Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 4, pp. 733–749
DOI : 10.47974/JSMS-863

Developing a mixture decision tree (MDT) algorithm with k-means clustering to improve classification accuracy

Md. Ashraful Islam *
Department of System Innovation
Graduate School of Engineering Science
Osaka University
Japan

Deluar Jahan Moloy
Md. Sifat Ar Salan
Department of Statistics
Mawlana Bhashani Science and Technology University
Bangladesh

Abstract

Statistical research based on data is one of the primary keys to the development of the world. But the size of the data is getting increased very rapidly. For this reason, statistical data analysis is getting complicated, and sometimes the existing data mining techniques are not providing significant and satisfactory results. On the other hand, traditional classification techniques provide results that are not statistically significant every time. In this research, a Mixture Decision Tree algorithm is being proposed that is a more powerful classification technique, and that may provide better results than the existing methods. In this Mixture Decision Tree, the Classification and Regression Tree (CART) algorithm is combined with k-means clustering, by which the primary data is filtered in two stages. Applying this method to some predefined data, significant enhancement is found in the classification precision for the current classification techniques. The Mixture Decision Tree algorithm is slightly more complex than the existing methods, but computer programming techniques can be done quickly.

Subject Classification: *62J05, 03C45.*

Keywords: *Data mining, Neural network, CART, Random forest, Mixture decision tree, Classification accuracy.*

* *E-mail:* ashraful.stat@yahoo.com

1. Introduction

In recent years, the globe has been running with modern science, where big data has become an essential part of the fashionable age. Enormous data refers to the improvement in the volume of cooperative and unstructured data, the speed at which it's created and received, and, as needed, the number of data centers available. Massive data frequently comes from a variety of sources and appears in numerous formats. Large-scale data analysis examines a large amount of data in order to identify underlying models, relationships, and varied encounters. With the current advancement, it's possible to investigate your data and acquire answers from it right away, an undertaking that is progressively more profitable with more standard business information courses of action. The term "information handling" perplexes many people, particularly those who are not involved with information technology; most people who hear the term "information preparation" conceive of excavators uncovering and searching for gold and precious stones. By examining data for ordinary plans, grouping, bunching, and association rules are created.

Using the assistants and surety norms to locate the primary critical associations inside the information, Support refers to how constantly things appear in the database, whereas conviction refers to how accurate the events and verbalization are. The orders play an important role in this investigation. Grouping and anticipation are two types of data analysis that can be used to segregate models depicting important information classes or predict future information patterns.

Grouping might be an information-preparing capacity that allots things in a very wide assortment to zero classes or classifications. The decision tree includes multidimensional applications such as weather forecasting, credit endorsement, finding, misrepresentation discovery, archive arrangement, image acknowledgment, signal grouping, among other examples, where multi-class is an important issue in example acknowledgment applications. The decision tree includes multidimensional applications such as weather forecasting, credit endorsement, finding, misrepresentation discovery, archive arrangement, image acknowledgment, signal grouping, among other examples, where multi-class is an important issue in example acknowledgment applications.

The decision tree has a few focal points; for example, it is not difficult to grasp and can manage large amounts of data. Other AI models will join the decision tree (DT). The dataset with numerous qualities and each

character having many traits of esteem will build a decision tree. It may be necessary to characterize visible and hidden preparation cases in the decision tree. However, these strategies do not always produce meaningful results. In these circumstances, the hybrid model is essential. In this paper, the feed-forward neural network (FFNN) of artificial networks (ANN) with optimization techniques such as particle swarm optimization (PSO) and support vector regression (SVR) are used in short-term PV power generation forecasting.

The error calculation in terms of the mean absolute error evaluated the result (MAE) and root mean squared error (RMSE) (Jyothi Varanasi & M. M. Tripathi, 2019). K-means simple (using Java code on MapReduce), K-means with IEC (Initial Equidistant Centres), K-means on Mahout (using the Machine learning library), and K-means on Spark were all utilized to evaluate the K-means method (using the Machine learning library). To apply these methods to static IP address data sets, we used the MapReduce framework (Satvik Vats & B. B. Sagar, 2019).

2. Literature Review

A literature review is likely to be an appraisal of conveyed sources or composition on a picked subject. It is a review of the composition that includes a diagram, portrayal, relationship, and assessment. At the postgraduate level, a piece could intertwine into a composition overview, an investigation report, or proposition. No new or imaginative test work was reported because writing audits are optional assets. Most writing surveys are associated with academic structured writing, which consists of a proposition or individual researched pieces; it generally takes an examination proposition and winds up in the section. Its significant goal is to find the ongoing investigation in the collection of writing and to respect a particular correctness with flexibility. The research proposed a parcel restrictive autonomous partial examination (PCICA) strategy for guileless Bayes order in microarray information investigation. PCICA split the small-sized information examinations into different segments, allowing for independent segment examinations (ICA) inside each component. PCICA also attempted to extract ICA-based elements from each portion, obliging a few classes (Polat & Güneş, 2006). Using a C4.5 blend, they likewise achieved 91.84 percent and 92.94 percent correctness.

A decision tree with a weighted pre-taking care of and a mix of a C4.5 decision tree with k-NN set up a weighted pre-planning concerning

the finding of erythrocyte at squalor infections, independently. A couple of assessments were declared to impress experts during the action of the picture division. Tin and Kwork achieved an 83 percent course of action accuracy in one of these tests (Kwok, 1999) by using a support vector machine (SVM) to depict visual division. Another study used the k-NN (k-closest neighbor) classifier and achieved an 85.2 percent success rate (Tolson, 2001).This paper also obtained 87.6% and 86.7% percent characterization exactness using PNN (probabilistic neural organization) and GRNN (summarized relapse organization).

The author proposed a fluffy choice tree Gini Index-based (G-FDT) calculation to fuzzily the choice limit without changing the numeric qualities in smooth etymological terms. To select the first suitable parting trait for each hub within the decision tree the Gini Index used as the split measure in the G-FDT tree. The Gini list processed for the choice tree using fluffy enrollment estimates of the quality love a split worth and fluffy participation estimates of the occurrences (Chandra & Varghese, Fuzzifying Gini Index based decision trees, 2009). The split centre's picked in light of the midpoints of quality regards and changed the class information. The paper presented a co-creating decision tree method and considered an inquisitively enormous number of elements in datasets. They proposed a single blend of DTs and formative procedures, equivalents to the terminating approach of a DT classifier and a back-inducing neural association system, to enhance the gathering precision (Aitkenhead, 2008).

To identify group network assaults, the authors used a veiled gullible Bayes (HNB) classifier in an organization interruption identification framework (NIDS). It significantly improves the recognition of denial-of-service (DOS) assaults (Koc, Mazzuchi, & Sarkani, 2012). Another study provided a fantastic Nave Bayes classifier to portray trademark wonder datasets while overcoming two major interferences: sub-current and over fitting. Mazzuchi, & Sarkani, 2012). Another study provided a fantastic Nave Bayes classifier to portray trademark wonder datasets while overcoming two major interferences: sub-current and over fitting. It did not require any previous component assurance methodologies in the context of micro array data analysis measured a larger than expected number of characteristics. (Chandra and Gupta, Robust way for estimating probability in Nave–Bayes Classifier for quality articulation information).

To arrange blended styles of information, this paper presented a grouping approach called expanded gullible Bayes (ENB). The mixed

information styles incorporated into Complete and numeric data ENB utilized a standard NB calculation to ascertain the probabilities of unmitigated ascribes (Hsu, Huang, & Chang, 2008). It received the quantifiable theory to discrete the numeric credits into images by considering the usual and change of numeric characteristics when managing numeric attributes. A Naive Bayes classifier is likely to a required probability-based approach for predicting arrangement interest probability. It's a couple of right conditions: (a) ease of use, and (b) Probability age required for just one compass of arrangement data (Chena, Huang, Tian, & Qu, 2009). Text characterization with Naive Bayes is an option (Kemal & Salih, 2009). An NB classifier may undoubtedly deal with missing trait esteems by essentially excluding the comparing probabilities for those characteristics when computing the probability of enrollment for each class. The classification restrictive freedom is also required by the NB classifier, i.e., the impact of quality on a particular type is independent of distinct ascribes (Farid & Rahman, Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm, 2010)

3. Methodology

Today, the world is running with big data. The data size is getting bigger very rapidly. There are several methods to handle this kind of big data. Data mining technique is broadly used to conduct with this type of big data. Data mining techniques, including Clustering and Classification, are two of these methods with a wide application in big data. Hierarchical and non-hierarchical methods solve the clustering problems, and the classification problems are solved by several methods, including ID3, Random Forest, C4.5, C5.0, CART etc. Sometimes these methods are unable to give the best performance. So, it is essential to think apart from the built forms. For this reason, it is badly needed to develop a Mixture Decision Tree when the usual methods give the results but not as expected. In this chapter, the main task is to show the development of a Mixture Decision Tree algorithm and how the classification accuracy compares the ordinary methods.

3.1 *Development of a Mixture Decision Tree Algorithm*

A Mixture Decision Tree Algorithm is the demand of the time. There are invented several methods and algorithms at the passes of time. In this section, we have developed a Mixture Decision Tree Algorithm which is applied to some existing datasets in from UCL (University College London) machine learning repository (UCL Machine Learning Repository, University of California., 2008) (Frank & Asuncion, 2010).

3.1.1 The classification and clustering methods used in the proposed Algorithm

Several methods are developed to solve classification problems. Some of these give higher classification accuracy compare the others. In this section, several important ordinary classification methods which are considered in this research given below.

1. **K-means clustering (K-Means):** K-means bunching is a kind of solo realizing, which is utilized when you have unlabeled information (i.e., information without characterized classes or gatherings). The objective of this calculation is to discover bunches in the information, with the number of gatherings, spoke to by the variable K. The calculation works iteratively to allocate every information highlight one of K bunches dependent on the highlights that are given (Farid, Zhang, Rahman, Hossain, & Strachan, 2014).
2. **Random Forest (RF):** A random forest is essentially an assortment of choice trees whose outcomes are accumulated into one conclusive result. Their capacity to restrict over fitting without significantly expanding blunder because of predisposition is the reason they are such ground-breaking models. Single direction Random Forests diminish difference is via preparing on various examples of the information (Farid, Zhang, Rahman, Hossain, & Strachan, 2014).
3. **Classification and Regression Tree (CART):** Classification and Regression Trees or CART for short is a term acquainted by Leo Breiman with alluding to Decision Tree calculations that can be utilized for order or relapse prescient displaying issues (Farid, Zhang, Rahman, Hossain, & Strachan, 2014).

3.1.2 The datasets that are used in the proposed algorithm

A bunch of information things called datasets which is very root idea of information mining and AI research. A dataset is generally identical

Table 1
Datasets Descriptions

Datasets	Number of Attributes	Number of Instances	Number of Classes
Iris Plants	4	150	3
beaver2	3	100	2
ToothGrowth	2	60	2

to a two-dimensional accounting page or information base table. Table 1 describes the datasets from r-cran and UCL machine learning repository which are used in experimental analysis (UCI Machine Learning Repository, University of California., 2008) (Frank & Asuncion, 2010).

3.2 *Conceptual Framework*

The main objective of this research is to reduce misclassification errors. As we know, the classification technique is supervised learning. For this reason, we have to identify all the given classes in which the data will be classified. Including the predefined classes but keeping the number of classes constant, we can conduct clustering to the datasets by different clustering techniques like hierarchical clustering (single linkage, complete linkage, average linkage, ward etc.) and non-hierarchical (k-means clustering, fuzzy k-means clustering, sequential k-means clustering etc.) (Chandra & Varghese, Fuzzifying Gini Index based decision trees, 2009). Using the cluster analysis output as the input of the classification technique, by keeping the predefined classes constant, we can find a better classification output. Keeping an eye on the classification techniques, we have first applied k-means clustering. After clustering, the classification rule is applied. The essential strides of k-implies bunching are basic. Initially, it decides the quantity of group k and expects the centroid or focuses of these bunches. Next, it accepts rare items as the underlying centroids or first k articles in an arrangement that can likewise serve the centroids. The methodological four (04) steps are described below:

To apply this method, the "Iris Plants Database" is used as example,
Step-1 : Identify and then remove the target classes of the original data We take 10 random observations from the "Iris Plants Database" where "Species" is the target class, the data set is as follow:

Table 2

10 random observation from Iris Plants Database.

SN	Sepal. Length	Sepal. Width	Petal. Length	Petal. Width	Species
57	6.3	3.3	4.7	1.6	versicolor
138	6.4	3.1	5.5	1.8	virginica
14	4.3	3	1.1	0.1	setosa
103	7.1	3	5.9	2.1	virginica
119	7.7	2.6	6.9	2.3	virginica
11	5.4	3.7	1.5	0.2	setosa
7	4.6	3.4	1.4	0.3	setosa
128	6.1	3	4.9	1.8	virginica
89	5.6	3	4.1	1.3	versicolor
118	7.7	3.8	6.7	2.2	virginica

In the first step, we remove the target class from the above data set in Table 2. After removing the target class, we obtain the following data set in Table 3,

Table 3

10 random observation from Iris Plants Database after removing target classes.

SN	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
57	6.3	3.3	4.7	1.6
138	6.4	3.1	5.5	1.8
14	4.3	3	1.1	0.1
103	7.1	3	5.9	2.1
119	7.7	2.6	6.9	2.3
11	5.4	3.7	1.5	0.2
7	4.6	3.4	1.4	0.3
128	6.1	3	4.9	1.8
89	5.6	3	4.1	1.3
118	7.7	3.8	6.7	2.2

Step-2 : Apply k-means clustering

Now apply k-mean clustering is applied to obtain new target classes.

Step-3: Assign obtained clusters to the original data

The gained classes are then assigned to the data set from which the target classes were removed in step-2 as follows in Table 4.

Table 4

Assignment of obtained clusters with original dataset.

SN	Sepal. Length	Sepal. Width	Petal. Length	Petal. Width	Species
57	6.3	3.3	4.7	1.6	versicolor
138	6.4	3.1	5.5	1.8	setosa
14	4.3	3	1.1	0.1	virginica
103	7.1	3	5.9	2.1	versicolor
119	7.7	2.6	6.9	2.3	versicolor
11	5.4	3.7	1.5	0.2	setosa
7	4.6	3.4	1.4	0.3	versicolor
128	6.1	3	4.9	1.8	virginica
89	5.6	3	4.1	1.3	versicolor
118	7.7	3.8	6.7	2.2	setosa

Step-4 : Finally Apply CART Algorithm

Finally, we apply CART algorithm to the dataset found in step-3 and measure the classification accuracy.

4. Results and Discussion

In this section, the obtained results are discussed with tables and figures.

4.1 *Experimental analysis*

To illustrate the proposed mixture decision tree algorithm, the Iris plants dataset is considered an experimental dataset. The first step in a mixture decision tree algorithm is to remove the target class (for iris data, "Species"). In the second step, the k-means clustering is performed to find out clusters, and the obtained clusters are assigned to the original dataset to find out the modified dataset. In the last step, the Classification and Regression Tree is applied to the modified data and finally obtained the results. Table 5 shows the original Iris Plants dataset (Frank & Asuncion, 2010).

Table 5
Dataset (Iris).

SN	Sepal. Length	Sepal. Width	Petal. Length	Petal. Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
. …					
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3	5.1	1.8	virgin

That given Iris dataset is modified by k-means clustering. Firstly, the target classes of the original data is removed and then k-means clustering is imposed to the information to discover the quantity of groups. Also, the acquired bunches are doled out to the objective classes rather than the predefined classes.

4.2 *Experimental Results*

The experimental results from different decision tree algorithms (CART, Random Forest and Mixture Decision Tree) for three types of data sets are shown in 3 different tables below.

Table 6
Results of CART (Classification and Regression Tree) decision tree

Datasets	Correctly Classified	Incorrectly Classified	Classification Rate (%)	Misclassification Rate (%)
Iris Plants	144	6	96	4
beaver2	96	4	96	4
ToothGrowth	40	20	66.67	33.33

Table 7
Results of Random Forest decision tree

Datasets	Correctly Classified	Incorrectly Classified	Classification Rate (%)	Misclassification Rate (%)
Iris Plants	143	7	95.33	4.76

Contd...

beaver2	98	2	98	2
ToothGrowth	40	20	66.67	33.33

Table 8

Results of the Mixture decision tree algorithm

Datasets	Correctly Classified	Incorrectly Classified	Classification Rate (%)	Misclassification Rate (%)
Iris Plants	146	4	97.33	20.67
beaver2	100	0	100	0
ToothGrowth	58	2	96.67	3.33

4.3 *Comparison of the results*

The study was revealed to explore the performance of the developed Mixture Decision Tree algorithm and to compare the results. The comparison of the results and the performances of different decision tree models are shown on different graphs below.

4.3.1 Comparison of the results among the usual methods and Mixture Decision Tree

Comparisons by observed results for using different decision tree methods on three different types of datasets are shown below.

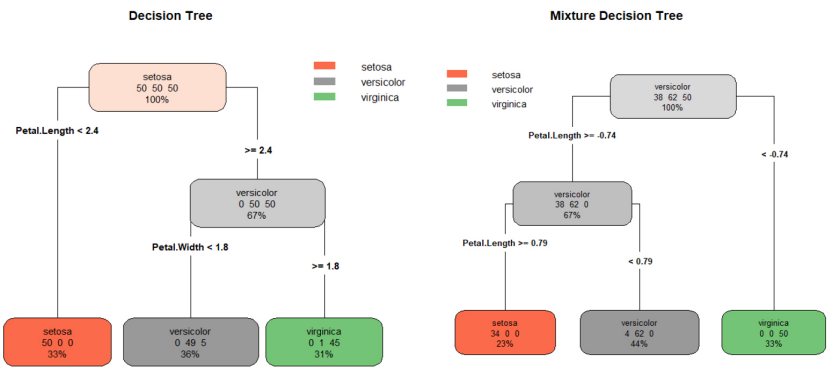

Figure 1

Comparison between CART and Mixture Decision Tree for Iris Plants

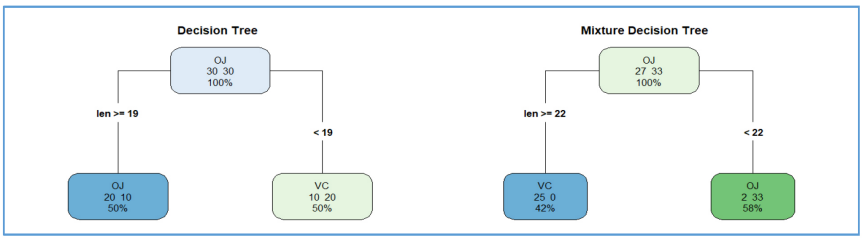

Figure 2

Comparison between CART and Mixture Decision Tree for ToothGrowth dataset

Figure 1 shows the comparison between the results obtained by the CART decision tree and the Mixture Decision tree algorithm for the Iris Plants dataset. The left part of Figure 1 reveals the result for the CART. It describes that no item is misclassified in setosa, five items are misclassified in versicolor, and 1 item is misclassified in virginica. On the other hand, the right part of Figure 1 shows the result for using the mixture decision tree. This part describes that no item is misclassified in setosa and virginica; only four items are misclassified in versicolor. Therefore, it is clear that the result is improved for using a Mixture Decision Tree.

Figure 2 compares the results obtained by the CART decision tree and the Mixture Decision tree algorithm for the ToothGrowth dataset. The left part of Figure 2 reveals the result for CART. It describes that ten items are misclassified in each OJ and VC. On the other hand, the right part of Figure 2 finds the result for the mixture decision tree. This part describes that no item is misclassified in VC, and two items are misclassified in OJ. It also implies that the result is improved for Mixture Decision Tree.

Figure 3 describes the comparison between the results obtained by the CART decision tree and the Mixture Decision tree algorithm

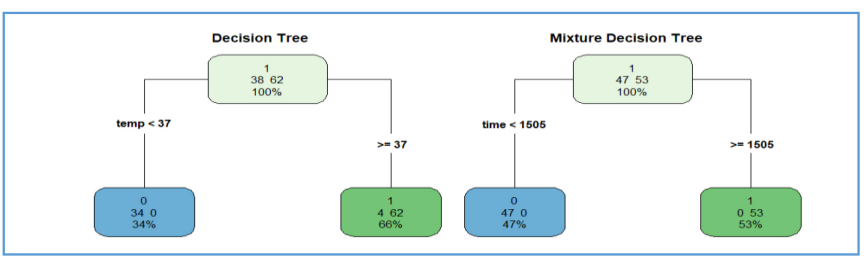

Figure 3

Comparison between CART and Mixture Decision Tree for beaver2 dataset

for the beaver2 dataset. The left part of Figure 3 shows the result for CART. It describes that no item is misclassified in 0 class, and 40 items are misclassified in 1 class. On the other hand, the right part of Figure 3 shows the result for using the Mixture Decision Tree. This part reveals that no item is misclassified in both 0 and 1 classes. It also implies that the result is improved for Mixture Decision Tree.

4.3.2 Comparison of the performances among usual methods and Mixture Decision Tree

Table 9

Comparison of the performances among usual methods and Mixture Decision Tree

Methods	Correctly Classified Items (Accuracy %)		
	Methods		
	CART	Random Forest	Mixture Decision Tree
Iris Plants	144 (96)	143 (95.33)	146 (97.33)
beaver2	96 (96)	98 (98)	100 (100)
ToothGrowth	40 (66.67)	40 (66.67)	58 (96.76)

The following Figure 4 shows the comparison of the performances obtained by different decision tree models. The proposed mixture decision tree gives the best performance for the given three sets of data compared to the CART and Random Forest decision trees model. It implies that the accuracy of classification problems improves by modelling with the mixture decision tree algorithm.

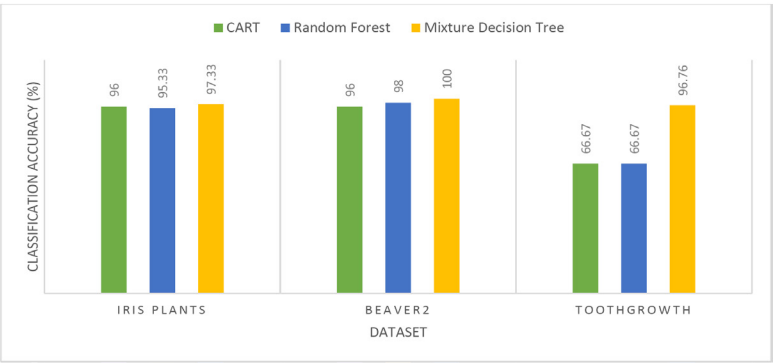

Figure 4

Performance of different types of decision trees.

Decision tree models play a vital role to perform classification problems. Till now, there have been built up countless decision tree models. But in many cases, all the decision tree models do not perform up to the mark. This case tends to develop a new model. For this reason, a new mixture model is developed by combining CART (Classification and Regression Tree) and k-means clustering algorithm, which reduces misclassification errors more than the other ordinary methods.

4.4 *Discussion*

After analyzing and contrasting the effects of various approaches, such as CART (Classification and Regression Method), ID3 (Iterative Dichotomiser 3), C4.5 (an expansion of Quinlan's earlier ID3 algorithm), Random Forest, and suggested hybrid decision tree algorithms (MDT-Mixture Decision Tree Algorithm), it is apparent that MDT improves classification accuracy. Tables 6, 7, and 8 display the outcomes of multiple decision tree algorithms for various datasets. The results show in Table 6 that Table 6 indicates that the classification precision of the CART model for various categories of datasets (Iris Plants, Beaver2, and ToothGrowth) is 96 per cent, 96 per cent, and 66.67 per cent, respectively, for Iris Plants, Beaver2, and ToothGrowth. Similarly, the classification accuracies for the Random forest model are 95.33 per cent, 98 per cent, and 66.67 per cent, respectively. Finally, the suggested Mixture Decision Tree Algorithm yields 97.33 per cent, 100 per cent, and 96.67 per cent. Related findings are seen in various graphs, such as figures 1, 2, and 3. Figure 4 and Table 9 display the results of the final comparisons. When the suggested Mixture Decision Tree Algorithm is used, the overall findings in Table 9 show a substantial improvement in classification accuracy.

5. Conclusion

The research was undertaken to establish an algorithm for data mining that may handle the classification problem more efficiently to reduce misclassification error. The proposed Mixture Decision Tree algorithm was developed by combining CART (Classification and Regression Tree) with K-means clustering. Then the algorithm has been applied to some pre-classified well-established dataset (*Iris, beaver2 and ToothGrowth*) to explore the change of accuracy or misclassification rates. The results show that the Mixture Decision Tree algorithm is providing a higher correctly classification rate compared to *Random Forest and Classification*

and Regression Tree (CART) classification methods for every dataset, and the correct classification rate improve almost 1.33% for *the Iris dataset*, 4% for beaver2 dataset and 30.09% *ToothGrowth* dataset.

References

[1] Aitkenhead, M. J., A co-evolving decision tree classification method. *Expert System with Application*, 34(1), 18-25 (2008). doi:https://doi.org/10.1016/j.eswa.2006.08.008

[2] Biswas, A., Farid, D. M., & Rahman, C. M., A New Decision Tree Learning Approach forNovel Class Detection in Concept Drifting Data Stream Classification. *Journal of Computer Science and Engineering*, 14(1), 1-8 (2012).

[3] Chandra, B., & Gupta, M., Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data. *Expert Systems with Applications*, 38(3), 1293-1298 (2011). doi:https://doi.org/10.1016/j.eswa.2010.06.076.

[4] Chandra, B., & Varghese, P. P., Fuzzifying Gini Index based decision trees. *Expert Systems with Applications*, 36(4), 8549–8559 (2009). doi:https://doi.org/10.1016/j.eswa.2008.10.053.

[5] Chena, J., Huang, H., Tian, S., & Qu, Y., Feature selection for text classification with Naïve Bayes. *Expert Systems with Applications*, 36(3), 5432-5435 (2009). doi:https://doi.org/10.1016/j.eswa.2008.06.054.

[6] Farid, A. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R., Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4), 1937-1946 (2014). doi:https://doi.org/10.1016/j.eswa.2013.08.089.

[7] Farid, D. M., & Rahman, M. Z., Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm. *Journal of Computers*, 5(1), 23-31 (2010). doi:DOI:10.4304/jcp.5.1.23-31.

[8] Farid, D. M., Maruf, G. M., & Rahman, C. M., A new approach of Boosting using decision tree classifier for classifying noisy data. International Conference on Informatics, *Electronics & Vision* (ICIEV) (2013). doi:DOI: 10.1109/ICIEV.2013.6572718.

[9] Frank, A., & Asuncion, A., UCI ((University College London) Machine Learning Repository (2010). Retrieved from https://archive.ics.uci.edu/ml/datasets/Iris.

[10] Hechmi, J. M., Khlaifi, H., Amine, B., & Zrelli, A., Intrusion Detection Using Data Fusion and Machine Learning. 26th International Conference on Software, *Telecommunications and Computer Networks* (SoftCOM) (2018). doi:10.23919/SOFTCOM.2018.8555800.

[11] Hsu, C.-C., Huang, Y.-P., & Chang, K.-W., Extended Naive Bayes classifier for mixed data. *Expert Systems with Applications,* 35(3), 1080–1083 (2008). doi:https://doi.org/10.1016/j.eswa.2007.08.031.

[12] Kaklamanis, M. M., & Filippakis, M. E., A comparative survey of machine learning classification algorithms for breast cancer detection. PCI '19: Proceedings of the 23rd Pan-Hellenic Conference on Informatics, pp. 93-103 (2019). doi:https://doi.org/10.1145/3368640.3368642.

[13] Kemal, P., & Salih, G., A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multiclass classification problems. *Expert Systems with Applications,* 36(2), 1587-1592 (2009). doi:https://doi.org/10.1016/j.eswa.2007.11.051.

[14] Koc, L., Mazzuchi, T. A., & Sarkani, S., A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. *Expert Systems with Applications,* 39(18), 13492-13500 (2012). doi:https://doi.org/10.1016/j.eswa.2012.07.009.

[15] Kwok, J. T.-Y., Moderating the Outputs of Support Vector Machine Classifiers. *IEEE Transactions On Neural Networks*, 10(5), 1018-1031 (1999).

[16] Liberman, N., Decision Trees and Random Forests, Towards Data Science (2017).

[17] Mingo, L., Aslanyan, L., Castellanos, J., Díaz, M., & Riazanov, V., FOURIER NEURAL NETWORKS: AN APPROACH WITH SINUSOIDAL ACTIVATION. *International Journal "Information Theories & Applications,* 11, 52-55 (2006).

[18] Mugambi, E. M., Hunter, A., Oatleyd, G., & Kennedy, L., Polynomial-fuzzy decision tree structures for classifying medical data. Knowledge-Based Systems, 17(2-4), 81-87 (2004). doi:https://doi.org/10.1016/j.knosys.2004.03.003.

Nagarajan, P. R., Sandhiya, K., Vinod, V. M., & Mekala, V., Offline Navigation:GPS based assisting system in Sathuragiri forests using Machine Learning. International Conference on Intelligent Computing and Communication for Smart World (I2C2SW) (2018). doi:10.1109/I2C2SW45816.2018.8997523.

[19] Polat, K., & Güneş, S., The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-NN based weighted pre-processing methods to diagnosis of erythemato-squamous diseases. *Digital Signal Processing*, 16(6), 922-930 (2006). doi:https://doi.org/10.1016/j.dsp.2006.04.007.

[20] Ram, N. P., Sandhiya, K., Vinod, V. M., & Mekala, V., Offline Navigation:GPS based assisting system in Sathuragiri forests using Machine Learning. International Conference on Intelligent Computing and Communication for Smart World (I2C2SW), pp. 326-331 (2018). doi:10.1109/I2C2SW45816.2018.8997523.

[21] Sabbirantor. K means clustering | K Means ++ (Slideshow) (2019).

[22] Siswantoro, J., Prabuwono, A. S., Abdullah, A., & Idrus, B. A linear model based on Kalman filter for improving neural network classification performance. *Expert Systems with Applications*, 112-122 (2016). doi:https://doi.org/10.1016/j.eswa.2015.12.012.

[23] Tolson, E., Machine Learning in the Area of Image. Advanced Undergraduate Project Spring 2001, 1-35 (2001).

[24] Tuovinen, T. S., Real-time classification of SMEs credit and risk ratings and the impact of financial indicators and payment behaviour. Bachelor's Thesis, University of Applied Sciences (2020).

[25] Jyothi Varanasi & M. M. Tripathi. K-means clustering based photo voltaic power forecasting using artificial neural network, particle swarm optimization and support vector regression, *Journal of Information and Optimization Sciences*, 40:2, 309-328 (2019). DOI: 10.1080/02522667.2019.1578091.

[26] Satvik Vats & B. B. Sagar, Performance evaluation of K-means clustering on Hadoop infrastructure, *Journal of Discrete Mathematical Sciences and Cryptography*, 22:8, 1349-1363 (2019), DOI: 10.1080/09720529.2019.1692444.

Received December, 2020

Revised September, 2021