

Fuzzy Clustering in Biochemical Analysis of Cancer Cells

Xiao-Ying Wang **Jonathan M. Garibaldi**
 Department of Computer Science and IT
 The University of Nottingham, United Kingdom
 {xyw, jmg}@cs.nott.ac.uk

Benjamin Bird **Mike W. George**
 Department of Chemistry
 The University of Nottingham, United Kingdom
 {pcxbb1,mike.george}@nottingham.ac.uk

Abstract

Currently there is considerable effort investigating whether infrared spectroscopy can be used as a diagnostic probe to identify early stages of cancer since these techniques are sensitive to biological changes within cells. Cluster analysis is often used to try and unravel complex vibrational images. In this paper, a Fuzzy C-Means (FCM) based model selection algorithm was used to automatically cluster sets of infrared spectral data taken from lymph node tissue sections. Initial results were often prone to the creation of excessive clusters in comparison with clinical diagnosis which is partly due to the complexities of biological tissue. A new method to merge clusters was developed with the ability to successfully find and combine the two most similar clusters in order to address this problem. This new method was applied to four sets of infrared spectral data, from two types of human cancers (lymph node and oral). The experimental results clearly show that this new method can successfully combine the most similar clusters together and potentially improve the accuracy of diagnosis.

Keywords: FCM, lymph node.

1 Introduction

Breast cancer is the most common malignancy found among women, with high death rates reported in the United Kingdom (13,000 p.a.) and the United States of America (40,000 p.a.) [1]. The ability to accurately identify the malignancy is crucial for prognosis and preparation of effective treatment. Disease staging in breast cancer includes the assessment of lymph nodes in the ipsilateral axilla. The presence of metastasis (cancer spread from its

original location) is an indicator for local disease recurrence and thus a method for identifying patients who are at high risk of developing disease that could spread throughout the body. Intra-operative diagnosis has therefore become increasingly important with the recent introduction of sentinel lymph node biopsy [2]. Present techniques employed to facilitate faster intra-operative diagnosis, such as imprint cytology and frozen section assessment, report wide variation in their sensitivity and specificity. This general lack of a consistently available and reliable intra-operative tool for sentinel lymph node diagnosis has given rise to a variety of different spectroscopic methods being investigated in order to determine whether such approaches could be used to generate a reliable aid for diagnosis [3-5].

Infrared microscopy has proven application in the optical assessment of complex biological systems in tissues since it is sensitive to chemical changes within cells [6]. In this paper we investigate using infrared imaging, malignancy that has spread from the breast to the lymph nodes. The distinct ability to identify small biochemical changes within tissue, coupled with recent advances in infrared imaging technologies, may allow high accuracy detection of malignant change within the timescale required for intra-operative diagnosis.

The complexity of biological systems have meant that multivariate clustering techniques have often been used to analyse complex spectra from tissue sections [7]. The analysis groups a set of unlabelled data (infrared spectra in this study) into separate clusters based on their characteristics, such that the data in the same cluster are as similar as possible and the data within different clusters are as dissimilar as possible. Therefore, different types of cells found within biological tissue can be separated. In previous work, Hierarchical Cluster Analysis (HCA) and FCM clustering were applied to seven infrared spectral data sets taken from oral cancers patients [8]. The experimental results showed that

the FCM clustering approach performed significantly better than HCA when compared to clinical analysis. Furthermore, the time complexity for HCA is $O(n^2)$, where n is the number of data points, therefore, for a larger dataset, HCA may need high computational requirement.

We previously reported [9,10] a Simulated Annealing Fuzzy Clustering algorithm (SAFC) which could be used to automatically obtain the optimal number of clusters from the given data sets by traversing the search space using three different neighbourhood operations: i) perturb centre, ii) delete centre and iii) split centre. The cluster solution that achieved the minimum cluster validity index value was returned. It was shown that the algorithm obtained the same clusters as clinical analysis in four out of the seven data sets. In the remaining data sets, the number of clusters differed from clinical analysis but achieved smaller Xie-Beni validity measures [11]. However, the objective of our research is to facilitate clinical analysis automation, which would require robust and consistent identification of the clinical clusters. Although the simulated annealing clustering process performed well on small datasets, it can become time consuming when applied to larger datasets. In order to address this we have changed to a FCM based model selection algorithm to find the optimal number of clusters.

As previously mentioned, both the SAFC and FCM based model selection algorithms occasionally identified an excessive number of clusters compared to clinical diagnosis. This was due in part to the cluster validity index, where all distances between data points and cluster centres are calculated using the Euclidean Distance. This resulted in inefficient clustering particularly when the shape of the clusters differed greatly from the ideal (spherical). The complexity and range of different cell types which may be present e.g. pre-cancerous and mature cancer can result in the algorithm identifying excessive clusters corresponding to the different severity of cancer cells present. Nevertheless, at this stage of study, we prefer to group all cells with the same clinical diagnosis into one cluster. In order to solve this problem, it is required to combine the clusters with the most similarity together, such as the pre-cancerous and mature cancer cell types. Although they are different stages of cancer, they may have properties that make them similar to one another. This information may be contained in the existing infrared spectra and data analysis.

In this paper, a new method is proposed that will enable similar separated clusters, produced via the FCM based model selection algorithm, to be merged together into one cluster. In order to achieve this, we have used the Xie-Beni index as the cluster validity index (V_{XB}) within our experiments.

2 FCM Based Model Selection Algorithm

It has been shown that fuzzy clustering such as FCM has clear advantages over crisp and probabilistic clustering methods [12]. This algorithm is based on the standard FCM method whereby the minimal and maximal number of clusters is referred to as c_{min} and c_{max} , and the best data structure (C) based on the cluster validity (V_{XB}) is returned. The algorithm can be performed in the following steps:

- 1) Set c_{min} and c_{max} (in the experiments, we set $c_{min}=2$ and $c_{max}=10$)
- 2) For $c = c_{min}$ to c_{max}
 - 2.1) Initialize the cluster centres
 - 2.2) Apply the standard FCM algorithm and obtain the new center C and new fuzzy partition matrix U .
 - 2.3) Calculate the cluster validity (V_{XB})
- 3) Return the best data structure (C) that corresponds to the minimal V_{XB}

During the clustering process, a data point is set to a specific cluster for which the degree of the fuzzy partition matrix U is maximal.

3 A New Merge Clustering Method

A set of infrared spectra from one tissue section were clustered and the best data structure obtained with the corresponding minimal value of cluster validity (V_{XB}). As mentioned in section 1, the FCM algorithm can occasionally identify an excessive number of clusters in comparison to clinical analysis. Based on this problem, a method that can find and merge two similar clusters has been developed. In the rest of this section, the following questions will be answered: "Given a set of c clusters, which two clusters are the most similar?" and "How should they be merged?" (specifically, how to calculate the new cluster centre).

There are 821 absorbance values corresponding to each wave-number for each datum in the given infrared spectral datasets. In order to reduce the number of variables in the clustering analysis, we initially identified the first 10 principal components

(PCs) that were extracted using Principal Component Analysis (PCA) [13]. Within these first 10 PCs, the majority of the variance (approximately 99%) is represented from the original data. The first 2 PCs incorporate around 80% of the variance in the original data. Thus plotting the original data in these two PC dimensions, the approximate data distribution can be visualised.

3.1 Choosing two similar clusters

Previously, there have been many algorithms for merging clusters [e.g. 14,15]. The different approaches can generally be divided into two groups. The first group are those that select the clusters which are 'closest' to each other [14], and the second those that choose the 'worst' two clusters (judging by some cluster validity function) [15]. However, neither of these is suitable for solving the problem described in this research, illustrated in Figure 1.

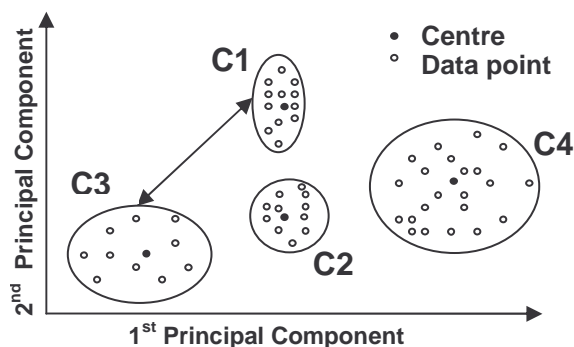


Figure 1 A plot of first and second principal components obtained from an infrared data set.

A set of infrared spectral data were plotted in first and second PCs after applying the FCM based model selection algorithm. Four clusters (C1, C2, C3, and C4) were formed as shown in Figure 1. In clinical analysis, C1 and C3 are all cancer cells although they were taken from different areas of the tissue section that may contain different stages of cancer. C2 and C4 are normal nodal and reticulum cells respectively. Obviously, the two similar clusters that need to be merged together are C1 and C3. Referring to the two types of techniques that merge clusters (described above), the closest two clusters in the data set are C1 and C2 (see the distances between each cluster centre). Normally, a good cluster is defined by the data points within the cluster being tightly condensed around the centre (compactness). In the sample data set, clusters C1 and C2 are more compact than the other two, so the

worst two clusters are C3 and C4. Neither technique chooses the desired two clusters, C1 and C3.

In order to tackle this problem, we examined the original infrared spectra rather than searching for a relationship using the data structure in the PCA plot. Plotting the mean spectra from the separate clusters allowed the major differences between them to be more clearly visualized. The similarity between clusters is more obvious at the wavelength where the biggest difference between any two mean spectra is located. Based on this observation, the proposed method to find two similar clusters is:

- 1) Obtain the clustering results from the FCM based model selection algorithm.
- 2) Calculate the mean spectra \bar{A}_i for each cluster,

$$\bar{A}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} A_{ij} \quad (i=1\dots c) \quad (1)$$

where N_i is the number of data points in the cluster i ; A_{ij} is the absorbance of the spectrum for each data point j in cluster i ; c is the number of clusters. The size of \bar{A}_i is p , the number of wave-numbers in each spectrum (each mean spectrum is a vector of p elements).

- 3) Compute the vector of pair-wise squared differences D_{ij} between all mean spectra,

$$D_{ij} = (\bar{A}_i - \bar{A}_j)^2 \quad (i=1\dots c, j=1\dots c) \quad (2)$$
- 4) Find the largest single element, d_{max} , within the set of vectors D .
- 5) Determine the wave-number corresponding to the maximal element d_{max} .
- 6) At the wave-number identified from 5), find the two mean spectra with minimal difference. The clusters corresponding to these spectra are merged.

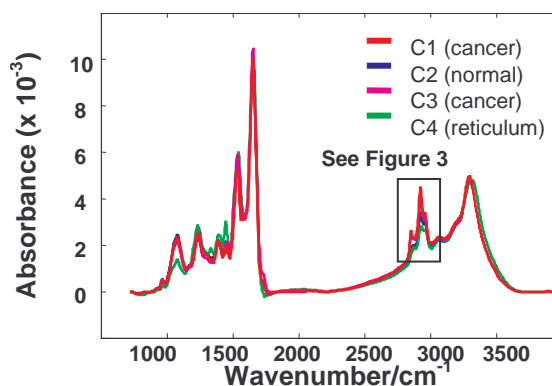


Figure 2: Example of mean infrared spectra obtained from different clusters.

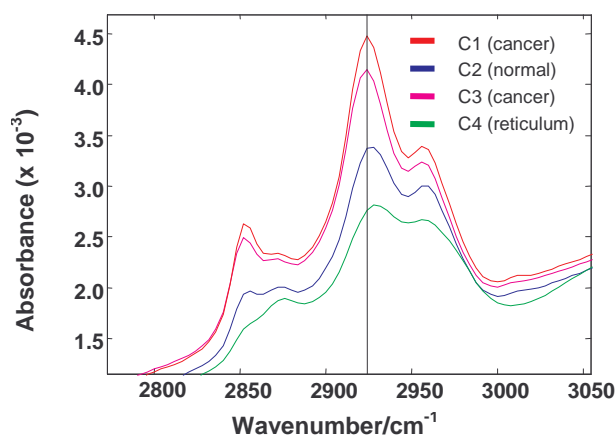


Figure 3: Enlarged region of Figure 2

The mean spectra for the four clusters are displayed in Figure 2. We calculated the set of differences, D , between each pair of mean spectra using equation (2). The largest difference d_{max} exists between C1 and C4 as shown in Figure 3. The wave-number that corresponds to d_{max} is 2924 cm^{-1} . Examining the four absorbance values at this wave-number we can clearly see that the two closest spectra are those belonging to clusters C1 and C3 (in red and pink). So these are the two clusters selected to be merged.

3.2 Merging Clusters

After finding the two most similar clusters, the next stage is to merge them together. Firstly, all of the data points within each of the clusters are assigned to their centre with a membership (μ) of 1. Note the chosen two clusters (C1 and C3) outlined above are now considered as one cluster. Data points within a particular cluster are assigned a membership value of 0 to other cluster centres. The following equation (3) was used to calculate the centres [16].

$$v_j = \frac{\sum_{i=1}^n \mu_{ij} x_i}{\sum_{i=1}^n \mu_{ij}}, \forall j = 1, \dots, c \quad (3)$$

where c is the number of clusters, and n is the number of data points. x is the data point.

4 Experiments and Results

4.1 Tissue Samples

Tissue specimens were collected during routine surgical resection for breast cancer with approval from Gloucestershire Research Ethics Committee and fully informed consenting patients. From each appropriate case, a small portion of one axillary

lymph node was collected for analysis. Samples were immediately snap frozen in liquid nitrogen to maintain their biochemical condition, and subsequently cut using a freezing-microtome, to obtain a $7\mu\text{m}$ thick tissue section. These sections were then placed onto barium fluoride discs and stored in a cryovial ready for infrared analysis. The remaining specimens were formalin-fixed and wax embedded to allow conventional histology. An additional tissue section was cut immediately adjacent to the frozen section, and further hematoxylin and eosin stained for comparative analysis by a Consultant Breast Histopathologist.

4.2 Mid-Infrared Imaging

Prior to IR analysis, specimens were removed from the cryovial and passively warmed to room temperature. Using the adjacent hematoxylin and eosin stained tissue sections, areas of interest were first located and then infrared images were obtained using a commercially available Perkin Elmer Spectrum Spotlight 300 FTIR Imaging System[®]. The system uses a linear array of small area detectors coupled with an electronic stage, to raster across the sample in both the X and Y directions, thus creating a microscopic IR image. Each pixel within the image is representative of one IR spectroscopic measurement. All spectra were collected using a pixel size of $6.25\mu\text{m} \times 6.25\mu\text{m}$ and a spectral resolution of 8cm^{-1} . In these experiments 16 spectra were coadded over the range $720\text{--}4000 \text{ cm}^{-1}$. Dependent upon sample size, analysis times varied from thirty minutes to several hours.

4.3 Data Pre-processing

All spectra were uniformly pre-treated before undergoing multivariate analysis. Possible contributions from atmospheric water vapour and CO_2 , which absorb in some regions of the IR spectrum, were removed by an atmospheric correction algorithm integrated into the Perkin Elmer Spotlight Software[®]. Sloping and curved baselines sometimes encountered in spectra, caused by irregularities in cell density across tissue sections, were corrected by applying a 6 base point linear interpolation to all spectra. Baseline points used in this process were located at $4000, 3744, 2200, 1836, 876$ and 720 cm^{-1} respectively. Finally spectra were scaled before cluster analysis such that the sum over the indicated wavelengths ($720\text{--}4000\text{cm}^{-1}$) equals unity (also known as peak area normalisation).

4.4 Clustering Results

In these experiments, we initially analysed two lymph node tissue sections named LNII5 and LNII7. The FCM based model selection algorithm was firstly used to generate initial clusters with good data structure. These clusters were then further combined using our new merge clustering method. Due to different initialization states may lead to different clustering results. We run the FCM based model selection algorithm 10 times on both datasets. The results are fairly stable, as shown in Figures 4-7.

Figures. 4 and 6 show the clustering results obtained from the FCM based model selection algorithm and Figures. 5 and 7 show the merged cluster results for LNII5 and LNII7 respectively. The results clearly show that the separate cancer clusters have now correctly merged using the new approach. In order to verify the merging clusters algorithm, the method was further applied to previous work [9,10], whereby the SAFC clustering algorithm obtained 3 clusters, rather than 2 found by histological analysis.

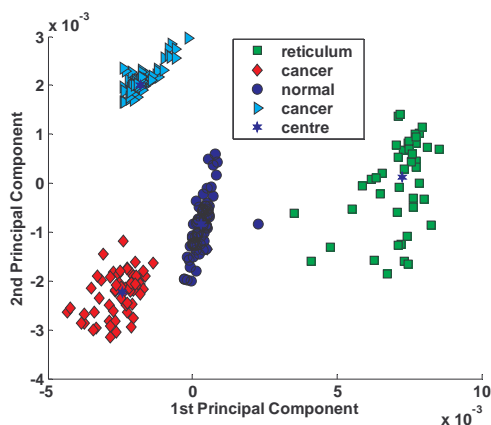


Figure. 4: LNII5 clustering results obtained from the FCM based model selection algorithm

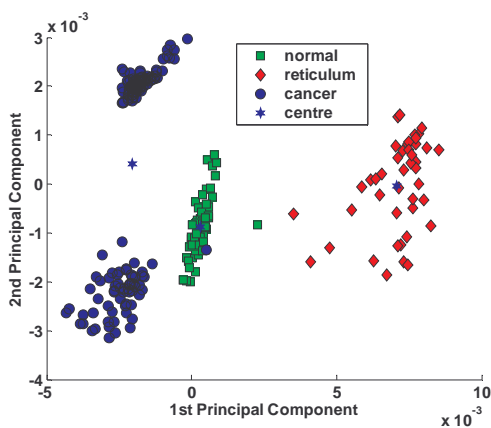


Figure. 5: LNII5 merged clusters results.

Due to space limitations, we only report results found after applying the new merging method, as shown in Figures 8 and 9. The circled regions display the initial clustering, whereas the coloured legend shows the merged results. It is apparent that the proposed method combines the correct clusters.

5 Discussion and Conclusion

Due to the distance measurement in FCM is Euclidean Distance, when we subjectively set the number of cluster to be the same as obtained from clinic analysis, the clustering results do not match clinical diagnosis. However, when we use the Xie-Beni index to select the best number of clusters from FCM and apply our cluster merging method to obtain the same number of clusters as in clinical diagnosis, then the clustering results are very close to clinical diagnosis.

We applied the proposed method to four separate infrared spectral datasets (for which previous approaches could not obtain the correct number of clusters). For each dataset, the proposed method

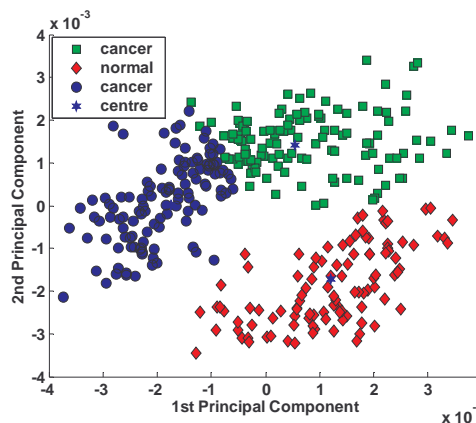


Figure. 6: LNII7 clustering results obtain from the FCM based model selection algorithm

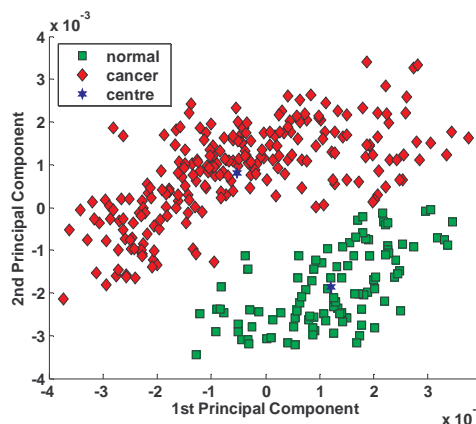


Figure. 7: LNII7 merged clusters results.

identified the correct amount of clusters. It should be noted that after merging clusters, there were some misclassified data points (approximately 3-4 for LNII7 and 1-2 for the remaining datasets). However, the clustering accuracy was significantly improved. In future work, we aim to further investigate the merging of clusters, such as the correct conditions needed to stop further combination of clusters. We are also attempting to collect and analyse a wider source of infrared spectra from varying types of cancerous tissues to further refine our algorithms.

Acknowledgments

The authors would like to thank Dr. N. Stone and Dr. J. Smith from the Gloucestershire Royal Hospital for their aid in sample collection and diagnosis. We are particularly grateful to Professor M. A. Chester and Mr. J. M. Chalmers for many helpful discussions.

References

- [1] S.H. Landis, T. Murray, S. Bolden, and P.A. Wingo, "Cancer Statistics", *CA: A Cancer Journal for Clinicians*, vol. 49(1), pp. 8-31, 1999
- [2] R. R. Turner, D. W. Ollila, D. L. Krasne and A. E. Giuliano, "Histopathologic Validation of the Sentinel Lymph Node Hypothesis for Breast Carcinoma", *Ann. Surg.*, vol.226(3), pp. 271-276, 1997
- [3] K. S. Johnson et al, "Elastic Scattering Spectroscopy for Intraoperative Determination of Sentinel Lymph Node Status in Breast", *J. Biomed. Opt.*, vol 9(6), pp. 1122-1128, 2004
- [4] A. Godavarty et al, "Diagnostic Image of Breast using Fluorescence Enhanced Optical Tomography: Phantom Studies", *J. Biomed. Opt.*, vol 9(3), pp. 486-496, 2004
- [5] J. Smith, C. Kendall, A. Sammon, J. Christie-Brown and N. Stone, " *Technol. Cancer. Res. T.*, vol 2 (4), pp. 327-331, Aug 2003
- [6] M. J. Tobin et al, "Investigating the potential for infrared microanalysis in cancer screening", *Europ. Clinical Laborator*, vol. 21 (4), pp.20-22, 2002
- [7] P. Lasch, W. Haensch, D. Naumann and M. Dien, "Imaging of Colorectal Adenocarcinoma Using FT-IR microspectroscopy and Cluster Analysis", *BBA-MOL BASIS DIS*, vol.1688 (2), pp.176-186, 2004
- [8] X.Y. Wang, J. Garibaldi, and T. Ozen, "Application of The Fuzzy C-Means Clustering Method on the Analysis of non Pre-processed FTIR Data for Cancer Diagnosis", Proc. of the ANZIIS Conference, pp. 233-238, Australia, 2003
- [9] X.Y. Wang and J. Garibaldi, "Simulated Annealing Fuzzy Clustering in Cancer Diagnosis", to appear *Euro. J. of Informatica*, 2005

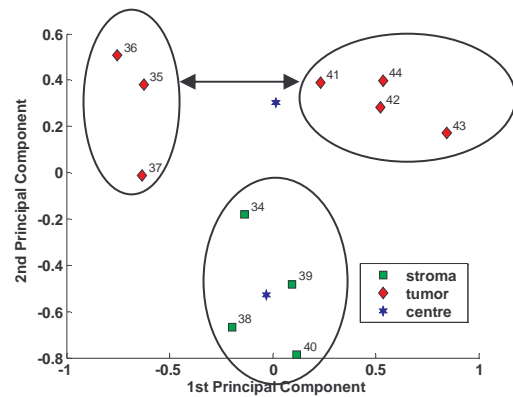


Figure. 8: Data set 3 merged clusters results.

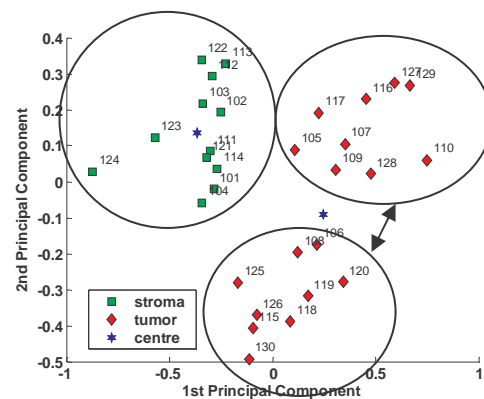


Figure. 9: Data set 5 merged clusters results.

- [10] X.Y. Wang, G. Whitwell, and J. Garibaldi, "The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis", Proc. of the IEEE 4th International Conference on Intelligent System Design and Application, pp. 467-472, Hungary, 2004
- [11] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering", *IEEE T Pattern Anal*, Vol. 13, pp. 841-847, 1991.
- [12] H. Sun, S. Wang, and Q. Jiang, "FCM-Based Model Selection Algorithms for Determining the Number of Clusters", *Pattern Recogn*, vol. 37, pp. 2027-2037, 2004
- [13] I.T. Jolliffe, "Principal Component Analysis", Springer-Verlag, New York, 1986
- [14] X. Gong and M.B. Richman, "On the application of cluster analysis to growing season precipitation in North America east of the Rockies", *Journal Climata*, vol. 8, pp. 897-931, 1995
- [15] Y. Xie, V.V. Raghavan, and X. Zhao, "3M Algorithm: Finding an Optimal Fuzzy Cluster Scheme for Proximity Data", Proc. of the IEEE International Conference, vol 1, pp. 627-632, 2002
- [16] J. Bezdek, "Pattern Recognition With Fuzzy Objective Function Algorithms", Plenum, New York, 1981