CHAPTER 1

Introduction

1.1 Background and Motivation

A pattern recognition has been wildly applied in many fields for a long time. It is an automatic system that aims to classify an input pattern into a suitable class by using several techniques. Techniques to solve pattern recognition can be grouped into two main fundamental approaches, i.e., a numerical (or decision theoretic) approach and a structural (or syntactic) approach [1]. Both approaches use different methods to implement description and classification tasks. The numerical approach is suited for tasks where patterns can be represented in vector form. The limit of this approach is the length of feature vectors must be fixed, so the length of each data must be the same for all data in the dataset. Moreover, this approach makes it difficult to classify among groups based on the structure of a pattern because it lacks a suitable process for handling structures and their relationships. In the past several years, the structural pattern recognition [1-4] has been receiving increased attention because it has the structure handling ability lacked by the numerical approach.

Our research treats the problems of pattern recognition and its applications by using the syntactic approach. The syntactic pattern recognition is the one of the important approaches in pattern recognition field, in which each data can be represented by sets of simple pattern primitives and grammatical rules. In particular, a sentence, e.g., a string or a tree or a graph, in a language is utilized for describing a pattern in syntactic approach [1-4]. These pattern primitives are utilized for representing pattern structures, taking into account more complex pattern such as in pictorial data analysis. The sample of a string representation is shown in Figure 1.1. These are equilateral triangles structures of various sizes. In this case, the length of primitives *a, b* and *c* are equal, we can represent such triangles by string of the form *aaaa….bbbb…cccc* [1]. String representation is sufficient for describing the objects which structure is based on relatively simple connectivity of primitives. Another example of string representation is shown in Figure 1.2 (b) which is the result of the encoded submedian chromosome structure in term of primitive defined in Figure 1.2 (a). Thus, the submedian chromosome can be represented by string *abcbabdbabcbabdb* [1].

The structural approach has two main advantages. Firstly, this approach can use small set of simple primitive for describing the large set of complex patterns [1-4] and another advantage is these string grammars can be transformed back to original image as well.

String is one of the representations in structural pattern recognition. We use a string rather than a feature vector because it has some advantages over than using feature vector. Specifically, such as the length of each string in set is variant and depends on the individual pattern under consideration, while we always use the same number of features in a feature vector, no matter how simple or complex a pattern is [1-4].

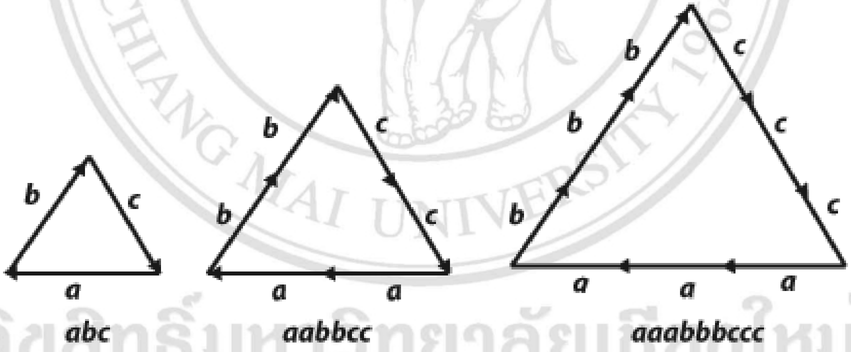

Figure 1.1 String representations of equilateral triangles.

In pattern recognition, clustering [5] is a widely used technique in the unsupervised learning field of Machine Learning, such as data mining, pattern recognition, image processing, etc. The clustering algorithm divides a set of samples into groups in such a way that the members of the same group are more similar to the members of other groups. This method can be natural for clustering of compact and well-separated groups of data. However, in practice, each data of clusters might be overlapped, and some data vectors belong partially to several clusters. There are several clustering methods available. The hard or crisp clustering methods restrict that each data belongs to exactly one cluster.

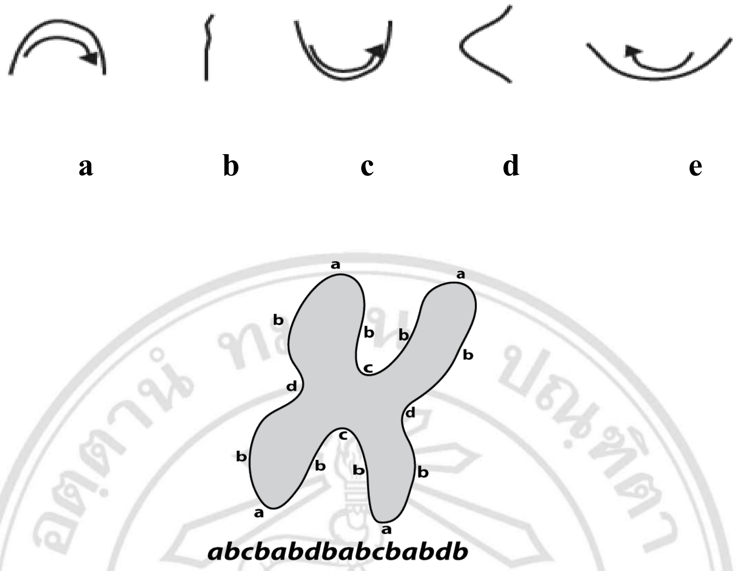

Figure 1.2 (a) Primitives (b) String representations of submedian chromosome.

In 1965, Zadeh gave the idea of fuzzy clustering [6] that provides more information by adding membership function. Fuzzy set theory is applied to use in cluster algorithm for a long time such as in work of [7] and [8]. Nowadays, fuzzy clustering has been widely used in a variety of areas and become the important tools to cluster analysis. Fuzzy clustering has been beneficial over hard clustering. There are several fuzzy clustering methods are proposed. Fuzzy C-means (FCM) clustering algorithm has been effectively applied to several problems [9]. However, the serious disadvantage of this method is that it cannot detect noisy data and outliers in data. There are several methods to solve the problem of FCM, including the possibilistic C-means (PCM) [10], fuzzy C-medians (FCMed) [11, 12], Fuzzy Possibilistic C-means (FPCM) [13], Possibilistic Fuzzy C-means (PFCM) [14], Unsupervised Possibilistic C-means (UPCM) [15], and Unsupervised Possibilistic Fuzzy C-means (UPFCM) [16], which solves the noise sensitivity defect of FCM. These methods only have been used in numeric method or vector form.

The string grammar clustering can be applied using the clustering techniques mentioned above. The objective of string grammar cluster analysis is to assign each string observation to cluster so that string observations within each group are more similar than one another. However, to measure the distance between two strings, we cannot use numeric distance metrics such as Euclidean distance or Manhattan distance. Since strings are not numeric vectors, we need to take the appropriate measure for string. The standard string distance has been widely used as the Levenshtein distance [17-19]. The Levenshtein distance function measures the minimum number of edit operations, i.e., insertion, deletion and substitution of symbols required to transform one string into the other string. This distance can be computed in the quadratic time with respect to the length of the two strings under consideration. The most analytic string clustering approach is the string grammar hard C-means (sgHCM) algorithm [2, 19]. The sgHCM is an extension of the ordinary hard C-means algorithm (HCM). As we all know, fuzzy clustering has benefited more than crisp clustering. The main drawback of string grammar hard clustering method is partitioning of the string data such that each string belongs to exactly one of the partitions, so the accuracy and efficiency of hard clustering have been very poor for some applications. Hence, we would like to propose a new fuzzy clustering technique for domain of string. This technique is designed for working with syntactic pattern, and eliminating the drawback of string grammar hard clustering method. Moreover, we also compare our results indirectly with the results from algorithms run on numeric vectors of these data sets.

1.2 Literature Review

This section reviews the research works on string grammar clustering and the mathematical theories related to these works. It begins with some researches related to syntactic pattern recognition, string grammar clustering, median string, and clustering based on the Levenshtein distance.

Fu, K. S. [2] proposed the definition of grammars to describe patterns in structural pattern recognition. String primitive is utilized for each pattern of dataset which relates to a sentence in a language. One grammar will generate all strings which belong to the same class. In addition, this string grammar is applied to Chinese handwriting application using Freeman chaincode and sgHCM algorithm.

Fu, K. S., et al. [17] proposed to a new distance measure for structural patterns in terms of error transformations. The nearest neighbor is used for structural pattern recognition with respect to samples in dataset and then proposed the suitably cluster algorithm procedure for structural pattern.

Bezdek, J. C., et al. [19] proposed a string grammar Hard C-means (sgHCM) which is an extension of the nearest neighbor and Hard C-means algorithms. The Levenshtein distance is used as a dissimilarity measure.

Cha, S.H., et al. [20] proposed to use contour analysis for string matching algorithm. The method reduces dimensional of image into a one-dimensional string and then measures the distance using the Levenshtein distance. Moreover, they applied the method to handwriting analysis for both online and offline character recognition.

Kruzslicz, F. [21] proposed a new method for finding the median string of a set using a new greedy median string algorithms. The method is a fast heuristic algorithm that provide an effectively approximation of the median string of a set.

Juan, A., and Vidal, E. [22] applied the normalized edit distance based on k-nearest neighbours. They claim that the use of these techniques give better results than other methods that have been applied to human banded chromosomes classification.

Martinez, C.D., et al. [23] improved the calculate of prototype of string classification using approximate median string. The approximate median string applies the edition operations, i.e., insertion, deletion and substitution over each symbol of the string, and then calculate the sum of the distance edition to each string. This algorithm was iterated until stable.

Zhang, C., et al. [24] presented a new distance for shapes matching based on the contour analysis which called shape edit distance (SED). In this method convert shapes to string sequences. A maximum a posteriori probability (MAP) method is utilized for string sequences alignment and then find the sequence that minimizes the string edit operation cost using shape edit distance. This method is successfully applied to image retrieval and clustering problems.

Neuhaus, M., and Bunke, H. [25] proposed the hybrid method using the kernel function based on the Levenshtein distance for syntactic pattern classification. The proposed approach is appropriate to both string and graph representations of patterns. Then, the syntactic classification is done on those kernels. Although the classification rate in this case is very good, the validity of the kernel method cannot be established.

Yeh, M.C., and Cheng, K.T. [26] presented an approach for scene classification and shape retrieval based on Levenshtein distance. The method extracts local features sequences from original image and then perform feature alignment using the Levenshtein distance.

Zhangl, S., and Wang H. [27] applied Levenshtein distance to the hand language video dataset. The hand feature vectors are converted to strings, and then matching the hand video using sgHCM, the test hand video string is assigned to the closest prototype belongs to according to the Levenshtein distance.

Deng, J., et al. [28] introduced a method for feature extraction based on fuzzy matching for genome data clustering. They translate the text data into vector of numeric numbers and inputs to FCM clustering. The Levenshtein distance is used to determine whether two original texts are in the same cluster.

Patil, N., et al. [29], improved fuzzy C-means clustering for text clustering and the Levenshtein distance algorithm is used to calculate Levenshtein distance between candidate and query sequences is used as input in the FCM.

Chanda, P., et al. [30] proposed a Thai sign language translation system using the upright speed-up robust feature (U-SURF) and the fuzzy C-means (FCM) for 42 words. They compared the result of the method with that from string grammar hard C-means.

Shen, W., et al. [31] presented the method for shape matching using an agglomerative hierarchical algorithm. This clustering method extracts the structure which captures the within-class structural information of the cluster of shapes. Moreover, this method is more robust distance measure between clusters.

He, Y., et al. [32] proposed the method for describing road marking using local junction feature. Each road marking is described by a junction string. those junctions are encoded within a range as the same code. A weighted edit distance is used to measure the similarity between detected junction string and ground truth junction string, and assign different deviation with different weight.

Putra, M. E. W., et al. [33] proposed the method for off-line structural handwriting character recognition. Graph structure is utilized for analyzing curve of character's structure. Moreover, for classification, an approximate subgraph matching is compared with Levenshtein distance. The Levenshtein distance give better performance than approximate subgraph matching in recognition accuracy.

Khan. S., et al. [34] Proposed a method for vehicle plate matching using license plate recognition based on modified Levenshtein Distance. The key module of this technique is Precision-Recall curve, which contains the conditional probabilities of observing one character at one node for a given observed character at an additional station. Therefore, the evaluation of the performance constraint relies on the by hand extracted position truth of a large number of plates, which is an unwieldy and deadly process. To beat this negative aspect, in this cram, they propose an inventive novel LPM-MLED (License Plate Matching - Modified Levenshtein Edit Distance) method that removes the need for extracting ground truth by hand.

Ratnasari, C.I., et al. [35] proposed the method for checking misspelled word for patient Complaints in Bahasa Indonesia. The data are stored in the form of free-text data or a medical narrative by the doctor when taking the medical history or conducting the medical interview. Based on data on patient complaints obtained from physicians, this study develops lexicon resource which is used as spell detection, and then using the Levenshtein distance for spell correction.

Kopel, M. [36] proposed a new method for automatically retrieve harmony from song by finding the shortest commonly repeated chord progression, which may be a riff. The Levenshtein distance is utilized to measure the dissimilarity between sequences.

From literature reviews above, these methods can be applied to several applications, i.e., handwriting character recognition [2, 20, 33], scene classification [26], shape matching and retrieval [24, 26] and sign language [27, 30], etc. The summary of syntactic pattern recognition these applications are concluded in Table 1.1.

However, in this thesis focus on the string grammar clustering algorithm [2, 19]. The string grammar clustering algorithm so far is the hard clustering in previous works. Partitioning of the data of hard clustering such that each data point belongs to exactly one of the partitions, so the accuracy and efficiency of hard clustering have been very poor for some applications.

Table 1.1 Summary of syntactic pattern recognition applications

	Year	**Applications**	**Technique**	**Distance measure**
1	1982	Chinese handwriting [2]	Chaincode + sgHCM	Levenshtein distance
2	1999	Handwriting analysis [20]	Contour analysis	Levenshtein distance
3	2000	Chromosome[22]	Nearest Neighbors	Normalized edit Distance
4	2006	Shape dataset [24]	Fourier descriptor	Shape edit distance (SED).
5	2006	Handwriting analysis and chromosome [25]	Chaincode + Median string	Levenshtein distance
6	2008	Scene analysis and shape dataset [26]	Gobally ordered and locally unordered representation	Levenshtein distance
7	2010	Hand Language Video [27]	Converting Time Serial Features to Strings	Weight edit distance
8	2010	Genome Data [28]	FCM	Levenshtein distance
9	2012	Text Mining [29]	FCM	Levenshtein distance
10	2012	Sign Language [30]	U-SURF + sgHCM	Levenshtein distance
11	2013	Shape dataset [31]	Skeleton graph +agglomerative shape clustering	Common structure skeleton graphs distance

Table 1.1 Summary of syntactic pattern recognition applications. (continue)

	Year	Applications	Technique	Distance measure
12	2014	Arrow Road Marking [32]	L- junction string	Levenshtein distance
13	2015	Handwriting character recognition [33]	Curves, and graph structure	Approximate subgraph matching and Levenshtein distance
14	2016	Vehicle Plate Matching [34]	Precision-Recall curve	Levenshtein distance
15	2017	Vehicle Plate Matching [35]	Precision-Recall curve	Levenshtein distance
16	2017	Automatically retrieve harmony from song [36]	Time series	Levenshtein distance

In sgHCM clustering, the prototype of each cluster is a median string of the strings in that cluster. Definition of the (generalized) median string is the string that minimizes the sum of the distances to each string of the set. No efficient algorithm can be designed to compute the median string. Hence, [23] proposed improved algorithms in finding the median of cluster. It has been proved that modified median string will give a better classification rate than the regular median string. In [28, 29] They use the algorithms that combining the FCM with the Levenshtein distance. However, in both algorithms, FCM uses the Eucliden distance not the Levenshtein distance as a dissimilarity measure. Hence, these two algorithms are not really a synthetic clustering.

In our work, we propose incorporating fuzzy clustering theory into string grammar clustering method. A pattern will be represented by a string in a language, which is encoded from pattern primitives [2] which are presumably easy to recognize. Moreover, string can be produced from contour of shape [20], SIFT or U-SURF [30]. An input string pattern is matched against strings representing each prototype based on a selected similarity criterion, the input pattern is classified in the same class as the prototype pattern, which is the best to match the input. Although in [22] proposed the normalized edit distance that leads to good performance in several applications, we will not apply this string distance in our work because it does not satisfy the triangle inequality.

Hence, Levenshtein distance [17, 19] is utilized the as a dissimilarity measure and the fuzzy median is utilized to calculate a cluster prototype which follows the median string concept [22, 23]. As we all know, fuzzy clustering has benefited more than crisp clustering. Consequently, the use of fuzzy theory combines with string grammar clustering can give a higher accuracy rate of classification.

1.3 Research objective

1.3.1 To develop a novel string grammar fuzzy clustering method.

1.3.2 To apply a novel string grammar fuzzy clustering algorithm on synthetic and real world string datasets.

1.4 Research Scope and Method

1.4.1 Implement the algorithm on synthetic and standard public datasets.

1.4.2 Compare the result with string grammar hard clustering method.

1.5 Education/Application Advantages

1.5.1 To obtain a new fuzzy clustering technique for string grammar.

1.5.2 To obtain a new string grammar fuzzy clustering technique for apply on synthetic and real world string datasets.

1.6 Research Methodologies

1.6.1 Study the theories and review the literatures.

1.6.2 Collect the data sets.

1.6.3 Design and implement the string grammar fuzzy clustering algorithms.

1.6.4 Test and improve the performance of algorithm.

1.6.5 Collect the experimental results and find the optimal parameters of the algorithm.

1.6.6 Discuss, conclude and write a thesis.

1.7 Organization of Dissertation

The thesis is divided into six chapters. Chapter 1 begins with the introduction. Chapter 2 reviews the fuzzy sets, the clustering method of interest, i.e. hard C-means, fuzzy C-means, fuzzy C-medians, possibilistic C-means, fuzzy possibilistic C-means, possibilistic fuzzy C-means, unsupervised possibilistic C-means and unsupervised possibilistic fuzzy C-means, string grammar, string grammar clustering, fuzzy clustering validation techniques and measuring of overlapping data. Chapter 3 describes the research designs and the proposed method. Chapter 4 describes the experimental results of the proposed method on real standard data sets. Chapter 5 describes our applications. Finally, conclusions are drawn in Chapter 6.