Newtonian Spectral Clustering

Konstantinos Blekas, K. Christodoulidou, and I.E. Lagaris

Dept. of Computer Science, University of Ioannina,
P.O. Box. 1186, 45110 Ioannina, Greece
{kblekas,kchristo,lagaris}@cs.uoi.gr

Abstract. In this study we propose a systematic methodology for constructing a sparse affinity matrix to be used in an advantageous spectral clustering approach. Newton's equations of motion are employed to concentrate the data points around their cluster centers, using an appropriate potential. During this process possibly overlapping clusters are separated, and simultaneously, useful similarity information is gained leading to the enrichment of the affinity matrix. The method was further developed to treat high-dimensional data with application to document clustering. We have tested the method on several benchmark data sets and we witness a superior performance in comparison with the standard approach.

1 Introduction

Given a set of data points, the problem of clustering is to discover a number of subsets, called *clusters*, that contain points with similar properties. In the literature there is a plethora of clustering approaches that have been proposed rather recently. In this work we concentrate on the class of methods which are based on spectral clustering [1], [2]. Spectral clustering has become increasingly popular during the last decade. Such algorithms are based on similarity information between data points. That is, similar data points (or points with high affinity) are more likely to belong to the same cluster than points with low affinity. These kind of algorithms have proved to be quite successful in numerous application domains, such as computer vision [3], [4], [5], speech recognition [6], bioinformatics [7], [8], text mining [9], etc.

Spectral clustering techniques make use of information obtained from an appropriately defined affinity matrix. Their primary strength is their ability to treat complex data shapes where other well-known methods (such as k-means) either cannot be directly applied, or fail. The similarity matrix must be built in such a way so as to reflect the topological characteristics of the data set. In addition *sparsity* is another desired property, since it offers computational advantages [2], [10]. In applications of computer vision and related problems, the similarity matrix is naturally sparse due to the local character of the similarities.

Methodologies leading to sparse affinity matrices have been proposed in the past [2]. For instance, the ϵ-neighborhood technique connects only points whose pairwise distances are smaller than ϵ. Another similar method is the (mutual)

C. Alippi et al. (Eds.): ICANN 2009, Part II, LNCS 5769, pp. 145–154, 2009.

k-nearest neighbor, where every point is connected only with its k nearest neighbors. However, these methods heavily depend on the choice of the control parameter (ϵ or k) that acts as a threshold for cutting some edges of the associated graph.

We present here an alternative spectral clustering method that consists of two phases. The data points are initially manipulated in a way suggested in the Newtonian clustering [11], where the original data set is transformed and the cluster appearance becomes more prominent. This is done via a dynamic procedure based on Newton's equation of motion using a properly constructed potential function. During the next phase, the affinity matrix is calculated not in the usual way, but with extra information embedded that was gained in the previous phase. At the same time this information has a sparsifying effect, and hence our affinity matrix is both sparser and richer. We further modified our method in order to treat problems of high dimensionality, such as those appearing in document clustering. The modification is carried out by choosing a different potential function and likewise a slightly different equation of motion. We have tested our method on a suite of well known benchmarks ranging from continuous feature data to image segmentation and document clustering problems. We compare to the standard spectral clustering method and the classical k-means algorithm.

In section 2 we lay out an algorithmic description of the proposed Newtonian spectral clustering while in section 3 we report experimental results for several data sets. Finally in section 4 we summarize and conclude with some remarks.

2 The Proposed Method

2.1 Spectral Clustering with a Dynamic Procedure

Let the set $X = \{x_1, \ldots, x_N\}$ denote the input set of N observations that we want to partition into K groups. We consider that the data points correspond to particles of unit mass, interacting via a two-body attractive, short-range potential. Let V_{ij} be the potential between particles located at points x_i and x_j. In this section we will consider a simple potential of Gaussian form given by:

$$V_{ij} = -\exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) , \tag{1}$$

where the scale parameter (σ) determines the range of the potential. The value of σ is important since it affects the dynamic procedure that shrinks the clusters, as well as the performance of the subsequent spectral clustering application. The determination of this parameter will be detailed later on.

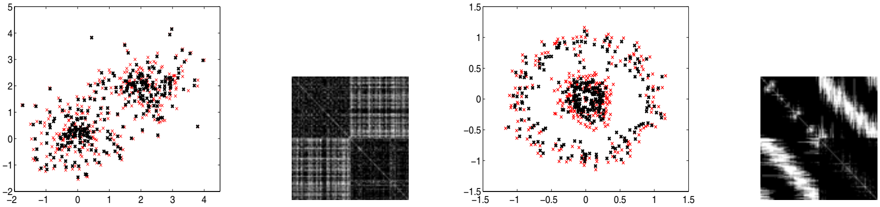
Under this consideration, the data points move under the influence of a *force*. Data that move toward different clusters either repel each other, or they are too far to interact. We expect that after an ample number of steps in time, points belong to the same cluster will come together forming to shrank clusters. The proposed dynamic procedure is governed by the Newton's equations of motion, which are:

$$\frac{d^2 x_i(t)}{dt^2} = -\nabla_i \sum_{\substack{j=1 \\ j \neq i}}^{N} V_{ij} \equiv F_i, \; \forall i = 1, 2, \cdots, N \; . \tag{2}$$

The initial positions are taken to be the original data points, i.e. $x_i(t=0) = x_i$ ($\forall i = 1, \ldots, N$), while the initial velocities ($v_i \equiv \frac{dx_i}{dt}$) are set to zero. We integrate the equations of motion in small time steps δt, considering that the forces F_i remain constant during this short time interval. At each step we reset the velocities to zero in order to avoid artifacts due to "heating". Hence we obtain the following motion scheme:

$$x_i(t + \delta t) = x_i(t) + \frac{1}{2} \delta t^2 F_i \; . \tag{3}$$

Since the interaction is attractive, after a time period T the particles belonging to the same neighborhood-cluster will concentrate around its center. So an initially spread–out cluster is being shrunk as a result of the dynamic procedure. The simulation terminates, after a certain number of steps or when the steps become too small and further iterations hardly make any difference. Two typical examples are presented in Fig. 1 (a) and (c), where the initial data points (red) are concentrated (black) after 100 steps.



(a) (b) (c) (d)

Fig. 1. Two typical examples of the effect of the dynamic procedure

The path traveled by each particle can offer useful information. In particular, at the end of the dynamic process every point $x_i = x_i(0)$ has been moved into a new position $x_i(T)$. Let $dist_{ij}(t)$ denote the distance between two points x_i and x_j at time t. The elements of the affinity matrix A are then given by:

$$A_{ij} = b_{ij} \exp(-\frac{dist_{ij}^2(T)}{2\sigma^2}) \; , \tag{4}$$

where $b_{ij} = 0$ if $dist_{ij}(T) > dist_{ij}(0)$ and $b_{ij} = 1$ otherwise. The above rule denotes that when two points move apart, they belong to different cluster and hence have zero affinity. Points that cluster together have a prominent affinity since $dist_{ij}(T) < dist_{ij}(0)$. Figure 1 (b) and (d) shows the sparsity of the affinity matrix in the case of the two artificial data sets of Figure 1 (a) and (c), respectively. More than 20% of the Affinity matrix elements are discarded (white pixels) due to the shrinking effect.

Spectral clustering is based on the data set's affinity matrix. In the literature there are several variations of the standard methodology described in [1], which we follow in our study. After having calculated the affinity matrix A, the *Laplacian* matrix L is then given by

$$L = D^{-1/2} A D^{-1/2} \ , \tag{5}$$

where D is a diagonal matrix with elements $D_{ii} = \sum_{j=1}^{N} A_{ij}$. The Laplacian matrix is known to be symmetric and positive semi-definite. Next, the K normalized eigenvectors $u_1, \ldots, u_K$ of matrix L (where K is the desired number of clusters) that correspond to the largest eigenvalues are computed and eventually fed into the k-means algorithm in order to estimate the final clustering solution.

Estimating the scale parameter $\boldsymbol{\sigma^2}$. As mentioned before, the determination of the scale parameter σ is crucial and has to be chosen carefully. Sparse data sets require a longer range than dense data sets. Hence σ depends on the data set. An automatic determination of its value was suggested in [1] by running the clustering algorithm repeatedly for a number of values of σ and selecting the one which provides the least distorted clustering solution.

In this direction, we present here a more systematic methodology. The average nearest-neighbor (NN) distance and order statistics are keys to our analysis. In particular, let the average NN distance of order m be given by

$$< d_m >= \frac{1}{N} \sum_{i=1}^{N} d_m^{(i)}, \ \ \forall \ m = 1, 2, \cdots, N-1 \ , \tag{6}$$

where $d_m^{(i)}$ is the distance between point at x_i and its m^{th} nearest neighbor. We studied its variance $\tilde{\sigma}_m^2$ as obtained by

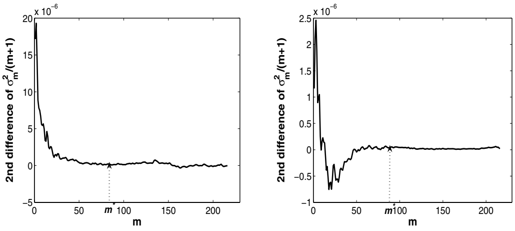
$$\tilde{\sigma}_m^2 = \frac{1}{m} \sum_{k=1}^{m} \left(< d_k^2 > - < d_k >^2 \right) \ , \tag{7}$$

using order statistics. It was found in [11] that in the case of a single cluster the functional form of $\tilde{\sigma}_m^2$ is given as

$$\tilde{\sigma}_m^2 = \alpha (m+1)^2 + \beta (m+1) \ . \tag{8}$$

When there are more than one clusters within the data set, $< d_m >$ acquires discontinuities and the cumulative quantity $\tilde{\sigma}_m^2$ is given by a superposition of translated quadratics. Then, the value for the range of the potential is estimated by finding the number of neighbors m^* for which the second difference of $\frac{\tilde{\sigma}_m^2}{m+1}$

(with respect to m) vanishes. Figure 2 illustrates this behavior by plotting the quantity $\frac{\tilde{\sigma}_m^2}{m+1}$ versus m, in the case of two typical examples of Fig. 1. As our experiments have shown, there is a wide stability region around m^* for estimating the range value ($\sigma^2 = \sigma_{m^*}^2$), where the performance of our approach was identical. A detailed description of the above method for estimating the proper value of σ can be found in [11].



(a) (b)

Fig. 2. Plots of second difference of quantity $\frac{\tilde{\sigma}_{M,m}^2}{m+1}$ with respect to m where the number m^* is estimated

2.2 Extension to Document Clustering

An important issue in clustering is treating high-dimensional data. Since spectral clustering is a common technique used for this purpose, we have tried to adjust the proposed method to deal with such problems. Document clustering is a very interesting application in text mining and information retrieval, aiming to the division of a collection of documents into groups based on their similarity.

In our study, each input document is transformed into a feature vector $x_i \in R^M$, where M is the size of the corpus vocabulary, such that every feature denotes the weight of the corresponding term. We have applied the TF-IDF (term frequency, inverse document frequency) weighting scheme for creating feature vectors. Moreover, the proximity between each pair of documents is computed used the cosine similarity metric. Since documents are normalized vectors, the similarity measure is reduced to the following simple rule:

$$V_{ij} = x_i^T x_j \ . \tag{9}$$

The above metric is also used as the potential function V_{ij} during the Newtonian dynamic procedure (see Eq. 1 - Eq. 3). The introduction of such kind of potential requires an alternative motion scheme of data points. Now, the interaction is not always attractive. Naturally, any particle is influenced positively from similar documents (that belong to the same cluster) and their interaction is attractive (positive force). In the opposite case, dissimilar document vectors have a repulsive effect to the particle and thus offering a negative sign force within its motion update rule. It can be easily found that the formulation of the force F_i now becomes as:

$$F_i = \sum_{\substack{j=1 \\ j \neq i}}^{N} c_{ij}(t) x_j \text{ , where } c_{ij}(t) = \begin{cases} +1 \text{ if } V_{ij} > \overline{V}_{ij}/2 \\ -1 \quad \text{otherwise} \end{cases} . \tag{10}$$

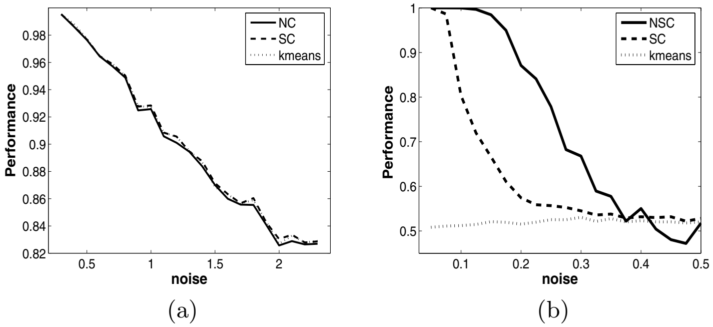
In fact the quantity $\overline{V}_{ij}/2$ acts as a threshold similarity value for distinguishing between attractive and repulsive documents.

3 Experimental Results

Several experiments have been performed in order to examine the effectiveness of the proposed Newtonian Spectral Clustering approach (NSC). We have considered both simulated data sets and other widely used benchmarks. We compare with the standard Spectral Clustering (SC) and the traditional k-means algorithm. During all experiments the number of Newtonian steps was fixed at $T = 100$, while the value of time step was set to $\delta t = 10^{-5}$. Moreover, both approaches, NSC and SC, used the same value of σ in the Gaussian similarity function as estimated by the proposed method. Finally, since we were aware of the true class label of data, all clustering methods were evaluated using the purity metric (classification accuracy), by assuming that all objects of a cluster are assigned to its dominant class.

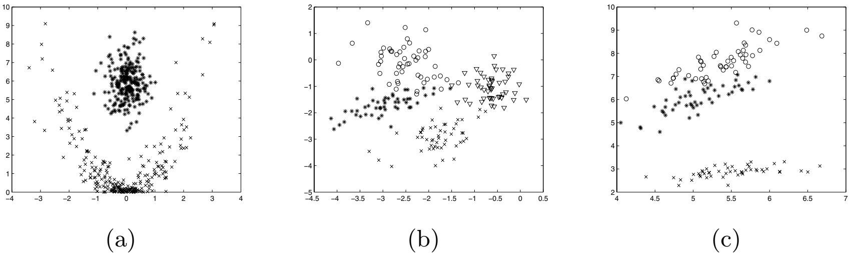
The first series of experiments was performed on two simulated datasets (150 points per class) with two class ($K = 2$) presented in Fig. 1 (a) and 1 (c). By considering different levels of noise, we performed 50 experiments for each noise value and kept record of the mean accuracy for every method. The depicted comparative results are illustrated in the two diagrams of Fig. 3 in terms of different noise values. As it was expected, in the first data set with two spheres all three methods displayed identical behavior, since data were generated by sampling from two Gaussian densities that have the same spherical-type covariance matrix of the form $\sigma^2 I$. In the second data set (Fig. 1 (c)) which is more complex with two concentric clusters, our method performs better than the standard SC method especially in high-level noisy environments. The traditional k-means algorithm fails in situations with non-spherical data shapes.

Additional experiments were made using four known benchmarks (Fig. 4). The first one Fig. 4(a) is a two-class problem with a moon and a sun shape, while the next Fig. 4(b) is the CRAB data set of Ripley [12], that contains $N = 200$ data belonging to four clusters ($K = 4$). Here, we have created a 2-dimensional data set by projecting the data on the plane defined by the second and third principal components. We have also studied two UCI benchmarks [13] the renowned Fisher-IRIS data set Fig. 4(b) with $N = 150$ points belonging to three clusters ($K = 3$) (projected on the plane using the first two principal components), and the wine set consisted of $N = 178$ $K = 3$-classes data with 13 features (were we have applied zero-mean normalization). Table 1 summarizes



(a) (b)

Fig. 3. Comparative results of NSC and its peers in terms of noise value using the two data sets of Fig. 1



(a) (b) (c)

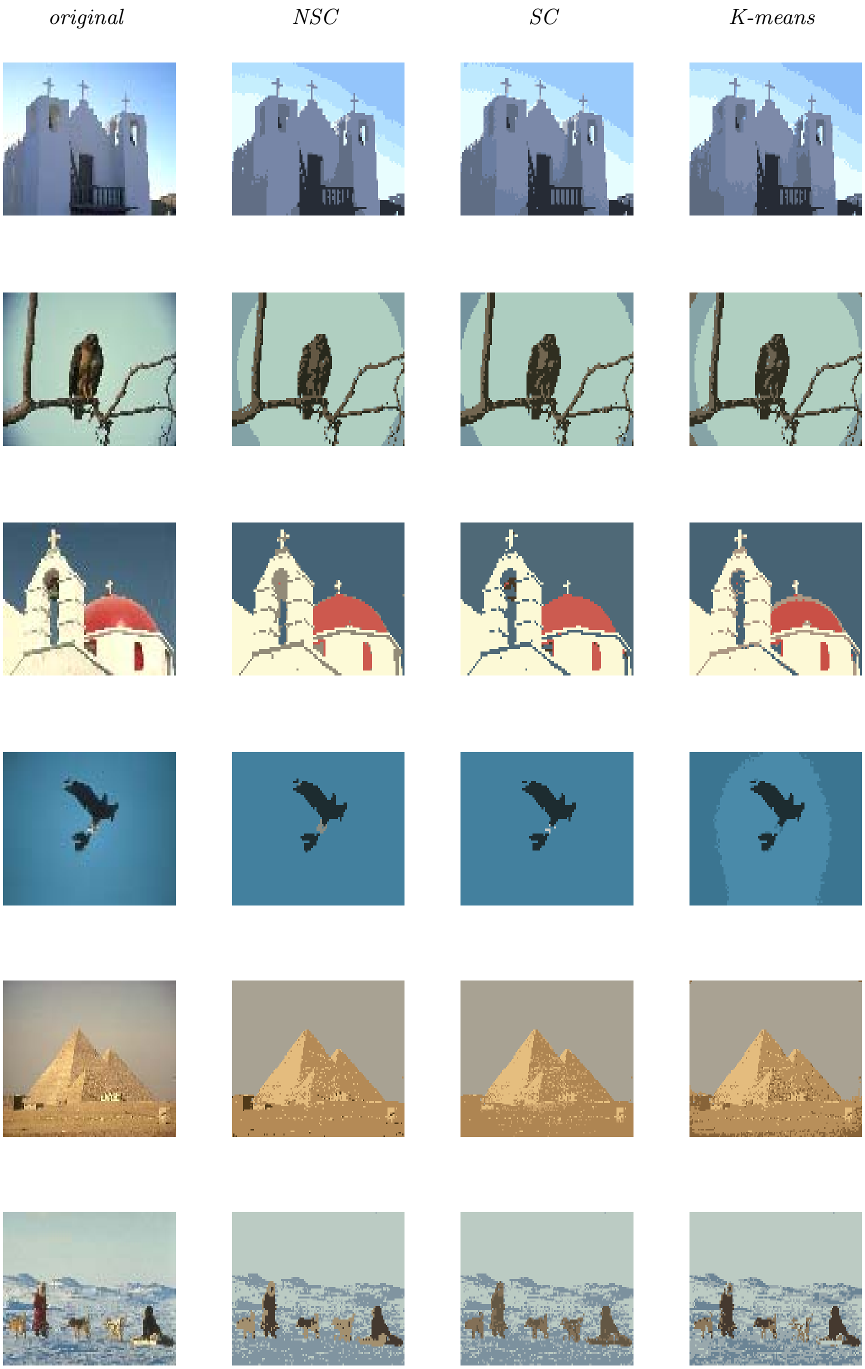
Fig. 4. Three known benchmarks used in our experiments: (a) the moon & sun, (b) the crabs and (c) the iris data set

the results obtained by the application of the three comparative approaches to the above mentioned data sets. In these cases the performance of NSC and the SC yielded comparable results, while being superior to k-means.

We have also applied our method to tackle the problem of image segmentation. For this purpose, we have selected six colored images from the Berkeley segmentation database[1] presented in Fig. 5, all with resolution around 150×150. We note here that in this series of experiments, since the number of input data is large, we have followed the Nyström method [14] for finding a numerical approximation to eigendecomposition. Fig. 5 illustrates the segmentation results of each method, where in the reconstructed images every pixel takes the intensity value of the cluster center that belongs. It is interesting to notice here that the NSC creates much smoother regions in comparison with the standard SC. We believe that if we take into account additional information, such as spatial, texture, etc. the resulting segmentation will be improved.

Finally, we have studied the performance of our method when dealing with high-dimensional spaces. For this purpose we have selected sets of documents

[1] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/



original *NSC* *SC* *K-means*

Fig. 5. Segmentation results obtained by three comparative clustering methods in six real colored images. NSC creates smoother regions.

Table 1. Comparative results using four known experimental data sets

Experimental dataset	*Performance of*		
	NSC	*SC*	*k-means*
moon & sun (Fig. 4(a))	0.94	0.94	0.92
crabs (Fig. 4(b))	0.94	0.93	0.93
iris (Fig. 4(c))	0.93	0.91	0.89
wine	0.98	0.98	0.97

and in particular four subsets of the popular 20-Newsgroup collection[2]. Their characteristics are presented in Table 2. The first set Talk$_3$ consists of documents of the talk subjects (politics.guns, politics.mideast, politics.misc), the next two of scientific documents (crypt, electronics, med, space), and the fourth set has documents from five newsgroups (comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast). Table 2 shows also the results from the above data obtained by both approaches, NSC and SC. As it can be observed, the performance of our method is significantly better showing that the proposed way of constructing affinity matrix is worthwhile in high-dimensional data. Several other experiments were made with other subsets from the same data collection with similar results.

Table 2. Document data used in our experiments and the accuracy results obtained by both NSC and SC methods

Document dataset		*Performance of*	
name	*description*	*NSC*	*SC*
Talk$_3$	$N = 300, K = 3, M = 4515$	0.78	0.71
Science$_4$-400	$N = 400, K = 4, M = 4855$	0.71	0.62
Science$_4$-2000	$N = 2000, K = 4, M = 10250$	0.74	0.73
Multi$_5$	$N = 500, K = 5, M = 5589$	0.75	0.63

4 Conclusions

In this study we presented a novel method, the Newtonian spectral clustering, that inherits from Newtonian clustering information such that renders possible the formation of a proper affinity matrix that is sparse and contains enriched information. An extension of this approach has also been presented in order to deal with high-dimensional data such as documents. We have applied the method to several benchmark problems and we noticed performance superior to the standard spectral clustering approach. It is our intention to further pursue and develop the method to handle different problems with complex type of data such as time-series, multimedia data, discrete sequences, etc. Finally, the persistent

[2] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/news20.html

issue of discovering the optimal number of clusters may be examined in the framework of this method as well.

References

1. Ng, A., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 14, pp. 849–864 (2001)
2. Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing 17, 395–416 (2007)
3. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Recognition and Machine Intelligence 22, 888–905 (2000)
4. Park, J., Zha, H., Kasturi, R.: Spectral clustering for robust motion segmentation. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 390–401. Springer, Heidelberg (2004)
5. Chang, H., Yeung, D.: Robust path-based spectral clustering with application to image segmentation. In: Proc. Intern. Conf. on Computer Vision, pp. 278–285 (2005)
6. Bach, F., Jordan, M.I.: Learning spectral clustering, with application to speech seperation. Journal of Machine Learning Research 7, 1962–2001 (2006)
7. Pentney, W., Meila, M.: Spectral clustering fof Biological sequence data. In: Proc. of the 25th Annual Conference of AAAI, pp. 845–850 (2005)
8. Higham, D.J., Kalna, G., Kibble, M.: Spectral clustering, and its use in bioinformatics. Journal of Computational and Applied Mathematics 204, 25–37 (2007)
9. Dhillon, I.S.: Co-clustering documents and words using bipartite spectral graph partitioning. In: Proc. seventh ACM SIGKDD Intern. Conf. on Knowledge Discovery and Data mining (KDD), pp. 269–274 (2001)
10. Chen, W., Song, Y., Bai, H., Lin, C., Chang, E.Y.: Parallel spectral clustering. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 374–389. Springer, Heidelberg (2008)
11. Blekas, K., Lagaris, I.E.: Newtonian clustering: an approach based on molecular dynamics and global optimization. Pattern Recognition 40, 1734–1744 (2007)
12. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge Univ. Press Inc., Cambridge (1996)
13. Merz, C.J., Murphy, P.M.: UCI repository of machine learning databases. Irvine, CA (1998), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
14. Fowlkes, C.S., Belongie, F., Chung, F., Malik, J.: Spectral grouping using the Nyström method. IEEE Trans. on Pattern Analysis and Machine Intelligence 26, 214–225 (2004)