Improving Classification Accuracy of Random Forest Algorithm Using Unsupervised Discretization with Fuzzy Partition and Fuzzy Set Intervals

Muhammad Nur Fikri Hishamuddin
Department of Computer and Information Science
Universiti Teknologi PETRONAS
32610, Seri Iskandar, Perak, Malaysia
+60-173918321
fikri_18002819@utp.edu.my

Mohd Fadzil Hassan
Department of Computer and Information Science
Universiti Teknologi PETRONAS
32610, Seri Iskandar, Perak, Malaysia
+60-123675199
mfadzil_hassan@utp.edu.my

Ainul Akmar Mokhtar
Department of Mechanical Engineering
Universiti Teknologi PETRONAS
32610, Seri Iskandar, Perak, Malaysia
+60-125195009
ainulakmar_mokhtar@utp.edu.my

ABSTRACT

It is known that certain classification algorithm requires continuous data to be discretized for it to produce better classification accuracy. Hence, many works have explored the pairing of classification algorithm and discretization techniques, yet tree-based classifier especially Classification and Regression Trees (CART) still have an issue with classification accuracy regardless of different pairing with existing discretization techniques. The role of fuzzy partition and fuzzy sets interval are not something new in data discretization but none yet to explore the pairing of fuzzy discretization with tree-based algorithm. This paper will be discussing on an approach of using fuzzy based discretization and a member of tree-based algorithm known as Random Forest, a better version of CART. In this study, continuous data are identified from a dataset and discretized through the fuzzy discretization. Then, 10-fold cross validation is done on the transformed dataset and seven well-known classifiers are used including the proposed approach. Based on the results, better classification accuracy is achieved when fuzzy discretization is paired with Random Forest algorithm compared to CART. On top of that, with the present of fuzzy discretization technique, an increased in the classification accuracy has been obtained compared to other classification algorithms.

CCS Concepts

• Information systems→Information systems applications→ Data mining • Computing methodologies→Machine learning → Machine learning approaches → Classification and regression trees

Keywords

Classification; Random Forest; Fuzzy; Discretization.

1. INTRODUCTION

In this paper, an unsupervised discretization technique with fuzzy partition has been proposed to be paired with random forest classifier to boost the classifier's classification accuracy. With the present of continuous data, most classifier produce inaccurate classification thus requiring it to be discretized. In recent years, many researches have been done to improve classification algorithm accuracy by trying different combination of classification algorithms and data discretization techniques. However, classification accuracy is still an issue especially for tree-based algorithm regardless of the different combinations with the existing discretization techniques.

In related studies, among the well-known decision algorithm is tree-based algorithm. With regards to that, C4.5 algorithm was chosen the most to be used as a classifier, followed by fuzzy decision tree and random forest. Due to the limitations of decision tree to handle continuous data, discretization techniques came into the picture.

Data discretization is not something new. In [3,10], a taxonomy of data discretizer has been built to help researchers to understand the types and functions of every data discretization techniques. Thus, many researches from various domain have been using those data discretization techniques in their data pre-processing phase.

In relation to this paper, many studies have been found trying to boost the classification accuracy by including tree-based algorithm as one of their classifiers and several data discretization techniques [5–7,9,11,14]. Most of the studies have not fully explore the decision tree algorithm family and majority focus on C4.5 algorithm. For the discretization techniques, majority use supervised discretization, followed by unsupervised discretization and both. The same result gathered from those study was the classification accuracy of the tree-based algorithm used is still an issue regardless of different combinations with discretization techniques.

Due to presence of uncertainty in decision making, fuzzy logic has been widely used to aid the decision process in the area of discretization and classification. In [12], fuzzy logic has been used for discretization and classification. In the study, the output for discretization is discrete categorical values which will be used for fuzzy classification using fuzzy inference system. However, when dealing with decision tree classification especially CART, the input for classification can be either categorical or numerical. In [2], fuzzy interval has been used as numerical input for classification phase. However, this study only Bayesian Network

ICSCA 2020, February 18–21, 2020, Langkawi, Malaysia

ACM ISBN 978-1-4503-7665-5/20/02…$15.00
https://doi.org/10.1145/3384544.3384590

is used as classifier as it mainly focusses on flood disaster model. As the focus of this paper is enhancing classification accuracy of tree-based classification algorithm, the implementation of fuzzy logic as fuzzy discretization which produces fuzzy interval for classification phases is proposed.

Thus, in this paper, to enhance the classification accuracy of tree-based algorithm through the pairing issue, an integrated discretization technique which consist of unsupervised discretization, fuzzy partition and fuzzy set intervals has been proposed to be paired with Random Forest. This combination has not been explored before based on the related studies. Random forest has many advantages in classification according to the studies [13]. Meanwhile, as the existing discretization techniques unable to enhance the classification accuracy of decision tree algorithm (further discussion in Section II), fuzzy partition and fuzzy sets interval is proposed to help in the data preprocessing phase and boosting the classification accuracy of random forest algorithm.

The paper is organized as follow. Section II describes previous studies on data discretization. In section III, the proposed data discretization technique is discussed. Next in section IV, the paper describes the performance of previous discretization techniques and the proposed discretization technique on random forest classifier. Section VI draws concluding observations of the study.

2. LITERATURE REVIEW

2.1 Data Discretization

Data discretization is one of data preprocessing techniques in data mining. It is performed to transform continuous data into discrete data as most decision-making model such as decision tree and Bayesian network only accepts discrete values from the data [8]. With discretization, infinite range of possible values can be transformed into finite values to be used by decision making model [12]. According to [8][1], there are 5 categories of data discretization, namely (1) unsupervised and supervised, (2) incremental or direct, (3) local or global, (4) dynamic or static and (5) bottom-up or top-down approach. In this study, the focus will be on unsupervised data discretization method. In unsupervised discretization method, class or instance labels are not being used during the discretization process and easy to implement compared to supervised discretization techniques [4][5]. Although most studies recommend supervised discretization method in data pre-processing, classification accuracy is still an issue when pairing supervised discretization techniques with tree-based algorithm.

2.2 Classification Accuracy Issues for Tree Based Algorithm and the Pairing with Discretization Techniques

Several studies have been identified implementing the pairing of classifier and discretization techniques in their study approach regardless of their study goals. The similar issue has been identified in those study whereby during the benchmarking phase, most unable to enhance the classification accuracy of tree-based algorithm such as C4.5, fuzzy decision tree and ID3 [3], [5], [6]. This shows that discretization techniques play an important role for tree-based algorithm and is still an issue to be solve.

2.3 Role of Fuzzy in Data Discretization

Based on [6,7,9,11,14] studies, as the existing discretization techniques unable to enhance the classification accuracy of tree-based algorithm when paired together, fuzzy partition is proposed in this study to assist the existing discretization techniques. As data discretization converts continuous numerical attributes to discrete nominal attributes, encoding techniques have been used in many studies to ease the computation of decision classifier such as decision trees. In [2], the author used fuzzy membership graph to discretize continuous data to be fed into Bayesian flood disaster model. Five fuzzy set intervals were used and represented as linguistic variables ranging from “very low”, “low”, “moderate”, “high” and “very high” with the intervals of 1 to 5 respectively. Meanwhile in [12], fuzzy discretization has been implemented on medical data as it stated that medical data consist of continuous attributes which are vague, imprecise and have multiple distribution of different classes. In the study, fuzzy discretization is derived from interval discretization using the Equal Width (EW) method. Then, fuzzy rule-based classification is implemented to classify the presence or absence of a disease where Mamdani-type Fuzzy Inference System (FIS) is used. The performance of the fuzzy discretization technique is measure with six others classifiers namely Associative Classifier (CBA), Decision tree classifier (c4.5), Support Vector Machine (SVM), Multi-layer Perceptron Classifier (MLP), Naïve Based classifier (NB), k-Nearest Neighbor classifier (KNN) and rule-based Fuzzy Inference System (FIS). The performance parameters were Classification Accuracy (CA), Sensitivity (SN) and Specificity (SP). The result shows that fuzzy discretization based fuzzy classification obtained the highest accuracy of 64.58. This shows that with the help of fuzzy in the discretization phase, classification accuracy of classifier can be enhanced.

2.4 Random Forest Classifier

Random Forest is widely used in developing prediction model due to its popularity in the machine learning paradigm. The idea of random forest came from decision tree whereby according to [13], random forest is collection of classification and regression trees. Due to inability of decision tree to handle complex datasets, thus it results in poor classification accuracy. With random forest, randomly selected training data and predictor variable are the key in the modelling of many classification and regression trees. The results of these trees are aggregated to give a better prediction. Random forest performs better in term of accuracy compared to a single decision tree. On top of that, random forest gives better accuracy compared to other prediction model in term of classification [13]. Therefore, in this study, random forest will be representing the tree-based algorithm family to be paired with discretization technique to achieve better classification accuracy for tree-based algorithm.

3. PROPOSED PAIRING OF FUZZY DISCRETIZATION TECHNIQUE WITH RANDOM FOREST CLASSIFIER

In this study, as continuous data will be discretized, fuzzy discretization is implemented. The proposed method has 3 phases namely discretization, classification and performance analysis as shown in figure 1. In order to simplify the discretization techniques, fuzzy partition and fuzzy sets interval, the processes are renamed into a single process called fuzzy discretization

.

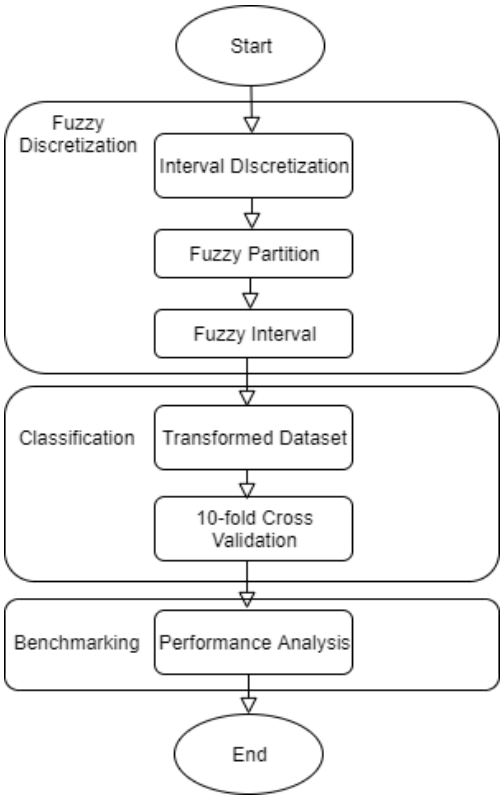

Figure 1. Pairing of Fuzzy Discretization and Random Forest Classification Algorithm.

3.1 Fuzzy Discretization

This phase consists of 3 phases namely interval discretization, fuzzy partitioning and data transformation using fuzzy membership graph and linguistic variable.

3.1.1 Interval Discretization

In this step, method by [12] is adopted where continuous data are identified from the dataset and discretized using Equal Width (EW) discretization method into k equal sized intervals (I1, I2… Ik). The width of intervals (w) are computed using these equations (1), (2) and (3) as below:

$$W = (Vmax - Vmin) / k \quad (1)$$

$$Vmax = \max \{V1,V2,V3…Vn\} \quad (2)$$

$$Vmin = \min \{V1,V2,V3…Vn\} \quad (3)$$

where Vmax and Vmin are maximum and minimum values of attributes respectively. K is the number of cut point specified by domain's expert. In this study, k = 5 as the intervals that will be used is five.

3.1.2 Fuzzy Partitioning

In this step, fuzzy membership graph is developed from the interval discretization. Fuzzy sets are constructed for each attribute of the dataset using triangular membership function and intervals values computed previously [12]. Three parameters known as (a,b,c) are required in this membership function where a and c represent the lower and upper limit of intervals and fuzzy sets respectively while b represents the center of the fuzzy sets. This results in non-overlapping of 5 interval fuzzy sets.

3.1.3 Data Transformation Using Linguistic Variable

In this step, linguistic variables are used to discretize and label the fuzzy membership interval graph. This method is used by [2] where linguistic variables represent the fuzzy set intervals as shown in table 1 below:

Table 1. Mapping of Linguistic Variables with Fuzzy Intervals

Linguistic Variables	Fuzzy Interval
Very Low	1
Low	2
Moderate	3
High	4
Very High	5

The fuzzy interval values are the output of the discretization phase. As the final step of data pre-processing, these fuzzy interval values will be replacing the original values in the dataset accordingly and in other words, the random forest classifier will receive these fuzzy intervals as an input for the classification process.

3.2 Random Forest Classification

In this phase, the discretized values which now known as fuzzy intervals values are used as an input for decision classification process using random forest. The steps implemented in this phase are as follows:

3.2.1 Identify Target Variable in a Dataset

Target variable is the output variable or decision class where all variable in the dataset will be classified into the target variable. On top of that, other variables in the dataset are known as predictive variable or predictive features.

3.2.2 Dividing the Data

In this step, all predictive variables in the dataset are divided into two parts, namely training data and testing data. Training data is used to train the model while testing data is used to make model validation and predictions.

3.2.3 Building the Forest

In this step, six sub-steps involved. Step 3.1 until step 3.4 is the process of building a decision trees while step 3.5 is the process of building multiple decision tress or known as the forest. Those sub-steps are as below:

Step 1:

Randomly select m variable from T, where m<T and m is the number of randomly selected predictive variable and T is the total number of predictive variables in the dataset.

Step 2:

For every node, d, in the decision tree, calculate the best split point among the m variable.

Step 3:

Split the node into two child nodes using the best split method.

Step 4:

Repeat step 1, 2 and 3 until n number of nodes is reached where no more classification can be done. The final node is known as leaf node.

Step 5:

Repeat step 3.1, 3.2, 3.3 and 3.4 until D number of times where D is the total number of trees in the forest.

Step 6:

Compile the result of all decision trees and take the majority voting on the decision predictions.

3.3 Performance Benchmarking

In this phase, 10-fold cross validation technique is implemented. The dataset along with the discretized data will be run by several classification algorithm. According to [12], performance evaluation parameters are Classification Accuracy (CA), Sensitivity (SN) and Specificity (SP) and are computed through the following equations (4), (5) and (6) :

$$CA = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

$$SN = TP / (TP + FN) \quad (5)$$

$$SP = TN / (TN + FP) \quad (6)$$

where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively. In this study only classification accuracy is taken, and it is taken as a mean classification accuracy.

4. SIMULATION

In this study, MATLAB is used for fuzzy discretization process while Jupyter-Notebook with Python is used for the classification and prediction process. Diabetes dataset from Kaggle is used for this study. From the dataset, two variables have been identified as continuous variable namely Body Mass Index (BMI) and Diabetes Pedigree Function (PED) as shown in figure 2. BMI and PED are discretized through the fuzzy discretization method mentioned in Section III. The output of the process are fuzzy interval values of 1 to 5.

Then, the dataset is updated where all continuous values of BMI and PED are replaced with the discretized values as shown in figure 3. To benchmark the result, classification process is then done through several well-known classification algorithm namely Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Classification and Regression Trees, Naïve Bayes, Support Vector Machine and Random Forest. According to scikit-learn documentations on classifier, tree-based algorithm like C4.5, ID3 and C5.0 are represented by Classification and Regression Trees (CART). 10-Fold cross validations have been implemented on the datasets and all classification models. The mean classification accuracy is taken.

Out[9]:

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1
10	4	110	92	0	0	37.6	0.191	30	0

Figure 2. Original Dataset with Continuous Values of BMI(mass) and PED(pedi).

Out[17]:

	preg	plas	pres	skin	test	mass	pedi	age	class
0	6	148	72	35	0	1.0	5.0	50	1
1	1	85	66	29	0	1.0	3.0	31	0
2	8	183	64	0	0	1.0	5.0	32	1
3	1	89	66	23	94	1.0	3.0	21	0
4	0	137	40	35	168	1.0	5.0	33	1
5	5	116	74	0	0	1.0	3.0	30	0
6	3	78	50	32	88	1.0	3.0	26	1
7	10	115	0	0	0	1.0	3.0	29	0
8	2	197	70	45	543	1.0	3.0	53	1
9	8	125	96	0	0	1.0	3.0	54	1
10	4	110	92	0	0	1.0	3.0	30	0

Figure 3. Transformed Dataset with after Discretization Process with Fuzzy Interval Values on BMI(mass) and PED(pedi).

5. RESULTS AND DISCUSSION

As the main objective of this study is to enhance classification accuracy of tree-based algorithm specifically Random Forest, only the results of tree-based classification algorithm are included for discussion. This is done to prove that in the tree-based classification algorithm family, the classification of Random Forest can be enhanced when paired with fuzzy discretization. Based on the observations on the mean classification accuracy obtained, the result draws several observations although Random Forest is ranked 4th among the seven classifiers as shown in table 2.

First, comparing the classification accuracy between tree-based algorithm which are Random Forest and CART, Random Forest perform better in term of mean classification accuracy with Random Forest being in the 4th place while CART being in the 6th place among the seven classifiers involved.

Table 2. Ranking of Tree-Based Classification Algorithm with Fuzzy Discretization

Rank	Classification Algorithm	Mean Classification Accuracy
1	LDA	0.773462
2	LR	0.769515
3	NB	0.755178
4	RF	0.751299
5	KNN	0.726555
6	CART	0.682194
7	SVM	0.651025

Table 3. Gap Difference and Total Gap Difference of Tree-Based Classification Algorithm Before and After Fuzzy Discretization

	RF	CART	Gap Difference
Before pairing with Fuzzy Discretization Technique	0.738260	0.695284	0.042976
After pairing with Fuzzy Discretization Technique	0.751299	0.682194	0.069105
Increase in Gap Difference			0.03

Table 4. Difference in Classification Accuracy for All Classification Algorithm Involved.

	LDA	LR	NB	RF	KNN	CART	SVM
Before Pairing with Fuzzy Discretization Technique	0.773462	0.769515	0.755178	0.738260	0.726555	0.695284	0.651025
After Pairing with Fuzzy Discretization Technique	0.773462	0.769515	0.755178	0.751299	0.726555	0.682194	0.651025
Difference in Mean Classification Accuracy	0	0	0	+0.013039	0	-0.01309	0
Status	No Effect	No Effect	No Effect	Increase	No Effect	Decrease	No Effect

Next, with fuzzy discretization proposed, the gap difference between Random Forest and CART before discretization and after discretization increased 0.03 % which bring to the last observations where with fuzzy discretization, the mean classification accuracy of Random Forest has increased 0.013 % while CART decreased with 0.013% and others remain constant as shown by table 3 and table 4 respectively. These observations conclude that fuzzy discretization is good combination with Random Forest algorithm.

6. CONCLUSION

In the goal of boosting classification accuracy, the pairing combination between data discretization techniques and classification algorithm is still an issue especially with tree-based algorithm. Hence, in this paper, an unsupervised discretization technique combined with fuzzy partitions and fuzzy set intervals or known as fuzzy discretization, is paired with a tree-based classification algorithm, Random Forest. After the discretization process, 10-fold cross validation is done with seven well known classification algorithms with two of them consist of tree-based algorithm known as CART and Random Forest. Overall result discussion shows that with fuzzy discretization technique, the classification accuracy of Random Forest algorithm can be enhanced compared to CART and other classification algorithms. Among future work that can be explored within this work is improving the classification accuracy of random forest with fuzzy discretization compared to other classification algorithms that currently leads the pack such as Linear Discriminant, Logistic Regression and Naïve Bayes.

7. REFERENCES

[1] Gennady Agre and Stanimir Peev. 2002. On Supervised and Unsupervised Discretization. *Methods* 2, 2 (2002).

[2] Nor Idayu Ahmad-Azami, Nooraini Yusoff, and Ku Ruhana Ku-Mahamud. 2018. Fuzzy Discretization Technique for Bayesian Flood Disaster Model. 2, 2 (2018), 167–189.

[3] Azuraliza Abu Bakar, Zulaiha Ali Othman, Nor Liyana, and Mohd Shuib. 2009. Building A New Taxonomy For Data Discretization Techniques. *2009 2nd Conf. Data Min. Optim.* October (2009), 132–140. DOI:https://doi.org/10.1109/DMO.2009.5341896

[4] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and Unsupervised Discretization of Continous Fearures. 0,.

[5] Michela Fazzolari, Rafael Alcalá, and Francisco Herrera. 2014. A multi-objective evolutionary method for learning granularities based on fuzzy discretization to improve the accuracy-complexity trade-off of fuzzy rule-based classification systems : D-MOFARC algorithm ☙. *Appl. Soft Comput. J.* 24, (2014), 470–481. DOI:https://doi.org/10.1016/j.asoc.2014.07.019

[6] Mehmet Hacibeyoglu and Ahmet Arslan. 2011. Improving Classification Accuracy with Discretization on Datasets Including Continuous Valued Features. June (2011).

[7] Ehsan Ali Kareem and Mehdi Duaimi. 2014. Improved Accuracy for Decision Tree Algorithm Based on Unsupervised Discretization. September (2014).

[8] Sotiris Kotsiantis and Dimitris Kanellopoulos. 2006. Discretization Techniques : A recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* 32, 1 (2006), 47–58.

[9] Simone A Ludwig. 2015. Analyzing gene expression data : Fuzzy decision tree algorithm applied to the classification of cancer data Analyzing Gene Expression Data : Fuzzy Decision Tree Algorithm applied to the Classification of Cancer Data. August (2015). DOI:https://doi.org/10.1109/FUZZ-IEEE.2015.7337854

[10] Sergio Ram, David Mart, and Manuel Ben. Data Discretization : Taxonomy and Big Data Challenge. 1–

26.

[11] Sahar Sardari, Mahdi Eftekhari, and Fatemeh Afsari. 2017. Hesitant fuzzy decision tree approach for highly imbalanced data classification. *Appl. Soft Comput. J.* 61, (2017), 727–741. DOI:https://doi.org/10.1016/j.asoc.2017.08.052

[12] M. Shanmugapriya, H. Khanna Nehemiah, R.S. Bhuvaneswaran, Kannan Arputharaj, and J. Dhalia Sweetlin. 2017. Fuzzy Discretization based Classification of Medical Data. *Res. J. Appl. Sci. Eng. Technol.* 14, 8 (2017), 291–298. DOI:https://doi.org/10.19026/rjaset.14.4953

[13] Jaime Lynn Speiser, Michael E Miller, Janet Tooze, and Edward Ip. 2019. A comparison of random forest variable selection methods for classification prediction modeling. *Expert Syst. Appl.* 134, (2019), 93–101. DOI:https://doi.org/10.1016/j.eswa.2019.05.028

[14] Chih-fong Tsai and Yu-chi Chen. 2019. The optimal combination of feature selection and data discretization : An empirical study. *Inf. Sci. (Ny).* 505, (2019), 282–293. DOI:https://doi.org/10.1016/j.ins.2019.07.091