Updated on 13 Aug 2019

M 1

Master Thesis Proposal

Faculty of Science, Chulalongkorn University

Name **Ms. Patcharasiri Fuangfoo** ID **6470121923**

Department **Mathematics and Computer Science**

Program **Applied Mathematics and Computational Science** Thesis Credits **18**

Academic Curriculum ☑ Regular ☐ Part-Time ☐ English Program ☐ International Program

Enrolled since the ☑ First ☐ Second Semester, Academic Year **2021**

Contact Address **126 Moo 20 Rop Wiang, Muang Chiang Rai, Chiang Rai, 57000**

Telephone Number **0956750521**

Thesis Title (Thai)

ตัวแบบจำแนกประเภทเพื่อนบ้านใกล้สุดแบบพลวัตโดยใช้ปัจจัยค่าผิดปกติแมส-เรโช-แวเรียนซ์ สำหรับปัญหาคลาสไม่ดุล

Thesis Title (English -- Capital Letters)

DYNAMIC NEAREST NEIGHBOR CLASSIFIER USING MASS-RATIO-VARIANCE OUTLIER FACTORS FOR CLASS IMBALANCE PROBLEM

Thesis Advisor **Associate Professor Dr. Krung Sinapiromsaran** Tel. **02-218-7123**

Signature.....พัชรสิริ เฟื่องฟู.....

Student

Signature.....

Thesis Advisor

Signature.....เพชรตภา บุญเสริม.....

(Associate Professor Dr. Petarpa Boonserm)

Program Chair

Proposal Examination Date **2 December 2022**

Expecting to Defend the Thesis in the ☐ First ☑ Second Semester, Academic Year **2022**

Approved by the Administrative Committee of Master Program in **Applied Mathematics and Computational Science** Meeting No. …**22**…/…**2022**…… Date…..**19 December 2022**…….

Signature.....เรวัต กนิดกิ่งไชย..... (Program Secretary)

Date..19../..12../..2022

Approved by the Faculty Administrative Committee Meeting No. …2/2023….Date…23/1/2023……..

Signature.....

Associate Dean, Faculty of Science

Date.24…../..1…../…2023

M 2

Master Thesis Proposal

Faculty of Science Administrative Committee

Name **Ms. Patcharasiri Fuangfoo** ID **6470121923**

Department **Mathematics and Computer Science**

Program **Applied Mathematics and Computational Science** Thesis Credits **18**

Academic Curriculum ☑ Regular ☐ Part-Time ☐ English Program ☐ International Program

Enrolled since the ☑ First ☐ Second Semester Academic Year **2021**

Thesis Title (Thai)
ตัวแบบจำแนกประเภทเพื่อนบ้านใกล้สุดแบบพลวัตโดยใช้ปัจจัยค่าผิดปกติแมส-เรโช-แวเรียนซ์ สำหรับปัญหาคลาสไม่ดุล

Thesis Title (English -- Capital Letters)
DYNAMIC NEAREST NEIGHBOR CLASSIFIER USING MASS-RATIO-VARIANCE OUTLIER FACTORS FOR CLASS IMBALANCE PROBLEM

Thesis Advisor **Associate Professor Dr. Krung Sinapiromsaran**

Objectives(s)
To improve the k-nearest neighbor classifier that can handle class imbalance problem

Rationales, Theories, or Assumptions

Classification is the process of classifying a class of an instance in a given dataset via a classifier. The classifier is built from a training dataset using the classification algorithm. One of the popular classifiers is k-nn stands for "k-nearest neighbor". A user specifies parameter 'k' before applying the classifier to an unknown instance. When dealing with an imbalanced dataset, the model normally predicts an instance as a majority which misclassifies most minority instances.

Mass-ratio-variance outlier factors (MOF) uses the density concept by utilizing the mass-ratio between a pair of data points. The variance of the mass-ratio distribution is used to assign MOF for each data point. The large variance is associated with outliers, while the small variance is associated with a normal data point.

In this work, the k-nearest neighbor classifier will be improved by dynamically identify the number of nearest neighbors to be used for classifying the current data point based on mass-ratio-variance outlier factors.

Updated on 13 Aug 2019

M 3

Research Plan

Use straight lines (no arrows) in numbered spaces to represent the period of time (in months) for each step to complete the thesis

Description	Month, Year (Thesis started in **January, 2022**)																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1. Study interesting topics	—	—	—	—	—	—	—	—										
2. Review articles and related works	—	—	—	—	—	—	—	—	—	—								
3. Design the dynamic nearest neighbor algorithm									—	—								
4. Implement code **- Mass-ratio-variance outlier factors** **- k nearest neighbor**									—	—	—							
5. Write thesis and defense											—	—	—	—	—	—		
6. Publish the finding in the conference																—	—	—

Expected beneficial outcome(s) from the thesis

Obtain the dynamic nearest neighbor classifier that provides a better accuracy than traditional k nearest neighbor for a class imbalanced problem.

Signature.....พัชรสิริ เพี่ยงพุ่.....

Student

M 4

Appendix

Master Thesis Committee Nomination

Name **Ms. Patcharasiri Fuangfoo** ID **6470121923**

Department **Mathematics and Computer Science**

Program **Applied Mathematics and Computational Science** Thesis Credits **18**

Academic Curriculum ☑ Regular ☐ Part-Time ☐ English Program ☐ International Program

Enrolled since the ☑ First ☐ Second Semester Academic Year **2021**

Thesis Title (Thai)
ตัวแบบจำแนกประเภทเพื่อนบ้านใกล้สุดแบบพลวัตโดยใช้ปัจจัยค่าผิดปกติแมส-เรโช-แวเรียนซ์ สำหรับปัญหาคลาสไม่ดุล

Thesis Title (English -- Capital letters)
DYNAMIC NEAREST NEIGHBOR CLASSIFIER USING MASS-RATIO-VARIANCE OUTLIER FACTORS FOR CLASS IMBALANCE PROBLEM

List of Master Thesis Committee

Program Chair or Appointed Personnel by the Chair	Chair
Associate Professor Dr. Krung Sinapiromsaran	Advisor
Assistant Professor Dr. Kitiporn Plaimas	Committee
Assistant Professor Dr. Chumphol Bunkhumpornpat [1]	External Examiner

Signature .....เพชรภา มุณสิม.....
Program Chair

[1] **Department of Computer Science, Faculty of Science, Chiangmai University**

Note: The thesis committee nomination must comply with Section 6 “Instruction and Examination”, Part 6 “Thesis examination”, Number 94 in Chulalongkorn University Graduate Studies Regulations, 2008

Updated on 13 Aug 2019

M 5

Master Thesis Proposal

Faculty of Science Administrative Committee

Name **Ms. Patcharasiri Fuangfoo** ID **6470121923**

Department **Mathematics and Computer Science**

Program **Applied Mathematics and Computational Science** Thesis Credits **18**

Academic Curriculum ☑ Regular ☐ Part-Time ☐ English Program ☐ International Program

Enrolled since the ☑ First ☐ Second Semester Academic Year **2021**

Thesis Title (Thai)

ตัวแบบจำแนกประเภทเพื่อนบ้านใกล้สุดแบบพลวัตโดยใช้ปัจจัยค่าผิดปกติแมส-เรโช-แวเรียนซ์ สำหรับปัญหาคลาสไม่ดุล

Thesis Title (English -- Capital letters)

DYNAMIC NEAREST NEIGHBOR CLASSIFIER USING MASS-RATIO-VARIANCE OUTLIER FACTORS FOR CLASS IMBALANCE PROBLEM

☐ The thesis is involved with human or animal testing.

(Student must request an approval from a committee on ethics in human or animal testing prior to the proposal approval.)

☑ The thesis is not involved with human or animal testing.

Biosafety regulation

☐ The thesis uses pathogens, animal toxins (Pathogens and Animal Toxins Act, B.E. 2015) or genetically modified organisms.

- ☐ Approved by Faculty of Science Institute Biosafety Committee (Project ID……)
- ☐ Exempt by Faculty of Science Institute Biosafety Committee (Project ID……)

☑ The thesis does not use pathogens, animal toxins (Pathogens and Animal Toxins Act, B.E. 2015) or genetically modified organisms.

Updated on 13 Aug 2019

Appendix

Name **Patcharasiri Fuangfoo** ID **6470121923**
Department **Mathematics and Computer Science**
Program **Applied Mathematics and Computational Science**

Thesis Title (Thai)
ตัวแบบจำแนกประเภทเพื่อนบ้านใกล้สุดแบบพลวัตโดยใช้ปัจจัยค่าผิดปกติแมส-เรโช-แวเรียนซ์ สำหรับปัญหาคลาสไม่ดุล

Thesis Title (English)
DYNAMIC NEAREST NEIGHBOR CLASSIFIER USING MASS-RATIO-VARIANCE OUTLIER FACTORS FOR CLASS IMBALANCE PROBLEM

Rationales, theories or assumptions (in detail)

Machine learning is a computer science field that studies algorithms, statistical models, and mathematical models [1] to perform learning. A learning task is not a programming construct with explicit instructions, but it relies on patterns to be captured by a model and inference to be able to apply it to unknown instances. A machine learning algorithm builds a mathematical model based on sample data (training data) in order to make predictions, decisions, etc. Machine learning is divided into three main categories [2]: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is machine learning paradigm for problems where the available data consists of labeled instances. On the other hand, unsupervised learning is machine learning paradigm to learn patterns from unlabeled data to group similar instances or reduce dimension of input data. Finally, reinforcement learning is an area of machine learning concerned with how intelligent agents ought to take actions in an environment in order to maximize the notion of cumulative reward.

Classification is the process of classifying a class of an instance in a given dataset via a classifier. The classifier is built from a training data using the classification algorithm. It is one of the supervised learning algorithms used to identify the category of new observations on the basis of training data. One of the popular classifiers is k-nn, stands for "k-nearest neighbor" [3].

The principle of the k-nearest neighbor algorithm assigns a class to each test instance using the majority-voting class from its k-nearest neighbor. The k-nearest neighbor algorithm assumes that all data correspond to points in the d-dimensional space. Let the test instance x_t be represented by the feature vector $[x_t^1, x_t^2, x_t^3, \dots, x_t^d]'$, where x_t^l denotes the value of the l-th attribute of the test instance x_t and x_t' is the transpose of x_t. An example of a distance used in the calculation is the Euclidean distance. The Euclidean distance of x_i and x_j is defined as

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_i^l - x_j^l)^2}.$$

If the number of training data is n, then n distances will be calculated, and the k closest training data will be identified as neighbors. If $k = 1$, then the class label of the test instance is the same as the class of the closest training instance. If $k > 1$, then the class label of the test instance is identified as the class label from most of the neighbors have [4].

A k-nearest neighbor algorithm is simple and does not require any model parameter except k, it has two main problems as well [9]:

1. The algorithm uses all the data as a model, which takes a long time to search for k-nearest neighbors of any instance.
2. Classification performance depends on a single parameter k as the number of neighbors to be used for any instance.

Many studies have attempted to solve the first problem using some structure search tree. Their strategy is to eliminate similar instances or find a representative of instances [10] or use that representative to reduce the sample size considered from the whole dataset to only the small portion in that representative region. Some researchers propose using the majority vote for all outcomes resulting from every classifier k-nearest neighbors when k is $1,3,5,\ldots,\sqrt{n}$ to deal with the second problem [9].

Currently, a classification problem may face with class imbalance problem [5], [6], which has gained a lot of attention lately because the imbalanced data is found in most real-world problems such as fraud detection, anomaly detection, and medical diagnosis. In this situation, the number of instances of a class called “minority” is much lower than the instances of other classes [6]. If there are two classes, then another class is called the majority.

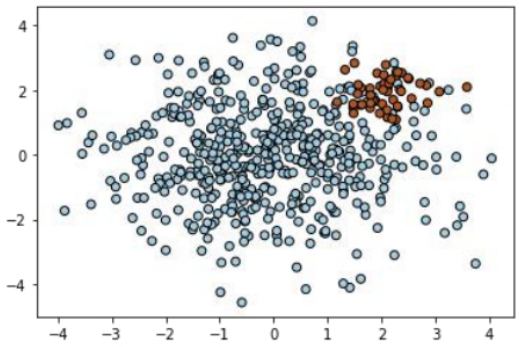

Figure 1: The red circle represents the minority, and the blue circle represents the majority.

When dealing with an imbalanced dataset, the classifier normally misclassifies minority instances. There are several methods to tackle the problem of imbalanced datasets. These methods are grouped into three categories [7]: the data-level approach, the algorithmic-level approach, and the ensemble approach. The first approach modifies the distribution of data to balance it, such as oversampling, which generates new instances of minorities, or under-sampling, which removes instances from majorities. The second approach directly modifies existing learning algorithms to reduce the bias towards majority objects. For example, improving k nearest neighbor with Exemplar Generalization, called k ENN [8], enlarge the positive instances in the training sample, which are referred to as “exemplar positive instances”, to expand the decision boundary of the positive class. It was shown that k ENN significantly improves the performance of k-nearest neighbors. Nearest neighbor classification with locally weighted distance [6] based on gradient ascent is proposed to improve the nearest neighbor algorithm on imbalanced data. The last approach combines the advantages of the two previous ones.

When using k-nearest neighbors (regardless of k) with the imbalance problem, poor performance is often obtained because the instance orientation may not be suitable for any fixed k. Therefore, selecting the k value according to the density of each instance would give better results.

Mass-ratio-variance outlier factors (MOF) [11] uses the density concept by utilizing the mass-ratio between a pair of data points. The variance of the mass-ratio distribution is used to assign MOF for each data point. The large variance is associated with outliers, while the small variance is associated with a normal data point.

This work proposed the algorithmic level approach which will use different k-nearest neighbors for each instance using different mass-ratio-variance outlier factors as the threshold to identify the number of appropriate nearest neighbors.

To identify the number of nearest neighbors, the MOF scores are divided into $\sqrt{n}$ ranges, where n is the number of instances, and each range is assigned a different number of nearest neighbors. If the MOF score is high, the number of nearest neighbors will be set to be low to prevent misclassification. If the MOF score is low, the number of nearest neighbors will be set to be high to increase the reliability of the neighbors.

Research procedures (in detail)

1. Study interesting topics
2. Review articles and related works
3. Design the dynamic nearest neighbor algorithm
4. Implement code
   - Mass-ratio-variance outlier factors
   - k nearest neighbor
5. Write thesis and defense
6. Publish the finding in the conference

References (article titles must be included)

[1] E. Alpaydın, “Introduction to Machine Learning (third edition)”, *The MIT Press*, Cambridge, Massachusetts London, England, 2014.

[2] Alejandro A. Torres-García, Carlos A. Reyes-García, Luis Villaseñor-Pineda, and OmarMendoza Montoya, “Biosignal Processing and Classification Using Computational Learning and Intelligence Principles, Algorithms, and Applications”, *Academic Press*, 2022.

[3] G. H. Chen and D. Shah, “Explaining the Success of Nearest Neighbor Methods in Prediction”, *Foundations and Trends® in Machine Learning*, Vol. 10, No. 5-6, pp 337-588, 2018.

[4] M. Sarkar and T. Y. Leong, “Application of K-Nearest Neighbors Algorithm on Breast Cancer Diagnosis Problem”, *AMIA Annual Symposium Proceedings Archive*, pp 759-763, 2000.

[5] Y. Li and X. Zhang, “Improving k Nearest Neighbor with Exemplar Generalization for Imbalanced Classification”, *PAKDD 2011*, Part II, LNAI 6635, pp. 321–332, 2011.

[6] Z. Hajizadeh, M. Taheri, and M. Z. Jahromi, “Nearest Neighbor Classification with Locally Weighted Distance for Imbalanced Data”, *International Journal of Computer and Communication Engineering*, Vol. 3, No. 2, 2014.

[7] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions”, *Progress in Artificial Intelligence*, Vol. 5, pages221–232, 2016.

[8] Y. Li and X. Zhang, “Improving k Nearest Neighbor with Exemplar Generalization for Imbalanced Classification”, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp 321–332, 2011.

[9] A. Hassanat, M. A. Abbadi, G. A. Altarawneh, and A. A. Alhasanat, “Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach”, *(IJCSIS) International Journal of Computer Science and Information Security*, Vol. 12, No. 8, pp. 33-39, 2014.

[10] L. J. Moreira and L. A. Silva, “Prototype Generation Using Self-Organizing Maps for Informativeness-Based Classifier”, *Computational Intelligence and Neuroscience*, Vol. 2017, Article ID 4263064, 15 pages, 2017.

[11] P. Changsakul, S. Boonsiri, and K. Sinapiromsaran, “Mass-ratio-variance based Outlier Factor”, *2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2021.