Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.125 – 135

RESEARCH ARTICLE

A Mean Deviation Based Splitting Criterion for Classification Tree

Syed Jawad Ali Shah[1], Qamruz Zaman[2]

[1]Department of Statistics, University of Peshawar, Pakistan

[2]Department of Statistics, University of Peshawar, Pakistan

[1] jawadalishah@hotmail.com; [2] ayanqamar@gmail.com

Abstract— For the learning of Classification Tree, many researchers have used different splitting criteria, in which most commonly impurity-based criteria are: Gini index, Entropy function and Exponent-based index. By comparing Misclassification rates, none of the splitting criterion can be declared as providing best results in every situation. In this study, a new Mean Deviation based index has been proposed and a simulation study is designed, to explore which measure gives best result in what type of situation? From simulation study, it is concluded that generally new proposed M.D-based index and Exponent-based index give excellent results in case of imbalanced data. While, in case of balanced data, Gini index and Entropy function have less misclassification rates.

Keywords— "Classification Tree", "Misclassification rate", "Impurity-based criteria", "Simulation", "Imbalanced data"

I. INTRODUCTION

Now-a-days, in the era of information technology, private/public organizations collect massive amounts of data for processing and analyzing to make best policy. Before analyzing the data set, the analyst looks at the nature of data as well as types of variables. In case of multivariate data, sometimes the analysts are interested to classify the data into various homogenous classes according to their resemblance. They usually apply the Discriminant Analysis technique, Logistic Regression, Neural network and Cluster analysis. Another technique for classifying the multivariate data is Classification and Regression Tree, which is a non-parametric technique having Tree-based structure and is extensively used in biological and social sciences field for classification or prediction purpose.

The Classification or Regression Tree is a Non-parametric model, represented in the form of binary or multiple Trees with the objective to make predictions for the unobserved data examples/units. The difference between Classification Tree and Regression Tree depends on the nature of dependent/response variable. If the response variable is categorical, the resultant Tree is known as Classification Tree. On the other hand, if the response variable is continuous, the resultant Tree is known as Regression Tree.

Tree based Classifiers are used successfully in many diverse areas such as radar signal classification, character recognition, medical diagnosis, species of plants, survival rate after a surgical treatment, speech recognition and many more.

Leo Breiman is the pioneer of Classification & Regression Tree [1]. To grow a Classification or Regression Tree, a Greedy algorithm is adopted by recursively splitting the Tree nodes into two or more than two child nodes. At each split, one variable is selected on the basis of a splitting criterion and a left/right branching decision is made according to some threshold value of variable. This process is repeated until a full Tree is constructed (i.e. all variables are utilized for splitting). After learning the Tree model, a new observation/unit is classified by simply following the top to bottom path of the Tree until a terminal node is reached.

There are different nodes splitting functions or evaluation criteria available in the literature for the construction of binary Classification Tree. For example, Gini index, Twoing rule, Entropy function, and Exponent-based index.

In this study, an attempt is made to propose a splitting measure, which produces less misclassification error as compared to existing splitting measures in every type of data, or if it is not so possible, then it might be explored which measure gives best result in which type of situation?

II. LITERATURE REVIEW

Decision Tree Classifiers (DTC's) are used successfully in many diverse areas such as radar signal classification, character recognition, medical diagnosis, speech recognition and many more. These Classifiers are not newer, but have been used since 1944. They have their origin in game theory, developed by Von Neumann and Morgenstern in 1944 [2].

In 1960, another Tree-based Classifier AID (Automatic Interaction Detection) was developed by Morgan and Sonquist [3], which laid down the foundation of Regression Tree. Later on, in the 1970s, Morgan and Messenger [4] created THAID (Theta AID), which played a vital role in the development of Classification Trees. AID and THAID were developed at the Institute for Social Research at the University of Michigan. With the passage of time, more development was done in the algorithm of Tree models, and in 1980's, a famous statistician Leo Breiman developed Classification & Regression Tree.

The growth of Classification Tree follows a process of recursive partitioning from top to end nodes. There are different nodes splitting functions or evaluation criteria available in the literature for the construction of binary Classification Tree. The most standard and commonly used splitting criteria are: Gini index, Entropy function, Twoing Rule and Exponent-Based index.

A. Gini Index

The Gini coefficient (also known as the Gini index or Gini ratio) developed by the Corrado Gini [5], and was used by Leo Breiman as a splitting measure for the learning of Classification Tree. For a given node ***t***, Gini index is given by:

$$i(t)_G = 1 - \sum_{j=1}^{J} p_j^2$$

where ***j*** is the total number of categories/classes, and p_j is the proportion of **j**th class in a node ***t***.

Reference [6] discussed that this measure looks for the largest class in the data set and tries to isolate it from all other classes, while Breiman explained that Gini index tends to choose a split that divides the data into two different size nodes (i.e. one small pure node and a large impure node). So, there was a need to utilize another splitting criterion, which reduces the impurity after dividing a node.

B. Entropy Function

Another commonly used function, the Entropy function was introduced by Shannon [7], while Quinlan [8] utilized it as a splitting measure for the construction of Classification Tree. For a given node ***t***, Entropy function is given by:

$$i(t)_Q = -\sum_{j=1}^{J} p_j \log_2 p_j$$

C. Twoing Rule

Another splitting rule known as Twoing Rule was introduced by Breiman [1]. It measured the difference in probabilities of a class appearing in the left and right child node. So this splitting criterion was thus based on separation rather on node impurity. The criterion ensured to select a best splitting value, which maximizes:

$$g(x,s,t)_T = \frac{p_L p_R}{4}\left[\sum_{j=1}^{J} | p_{j/L} - p_{j/R} |\right]^2$$

where t_L and t_R are two left and right child nodes, $P_{j/L}$ and $P_{j/R}$ are the proportions of **j**th class on left and right child node respectively.

Although, Breiman found that both Twoing and Gini generally produced the same initial splits, but there was a difference. The Twoing rule had tendency to choose a split that produce equal size nodes, while Gini tends to choose a split that divide the data into two different size nodes (i.e. one small pure node and a large impure node).

D. Exponent-based index

Recently, another splitting measure based on exponent, has been proposed by Azam [9]. The Exponent based index is given by:

$$i(t)_{exp} = 1 - \frac{1}{e}\sum_{j=1}^{J} p_j e^{p_j}$$

III. PROPOSED MEAN DEVIATION BASED SPLITTING CRITERION

Suppose there are N observations in the learning/Training data set and N_j is the overall number of observations belonging to class j, where $j=1,2,\ldots,J$. Then the **j**th class probability can be defined as:

$$\pi_j = \frac{N_j}{N}$$

That is, the proportion of units belonging to **j**th class relative to overall number of units in the Training sample.

Let $N(t)$ be the number of observations in node t and $N_j(t)$ denotes the number of observations belonging to **j**th class in the same node t. Then a joint probability of the event that an observation of **j**th class falls into node t is:

$$p(j,t) = \pi(j)\frac{N_j(t)}{N_j}$$

Therefore,

$$p(t) = \sum_{j=1}^{J} p(j,t)$$

and conditional probability of a unit belonging to node t, given that its class is j, can be computed as:

$$p(j \mid t) = \frac{p(j,t)}{p(t)} = \frac{N_j(t)}{N(t)}$$

That is, the proportion of class j in node t. It is obvious that

$$\sum_{j=1}^{J} p(j \mid t) = 1$$

By definition, a sample mean deviation estimate for node t observations will be:

$$\text{sample } M.D = \frac{1}{N}\sum_{i=1}^{N}\left| x_i - \frac{\sum x_i}{N}\right|$$

$$= \frac{1}{N}\left[\left|x_1 - \frac{\sum x_i}{N}\right| + \left|x_2 - \frac{\sum x_i}{N}\right| + \ldots + \left|x_N - \frac{\sum x_i}{N}\right|\right]$$

Suppose some observations belong to **j**th class (assigning one (**1**) for it) and some belongs to ***j**** (assigning zero (**0**) for it). Let total number of observations belonging to **j**th class is ***k***. Then,

$$= \frac{1}{N}\left[\left|1 - \frac{k}{N}\right| + \left|0 - \frac{k}{N}\right| + \ldots + \left|1 - \frac{k}{N}\right|\right]$$

$$= \frac{1}{N}\left[k\left|1 - \frac{k}{N}\right| + (N-k)\left|0 - \frac{k}{N}\right|\right]$$

which, after simplification becomes,

$$= 2(1 - p_{j/t})p_{j/t}$$

where $p_{j/t} = \frac{k}{N}$

taking summation for all ***j***'s (i.e. more than two classes), it becomes

$$\text{Proposed M.D based index} = \sum_{j=1}^{J} 2(1 - p_j)p_j$$

or

$$\text{Proposed M.D based index} = 2\left[1 - \sum_{j=1}^{J} p_j^2\right]$$

So the proposed M.D-based index is a function of P_j i.e. $f(p_1, p_2, \ldots, p_j)$.

Now the question arises, whether the proposed M.D-based index will reduce the impurity after splitting node or not? The answer to this question is "Yes", because the proposed index is a Convex function, and it can be proved by the following property of Convex function:

Any function will be considered as Convex function, if it satisfies the property:

$$f\{tx_1 + (1-t)x_2\} > tf(x_1) + (1-t)f(x_2) \quad \ldots\ldots(1)$$

To show that the proposed index is a convex function, proceed by taking left hand side of inequality (1), as:

$$f\{tp_1 + (1-t)p_1^*, \ldots, tp_j^* + (1-t)p_j^*\}$$
$$= 2[1 - \sum tp_j^2] + 2[1 - (1-t)\sum p_j^{*2}]$$
$$= 2[2 - t\sum p_j^2 - (1-t)\sum p_j^{*2}] \quad \ldots\ldots(2)$$

Now taking the Right hand side of inequality,

$$tf(p_1, p_2, \ldots, p_j) + (1-t)f(p_1^*, p_2^*, \ldots, p_j^*)$$
$$= t\{2[1 - \sum p_j^2]\} + (1-t)\{2[1 - \sum p_j^{*2}]\}$$
$$= 2[1 - t\sum p_j^2 - (1-t)\sum p_j^{*2}] \quad \ldots\ldots(3)$$

Comparing results (2) and (3), it is obvious that the condition for convex function given in (1) holds. Hence the proposed M.D-based index is a convex function, and guarantees that the impurity will reduce after splitting.

A. Properties of Splitting Criterion

Reference [1] illustrated that any function **φ** will be called as impurity function if it satisfies the following properties:

1. φ must attains its maximum value at point $\left(\frac{1}{j},\frac{1}{j},\ldots,\frac{1}{j}\right)$.
2. φ must attains its minimum value at points (1,0,…,0), (0,1,...,0),...,(0,0,…,1).
3. φ must be a symmetric function of $p_1,p_2,\ldots,p_j$.

The proposed splitting criterion is also an impurity function, because it satisfies the above conditions/properties. The proof of each property is as under:

1) The proposed function attains its maximum value at point $\left(\frac{1}{j},\frac{1}{j},\ldots,\frac{1}{j}\right)$.

The lagrange's function under constraint $\sum p_j = 1$ will be:

$$L = 2[1-\sum p_j^2] + \lambda[\sum p_j - 1]$$

$$\frac{\partial L}{\partial p_j} = \frac{\partial}{\partial p_j}[2(1-p_1^2-p_2^2-\ldots\ldots-p_j^2)+\lambda(p_1+p_2+\ldots.+p_j-1)] = 0$$

$$\Rightarrow -4p_j + \lambda = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

$$\Rightarrow p_j = \frac{\lambda}{4}$$

Also the second derivative is negative, so the function maximized on value of p_j. Now taking summation both the sides of (4) over all j's , and putting value of λ, we get

$$p_j = \frac{1}{j}$$

2) The proposed function attains minimum at points (1,0,…,0), (0,1,…,0),…,(0,0,…,1).

First, let us find out the minimum value of the function.

$$\text{Proposed M.D based index} = 2\left[1-\sum_{j=1}^{J} p_j^2\right]$$

So it depends on $\sum p_j^2$. And

$$\sum p_j^2 = p_1^2 + p_2^2 + \ldots + p_j^2$$

$$= \left(\frac{N_1}{N}\right)^2 + \left(\frac{N_2}{N}\right)^2 + \ldots + \left(\frac{N_j}{N}\right)^2$$

Therefore,

$$1-\sum p_j^2 \geq 0 \quad \Rightarrow \quad 2\left(1-\sum p_j^2\right) \geq 0$$

It means the possible minimum value of the proposed index will be zero. Now to examine, at what values of P_j it attains the minimum value, consider:

$$2\left(1-\sum p_j^2\right) = 0$$

$$\Rightarrow \sum p_j^2 = p_1^2 + p_2^2 + \ldots + p_j^2 = 1$$

Here only one equation and j unknowns, so a unique solution does not exist.

Therefore keeping in view, the restriction $\sum p_j = 1$, if taking p_1=1, then definitely the remaining p's will be zero (i.e. p_2=0,..., p_j=0). So, one possible solution is (1, 0, 0, …, 0). Similarly, the other possible solutions are (0, 1, 0, …, 0), …, (0, 0, 0, …, 1).

3) ***φ*** *should be a symmetric function of* $p_1, p_2, \ldots, p_J$.

To show mathematically, that a given function ***f(x)*** is symmetric; it is to prove that $f(\alpha - x) = f(\alpha + x)$ or $f(x - \alpha) = f(x + \alpha)$, where ***α*** is any constant and the function will be symmetric at ***x*** (usually the mode of the function). Here, it is already proved that $p_j = \frac{1}{j}$. So by mathematical induction, proceed as:

First, let us take the case of only two classes (p_1, p_2), our proposed formula will be

$$f(p_1, p_2) = 2(1 - p_1^2 - p_2^2)$$

$$f\left(\frac{1}{2} - \alpha, \frac{1}{2} + \alpha\right) = 2\left[1 - \left(\frac{1}{2} - \alpha\right)^2 - \left(\frac{1}{2} + \alpha\right)^2\right]$$

and

$$f\left(\frac{1}{2} + \alpha, \frac{1}{2} - \alpha\right) = 2\left[1 - \left(\frac{1}{2} + \alpha\right)^2 - \left(\frac{1}{2} - \alpha\right)^2\right]$$

Therefore

$$f\left(\frac{1}{2} - \alpha, \frac{1}{2} + \alpha\right) = f\left(\frac{1}{2} + \alpha, \frac{1}{2} - \alpha\right)$$

Now, consider the case for three classes (p_1, p_2, p_3), the proposed formula will be

$$f(p_1, p_2, p_3) = 2(1 - p_1^2 - p_2^2 - p_3^2)$$

$$f\left(\frac{1}{3} + \alpha, \frac{1}{3} - \alpha, \frac{1}{3}\right) = 2\left[1 - \left(\frac{1}{3} + \alpha\right)^2 - \left(\frac{1}{3} - \alpha\right)^2 - \left(\frac{1}{3}\right)^2\right]$$

and

$$f\left(\frac{1}{3} - \alpha, \frac{1}{3} + \alpha, \frac{1}{3}\right) = 2\left[1 - \left(\frac{1}{3} - \alpha\right)^2 - \left(\frac{1}{3} + \alpha\right)^2 - \left(\frac{1}{3}\right)^2\right]$$

and

$$f\left(\frac{1}{3}, \frac{1}{3} - \alpha, \frac{1}{3} + \alpha\right) = 2\left[1 - \left(\frac{1}{3}\right)^2 - \left(\frac{1}{3} - \alpha\right)^2 - \left(\frac{1}{3} + \alpha\right)^2\right]$$

and for all other combinations, these all are equal. Hence, the function is symmetric about $\frac{1}{3}$.

Now, generalizing for ***j*** classes ($p_1, p_2, p_3, \ldots, p_j$), we have

$$f\left(\frac{1}{j} + \alpha, \frac{1}{j} - \alpha, \frac{1}{j}, \ldots, \frac{1}{j}\right) = 2\left[1 - \left(\frac{1}{j} + \alpha\right)^2 - \left(\frac{1}{j} - \alpha\right)^2 - \left(\frac{1}{j}\right)^2 - \ldots - \left(\frac{1}{j}\right)^2\right]$$

and

$$f\left(\frac{1}{j} - \alpha, \frac{1}{j} + \alpha, \frac{1}{j}, \ldots, \frac{1}{j}\right) = 2\left[1 - \left(\frac{1}{j} - \alpha\right)^2 - \left(\frac{1}{j} + \alpha\right)^2 - \left(\frac{1}{j}\right)^2 - \ldots - \left(\frac{1}{j}\right)^2\right]$$

and similarly for all other combinations, these all will be equal.

Hence, for ***j*** classes, the proposed function is symmetric about $\frac{1}{j}$.

IV. SIMULATION RESULTS

In this section, a simulation study is designed in R for comparing the new proposed M.D-based splitting criterion and other three impurity based criterion. In the first stage, data has been generated (consisting of 100 observations, because [1] has suggested that 100 cases are sufficient for Classification Tree), from three continuous distributions: (i) Normal distribution; (ii) Exponential distribution; (iii) Uniform distribution; and three discrete distributions: (iv) Binomial distribution (v) Poisson distribution (vi) Hypergeometric distribution, and the Misclassification Error has been recorded.

In the second stage, the threshold values have been adjusted to generate Imbalanced data from the above distributions and the Misclassification Rates are compared. Further, model is initially fitted by using 3 variables (2 independent variables and 1 dependent variable) and then it is extended for 4 variables (3 independent variables and 1 dependent variable) just like step-wise regression techniques.

The simulation results are presented in the next eight tables (i.e. from Table I to Table IV).

TABLE I

SIMULATION RESULTS OF BALANCED DATA GENERATED THROUGH CONTINUOUS DISTRIBUTIONS

Continuous Distribution	**Number of iterations**	**Misclassification Rate (03 variables)**				**Misclassification Rate (04 variables)**			
		Gini Index	**Entropy function**	**Exponent-based function**	**Proposed M.D-based index**	**Gini Index**	**Entropy function**	**Exponent-based function**	**Proposed M.D-based index**
Normal with parameters (0,1)	**10**	0.5567	0.5200	0.4433	0.5233	0.4927	0.4800	0.4700	0.4733
	20	0.5067	0.5067	0.5017	0.5333	0.4883	0.4783	0.5283	0.5350
	50	0.4947	0.4887	0.5087	0.4653	0.5027	0.4987	0.4840	0.5140
	100	0.5007	0.5017	0.4983	0.4903	0.5033	0.4903	0.4873	0.4927
	500	**0.4902**	**0.4997**	0.5060	0.5025	**0.4979**	**0.4907**	0.4955	0.4998
Exponential with parameter (1)	**10**	0.3333	0.3533	0.4000	0.3700	0.3900	0.4300	0.4133	0.3500
	20	0.3800	0.4167	0.4013	0.4200	0.3850	0.3600	0.4400	0.3930
	50	0.3600	0.4207	0.3940	0.3587	0.3993	0.3944	0.4300	0.4067
	100	0.3993	0.3800	0.3867	0.3937	0.4240	0.3953	0.4000	0.4167
	500	0.3842	**0.3820**	**0.3795**	0.3884	0.4140	**0.4104**	**0.4024**	0.4283
Uniform with parameters (-1,+1)	**10**	0.5100	0.4933	0.4367	0.4833	0.4933	0.4800	0.4833	0.5133
	20	0.5050	0.4867	0.4733	0.4800	0.5350	0.4733	0.4717	0.5117
	50	0.5007	0.4947	0.5213	0.5113	0.5247	0.4927	0.4600	0.5140
	100	0.5110	0.5049	0.5007	0.5040	0.5062	0.5124	0.4890	0.4952
	500	0.5057	**0.4890**	**0.5027**	0.5045	0.5123	**0.4940**	**0.4817**	0.4987

TABLE II

SIMULATION RESULTS OF BALANCED DATA GENERATED THROUGH DISCRETE DISTRIBUTIONS

Discrete Distribution	***Number of iterations***	***Misclassification Rate (03 Variables)***				***Misclassification Rate (04 Variables)***			
		Gini Index	***Entropy function***	***Exponent-based function***	***Proposed M.D-based index***	***Gini Index***	***Entropy function***	***Exponent-based function***	***Proposed M.D-based index***
Binomial with parameters (1,0.5)	***10***	0.5000	0.4767	0.5267	0.5233	0.5300	0.4267	0.5067	0.5133
	20	0.5117	0.5150	0.4717	0.5250	0.5017	0.4850	0.4950	0.4900
	50	0.4887	0.5013	0.4987	0.4887	0.4947	0.5073	0.5087	0.5040
	100	0.5003	0.5087	0.4860	0.5057	0.4876	0.4967	0.4952	0.4933
	500	**0.4939**	**0.4944**	0.4991	0.4990	**0.4950**	**0.4983**	0.5043	0.5133
Poisson with parameter (10)	***10***	0.4867	0.4833	0.4667	0.4667	0.4833	0.4867	0.4533	0.5133
	20	0.4983	0.4867	0.4850	0.4817	0.4600	0.4850	0.4900	0.4767
	50	0.4953	0.4727	0.4807	0.4660	0.4627	0.4773	0.4787	0.4753
	100	0.4847	0.5007	0.4773	0.4863	0.4850	0.4723	0.4733	0.4777
	500	**0.4680**	0.4858	**0.4747**	0.4828	**0.4832**	0.4862	**0.4802**	0.4871
Hypergeometric with parameters (1,10,5)	***10***	0.5067	0.4667	0.4967	0.4567	0.4533	0.4933	0.4933	0.5100
	20	0.4433	0.5117	0.4767	0.4483	0.4517	0.4950	0.5217	0.4867
	50	0.4880	0.4773	0.4700	0.4853	0.4827	0.4927	0.4927	0.4833
	100	0.4827	0.4983	0.4737	0.4747	0.4670	0.4890	0.4760	0.4887
	500	0.4785	0.4857	**0.4783**	**0.4782**	0.4795	0.4793	**0.4774**	**0.4769**

In Table I and Table II, comparison of Misclassification rates is depicted and results revealed that in case of balanced and data having: (i) Normal distribution structure, Gini index and Entropy function produces less Misclassification rates; (ii) Exponential distribution structure, Entropy function and Exponent-based index produces less Misclassification rates; (iii) Uniform distribution structure, Entropy function and Exponent-based index produces less Misclassification rates; (iv) Binomial distribution structure, Gini index and Entropy function produces less Misclassification rates; (v) Poisson distribution structure, Gini index and Exponent-based index produces less Misclassification rates; (vi) Hypergeometric distribution structure, Exponent-based index and new proposed M.D-based index produce less Misclassification rates.

TABLE III

SIMULATION RESULTS OF IMBALANCED DATA GENERATED THROUGH CONTINUOUS DISTRIBUTIONS

Continuous Distribution	***Number of iterations***	***Misclassification Rate (03 Variables)***				***Misclassification Rate (04 Variables)***			
		Gini index	***Entropy function***	***Exponent-based function***	***Proposed M.D-based index***	***Gini index***	***Entropy function***	***Exponent-based function***	***Proposed M.D-based index***
Normal with parameters (0,1)	***10***	0.1467	0.1533	0.1667	0.1600	0.2330	0.2000	0.2200	0.1733
	20	0.1517	0.1717	0.1383	0.1817	0.1300	0.1933	0.1833	0.1617
	50	0.1460	0.1747	0.1480	0.1587	0.1520	0.1900	0.1633	0.1600
	100	0.1717	0.1687	0.1493	0.1600	0.1633	0.2300	0.1847	0.1640
	500	**0.1573**	**0.1544**	0.1608	0.1603	**0.1697**	**0.1667**	0.1879	0.1690
Exponential With parameter (1)	***10***	0.1733	0.1233	0.1167	0.1333	0.1900	0.2330	0.1567	0.1333
	20	0.1250	0.1367	0.1417	0.1333	0.1233	0.2333	0.1150	0.1417
	50	0.1367	0.1313	0.1340	0.1400	0.1420	0.2733	0.1380	0.1460
	100	0.1396	0.1276	0.1382	0.1423	0.1473	0.2489	0.1503	0.1373
	500	0.1367	0.1331	**0.1329**	**0.1297**	0.1495	0.2000	**0.1447**	**0.1487**
Uniform with parameters (-1,+1)	***10***	0.1133	0.1067	0.1100	0.1233	0.1333	0.2433	0.1367	0.1233
	20	0.1117	0.1083	0.1117	0.1533	0.1267	0.2333	0.1200	0.1270
	50	0.1300	0.1120	0.1133	0.1260	0.1207	0.2560	0.1313	0.1480
	100	0.1330	0.1310	0.1137	0.1260	0.1320	0.2867	0.1320	0.1338
	500	**0.1162**	0.1251	**0.1198**	0.1306	**0.1256**	0.2650	**0.1299**	0.1302

TABLE IV

SIMULATION RESULTS OF IMBALANCED DATA GENERATED THROUGH DISCRETE DISTRIBUTIONS

Discrete Distribution	***Number of iterations***	***Misclassification Rate (03Variables)***				***Misclassification Rate (04 Variables)***			
		Gini index	***Entropy function***	***Exponent-based function***	***Proposed M.D-based index***	***Gini index***	***Entropy function***	***Exponent-based function***	***Proposed M.D-based index***
Binomial with parameters (1,0.5)	***10***	0.0967	0.0833	0.1067	0.1100	0.1833	0.2400	0.1293	0.0867
	20	0.1150	0.0917	0.1183	0.0867	0.1133	0.2433	0.0967	0.0883
	50	0.1033	0.0960	0.1053	0.0973	0.1073	0.2333	0.0886	0.1167
	100	0.1000	0.1130	0.1003	0.1023	0.1073	0.2633	0.1233	0.0907
	500	0.1043	0.1033	**0.0988**	**0.1015**	0.1352	0.2067	**0.1033**	**0.1092**
Poisson with parameter (10)	***10***	0.0533	0.0800	0.0700	0.0567	0.0733	0.2033	0.0567	0.0800
	20	0.0750	0.0967	0.0933	0.0867	0.1117	0.2317	0.0883	0.1000
	50	0.0853	0.0773	0.0866	0.0700	0.0833	0.2238	0.0993	0.0933
	100	0.0710	0.0907	0.0907	0.0887	0.0786	0.2400	0.0827	0.1010
	500	0.0847	0.0917	**0.0840**	**0.0814**	0.1137	0.2233	**0.0883**	**0.0873**
Hypergeometric with parameters (1,10,5)	***10***	0.5400	0.4733	0.5400	0.4916	0.4467	0.4633	0.4833	0.4833
	20	0.5033	0.5067	0.4700	0.4900	0.4767	0.4850	0.4917	0.5083
	50	0.4753	0.4820	0.4880	0.5007	0.4827	0.4707	0.4933	0.4907
	100	0.4769	0.4947	0.4820	0.4877	0.4823	0.4993	0.4957	0.4783
	500	0.4890	**0.4849**	0.4878	**0.4800**	0.4908	**0.4867**	0.4959	**0.4883**

In Table III and Table IV, comparison of Misclassification rates for imbalanced data is depicted and results revealed that data having: (i) Normal distribution structure, Gini index and Entropy function produces less Misclassification rates; (ii) Exponential distribution structure, Exponent-based index and new proposed M.D-based index produces less Misclassification rates; (iii) Uniform distribution structure, Gini index and Exponent-based index produces less Misclassification rates; (iv) Binomial distribution structure, Exponent-based index and new proposed M.D-based index produces less Misclassification rates; (v) Poisson distribution structure, Exponent-based index and new proposed M.D-based index produces less Misclassification rates; (vi) Hypergeometric distribution structure, Entropy function and new proposed M.D-based index produce less Misclassification rates.

V. CONCLUSIONS

In this study, a new Mean Deviation based index has been proposed and the properties of an impurity measure for proposed criterion have been verified. Also, a simulation study is conducted to explore, which measure gives best result in what type of situation? After comparing the Misclassification Rates obtained from simulation study, it can be concluded that no one splitting criterion can perform best in every situation, but

generally new proposed M.D-based index and Exponent-based index give excellent results in case of imbalanced data. While, in case of balanced data, Gini index and Entropy function produces less misclassification rates.

REFERENCES

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

[2] S. R. Safavian, and D. Landgrebe, "A survey of decision tree classifier methodology", *IEEE Trans. Systems, Man Cybernet*, vol. 21, Issue 3, pp. 660–674, 1991.

[3] J. N. Morgan, and J.A. Sonquist, "Problems in the analysis of survey data, and a proposal", *Journal of the American Statistical Association*, vol. 58, pp. 415–435, 1963.

[4] J. N. Morgan, and R. C. Messenger, "THAID: A sequential analysis program for the analysis of nominal scale dependent variables", Institute for Social Research, University of Michigan, Ann Arbor Tech. Rep., 1973.

[5] C. Gini, *Variability and Mutability*, Cuppini, Bologna, Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T), Rome: Libreria Eridi Virgilio Veschi 1, 1912.

[6] A. P. Bremner, "Localised Splitting Criterion for Classification and Regression Trees", Ph.D thesis, Murdoch University, 2004.

[7] C. .E. Shannon, "A mathematical theory of communication", *Bell System Technical Journal*, vol. 27, issue 3, pp. 379–423, 1948.

[8] R. Quinlan, "Discovering rules by induction from large collections of examples", Expert Systems in the Micro-electronic Age, Edinburgh University Press, pp. 168-201, 1979.

[9] M. Azam, Q. Zaman, and K. P. Pfeiffer, "Improved classification trees with two or more classes", *in Proc. 9th Islamic Countries Conference on Statistical Sciences*, 2007.

