PERFORMANCE COMPARISON OF ADAPTIVE NEURO FUZZY INFERENCE SYSTEM AND SUPPORT VECTOR MACHINE ALGORITHM IN BALANCED AND UNBALANCED MULTICLASS DATA CLASSIFICATION

Muhammad Irfan Saputra[1], Irwan Budiman[2], Dwi Kartini[3], Dodon Turianto Nugrahadi[4], Mohammad Reza Faisal[5]
[12345]Ilmu Komputer FMIPA ULM
A. Yani St. KM 36 Banjarbaru, South Kalimantan
Email: m.irfansputra@gmail.com

Abstract

Data is a record collection of facts. At first the data in the real world were largely unbalanced. Although, the existence of data on fewer categories is much more important to know data on more categories. However, there are some balanced data. This balanced data is the possibility of a ratio of 1:1 in which, the data in the dataset is the same. In this study, using the ANFIS algorithm and SVM to see affected performance on balanced and imbalanced data with multiclass. Data is taken from the UCI Machine Learning. From the research conducted, it is known that the SVM method on the Wine dataset has an accuracy of 96.6% and the ANFIS method on the Iris dataset has an accuracy of 94.7%.

Keywords*: ANFIS, SVM, Balanced Data, Imbalanced Data, Multiclass.*

1. INTRODUCTION

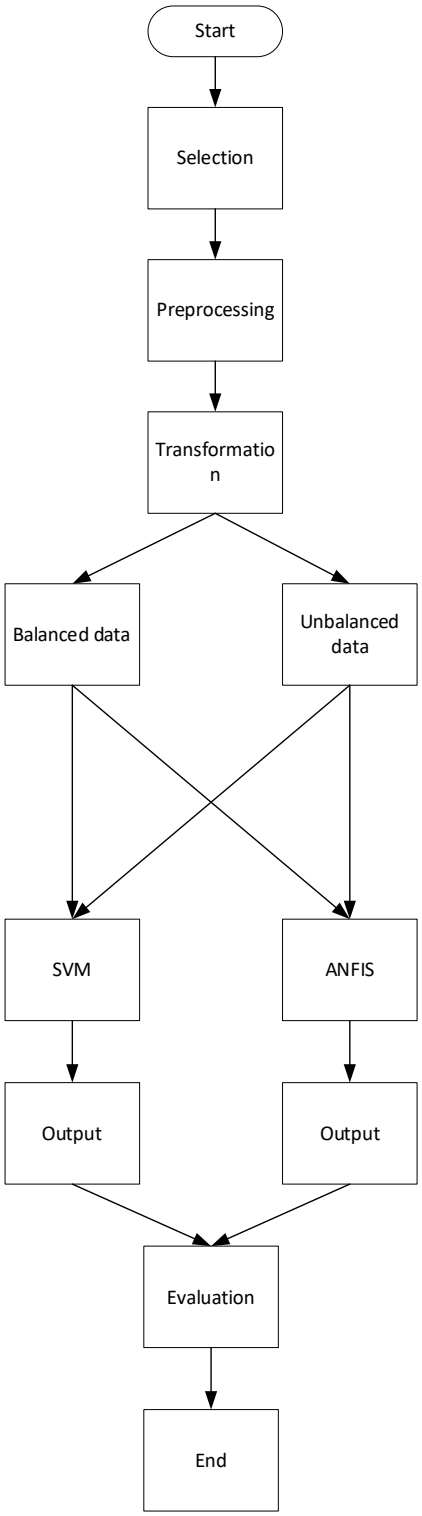
Data is a record of a collection of facts. In its use, data means statements that are accepted as they are. The result of the data is the formation in the form of numbers, words, or images/pictures.

Unbalanced data or imbalanced class is a condition of unbalanced data class division, the number of data classes (instances) is one less or more than the number of other data classes. The data class that has fewer data is called the minority group (minority), while the more data class group is called the majority group [5]. However, there are some data that are balanced. This balanced data usually has a ratio of 1:1 in which the data in the dataset is the same. For example, for class a there are 60 data and class b also has 60 data.

Unbalanced data conditions can make it difficult for classification methods to process data mining. Class imbalance in the data has a bad effect on the classification performance where the minority class is sometimes misclassified as the majority class. In some cases, minority classes are more important to identify than other classes [5].

2. RESEARCH METHODOLOGY

The procedure of this research is shown in Figure 1.



The following is the flow of this research:

a. Selection
   This stage consists of making data collection. Data used for research obtained from UCI Machine Learning, Iris for balanced data and Wine for unbalanced data.
b. Preprocessing
   This stage consists of data cleaning and initial processing to obtain data that is consistent. At

this stage the data is converted into a suitable form so that it can be done data mining process using the 10-fold Cross-Validation method to divide training data and testing data.

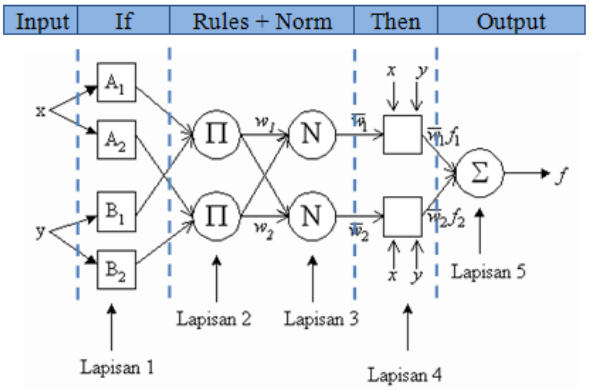
c. Transformation
This stage usually consists of selecting the attributes that will be used in the data mining stage.

d. Data Mining
This stage consists of searching for patterns of interest in certain forms of representation, depending on the goals (usually predictions). Research that will carried out consists of testing the classification method, namely the SVM and ANFIS to get the values of accuracy, sensitivity, and specificity.

e. Interpretation/Evaluation
This stage consists of explaining and evaluating the knowledge obtained. The knowledge gained will be analyzed and explained, such as the value of accuracy, sensitivity, and the specificity generated from the data mining stage using the confusion matrix.

2.1. Adaptive Neuro-Fuzzy Inference System (ANFIS)

ANFIS is a combination of Adaptive Neural Network and Fuzzy Logic. ANFIS has the same architecture as Fuzzy Sugeno systematically. Steps in the method Sugeno in his inference that is fuzzyfication, formation of fuzzy knowledge base, machine inference, and defuzzification. ANFIS method is explained in Figure 2.



2.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a related algorithm for a method learning, for both classification and regression problems. By being oriented towards task, power, and calculations are easy to do, SVM is a successful method and is considered as the current state-of-the-art classifier. Data classes represented as circles and points outlined for decision making. After observing, there are many decisions which can be used to separate the two groups of data [4].

2.3. Cross Validation

Cross Validation method is used as a performance analysis to ensure credibility the result of the prediction. Cross Validation process consists of randomly dividing the dataset into k-part. One part is used to validate the model and the rest is used to perform classification process to train training data. This process is repeated k-times with a selection of subsets different validations [3].

2.4. Confusion Matrix

In this research, Confusion Matrix is used to measure accuracy, specificity, and sensitivity for each dataset used. Confusion Matrix is represented by a table that states the amount of testing data that is correctly classified and the amount of testing data misclassified [6].

Table 1 Confusion Matrix

Classification		Observed Class	
		Class = Yes	Class = No
Predicted Class	Class = Yes	true positive-TP	false positive-FP
	Class = No	false negative-FN	true negative-TN

Based on the Confusion Matrix table above:

1. True positive is the number of positive data classified as positive values.
2. False positive is the number of negative data classified as positive values.
3. False negative is the number of negative data classified as positive values.
4. True negative is the number of negative data classified as negative values.

Prediction accuracy is measured using the accuracy formula as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Sensitivity is used to measure the proportion of true positive correctly identified. The sensitivity formula is as follows:

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

Meanwhile, specificity is used to measure the proportion of true negative correctly identified. The specificity formula is as follows:

$$\text{specificity} = \frac{TN}{TN + FP}$$

3. RESULTS AND DISCUSSIONS

In testing the ANFIS and SVM. The Iris dataset has 3 classes and has 5 columns, while the Wine dataset has 3 classes has a total of 3 classes and has 10 columns, because it uses the 10-fold Cross Validation method means that the entire data is divided into 10 parts where one part is used to validate the model and the remainder is used to carry out the classification process for training data.

The following are the classification results from the ANFIS and SVM tested by Confusion Matrix method.

Table 2 SVM classification results

Classification	SVM					
	Dataset Iris			Dataset Wine		
Class	Iris-setosa	Iris-versicolour	Iris-Virginica	1	2	3
Sensitivity	0,98	0,96	0,92	0,95	0,97	0,98
Specificity	1,0	0,96	0,97	0,99	0,96	0,99
Accuracy		95,3%			96,6%	

Table 3 ANFIS classification results

Classification	ANFIS					
	Dataset Iris			Dataset Wine		
Class	Iris-setosa	Iris-versicolour	Iris-Virginica	1	2	3
Sensitivity	1,0	0,88	0,96	0,95	0,92	0,90
Specificity	1,0	0,98	0,94	0,97	0,93	0,99
Accuracy	94,7%			92,1%		

4. CONCLUSION

The conclusion of this research is the highest accuracy value in SVM method is 96,6% on the Wine dataset and in the ANFIS method, the highest accuracy value is obtained from the Iris dataset is 94,7%.

The highest sensitivity value is obtained by Class Iris-setosa from Iris dataset and Class 3 from Wine dataset with the SVM method of 0,98. In the ANFIS, the highest sensitivity value is obtained in the Class Iris-setosa from Iris dataset with with a sensitivity value of 1,0.

The highest specificity value in the SVM method is obtained by Class Iris-setosa from Iris dataset with a specificity value of 1.0. In the ANFIS, the highest specificity value is obtained in the Class Iris-setosa from Iris dataset of 1.0.

REFERENCES

[1] Armayani, C., Fauzi, A. & Sembiring, H., 2021. Implementasi Data Mining Pengelompokan Jumlah Data Produktivitas Ubinan Tanaman Pangan Berdasarkan Jenis Ubinan dengan Metode Clustering Dikab Langkat (Studi Kasus: Badan Pusat Statistik Langkat). Jurnal Informatika Kaputama (JIK), 5(1), pp.185-196.

[2] Indriani, A., 2014. Klasifikasi Data Forum dengan Menggunakan Metode Naive Bayes Classifier. Seminar Nasional Aplikasi Teknologi Informasi (SNATI) (Vol. 1, No. 1).

[3] Khammassi, C. & Krichen, S., 2017. A GA-LR Wrapper Approach For Feature Selection In Network Intrusion Detection. Computers & Security, 70, pp.255-277.

[4] Octaviani, P.A., Wilandari, Y., & Ispriyanti, D., 2014. Penerapan Metode Klasifikasi Support Vector Machine (SVM) pada Data Akreditasi Sekolah Dasar (SD) di Kabupaten Magelang. Jurnal Gaussian, 3(4), pp.811-820.

[5] Siringoringo, R., 2018. Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan k-Nearest Neighbor. Jurnal ISD. **3**: 44-49.

[6] Suwarno. 2016. Penerapan Algoritme Bayesian Regularization Backpropagation untuk Memprediksi Penyakit Diabetes. Jurnal MIPA. 32: 150-158s.