PAPER NAME

1 Prosiding Hartono IOP.pdf

AUTHOR

Hartono

WORD COUNT

3501 Words

CHARACTER COUNT

17511 Characters

PAGE COUNT

10 Pages

FILE SIZE

1.3MB

SUBMISSION DATE

Feb 23, 2024 12:45 AM GMT+7

REPORT DATE

Feb 23, 2024 12:56 AM GMT+7

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 0% Publications database
- Crossref Posted Content database
- 18% Submitted Works database

Excluded from Similarity Report

- Internet database
- Crossref database
- Bibliographic material
- Small Matches (Less then 17 words)

Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem

To cite this article: Hartono *et al* 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* **288** 012075

Optimization Model of K-Means Clustering Using Artificial Neural Networks to Handle Class Imbalance Problem

Hartono[1,2*], O S Sitompul[2], Tulus[3] and E B Nababan[2]

[1]Department of Computer Science, STMIK IBBI, Medan, Indonesia
[2]Department of Computer Science, University of Sumatera Utara, Medan, Indonesia
[3]Department of Mathematics, University of Sumatera Utara, Medan, Indonesia

*hartonoibbi@gmail.com

Abstract. Class imbalance is a situation where instances in one class much higher than instances in other classes. In clustering, this problem not only affects the accuracy of a prediction but also introduces bias in decision-making process. In this case, a machine learning technique will yield a good prediction accuracy from training data class with a large number of instances, but give a poor accuracy in classes with the small number of instances. In this research, we propose an approach for optimizing K-Means clustering in handling class imbalance problem. The approach uses the perceptron feed-forward neural network to determine coordinates of the centroid of a cluster in K-Means clustering processes. Data used in this research are datasets from the UCI Machine Learning Repository. From the experimental results obtained, the proposed approach could optimize the result of K-Means clustering in terms of minimizing class imbalance.

1. Introduction

Class imbalance is a situation where a number of instances in at least one class much higher than the number instances in other classes. This situation creates a majority versus minority class problem. In some particular cases, such as anomaly detection of computer network access patterns, the focus of interest is given more to the minority class since it may contain unusual behaviour differs from the general access patterns [1]. However, in other cases, such as clustering, the focus would be placed on the majority class because it could greatly affect the accuracy of a prediction. A machine learning technique will yield a good prediction accuracy from training data class with a large number of instances, but give a poor accuracy in classes with the small number of instances [2]. In addition, the class imbalance problem also introduces bias in decision-making process as this situation tends to skew rigorously toward the majority class [3].

The focus of this research is to determine the coordinates of the centroid of a cluster in a K-Means clustering process and to analyze its effect to class imbalance. In determining the centroid in the K-Means clustering we proposed the use of perceptron feed-forward neural network. As a supervised learning algorithm, this method is well-known at handling complex pattern recognition. The Various approach had been proposed to minimize the mean square error (MSE) from a simple gradient descent algorithm to more complex distributed autonomous neuro-gen learning engine with distributed adaptive neural network as one of its component [4, 5, 6].

The rest of this paper is organized as follows. In Section 2 we will provide related works concerning the determination of centroid for clustering process as well as the class imbalance problems. In Section

3 we describe the methodology used in this research and in Section 4 we provide the experimental process performed in this research. Results and discussion are given in Section 5 and finally, we conclude the research in Section 6.

2. Related Works

In the k-means algorithm, deciding the number of clusters and determining the centroid for each cluster are always cumbersome and important tasks since these tasks could directly affect the quality of the resulted clusters. Research works on the determination of the centroid based on fuzzy approach had been done in [7], however, the fuzzy approach has some disadvantages in terms of accuracy for large-size datasets. An approach using new cost function and distance measure based on co-occurrence of values had also been proposed [8]. The intended results are to overcome the limitation of k-means in dealing with numeric data, whereby a modified description of cluster center was presented. Another approach using Fuzzy C-Means (FCM) had been proposed in [9]. The clustering results obtained are integrated into a judgment matrix, which is then iteratively partitioned to identify the desired cluster number and the final result.

Further research using neural networks approach had been proposed in [10] using a modified back-propagation in order to accelerate the learning rate for two-class imbalance problems. This consists of calculating a direction in weight-space which decreases the same amount of errors for each majority and minority classes. Later, [11] corrected errors in the placement of clustering class, despite the correction could overcome the class imbalance. The modified learning algorithm has been proposed for dealing with this problem. A Modified Back Propagation had been proposed in [12] to avoid the ignoring of minority class in a training process in addition to accelerating the convergence of the neural network. More recent work can be found in [13] in which deep neural network training is implemented to handle imbalance datasets. This method equally captured classification errors from both majority class and minority class.

A novel technique with Under Sampled K-Means in handling class imbalance problem had been proposed in [14]. The technique intelligently removed noisy and weak instances from overwhelming (majority) class. However, as formally illustrated in [15], even though the input data for K-Means comes from various true cluster sizes, the resulting clusters are relatively uniform in their sizes. To further dealing with the uniform effect in K-Means, [16] used the Visual K-Means algorithm for skewed distributed data sources. They argued that the proposed algorithm could effectively handle the imbalanced datasets.

3. Methods

The data used in this research are the Balanced Scale and Abalone Datasets from the UCI Machine Learning Repository. In general, the methodology consists of two stages: preparation stage and the clustering stage. In the preparation stage, the dataset is trained by a perceptron neural network to generate *k* initial cluster centers (centroids) in which for the Balanced Scale dataset there are three classes, namely Class B, Class L, and Class R while in the Abalone dataset the clusters are Sex F, Sex I, and Sex M. In this research, the generated *k* centroids are used as the initial cluster centers for the k-means clustering. The general architecture of the proposed method used is depicted in Fig. 1.

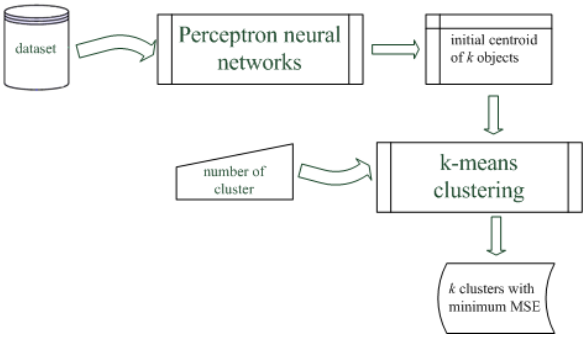

Figure 1. The general architecture

The modification of K-means clustering algorithm in determining centroid using perceptron are shown as follows.

Input: The Number of Cluster *k* and a database containing *n* objects
Output: A set of *k* clusters which minimize the squared-error criterion
Method:

Step 1. Calculate the initial cluster centers of *k* objects using perceptron.
Step 2. Repeat until convergence {
Step 3. (Re) assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster.
Step 4. Update the cluster means, i.e., calculate the mean value of the objects for each cluster.
Step 5. }

As for the original K-Means Algorithm and the modified K-Means Algorithm using Perceptron are shown in Fig. 2 and 3.

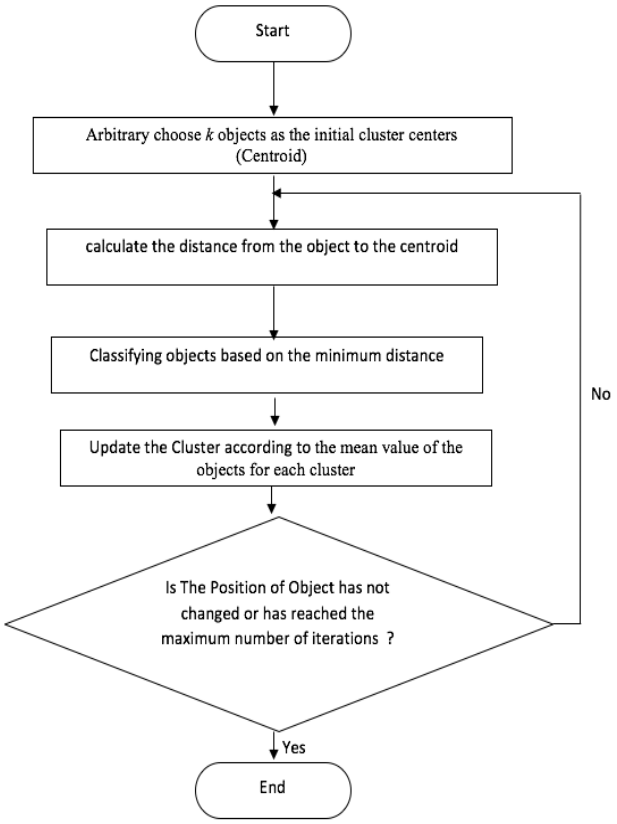

Figure 2. The original K-Means algorithm

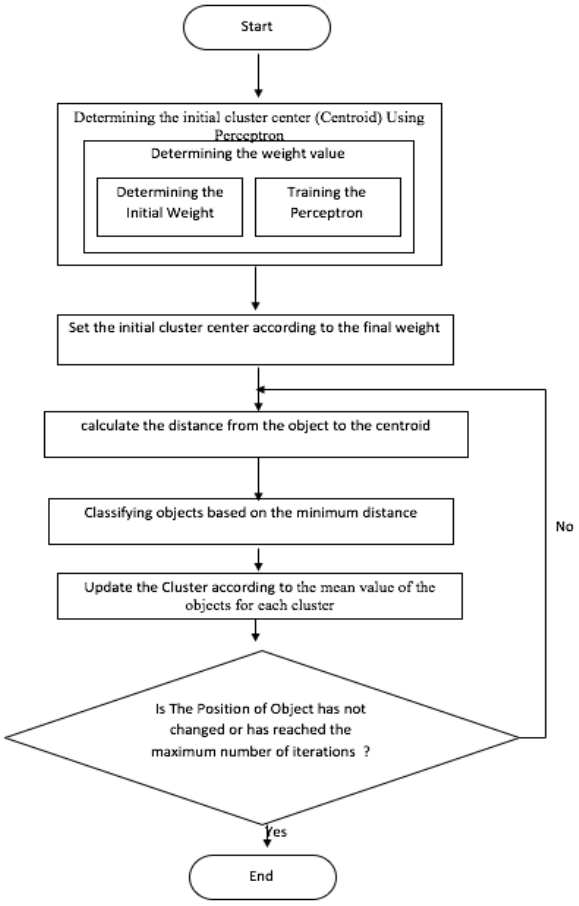

Figure 3. The modified K-Means algorithm using Perceptron

The algorithm of calculating the initial cluster centers (centroid) of *k* objects using perceptron are shown as follows.

Step 1. Initialize weight and bias
Set learning rate
Set *input* according to dataset

Set *target* according to cluster of data from dataset
Step 2. While stopping condition is false, do steps 3-6
Step 3. For each training pair s:t, do steps 4-5
Step 4. Set activations of input units:

$$x_i = s_i$$

Step 5. Compute response of output unit:

$$y\_in = b + \sum_i x_i w_i$$

$$y = \begin{cases} 1, if\ y\_in > 0 \\ 0, if\ -0 \le y\_in \le 0 \\ -1, y\_in < -0 \end{cases}$$

Step 6. Update weights and bias if an error occurred for this pattern.

If $y \neq t$

$w_i\,(new) = w_i\,(old) + \alpha.t.xi$

$b\,(new) = b\,(old) + \alpha.t$

else

$w_i\,(new) = w_i\,(old)$

$b\,(new) = b\,(old)$

Step 7. Test stopping condition
if no weights changed in Step 3, stop; else, continue Step 2.

4. Experimental Process

4.1. Dataset Description

The dataset used in this experiment are the Balanced Scale and Abalone datasets from the UCI Machine Learning Repository.

4.2. Determining the Number of Clusters

The number of clusters K in this study is 3 in accordance with the classes in the Balanced Scale dataset and the Abalone dataset. There are 4 (four) attributes in Balanced Scale dataset: Left-Weight, Left-Distance, Right-Weight, and Right-Distance and the Abalone dataset has 8 (eight) attributes: Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight, and Rings.

4.3. Determining the Centroid of K-Means

Determining the centroid of K-Means using Artificial Neural Network can be divided into three steps: determining the centroid for each cluster, and each step has a step as follows (in this section, the example is given for Cluster Class B in Balanced Scale Dataset)

Step 1. Initialize weight and bias
Set learning rate
Set *input* according to dataset
Set *target* according to cluster of data from dataset
Step 2. Step 1-6. Training process for determining the centroid for Class B

The Training process of determining the centroid for Class B can be seen in Table 1.

Table 1. Training process in determining the centroid of Class B (Epoch 1)

Input					Weight						Final Weight			
X1	X2	X3	X4	Target	W1	W2	W3	W4	Actual	Error	W1	W2	W3	W4
1	1	1	1	1	0	0	0	0	0	1	0.1	0.1	0.1	0.1
1	2	1	2	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
1	3	1	3	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
1	4	1	4	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
2	3	2	3	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
3	4	4	3	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
4	1	1	4	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
3	5	5	3	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
5	2	2	5	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1
5	5	5	5	1	0.1	0.1	0.1	0.1	1	0	0.1	0.1	0.1	0.1

The maximum number of epochs in this study is 200. The training process were done using the R Language. Fig. 4, Fig. 5, and Fig. 6 are the training process in determining the centroid of the Class B, Class L, and Class R. Fig. 7 is the result of Centroid from training process.

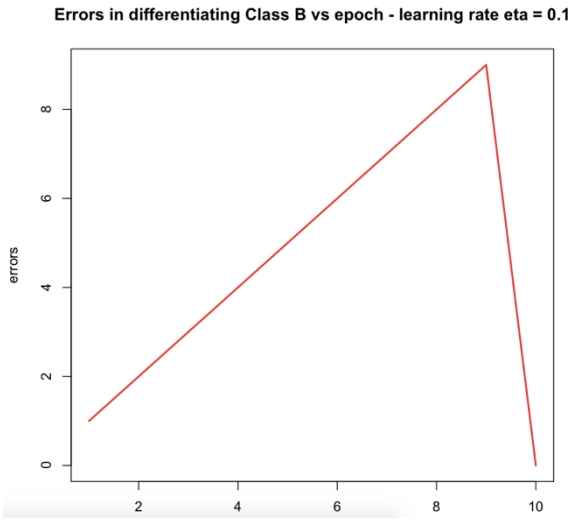

Figure 4. Training Process in Determining Class B

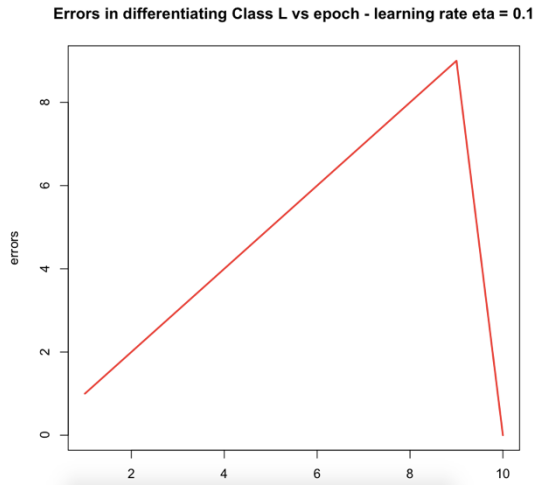

Figure 5. Training Process in Determining Class L

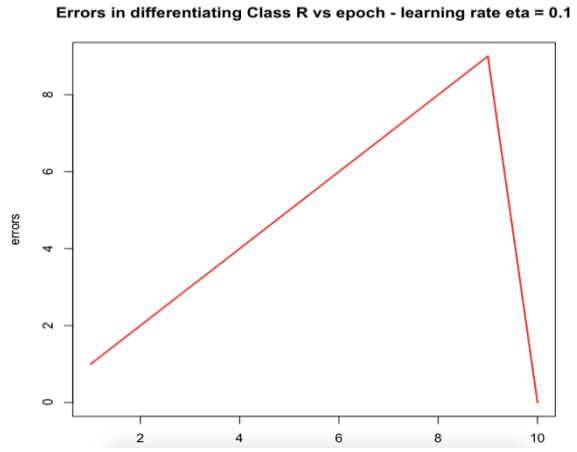

Figure 6. Training Process in Determining Class R

```
[1] "Centroid Left Weight for Class B is "
[1] 0.9
[1] "Centroid Left Distance for Class B is "
[1] 3.2
[1] "Centroid Right Weight for Class B is "
[1] 2
[1] "Centroid Right Distance for Class B is "
[1] 2
[1] "Centroid Left Weight for Class L is "
[1] 0.9
[1] "Centroid Left Distance for Class L is "
[1] 3.4
[1] "Centroid Right Weight for Class L is "
[1] 1.3
[1] "Centroid Right Distance for Class L is "
[1] 1.3
[1] "Centroid Left Weight for Class R is "
[1] 0.9
[1] "Centroid Left Distance for Class R is "
[1] 0.9
[1] "Centroid Right Weight for Class R is "
[1] 1.6
[1] "Centroid Right Distance for Class R is "
[1] 4.5
```

Figure 7. The Result of Centroid from Training Process

From the Fig. 4, 5, and 6 can be seen that the training process get the minimum error in epoch 10, actually, the maximum number of epochs is 200. The Fig. 7 show us the Centroid of Left-Weight, Left-Distance, Right-Weight, and Right-Distance in determining Class B, Class L, and Class R.

The training process were done using the R Language. Fig. 8, Fig. 9, and Fig. 10 are the training process in determining the centroid of the Abalone Class Sex F, Abalone Class Sex I, and Abalone Class Sex M. Fig. 11 is the result of Centroid from training process.

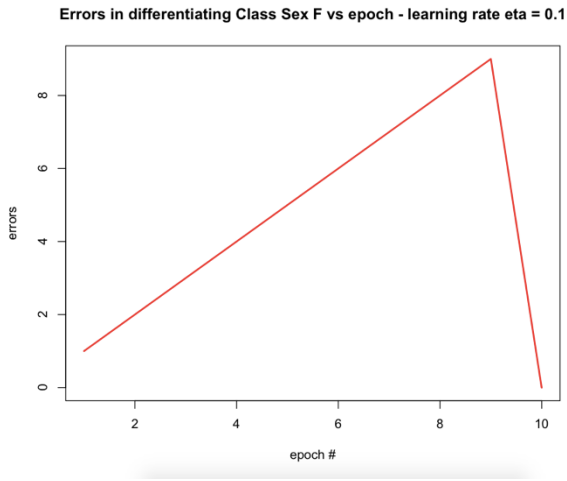

Figure 8. Training Process in Determining Abalone Sex F

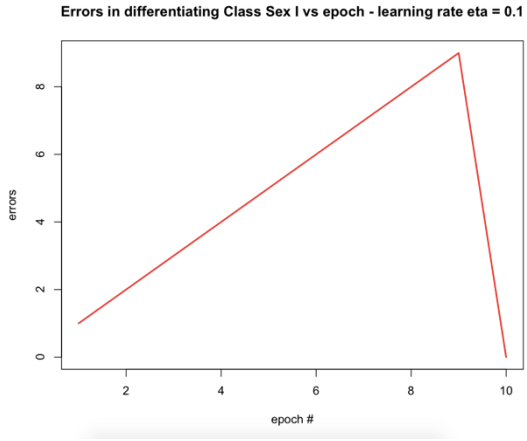

Figure 9. Training Process in Determining Abalone Sex I

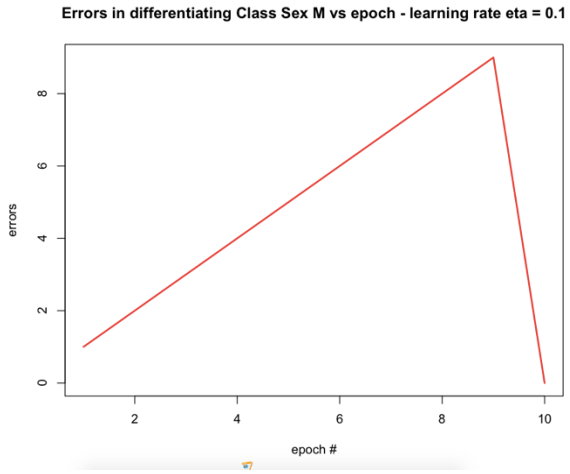

Figure 10. Training Process in Determining Abalone Sex M

```
[1] "Centroid Length for Sex F is="
[1] 0.53
[1] "Centroid Diameter for Sex F is="
[1] 0.415
[1] "Centroid Height for Sex F is="
[1] 0.15
[1] "Centroid Whole Weight for Sex F is="
[1] 0.7775
[1] "Centroid Shucked Weight for Sex F is="
[1] 0.237
[1] "Centroid Viscera Weight for Sex F is="
[1] 0.1415
[1] "Centroid Shell Wight for Sex F is="
[1] 0.33
[1] "Centroid Rings for Sex F is="
[1] 20
[1] "Centroid Length for Sex I is="
[1] 0.425
[1] "Centroid Diameter for Sex I is="
[1] 0.3
[1] "Centroid Height for Sex I is="
[1] 0.095
[1] "Centroid Whole Weight for Sex I is="
[1] 0.3515
[1] "Centroid Shucked Weight for Sex I is="
[1] 0.141
[1] "Centroid Viscera Weight for Sex I is="
[1] 0.0775
[1] "Centroid Shell Wight for Sex I is="
[1] 0.12
[1] "Centroid Rings for Sex I is="
[1] 8
[1] "Centroid Length for Sex M is="
[1] 0.35
[1] "Centroid Diameter for Sex M is="
[1] 0.265
[1] "Centroid Height for Sex M is="
[1] 0.09
[1] "Centroid Whole Weight for Sex M is="
[1] 0.2255
[1] "Centroid Shucked Weight for Sex M is="
[1] 0.0995
[1] "Centroid Viscera Weight for Sex M is="
[1] 0.0485
[1] "Centroid Shell Wight for Sex M is="
[1] 0.07
[1] "Centroid Rings for Sex M is="
[1] 7
```

Figure 11. The Result of Centroid from Training Process

5. Results and discussion

Results of the original K-Means clustering for Balanced Scale and Abalone datasets, can be seen in Table 2.

Table 2. Result of Balanced Scale and Abalone dataset with original K-Means clustering

Testing Number	Number of Error		Number of Data in Class 1		Number of Data in Class 2		Number of Data in Class 3	
	Balanced Scale	Abalone	Balanced Scale	Abalone	Balanced Scale	Abalone	Balanced Scale	Abalone
1	383	2430	276	910	196	3073	153	194
2	435	2499	200	824	149	3217	276	136
3	480	3262	281	2722	145	1194	199	261
4	493	2488	152	2281	201	449	272	1447
5	481	2486	266	194	194	3212	165	771
6	383	2937	155	2722	196	261	274	1194
8	401	3262	211	2722	135	1194	279	261
9	396	3331	165	2095	256	1592	204	490
10	403	2771	244	1194	225	261	156	2722
Average	385.5	2546.6						

From Table 2, it can be seen that in both Balanced Scale and Abalone datasets clustering there are class imbalance problem. The number of Data in Class 1, Class 2, and Class 3 in Balanced Scale and Abalone datasets can be very different in Size. As the result of this class imbalance, we can see that the number of error that occur is quite large. The average of error in Balanced Scale is 385.5 (61.68%) and the average of error in Abalone is 2546.6 (60.96%)

Results of the clustering using the optimization model of K-Means in determining the centroid of both Balanced Scale and Abalone datasets we can see the in Table 3.

Table 3. Results of Balanced Scale and Abalone datasets with the optimization model of K-Means clustering

Dataset	Number of Error	Numberr of Data in Class 1	Number of Data in Class 2	Number of Data in Class 3
Balanced Scale	289	202	186	237
Abalone	2166	1182	2035	960

From the Table 3, it can be seen that the optimization model of K-Means Clustering can reduce the number of Errors. For Balanced Scale dataset the number of errors were decreased to 289 or about 46.24%., whereas clustering result of Abalone dataset has reduced the number of errors to 2166 or about 51.86%.

By comparing the results of Table 2 and Table 3, it can be seen that the number of errors of both Balanced Scale and Abalone datasets has been reduced using the optimization model of K-Means clustering and it has a better performance than the result of [17]. It gave the number of error of 50% using K-Modes in Balanced Scale dataset and in Abalone dataset has the slightly below performance with another result from [18], that the result of [18] gave the number of error of 38.22 using modified K-Means but still has a better performance than the result of [18] that using Hierarchical Clustering that gave the number of error of 93.77%.

6. Conclusion

The conclusion of this research are as follows. First, the Artificial Neural Network can determine the centroid of K-Means Clustering, it was indicated by decreasing the number of errors. Second, it is confirmed that the Optimization Model of K-Means Clustering using Artificial Neural Network also can handle the class imbalance problem.

Our case study is using numerical datasets and in the future, it should be another study on non-numerical datasets and the average number of error can be reduced using more iteration. The importance of this research for future studies is that the results indicate that proper determination of the centroid can reduce the occurrence of class imbalance.

References

[1] A Erhahman S M and Abraham A 2014 A Review of Class Imbalance Problem *J. of Network and Innovative Computing* **1** 332-340

[2] Galar M, Fernandez A, Barrenechea E and Bustince H 2012 A Review on Ensembles for the Class Imbalance Problem: Bagging, Boosting, and Hybrid-Based Approachs *IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews* **42** 463-484

[3] Ertekin S, Huang J and Giles C L 2007 Active Learning for Class Imbalance Problem *Proceedings of the 30th annual international ACM SIGIR conference on Research and Development in Information Retrieval* 823-824

[4] Riedmiller M 1994 Advanced Supervised Learning in Multi-Layer Perceptrons - From Backpropagation to Adaptive Learning Algorithms *J. Computer Standards & Interfaces* **1** 265-278

[5] Calvert D and Guan J 2005 Distributed Artificial Neural Network Architectures *The 19th International Symposium on High Performance Computing Systems and Applications (HCPS)* 2-10

[6] Aaron, Sitompul O S and Rahmat R F 2014 Distributed Autonomous Neuro-Gen Learning Engine for Content-Based Document File Type Identification *International Conference on Cyber and IT Service Management (CITSM)* 63-68

[7] Rahman M A and Islam M Z 2012 CRUDAW: A Novel Fuzzy Technique for Clustering Records Following User Define Attribute Weights *Proceedings of the Tenth Australasian Data Mining Conference (AusDM)* 27-41
[8] Ahmad A and Dey L 2007 A K-Means Clustering Algorithm for Mixed Numeric and Categorical Data *Data & Knowledge Engineering* **63** 503-527
[9] Mok P Y, Huang H Q, Kwok Y L and Au J S 2012 A Robust Adaptive Clustering Analysis Method for Automatic Identification of Clusters *Pattern Recogn* **45** 3017-3033
[10] Zhou Z H and Liu X Y 2006 Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem *IEEE Trans. Knowledge Data Eng.* **18** 63-77
[11] Anand R, Mehrotra K, Mohan C and Ranka S 1993 An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets *IEEE Trans. Neural Networks* **4** 962-969
[12] Alejo R, Valdovinos R M, Garcia V and Sanchez J H P 2013 A Hybrid Method tof Face Class Overlap and Class Imbalance on Neural Networks and Multi-Class Scenarios *Pattern Recognition Letters* **34** 380-388
[13] Wang S, Liu W, Cao L M Q and Kennedy P J 2016 Training Deep Neural Networks on Imbalanced Data Sets *International Joint Conference on Neural Networks (IJCNN)* 4368-4374
[14] Kumar N S, Rao K N, Govardhan A, Reddy K S and Mahmood A M 2014 Undersampled K-Means Approach for Handling Imbalanced Distributed Data *Progress in Artificial Intelligence* **3** 29-38
[15] Wu J 2012 The Uniform Effect of K-Means Clustering in *Advances in K-Means Clustering: A Data Mining Thinking* ed Wu J (Berlin Heidelberg: Springer Theses) **2** 17-35
[16] Kumar N S, Rao K N and Govardhan A 2015 Visual K-Means Approach for Handling Class Imbalance Learning *Proceedings of the Second International Conference on Computer and Communication Technologies* 389-396
[17] Aranganayagi S and Thangavel K 2009 Improve K-Modes for Categorical Clustering using Weighted Dissimilarity Measure *World Academy of Science, Engineering and Technology* 991-997
[18] Mayukh H 2010 Age of Abalones using Physical Characteristics: A Classification Problem *ECE 539 Fall 2010 Project Report University of Wisconsin-Madison* 1-4

18% Overall Similarity

Top sources found in the following databases:

- 0% Publications database
- Crossref Posted Content database
- 18% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	**Universitas Negeri Surabaya The State University of Surabaya on 2020...** Submitted works	8%
2	**Higher Education Commission Pakistan on 2019-08-04** Submitted works	3%
3	**UIN Sunan Gunung DJati Bandung on 2018-08-06** Submitted works	2%
4	**Marmara University on 2020-06-22** Submitted works	2%
5	**Universiti Teknologi MARA on 2021-07-27** Submitted works	1%
6	**Universiti Teknologi MARA on 2020-11-24** Submitted works	1%
7	**University of Sunderland on 2007-02-15** Submitted works	<1%
8	**Universitas Diponegoro on 2018-08-18** Submitted works	<1%
9	**Institute of Technology, Nirma University on 2016-05-10** Submitted works	<1%