CHAPTER 1

Introduction

1.1 Background and motivation

For decades, many researchers have tried to create a new algorithm for classification and partitioning of objects, images, etc. One popular algorithm is the so-called Nearest Neighbor. The nearest neighbor (NN) was introduced by Fix and Hodges in 1951 [1], the NN rule gained enormous popularity after 1967 when some of its formal properties were described by Cover and Hart [2]. The main concept of the nearest neighbor method is to find the minimum or the closest distance from a test sample to the training samples and then predict the corresponding label of the test sample. This algorithm is easy to use, provides quite high accuracy rates, and can be applied practically in a variety of fields. In pattern recognition, the nearest neighbor [3] is used to predict the labels of test samples from training samples. Euclidean distance is normally used to find the nearest distance between two training samples $\vec{x}$ and $\vec{y}$ as

$$\left\|\vec{x}-\vec{y}\right\| = \sqrt{\sum_{i}^{n}(x_i - y_i)^2} \tag{1.1}$$

where $\vec{x} = (x_1, x_2, ..., x_n)$ and $\vec{y} = (y_1, y_2, ..., y_n)$. For an efficiency trick, we can avoid computing the squared root by computing

$$\left\|\vec{x}-\vec{y}\right\|^2 = \sum_{i}^{n}(x_i - y_i)^2 \text{ instead.} \tag{1.2}$$

In many situations, the feature vector $\vec{x}$ and the true class C_i to which it belongs to, the underlying joint distribution of the observation may not have complete statistical knowledge.

However, K-nearest neighbor rule [1, 2] is a nonparametric pattern classifier, which is simple, yet yields good classification accuracy.

For a test sample, the K-nearest neighbor rule is that a test vector is assigned to the class to which the majority of its K-nearest neighbors belong to. The advantages and disadvantages of K-NN are described below [4]:

Advantages

- Easy to use.
- Robust to noises in the training data if the inverse square of weighted distance is used as the "distance" measure.
- Efficient when the training data are large.

Disadvantages

- Calculating distances from each test sample to all training samples causes the waste of time and computation cost is quite high.
- The need for greater accuracy required for data training consumes a lot of memory.
- Low accuracy rate when computed in multidimensional datasets.
- For more accuracy, it is needed to determine the appropriate value of parameter K (the number of nearest neighbors).
- It is difficult to choose an appropriate choice of distances to use for the algorithm which depends on the type of data.

Fuzzy K-nearest neighbor is another popular classification algorithm. The fuzzy form provides results differently from the crisp version. The advantage of the algorithm is that there is no arbitrary assignment. In addition, the membership values of a sample vector in a particular class provide a level of assurance to accompany the true classification. For example, if we assign the membership value in one class as 0.95 and in another class as 0.05, then we might be able to say that the vector belongs to the first class with the highest degree than the second class. Likewise, if a vector has the

membership value in class one as 0.54 and its membership values in class two and three are 0.45 and 0.01, then we can say that the vector does not belong to class 3 and it might be around the boundary of the first 2 classes.

Although fuzzy K-Nearest Neighbor is a popular algorithm, it cannot be used directly in the case of structural datasets, Syntactic pattern recognition or structural pattern recognition should be more appropriate. Syntactic pattern recognition is a form of pattern recognition, in which each object is represented by a variable-cardinality set of symbolic, nominal features.

Moreover, syntactic pattern recognition can be used instead of statistical pattern recognition if there is a clear structure in the patterns. One way to present such structures is by means of strings of symbols from a formal language.

As a result, in order to solve the above-mentioned problems, we would like to present a new technique for classification and partitioning of objects or images, by incorporating uncertainty into string grammar K-nearest neighbor.

1.2 Literature review

The research works involving to the incorporating uncertainty into string grammar K-nearest neighbor in recognition and object detection algorithm are reviewed in this section.

1.2.1 K-Nearest Neighbor

Cunningham, P. and Delany, S. J. [5] presented a new technique for nearest neighbor classification focusing on, mechanisms for assessing similarity (distance), computational issues in identifying nearest neighbor and mechanisms for reducing the dimension of the data.

Nivre, J. [6] proposed the new method that used the K-nearest neighbor approximation where the distance is measured by the Euclidean distance between the points in n-dimensional space.

Charles, E. [7] proposed the new method for predicting the class of a test sample using K-nearest neighbor which is used for computing the closely distance between the testing sample string and all training sample strings.

1.2.2 Fuzzy K-Nearest Neighbor

Keller, J. M., Gray M. R. and Givens Jr., J.A. [8] presented the new method of K-nearest neighbor in fuzzy term and three methods of assigning fuzzy memberships to the labeled samples. The experimental results are compared to the crisp version.

Keller, J. M. and Hunt, D. J. [9] presented the new method that is guaranteed a solution converges in a finite number of steps of the linearly separable or non-linearly separable dataset.

1.2.3 String grammar

Phitakwinai, S., Auephanwiriyakul, S. and Theera-Umpon, N. [10] proposed an automatic Thai finger-spelling sign language translation system using FCM and SIFT algorithm. They build a dynamic hand gesture translation system with video caption without prior hand region segmentation. In particular, they apply the fuzzy C-means algorithm to find a representative frame that represents a set of frames with little significant difference forming a scene or shot. After that, they also apply the scale invariant feature transform (SIFT) in the recognition process of the dynamic hand gesture in the Thai finger-spelling sign language for 10 words, i.e., “Far”, “Good”, “Leaf”, “Meet”, “Fire”, “Rich”, Erase”, “Float”, “Color”, and “Comb”.

Bacon, D. [11] proposed a method with a useful form for dealing with context free grammars (CFG) (the Chomksy normal form). This is a particular form of writing a CFG which is useful for understanding CFGs and for proving things about them.

Fu, K.S. [12] proposed the phrase structure grammars describing patterns in syntactic pattern recognition. Each pattern is represented by a string of primitives which corresponds to a sentence in a language (tree or graph in

high dimensional grammars). All strings which belong to the same class are generated by one grammar.

Lee, M. H., Kim, S. H., Lee, G. S., Kim, S. H. and Yang, H. J. [13] proposed a correction method for misrecognition of Korean Texts in signboard images using improved Levenshtein metric. The proposed method calculates distances of five recognized candidates and detects the best match texts from signboard text database. For verifying the efficiency of the proposed method, a database dictionary is built using 1.3 million words of nationwide signboard through removing duplicated words. They compared the proposed method to Levenshtein Metric which is one of the representative text string comparison algorithms. As a result, the proposed method based on improved Levenshtein metric represents an improvement in recognition rates 31.5% on average compared to that of conventional methods.

Liu, H. H. and Fu, K. S. [14] proposed two syntactic methods for the recognition of seismic waveforms. The seismic waveforms are represented by strings of primitives. Primitive extraction is based on cluster analysis. Finite-state grammars are inferred from the training samples. The nearest-neighbor decision rule and error-correcting finite-state parsers are used for pattern classification. While both show equal recognition performance, the nearest-neighbor rule is much faster in computation speed. The classification of real data for earthquake/explosion is presented as an application example.

Heryadi, Y., Fanany, M. I. and Arymurthy, A. M. [15] proposed a simple and computationally efficient framework for 3D dance basic motion recognition based on syntactic pattern recognition.

Clark, A. [16] proposed the new algorithm for inducing a context-free grammar from a set of strings. This algorithm comes with a strong theoretical guarantee, it works in polynomial time and for any grammar in a certain class it will converge to a grammar which is isomorphic/strongly equivalent to the target grammar.

From literature review above, some papers cannot use in the structural data and used the old technique for the classification.

Thus, we propose the method that can handle synthetic pattern recognition and incorporating uncertainty into the string grammar K-nearest neighbor. It can be ameliorated by generating the proposed the membership functions based on the structured data using the fuzzy set theory or the possibilistic theory. This technique is designed for working with structured data to attack the disadvantage of K-NN algorithm and being able to handle validity index problem to give a high accuracy rate of classification, and provide a clear distance based learning.

1.3 Research objectives

To develop a novel classification method by incorporating uncertainty into string grammar K-nearest neighbor.

1.4 Research scopes and method

1.4.1 Research Scope

1.4.1.1 Implement the algorithm on synthetic and at least 10 standard public datasets.

1.4.1.2 Compare the result with its numeric counterpart.

1.4.2 Research Method

1.4.2.1 Study the theories and review the literatures

1.4.2.2 Collect the datasets

1.4.2.3 Design algorithms by incorporating uncertainty into string grammar K-nearest neighbor method and develop software

1.4.2.4 Test and improve the performance of algorithm

1.4.2.5 Collect the experimental results and find the optimal parameters of the algorithm

1.4.2.6 Discuss, conclude and write a thesis

1.5 Educational advantages

To obtain a novel classification method by incorporating uncertainty, e.g., using the fuzzy set theory or the possibilistic theory, into string grammar K-nearest neighbor.

1.6 Research location

The research is conducted at the Computational Intelligence Research Laboratory (CIRL), Faculty of Engineering, Chiang Mai University, Thailand.

1.7 Thesis organization

The thesis consists of 5 chapters. In chapter 2, we describe how to incorporate uncertainty into string grammar K-nearest neighbor. Chapter 3 explains the research designs and the proposed methodology of incorporating uncertainty into string grammar K-nearest neighbor. Chapter 4 describes the experimental results and discussion of the

proposed method on the standard real-world datasets. Finally, conclusions are stated in chapter 5.