ISOMAP based metrics for clustering

Ariel E. Bayá[1] and Pablo M. Granitto[1]

Instituto de Física Rosario, CONICET/UNR,
Bv 27 de Febrero 210 Bis, 2000 Rosario, República Argentina
{abaya,granitto}@ifir.edu.ar

Abstract. Many successful clustering techniques fail to handle data with a manifold structure, i.e. data that is not shaped in the form of compact point clouds, forming arbitrary shapes or paths through a high-dimensional space. In this paper we present a new method to evaluate distances in such spaces that naturally extend the application of many clustering algorithms to these cases. Our algorithm has two stages. Following ISOMAP, it searches for sets of locally-uniform manifolds, which could be disjoint. These manifolds are then connected using two slightly different strategies. We compare these strategies between them and with a state of the art algorithm using three artificial problems, obtaining encouraging result. Both new metrics allow diverse algorithms to easily find clusters of arbitrary shape.

1 Introduction

Clustering is a fundamental topic in machine learning and pattern recognition. Its final aim is to find any arbitrary structure hidden in a set of data. Typical applications are, for example, characterizing customer groups based on purchasing patterns, categorizing Web documents, grouping genes and proteins that have similar functionality, grouping spatial locations prone to earthquakes based on seismological data, etc.

The problem of finding clusters efficiently has been widely studied in the literature [1]. Clustering algorithms are usually divided into hierarchical and partitional methods. Hierarchical algorithms (HA) find successive clusters using previously defined ones, in an agglomerative ("bottom-up") or divisive ("top-down") way. Divisive algorithms begin with the whole dataset and proceed to divide it into successively smaller clusters. Agglomerative algorithms, on the other side, begin with each element as a separate cluster and merge them into successively larger clusters. The result of this process is a binary tree, called a dendogram. Each level of the dendogram represents a particular clustering of the data, so it is (usually) left to the user to decide at which level the process will be stopped, and consequently how many clusters will be obtained. Partitional algorithms determine a fixed number of clusters, all at once. One of the most used approaches is the K-Means [2] algorithm (or its improved version PAM[3]) that, starting from k random clusters, search iteratively for a locally optimal solution of the clustering problem. Recently, Frey and Dueck[4] proposed the innovative

and computationally efficient Affinity Propagation (AP) algorithm. With this method each data point is viewed as a node in a network. Nodes exchange messages until a set of cluster-centers emerges as a solution. The algorithm shares characteristics with both hierarchical and partitioning methods.

Despite the success all these methods showed in several artificial and real life datasets, they fail to handle data that exposes a manifold structure, i.e. data that is not shaped in the form of compact point clouds, but forms arbitrary shapes or paths through a high-dimensional space. Recently, several methods for characterizing the possible non-linear manifold where a dataset would lie were developed, such as Locally Linear Embedding[5], Laplacian Eigenmaps[6] or ISOMAP[7]. All these methods look for local neighbourhood relations that can be used to produce low dimensional projections of the data at hand. Ham et al.[8] showed that they can also be interpreted as particular kernels.

In this paper we present a new strategy to evaluate distances in manifold spaces that naturally extend the application of the previously described clustering methods to these cases. Our algorithm has two stages. Following ISOMAP, it first searches for locally uniform manifolds (which could be disjoint). Two slightly different strategies are then used to connect the disjoint manifolds found by the first stage.

The rest of this paper is organized as follow. In Section 2 we introduce our new method and compare the two suggested strategies to construct a fully connected manifold. In section 3 we show results on several artificial datasets. Finally we draw some conclusions and discuss future lines of research.

2 Measuring distances in arbitrarily manifolds

Our strategy for creating a new metric useful for non-linear clustering has two stages. First we search the original dataset space for locally dense structures. As we are trying to find more than one cluster in our dataset, we set our search appropriately in order to find several disjoint structures. Naturally, when working with finite noisy samples, this would probably create too many separated structures, maybe more than one for each real cluster. We then use another (penalized) metric to join these structures and let the final decision about how to form the clusters to any selected clustering algorithm.

As we stated in the introduction, several algorithms for discovering low dimensional manifolds were recently introduced. Among them, ISOMAP has strong theoretical properties and is also easy to understand. We follow the main ideas from ISOMAP to search for locally connected structures. As explained by Tenenbaum, de Silva and Langford[7], in a curved manifold the geodesic distance between neighbouring points can be correctly approximated by the Euclidean input space distance, but for faraway points geodesic distances are better approximated by adding a series of short hops between neighbouring points. To this end, we construct the k-nearest neighbours graph of the data, i.e. the graph with one vertex per observed example, and arcs between each vertex and its k near neighbours with weights equal to the Euclidean distance between them. In

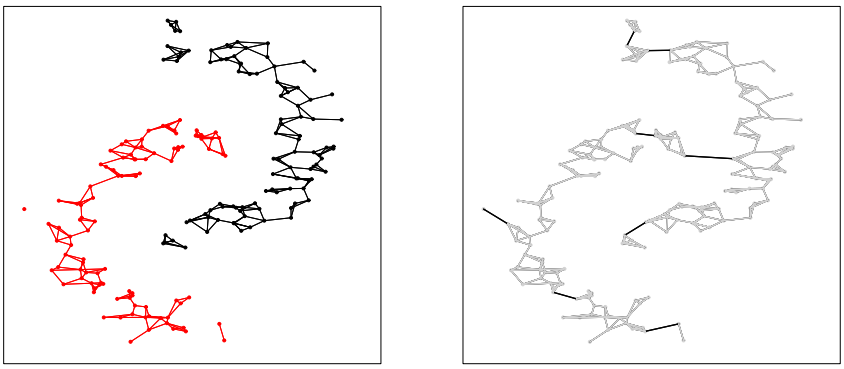

Fig. 1. Two-arcs dataset. Left Panel: Original data with the corresponding 3-nearest neighbours graph, showing edges between connected vertices. Right Panel: Thick black lines show added between-sub-graphs connections.

Figure 1 we show an example of this process. Left panel shows the original data with the corresponding 3-nearest neighbours graph. The use of a low number of neighbours (3 to 7 in all our cases) produces graphs that follow the curved structure of our data. With this graph we can know compute geodesic distances between faraway points using computational efficient algorithms like Floyd[9] or Dijkstra[10].

As can be seen in the Figure, one problem is that the use of a low k produces several disjoint sub-graphs. In the ideal case, each of these sub-graphs should correspond to a unique cluster, and we could left them disjoint and interpret distances between disjoint vertices as equal to infinite. In the real case, however, some sub-graphs correspond to the same cluster (as can be seen in Figure 1, left panel), so we need to evaluate also geodesic distances between points lying in different sub-graphs. With that purpose, we add to the graph the minimum number of edges, each of them of minimum length, which fully connect the graph. On Figure 1, right panel, we show the result of this last process (we will call these added edges "between-sub-graphs connections"). The key point of our method is that we use for all these new connecting edges the original Euclidean metric but penalized with an exponential factor, in the form:

$$w = d \quad e^{d/\sigma}, \tag{1}$$

where w is the graph weight corresponding to the added edge, d is the Euclidean distance between the points being connected by that edge and σ is the mean Euclidean distance between nearest neighbours in the graph. Using this metric we can connect sub-graphs corresponding to the same cluster with a relatively small cost, because connections in the same order of magnitude of σ will get a

low penalization. On the other hand, edges connecting faraway sub-graphs will be strongly penalized. We call this metric ISOMAP-min.

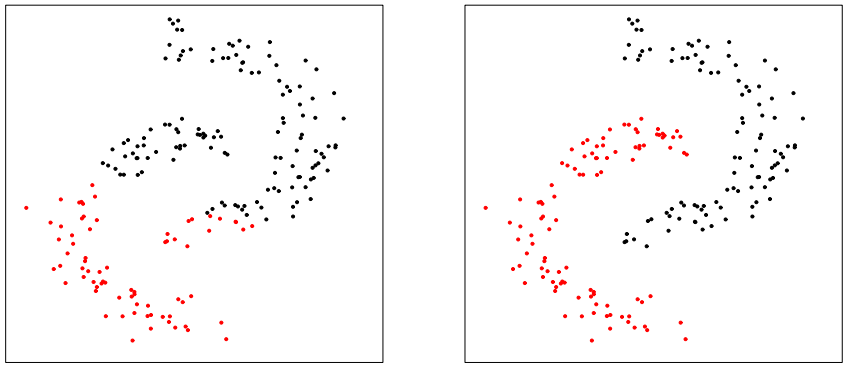

Fig. 2. Clustering results for the Two-arcs dataset. Left Panel: PAM algorithm with Euclidean metric. Right Panel: PAM algorithm with ISOMAP-min metric. Assigned clusters are shown with colors.

Figure 2 shows the result of this process. The dataset, called two-arcs, correspond to points uniformly sampled from two arcs of circumference, with fixed Gaussian noise added to the radial direction. Both panels of the Figure show the clusters produced by PAM , left panel based on Euclidean distances and right panel corresponding to our new metric. Figure 3 show another artificial dataset, called three-spirals. In this case the data was generated sampling uniformly from three equally separated spirals, again with added Gaussian noise but this time proportional to the radial distance. The result is a clustering problem with a non-uniform density, which can confuse some algorithms. PAM algorithm was applied to Euclidean distances (left panel) and to our new metric (right panel). Again, our method produces a perfect clustering, showing that is not affected by the changing density.

In Figure 4 we show the results on the three-rings dataset. On this dataset the central cluster correspond to a uniform sampling of a circle, which is surrounded by two ring, also sampled uniformly but with constant Gaussian noise added to the radial component. This third dataset also has zones with different densities. Our method produces a perfect separation of the three clusters, not achieved by the Euclidean metric.

This dataset, in fact, presents a particular challenge to our method. In Figure 5, left panel, we show the 5-nearest neighbours graph for a different sample of this dataset. The data has complete axial symmetry. Points from any ring have

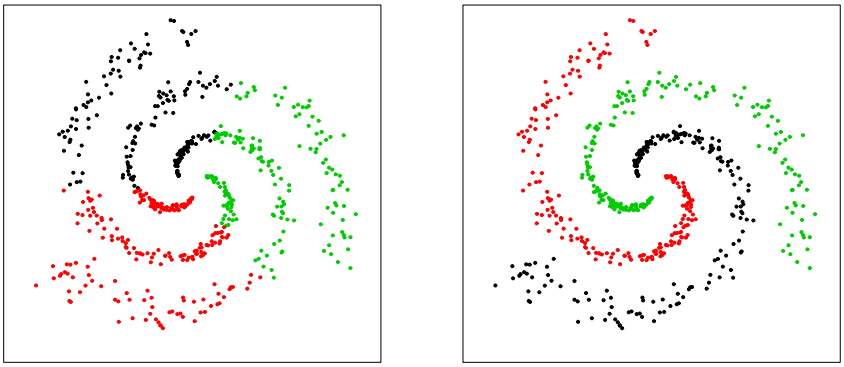

Fig. 3. Clustering results for the Three-spirals Dataset. Left Panel: PAM algorithm with Euclidean metric. Right Panel: PAM algorithm with ISOMAP-min metric. Assigned clusters are shown with colors.

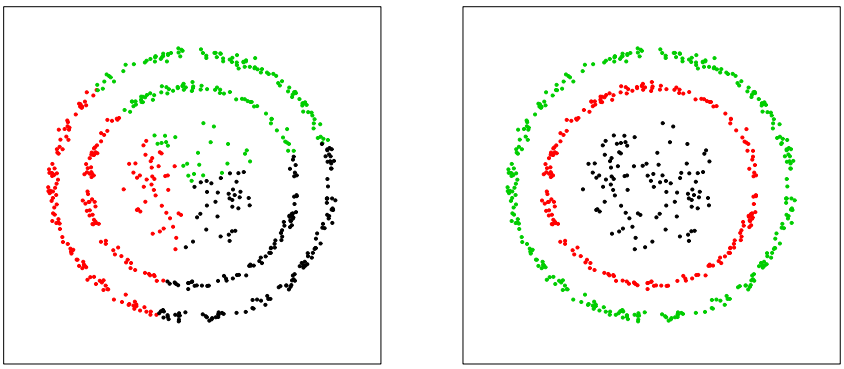

Fig. 4. Clustering results for the Three-rings Dataset. Left Panel: PAM algorithm with Euclidean metric. Right Panel: PAM algorithm with ISOMAP-min metric. Assigned clusters are shown with colors.

almost the same distance to points in a different ring. In the right panel we show the few edges added by our procedure. As can be seen, the upper-half of the middle-ring (red points in the left panel) has been connected to the graph trough the upper-half of the outer-ring. As a consequence, the distance between points in the upper-half of the middle-ring and points in the inner-circle (black) will be completely different to the distance between the lower-half and the inner-circle, which is an arbitrary arrangement guided by the noise in the dataset. Another sample from the same distribution could connect the half-rings in an opposite way, leading to a completely different distances scheme. We can see this problem graphically using a Multi-Dimensional Scaling (MDS) [11] of the distances between points (MDS produces projections on low dimensional subspaces that preserve as much as possible the original distances between vertices). In Figure 6, left panel, we show the first two components of the MDS projection of the sample in Figure 5. To this metric, the outer-ring (green) is located between both halves of the medium-ring. As this arrangement is driven by noise, it is clearly unstable. We will show in the next section that this instability does not affect the performance of the method.

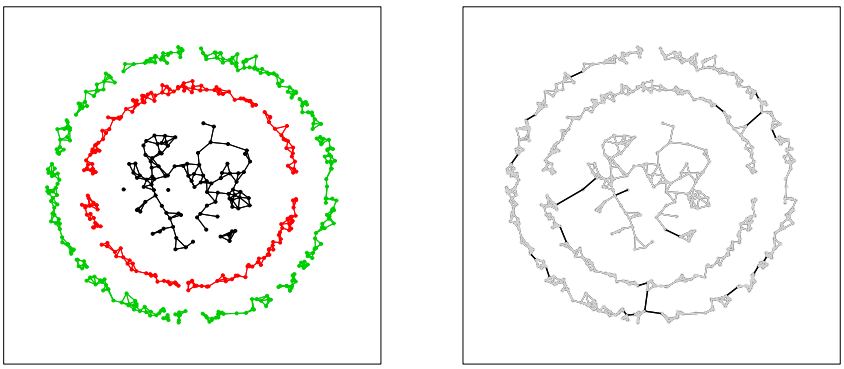

Fig. 5. Three-rings dataset. Left Panel: Original data with the corresponding 3-nearest neighbours graph, showing edges between connected vertices. Right Panel: Thick black lines show added between-sub-graphs connections.

We can make our metric more stable changing only the way we connect the different sub-graphs. Instead of searching for the minimum distance edges, we can add edges between all not-connected vertices in the graph using the penalized metric of eq. 1 in all cases. Then, using again Floyd's[9] or Dijkstra's[10] algorithm, we can compute geodesic distances using the fully connected graph. We call this second metric ISOMAP-all. For most datasets there will be no real

differences between the two connection methods, because ISOMAP-all is an extension of ISOMAP-min. In problems with high symmetry like three-rings there will be a lot of similar connections between the rings, making the ISOMAP-all metric less dependent of noise. In Figure 6, right panel, we show the same MDS projection as before but now using this new connection scheme. For ISOMAP-all metric all the medium-ring (red) is located in the right place, between the inner-circle and the outer-ring.

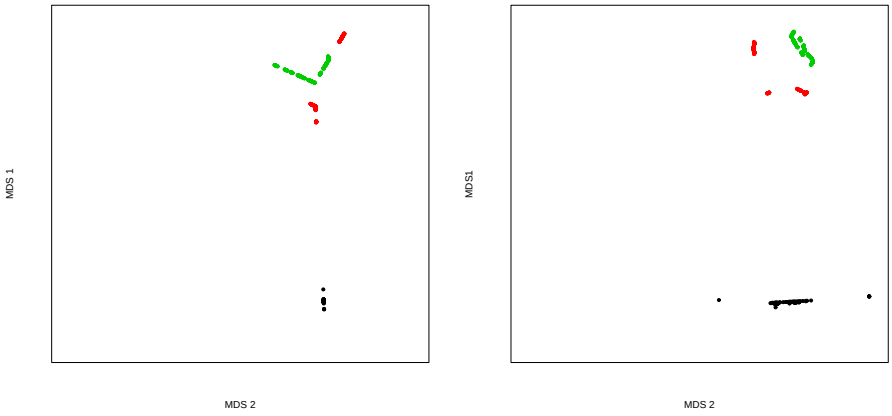

Fig. 6. MDS projections of the sample of the Three-rings Dataset analyzed in Figure 5. Left Panel shows the first two components of the MDS projection of the sample with the ISOMAP-min metric. Right panel shows the same information for the ISOMAP-all metric. The color code is the same as in Figure 5.

We can compare the stability of our two metrics with the following experiment. For each of the 3 datasets, we generate a sample of a given length. Then we repeat 100 times the following procedure: i) we generate another sample of the same length and add it to the first (fixed) sample and ii) using the two metrics we evaluate the inter-points distance in the first (fixed) sample (using the connected graph constructed with both the fixed and variable samples together). We then evaluate the variation over the 100 evaluations measuring for each pair of points the normalized standard deviation of the distance between them (the standard deviation divided by the mean value). As we are evaluating distances over fixed points the normalized standard deviation should be as lower as possible for a stable method. In Figure 7 we show the difference between the normalized standard deviation of both metrics (ISOMAP-min minus ISOMAP-all) for each point as boxplots. In the left and central panels we analyze the first two datasets, two-arcs and three-spirals, for three different noise levels. In this cases all values are very low but all distributions have a right-hand tail of outliers, showing that, as

expected, both metrics work fine for these datasets, being ISOMAP-all slightly more stable than ISOMAP-min. In the right panel of Figure 7 we show the same statistic for three different noise levels of the three-rings dataset. In this case the differences between both metrics are notorious. Interestingly, the difference decreases with an increase in noise levels (the decrease between medium and high noise is difficult to appreciate in the Figure due to the reduced scale). This can be explained taking into account that a bigger spread of the rings produces smaller between-sub-graphs distances (which are exponentially penalized).

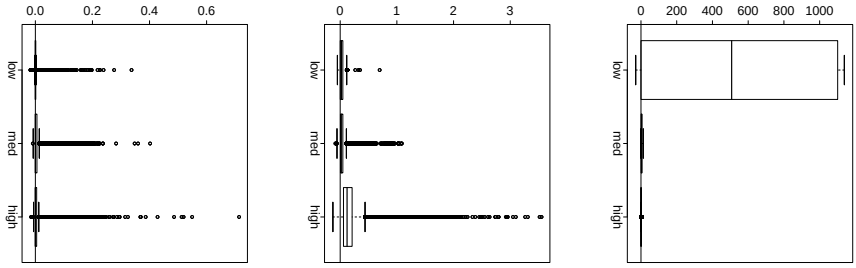

Fig. 7. Comparative stability analysis of ISOMAP-min and ISOMAP-all, see main text. From left to right, for the Two-arcs, Three-spirals and Three-rings datasets. Note the different scales for each boxplot.

3 Results

We evaluate our new metrics on the three previously introduced datasets, using for each one three different noise levels that increase the difficulty of the clustering problem. For the Three-rings dataset we split both middle- and outer-rings in halfs (adding a small gap), to consider the dataset as a more difficult 5 clusters problem. The quality of the solutions is evaluated in terms of the clustering accuracy, i.e. the percentage of the data set assigned to the right cluster. We use the 3 clustering algorithm described in the introduction, PAM, HC and AP, to evaluate the performance of our new metrics with diverse types of clustering methods. In Tables 1, 2 and 3 we show the corresponding results. It is clear that both new metrics show similar very good results, even for the Three-rings datasets where ISOMAP-min was shown to be unstable. Even more, the good performance sustains for all three clustering methods.

In order to compare these results with the state of the art in clustering methods, we repeated all the experiment with the Spectral Clustering[12] algorithm. This method has recently received a lot of attention and is considered to be effective in arbitrary manifolds. In each run we set the σ parameter proportional to the mean distance between points in the datasets. Table 4 show the corresponding results. For the Two-arcs dataset the results are similar to ISOMAP based

metrics. For the other two datasets, however, our two ISOMAP based metrics show better results than Spectral Clustering.

Metric	Algorithm	Low noise	Medium noise	High noise
Euclidean	AP	53 ± 5	53 ± 6	55 ± 5
	PAM	76 ± 2	77 ± 3	77 ± 3
	HC	80 ± 4	81 ± 5	79 ± 5
Isomap Min	AP	100 ± 0	100 ± 0	99.97 ± 0.06
	PAM	100 ± 0	100 ± 0	99.97 ± 0.06
	HC	100 ± 0	100 ± 0	99.97 ± 0.07
Isomap All	AP	100 ± 0	100 ± 0	99.97 ± 0.06
	PAM	100 ± 0	100 ± 0	99.97 ± 0.06
	HC	100 ± 0	100 ± 0	99.97 ± 0.08

Table 1. Clustering results for the Two-arcs Dataset, for the Euclidean and our two newly introduced metrics and three different clustering algorithms: PAM, Affinity Propagation (AP) and Hierarchical Clustering (HC). The results are mean ($\pm$ standard deviation) accuracies on 100 runs of each dataset.

Metric	Algorithm	Low noise	Medium noise	High noise
Euclidean	AP	33.33 ± 0	41 ± 6	40 ± 6
	PAM	42 ± 3	42 ± 2	42 ± 3
	HC	46 ± 5	45 ± 5	46 ± 6
Isomap Min	AP	100 ± 0	100 ± 0	99 ± 3
	PAM	100 ± 0	100 ± 0	98 ± 3
	HC	100 ± 0	100 ± 0	98 ± 4
Isomap All	AP	100 ± 0	100 ± 0	99 ± 3
	PAM Isomap	100 ± 0	100 ± 0	99 ± 3
	HC	100 ± 0	100 ± 0	99 ± 2

Table 2. Clustering results for the Three-spirals Dataset, for the Euclidean and our two newly introduced metrics and three different clustering algorithms: PAM, Affinity Propagation (AP) and Hierarchical Clustering (HC). The results are mean ($\pm$ standard deviation) accuracies on 100 runs of each dataset.

4 Conclusions

In this work we introduced the ISOMAP-min metric for clustering arbitrary manifolds. This metric is based on a combination of the use of the Euclidean distance measured on a neighbours graph (following the ISOMAP algorithm)

Metric	Algorithm	Low noise	Medium noise	High noise
Euclidean	AP	$- \pm -$	40 ± 3	40 ± 3
	PAM	23 ± 3	24 ± 4	26 ± 5
	HC	50 ± 0.2	50 ± 0	50 ± 0
Isomap Min.	AP	100 ± 0	86 ± 6	86 ± 4
	PAM	100 ± 0	87 ± 7	80 ± 10
	HC	100 ± 0	85 ± 7	82 ± 6
Isomap All	PAM	100 ± 0	87 ± 7	87 ± 5
	AP	100 ± 0	86 ± 5	80 ± 10
	HC	100 ± 0	81 ± 7	79 ± 9

Table 3. Clustering results for the Three-rings Dataset, for the Euclidean and our two newly introduced metrics and three different clustering algorithms: PAM, Affinity Propagation (AP) and Hierarchical Clustering (HC). The results are mean ($\pm$ standard deviation) accuracies on 100 runs of each dataset. The result for Euclidean-AP-Low noise is empty because we were not been able to find converging parameters for this case.

and the use of an exponentially penalized distance for connecting disjoint sub-graphs. We showed that this strategy is effective in several artificial problems, but can be unstable for problems with high symmetry. To improve the stability of the metric we introduced the ISOMAP-all metric, an extension of ISOMAP-min using multiple connection between sub-graphs. Measuring distances on perturbed realization of the artificial datasets we verified the increased stability of ISOMAP-all. We also compared the clustering accuracy of both new metrics in association with 3 different clustering algorithms, showing similar (high) performance. A comparison with Spectral Clustering, a state of the art algorithm for clustering, confirms the validity of these encouraging results.

We plan to extend our study including real genomic data, datasets from proteomics, face recognition and others, also looking for efficient implementations (both in time and space) of our strategy. We also plan to analyze in depth the interesting properties of the Affinity propagation algorithm.

Acknowledgements

We acknowledge partial support for this project from ANPCyT (grant PICT 2003 11-15132).

References

1. R. Xu, D. Wunsch II: Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, **16:3**, (2005) 645–678.
2. Bottou L., Bengio Y.: Convergence properties of the K-means algorithms. In G. Tesauro, D. Touretzky,and T. Leen, editors, Advances in Neural Information Processing Systems, **volume 7**, The MIT Press, (1995), 585–592.

Dataset	Low noise	Medium noise	High noise
Two-arcs	100 ± 0	100 ± 0	99.97 ± 0.06
Three-spirals	90 ± 20	91 ± 17	93 ± 15
Three-rings	78 ± 13	70 ± 10	68 ± 8

Table 4. Clustering results for the Spectral Clustering algorithm on the three datasets and three noise levels analyzed in this work. The results are mean ($\pm$ standard deviation) accuracies on 100 runs of each dataset.

3. Kaufman L., Rousseeuw P. J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, (1990), Brussels, Belgium.
4. Frey, B. J., Dueck, D.: Clustering by passing messages between data points. Science, **315**,(2007) 972–976.
5. Roweis S., Saul L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, (2000), 2323–2326.
6. Belkin M., Niyogi P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In Dietterich T. G., Becker S., and Ghahramani Z., editors, Advances in Neural Information Processing Systems **14**, Cambridge, MA, MIT Press, (2002), 585–591.
7. Tenenbaum, J., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science, **290**,(2000), 2319–2323.
8. L. K. Saul and S. T. Roweis.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research **4**. (2003), 119–155.
9. Cormen, T. H., Leiserson, C. E., Rivest, R. L.: Introduction to Algorithms, first edition, MIT Press and McGraw-Hill, (1990), 558–565.
10. Cormen, T. H., Leiserson, C. E., Rivest, R. L., Stein C.: Introduction to Algorithms, Second Edition. MIT Press and McGraw-Hill, (2001), 595–601.
11. Cox T., Cox M.: Multidimensional Scaling, Chapman & Hall, London, (1994).
12. Ng A.,Jordan M., Weiss Y.:On spectral clustering: Analysis and an algorithm. In Dietterich T. G., Becker S., and Ghahramani Z., editors, Advances in Neural Information Processing Systems **14**, (2002).