International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Best Classification of Continuous Data Based on Hybrid Decision Tree

Noor. Fadel*[1], Iman K. Abbood[2], Hadeel Qasem Gheni[3]

Submitted: 06/06/2022 **Accepted**: 10/09/2022

Abstract: The correct classification of continuous data and finding classification algorithms with high accuracy to classify previously invisible records is one of the most important real problems facing researchers in the computing world, especially with continuous data, which always suffers from classification errors due to the slitting of the data into periods or discretization. In this paper, we present a hybrid algorithm that uses the principle of a decision tree, excluding the concepts of entropy and information gain in order to avoid splitting continuous data and replacing them to form the period based on concepts like the counters In the training process, the periods of continuous data will be self-dividing, meaning that the periods of continuous data are self-forming, which takes into account the composition of the period or split data based on the first and last element of the larger counter. And so on, right up to the counter with thefewest number of data points. The experimental result gave a very good rating of 95.56%, which means better results in the handling and classification of continuous data.

Keywords: *Classification, Continuous Data, DBSCAN, Hybrid Decision Tree.*

1. Introduction

In general, the correct and accurate classification of data is a real problem, especially with numerical and continuous data. The need for reliable classification algorithms has become urgent,especially with the huge increase in data preparation, which is themain concern of most researchers. Since the classification processis not limited to a specific category and field but involves interference in almost all fields, such as the medical classificationof tumors, whether benign or malignant or the field of industries, such as product classification as defective or non-defective, etc. [1].

It is worth mentioning that images of all kinds have entered the field of classification and can also get useful data by methods of image processing and configuration of a database.

Finding the best split for continuous data is computationally expensive and takes O(N2). Thus, a candidate value is determined and the number of the larger and smaller classes is counted from the candidate value after the data is sorted. This process takes O(N log N) [2].

This process minifies the accuracy of the correct classification, and the loss of information increases the rate of error classification. Therefore, the rating ratio was low compared to theproposed method.

One common method of dealing with continuous data isconverting it to intervals by using an estimate of the amount of data and the interval. The whole process is called (Discretization) [3].

Because of the combination, data from two different categories in the same period were treated by a single decision. So the main disadvantage of converting continuous data into intervals or many periods is the loss of information in the original data. For example, both normal and abnormal objects in the same periodare treated by a single decision [4]. As shown in Fig .1.

The discretization aims to reduce the number of continuous feature values by aggregating them into several time slots. Two major problems with discretization are determined by the number of slots and the width of those slots [5].

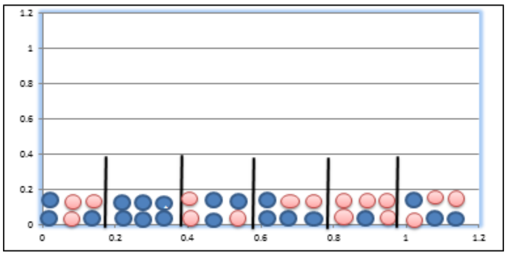

Figure 1. Discretization, blue circle is normal objects, a pink circle is abnormal objects.

Define A decision tree algorithm is one of the most popular and important machine learning algorithms. It can deal with large amounts of data and is a robust data analysis form that can predict, classify, and describe a target [4].

The decision tree model is a flowchart with two or more branches, where each branch represents a situation or condition while the node represents a feature or attribute. The number of tree levels depends on the number of classified data. The importance of the classification process is the ability to classify a large amount of data that was not visible before [5].

The traditional way of determining the amount of data classified at each level is to calculate the entropy and the information gain. After splitting the data into periods, it then computes the entropy for each period, where the value of the entropy (low random- purest period) is better than the maximum entropy. The next step

[1&2] *Information Technology,, University of Babylon, Iraq*
ORCID ID : 0000-0003-2926-5216
ORCID ID : 0000-0002-7439-6752
[3]*College of Science for Women - Department of Computer Science, Iraq*
ORCID ID : 0000-0002-8875-3677
** Corresponding Author Email: noorfadel75@email.com*

calculates the information gain, and the greater value of information gain means the purer the data than if the value of information gain is low [6].

After testing the features, the first root in the tree (parent) of the features has the lowest randomness, i.e. the lowest entropy. If the data meets the conditions, it is classified as a leaf node. If the condition is not met, the condition turns to the second level to a new node and condition, which is called the children. The tree stops when all the nodes have a leaf [7].

Valuable information can be obtained from images by applying image processing methods such as preprocessing, which include filtering the image from noise or improving the context of the image. The second stage is segmentation techniques, where the image is segmented into multiple parts that have similar qualities. At the same time, the important stage is the feature extraction of the image for each segment, which determines what features are important to extract or help with accurate classification. [8]. Image pixel information is transformed into a database like atable that allows different data mining algorithms to make exploration easier [9].

This paper is organized as follows. Related work in section 2. Segmentation methods and feature extraction are explained in section 3. The classification used a hybrid decision tree with counter (HDTC) is explained in section 4. Finally, evaluation and results are discussed in section 5.

2. Related Works

In 2015, researchers introduced the HDTKM algorithm as a hybrid algorithm that combines a decision tree with K-Means. Nine standard datasets were tested from the UCI repository and persistent values were processed into a node using the K-Means algorithm. The rating accuracy was moderate for the HDTKM algorithm at 78.1%, compared with the J48 algorithm, which gave 76.1% accuracy. It is noteworthy that the hybrid algorithm collected a decision tree of type C4.5 with K-sets for continuous attributes [10].

In 2018, researchers have suggested an approach of mixed classification algorithms, The research presented the application of several algorithms to categorical or continuous data values. The paper included the application of three different algorithms to the decision tree, namely ID3, CART, and C4.5 The paper discusses the characteristics of handling different types of data so it can be applied to different datasets. A comprehensive study was done on decision tree algorithms and this paper concludes that CART is the most accurate algorithm among the others [11].

In 2019, the researchers introduced the technique of detecting internal defects in X-ray images by categorizing data into two types of classification: defective and non-defective images. Wheel car images are used in the work. The researchers began preprocessing the images by improving the contrast of the images and then the segmentation phase using the FCM algorithm. The third phase is extracting the geometry features. While the classification stage used an optimization threshold decision tree OTDT, which selected the best feature of the decision tree based on entropy, gaining information and then dividing it to ensure as homogeneous a range as possible. The result is a rating of 98.16 compared with the methods Naïve Bayes 85.77 and SOM 77.79 [12].

3. Segmentation And Feature Extraction

Images often contain valuable information that can be obtained by common image processing methods. It is known that X-ray images consist of three gradations of black, white, and gray,where the thicker areas are dark black and light places are grayscale depending, As for the vacuum, it will be white in color[13].

The DBSCAN algorithm was utilized for segmentation, and then the length, width, average, standard deviation, area, and perimeter of each object were determined to create a continuous-type database for the research. to be ready for entry into the classification stage. The focus of the research is on the classification of continuous data that has long been sufferingfrom the problem of forming periods.

Density-based spatial cluster analysis of noisy applications abbreviated as DBSCAN. Unlike k-means, it is an algorithm for clustering data based on density.

This technique performs admirably for locating outliers in a dataset. By analyzing the density of data points in various places, it is able to identify clusters of arbitrary shapes. It does this by dividing the environment into high-density clusters and low-density zones, where it can more easily spot anomalies.

When dealing with irregularly shaped data, this algorithm performs better than k-means, see Fig. 2.

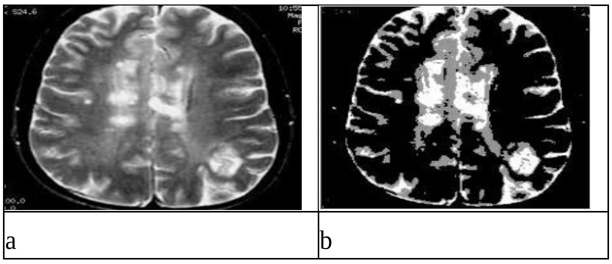

Figure 2: a: Original Image, b: Segmented Image.

Use DBSCAN employs minPts (the minimal number of data points that need to be clustered together for an area to be considered high-density) and eps (the expected probability of finding a cluster) to define clusters (the distance used to determine if a data point is in the same area as other data points).

It is crucial to correctly set the initial parameters for this method. In the X-ray images, the focus will be on the geometric features of objects (length, width, average, standard deviation, area, and perimeter) and the mean gray level feature.

Features such as the area, length, width, perimeter, average, standard deviation, and mean gray level are extracted for each region that was obtained from the image segmentation phase. These features are useful and required to determine the specifications of normal and abnormal objects. The calculation of the first four features (geometric features) gives the possibility of identifying the shape of objects and their size, whether round or longitudinal. Because of irregular shapes, it is necessary to know features that give an indication of the brightness of a pixel by computing mean gray level features, standard deviation, and average.

The area, length, width, perimeter feature gives a measure of the shape region. Data has been extracted from 125 images, a subset of images from the cancer imaging archive. The total amount of data extracted was 3000 objects or records.

4. Hybrid Decision Tree With Counter

It is known that the decision tree is one of the strongest and most reliable algorithms in the classification, as well as being easy to understand and apply, depending on the impurity and information gain standards in choosing the best part of the classification.

In the brute force of classification using the decision tree algorithm, the data is divided into periods and the purity of all periods is calculated using impurity metrics (entropy, error classification, and Gini) and the selection of the purest period is in addition to the calculation of the information gain, we choose the period with the lowest value of the information gain. The relationship between the measures of purity and information gain is inverse.

$$\boldsymbol{Entropy}\ (\boldsymbol{E}) = -\sum_{i=0}^{c-1} \boldsymbol{p(i/t)}\ \boldsymbol{log_2}\ \boldsymbol{p(i/t)} \qquad (\mathbf{1})$$

Where $p(i/t)$ is referred to the fraction of sample that belongs to a class i for a particular node ,c is the number of classes.

$$\text{Gain} = X(\text{parent}) - \sum_{i=1}^{t} \frac{R(h_i)}{R} X(h_i) \qquad (2)$$

X is refers to the impurity of the parent node, h is the number of attribute values, R is the total number r of records of the parent node, (h_i) is a number of records related to the child node.

The first equation is uses to measure the impurity of the period. If the value of entropy is small, it means that the period is purer and belongs to one class and vice versa. The brute force method used a second equation to estimate the information contained by each attribute by using information gain as a criterion. We are going to use some points deducted from information theory to measure the randomness or doubt of a random variable i is defined by Entropy. In this research, the optimal method was used in dealing with continuous data and the possibility of classifying the feature in the correct way was explored using the proposed algorithm HDTC. Where we completely move away from dividing the data and discretization problems meaning that the periods of continuous data are self-forming. The division of data was by using the counter and not a specific threshold limit. At the same time, an advantage of the larger counter feature has been taken in determining the pure period on the basis of which the next childof the tree.

This paper makes use of 3000 records, 2000 records for learning, and 1000 records for testing. To classify any record that starts from the root node and applies the condition in the root to the record, If the predicate is correct, go to the leaf node. If it is wrong, then go to the left child to be classified at other levels of the tree. This process is repeated until the record matches one of the conditions in the tree nodes or until the leaf node is reached. This leaf node classifies the item as either a normal record or abnormal.

To determine the best attribute which categorizes the largest number of records and has a correct classification. As a first step, the data has been arranged in ascending order. The best period (the best attribute that can be adopted as a tree root) is chosen based on the counters in which the data was divided on the basis of its size, from the largest to the smallest counter. After that, a period for the attribute is created and chosen as the paper node bytaking the first and last value from the largest (counter) appearingat the current level. That is the condition for the parent node consists of (F <= R >=L) where R is the record, F is the first element of the counter and L is the last element of the counter. Finally, the data is arranged again for all features and the largest counter is also selected, thereby creating the best feature and choosing it as a paper node for the current level and so on for each level. The same process is repeated in each level, it is necessary that each record has a label because the composition of the period depends on it mainly.

This method differs from C4.5 and decision tree algorithms, for the period to be dynamic depends mainly on similar labels to form the period, and it is not fixed in the number of records on which other methods depend on creating the period. The reason is that the proposed method will interrupt any period in the event that a label different from the previous label appears. It is worth noting that the elements of the period are compared with the label or class, and the counter is cut off in the event that any element of another class is received for that period. So the periods are completely pure and the error classification rate is zero.

In this way, the implication is that the formation of the period does not depend on the division of data that leads to errors in classification. Also, the time complexity has been reduced by avoiding having to calculate the entropy of the whole traits and choosing the lowest entropy. Fig. 3 shows the proposed method steps.

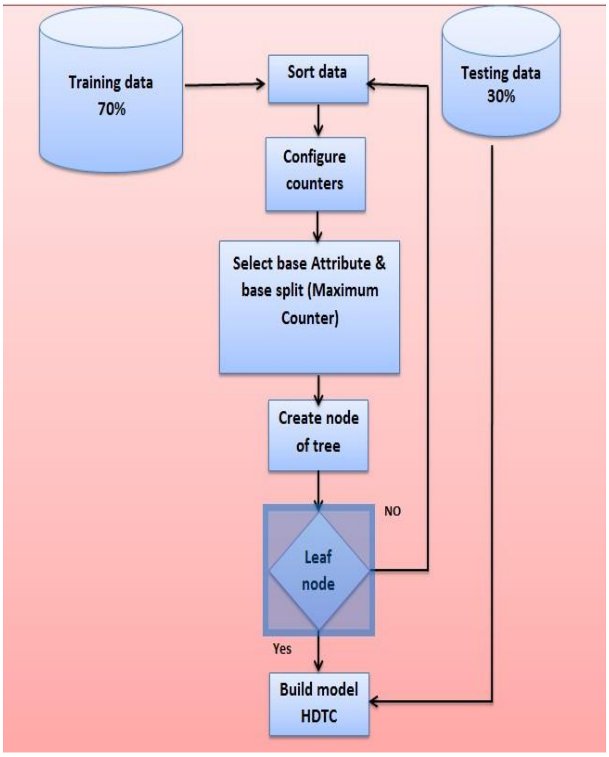

Figure 3. Block Diagram HDTC

It should be noted that each period has been determined to be a (node) condition that applies to a previously invisible record in which the error rating ratio is 0 because we have a period consisting of the largest counter of the existing features and the period is pure as a substitute for the specified period as it is discretization case.

If the result of the condition is correct, a previously invisible record will be classified, but if the condition in the node does not apply to the record, it goes to another level of the tree. Fig. 4 shows the proposed method algorithm.

```
HDTC Algorithm
Input Dataset ,  Number of used attributes
Output HDTC  Model.
Step 1: Split Dataset into Training dataset  and    Testing
Dataset
Step 2: Call Training dataset  to build HDTC  model
Step 3: Sort the Training dataset
Step 4: Configuration counter for each attributes
Step 5: For each attribute find  maximum counter  with
            the same label
Step 7: Determined the first and last value from the
largest (Counter) appeared at the current level
Step 8: Create New Internal node
Step 10: Set direction of node Left
Step 11: If all the node is Leaf node
               Stop
            Else
               Return to step 3
Step 12: Return  HDTS  binary spiting
```

Figure 4 HDTC Algorithm

5. Evaluation of A Proposed Method

GDX- ray is public database images are organized in, they are used for scientific and research purposes. The database used in this research is taken from a set: https://www.kaggle.com/datasets/navoneel/brain-mri-images-for-brain-tumor-detection. The data are a subset of images from the cancer imaging archive.

HDTC Classifier	
Total Number of Instances	3000
Correctly Classified Instances	2867
Incorrectly Classified Instances	133
Accuracy = 95.56 %	

After obtaining the rules resulting from the construction of HDTC we test the model with testing data. The table (1) below shows the confusion matrix for HDTC.

Table .1 Confusion Matrix For Classifier

predicate	Normal	abnormal
normal	TN =2069	FN*=68
abnormal	FN=65	TN*=798

Where FN, TN, FN*, and TN* means are a false normal, true normal, false abnormal, and true abnormal respectively.

Depending on the matrix of confusion we can calculate both accuracy and sensitivity or recall measures. The accuracy is the most widely-used measure of the proximity of the true value and to know the predictive capability of a model.

$$\mathbf{Ac} = \frac{(TN+TN*)}{(TN+TN*+FN+FN*)} \qquad (3)$$

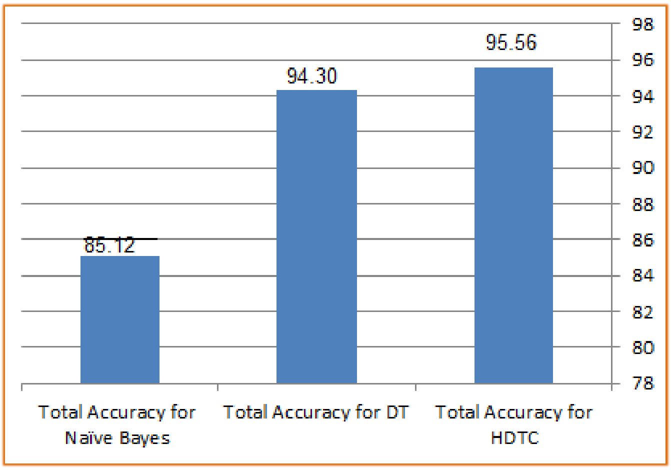

Figure 5. The total percentage of the accuracy classification.

Fig. 5 shows demonstrates the HDTC method to find the best results in classification for continuous data, as it exceeded the correct classification rate compared to the Naïve Bayes and DT method that relies on probability and other methods mentioned in previous works.

Where the Naïve Bayes algorithm and DT mentioned above were applied to the same data for the X-ray images and produced the highest rating for the proposed method.

It is worth noting that the high accuracy in this proposed method is directly proportional to the number of levels of the tree, meaning that the accuracy that we obtained in the training phase was for a tree with 21 levels. It should be noted that the lower levels of the tree classify fewer records compared to the higher levels. To reduce overfitting in the context of decision tree learning, we perform pruning of the tree, i.e., to delete some branches or sub trees and replace them with leaves of majority classes.

This method was applied to data extracted from images, and the implementation of the method on inconsistent data does not result in a high accuracy rate, due to the difficulty of creating periods based on the counter and the label of the entered record and it is very important that all records have a label.

It is worth mentioning that the scope of the research is to focus on the classification and evaluation of continuous data. As for the two stages of segmentation and extracting features, for the research, it was a method of obtaining data, and we did not focus on the evaluation for these two stages because it was not part of the scope of the research.

6. Conclusion

Classification of continuous data is the major problem facing researchers in their choice of classification algorithms with high accuracy. Dividing continuous data into periods has two main problems. One is the discretion to choose the number of periods and the width of each period. In addition to the possibility of classifying two different records in the same period into one category. In this research, we dealt with the classification of continuous data in an ideal method by using an HDTC algorithm that gave a very high classification rate and moved away from the need to divide the data into periods or categories that are accompanied by a mistake in the classification due to the division of data. Where the user relies on simple computer programming with the counter feature, fantastic results are achieved with the hybridization of the decision tree. Where the concept of the tree was used for classification, the best feature and the best division were chosen based on the first and last component of the largest counter of all features. where the period is dynamically dependent on the number of labels or classes that appear sequentially together. It is very important that all records have a label. The proposed algorithm gave an impressive classification accuracy of 95,56 compared to the DT and Naïve Bayes methods that rely on probability, where the classification accuracy is 94.30 and 85.12 sequentially.

Conflicts of interest

The author's no conflicts of interest.

References

[1]D. Computing and I. Santos I. Salazar M. Santamar A. and Bringas PG "Collective Classification for the Detection of Surface Defects in Automotive Castings" 8th IEEE conference 2013:941-946. 2017.

[2]Y. Y Song, and L. U Ying, "Decision tree methods: applications for classification and prediction" Shanghai archives of psychiatry, 27(2): 130. 2015.

[3]G, Mitra S, and S. BKA,"simple data discretizer" ,http://arxiv.org/abs/1710.05091. 2017.

[4]Z. Cebeci, & F. Yıldız, "Unsupervised discretization of continuous variables in a chicken egg quality traits dataset", Turkish Journal of Agriculture-Food Science and Technology, 5(4): 315-320. 2017.

[5]P. Kanmani, P. Marikkannu, and M. Brindha, "A Medical Image Classification using Id3 Classifier", Sri Ramakrishna Institute of Technology, Coimbatore, India 3(10): 27- 30. 2016.

[6]X. Wang, X. Liu, W. Pedrycz, & L. Zhang, "Fuzzy rule based decision trees", Pattern Recognition, 48(1): 50-59. 2015.

[7]M. Norouzi, M. Collins, M. A. Johnson, D. J. Fleet, & P. Kohli, ,"Efficient non-greedy optimization of decision trees", In Advances in neural information processing systems, pp. 1729-1737. 2015.

[8] F. Riaz, K. Kamal, T. Zfar, and R. Qayyum, "An inspection approach for casting defects detection using image segmentation", In Mechanical, System and Control Engineering (ICMSC), pp. 101-105. 2017.

[9] J. Naik, and S. Patel, "Tumor detection and classification using decision tree in brain MRI", International Journal of Computer Science and Network Security (IJCSNS), 14(6): 87. 2014.

[10] P. K. Aparna, & D. R. Shettar, "Hybrid Decision Tree using K- Means for Classifying Continuous Data", International Journal of Innovative Research in Computer and Communication EngineerinG 3(10): 32-37. 2015.

[11] H. H. Patel, and P. Prajapati, "Study and analysis of decision tree based classification algorithms", International Journal of Computer Sciences and Engineering, 6(10): 74-78. 2018.

[12] W. Al-Hameed, and N. Fadel, "Defect Detection Based Optimized Threshold of Decision Tree", Journal of Computational and Theoretical Nanoscience, 16(3): 914-919. 2019.

[13] W. Al-Hameed, and N. Fadel, "Fuzzy Logic for Defect Detection of Radiography Images", Journal of Computational and Theoretical Nanoscience, 16(3**):** 1023-1028. 2019.

[14] F. Zunlin, Bi. DUYAN," Low-level structure feature extraction for image processing via stacked sparse denoising auto encoder" , Elsevier 243(21): 12-20. 2017.