Performance Assessment of Some Clustering Algorithms Based on A Fuzzy Granulation-degranulation Criterion

Sriparna Saha and Sanghamitra Bandyopadhyay
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India-700108
Email: {sriparna_r, sanghami}@isical.ac.in

Abstract

In this paper a fuzzy quantization dequantization criterion is used to propose an evaluation technique to determine the appropriate clustering algorithm suitable for a particular data set. In general, the goodness of a partitioning is measured by computing the variances within it, which is a measure of compactness of the obtained partitioning. Here a new kind of error function, which reflects how well the formed cluster centers represent the whole data set, is used as the goodness of the obtained partitioning. Thus a clustering algorithm, providing a good set of centers which approximate the whole data set perfectly, is best suitable for partitioning that particular data set. Five well-known clustering algorithms, GAK-means (genetic algorithm based K-means algorithm), a newly developed genetic point symmetry based clustering technique (GAPS-clustering), Average Linkage clustering algorithm, Expectation Maximization (EM) clustering algorithm and Self Organizing Map (SOM) are used as the underlying partitioning techniques. Five artificially generated and three real-life data sets are used to establish that the proposed methodology is able to correctly identify appropriate clustering algorithm for a particular data set.

Keywords: Unsupervised classification, clustering algorithms, algorithm identification, fuzzy vector quantization.

1 Introduction

Clustering [6] is a core problem in data-mining with innumerable applications spanning many fields. The three fundamental questions that need to be addressed in any typical clustering scenario are: (i) what is a good clustering technique suitable for a given data set, (ii) how many clusters are actually present in the data, and (iii) how real or good is the clustering itself. It is well-known in the pattern recognition community that different algorithms are applicable for data with different characteristics. For example, while K-means [6] is widely used for hyperspherical, convex, equisized clusters, it is known to fail where the clusters are not hyperspherical and also significantly unequal in size. Similarly average (or, single) linkage clustering algorithms [6] work well for non-overlapping clusters of any shape, but fail if the clusters overlap. The expectation maximization algorithm (EM) [5] is considered to be an appropriate optimization algorithm for constructing statistical models of data when the type of data distribution (e.g., Gaussian) is known. It provides a probabilistic clustering where each data element has a certain probability of being a member of any cluster. Unlike the K-means it does not depend on any distance measure, and accommodates categorical and continuous data in a superior manner. Recently, some clustering methods have been proposed that exploit the symmetry property within the clusters [3]. These methods are found to be superior to several other techniques when the clusters do offer a symmetric structure. Thus given a wide choice of methods, determining an appropriate clustering algorithm presents a challenge.

In this paper a fuzzy granulation and degranulation criterion is used to determine the appropriate clustering algorithm suitable for a particular data set. In the fuzzy vector quantization technique [10], the vectors in the code book are used to encode the original data in terms of the membership values. During decoding, a given vector is expressed as a function of the membership values and the cluster centers. In this paper the final cluster centers formed by the respective clustering algorithm are regarded as the representatives of the whole data set. Next, the final membership values of all data points present in the data set with respect to all the cluster centers are calculated. Now based on this membership vector and the final cluster centers, each data point is approximated. The Euclidean distance between the approximated point and the original point is the quantization error for that particular point. The total quantization error (*quan_error*) of the entire data set represents how well the clustering algorithm is. The clustering algorithm with minimal total quantization error is the best suitable for segmenting that particular data set. It is easy to understand that if the cluster centers formed by a particular clustering algorithm represent the whole data set properly, then the total quantization error will be small. Thus in the absence of original data points, they can be approximated by some combination of cluster centers and the corresponding membership values.

The performance of five well-known clustering algorithms are assessed on five artificially generated and three real-life data sets of varying complexities. Based on the proposed performance index the appropriate clustering algorithm for a particular data set is determined. The five clustering algorithms are the well-known genetic algorithm based K-means clustering technique (GAK-means) which is developed in [9], in order to overcome the limitation of K-means algorithm to get stuck at sub optimal solution, a newly developed genetic point symmetry based clustering technique (GAPS) [3], the well-known Expectation Maximization clustering algorithm [5], the Average Linkage clustering algorithm [7] and Self Organizing Map (SOM) [8]. Let the clustering algorithms be denoted by $\mathcal{A}_i$ where $i = 1, \ldots A$, A is the total number of clustering algorithms those are to be evaluated. Let after application of $\mathcal{A}_i$ on a particular data set (with number of clusters equal to that actually present in the data set), total quantization error formed be denoted by V_i. Then the suitable clustering algorithm for that particular data set is denoted by $\mathcal{A}_i = argmin_{i=1,\ldots A} V_i$. Thus, according to the proposed criterion, $\mathcal{A}_i$ is the most suitable algorithm for that particular data set.

2 The Proposed Method of Determining the Proper Clustering Algorithm for A Particular Data Set

The proposed method of detecting the appropriate clustering algorithm for a particular data set is inspired by the fuzzy vector quantization-dequantization technique proposed in Ref.[10]. In fuzzy vector quantization, the code book is formed by optimizing some error function by using some optimization techniques. Here code book consists of elements of the data which approximate the whole data set appropriately. In fuzzy data clustering the cluster centers and the membership values are the representatives of data points present in the data set. Thus, codebook consists of cluster centers formed by a particular clustering algorithm. The membership values of all points to all cluster centers are computed based on the available cluster centers. Any point can then be approximated by these membership values and the cluster centers present. The cluster centers which approximate the whole data set well are desired to be found out by a clustering algorithm. Let a particular clustering algorithm be $\mathcal{A}_i$, $i = 1, \ldots A$, where A is the total number of clustering algorithms those are to be evaluated. Let, total number of clusters present in a data set that is known *apriori* be K. The data set consists of N number of points represented by $\overline{x}_i$, $i = 1, \ldots N$. The final cluster centers provided by the particular algorithm $\mathcal{A}_i$ for this particular data set are $\{\overline{v}_1, \overline{v}_2, \ldots, \overline{v}_K\}$. Then a way of encoding a particular data point $\overline{x}$ in the data set can be represented by the collection of membership values to different clusters. We require that the corresponding membership degrees $u_i(\overline{x}), i = 1, 2, \ldots K$ are confined to the unit interval and sum up to 1. The membership values are calculated by minimizing the following performance index

$$Q_1(x) = \sum_{i=1}^{K} u_i^m(x) \|\overline{x} - \overline{v}_i\|^2 \quad (1)$$

subject to the following constraints already stated above, that is

$$u_i(\overline{x}) \in [0, 1], \quad \sum_{i=1}^{K} u_i(\overline{x}) = 1 \quad (2)$$

Here the Euclidean distance function which is denoted by $\|\|^2$ is used. The fuzzification coefficient $(m, m > 1)$, shown in the above expression is used to adjust a level of contribution of the prototypes to the result of representation. The collection of K weights $\{u_i(\overline{x})\}$, $i = 1, \ldots K$ along with the cluster centers are used to represent a particular data point $\overline{x}$.

The minimization of Equation 1 is straightforward and follows a standard way of transforming the problem to unconstrained optimization using Lagrange multipliers. Now rewriting Equation 1 by accommodating the constraint in the form of the Lagrangian multiplier (λ), we obtain

$$Q_1(x) = \sum_{i=1}^{K} u_i^m(x) \|\overline{x} - \overline{v}_i\|^2 - \lambda(\sum_{i=1}^{K} u_i(\overline{x}) - 1) \quad (3)$$

The resulting system of equations leading to the minimum of Q comes in the form

$$\frac{dQ}{d\lambda} = 0, \quad \frac{dQ}{du_i(\overline{x})} = 0 \quad (4)$$

After solving the equations with respect to λ and $u_i(\overline{x})$, the resulting weights (membership degrees) become

$$u_i(\overline{x}) = \frac{1}{\sum_{i=1}^{K} (\|\overline{x} - \overline{v}_i\| / \|\overline{x} - \overline{v}_j\|)^{2/(m-1)}} \quad (5)$$

where, $i = 1, 2, \ldots K$. Here, the fuzzification coefficient, m is chosen equal to 2 though the importance of its proper choice is studied in [10].

Thus each data point is represented by the K membership values $u_i(\overline{x}),\ i = 1, \ldots K$ computed by Equation 5 and with the help of K cluster prototypes.

Now, these computed membership values and the cluster prototypes are used to approximate each data point $\overline{x}$. Approximation is based on some aggregation of the cluster prototypes and the associated membership grades $u_i(\overline{x})$. The way of forming $\overline{x'}$ is accomplished through a minimization of the following expression.

$$Q_2(\overline{x'}) = \sum_{i=1}^{K} u_i^m \|\overline{x'} - \overline{v}_i\|^2 \qquad (6)$$

If the Euclidean distance is used to measure the distance between the prototypes and $\overline{x'}$, the problem of unconstrained optimization leads to a straightforward solution expressed as a convex combination of the prototypes

$$\overline{x'} = \frac{\sum_{i=1}^{K} u_i^m \overline{v}_i}{\sum_{i=1}^{K} u_i^m} \qquad (7)$$

where the corresponding prototypes are weighted by the membership degrees. Then the total error due to clustering is calculated as follows.

$$\textit{quan\_error} = \sum_{i=1}^{N} \|\overline{x}_i - \overline{x'}_i\|^2 \qquad (8)$$

where N is the total number of points in the data set. It is shown in [10] that the quality of reconstruction depends on a number of essential parameters of the scheme including the size of the codebook (i.e., here the number of cluster centers) as well as the value of the fuzzification coefficient m.

Thus it is easy to understand that if the final cluster centers formed by a particular clustering algorithm represent the whole data set appropriately, then each data point should be well-approximated by using the cluster prototypes and the corresponding membership values. Then, total distance between the approximated point and the original point would be less, resulting in a smaller value of *quan_error*. Thus the proposed method of determining the appropriate clustering algorithm suitable for a particular data set relies on the calculated *quan_error* produced by all the algorithms for that particular data set. The clustering algorithm which corresponds to the minimum *quan_error* is considered to be best suitable for that particular data set.

3 Experimental Results

This section deals with the experimental results which reveal the superiority of the proposed method in detecting the appropriate clustering algorithm for a particular data set.

3.1 Clustering Algorithms Used for Comparison

Here, performance of five clustering algorithms on five artificial and three real-life data sets are evaluated in terms of the above mentioned criterion. The algorithm which provides the smallest *quan_error* is regarded as the best suitable for that particular data set. Five clustering algorithms, viz., a newly developed point symmetry based genetic clustering technique (GAPS) [3], GAK-means algorithm [9], Average-linkage clustering algorithm [6] (source code was obtained from http://bioinformatics.oxfordjournals.org/cgi/content/abstract), Self Organizing Map (SOM) [8] (source obtained from http://www.cs.tau.ac.il/~rshamir/expander), Expectation Maximization (EM) algorithm [5] (matlab source code was obtained from http://www.mathworks.com/matlabcentral/fileexchange/) are used as the underlying partitioning techniques. The parameters of the genetic clustering algorithms (GAPS and GAK-means) are as follows: population size is equal to 100 and maximum number of generations is equal to 30. For GAPS, the crossover and mutation probabilities are chosen adaptively as in [3]. For GAK-means, the crossover and mutation probabilities are chosen as 0.8 and 0.01, respectively. As already mentioned, the codes for Average Linkage, EM-algorithm and SOM were obtained from different sources and were executed using default parameters. The results reported in the tables are the average values obtained over ten runs of the algorithms.

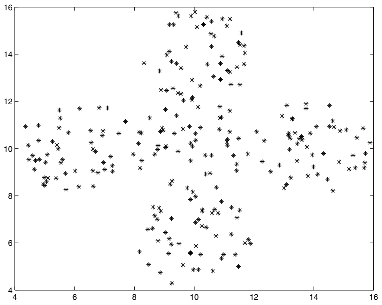


Figure 1. *Data1*

3.2 Data Sets Used

The data sets that are used for the experiments are divided into 2 different groups.

Group 1: Consists of five artificially generated data sets. These data sets are used in [2].

Data1: This data set, shown in Figure 1, consists of 250 data points distributed over 5 spherically shaped clusters in 2-dimensional space. The clusters present here are highly overlapping, consisting of 50 data points each.

Data2: This data set, shown in Figure 2(a), consists of 400

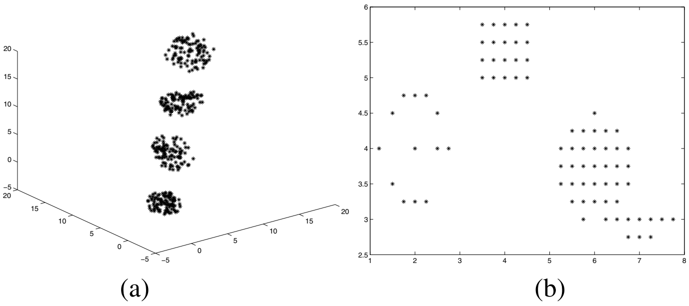


Figure 2. (a) *Data2* **(b)** *Data3*

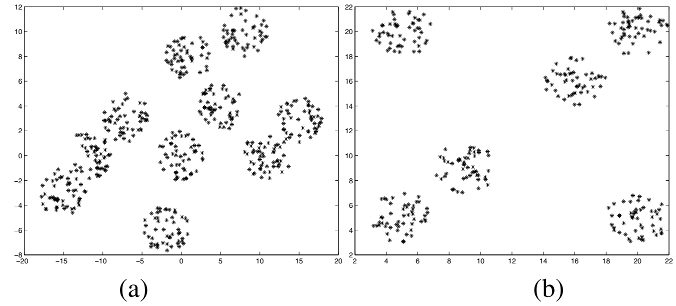


Figure 3. (a) *Data4* **(b)** *Data5*

Table 1. Total *quan_error* **obtained by the five clustering algorithms for eight data sets (entries in bold face indicate the optimal error for respective data)**

Data Set	GAPS	GAK -means	Avg. Link.	EM	SOM
Data1	207.96	**204.95**	212.27	211.29	208.60
Data2	**513.12**	**513.12**	**513.12**	**513.12**	**513.12**
Data3	**15.47**	**15.47**	**15.47**	**15.47**	25.44
Data4	634.08	**462.74**	465.69	656.94	860.93
Data5	**272.13**	**272.13**	**272.13**	**272.13**	**272.13**
Iris	**7.48**	7.89	7.56	7.719	7.719
Cancer	1690.60	**1685.85**	1693.03	1702.45	1694.13
Newthy.	7364.57	6718.45	**2566.79**	8963.22	8471.56

data points in 3-dimensional space distributed over 4 hyperspherical disjoint clusters. Each cluster contains 100 data points.

Data3: This data set, shown in Figure 2(b), consists of 76 data points distributed over 3 clusters.

Data4: This data set, shown in Figure 3(a), is consisting of 500 data points distributed over 10 different clusters. Some clusters are overlapping in nature. Each cluster consists of 50 data points.

Data5: This data set, shown in Figure 3(b), consists of 300 data points distributed over 6 clusters in 2-dimensional space. The clusters are of same sizes.

Group 2: Consists of three real life data sets, obtained from [1]. These are *Iris*, *Cancer* and *Newthyroid* data sets.

Iris: Iris data set consists of 150 data points distributed over 3 clusters. Each cluster consists of 50 points. This data set represents different categories of irises characterized by four feature values.

Breast Cancer: Here we use the Wisconsin Breast Cancer data set consisting of 683 sample points. Each pattern has nine features. There are two categories in the data: malignant and benign. The two classes are known to be linearly separable.

Newthyroid: The original database from where it has been collected is titled as Thyroid gland data ('normal', 'hypo' and 'hyper' functioning). Five laboratory tests are used to predict whether a patient's thyroid belongs to the class euthyroidism, hypothyroidism or hyperthyroidism. There are a total of 215 instances. Total number of attributes is five.

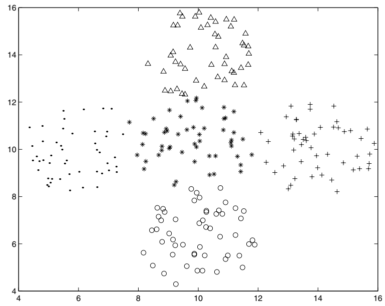
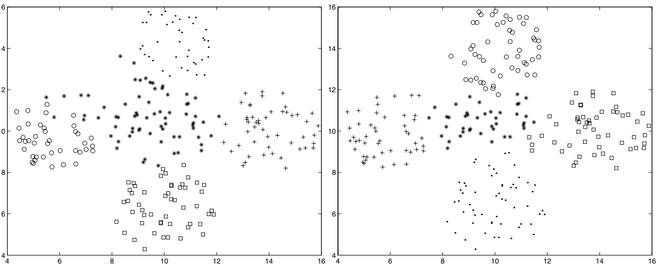


Figure 4. Clustered *Data1* **after application of GAK-means for** $K = 5$

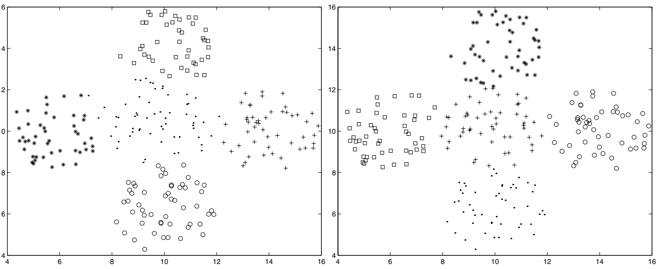
4 Discussion of Results

Here, each of the above mentioned five clustering algorithms are executed on each of the eight data sets with number of clusters kept equal to the original number of clusters present in the data set. The obtained cluster centers (prototypes) are then used to find the membership values of the points to all the clusters. After that, each point is realized in terms of the cluster prototypes and the membership values computed. The error between the approximated point and the original point is calculated. Then the total error of the entire data set gives an idea that how well the clustering algorithm estimates the clusters in terms of its prototypes and the membership values of the points to different clusters. The total error, *quan_error*, obtained by each of the above mentioned five algorithms for the five artificial and three real-life data sets are provided in Table 1.



(a) (b)

Figure 5. Clustered *Data1* after application of (a) GAPS for $K = 5$ (b) Average Linkage for $K = 5$



(a) (b)

Figure 6. Clustered *Data1* after application of (a) EM algorithm for $K = 5$ (b) SOM for $K = 5$

For *Data1*, the obtained *quan_error* after application of GAK-means algorithm is the minimum among all the five algorithms. This implies that GAK-means is the best for this particular data set. This is also verified after visually inspecting the partitioning obtained by all the five clustering algorithms for this particular data set. Figures 4, 5(a), 5(b), 6(a) and 6(b) show, respectively, the partitionings obtained after application of GAK-means, GAPS, Average Linkage, Expectation Maximization and SOM on *Data1*. It is easy to conclude that GAK-means partitions the data in proper five clusters (shown in Figure 4). For *Data2*, all the algorithms provide the same *quan_error*. This implies that all the algorithms perform similarly for this particular data set. As a result, all the algorithms provide the same partitioning for this data set. The corresponding partitioning is shown in Figure 7. For *Data3*, GAPS, GAK-means, Average Linkage and EM algorithms perform equally well, providing the lowest total error, *quan_error*. SOM performs badly for this particular data set providing the largest *quan_error*. The partitionings obtained by GAPS, GAK-means, Average Linkage and EM are the same. It is shown in Figure 8(a). Figure 8(b), providing the partitioning obtained by SOM for *Data3* shows the worst performance of SOM. For *Data4*,

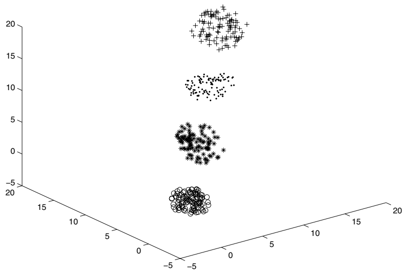
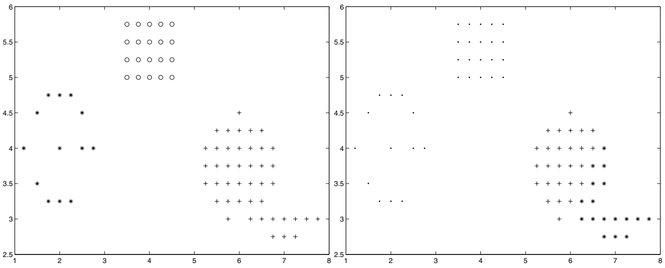


Figure 7. Clustered *Data2* after application of GAK-means/GAPS/Average Linkage/Expectation Maximization/SOM for $K = 4$



(a) (b)

Figure 8. Clustered *Data3* after application of (a) GAK-means/GAPS/Average Linkage/Expectation Maximization for $K = 3$ (b) SOM for $K = 3$

the *quan_error* obtained after application of GAK-means is the minimum. This implies that partitioning provided by GAK-means is the best. This is also verified visually from Figure 9 which contains the partitioning provided by GAK-means after application on *Data4* for $K = 10$. The performance of SOM is the worst among the five algorithms. The corresponding segmentation result is shown in Figure 10. For *Data5*, again all the algorithms perform equally well providing the same quantization error. The partitionings provided by all the algorithms are the same and it is shown in Figure 11.

For the real-life data sets, no visualization is possible as these are higher dimensional data sets. In order to measure the goodness of the partitioning for these three real-life data sets, the *Minkowski Score* (MS) [4] is calculated. This is a measure of the quality of a solution given the true clustering. For MS, the optimum score is 0, with lower scores being "better". For *Iris* data set, as seen from Table 1, GAPS clustering performs the best. The corresponding MS score is 0.62. The other algorithms perform quiet similarly for

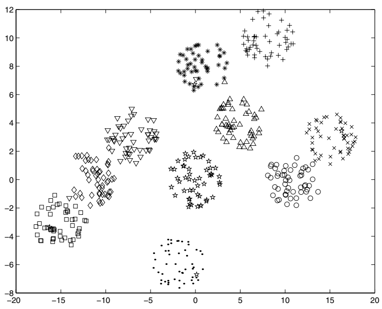


Figure 9. Clustered $Data4$ after application of GAK-means algorithm for $K = 10$

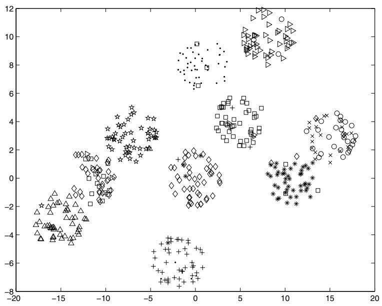


Figure 10. Clustered $Data4$ after application of SOM algorithm for $K = 10$

this data set. For *Cancer* data set, GAK-means performs the best in terms of the total *quan_error* obtained. The corresponding MS score is 0.367. The performances of GAPS, Average Linkage and SOM are quite similar. For *Newthyroid* data, Average Linkage performs the best in terms of total $quan\_error$ obtained. The MS score corresponding to the obtained partitioning is 0.878716.

5 Discussion and Conclusion

In this paper, a new criterion for determining the appropriate partitioning algorithm for a given data set is proposed which uses a new error function other than total variance/compactness of the clusters. The error function is based on the fuzzy vector quantization-dequantization criterion. This error function gives an quantitative measurement of how well the cluster centers obtained by a particular clustering algorithm after application on a particular data set, represent the whole data set. The clustering algorithm which provides the minimal total error is regarded as the suitable partitioning technique for that particular data set. The effectiveness of the proposed criterion in detecting the proper partitioning technique among the five well-known data partitioning algorithms is established on five artificially generated and three real-life data sets.

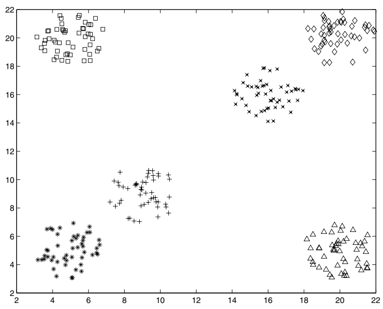


Figure 11. Clustered $Data5$ after application of GAK-means/GAPS/Average Linkage/Expectation Maximization/SOM for $K = 6$

Future work includes the use of some other distances in place of Euclidean distance while calculating the membership values of different points to different clusters, so that the proposed criterion is able to detect proper partitioning technique for some non convex/ convex symmetrical clusters other than hyperspherical ones.

References

[1] http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[2] S. Bandyopadhyay and U. Maulik. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition*, (2):1197–1208, 2002.

[3] S. Bandyopadhyay and S. Saha. GAPS: A clustering method using a new point symmetry based distance measure. *Pattern Recognition*, 40:3430–3451, 2007.

[4] A. Ben-Hur and I. Guyon. *Detecting Stable Clusters using Principal Component Analysis in Methods in Molecular Biology*. Humana press, 2003.

[5] P. Bradley, U. Fayyad, and C. Reina. Scaling EM (expectation maximization) clustering to large databases. Technical report, Microsoft Research Center, 1998.

[6] B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. London: Arnold, 2001.

[7] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[8] T. Kohonen. *Self-Organization and Associative Memory*. Springer-Verlag, New York, Berlin, 3rd edition, 1989.

[9] U. Maulik and S. Bandyopadhyay. Genetic algorithm based clustering technique. *Pattern Recog.*, 33:1455–1465, 2000.

[10] W. Pedrycz and K. Hirota. Fuzzy vector qunatization with the particle swarm optimization: A study in fuzzy granulation-degranulation information processing. *Signal Processing*, accepted, doi:10.1016/j.sigpro.2007.02.001, 2007.