CHAPTER 6

Conclusion

This thesis presents the algorithm for clustering string data. Syntactic pattern recognition is one of pattern recognition form, in which each object can be represented by sets of simple pattern primitives that can be described by string in a language. This allows for representing pattern structures, taking into account more complex pattern such as in pictorial data analysis. The main advantages of the syntactic approach are that it provides a capability for describing the large set of complex patterns by using small set of simple pattern primitives and these string grammars can be transformed back to original image as well. In string clustering, the objective of cluster analysis is to assign observations to cluster so that string observations within each group are similar to one another. The Levenshtein distance is utilized for distance measure between strings.

We propose new fuzzy clustering techniques for string. This technique is designed for working with syntactic pattern, and eliminating the drawback of string grammar hard clustering method. There are five string grammar fuzzy clustering methods. The first string grammar fuzzy clustering method is based on the fuzzy C-medians algorithm, that is string grammar fuzzy C-medians. The second string grammar fuzzy clustering method is based on the fuzzy posibilistic C-means algorithm, that is string grammar fuzzy posibilistic C-medians. The third string grammar fuzzy clustering is string grammar possibilistic fuzzy C-medians. The fourth string grammar fuzzy clustering is string grammar unsupervised possibilistic C-medians. The last is string grammar unsupervised possibilistic fuzzy C-medians. We also proposed a fuzzy median string which incorporates the concept of belongingness of each string in clusters, membership values are introduced into the calculation of a fuzzy median string for all string grammar fuzzy clustering methods.

We can conclude advantages and disadvantages of each algorithm below:

1) The sgFCMed has benefited more than sgHCM clustering. However, sgFCMed is not good enough for some applications with large overlapping data and noisy data with outlier.

2) The sgFPCMed can solves the noise sensitivity deficiency of sgFCMed. It is a combination of possibilistic theory and sgFCMed clustering. However, the row sum constraints must be equal to one that may be problematic for big data set.

3) The sgPFCMed can reduce the effect of outliers and prefer better recognition rates than sgFPCMed by eliminating the row sum constraints of sgFPCMed. Hence, the sgPFCMed can solve the noise sensitivity deficiency problem better than sgFCMed and sgFPCMed. However, the accuracy depends on several parameters such as m, η, a, b and γ.

4) The sgUPFCMed use the exponential membership functions to describe the degree of belonging. The membership value tends to be zero where the string is too far from its prototype. Hence, the algorithm can detect the outlier of the dataset as well. However, sgUPCMed might produce some prototypes with the same location because the algorithm can generate the same memberships values from relaxing membership constraint $\sum u_{ik} = 1$.

5) In the last proposed algorithm, the sgUPFCMed give the best classification accuracy rate for all of experiments. It can solve the problems of previous algorithms since the sgUPFCMed take advantages from the PFCMed and the sgUPCMed algorithms. Hence, the sgUPFCMed can detect the outlier of dataset as well and can solve coincident cluster of sgUPCMed. However, the accuracy depends on several parameters such as m, η, a, and b.

We used six real-world data sets, i.e., MPEG-7 data set, Copenhagen chromosomes data set, MNIST database of handwritten digits, USPS database of handwritten digits, and Thai sign language translation and identification of cardio-pulmonary resuscitation activity in medical simulation videos to evaluate the performance of our clustering

methods. For the first four real standard data sets, we implement the ten-fold cross validation of the experiments in training process. In testing process, a test string was assigned to the class of the nearest prototype. We then evaluate the quality of partitions using three cluster validity indices, i.e., PE, PC, and XB. We can see that all of algorithms are better than string grammar hard C-means and several numeric methods. Moreover, when we compared our string grammar clustering algorithms, the sgUPFCMed gave the best in term of classification accuracy rate in all experiments.

For dynamic Thai sign language translation system, we improve the system with video caption without prior hand region detection and segmentation using SIFT and our five string grammar fuzzy clustering methods. The SIFT was used to match test frame with symbols in the signature library. String grammar fuzzy clustering and fuzzy K-nearest neighbor (FKNN) are used to find a matched sign words. We compare the performance of our algorithms of Thai sign language translation system with HMM on the same dataset. In the TSL dataset. Our algorithms yielded a pretty good result for all of experiments.

For identification of cardio-pulmonary resuscitation activity in medical simulation videos, video frame that involves CPR can be identified by using patio temporal three-dimensional gradients and self-organizing map for string representation and by using string grammar fuzzy clustering model for finding prototypes of CPR and non-CPR. We then used fuzzy K-nearest neighbor to classify the test dataset. We compared our results with those from the SVM and HMM, the best of our results is better than other two methods. From the result on six real-world data sets, the sgUPFCMed give the best classification accuracy rate. We recommend to use the sgUPFCMed for clustering or classify syntactic dataset.

The advantages of our methods can be described as follows:

1) Our methods are more appropriate for applications where the structure of a pattern can be easily translated into string.

2) The length of each string can be varied. This has an advantage over the numerical vector that the length is always fixed.

3) The use of string grammars representative can be transformed back to original image as well.

4) Our algorithms are more resistant to outliers and can handle some strings belong partially to various clusters (overlapping data), especially when we use our sgPFCMed, sgUPCMed and sgUPFCMed.

The disadvantages of our methods can be described as follows:

1) Normally, clusters in real world dataset are different shapes. it is difficult to visualize shape of string cluster. In this thesis use only median for prototype calculation. It is the one limit of our algorithms and the median cluster center only suits for symmetric distribution data set, it may not be suitable for largely skewed distribution.

2) In this thesis, we use the Levenshtein distance as the string distance measure, we found some problems of the Levenshtein distance. The Levenshtein distance takes all transform operations in the same way without taking into account the character that is used in the transform operation. This drawback might affect to the result of classification.

Hence, these detected problems should be solved and prevented in the future such that we should consider other methods that be suitable for finding the cluster center and use other distance that may be more appropriate for calculating the distance than Levenshtein distance.