http://infor.seaninstitute.org/index.php/infokum/index
JURNAL INFOKUM, Volume 10, No.1, Desember 2021 ISSN : 2302-9706

ENHANCE THE ACCURACY OF K-NEAREST NEIGHBOR (K-NN) FOR UNBALANCED CLASS DATA USING SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE) AND GAIN RATIO (GR)

Khairul Umam Syaliman
Program Studi Teknik Informatika, Politeknik Caltex Riau, Pekanbaru, Indonesia
khairul@pcr.ac.id

Article Info
Received : 10 November 2021
Revised : 01 December 2021
Accepted : 07 December 2021

Abstract

k-Nearest Neighbor (k-NN) has very good accuracy results on data with almost the same class distribution, but on the contrary for information whose class distribution is not the same, the accuracy of k-NN will generally be lower. In addition, k-NN does not separate information for each class, implying that each class has an equal influence in determining the new information class, so it is important to choose a class that generally applies to information before characterizing the class assignments process. To overcome this problem, we will propose a structure that uses the Synthetic Minority Oversampling Technique (SMOTE) strategy to address class distribution problems and Gain Ratio (GR) to perform attribute selection to generate a new dataset with a reasonable class spread and significant class information attributes. E-Coli and Glass Identification were among the datasets used in this review. For objective results, the 10-fold-cross validation method will be used as an evaluation method with k values 1 to 10. The results of the research prove that SMOTE and GR can increase the accuracy of the k-NN method, where the highest increase occurred in the Glass Identification dataset by a difference increase of 18.5%. The lowest increase in accuracy occurred in the E-Coli dataset with an increase of 11.4%. The overall proposed method has given the better performance, although the value of precision, recall, and F1-Score is not better than original k-NN when used in dataset E-Coli. To all datasets, an improvement from precision is 41.0%, recall is 43.4% and F1-Score is 41.5%.

Keywords: k-NN, SMOTE, Gain Ratio, Unbalanced Class Data

1. INTRODUCTION

k-Nearest Neighbor (k-NN) is a well-known Machine learning technique because it is simple to apply to a variety of problem spaces, is intuitive and clear, and has a surprising level of accuracy. However, because it is affected by unbalanced scattering of data classes, k-NN has low performance on unbalanced data classes. The difference in the number of events between classes isn't addressed in the same way. The k-NN is also a distance-based system, with each k-NN having the same effect on data attribute selection.

When a class or many classes become underrepresented, it is referred to as a minority class since its numbers are substantially lower than those of other classes. [1]. Many traditional machine learning

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

algorithms do poorly when it comes to forecasting minority groups. One method of addressing the class unbalance problem is to use resampling. Resampling is a technique for altering the minority class distribution. A well-known strategy for this problem is the Synthetic Minority Oversampling Technique (SMOTE). To fit the argument against the majority class, SMOTE focuses on other examples of the minority class [2].

Feature selection is a critical topic in the efficiency of machine learning systems. Feature selection is tasked with removing unimportant features so that they can contribute significantly by improving categorization results. [3]–[8]. Principal Component Analysis (PCA) and Gain Ratio are two methods for selecting features.

PCA is a statistical technique that uses as much original data as feasible to achieve feature selection. According to [9], because PCA is a linear combination of numerous factors, interpreting the results can be tricky at times. Gain Ratio, unlike PCA, selects features depending on how relevant they are to data classes. Up to the chosen threshold, the Gain Ratio will select characteristics that are highly relevant to the data class. According to [10], GR was able to give data with fewer features while maintaining intrinsic information, and it was able to improve the accuracy value of the k-NN approach.

Inaccurate results are caused by an unbalance of irrelevant data and features. To improve the performance of k-NN classification, we conducted experiments on datasets with diverse unbalance classes, such as e-Coli and Glass Identification. We'll use the SMOTE approach to resample the unbalanced data, which will then be followed by feature selection using GR to find new datasets. It is hoped that by combining the SMOTE and GR approaches, the k-NN method will be able to produce more dependable data sets and improve its accuracy.

2. Related Works

One of the issues in categorization is class unbalance learning. In comparison to other courses, some are significantly underrepresented. As a result of this unbalance, data dispersion is uneven. Because most machine learning approaches assume that data is evenly distributed, many machine learning methods are less effective, especially in predicting minority classes. When there is a class unbalance in the data, the classifier tends to favor the majority class, resulting in poor categorization of the minority class. Several studies have looked into ways to improve the k-NN method's performance on unbalanced data. Using the SMOTE technique, [11] examined attribute selection to improve classification performance and class unbalances concerns. For unbalanced data, the results of this study show that the under-sampling strategy is more effective than the SMOTE approach. [12] provides a new strategy using the SMOTE method to balance the categorization training data. The proposed method artificially generates new sample data points from the original data sample, and it has been demonstrated that the k-NN algorithm's classification performance improves. [13] Compares the performance of the K-Nearest Neighbors algorithm with SMOTE to the K-Nearest Neighbors technique without SMOTE. In the situation of unbalanced data sets, the Accuracy value provided by SMOTE is better than the Accuracy value produced without SMOTE. [14] used a data pre-processing strategy that improved the performance of k-NN classifiers in uneven data stacks by balancing training data. The results show that applying multiple methodologies to training data, including random undersampling, random oversampling, and ensemble oversampling, can increase the precision performance of the k-NN classifier on unbalanced data.

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

3. Methods

2.1 Synthetic Minority Oversampling Technique (SMOTE)

In his publication "SMOTE: Synthetic Minority Oversampling Approach," NV Chawla, et al. proposed the synthetic Minority Oversampling (SMOTE) technique for the first time in 2002. SMOTE is a resampling approach that extracts and creates new data (synthetic data) depending on the value of data from minority classes to achieve a balanced distribution of data amongst classes. The SMOTE technique favors oversampling with the k-NN approach since most oversampling methods work on the premise of replicating fresh data at random.

Synthetic data processing on data with numeric feature values differs from data with categorical feature values, according to [15]. The similarity of numerical data to the Euclidean equation is measured, whereas categorical data is measured using the Value Difference Metric (VDM) formula, which is:

$$d(V_1, V_2) = \sum_{i=1}^{N} \left| \frac{C_{1i}}{C_1} - \frac{C_{2i}}{C_2} \right| \quad (1)$$

2.2 Gain Ratio (GR)

Gain Ratio (GR) is a method of modifying Gain Information by removing bias. In the Decision Tree approach, GR is the deciding parameter in picking the most relevant characteristics to the outcomes by analyzing the intrinsic information from the data. [10] Describe the processes in the GR calculation as follows:

Step-1: Calculate the entropy value using the equation (1)

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * log_2 pi \quad (2)$$

Step-2: Calculate the gain information with the equation (2)

$$Information\ Gain\ (S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \quad (3)$$

Step-3: Calculate split info with equation (3)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{D_j}{D}) \quad (4)$$

Step-4: Calculate GR with equations (4)

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{SplitInfo(A)} \quad (5)$$

2.3 k-Nearest Neighbor (KNN)

The supervised technique or method k-Nearest Neighbor (k-NN) can be used to categorize data. k-NN is one of the most prominent Machine Learning techniques in fact, [16] lists it among the top 10 data mining algorithms in his book. The distance-based algorithm k-NN is another example of a distance-based algorithm. Distance-Based Algorithms identify data similarity based on data distance and create the majority class a class for new data [17]–[21]. Because it can be used in a variety of application areas and is easy to develop, intuitive, and simple, k-NN may produce reasonably acceptable classification results [19], [22].

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

3. Results and Discussion

In this section, we will establish the foundation for this study; in general, this research will combine the SMOTE approach for dealing with unbalanced data and the Gain Ratio for selecting significant characteristics, as shown in Figure 1.

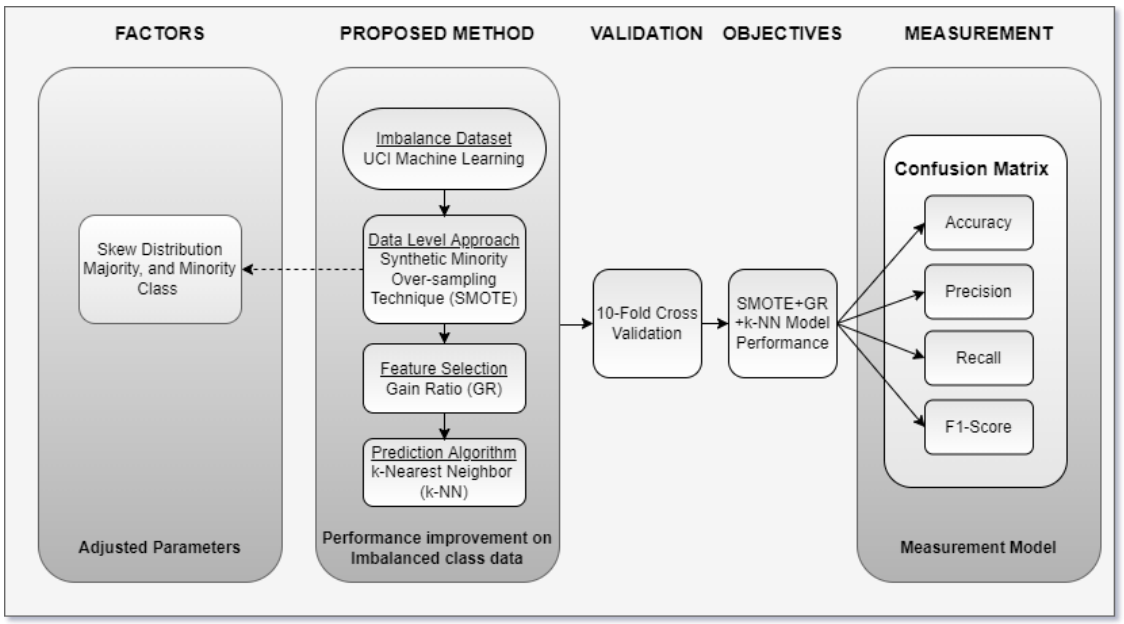

Figure 1. The Proposed Method

3.1 Data requirements

The first stage is data collecting, in which a dataset with an unbalanced class with a varying amount of features, such as e-Coli and Glass Identification, is employed, as indicated in Table 1. The data will be used to determine how many minority and majority classes there are.

To see if the suggested technique can produce higher accuracy results, this study will compare it to the k-NN method utilizing an unbalanced dataset and a 10-fold Cross-Validation assessment method with a value of k 1 to 10. Table 1 shows the dataset that was used:

Table 1. Detail of Data

Data	*Attributes*	*Class*	*Distribution Class*							
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
E-Coli	7	8	143	77	52	35	20	5	2	2
Glass Identification	9	6	76	70	29	17	13	9	-	-

3.2 Oversampling Process

The dataset will be oversampled using the Synthetic Minority Oversampling Technique (SMOTE) in the second step to balance the quantity of data that is not balanced between the positive and negative

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

classes. The unbalanced data will be first oversampled using SMOTE, and then features will be identified using the Gain Ratio in the suggested technique. Table 2 contains the details for the new dataset.

Table 2. Detail of Data After SMOTE

Data	*Attributes*	*Class*	*Distribution Class*							
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
E-Coli	7	8	143	143	143	143	143	143	143	143
Glass Identification	9	6	76	76	76	76	76	76	-	-

3.3 Data Preprocessing

In the third stage, data is preprocessed by using the Gain Ratio (GR) to pick data properties. GR is used to choose the most relevant features by taking into account intrinsic data from the data, resulting in a greater accuracy rating.

3.4 Testing

The next step is to use k-Nearest Neighbor to assess the training and test data on each dataset (k-NN). In this study, we used SMOTE to test the data set with attribute selection and SMOTE to test the data set without attribute selection.

3.4 Evaluation

Finally, we compare the unbalanced dataset and the new dataset using the 10-fold cross-validation evaluation technique with k values ranging from 1 to 10 to examine if SMOTE and GR can increase the k-NN method's accuracy.

3.5 Result

According to Figure 2 Accuracy value of k-NN, the accuracy value for k-NN in the E-Coli dataset is 75.94%, whereas the suggested technique is 87.37%. The accuracy value for k-NN is 59.22 % in the Glass Identification dataset, whereas the suggested technique is 77.67 %. Finally, the glass dataset saw the largest rise of 18.5%, while the E-Coli dataset saw the least increase of 11.4 %. Table 3 shows that the suggested technique outperforms the original k-NN in terms of precision, recall, and F1-Score when applied to the E-Coli dataset, even though the precision, recall, and F1-Score values are not superior. Precision is up 41.0 %, recall is up 43.4 %, and F1-Score is up 41.5 % across all datasets.

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

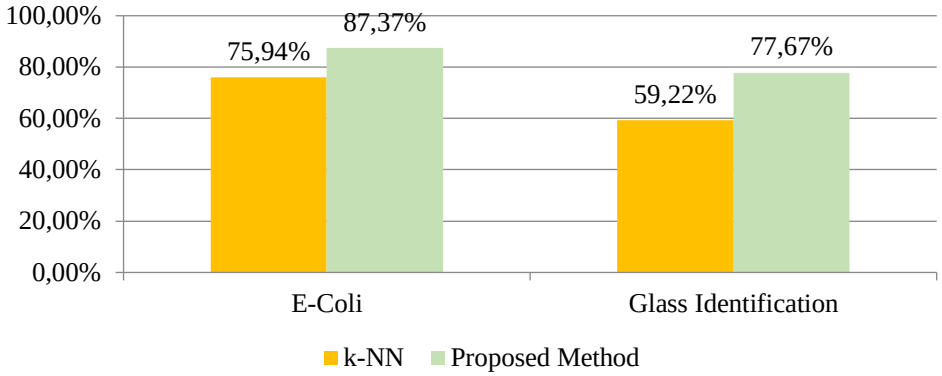

Figure 2. Average Accuracy from All Data

The value of performance from the k-NN technique will be compared with unbalanced data after discovering the new dataset, and the New Dataset will employ the 10-fold cross-validation evaluation method with k values 1 to 10 to check if SMOTE and GR can increase performance. The outcomes Figure 2 shows a comparison of accuracy, whereas table 3 shows precision, recall, and F1-score:

Table 3. Detail Performance

Performance	*E-Coli*		*Glass Identification*		*Average*		
	k-NN	*Proposed Method*	*k-NN*	*Proposed Method*	*E-Coli*	*Glass Identification*	*Avg Increase*
Accuracy	75.94%	87.37%	59.22%	77.67%	11.4%	18.5%	14.9%
Precision	48.04%	63.46%	33.75%	55.96%	59.7%	22.2%	41.0%
Recall	42.83%	64.72%	25.51%	53.28%	59.0%	27.8%	43.4%
F1-Score	44.45%	63.01%	28.06%	53.03%	58.0%	25.0%	41.5%

4. Conclusions

Based on the previous section's explanation, it can be concluded that combining SMOTE and Gain Ratio can improve the accuracy of the k-NN method for unbalanced class data problems, with the lowest increase inaccuracy (11.4%) occurring in the E-Coli dataset and the highest (18.5%) occurring in the Glass Identification dataset. Although the suggested technique outperformed the original k-NN in terms of precision, recall, and F1-Score when applied to the E-Coli dataset, it did not outperform the original k-NN in terms of precision, recall, or F1-Score. Precision is up 41.0 %, recall is up 43.4%, and F1-Score is up 41.5 % across all datasets. In the future, we'll experiment with other datasets with varied characteristics and class distribution numbers, compare method performance with other classification techniques, and do feature selection to improve the performance of the KNN method on unbalanced data.

Reference

[1] L. Cen, "Classifying unbalanced data using a Bagging Ensemble Variation (BEV)," *Proc. Annu. Southeast Conf.*, vol. 2007, pp. 203–208, 2007, doi: 10.1145/1233341.1233378.
[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. June, pp. 321–357, 2002, doi: 10.1613/jair.953.
[3] N. Laopracha, K. Sunat, and S. Chiewchanwattana, "A Novel Feature Selection in Vehicle Detection

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

Through the Selection of Dominant Patterns of Histograms of Oriented Gradients (DPHOG),” *IEEE Access*, vol. 7, pp. 20894–20919, 2019, doi: 10.1109/ACCESS.2019.2893320.

[4] D. Han, S. Member, and J. Kim, “Unified Simultaneous Clustering and Feature Selection for Unlabeled and Labeled Data,” pp. 1–16, 2018.

[5] A. U. Haq, D. Zhang, H. Peng, and S. U. Rahman, “Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection,” *IEEE Access*, vol. 7, pp. 151482–151492, 2019, doi: 10.1109/ACCESS.2019.2947701.

[6] C. Liang, H. Wu, H. Li, Q. Zhang, Z. Li, and K. He, “Efficient data preprocessing, episode classification, and source apportionment of particle number concentrations,” *Sci. Total Environ.*, vol. 744, pp. 1–17, 2020, doi: 10.1016/j.scitotenv.2020.140923.

[7] Z. Wang, X. Xiao, and S. Rajasekaran, “Novel and Efficient Randomized Algorithms for Feature Selection,” vol. 3, no. 3, pp. 208–224, 2020, doi: 10.26599/BDMA.2020.9020005.

[8] H. Benhar, A. Idri, C. Methods, A. Idri, and C. Methods, “Data preprocessing for heart disease classification: A systematic literature review,” *J. Pre-proof Data*, 2020, doi: 10.1016/j.cmpb.2020.105635.

[9] H. Shen and J. Z. Huang, “Sparse principal component analysis via regularized low rank matrix approximation,” *J. Multivar. Anal.*, vol. 99, no. 6, pp. 1015–1034, 2008, doi: 10.1016/j.jmva.2007.06.007.

[10] A. A. Nababan, O. S. Sitompul, and Tulus, “Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio,” 2018.

[11] N. Qazi and K. Raza, “Effect of feature selection, Synthetic Minority Over-sampling (SMOTE) and under-sampling on class unbalance classification,” *Proc. - 2012 14th Int. Conf. Model. Simulation, UKSim 2012*, no. May 2014, pp. 145–150, 2012, doi: 10.1109/UKSim.2012.116.

[12] A. M. De Carvalho and R. C. Prati, “Improving kNN classification under Unbalanced Data. A New Geometric Oversampling Approach,” *Proc. Int. Jt. Conf. Neural Networks*, vol. 2018-July, no. April, 2018, doi: 10.1109/IJCNN.2018.8489411.

[13] A. G. Pertiwi, N. Bachtiar, R. Kusumaningrum, I. Waspada, and A. Wibowo, “Comparison of performance of k-nearest neighbor algorithm using smote and k-nearest neighbor algorithm without smote in diagnosis of diabetes disease in balanced data,” *J. Phys. Conf. Ser.*, vol. 1524, no. 1, 2020, doi: 10.1088/1742-6596/1524/1/012048.

[14] Z. Shi, “Improving k-Nearest Neighbors Algorithm for Unbalanced Data Classification,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 719, no. 1, 2020, doi: 10.1088/1757-899X/719/1/012072.

[15] C. Li, L. Jiang, H. Li, and S. Wang, “Attribute weighted value difference metric,” *Proc. - Int. Conf. Tools with Artif. Intell. ICTAI*, pp. 575–580, 2013, doi: 10.1109/ICTAI.2013.91.

[16] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[17] A. Kataria and M. D. Singh, “A Review of Data Classification Using K-Nearest Neighbour Algorithm,” *Int. J. Emerg. Technol. Adv. Eng.*, vol. 3, no. 6, pp. 354–360, 2013.

[18] Z. Lei, S. Wang, and D. Xu, “Protein sub-cellular localization based on noise-intensity-weighted linear discriminant analysis and an improved k-nearest-neighbor classifier,” *Proc. - 2016 9th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2016*, no. 3, pp. 1871–1876, 2017, doi: 10.1109/CISP-BMEI.2016.7853022.

[19] J. Wang, P. Neskovic, and L. N. Cooper, “Improving nearest neighbor rule with a simple adaptive distance measure,” *Pattern Recognit. Lett.*, vol. 28, no. 2, pp. 207–213, 2007, doi: 10.1016/j.patrec.2006.07.002.

[20] Y. Yuliska and K. U. Syaliman, “Peningkatan Akurasi K-Nearest Neighbor Pada Data Index Standar Pencemaran Udara Kota Pekanbaru,” *IT J. Res. Dev.*, vol. 5, no. 1, pp. 11–18, 2020, doi: 10.25299/itjrd.2020.vol5(1).4680.

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)

[21] K. U. Syaliman, E. B. Nababan, and O. S. Sitompul, "Improving the accuracy of k-nearest neighbor using local mean based and distance weight," *J. Phys. Conf. Ser.*, vol. 978, no. 1, pp. 1–6, 2018, doi: 10.1088/1742-6596/978/1/012047.

[22] Y. Song, J. Liang, J. Lu, and X. Zhao, "An efficient instance selection algorithm for k nearest neighbor regression," *Neurocomputing*, vol. 251, pp. 26–34, 2017, doi: 10.1016/j.neucom.2017.04.018.

INFOKUM is licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC 4.0)