Analyzing the Distribution of a Large-scale Character Pattern Set Using Relative Neighborhood Graph

Masanori Goto*, Ryosuke Ishida†, Yaokai Feng† and Seiichi Uchida†
*GLORY LTD., Hyogo, Japan
Email: gotou.masanori@mail.glory.co.jp
†Kyushu University, Fukuoka, Japan
Email: uchida@ait.kyushu-u.ac.jp

***Abstract*—The goal of this research is to understand the true distribution of character patterns. Advances in computer technology for mass storage and digital processing have paved way to process a massive dataset for various pattern recognition problems. If we can represent and analyze the distribution of a large-scale character pattern set directly and understand its relationships deeply, it should be helpful for improving character recognizer. For this purpose, we propose a network analysis method to represent the distribution of patterns using a relative neighborhood graph and its clustered version. In this paper, the properties and validity of the proposed method are confirmed on 410,564 machine-printed digit patterns and 622,660 handwritten digit patterns which were manually ground-truthed and resized to 16 times 16 pixels. Our network analysis method represents the distribution of the patterns without any assumption, approximation or loss.**

I. Introduction

The purpose of this paper is to understand the "true distribution" of character patterns. For this ambitious purpose, we need to satisfy the following two requirements. First, we need to prepare real ground-truthed character patterns as many as possible. It is also important to prepare patterns having different properties, such as handwritten patterns and machine-printed patterns, for understanding their different distributions. Second, we need to use an analysis tool which is free from any assumption and approximation, while it should provide multiple observations enough to understand both of intra-class distribution and inter-class relationship.

For the first requirement, we have prepared 410,564 machine-printed digit images and 622,660 handwritten digit images. All of them are ground-truthed carefully, binarized, and resized into 16×16 pixels. Although it is, of course, impossible to cover all possible digit patterns by them (in the 256-dimensional binary feature space), we believe the amount of our patterns is enough to understand the distributions of 10 character classes, i.e., digits.

For the second requirement, we represent the pattern distribution through network representations, where each node corresponds to a single or multiple patterns and edge shows some relationship such as neighborliness. Different from low-dimensional representations such as principal component analysis and multi-dimensional scaling, network representations will cause neither approximation nor representation error. In addition, network representations are different from parametric distribution representation, such as Gaussian mixture modeling, and thus free from any assumption.

As the network representations, we used relative neighborhood graph (RNG) and its clustered version (Clustered-RNG). RNG has suitable for representing the neighboring relationship among patterns, and consequently, among classes. Clustered-RNG, which is newly introduced in this paper, provides a rough view of RNG, without any loss at representing inter-class relationship.

The distribution of the massive machine-printed or handwritten digit pattern images is analyzed by RNG and Clustered-RNG representations at the following aspects. (i) Inter-class relationship, especially, the neighboring relationship among "multiple classes" (i.e., not just a class pair). (ii) Intra-class distribution represented by how patterns of a certain class are scattered in RNG. (iii) Difference of (Clustered-)RNG for different digit image datasets (i.e., handwritten and machine-printed). Most analysis results matches with our intuitive expectation — this fact is important because the expectation is proved by our network analysis in a far more reliable and objective way.

II. Related Work

Datasets containing massive patterns have become indispensable for pattern recognition. For example, Torralba et al. [1] prepared a huge dataset with 80 million images gathered from the Internet, and they showed that high recognition accuracy was achieved just by the simplest 1-NN (nearest neighbor) rule with their massive image dataset. The quantity of a character pattern dataset also tends to increase. The classic MNIST dataset contains only 70,000 handwritten digit patterns. Smith et al. [2] have done one of the largest-scale researches with 223,000 handwritten digit patterns. Nowadays, Uchida et al. [3] analyzed our 822,000 handwritten digit pattern dataset with 1-NN analysis.

In past researches, these datasets are used mainly focusing on recognition accuracy rather than distribution analysis (except [3]). Exceptionally, Uchida et al. [3] employed minimum spanning tress (MST) for representing the structure of the digit pattern distribution and analyzed the network of MST. Although a network analysis with MST can represent the structure of the pattern distribution, MST has a strong constraint that a network of MST is a tree. This constraint causes that the

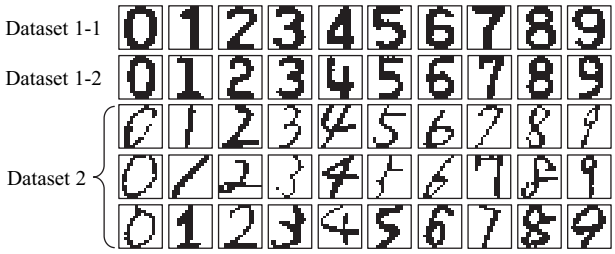


Fig. 1. Examples of digit images.

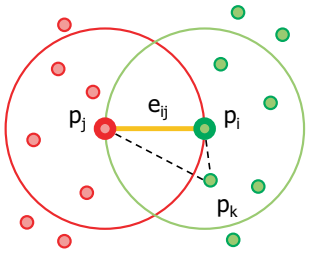


Fig. 2. An example of connected node pair (p_i, p_j).

nodes with the same class label come apart on the MST tree in case the patterns of the class have a great variation. Thus, instead of MST, we propose RNG network analysis, which is more suitable to understand the real distribution.

III. Experimental Setup for Pattern Distribution Analysis

In this paper, we deal with large-scale numeric character image pattern sets (digit patterns). For the purpose of analyzing the real pattern distribution, digit patterns possess the following merits over general image patterns. (i) Since there are only 10 classes for digit patterns, it is possible to have an enough number of patterns per class for understanding the precise distribution of each class. (ii) Small and binary character images can form a compact feature space. (iii) The classes of character patterns can be defined with far less ambiguity than visual objects.

A. Dataset

Our character image dataset is comprised of 410,564 machine-printed digit patterns (*Dataset 1*) and 622,660 handwritten digit patterns (*Dataset 2*). Fig. 1 shows several patterns from the dataset. All of the digit patterns were first isolated from their original scanned images and centered in the isolated image. Then the ground-truth, i.e., correct class label ("0",. . ., "9"), was attached to each pattern carefully by manual inspections by several professional operators.

1) Dataset 1: The images of Dataset 1 are machine-printed digit patterns of serial numbers of banknotes. Dataset 1 consists of two subsets (*Dataset 1-1* and *Dataset 1-2*). Each subset is comprised of patterns with the one kind of font which is issued in different countries. The distribution of these patterns is expected to be simple, because those machine-printed patterns are generally identical and just varied slightly. The main variation factors of machine-printed digit patterns are only dirt, blurred and binarization error. The number of machine-printed digit patterns is Dataset 1-1: 199,504, Dataset 1-2: 211,060. Only class "1" of Dataset 1-1 has 9,550 patterns, and the other classes have 21,106 patterns.

2) Dataset 2: The images of Dataset 2 are handwritten digit patterns written by a number of unknown people. Different from Dataset 1, Dataset 2 is comprised of patterns with large varieties. The number of handwritten digit patterns is 622,660 and all the classes have 62,266 patterns.

B. Feature and Distance Metric

Each pattern is represented as a 256-dimensional binary vector and thus corresponds to a corner of the 256-dimensional hypercube. [1]

As the distance metric, we employ Hamming distance. The Hamming distance can be interpreted intuitively. For example, if the Hamming distance is 25 between two 16×16 binary patterns, this indicates that the two patterns have different black/white value at 25 pixels (about 10% among 256 pixels).

IV. Relative Neighborhood Graph (RNG) and Clustered-RNG

A. RNG and Its Properties

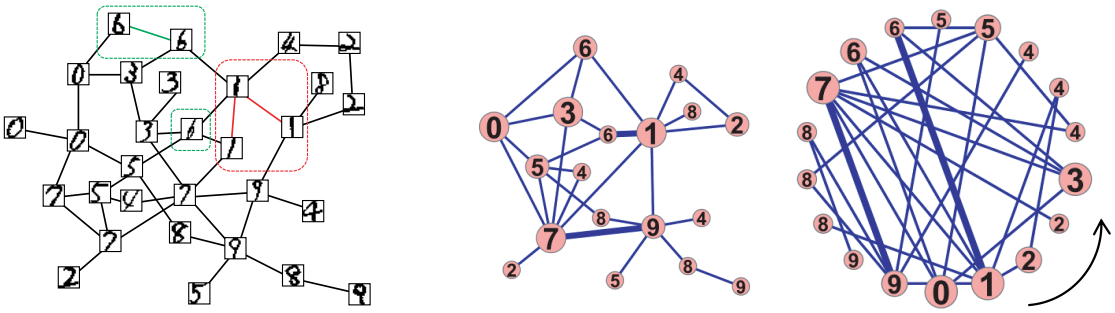
RNG [4], [5] is undirected graph where neighboring patterns tend to be connected by an edge. RNG is more intuitive than the nearest neighbor graph, which is a directed graph[2]. Because of this good property, RNG has been used for many researches, such as computer vision, geographic analysis, pattern classification, etc[6]. For pattern classification problems, Urquhart [7], Sanchez et al. [8] and Zighed et al. [9] proposed clustering methods with RNG. Ichino et al. [10] used RNG to select globally effective feature and achieve better classification performances. Different from those past trials, this paper utilizes RNG and its clustered version, called Clustered-RNG, for analyzing the distribution of massive character patterns.

Let $\boldsymbol{P} = \{p_1, p_2, ..., p_n\}$ and $\boldsymbol{E} = \{e_{ij}\}$ denote the sets of nodes and edges of an RNG, respectively. In this paper, each node corresponds to a character pattern represented by a d-dimensional feature vector. Each edge e_{ij} is prepared iff $d(p_i, p_j) \leq \max_{k=1,...n,k\neq i,j}[d(p_i, p_k), d(p_j, p_k)]$, where $d(p_i, p_j)$ is the distance between p_i and p_j. Fig. 2 illustrates this condition; intuitively, the edge $e_{i,j}$ exists iff there is no pattern in the intersection part between two hyper-spheres centered at p_i and p_j. The computational complexity for building an RNG is $O(n^3)$ [4]. Fig. 3 (a) shows a tiny example of RNG for handwritten digit patterns.

RNG has three properties suitable for analyzing the distribution of a large-scale pattern set. The first property is that the proximity between similar patterns is preserved. Consequently,

[1] Note that the analysis on this binary pattern distribution is directly related to a continuous distribution by some feature extraction. This is because most feature extraction methods are based on some linear operations and thus properties in the original binary pattern distribution are mostly preserved even in the continuous distribution.

[2] Even if X is the 1NN of Y, Y may not be the 1NN of X. Thus, NN graph is a directed graph.



(a) (b) (c)

Fig. 3. A tiny example of relative neighborhood graph (RNG) and its Clustered-RNG of handwritten digit patterns. (a) RNG of handwritten digit patterns, where Hamming distance was used for measuring the distance between to nodes. Nodes are shown with image patterns. (b) Clustered-RNG. Connected nodes with the same class label are clustered and shown as a circle symbol with the their label. The sizes of circle and the width of lines are proportional to the number of clustered nodes and the number of connection between each different clustered node. (c) Circular layout of Clustered-RNG. Clustered nodes are aligned as counter clockwise on the circumference as the number of clustered nodes.

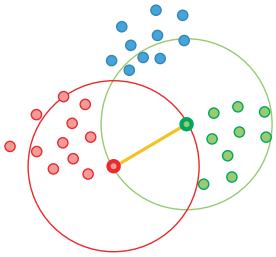


Fig. 4. An RNG edge linking nodes of distant patterns.

similar patterns will form a "cluster" on RNG. The cluster is defined as a set of connected nodes with the same class label. For example, even the small RNG of Fig. 3 (a), images of "0", "1", "3", and "7" form one cluster. The second property is that RNG is an undirected graph; this property is appropriate because generally a pattern distribution does not define any direction between patterns.

The third property of RNG is the most important and distinctive property for the purpose of this paper, i.e., the distribution analysis; namely, RNG can have edges between nodes of distant patterns. Consider patterns of Fig. 4. If edges are prepared only for nearest neighbor pattern pairs, only the nodes with the same color are connected by the edge. In contrast, according to the rule of RNG, there is a chance that the nodes with different colors are connected (as indicated by an orange edge). This edge is very useful; if we consider that Fig. 4 shows a pattern distribution of three classes, the edge represents inter-class boundary. Consequently, using RNG, we can analyze inter-class boundary. In fact, this property will be fully utilized in the later analysis.

B. Clustered-RNG and Its Properties

We also use Clustered-RNG, which is a compressed representation of RNG. As shown Fig. 3 (a), RNG has a node for each individual pattern and thus it is practically impossible to plot an RNG for a massive pattern set. In addition, even though we want to focus inter-class relationship by inter-class edges, it is also difficult; this is because such edges are massive and scattered in the RNG.

Clustered-RNG converts a set of connected RNG nodes with the same class label into a new node. As shown in Fig. 3 (b), three connected nods of "1" on RNG becomes a single node in the Clustered-RNG. On the other hands, the class "6" has two nodes in the Clustered-RNG, because one "6" is not connected directly to the other two "6"s, which are mutually connected in the RNG. In the following experiment, we will show a Clustered-RNG using circular layout as shown in Fig. 3 (c).

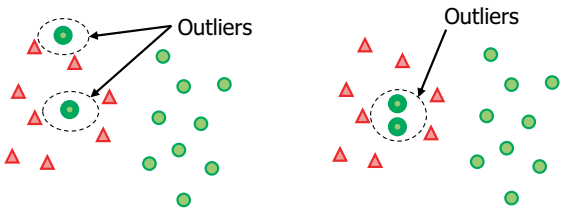
Clustered-RNG inherits the good properties of RNG mentioned above and also have merits by itself. The properties of the Clustered-RNG are summarized as follows.

Property (i): Clustered-RNG reveals inter-class relationship. An important point is that Clustered-RNG does not loose any edge of inter-class boundary during its building process. Thus, by observing Clustered-RNG we can understand the inter-class relationship without any loss. It is interesting to note that, as indicated by Fig. 4, Clustered-RNG as well as RNG will have the inter-class boundary edge between "distant" classes when there is no obstacle between the classes. Also note that the those inter-class edges are strongly related to support vectors by support vector machines (SVM).

Property (ii): Clustered-RNG reveals multi-class relationship. For example, in case of Dataset 1-1(Fig. 1), the class pairs "3"-"8" and "6"-"8" have the closer neighboring relationships and class "8" relay these neighboring relationships of the class "3" and "6". [3] This multi-class neighboring relationship will be useful, for example, for designing a multi-class recognizer by using a set of two-class classifiers (like SVM). In the above three-class case, we can understand that it is better to use a 3-vs-others (6 and 8) classifier and then a 6-vs-8 classifier instead of a 3-vs-6 classifier.

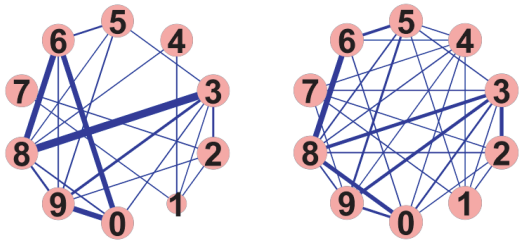
Property (iii): Clustered-RNG detects outliers. As shown Fig. 5 (a), we can detect outliers by measuring the nearest neighbor distance to the same class. However, for the case of (b), we cannot detect them because the two outliers are close

[3]It is also interesting to note that if class "8" is removed, some inter-class edges will appear between "3" and "6". This means that all the inter-class edges are determined by considering those three classes, even though each individual edge only links two classes. This fact also indicates that Clustered-RNG shows multi-class relationship.



(a) (b)

Fig. 5. Outliers can be detected (a) or cannot be detected (b) by the nearest neighbor distance to the same class.



(a) (b)

Fig. 6. Clustered-RNG. A thicker line between a pair of clustered nodes indicates that there are more edges between them. (a) Dataset 1-1. (b) Dataset 1-2.

to each other. [4] Clustered-RNG detects both of these outliers as other clustered nodes, because only the connecting nodes with same class label is clustered on a RNG network. According to increase of the patterns, outliers will increase and distribute closer to each other. Thus, this property is useful to analyze the distribution of a large-scale pattern set.

V. Experiment and Network Analysis

A. Printed Digits (Dataset 1-1, 1-2)

We first confirmed several properties of network analysis using RNG and Clustered-RNG, through the experiments with the large-scale dataset of machine-printed digit patterns. The machine-printed digit patterns are expected to have a simplest pattern distribution.

Fig. 6 (a) shows the Clustered-RNG of Dataset 1-1 (199,504 digit patterns). It was observed that each class forms only one clustered node. This means that machine-printed digit patterns from a class are very similar to each other. In fact, it is rather surprising that even in such a large dataset, there is no "lonely" pattern (like "9" at the rightmost in the RNG of Fig. 3 (a)) and no "isolated" small pattern set.

Fig. 6 (a) also shows that several class pairs (e.g., "0" and "1") had no edge and similar class pairs had many edges. Actually, the class pair "3" and "8" had 2,214 edges. It is interesting to note that the class pairs "3"-"8" and "6"-"8" had many edges, but the class pair "3" and "6" had no edge. It means that class "8" relays the neighboring relationship of class "3" and "6". This observation proves that Clustered-RNG can represent the neighboring relationship among multiple classes directly. (One may consider that the confusion matrix can show such a relationship — this consideration is wrong as proved later.)

The edge width also provides knowledge on the inter-class relationship of the pattern distribution. A thicker edge between "3" and "8" represents that there are many edges between those classes and thus the distributions of "3" and "8" are facing each other "widely".

Fig. 6 (b) shows the Clustered-RNG of Dataset 1-2 (211,060 digit patterns). Although its shape was similar to the Clustered-RNG of Dataset 1-1, there were more edges than Fig. 6 (a). This fact indicates that the patterns of Dataset 1-2 are distributed a bit wider than that of Dataset 1-1. Probably, the reason of it is that Dataset 1-2 has more binarization errors caused from the design of background. Several images of the nodes connecting the clustered node pair "4" and "5" are shown Fig. 7. (Note that there is no edge between "4" and "5" in Fig. 6 (a).) The influence of binarization errors are clearly observed from those images.

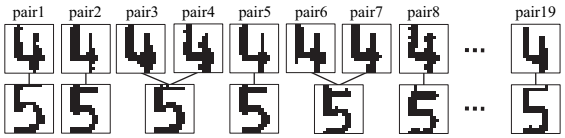


Fig. 7. The images of the node connecting the clustered node pair "4" and "5" on the Clustered-RNG of Dataset 1-2

B. Handwritten Digits (Dataset 2)

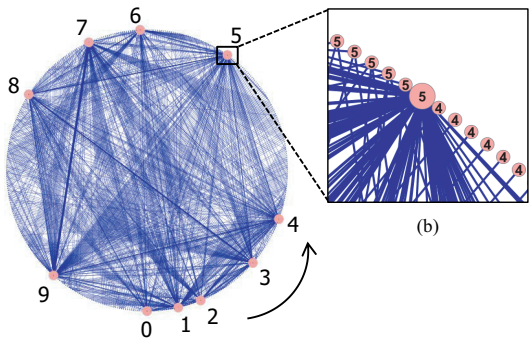
Fig. 8 shows the Clustered-RNG of Dataset 2 (622,660 digit patterns) and it is largely different from the Clustered-RNGs for machine-printed digits. Each of all the classes was separated to several clusters. Generally, the size of a clustered node was very large or very small. Specifically, each class had only one very large clustered node and other very small clustered nodes. The total number of the small cluster nodes was 370.

The small clustered nodes were comprised of patterns with large deformations. Fig. 9 shows several image examples of the small clustered nodes. This fact indicates that the Clustered-RNG is useful to extract "outliers" automatically as a small cluster (i.e., a small pattern set) surrounded by nodes of other classes.

Clustered-RNG represents the complex distribution of the large-scale handwritten digit patterns as a simple network. Fig. 10 shows a simplified version of Fig. 8 by only plotting large clustered nodes. Different from Fig. 6 (a) and (b), this graph is a complete graph; this means that for handwritten digits, any pair of digit classes have some neighboring relationship in their distributions. Most of the class pairs had many edges more than Clustered-RNG of the machine-printed patterns (e.g., class pair "7" and "9" had 36,287 edges).

A deeper analysis on Fig. 10 reveals that 34,935 nodes of class "9" (about 56%) had an edge connected to the node with different label. This fact indicates that many nodes are

[4] Even if we use kNN for outlier detection, it is useless if the more than k outliers are close to each other.



(a)

Fig. 8. Clustered-RNG of Dataset 2. (a) Clusters are aligned as counter clockwise on the circumference as the number of clustered nodes. (b) Enlarged view around the large cluster of class "5". The outliers are shown as small clusters.

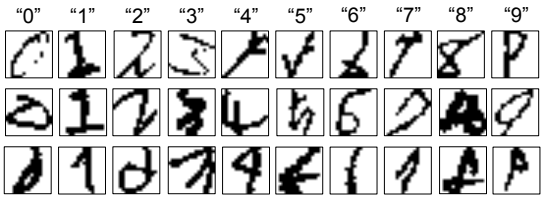


Fig. 9. Image examples from the small clustered nodes.

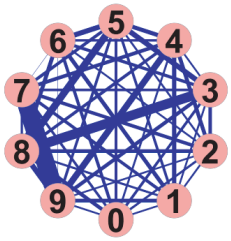


Fig. 10. Clustered-RNG with large clustered nodes of Dataset 2. A thicker line between a pair of clustered nodes indicates that there are more edges between them.

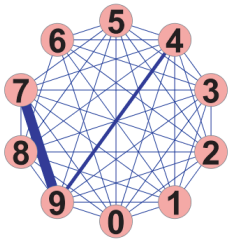


Fig. 11. The network of the class relationship by 1-NN classification of Dataset 2. A thicker line between a pair of nodes indicates that there are more misrecognitions between them.

located around the class boundary and the distribution is much complex. Even with such a complex distribution, the inter-class complex relationship is compressed and represented simply on Clustered-RNG network.

C. Difference from Confusion Matrix

Although a conventional confusion matrix also can provide knowledge on inter-class neighboring relationship, it is less informative than our RNG representation. Fig. 11 is a graph created from the confusion matrix of the 1-NN classification of Dataset 2. It is much simpler than Fig. 10. As shown in Fig. 4, the Clustered-RNG can show the "distant" neighboring relationship and also multi-class relationship and thus will have more edges. The pattern distribution of Dataset 2 is complex as observed above, but the network of 1-NN classification was rather simple. (i.e., a smaller number of thick lines). For example, class "0" had only 100 or less misrecognition patterns for each class pair. Clustered-RNG represents the distribution of the patterns without any loss at representing the inter-class relationship.

VI. Conclusion and Future Work

This paper proposes the method to represent and analyze the distribution of the patterns using RNG and Clustered-RNG without any assumption, approximation or loss. The properties and validity of this Clustered-RNG analysis is confirmed on 410,564 machine-printed digit patterns and 622,660 handwritten digit patterns. Clustered-RNG analysis represents the complex distribution of the patterns as a network of clustered nodes simply, and reveals not only inter-class relationship but also multi-class relationship of the pattern distribution. In addition, Clustered-RNG analysis extracts outliers automatically as a small cluster. As future work, the intra-class distribution of patterns within the clustered nodes of Clustered-RNG will be studied, such as how patterns of a certain class are scattered in RNG. Further analysis of the relationship between inter-class edges of RNG and support vectors of SVM will be interesting. It is also interesting to observe how RNG changes by the use of a kernel function in the distance metric. It is promising to observe the intrinsic dimensionality of pattern distribution via "graph embedding" into a higher dimensional space.

References

[1] A. Torralba, R. Fergus, and W. T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," IEEE Trans. PAMI, vol. 30, no. 11, pp. 1958-1970, 2008.

[2] S. J. Smith, M. O. Bourgoin, K. Sims, and H. L. Voorhees, "Handwritten Character Classification Using Nearest Neighbor in Large Databases," IEEE Trans. PAMI, vol. 16, no 9, pp. 915-919, 1994.

[3] S. Uchida, R. Ishida, A. Yoshida, W. Cai, Y. Feng, "Character Image Patterns as Big Data," Proc. ICFHR, pp. 477-482, 2012.

[4] G.T. Toussaint, "The Relative Neighbourhood Graph of a Finite Planar Set," Pattern Recognition, vol. 12, pp. 261-268, 1980.

[5] J. O'Rourke, "Computing the relative neighborhood graph in L_1 and L_∞ metrics," Pattern Recognition, vol. 15, no 3, pp. 189-192, 1982.

[6] J.W. Jaromczyk, G.T. Toussaint, "Relative neighborhood graphs and their relatives," Proc. IEEE, vol. 80, no 9, pp. 1502-1516, 1992.

[7] R. Urquhart, "Graph Theoretical Clustering Based on Limited Neighborhood Sets," Pattern Recognition, vol. 15, no 3, pp. 173-187, 1982.

[8] J.S. Sanchez, F. Pla, F.J. Ferri, "On the Use of Neighborhood-Based Non-Parametric Classifiers," Pattern Recognition Letters, vol. 18, pp. 1179-1186, 1997.

[9] D.A. Zighed, S. Lallich and F. Muhlenbach, "A Statistical Approach to Class Separability," Applied Stochastic Models in Business and Industry, vol. 21, pp. 187-197, 2005.

[10] M. Ichino, H. Yaguchi, "An Apparent Simplicity Appearing in Pattern Classification Problems," Pattern Recognition, vol. 33, pp. 1467-1474, 2000.