Clustering Based on Principal Curve

Ioan Cleju
Multimedia Signal Processing Group
Department of Computer and Information Science
University of Konstanz
78457 Konstanz, Germany
cleju@inf.uni-konstanz.de
http://www.inf.uni-konstanz.de/cgip/index.shtml.en

Pasi Fränti
Speech and Image Processing Research Group
Department of Computer Science
University of Joensuu
P.O. Box 111, 80110 Joensuu, Finland
franti@cs.joensuu.fi
http://cs.joensuu.fi/pages/franti/vq/

Xiaolin Wu
Multimedia Computing and Communications Laboratory
Department of Electrical and Computer Engineering
McMaster University
L8G 4K1, Hamilton, Ontario, Canada
xwu@mail.ece.mcmaster.ca
http://www.ece.mcmaster.ca/~xwu/

Abstract

Clustering algorithms are intensively used in the image analysis field in compression, segmentation, recognition and other tasks. In this work we present a new approach in clustering vector datasets by finding a good order in the set, and then applying an optimal segmentation algorithm. The algorithm heuristically prolongs the optimal algorithms for scalar quantization to vector space. Scalar quantization can be optimally solved in linear time. The same technique can be used as a heuristic to vector spaces, using the principal axis as a projection space. The main drawback of this method is that the principal axis model is too rigid to preserve the adjacency of the points. For better clustering we present a way to refine the order given by the projections on the principal axis by shortening the Hamiltonian path in the data graph. Then we propose to use the principal curve to sequence the data as a one-dimensional space projection that can better model the non-linearity of the data. The experimental results show that the principal curve based clustering method can be successfully used in cluster analysis.

1. Introduction

Clustering is a general classification procedure that divides a set of objects in different classes. Objects from the same class should be similar to each other. There is no indication about their number but only the properties of the objects in the set. The interest for clustering algorithms in image analysis comes from tasks as compression, segmentation and recognition.

The research in clustering algorithms is motivated by the importance of the problem and the necessity for good and fast algorithms. The clustering problem is known to be NP-complete [17]. Scalar quantization, a special case of clustering problem, can be optimally solved in linear time [20, 18]. The main difference to vector quantization is that the scalar space is natural ordered and optimal clusters are always subsequences of the data set. The main contribution of this work is a new method based on principal curve to find a good order in the vector space and apply the segmentation algorithm similar to the one used in scalar quantization.

1.1 Problem definition

The result of the clustering is a partitioning of the original data set that maps each data point to its class, or cluster. The quality of the clustering is defined by the objective function *f* that assigns a real number to each possible clustering. Thus, the clustering problem refers to finding a clustering that optimizes the objective function. The size of the clustering is defined as the number of different clusters. The K-clustering problem is a simplification of the original clustering problem and refers to finding a clustering of a given size *K*.

In this work we will consider only the K-clustering problem, although for simplicity we might refer to it as just the clustering problem. As the objective function, we will use the mean squared error (MSE). In this sense, it is of interest to designate to each cluster a unique representative, and assign the value of the representative to the mean vector of the cluster (centroid). The set of representatives for every cluster defines the codebook. The elements of the codebook are called code-vectors. The error representation for a point is defined as the distance from the point to its corresponding code-vector. The goal is finding a clustering that minimizes the MSE. Notice that on these considerations, a specification for one of partitioning, clustering or codebook will uniquely determine the other two.

1.2 Related Work

The clustering algorithms are very diverse, a good overview can be found in [11]. At the top level, the different approaches can be classified as hierarchical and partitional. Hierarchical algorithms produce a series of partitions, known as dendogram, while partitional ones produce one partition. The dendogram can be constructed top-down, by hierarchical divisive algorithms, or bottom-up by agglomerative clustering algorithms. The most used algorithms in cluster analysis are squared error algorithms, such as K-means [15]. The resulted clustering is highly dependent on the initialization and therefore K-means is usually used to fine-tune a solution given by another algorithm. Graph theoretic algorithms model the data as a graph (e.g. minimum spanning tree) and delete the edges that are too expensive [22]. Mixture resolving approaches assume that the data was generated by an unknown distribution and try to determine its parameters [10]. Fuzzy algorithms [2] and artificial neural networks [14] have been successfully used in clustering. New approaches using genetic algorithms give also good results [5].

1.3 Overview of Our Work

In this work we study different possibilities to reformulate the clustering problem as order constrained clustering [8]. The latter problem can be optimally solved in polynomial time. Section 2 defines order constrained clustering, shows that scalar quantization is a special case of order constrained clustering and explains a method for solving it optimally. A possibility to apply a similar method on vector spaces using the clustering based on principal axis algorithm is described in subsection 2.1. The approach has been applied before, but we extend it introducing a sequence tuning method based on minimum weight Hamiltonian path, in subsection 2.2. The main contribution of this work is in section 3, which describes our new clustering algorithm based on the principal curve. Section 4 presents the results and section 5 provides the conclusions and future development possibilities.

2. Order Constrained Clustering

Order constrained clustering is a special case of clustering [8]. The data set is ordered and the clusters have to be found as subsequences of the dataset sequence. Optimal order constrained clustering refers to optimal segmentation of the data sequence, relative to the objective function. Opposed to the unconstrained clustering problem, the order constrained clustering can be optimally solved in polynomial time [8].

Scalar quantization can be formulated as a constrained clustering problem, the order being defined by the natural order in the scalar space. The solution for order constrained scalar quantization gives the optimal solution for the unconstrained problem as well, due to the convex hull of the optimal clusters. The scalar quantization is solved using the minimum weight K-link path problem [1]. An oriented graph is constructed for the ordered data set, having edges from any node to all the nodes that appear later in the sequence. The weight of an edge is equal to the distortion of one cluster that contains all the data points between the corresponding nodes. The minimum weight path from the first to the last node in the sequence, consisting of *K* edges, corresponds to the optimal clustering of the set. It can be optimally found by a dynamic programming procedure in linear time [20, 18].

The order constrained clustering problem can be formulated as the minimum weight K-link problem as well. The weights are computed accordingly in the vector space. One should notice that in the vector space the optimal clusters of the order constrained problem do not coincide with the optimal clusters for clustering problem. The quality of the result strongly depends on the order relation. Different possibilities to order the vector space will be studied.

2.1 Clustering Based on Principal Axis

One way to sequence the vector data set to obtain a good order similar order to the scalar one is to use the projections over a one-dimensional space. The principal axis of the data set, computed using the direction given by first principal component [12], is considered to be the linear projection space that probably has the most information, as it maximizes the dispersion of data. The method has been applied to color quantization in [21] to simultaneously make the first cuts in a hierarchical divisive clustering algorithm.

In clustering, the method gives good results only if the dispersion of the data along the principal axis is significantly higher than the dispersion in the other directions and data set fits the linear model. Otherwise, it can be observed that the code-vectors are too close to the principal axis (see Figure 1). One way of improving the quality is to refine the solution by the K-means algorithm. The improvement is considerable but the solution is viable only for relative simple data sets.

2.2 Revising the Order By Hamiltonian Shortest Path

The quality of the order depends on how well the spatial relations of the data points are maintained during the projection to the one-dimensional space. It directly influences the solution of the clustering, since the order constrained clusters are always found optimal. The projection on the principal axis introduces great distortions in preserving the spatial relations. We therefore aim at tuning the order by shortening the length of the path from the first element to the last element of the sequence. Finding the minimum weight Hamiltonian path (MWHP) in a graph is also an NP-complete problem [7].

We will consider the first approximation of the minimum weight Hamiltonian path as the sequence found by the principal axis projection and try to improve it by simple heuristics. First we consider short subsequences for which the minimum weight path can be found easily, and we minimize the subsequence path length correspondingly. Then, as the size of the subsequences iteratively increases, we do not look for the optimal path but for approximations.

The clustering algorithm applied to the tuned order shows significant improvements. However, the global settlement does not always improve when clusters are fine-tuned by K-means, as compared to the basic algorithm tuned by K-means (see Figure 1). Clustering along the principal axis assures a good dispersion of code-vectors along this direction. This advantage disappears as the variance on other directions and the dimensionality increase.

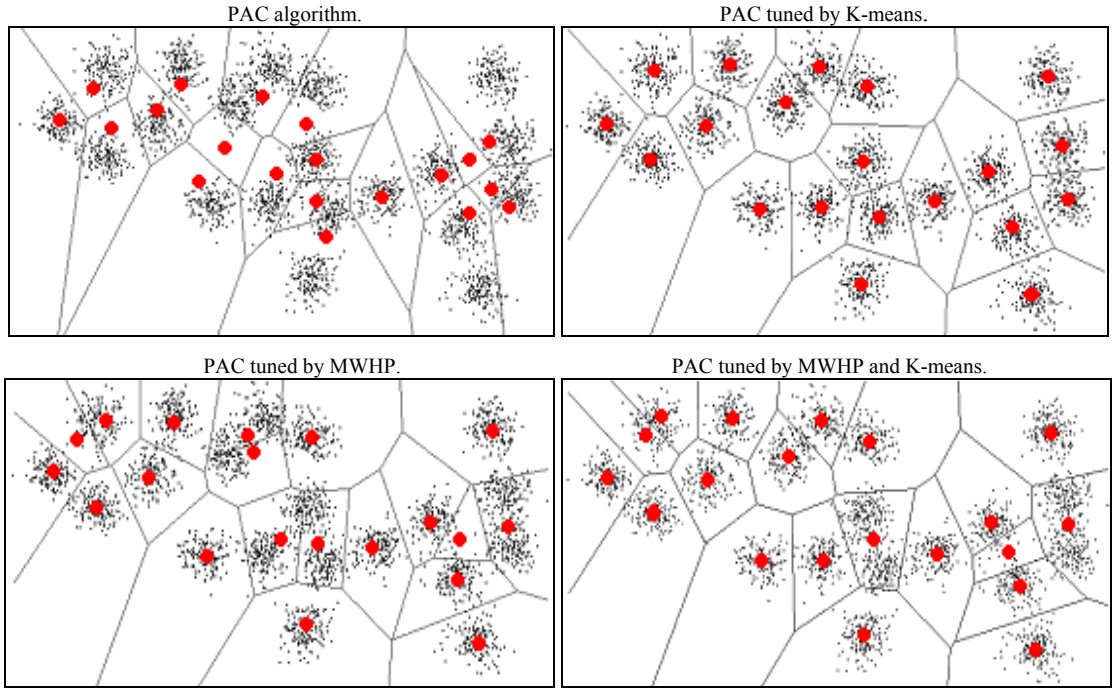

Figure 1: Results of the algorithms based on principal axis projection (PAC).

If the clusters are clearly separable, the minimum weight Hamiltonian path sequence might contain each cluster as subsequence. This does not necessarily happen if the clusters are overlapping; in this case, it is possible that the path would go through the same cluster more times. Therefore, it seems that the best order should be flexible enough to

capture the global layout of the clusters but not too detailed, in order to prevent the path jumping between clusters, or going through the same cluster several times.

3. Clustering Based on Principal Curve

As the principal axis is a linear model too rigid to fit the data, we propose to use the principal curve to project and order the dataset. The principal curve is a one-dimensional space that can capture the non-linearity from the data and can better preserve the adjacency from the space in the generated order. Figure 2 shows an example of the new algorithm. The main steps are:

- construction of the curve,
- projection over the curve and sequence the data,
- order constrained clustering, and
- form the Voronoi cells.

There are different approaches to principal curves. The paper that introduces the principal curves describes them intuitively as smooth one-dimensional curves that pass through the "middle" of data [9]. The curve is self-consistent, smooth and does not intersect itself. Next years, several other approaches have been considered. An application that uses closed principal curves to model the outlines of ice floes in satellite images is developed in [3]. In [4] an improved variant that combines the former approaches is applied to classification and feature extraction. Principal curves are defined in [13] as continuous curves of a given maximal length, which minimize the expected squared distance to the points of the space randomly chosen according to a given distribution. An incremental method similar to K-means for finding principal curves is introduced in [19]. Instead of considering the total length of the curve as in [13], the method in [16] constrains the sum of the angles along the curve (the total turn).

We propose to use the principal curve with length constraint, as it is defined in [13], to project and order the data; from now on we will simply refer to this curve as principal curve. The principal curve minimizes the distortion of the points to the curve. This assures that the adjacency of the points can be preserved on the curve.

The practical learning algorithm provided for the principal curve constructs a sub-optimal polygonal line approximation of the curve and is applied to finite discrete datasets. The main difference to the theoretical algorithm is that the length is not provided as a parameter, but the algorithm optimizes it. The procedure is iterative, each step a segment is split and then all the segments of the polygonal line are optimized.

In the segment optimization step, each vertex position is separately optimized, keeping all the other vertexes fixed. A Lagrangian formulation that combines the squared error of the data points that project on adjacent edges and the local curvature is minimized. The smoothing factor that weights the local curvature measure is heuristically found, dependent on the error, the current number of segments, the size of the dataset and the radius of the dataset. It can be controlled by the penalty coefficient. The modification of the penalty coefficient determines the shape and indirectly controls the length of the curve. A small value will determine a very long curve, at the limit just a path in the data set that has null error. A very large value will determine the principal curve to be very similar to the principal axis. The stopping criterion uses a heuristic test that can be also controlled by a parameter.

The steps that follow the curve construction after the data is sequenced are the same as for the principal axis based clustering. Dynamic programming is applied and the optimal clustering for the constrained clustering is found. Voronoi cells are then formed, and optionally K-means can be iterated. As shown in Figure 3, for a good parameterization of the curve the results are very close to the local optimum and the improvements by K-means are negligible.

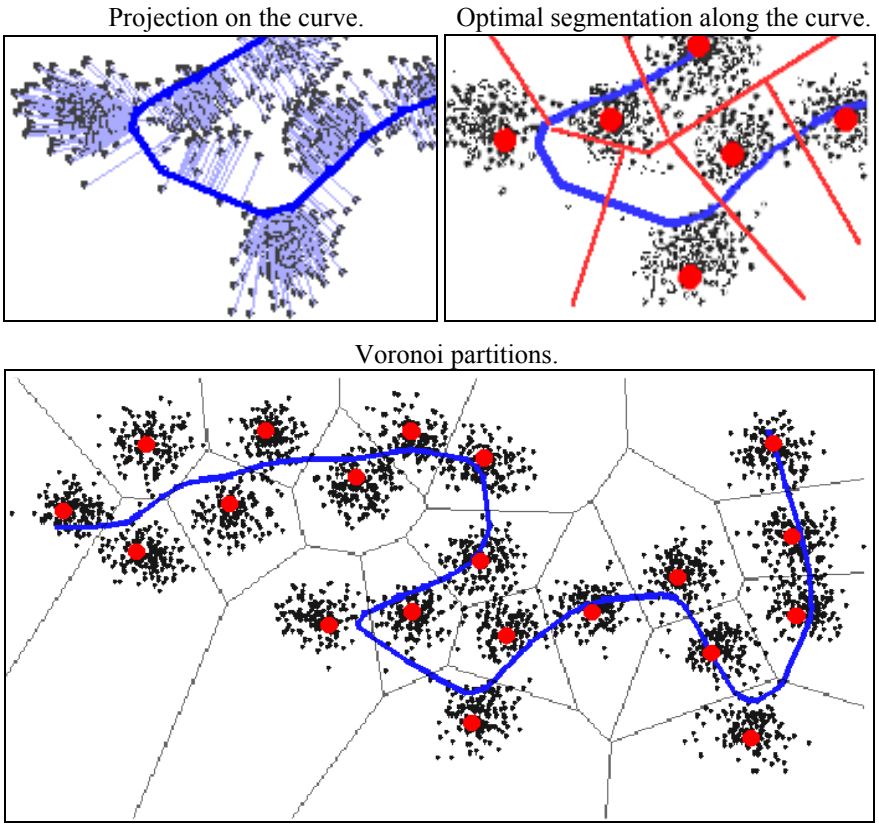

Figure 2: Projection of data points over the principal curve, segmentation of the sequence and final clusters.

3.1 Choice of Parameters

Except for the number of clusters, the other parameters for the clustering algorithm are used for the curve construction. One parameter influences the stopping criterion and the other one influences the curvature (the penalty coefficient). The parameter that controls the stopping criterion does not influence the clustering result, so we kept the default one from [13].

Changes of the penalty coefficient have much influence on the shape of the curve and on the clustering (Figure 3). Longer curves make possible for points from one cluster to project on several regions of the curve. On the other hand, shorter curves allow different clusters to project on overlapping regions. This coefficient must therefore be tuned depending on the data. Our experiments showed that the value proposed in [13] to 0.13 does not provide the best results in clustering (see Section 4).

3.2 Complexity of the Method

The principal curve algorithm has the complexity of $O(N^{5/3})$. It can be reduced if we consider the polygonal approximation of the curve with a constant number of segments S to $O(S\,N^{4/3})$. The standard algorithm considers the number of segments proportional to $N^{1/3}$. The overall complexity for clustering is $O(K\,N^2)$, given by the dynamic programming technique that is applied to order constrained clustering.

4. Experimental Results

For the experiments we have used three types of data sets. The A data sets (A1, A2, A3) are artificial and contain different numbers of two-dimensional Gaussian clusters having about the same characteristics. The S sets (S1, S2, S3, S4) are two-dimensional artificial datasets with varying complexity in terms of spatial data distributions. The real data sets (House, Bridge, Camera, Missa) come from images, representing color (House), 4×4 non-overlapping blocks of gray image (Bridge and Camera) and 4×4 difference blocks of two subsequent frames in the video sequence (Missa). Correspondingly, the data sets have 3 and 16 dimensions.

The experiments have been carried out for 15 different values for the penalty coefficient, ranging from 0.001 to 0.22. They show that for the artificial data sets that present clusters in the data, the MSE as a function of penalty coefficient clearly has a minimum. Good values are obtained for the penalty coefficient in the range 0.01 to 0.08 (see Figure 3). For data that does not have the evidence of clusters, the minimum is not clear and good clustering can be obtained for lower values of the penalty coefficient as well.

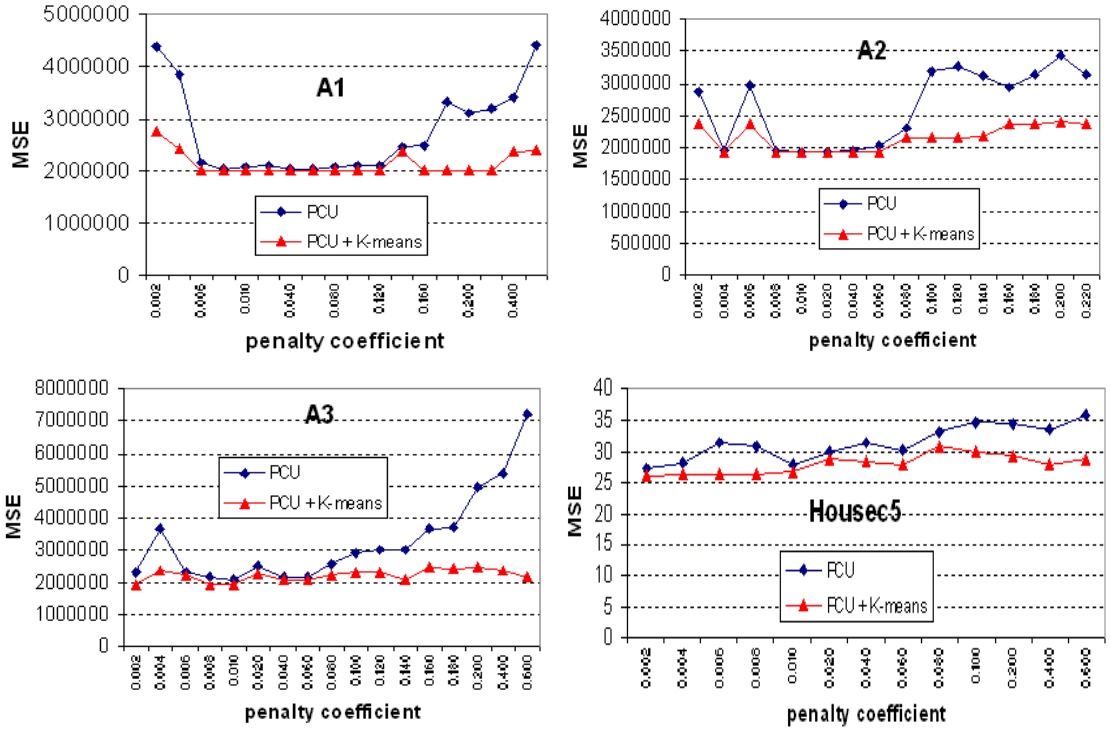

Figure 3: Results (MSE) for principal curve clustering (PCU), as a function of the penalty coefficient.

The comparative results include the popular K-means and randomized local search (RLS) [6]. The K-means MSE values are the best results obtained by 10 repeated trials. The MSE-values for the RLS method have been considered when the value of the MSE stabilizes; a slightly better solution is found after a larger number of iterations. Clustering results based on principal axis (PAC) and principal curve (PCU), with and without the K-means tuned versions, are then compared. The MSE value for the principal curve clustering was chosen as the best for the penalty coefficient in the range 0.001 to 0.22. Numerical results are shown in Tables 1, 2 and 3 respectively for the three dataset types.

The results of the principal curve clustering are significantly better than those based on principal axis. This is especially observed in the more complicated data sets A and S, where PCU performs better than PAC + K-means. The difference between the two approaches reduces for the real data sets. The real data sets present high linear correlation that enables the PAC algorithm to obtain a good result.

The comparison with repeated K-means and RLS show that the MSE values obtained are very close to each other and also to the global optimum.

Table 1: Comparison of the results for the A datasets.

Method	A sets		
	A1	A2	A3
K-means	$20.24*10^5$	$19.32*10^5$	$19.29*10^5$
RLS	$20.24*10^5$	$19.32*10^5$	$19.29*10^5$
PAC	$83.00*10^5$	$156.57*10^5$	$176.59*10^5$
PAC + K-means	$20.24*10^5$	$27.41*10^5$	$36.95*10^5$
PCU	$20.30*10^5$	$19.33*10^5$	$20.59*10^5$
PCU + K-means	$20.24*10^5$	$19.32*10^5$	$19.29*10^5$

Table 2: Comparison of the results for the S datasets.

Method	S sets			
	S1	S2	S3	S4
K-means	$134.44*10^7$	$13.27*10^8$	$16.88*10^8$	$15.70*10^8$
RLS	$89.17*10^7$	$13.27*10^8$	$16.88*10^8$	$15.70*10^8$
PAC	$840.48*10^7$	$77.34*10^8$	$57.11*10^8$	$63.40*10^8$
PAC + K-means	$143.54*10^7$	$18.65*10^8$	$16.88*10^8$	$15.70*10^8$
PCU	$89.18*10^7$	$13.29*10^8$	$16.94*10^8$	$15.91*10^8$
PCU + K-means	$89.17*10^7$	$13.27*10^8$	$16.88*10^8$	$15.70*10^8$

Table 3: Comparison of the results for the image datasets.

Method	Real data sets			
	Housec	Bridge	Camera	Missa
K-means	36.4	365	278	9.64
RLS	35.6	364	270	9.50
PAC	51.6	430	355	13.07
PAC + K-means	39.3	366	276	10.05
PCU	37.3	377	295	9.99
PCU + K-means	36.1	365	273	9.69

5. Conclusions

In this work we have considered solving the clustering problem using one-dimensional projections of the data set. We have continued studying the principal axis clustering and provide a possibility to revise the sequence in the sense of minimum weight Hamiltonian path.

The main part of the work concentrates on clustering along the principal curve. The principal curve has the advantage of being a non-linear projection that can model diverse types of data. The tests have shown that the principal curve can be successfully applied in clustering for complex datasets. The method proves to perform also on multidimensional datasets.

For the future, the parameterization of the principal curve should be considered by developing an algorithm that automatically determines or optimizes the value of the penalty coefficient.

Acknowledgments

The work was accomplished at the University of Joensuu during the M.Sc. degree studies of Ioan Cleju and was possible due to the IMPIT program financed by the mentioned university. The work of Xiaolin Wu was supported in part by NSERC, NSF and Nokia Research Fellowship.

References

[1] A. Aggarwal, B. Schieber and T. Tokuyama, Finding a Minimum Weight K-link Path in Graphs with Monge Property and Applications, In: *Proceedings of the 9th Annual Symposium on Computational Geometry*, San Diego, California, United States, May 1993, pp. 189-197.

[2] J. C. Bezdek, R. Ehrlich and W. Full, FCM: the Fuzzy c-Means Clustering Algorithm, *Computers and Geosciences*, vol. 10, 1984, pp. 191-203.
[3] J. D. Banfield and A. E. Raftery, Ice Floe Identification in Satellite Images Using Mathematical Morphology and Clustering about Principal Curves, *Journal of the American Statistical Association*, vol. 87, no. 417, March 1992, pp. 7-16.
[4] K. Chang and J. Ghosh, Principal Curves for Non-Linear Feature Extraction and Classification, In: *Proceedings SPIE*, vol. 3307, April 1998, pp. 120-129.
[5] P. Fränti, Genetic Algorithm with Deterministic Crossover for Vector Quantization, *Pattern Recognition Letters*, vol. 21, no. 1, January 2000, pp. 61-68.
[6] P. Fränti and J. Kivijäri, Randomized Local Search Algorithm for the Clustering Problem, *Pattern Analysis and Applications*, vol. 3, 2000, pp. 358-369.
[7] M. Garey and D. Johnson, *Computers and Intractability*, W. H. Freeman, January 1979, San Francisco, US.
[8] A. D. Gordon, *Classification*, Chapman and Hall, 1980, London, UK.
[9] T. Hastie and W. Stuetzle, Principal Curves, *Journal of the American Statistical Association*, vol. 84, no. 406, June 1989, pp. 502-516.
[10] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, March 1988, New Jersey, US.
[11] A.K. Jain, M. N. Murty and P.J. Flynn, Data Clustering: A review, *ACM Computing Surveys*, vol. 31, no. 3, September 1999.
[12] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, 1988, New Jersey, US.
[13] B. Kegl, A. Krzyzak, T. Linder and K. Zeger, Learning and Design of Principal Curves, *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 22, no. 3, March 2000, pp. 281-297.
[14] T. Kohonen, Self-Organizing Maps, *Springer Series in Information Sciences*, vol. 30, December 2000, Berlin, Germany.
[15] J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, vol. 1, pp. 281-296.
[16] S. Sandilya and S. R. Kulkarni, Principal Curves with Bounded Turn, *IEEE Transactions on Information Theory*, vol. 48, issue 10, October 2002, pp. 2789-2793.
[17] J. L. Slagle, C. L. Chang and S. L. Heller, A Clustering and Data-Reorganization Algorithm, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 5, pp. 121-128, 1975.
[18] F. K. Soong and B. H. Juang, Optimal Quantization of LSP Parameters, *IEEE Transactions on Speech and Audio Processing*, vol. 1, issue 1, January 1993, pp. 15-24.
[19] J.J. Verbeek, N. Vlassis and B. Krose, A k-Segments Algorithm for Finding Principal Curves, *Pattern Recognition Letters* vol. 23, issue 8, June 2002, pp. 1009-1017.
[20] X. Wu, Optimal Quantization by Matrix Searching, *Journal of Algorithms*, vol. 12, issue 4, December 1991, pp. 663-673.
[21] X. Wu, Color Quantization by Dynamic Programming and Principal Analysis, *ACM Transactions on Graphics*, vol. 11, issue 4, October 1992, pp. 348-372.
[22] C.T. Zahn, Graph-Theoretical Methods for Detecting and Describing Gestalt Clusters, *IEEE Transactions on Computers*, C-20, January 1971, pp. 68-86.