A Novel Statistical Theoretical Split Metric for Decision Tree Classification

Mainak Biswas

School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneshwar-751024, India

Corresponding author: Mainak Biswas, Email: mainakmani@gmail.com

Decision trees (DTs) are a significant category of logical tools in machine learning (ML), used to classify both text and numerical data. Over the years, two primary criteria for splitting DTs have been prevalent: information gain, which hinges on Shannon's entropy, and the Gini index. Both these criteria rely on the empirical probabilities of classes within the attribute space of the dataset. In this study, a novel split criteria is introduced, rooted in the principles of statistical mechanics. This measure draws inspiration from the second law of Thermodynamics, which stipulates that in a closed system with unchanging external conditions and entropy, the internal energy of the system will decrease and reach a minimum at equilibrium. This novel split criterion was tested on four datasets, each containing at least 100 instances. The results demonstrated a comprehensive enhancement in accuracy, precision, recall, and F1-score.

Keywords: Statistical mechanics, splitting criteria, decision tree

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 65–80. DOI: https://doi.org/10.56155/978-81-955020-7-3-6

1 Introduction

The study of machine learning (ML) [1] entails understanding of algorithms designed in the context of improving learning of machines with experience. ML algorithms are further categorized into two distinct groups of learning: supervised and unsupervised. The supervised learning are designed on the premise that machines trained on labeled instances would be able to predict labels of unknown ones. Generally a part of the dataset is used for training and other part for testing the ML model. The performance parameters of testing is used for evaluating the particular ML model. This supervised approach is also known as classification if the output is labels and regression if the output is numeric. Whereas, unsupervised learning algorithms are designed around the concept of learning complex pattern within unknown instances. In classification, the input is a group of instances, also called as instance space $\mathbf{X}$ and their labels, also known as output space, $\mathbf{Y}$. The input space is defined by the number of instances in it $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ and it's output space is defined by the number of labels $\mathbf{Y} = \{y_1, y_2, \cdots, y_m\}$, where $m \leq n$. The relationship from input-to-output space is many-to-one, where, more than one instances can belong to a single label. Each i^{th} instance x_i is further defined by a feature set $x_i = \{f_{i1}, f_{i2}, \cdots, f_{ij}\}$. The classification, therefore, entails finding the best approximation ($F'(\mathbf{X})$) of the true labeling function which maps $\mathbf{X}$ to $\mathbf{Y}$. The approximation function is given by $F'(\mathbf{X}) : \mathbf{X} \rightarrow \mathbf{Y}$ where, $F(\mathbf{X})$ is the true labeling function. Therefore, classification involves finding the best approximate function which will resemble the true one. In this paper, the main focus is on Decision Tree (DT) [2] which is one of the important supervised learning algorithms for classification. In order to deduce the class/ label of a particular instance, a path is followed from the root to the leaf of a DT. The leaf represents the class, whereas each internal node including the root represents a distinct feature. Each edge of the DT represents an action on the node or feature. Therefore,given an unknown instance, a test is carried out on its features from the root until it reaches a particular leaf representing its label as shown in Figure 1. Each node in a decision tree is selected from features based on the concept of homogeneity. The feature which achieves the highest homogeneity for the instances is selected as the best suited for splitting. The training dataset is therefore divided based on best feature selected through the splitting measure. The best feature becomes root node. This process is applied recursively on the divided sets until absolute homogeneity is achieved. The splitting measure computes whether the division makes the instances in the dataset homogeneous to be assigned a label. The most well known splitting measures are Information Gain [3] and Gini Index [4]. All these measures are probabilistic and are derived from finding ratio of number of distinct particles to all the particles in the universe of discourse. In this work, a new splitting criteria is proposed which is based on statistical mechanics. In ML, a huge number of instances are dealt within a data corpus [5] [6]. Similarly, statistical mechanics deals with the behavior of large number of particles in confined area [7] [8]. In statistical mechanics, it is stated that, for a confined area, under constant temperature and

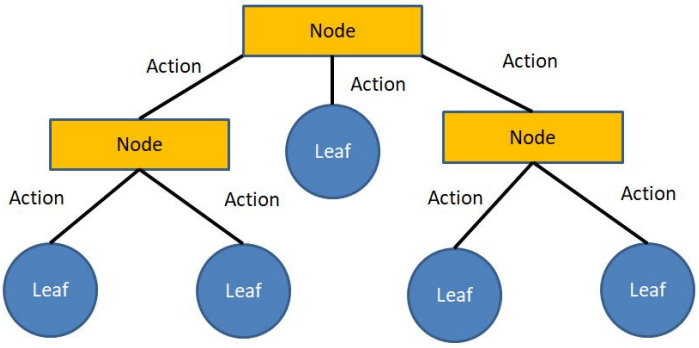

Figure 1: Decision tree conceptual diagram.

pressure, the internal energy of a system will approach a minimum value. In ML-DT, we can apply the same principle of maximum energy decay for best feature selection. The energy for a feature in here is directly proportional to the total number of classes under consideration, and inversely proportional to instances belonging to these classes. If feature values are unique, then individual energy is computed for each and then summed to find the total energy. The results clearly show that the current split measure performs better than contemporary splitting criteria.

The paper is further divided as follows: related works are discussed in section 2. The analogy between instances and particles is discussed in section 3 and the energy-based splitting criteria is discussed in section 4, results in section 5, discussion in section 6 and finally conclusion in the last section 7.

2 Related Works

Two of the most popular split functions: Information Gain (IG) and Gini Index (GI) are discussed here.

2.1 Information Gain

Shannon's entropy is the best example of application of statistical mechanics principle in information technology, networking and AI. The entropy refers here to amount of information contained in a message. Mathematically, it computes the number of bits required to represent the information by summing the proportions of bits required to represent the message. It is given by:

$$H(x) = -\sum p_i \log_2 p_i \tag{6.1}$$

where, p_i represents the probability of occurrence of an unique event. When applied as a split function to choose the best features for the decision tree, it computes the information gain. It does so, by computing the entropy of the class variable. Thereafter, for each feature, it computes the entropy of unique feature elements. The weighted average entropy of all unique feature elements are computed, which becomes the feature entropy. The IG for the feature is computed by subtracting the feature entropy from the class entropy. It is given by:

$$IG(D, f) = H(D) - H(D, f) \tag{6.2}$$

where, D is the dataset and f is the given feature. The IG is computed for all the features. The feature with the highest IG becomes the split feature.for the given dataset. The highest IG represents the greatest reduction in entropy for a given feature. The highest IG feature also states that it has the highest homogeneity among all others.

2.2 Gini Index

The Gini Index or GI is the estimation of the impurity when the features of instances belong to a single class. Mathematically, GI is computed as:

$$GI = 1 - \sum p_i^2 \tag{6.3}$$

which is subtracting the summation of squared probabilities of the classes from one. It varies between zero and one. The zero represents purity of classes while one represents random distribution of instances in classes.

In the next section we will look into the analogy of features in a dataset and molecules in a closed system.

3 Instances and Particles: An Analogy

Machine learning like statistical mechanics deals with probabilities of distributions of particles in the given universe of discourse. Therefore, principles of statistical mechanics can be applied in the case of distributions of instances within the discourse. In this work, we formulate a new splitting criteria based on second law of thermodynamics which states that under constant external parameters, the internal energy of a closed system will decrease and approach a minimum value at equilibrium [7]. We apply the same criteria to the features in the decision tree model, to select the best feature for splitting based on highest energy decay. We will use the given terminologies interchangeably throughout the whole text.

1. *Particles* in closed system are equivalent to *instances* in dataset

2. *Energy levels* in closed system are same as energy of *attributes/features* in dataset

3. *Degenerate states* in closed system and *classes or labels* in dataset are equivalent

Supposedly, there are M particles in a closed system. Each of these M particles or instances belong to one of the n energy levels. Therefore, the total such arrangements are $\frac{M!}{M_1!,M_2!,\cdots,M_n!}$. Within each energy levels there can be multiple degenerate states Q having same energy level. In here degenerency corresponds to different measurable states of a quantum system [1]. Therefore, the total number of arrangements become One such energy level "i" is shown in Figure 2. When equated with a given data corpus, the arrangement of instances in several classes for a feature i is shown in Figure 3.

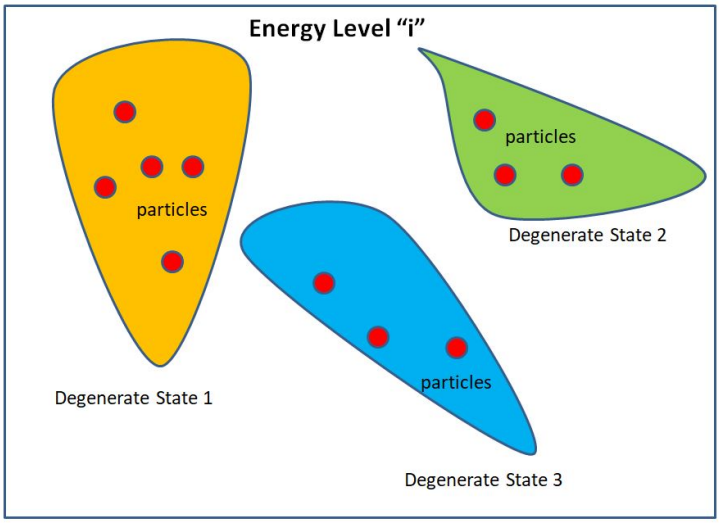

Figure 2: Distribution of particles in several degenerate states within the i^{th} energy level.

When, repeated for all energy levels, the total possible combinations is given by:

$$\delta = \frac{M!}{\prod_{i=1}^{n} M_i!} \cdot \prod_{i=1}^{n} Q_i^{M_i} \tag{6.4}$$

where, δ denotes the function representing the possibilities and M_i denotes number of particles in the i^{th} energy level.

The feature energy level must be deduced when it reaches equilibrium to find the best splitting feature. The equilibrium state is equivalent to the most probable situation given

[1] tinyurl.com/7qt9dcu

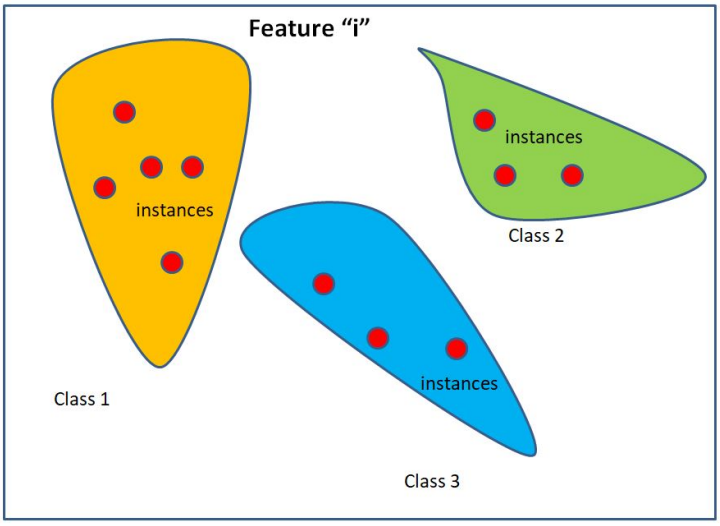

Figure 3: Distribution of instances in several classes for the i^{th} feature.

by Equation 6.4. There are two laws that must be applied. The first law states that the total number of particles must be conserved which is given by:

$$\sigma = \sum M_i = M(constant) \tag{6.5}$$

The second law states that the total energy of the system must be conserved. This is given by:

$$\theta = \sum E_i.M_i = U(constant) \tag{6.6}$$

where, E_i represents i^{th} energy level/feature. Next logarithm principle is applied to Equation 6.4, which is given by:

$$\log(\delta) = \log M! - \sum^{n} M_i! + \log \sum_{i=1}^{n} Q_i.M_i \tag{6.7}$$

Next Stirling's Approximation[2] is applied, output of which is given by:

$$\log(\delta) = M \log M - M - \sum_{i=1}^{n} M_i log M_i + \sum_{i=1}^{n} M_i + \log \sum_{i=1}^{n} Q_i.M_i \tag{6.8}$$

At this juntion, Lagrange's multiplier is applied to the Equation 6.8 using laws shown in Equations 6.5 and 6.6. The new parameters α and β are introduced for implementing Lagrange's multiplier which is given by:

$$\log(\delta) + \alpha\sigma - \beta\theta = 0 \tag{6.9}$$

Differentiating Equation 6.9 with respect to M_j, :

$$\frac{d\{\log(\delta) + \alpha\sigma - \beta\theta = 0\}}{dM_j} = 0 \tag{6.10}$$

After expansion of Equation 6.10 we get :
$d\{\log(M \log M - M - \sum_{i=1}^{n} M_i \log M_i + \sum_{i=1}^{n} M_i + \log \sum_{i=1}^{n} Q_i.M_i) + \alpha(\sum M_i) - \beta(\sum E_i.M_i)\}/dM_j = 0$
Since the total number of particles M is constant and the only terms which are non-zero when $i = j$ are:

$$\log Q_j - \log M_j + \alpha - \beta E_j = 0 \tag{6.11}$$

The value of $\alpha - \beta E_j$ is derived from Equation 6.11 :

$$\alpha - \beta E_j = \log \frac{M_j}{Q_j} \tag{6.12}$$

Therefore, the energy of a given feature j is given by:

$$E_j = \frac{\alpha - \log \frac{M_j}{Q_j}}{\beta} \tag{6.13}$$

or

$$E_j = \frac{\alpha + \log \frac{Q_j}{M_j}}{\beta} \tag{6.14}$$

[2] $logx! = xlogx - x$

Now, the proposed splitting criteria is discussed formally in the next section.

4 Proposed Energy-based Splitting Criteria

It is learned from Equation 6.14 that the energy level is directly proportional to degenerency of the particular level and inversely proportional to the number of particles in that level. This relationship forms the basis of the proposed splitting criteria where, features are represented as different energy levels, instances as particles and classes as degenerency states. The feature with the minimum energy is selected for splitting in the DT. The constant α is directly proportional to chemical potential and β is related to the number of instances under consideration. Since the proposed measure is inspired, we have only considered $\beta = \frac{1}{N}$ The proposed measure is therefore given as:

$$E_f = \frac{1}{\beta} \log \frac{Q_j}{M_j} \tag{6.15}$$

where,

$$Q_j = \Sigma(L) \tag{6.16}$$

and

$$M_j = \prod_{k=1}^{Q_j} I_k \tag{6.17}$$

Here, Q_j in represents the total number of labels in the context of feature j. As degenerencies are multiplicative, M_j in Equation 6.17 represents product of instances belonging to different labels/degenerate states. Alternately, Equation 6.15 can be also written as:

$$E_f = \frac{1}{\beta} \log \frac{\Sigma(L)}{\prod_{k=1}^{\Sigma(L)} I_k} \tag{6.18}$$

If there are distinct sub-features ($f = \{sf^1, sf^2, \cdots, sf^t\}$) within a feature, each energy levels of sub-features are computed and summed.

$$E_f = \Sigma_{i=1}^{t} E_{sf^i} \tag{6.19}$$

Since we have considered quantum system [8], negative energy values are accepted, and minimum energy is the greatest negative value.

5 Results

This section is divided into three sub-sections: datasets, performance parameters and results.

5.1 Datasets

We have considered four categorical datasets for the experiment. All the datasets are public and consists at least 100 instances. In this work we have used K5 cross-validation (80% training and 20% testing). The datasets are given as follows:

1. Zoo dataset [9] The dataset consisted of 101 instances, with 17 attributes, and seven classes of animals such as amphibians, insects etc. One attribute animal name was dropped,for better generalization.

2. Car evaluation dataset [10] The dataset consisted of 1728 instances, six attributes and four classes of evaluation from unacceptable to very good.

3. Breast Cancer Wisconsin Data Set [11] The dataset consisted of 569 instances, 32 attributes and two classes: benign and malignant.

4. Hayes Roth dataset [12] The dataset consists 160 instances, five attributes and three classes on categorization of human subjects.

5.2 Performance Parameters

1. Accuracy: It is defined as the percentage of correct predictions to the total number of predictions.

2. Precision is the fraction of relevant instances among the retrieved instances

3. Recall in this context is also referred to as the true positive rate or sensitivity

4. Height: The height is defined as the longest path from the root node to the leaf in a decision tree.

5.3 Performance

The accuracy, precision, recall, f1-score and height of the decision trees using proposed statistical split measure, entropy and Gini-Index are provided in Tables 1, 2, 3, 4 and 5. It is clearly seen that for accuracy, the proposed measure outperforms all other measures and equal for Gini-index in Hayes-Roth dataset. The accuracy for zoo-dataset, Breast, Hayes-Roth and Car evaluation dataset is 90%, 95%, 81%,and 77%, respectively. For precision, the proposed measure is better than all (Table 2). The precision values for zoo-dataset,

Breast, Hayes-Roth and Car evaluation dataset are 0.88, 0.98, 0.87,and 0.82, respectively. For recall (Table 3), the proposed measure is better than or equal (IG and proposed measure recall values is same for Breast, Gini and proposed measure recall values are same Hayes-Roth)to all contemporary metrics. The results of recall for zoo-dataset, Breast, Hayes-Roth and Car evaluation dataset are 0.90, 0.96, 0.81 and 0.77, respectively. For F1-score (Table 4) except Gini-index for Hayes-Roth dataset (0.81), the proposed measure outperforms IG and Gini: 0.88 for zoo, 0.97 for Breast, 0.79 for Car evaluation. In Height Table 5, proposed is equal to IG and Gini-index for Hayes-Roth (8) and Car Evaluation (12). However, for zoo (6) and breast dataset (6), IG is better. The corresponding accuracy, precision, recall, f1-score and height table are provided in Figures 4, 5, 6, 7 and 8, respectively.

Table 1: Accuracy (%) table.

Sl	Datasets/Models	Information Gain	Gini-Index	Proposed
01	Zoo dataset	85.00	50.00	**90.00**
02	Breast	95.68	83.45	**95.68**
03	Hayes-Roth	65.38	**80.77**	**80.77**
04	Car Evaluation	59.42	10.43	**76.52**

Table 2: Precision table.

Sl	Datasets/Models	Information Gain	Gini-Index	Proposed
01	Zoo dataset	0.8250	0.5333	**0.8750**
02	Breast	0.9779	0.9619	**0.9779**
03	Hayes-Roth	0.6538	0.8192	**0.8678**
04	Car Evaluation	0.5370	0.3115	**0.8181**

Table 3: Recall table.

Sl	Datasets/Models	Information Gain	Gini-Index	Proposed
01	Zoo dataset	0.8500	0.5000	**0.9000**
02	Breast	**0.9568**	0.8345	**0.9568**
03	Hayes-Roth	0.6538	**0.8077**	**0.8077**
04	Car Evaluation	0.5942	0.1043	**0.7652**

Table 4: F1-score table.

Sl	Datasets/Models	Information Gain	Gini-Index	Proposed
01	Zoo dataset	0.8333	0.5100	**0.8833**
02	Breast	0.9672	0.8915	**0.9672**
03	Hayes-Roth	0.6407	**0.8089**	0.8057
04	Car Evaluation	0.5622	0.1563	**0.7900**

Table 5: Height table.

Sl	Datasets/Models	Information Gain	Gini-Index	Proposed
01	Zoo dataset	**6**	28	20
02	Breast	**6**	18	10
03	Hayes-Roth	**8**	**8**	**8**
04	Car Evaluation	**12**	**12**	**12**

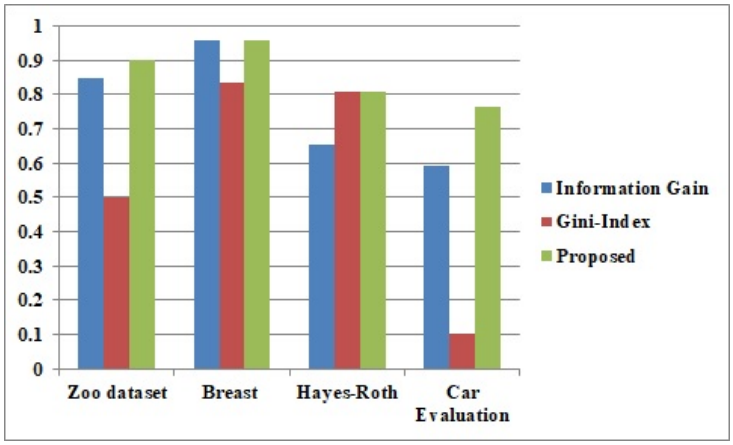

Figure 4: Accuracy bar-chart for proposed, entropy, and Gini split criteria using different datasets.

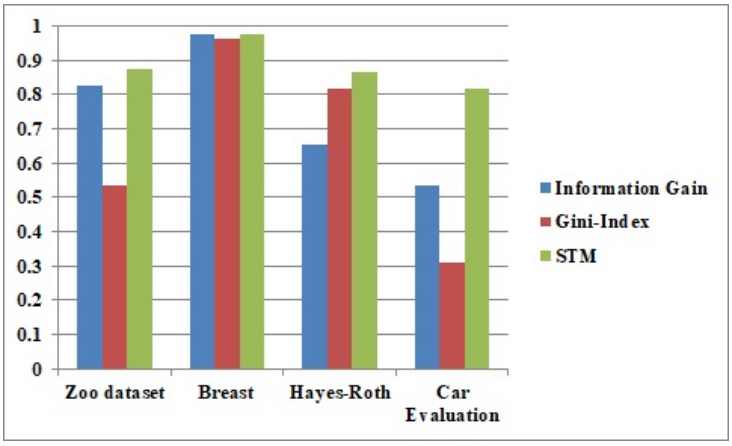

Figure 5: Precision bar-chart for proposed, entropy, and Gini split criteria using different datasets.

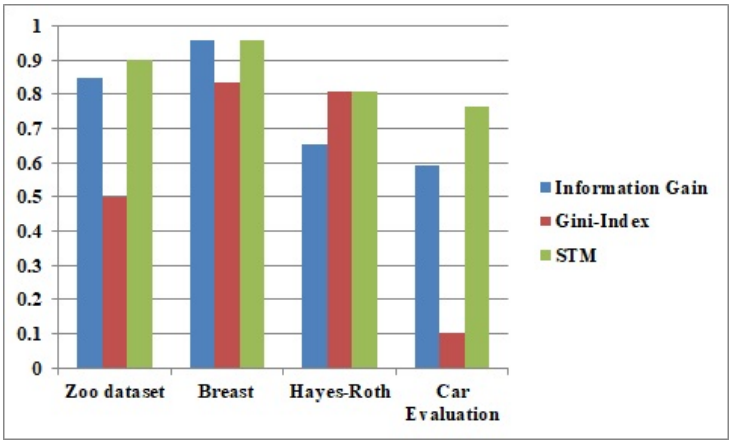

Figure 6: Recall bar-chart for proposed, entropy, and Gini split criteria using different datasets.

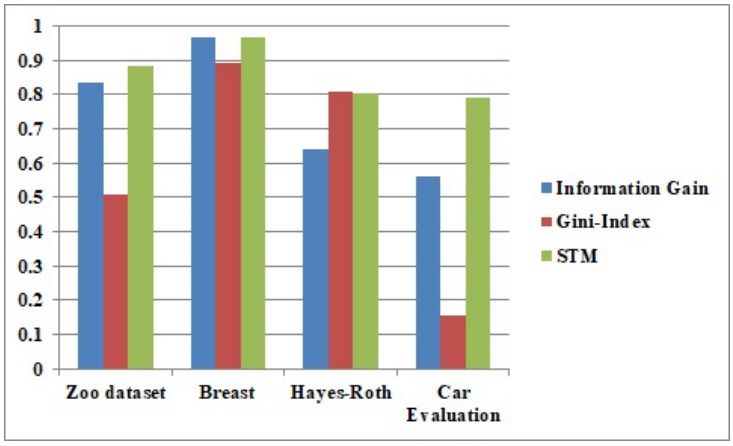

Figure 7: F1 score bar-chart for proposed, entropy, and Gini split criteria using different datasets.

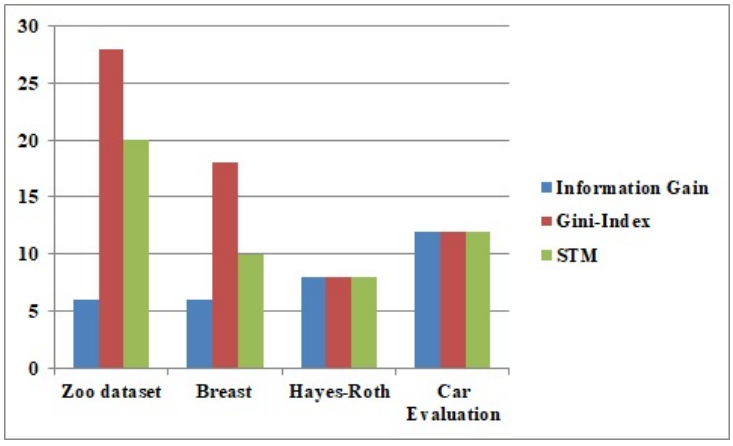

Figure 8: Decision tree height for proposed, entropy, and Gini split criteria using different datasets.

6 Discussion

In this paper, we have provided a novel split criteria based on statistical mechanics. The model considers dataset as closed system, instances as particles, attributes as energy levels and degenerate states as classes. The model computes energy levels of different attributes for the most plausible distribution of instances in the dataset. The model assumes the best feature as the one with the minimum energy in equilibrium state. Once all energy levels are computed the feature with minimum value is considered the best attribute for split. In the future works, it is proposed to apply in different applications such as page ranking [13], communication theory [3], , analysing epidemic datasets i.e., COVID-19 [14], image text classification [15], biomedical applications [16] [17], price monitoring [18], sentiment classification [19], text comparison [20] etc.

7 Conclusion

The paper presented a novel split criteria for decision trees for selecting the best attribute for splitting. The split criteria is derived from statistical mechanics. It is based on the idea of the second law of thermodynamics which states that at equilibrium the internal energy will decrease and approach a minimum value. This energy level is equal to the log of ratio of sum of classes and product of number of instances belonging to these classes. The proposed split criteria provides better performance with respect to accuracy, precision, recall and f1-score for different datasets. In the future, it is proposed to apply the criteria for classification in numerical datasets and regression.

Acknowledgement

The author would like to thank his father Shree Madhabendu Biswas and mother Shreemati Nipuna Biswas for being the pillar of inspiration and confidence.

References

[1] Peter F. *Machine learning: the art and science of algorithms that make sense of data*. Cambridge university press, 2012.

[2] Quinlan JR. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72, 1996.

[3] Shannon CE. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[4] Gini C. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, 208(1):73–79, 1936.

[5] Kuppili V., Biswas M, et al. A mechanics-based similarity measure for text classification in machine learning paradigm. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(2):180–200, 2018.

[6] Biswas M. A novel statistical mechanics-based metric for characterization of text documents. *COMPUTERS and ELECTRICAL ENGINEERING*, 87, 2021.

[7] Callen HB. Thermodynamics and an introduction to thermostatistics, 1998.

[8] Everett A., Thomas Roman T. *Time travel and warp drives: a scientific guide to shortcuts through time and space*. University of Chicago Press, 2012.

[9] Ray R., Dash SR. Comparative study of the ensemble learning methods for classification of animals in the zoo. In *Smart Intelligent Computing and Applications*, pages 251–260. Springer, 2020.

[10] Awwalu J. Performance comparison of data mining algorithms: A case study on car evaluation dataset. *Int. J. Comput. Trends Technol*, 13(2):78–82, 2014.

[11] Borges LR. Analysis of the wisconsin breast cancer dataset and machine learning for breast cancer detection. *Group*, 1(369):15–19, 1989.

[12] Nababan AA., Sitompul OS., et al. Attribute weighting based k-nearest neighbor using gain ratio. In *Journal of Physics: Conference Series*, volume 1007, page 012007. IOP Publishing, 2018.

in: *Advancements in Communication and Systems*. Ed. by Ashish Kumar Tripathi and Vivek Shrivastava. Computing and Intelligent Systems, SCRS, India., 2023, pp. 65–80. DOI: https://doi.org/10.56155/978-81-955020-7-3-6

[13] Bhagawati R., Subramanian T. An approach of a quantum-inspired document ranking algorithm by using feature selection methodology. *International Journal of Information Technology*, pages 1–13, 2023.

[14] Gupta S, Sharaff A., Nagwani NK. An efficient hybrid textual similarity technique for analyzing covid-19 dataset. 2022.

[15] Kumar TP, Florence ML, and Fathima G. A similarity measure for classification of text extracted from image using machine learning. In *2022 6th International Conference on Computing Methodologies and Communication (ICCMC)*, pages 964–968. IEEE, 2022.

[16] Kuppili V., Biswas M. et al,. Extreme learning machine framework for risk stratification of fatty liver disease using ultrasound tissue characterization. *Journal of medical systems*, 41:1–20, 2017.

[17] El-Mahelawi JK. et al., Bassem S Abu-Nasser, and Samy S Abu-Naser. Tumor classification using artificial neural networks. *International Journal of Academic Engineering Research (IJAER)*, 4(11), 2020.

[18] Nurcahyawati V., Mustaffa Z.. Online media as a price monitor: Text analysis using text extraction technique and jaro-winkler similarity algorithm. In *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pages 1–6. IEEE, 2020.

[19] Ramya P, Karthik B. Word sense disambiguation based sentiment classification using linear kernel learning scheme. *Intelligent Automation & Soft Computing*, 36(2), 2023.

[20] Xiao P. et al., Fast text comparison based on elasticsearch and dynamic programming. In *International Conference on Web Information Systems Engineering*, pages 50–64. Springer, 2023.