Allbp) | Life | 2800 | 27 | Six attributes numeric | 3 (9:124:2667) | Yes | DR |
| Thyroid (Allhyper) | Life | 2800 | 27 | Six attributes numeric | 4 (7:62:2723:8) | Yes | DR |
| Thyroid (Allhypo) | Life | 2800 | 27 | Six attributes numeric | 4 (154:2580:64:2) | Yes | DR |
| Thyroid (Allrep) | Life | 2800 | 27 | Six attributes numeric | 4 (2713:23:29:35) | Yes | DR |
| Bank Marketing | Business | 4521 | 15 | Seven numeric attributes | 2 (4000:521) | No | DR |
| Bankruptcy | Financial | 250 | 6 | All nominal attributes | 2 (143:147) | No | DR |
| Credit Approval | Financial | 690 | 15 | Six continuous attributes | 2 (307:383) | Yes | DR |
| Balance Scale | Social | 625 | 4 | All nominal attributes | 3 (49:288:288) | No | DR |

The continuous attributes contained in different datasets are converted to discrete type using Algorithm 3. The characteristics of the datasets after discretization has been shown in Table 9. The traditional Correlation Ratio method can be applied in some of these datasets as mentioned in Table 9. For most of these datasets k-fold cross validation has been performed in which the dataset is splitted into $k$ disjoint subsets and $(k-1)$ subsets are considered for training and for testing the model, the remaining subset is used. This process is iterated overall k-number of times and the results of all the iterations are integrated together(Han et al., 2006). The Spect-heart dataset is already available in two subsets - separate training (80 instances) and test sets (187 instances). The four different Thyroid datasets are also divided into training and test sets having 2800 instances and approximately 972 instances each respectively. Since Mushroom and Bank Marketing datasets are quite huge, so they are divided into separate training and test sets (Mushroom - 80%:20%, Bank Marketing - 70%:30%).

### *4.2. Results and Analysis*

Next we have applied our proposed DR technique on the datasets discussed in Subsection 4.1. Among the sixteen datasets we have considered here, five are common with Correlation Ratio based Decision Tree (CRDT) approach as shown in (Roy et al., 2016). For those five cases, we have provided the comparative analysis with DR based approach in Table 10. Among these five datasets, in comparison to the method in (Roy et al., 2016), increase in performances of the proposed DR method have been observed for three datasets, marginal drop in performance for one dataset and for the remaining one the performance is same as shown in Table 10. So, the other remaining datasets of (Roy et al., 2016) have not been considered here. The result clearly indicates that in most of the cases, DR performs better than CRDT approach.

Out of the total sixteen datasets, DR model has performed best in ten cases, in four other cases GR model has given best accuracies, GI has given highest accuracy for the one dataset and all the three other models - IG, GR and GI have given highest performance in one remaining case.

Table 11 shows that DR approach has performed well for small datasets (like Spect-heart, Hayes-Roth) and also for large datasets (like Allbp, Allhypo, Mushroom etc.) For Diabetic Retinopathy dataset, even though DR model

Table 10: Comparison with CR based method

| **Dataset** | **DR Accuracy** | **Modified CR Accuracy (Roy et al., 2016)** |
|---|---|---|
| Diabetic Retinopathy | **62.09**% | 61.25% |
| Mammography | **90.82**% | 80.96% |
| Pima Indian Diabetes | 70.96% | **71.09**% |
| Spect-heart | **78.61**% | **78.61**% |
| Statlog (heart) | **76.62**% | 74.69% |

given maximum performance among all, the accuracy percentage is not good. The reason for this is all the attributes were originally numeric (15 attributes) except 4 attributes and the numeric attributes were converted to discrete attributes using K-means clustering algorithm in the preprocessing phase. So, this discretization method may not be always proper. Similar is the case for Pima Indian Diabetes dataset where all the attributes are numeric. Despite that DR model has given the highest accuracy for this dataset also. All the attributes were discretized into equal number of categorical values. The dataset has no missing value. Actually some attributes, for example blood pressure contains zero values and possibly here the missing data are encoded as zero values.

DR approach has given highest accuracy in datasets having all nominal attributes like Spect-heart, Mushroom, Hayes Roth and Balance Scale and the difference between the accuracies with other approaches is comparatively high. On the other hand for the datasets having all nominal attributes like Bankruptcy- though DR model did not perform best but the difference between the accuracies of DR based method and the best performer is less. Even though there are combination of numerical attributes(originally) and more number of nominal attributes in the datasets like - Bank Marketing, Thyroid (Allbp), Thyroid (Allhypo), Thyroid (Allrep) and Mammography but since, these datasets have large number of data points, so our proposed approach have shown best or almost same accuracy as the best performing model.

For most of the datasets with more than two classes (like Ecoli, Hayes Roth, Balance Scale, All four Thyroid datasets) our proposed approach has resulted in highest or almost same performance as the best performing model. For

Table 11: Experimental Results

| Dataset | Cross validation | DR Accuracy | IG Accuracy | GR Accuracy | GI Accuracy | Difference in Performance between DR approach and nearest best performer |
|---|---|---|---|---|---|---|
| Diabetic Retinopathy | 5-fold | **62.09**% | 61.42% | 61.66% | 60.46% | 0.43% |
| Ecoli | 5-fold | **82.78**% | 81.04% | 80.37% | 78.32% | 1.74% |
| Hayes Roth | 5-fold | **76.52**% | 70.46% | 71.97% | 46.97% | 4.55% |
| Mammography | 5-fold | **90.82**% | 89.98% | 90.58% | 68.48% | 0.24% |
| Mushroom | 6499 training instances and 1625 test instances | **95.24**% | 91.96% | 81.29% | 87.19% | 3.28% |
| Pima Indian Diabetes | 5-fold | **70.96**% | 70.31% | 70.31% | 67.45% | 0.65% |
| Spect-heart | 80 training and 187 test instances | **78.61**% | 74.33% | 75.93% | 74.33% | 2.68% |
| Statlog (heart) | 5-fold | 75.84 % | 71.4% | **76.62**% | 75.11% | -0.78% |
| Thyroid (Allbp) | 2800 training instances and 972 test instances | **95.99**% | 94.86% | 95.58% | 91.98% | 0.41% |
| Thyroid (Allhyper) | 2800 training instances and 971 test instances | 97.12% | 97.94% | **98.15**% | 97.74% | -1.03% |
| Thyroid (Allhypo) | 2800 training instances and 972 test instances | **91.87**% | 90.33% | 91.77% | 70.27% | 0.1% |
| Thyroid (Allrep) | 2800 training instances and 972 test instances | 95.99% | 96.4% | 96.5% | **96.6**% | -0.61% |
| Bank Marketing | 3166 training instances and 1355 test instances | 86.28% | 85.18% | **86.43**% | 85.31% | -0.15% |
| Bankruptcy | 5-fold | 98.4% | **99.6**% | **99.6**% | **99.6**% | -1.2% |
| Credit Approval | 5-fold | 78.55% | 78.26% | **79.13**% | 78.4% | -0.58% |
| Balance Scale | 5-fold | **37.9**% | 33.6% | 36.32% | 33.92% | 1.58% |

very small datasets with less number of attributes where all the attributes are nominal like spect-heart and Hayes-Roth datasets, DR approach has given best performances amongst all.

IG based model has exhibited highest accuracy for Bankruptcy dataset as it is having same number of categorical values in all the attributes. So, the inclination of IG towards many different valued attributes had no effect for this dataset.

The Gini Index (GI) based approach have given best performances in few (4) cases and showed much lower accuracy in most of the cases as shown in Table 11. Since GI has a bias towards attributes with many different values, so it is not the best performer for datsets like Diabetic Retinopathy, Mammography etc. as there is the possibility to choose non-informative attributes over informative ones. But it has given best performances in case of Bankruptcy dataset because this dataset is having same number of distinct values for all

attributes and due to this there is no influence of the above mentioned bias. Since the dataset is almost balanced, it is a favourable condition for IG based decision tree learner.

The GR based model overcomes the disadvantage of Information Gain and GI models by reducing the bias towards many distinct valued attributes. So, it stands as the highest performing model for the datasets like Statlog (Heart), and Thyroid (Allhyper) datasets both of which are having lots of distinct valued attributes.

Overfitting is a concern for Decision Tree algorithm and there are different methods to handle this like Decision Tree pruning, cross-validation (Domingos, 2012) (Kane, 2017) etc. which are applicable to our model also. To make sure during our experiment that the model is not overfitting the training data, we have used cross-validation.

We have conducted non-parametric Sign test (Dixon and Mood, 1946) to find out whether there is a significant difference in accuracies between the DR based model and other models considered in the experiment. This test does not require any assumption on the data distribution. The test was conducted to compare the accuracies over sixteen datasets between the DR based model and each one of the IG, GR and GI based models. The test was performed considering two models at a time. The null hypothesis tells that the performances of the models considered for comparison are equivalent and the alternate hypothesis in this case is that there is significant difference in the performances of the models. The significance level is considered to be 5%. The results of Sign test are shown in the Table 1:

Table 12: Sign Test Results

| **Model considered** | $p$ **value** |
|---|---|
| DR model and IG model | **0.01242** |
| DR model and GR model | 0.31731 |
| DR model and GI model | **0.01242** |

So, Table 12 shows that indicates that the difference in the performances between DR and IG based models as well as DR and GI based models are significant at $p < 0.05$ but there is no statistical difference in the performances between DR and GR based models at $p < 0.05$. Since GR is also

not suffering from the bias towards attributes with many distinct values, so performances of the DR and GR based models are nearly equivalent as it is also evident from the Sign test results.

### *4.3. General Observation*

The observation made on the basis of analysis of the result is that generally our proposed approach is able to handle datasets with less as well as large no of distinct valued attributes with almost equal efficiency because it is not inclined towards attributes with more number of distinct values. On the other hand, IG and GI approaches prefer attributes with lots of distinct values. In healthcare datasets where there are many attributes and different attributes possess different numbers of distinct values, the IG and GI based approaches prefer attributes with more number of distinct values in place of attributes which have few distinct values even if the latter one may be comparatively more significant from the perspective of classification. This gives a reason for lesser performance of the IG and GI based models than our proposed approach in those cases where some not so important attributes are having many distinct values. But if the attributes are highly relevant then IG/GI based approach performs well. Another major observation is that our approach exhibits best performance for mainly those datasets which do not contain any missing values in the original dataset.

Though our main aim is to propose a better alternative for Information Gain based Decision Tree model, yet for comparison purpose we have depicted the results obtained from Gain Ratio and Gini Index based Decision Tree. After applying the Gain Ratio and Gini Index based models on the same set of datasets we obtained the results shown in Table 11.

It is evident that our proposed approach has exhibited better results in terms of accuracy compared to Gain Ratio and Gini Index based approaches. The Gain Ratio based approach has given almost same accuracies as IG based model in many cases as it is a variant of IG based method.

GI based technique has also an inclination towards attributes having many distinct values. And the problem is higher when there are more number of classes. So GI based model has given less performance for cases having more number of classes.

Some more general observations are listed below:

- Bankruptcy dataset is a well-balanced dataset, so all performed well. Also it has no missing values.
- Mushroom dataset is not fully balanced (Class 'e': 4208 and class 'p': 3916) but it is a large dataset. So most models have performed well.
- All the four Thyroid datasets are having few numerical and more number of nominal attributes and each of the datasets is quite large. So, all have have performed well on these datasets.
- For the datasets having more number of (or all) nominal/categorical attributes compared to numeric attributes, all the methods have provided more than 85% Accuracy. Since it is difficult to find the exact value of K in case of K-means discretization technique, the proposed model has achieved comparatively lesser accuracy for datasets having more number of numeric attributes than nominal attributes.

## 5. Conclusion

Now data are available in plenty. Datasets may contain many heterogeneous attributes and many class labels. Some datasets may have large number of attributes and different attributes have different numbers of distinct values and sometimes there may be very less number of instances. Because of such diverse nature of datasets, existing IG, GR or GI based splitting criterion may not always perform well. So, in this paper, we have presented another alternate splitting criterion based on Dispersion Ratio. The proposed method is known as DR based Decision Tree learner. We have also presented a discretization module for numeric dataset. The result shows that the proposed DR method along with the integrated discretization module performs quite well compared to IG, GI, GR, CRDT based techniques in many cases. Our proposed model has the following main advantages - it has no inclination towards attributes with large number of distinct values, it works well with datasets having very few instances and less number of attributes, it also handles well the datasets with more than two classes and not affected by imbalance class distribution. Thus, the method provides a nice complement of the existing approaches. In future we would like to evaluate our model using some more datasets from different other domains and also we would like to develop better discretization technique for numerical type attributes.

## 6. Acknowledgements

We are thankful to the University of California, Irvine, School of Information and Computer Sciences for providing different datasets through their repository for research purpose.

## References

Battiti, R., Brunato, M., and Mascia, F. (2009). *Reactive Search and Intelligent Optimization*. Springer US, US.

Boulle, M. (2005). Optimal bin number for equal frequency discretizations in supervized learning. *Intelligent Data Analysis*, 9(2):175–188.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC, Boca Raton.

Brueckner, R. (2018). Inside Bigdata. Retrieved from https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/. Accessed 05 April 2018.

Chandra, B., Kothari, R., and Paul, P. (2010). A new node splitting measure for decision tree construction. *Pattern Recognition*, *43*(8):2725–2731. doi:https://doi.org/10.1016/j.patcog.2010.02.025.

Crathorne, A. R. (1922). Calculation of the Correlation Ratio. *Journal of the American Statistical Association*, *18*(139):394–396. https://doi.org/10.1080/01621459.1922.10502484.

Dash, R., Paramguru, R. L., and Dash, R. (2011). Comparative analysis of supervised and unsupervised discretization techniques. *International Journal of Advances in Science and Technology*, 2(3):175–188.

Dheeru, D. and Karra Taniskidou, E. (2017). UCI machine learning repository. Retreived from http://archive.ics.uci.edu/ml. Accessed 05 April 2018.

Dianhong, W. and Liangxiao, J. (2007). An improved attribute selection measure for decision tree induction. In *Fourth International Conference Proceedings on Fuzzy Systems and Knowledge Discovery_FSDK 2007*, pages 654–658, Haikou, China. IEEE.

Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566. https://doi.org/10.2307/2280577.

Domingos, P. (2012). A Few Useful Things to Know about Machine Learning. *Communications of the ACM*, 55(10):78–87.

Domingos, P. and Hulten, G. (2000). Mining high-speed data streams. In *KDD 2000, Boston, MA USA*, pages 71–80. ACM.

Domingos, P. and Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning*, *29*(2-3):103–130. doi:https://doi.org/10.1023/A:1007413511361.

Gama, J. and Rocha, R. (2003). Accurate Decision Trees for Mining High-Speed Data Streams. In *KDD '03 Proceedings of the ninth ACM SIGKDD International Conference on Knowledge discovery and data mining*, pages 523–528, Washington, D.C. ACM New York, NY, USA ©2003.

Han, J., Kamber, M., and Pei, J. (2006). *Data Mining Concepts and Techniques (3rd ed.)*. Morgan Kaufman, USA:Waltham.

John, G. H. and Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, Montral, Qu, Canada. Morgan Kaufmann Publishers Inc.

Kane, F. (2017). *Hands-On Data Science and Python Machine Learning*. Packt Publishing Limited.

Kantardzic, M. (2002). *Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition*. Wiley. https://doi.org/10.1002/9781118029145.

Kozak, J. and Boryczka, U. (2016). Collective data mining in the ant colony decision tree approach. *Information Sciences*, 372:126–147. https://doi.org/10.1016/j.ins.2016.08.051.

Liu, W., Chawla, S., Cieslak, D. A., and Chawla, N. V. (2010). A Robust Decision Tree Algorithm for Imbalanced Data Sets. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 766–777, Columbus, Ohio. Society for Industrial and Applied Mathematics.

Maslove, D. M., Podchiyska, T., and Lowe, H. J. (2013). Discretization of continuous features in clinical datasets. *Journal of the American Medical Informatics Association: JAMIA*, 20(3):544–553. http://doi.org/10.1136/amiajnl-2012-000929.

Murata, N., Yoshizawa, S., and Amari, S. (1994). Network Information Criterion-determining the Number of Hidden Units for an Artificial Neural Network Model. *IEEE Transactions on Neural Networks*, *5*(6):865–872. doi:https://doi.org/10.1109/72.329683.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

Roy, S., Mondal, S., Ekbal, A., and Desarkar, M. (2016). CRDT : Correlation Ratio Based Decision Tree Model for Healthcare Data Mining. In *IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 36–43, Taichung, Taiwan. IEEE.

Sun, X., Liu, Y., Xu, M., Chen, H., Han, J., and Wang, K. (2013). Feature Selection Using Dynamic Weights for Classification. *Knowledge-Based Systems*, *37*:541–549. https://doi.org/10.1016/j.knosys.2012.10.001.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory (2nd ed.)*. Springer, New York.

Wang, J., Zhou, S., Yi, Y., and Kong, J. (2014). An Improved Feature Selection Effective Range for Classification. *The Scientific World Journal*, *2014*:8 pages. http://dx.doi.org/10.1155/2014/972125.

Weisstein and Eric, W. (1951). Correlation Ratio.From MathWorld–A Wolfram Web Resource. Retrieved from http://mathworld.wolfram.com/CorrelationRatio.html. Accessed 05 April 2018.