



# INTEGRATION OF CLASSIFICATION TECHNIQUES UNDER SIZE CONSTRAINT: A CASE STUDY OF ELECTORAL DISTRICTING

Nattapong Musikauppatum\* , Surapong Uttama

School of Information Technology, Mae Fah Luang University 333 Moo1, Thasud, Muang, Chiang Rai, 57100, Thailand

\*e-mail: socoolidea@hotmail.com

---

## Abstract

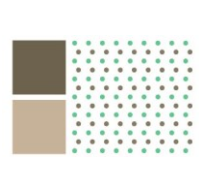
The electoral districting is a problem of classification under size constraint. Its objective is to classify voters according to their traveling distances to fixed election units which are limited in capacity to support voters. Thus the goal of this paper is to propose a new classification algorithm to satisfy two conditions: size constraint and minimum traveling distance to election units. We modify k-nearest neighbor (k-NN) classification to support size constraint and combine with discriminant analysis. The experiment was tested on electoral districting of a district in Chiang Rai, Thailand. The result showed that our proposed algorithm satisfied the size constraint and the average traveling distance was 10% better than the manual classification and 0.82% better than modified k-NN and SVM with size constraint.

---

**Keywords:** classification, size constraint, k-nearest neighbor, support vector machine, discriminant analysis,

## Introduction

The main focus of election system is voters and election units. Usually election units must be close to voters' locations and can support a limited number of voters. Allocating voters properly to election units is related to voters' traveling distance and can be considered as the problem of supervised classification. Supervised classification is a famous concept of machine learning and various algorithms were proposed by Paul (2012), Veenman et al (2005), Sahiner et al (2007). Each of them has different strengths and weaknesses. However, the classical supervised classification techniques are not available for our work because they do not support the size constraint. In this work, we need the classification that supports both size constraints together with minimizing distance. From literature, the classification under constraint was studied in different goals. For example, we found geometric constraint by Yongzhu et al (2010), training data size constraint by Yeran et al (2010), and neighborhood constraint by Idbraim et al (2009). The classification under size constraint was suggested by Nattapong (2012) which focused on the application of k-nearest neighbor (k-NN) and support



vector machine (SVM) under class size constraint for solving the problem of the electoral districting. The result was interesting but the process of SVM is complicated and costly. From our observation, the problem of electoral districting deals with a large number of voters. Naturally a distribution of voters' locations tends to be a normal distribution. Taking this into account, there is a possibility to replace SVM with simpler and faster algorithm that supports normal distribution of data which is a discriminant analysis.

The classification in this work is inspired by Nattapong (2012). We follow the same classification scheme and propose to replace the SVM by discriminant analysis.

## Methodology

The propose algorithm is a new combination of modified k-NN and discriminant analysis to increase the classification efficiency. This is because we know that discriminant analysis is powerful in classification but is lack of size constraint and requires sufficient number of training data. Integration of these two techniques would help to satisfy the size constraint with expectation of higher classification efficiency.

The proposed algorithm starts from generating training data. We generate a new training set by selecting empirically 55 nearest points for each centroid. This is done using the modified k-NN mentioned in Nattapong (2012) and set the size constraint to be 55 which is the best value for training of both the average traveling distance ( $\mu$ ) and the standard deviation ( $\sigma$ ). Then we train these data via discriminant analysis and again use it to classify the remaining data. Certainly it is possible that the result could have the overloaded groups. That is, the sizes of some groups may be larger than our size constraint. To correct this problem, we extract the redundant data from overloaded groups according to the distances to their centroids. Then we repeat the modified k-NN classification to assign them to the available groups. The entire processes can be described by the following algorithm.

**Algorithm:** Integration of modified k-NN with size constraint and discriminant analysis

**Input:** location of election units (centroids), location of voters (data)

**T**  $\leftarrow$  call modified k-NN to get 55 nearest training data

Train data **T** to discriminant analysis

Classify remaining data by discriminant analysis

FOR EACH classified group

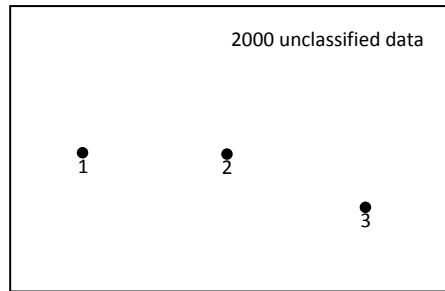
    IF group size  $>$  constraint

**R**  $\leftarrow$  Extract redundant data by distance

    call modified k-NN to classify **R** to not full groups (groups that do not reach the size limit)

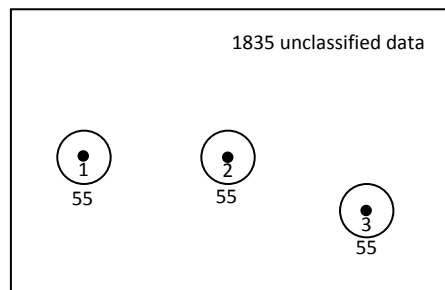
**Output:** group members

We illustrate the implementation of our proposed algorithm by the following example. Assume that there are three election units (centroids), the total number of data is 2000 and the size constraint for each unit is 700.



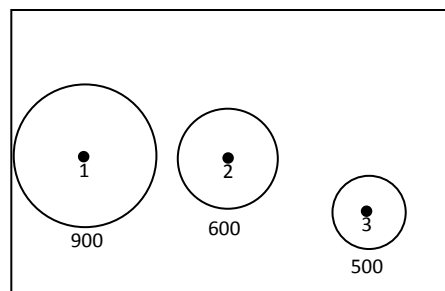
**Figure 1** Three sample election units.

1. Generate training data using k-NN with a fixed number of points. Empirically 55 points of training data is the best candidate in terms of speed and accuracy. The remaining unclassified data is 1835 ( $2000 - 55 \cdot (3)$ ).



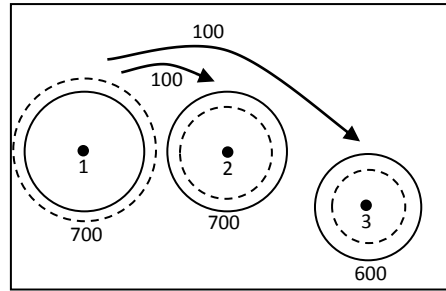
**Figure 2** Training stage.

2. Use the training data to train discriminant analysis and classify the remaining data. Assume that the classified result is 900, 600 and 500 points for each group as shown in figure 4.



**Figure 3** Result of discriminant analysis classification.

3. Observe that the first group is overloaded. Its size is larger than the limit (700). So we extract the redundant data and apply the modified k-NN again to distribute them to available groups (2 and 3). Note that the redundant data are distributed to the second group first because it is closer. After it is full, the rest will be assigned to the third group. The final result is shown in figure 5.



**Figure 4** Final classification result.

## Results

The proposed methods were tested covering 12 election units and 6,382 voters in downtown Chiang Rai, Thailand. The number and locations of voters were announced by the election committee and are available for public. Then we apply two sets of algorithm. The first set is the combination of modified k-NN and SVM from Nattapong (2012). The second set is our proposed algorithm which is the combination of modified k-NN and discriminant analysis. Then we measure the average traveling distance ( $\mu$ ) and its standard deviation ( $\sigma$ ). The result is shown in Table 1.

**Table 1** The comparison of the manual with the proposed method.

	<b>Manual</b>	<b>Modified k-NN + SVM</b>	<b>Modified k-NN + discriminant analysis</b>
$\mu$ (km)	0.604	0.548	0.5433
$\sigma$ (km)	0.363	0.441	0.4371

It is shown from table 1 that both k-NN and discriminant analysis algorithms give better outcome than the manual classification. The k-NN classification provides the smallest average traveling distance ( $\mu$ ) while keeping the low standard deviation ( $\sigma$ ). Smallest  $\mu$  represents that most voters are allocated to the nearest election unit. Low  $\sigma$  explains that the distribution in each election unit is appropriate. In other words, some voters are not too far while others are too close to their election unit.



**Table 2** The number of voters per election unit for each method.

<b>Election unit</b>	<b>Manual</b>	<b>Modified k-NN + SVM</b>	<b>Modified k-NN+Discriminant Analysis</b>
1	300	300	301
2	513	700	700
3	699	700	700
4	677	700	700
5	645	537	558
6	<b>743</b>	700	700
7	522	579	651
8	579	470	466
9	218	279	287
10	539	700	700
11	495	358	375
12	452	359	244

Table 2 illustrates the number of voters for 12 election units. It is clearly seen that the manual classification cannot satisfy the size constraint. That is, the election unit 6 has 743 voters which are larger than the limit (700). In turn, both two tested algorithms satisfy this condition.

### **Discussion and Conclusion**

This work proposed the combination algorithms to solve the problem of classification under size constraint which is applied to the allocation of voters to election units. The proposed algorithm is the integration of modified k-NN and discriminant analysis. The methods is proved to be efficient in terms of reducing average traveling distance of voters and keeping the number of voters for each election unit under limit. We expect to have further development for an improvement of our algorithm to have better efficiency and less complexity. Another is more experiments with different supervised classification algorithms to find the better integration and comparison. For the future work we look forward to use more classification methods for better results.



## References

1. Paul H (2012) A review and comparison of classification algorithms for medical decision making. *Health Policy*: 315-331.
2. Veenman C, Reinders M (2005) The Nearest Subclass Classifier: A Compromise between the Nearest Mean and Nearest Neighbor Classifier. *Pattern Analysis and Machine Intelligence*:1417-1429.
3. Sahiner B, Heang-Ping C, Hadjiiski L (2007) Classifier Performance Estimation Under the Constraint of a Finite Sample Size: Resampling Schemes Applied to Neural Network. *Neural Network*:1762-1766.
4. Zhu, M., & Martinez, A. M. (2006). Subclass discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*:1274 –1286.
5. Yongzhu X, Zhengdong Z, Feng C (2010) Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *Computer application and System Modeling*: 52-56.
6. Yeran S, Yunyan Y (2010) An Effective Bayesian Neural Network Classifier with a Comparison Study to Support Vector Machine. *Artificial Intelligence and Computational Intelligence*: 8-12.
7. Idbraim S, Aboutajdine D, Mammass D, Ducrot D (2009) Classification of Remote Sensing Imaging by an ICM Method With Constraints: Application in Land Cover Cartography. *Multimedia Computing and Systems*: 472-477.
8. Nattapong M (2012) Classification under class size constraint: application to electoral districting. *ICSTE Embedded Systems and Intelligent Technology*: 23