Preprocessing Unbalanced Data using Support Vector Machine with Method K-Nearest Neighbors for Cerebral Infarction Classification

A G M Sari1*, A M Putri1, Z Rustam1, and J Pandelek2
1Department of Mathematics, University of Indonesia, Depok 16424, Indonesia
2Department of Radiology, Cipto Mangunkusumo Hospital, Jakarta 10430, Indonesia
*e-mail: anomgaluhmustikasari@sui.ac.id

Abstract. Cerebral infarction is focal brain necrosis due to complete and prolonged ischemia that affects all tissue elements, neurons, glia, and vessels. Stroke infarction or known as cerebral infarction is a condition of damage in the brain due to insufficient oxygen supply, due to obstruction of blood flow to the area. Research shows stroke infarction does not only occur in the elderly, but occurs at a young age of around 15-55 years, especially with certain risk factors, such as diabetes, hypertension, heart disease, smoking, and long-term alcohol consumption. In diagnosing the presence of cerebral infarction in the brain, machine learning is used because it is not enough just to use a CT scan to diagnose. Therefore, it requires timely detection and more accurate methods of classification. This study aims to use Support Vector Machine (SVM) as preprocessing and K-Nearest Neighbors (KNN) algorithm to classify Infarction Cerebral. In this study, discusses the application of SVM to deal with class imbalances. The first strategy is to balance data using SVM as a preprocessor and the actual target value of the training data is then replaced by trained SVM predictions. Then, the modified training data is used to classify with K-NN method. We use data CT scan result from a Department of Radiology at Dr. Cipto Mangunkusumo Hospital (RSCM). This accuracy in this paper shows around 69.85 %.

Keyword: Machine, K-nearest neighbor, classification

1. Introduction
In Indonesia, stroke is the third most deadly disease, after heart disease and cancer. Stroke is a disease that occurs due to a blockage in the blood vessels of the brain or sudden rupture of blood vessels in the brain [1]. So as a result of blockage or rupture of this blood vessels, certain parts of the brain decrease and even stop the supply of oxygen. So that it becomes damaged or even dead [1].

The type of stroke when viewed from the cause is divided into two namely ischemic strokes and hemorrhagic stroke [1]. Ischemic stroke occurs if the bloods supplied stops due to blood clots and hemorrhagic stroke occurs if the blood vessels that supplied bloods to the brain burst [1]. Ischemic stroke is the most common type of stroke in Indonesia, accounting for 52.9% of all strokes patients.

Cerebral infarction is focal brain necrosis due to complete and prolonged ischemia that affects all tissue elements, neurons, glia, and vessels [2]. Stroke infarction or known as cerebral infarction is a
condition of damage in the brain due to insufficient oxygen supply, due to obstruction of blood flow to the area [3]. Research shows stroke infarction does not only occur in the elderly, but occurs at a young age of around 15-55 years, especially with certain risk factors, such as diabetes, hypertension, heart disease, smoking, and long-term alcohol consumption [2].

For patients with ischemic stroke, be a cerebral infarction can be seen in the brain through detection with a CT scan. However, the results of a CT scan are not enough to detect and diagnose the presence of infarction in the brain. In diagnosing the presence of cerebral infarction in the brain, machine learning is used because it is not enough just to use a CT scan to diagnose. Therefore, it requires timely detection and classification of infarcts in the brain using labels and features available from the results of CT scans.

This study proposes preprocessing unbalanced data using a support vector machine with the K-Nearest Neighbor method for classification of brain violations. The dataset used is CT scan data from a Department of Radiology at Dr. Cipto Mangunkusumo Hospital (RSCM). However, due to unbalanced data infarction, the tendency of imbalance class data will cause instability. The data will be more likely to classify as a class consisting of a larger number.

The problem of imbalanced data is solved by modifying the infarction dataset through the duplication of minority data, or data with a small number of classes, to be balanced data with a large number of data classes [6]. This process is also called oversampling. Other datasets are modified by reducing majority data, or reducing data with large number of classes, to be balanced with a smaller number of data classes [6]. This process is also called undersampling.

Researchers have never reported any preprocessing using intelligent methods to balance the data [4]. In this paper we employ SVM as a preprocessor. SVM is one of the best intelligent algorithms used for classification and regression purposes [4]. The best property of SVM is that it always yields global optimal solution, whereas other intelligent algorithms suffer from getting stuck with a local minimum. SVM tries to find the decision boundary between various classes without actually worrying about the number of instances available for a class [4]. SVM is suitable for high dimensional problems and works with a small number of observations as well [4]. Hence, trained SVM is proposed as a preprocessor in this paper. After using SVM as preprocessing, data modification will be done without reducing the accuracy with undersampling. Then, proceed with the classification of data using K-NN to get accuracy on cerebral infraction data.

This study aims to use Support Vector Machine (SVM) as preprocessing and K-Nearest Neighbors (KNN) algorithm to classify Cerebral Infarction. In this study, discussing the application of SVM to overcome class imbalances. The first strategy is to balance the data using SVM as a preprocessor and the actual target value of the training data is then replaced by trained SVM predictions. In the SVM process, we still get unbalanced trained data, so the data are modified to make the data balanced through undersampling. Then, the modified training data is used to classify the K-NN method.

2. Materials and Methods

2.1 Data

The data used in this study is data from ischemic stroke patients who have cerebral infarction in the brain. There are 136 data with seven proportional features which are used as 76% training data and 24% testing data from the original data with 103 major data and 33 minor data. Major data labeled '0' that represents the class data indicated in the brain and minor infarction of data labeled ‘1’ representing the class data indicated that no infarction. This data was taken from January to November 2018 from Dr. Cipto Mangunkusumo Hospital (RSCM). Table 1 explains the infarction data features that will be examined. Table 2 shows the display of the data provided.
Table 1. The features of cerebral infraction dataset

| No | Feature | Definition of feature                                      |
|----|---------|------------------------------------------------------------|
| 1  | Area    | The size of the area from the infarction point             |
| 2  | Min     | The minimum value of infraction                            |
| 3  | Max     | The maximum value of infraction                            |
| 4  | Average | The average value of infraction                            |
| 5  | SD      | Standard error value of infraction                         |
| 6  | Sum     | The sum value of infraction                                |
| 7  | Length  | Length of infraction point                                |

Table 2. The display of data cerebral infraction

| Area | Min | Max | Average | SD  | Sum  | Length | Target |
|------|-----|-----|---------|-----|------|--------|--------|
| 0.1  | 15  | 44  | 30.64   | 7.37| 7722 | 1.8    | 0      |
| 0.1  | 18  | 51  | 32.29   | 7.84| 8134 | 1.8    | 0      |
| 0.1  | 25  | 61  | 38.99   | 7.37| 6122 | 1.5    | 0      |
| 0.1  | 32  | 58  | 42.98   | 5.49| 7736 | 1.6    | 0      |
| 0.2  | 0   | 37  | 17.39   | 7.3 | 3965 | 2.1    | 1      |
| 0.1  | 5   | 38  | 17.93   | 8.41| 1506 | 1.4    | 1      |
| 0.1  | 8   | 43  | 28.55   | 7.85| 2198 | 1.3    | 1      |

2.2 Preprocessing Data

Data preparation or also called data preprocessing is a process or step taken to make raw data into quality data (good input for data mining tools).

Data need to be preprocessed because raw data still contains data that:
- Incomplete,
  Data that lacks attribute values or only contains aggregate data (example: address).
- Noisy,
  Data that still contains errors and outliers (example: salary = -10).
- Inconsistent
  Data containing discrepancies in the code and the name or brevity of the data are inconsistent (for example: first rating = 1,2,3 but now it becomes a, b, c).

So data preparation must be done because:
- If the input data is not quality, then the data mining results will also not be quality.
- Quality decisions must come from quality data.
- Data warehouse requires consistent integration of quality data.

Data preprocessing is one of the most data mining tasks which includes preparation and transformation of data into a suitable form to mining procedure [8]. Data preprocessing aims to reduce
the data size, find the relations between data, normalize data, remove outliers and extract features for data [8]. It includes several techniques like data cleaning, integration, transformation and reduction [8].

2.3 Support Vector Machine (SVM)

The SVM is a learning procedure based on the statistical learning theory and it is one of the best machine learning techniques used in data mining [7]. One of the strengths from Support Vector Machine is to predict, classify, and advance in a case. In its application, SVM has the basic principle of a linear classifier, which means this casing can be linearly separated [5]. However, currently SVM has also been developed [5]. So it can work on non-linear problems by adding kernel concepts to high-dimensional workspaces [5]. SVM in high dimensional space, we will look for something called hyperplane which can maximize the distance between several classes of data [5].

For solving a two-class classification problem, the main objective of SVM is to find an optimal separating hyperplane that correctly classifies data points as much as possible and separates the points of the two classes as far as possible, by minimizing the risk of misclassifying the training samples and unseen test samples [4].

The optimization problem for the SVM can be depicted as follows:

\[
\min \frac{1}{2} (w, w) \\
\text{Subject to } y_i (w \cdot x_i + b) \geq 1 \quad \forall x_i. 
\]

(1)

The SVM classification function for classifying linearly separable data can be written as:

\[
f(x) = \langle w, x \rangle + b = \sum_{i=1}^{l} y_i \alpha_i \langle x_i, x \rangle + b 
\]

(2)

This is also known as hard margin, where no room is given for errors. It is observed that most of the time it is linearly non separable. Hence slack variable \( \xi \) is introduced to allow \( \xi \) error and the optimization function takes the form of Eq. (3) as shown below:

\[
\min \frac{1}{2} (w, w) + C \sum_{i=1}^{l} \xi_i \\
\text{Subject to } y_i (w \cdot x_i + b) \geq 1 \quad \forall x_i. 
\]

(3)

To deal with the problem of non-linearly separable dataset, SVM first projects the data into a higher dimensional feature space using various kernels and tries to find the linear margin in the new feature space. The optimization function can be depicted as shown below:

\[
\min \frac{1}{2} (w, w) + C \sum_{i=1}^{l} \xi_i \\
\text{Subject to } y_i (w \cdot \phi(x) + b) \geq 1 \quad \forall x_i. 
\]

(4)

The optimal hyperplane separating the binary decision classes is given by Eq. (5):

\[
f(x) = \sum_{i=1}^{l} y_i \alpha_i K(x_i, x) + b 
\]

(5)

where \( K(x_i, x) = \phi(x_i)\phi(x) \) is taken with a semipositive definite kernel.

The main objective of SVM is to find an optimal separating hyperplane that correctly classifies data points as much as possible and separates the points of two classes as far as possible, by minimizing the
risk of misclassifying the training samples and unseen test samples [4]. C and kernel are the only parameters for training SVM [4]. In other words, more misclassifications for majority class instances mean more number of instances for minority class instances. As a result, the modified data has more instances for minority class. This is turn yields data which is not only balanced but also provides more number of instances for minority class without compromising the prediction accuracy of SVM. The reason behind using SVM for such preprocessing is that, it predicts similar instances from majority class instances as minority class instances, instead of randomly selecting them from training data as minority class instances.

2.4 Undersampling

A simple way to fix imbalanced datasets is simply to balance them, either by oversampling instances of the minority class or undersampling instances of the majority class. This simply allows us to create a balanced dataset that, in theory, should not lead to classifiers biased toward one class or the other. Undersampling is a process to balance class distribution by reducing the majority of class instances. Disadvantages of undersampling is the loss of data that are considered important for the survival of the decision making process by machine learning. Though this seems to work for the majority of cases, no detailed analysis exists about the impact of undersampling on the accuracy of the final classifier [9]. It emerges that the impact of undersampling depends on the number of samples, the variance of the classifier, the degree of imbalance and more specifically on the value of the posterior probability [9].

2.5 K-Nearest Neighbors

K-Nearest-Neighbors (KNN) is a simple, but effective classification method [10]. K-Nearest Neighbor or often abbreviated as KNN is one algorithm used to classify objects based on learning data (training data) which is the closest distance to the object [10]. This method builds models from data and classifies new data using models. This model is a representative set of training data, as an area in the data space. KNN is widely used in data mining applications, pattern recognition, image processing, and others.

KNN is a supervised learning algorithm whose algorithm uses available data and known outputs. The purpose of the KNN algorithm is to classify new objects based on attributes and training samples. Where the results of the newly approved test sample are based on the KNN category. In the classification process, this algorithm does not use any model to be matched and is only based on memory. KNN algorithm uses environmental classification as the estimated value of the new sample test data.

2.6 Proposed method

SVM is one of the most effective classification techniques proposed in literature and it is very efficient in solving two class classification problems [4]. The proposed approach first builds a SVM model and the actual target values of the training instances are then replaced by the prediction of the trained SVM. Later, this modified data is used to train the K-Nearest Neighbor algorithm. It is observed that use of SVM prediction accuracy of the classifiers for the minority class instances.

The steps involved in the proposed two phase balancing approach are as follows: steps 1 and 2 below make up phase 1 of the proposed approach, and steps 3 and 4 make up phase 2 of the proposed approach.

Step 1 SVM training – using the available unbalanced data SVM is trained, and the SVM model with the best prediction accuracy is selected and used for prediction purposes.

Step 2 Prediction of training data using trained SVM – after successfully obtaining the trained SVM, the target values in the training data are replaced by the predictions of the trained SVM. At this stage available data is modified and balanced data is obtained.
Step 3 Modified data is then used to train various intelligent algorithms using the modified data.

Step 4 Trained intelligent algorithms are used for prediction — predictions are obtained using the trained intelligent algorithms using the modified data and empirical analysis of the accuracy of prediction is carried out.

The steps involved in the proposed two phase balancing approach are presented in Fig. 1.

![Phase 1: Preprocessing](image1)

**Figure 1.** Proposed method step

### 3. Experimental Results

This research uses Google Colab, which is the coding environment of the Python programming language with the format "notebook" (similar to Jupyter notebook).

The results of the program created are:

- Out of 136 data with seven proportional features, 76% of training data and 24% of testing data with 103 major data labeled '1' and 33 minor data labeled '0'.
- Preprocessing is done using SVM, the ratio result is 114: 22.
- The next step is modifying training data using undersampling. So, we get the undersampled data, which is 22: 22.
- The final step is to classify the data using KNN, which obtained an accuracy of 69.85%

|                  | Preprocessing | Accuracy  |
|------------------|---------------|-----------|
| No SVM in preprocessing | 91.91%     |
| SVM preprocessing + KNN      | 69.85%     |

**Table 3.** The results of accuracy from the program

### 4. Discussion

In this paper discusses the use of SVM as preprocessing in infarction data. Where SVM is very easy to use because, only by finding the best hyperplane – hyperplane, which is useful for classifying two classes in the input space. The classification program can be completed by finding a line or hyperplane that separates the two groups. However, different from this research. SVM is only used as a preprocessor,
which will then be performed undersampling and classifying it using the KNN algorithm. Using SVM as preprocessing, to predict by randomly selecting it from training data as a minority class instance. This can be used as a comparison to determine the accuracy of further research, when using SVM as preprocessing or using SVM as a classification. This cannot be generalized using other data or other optimization parameters; consequently, the problem is limited in terms of the data used and optimization parameters.

5. Conclusion

From the results of the above experiments, it was produced that using SVM as preprocessing, then modifying the data by undersampling and classifying it using KNN, did not make the results better when compared to using SVM alone. But, we can still say that the accuracy of 69.85% is a good accuracy in this study. Because using SVM as a preprocessor does not make changes to the unbalanced data. The use of SVM is in vain because after doing SVM, trained data must be slightly modified with undersampling to make the data balanced. After obtaining a balanced balanced trained data, the classification is done using the KNN algorithm. We suggest that in future studies it is better to compare the results of classifications using SVM or the use of SVM, which is only as a preprocessor.

Acknowledgements
This research supported financially by the Ministry of Research, Technology, and Higher Education Republic of Indonesia (KEMENRISTEKDIKTI) with PTUPT 2020 research grant scheme. We would like to thank Dr. Dr. Jacub Pandelaki, Sp.Rad(K) from Department of Radiology, Cipto Mangunkusumo for the dataset.

References
[1] Honestdocs, accessed 29 September 2019, see https://www.honestdocs.id/stroke-pengertian-jenis-gejala-stroke.
[2] Neuropathology, accessed 29 September 2019, see http://neuropathology-web.org/chapter2/chapter2bCerebralinfarcts.html.
[3] Alodokter, accessed 30 September 2019, see https://www.alodokter.com/waspadai-stroke-infark-yang-mengancam-usia-muda.
[4] Farquad M A H and Indranil B 2012 Decisision Support System Preprocessing unbalanced data using support vector machine 53 226
[5] Z. Rustam, Dea A U, Rahmat H, Jacub P, and Widyo 2019 A N Hybrid Preprocessing Method for Support Vector Machine for Classification of Imbalanced Cerebral Infraction Datasets 9(2) 685
[6] V.N. Vapnik 1995 The Nature of Statistical Learning Theory (Springer Verlag: New York USA)
[7] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg 2008 Top 10 algorithms in data mining Knowledge and Information Systems 14 1
[8] Suad A A and Wesam S B 2017 Review of Data Preprocessing Techniques in Data Mining: Journal of Engineering and Applied Sciences 12(16) pp. 4102-4107
[9] Andrea D P, Olivier C, and Gianluca B When is undersampling effective in unbalanced classification tasks?
[10] Gongde G, Hui Wang, David B, Yaxin B, and Kieran G 2003 KNN Model-Based Approach in Classification pp 986-996