Report Number 09/48

DifFUZZY: A fuzzy spectral clustering algorithm for complex data sets

by

Ornella Cominetti, Anastasios Matzavinos, Sandhya Samarasinghe, Don Kulasiri, Sijia Liu, Philip K. Maini, and Radek Erban

Oxford Centre for Collaborative Applied Mathematics
Mathematical Institute
24 - 29 St Giles'
Oxford
OX1 3LB
England

BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–7

DifFUZZY: A fuzzy spectral clustering algorithm for complex data sets

Ornella Cominetti [1,*], Anastasios Matzavinos [2], Sandhya Samarasinghe [3], Don Kulasiri [3], Sijia Liu [2], Philip K. Maini [1,4], and Radek Erban [5]

[1] Centre for Mathematical Biology, Mathematical Institute, University of Oxford, 24–29 St. Giles', Oxford, OX1 3LB, United Kingdom
[2] Department of Mathematics, Iowa State University, Ames, IA 50011, USA
[3] Centre for Advanced Computational Solutions (C-fACS), Lincoln University, P.O.Box 84, Christchurch, New Zealand
[4] Oxford Centre for Integrative Systems Biology, Department of Biochemistry, University of Oxford, South Parks Road, Oxford, OX1 3QU, United Kingdom
[5] Oxford Centre for Collaborative Applied Mathematics, Mathematical Institute, University of Oxford, 24–29 St. Giles', Oxford, OX1 3LB, United Kingdom

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Soft (fuzzy) clustering techniques are often used in the study of high-dimensional data sets, such as microarray and other high-throughput bioinformatics data. The most widely used method is the Fuzzy C-means algorithm (FCM), but it can present difficulties when dealing with some data sets.

Results: A spectral fuzzy clustering algorithm, DifFUZZY, applicable to a larger class of clustering problems than other fuzzy clustering algorithms is developed. Examples of data sets (synthetic and real) for which this method outperforms other frequently used algorithms are presented, including two benchmark biological data sets, a genetic expression data set and a data set that contains taxonomic measurements. This method is better than traditional fuzzy clustering algorithms at handling data sets that are "curved", elongated or those which contain clusters of different dispersion.

Availability: The algorithm has been implemented in Matlab and is available at `http://www.maths.ox.ac.uk/cmb/diffuzzy`.

Contact: cominetti@maths.ox.ac.uk

Supplementary information: Supplementary information is available at *Bioinformatics* online.

1 INTRODUCTION

The need to interpret and extract possible inferences from high-dimensional bioinformatic data has led over the past decades to the development of dimensionality reduction and data clustering techniques. One of the first studied data clustering methodologies is the K-means algorithm, which was introduced by Macqueen (1967) and is the prototypical example of a non-overlapping, hard (crisp) clustering approach (Gan *et al*., 2007). The applicability of the K-means algorithm, however, is limited by the requirement that the clusters to be identified should be well-separated and "convex-shaped" (such as those in Fig. 1(a)) which is often not the case in biological data. Two fundamentally distinct approaches have been proposed in the past to address these two restrictions.

Bezdek *et al*. (1984) proposed the Fuzzy C-means (FCM) algorithm as an alternative, soft clustering approach that generates fuzzy partitions for a given data set. In the case of FCM the clusters to be identified do not have to be well-separated, as the method assigns cluster membership probabilities to undecidable elements of the data set that cannot be readily assigned to a specific cluster. However, the method does not exploit the intrinsic geometry of non-convex clusters, and, as we demonstrate in this article, its performance is drastically reduced when applied to some data sets, for example those in Figs. 2(a) and 3(a). This behavior can also be observed in the case of the standard K-means algorithm (Ng *et al*., 2002). These algorithms have been very successful in a number of examples in very diverse areas (such as in image segmentation (Trivedi *et al*., 1986), analysis of genetic networks (Stuart *et al*., 2003), protein class prediction (Zhang *et al*., 1995), epidemiology (French *et al*., 2008), among many others), but here we also explore data sets for which their performance is poor.

To circumvent the above problems associated with the geometry of data sets, approaches based on spectral graph theory have been recently devised. However, these algorithms are generally hard clustering methods which do not allow data points to belong to more than one cluster at the same time. This limits their applicability in clustering genetic expression data, where alternative or superimposed modes of regulation of certain genes would not be identified using partitional methods (Dembele & Kastner, 2003). In this paper we present DifFUZZY, a fuzzy spectral clustering algorithm that is applicable to a larger class of clustering problems than the FCM algorithm (Bezdek *et al*., 1984). For data sets with "convex-shaped" clusters both approaches lead to similar results, but DifFUZZY can better handle clusters with a complex, nonlinear

*to whom correspondence should be addressed

geometric structure. Moroever, DifFUZZY does not require any prior information on the number of clusters.

The paper is organized as follows. In Section 2, we present the DifFUZZY algorithm and give an intuitive explanation of how it works. In Section 3 we start with a prototypical example of a data set which can be successfully clustered by FCM, and we show that DifFUZZY leads to consistent results. Subsequently, we introduce examples of data sets for which FCM fails to identify the correct clusters, whereas DifFUZZY succeeds. Then, we apply DifFUZZY to biological data sets, namely, the Iris taxonomic data set and cancer genetic expression data sets.

2 METHODS

DifFUZZY combines ideas of fuzzy clustering and spectral clustering in a way that uses the strength of each method and avoids their weaknesses[1]. The input of the algorithm is the data set in the form

$$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N \in \mathbb{R}^p \tag{1}$$

where N is the number of data points and p is their dimension, plus four parameters, one of which is external, M, and the rest are the internal and optional parameters γ_1, γ_2 and γ_3. M is an integer which represents the minimum number of data points in the clusters to be found. This parameter is necessary, since in most cases only a few data points do not constitute a cluster, but a set of soft data points or a set of outliers. There are three optional parameters: γ_1, γ_2 and γ_3 whose default values (0.3, 0.1 and 1, respectively) have been optimized and used successfully in all the data sets analyzed. A regular user can use these values with confidence. However, more advanced users can modify their values, with the intuitive explanation provided in Section 2.4.

DifFUZZY returns a number of clusters (C) and a set of membership values for each data point in each cluster. The membership value of data point $\mathbf{X}_i$ in the cluster c is denoted as $u_c(\mathbf{X}_i)$, and it goes from zero to one, where this latter case means that $\mathbf{X}_i$ is very likely a member of the cluster c, while the former case ($u_c(\mathbf{X}_i) \sim 0$) corresponds to the situation in which the point $\mathbf{X}_i$ is very likely not a member of the cluster c. The membership degrees of the i-th point, $i = 1, 2, \ldots, N$, sum to 1, that is

$$\sum_{c=1}^{C} u_c(\mathbf{X}_i) = 1. \tag{2}$$

DifFUZZY has been implemented in Matlab and can be downloaded from `http://www.maths.ox.ac.uk/cmb/diffuzzy`. The algorithm can be divided into three main steps, which will be explained in the following Sections 2.1–2.3. The reader who is not particularly interested in understanding the details of the algorithm can skip this part of the paper.

2.1 Identification of the core of clusters

To explain the first step of the algorithm, we define the auxiliary function $F(\sigma) : (0, \infty) \to \mathbb{N}$ as follows. Let $\sigma \in (0, \infty)$ be a positive number. We construct the so called σ-neighborhood graph where each node represents one data point from the data set (1), *i.e.* the σ-neighborhood graph has N nodes. The i-th node and j-th node will be connected by an edge if $||\mathbf{X}_i - \mathbf{X}_j|| < \sigma$, where $||\cdot||$ represents the Euclidean norm. Then $F(\sigma)$ is equal to the number of components of the σ-neighborhood graph which contain at least M vertices, where M is the mandatory parameter of DifFUZZY introduced above.

Fig. 1(b) shows an example of the plot of $F(\sigma)$, which was obtained using the data set presented in Fig. 1(a). We can see that $F(\sigma)$ begins from zero, and then increases to its maximum value, before settling back down to a value of 1. The final value will always be one, because the σ-neighborhood graph is fully connected for sufficiently large σ values, *i.e.* it only has one component.

DifFUZZY computes the number, C, of clusters as the maximum value of $F(\sigma)$, *i.e.*

$$C = \max_{\sigma \in (0,\infty)} F(\sigma).$$

For the example in Fig. 1(b), we have $C = 3$, which corresponds to the three clusters shown in the original data set in Fig. 1(a).

In Fig. 1(b) we see that there is an interval of values of σ for which $F(\sigma)$ reaches its maximum value C. As the next step DifFUZZY computes σ^*, which is defined as the minimum value of σ for which $F(\sigma)$ is equal to C. Then the σ^*-neighborhood graph is constructed. The components of this graph which contain at least M vertices will form the "cores" of the clusters to be identified. Each data point $\mathbf{X}_i$ which lies in the c-th core is assigned the membership values $u_c(\mathbf{X}_i) = 1$ and $u_j(\mathbf{X}_i) = 0$ for $j \neq c$, as this point fully belongs to the c-th cluster. Every such point will be called a hard point in what follows. The remaining points are called soft points. Since we already know the number of clusters C and the membership functions of hard points, it remains to assign a membership function to each soft point. This will be done in two steps. First we compute some auxiliary matrices in Section 2.2 and then we assign the membership values to soft points in Section 2.3.

2.2 Computation of auxiliary matrices W, D and P

To compute the auxiliary matrices we define a family of matrices $\widehat{W}(\beta)$ with entries

$$\widehat{w}_{i,j}(\beta) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are hard points in the same core cluster,} \\ \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{\beta}\right) & \text{otherwise,} \end{cases} \tag{3}$$

where β is a positive real number. We define the function $L(\beta) : (0, \infty) \to (0, \infty)$ to be the sum

$$L(\beta) = \sum_{i=1}^{N} \sum_{j=1}^{N} \widehat{w}_{i,j}(\beta). \tag{4}$$

The log-log plot of function $L(\beta)$ is shown in Fig. 1(c) for the data set given in Fig. 1(a). We can see that it has two well defined limits

$$\lim_{\beta \to 0} L(\beta) = N + \sum_{i=1}^{C} n_i \left(n_i - 1\right) \qquad \text{and} \qquad \lim_{\beta \to \infty} L(\beta) = N^2,$$

where n_i corresponds to the number of points in the i-th core cluster. As explained in Section 2.4, we are interested in finding the value of

[1] The formulation of both the FCM and standard spectral clustering algorithm is given in the Supplementary Information.

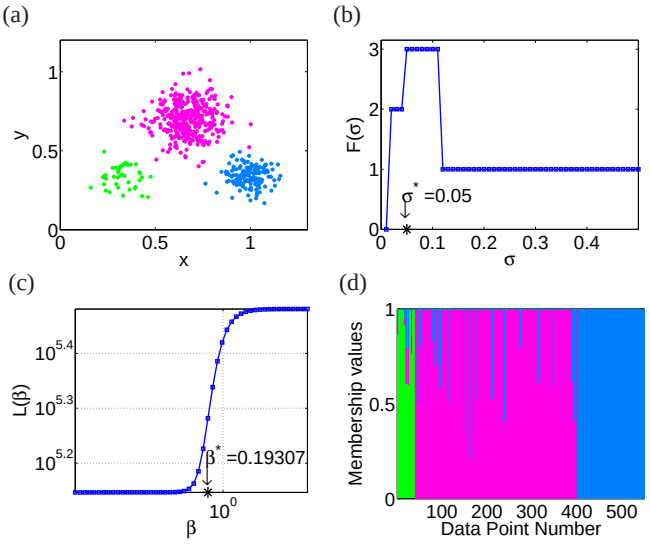

Fig. 1. (a) *"Globular clusters" data set.* (b) $F(\sigma)$ *for the data set in (a). For this data set we determined the number of clusters* C *to be* 3*, and* $\sigma^* = 0.05$*, for the parameter* $\mathrm{M} = 35$*.* (c) $L(\beta)$*, given by Eq.* (4) *plotted on a logarithmic scale, for the data set in (a).* $\beta^* = 0.068665$ *was obtained using Eq. (5).* (d) *DifFUZZY membership values for this data set. Each data point is represented by a bar of total height equal to* 1 *(from Eq.* 2*).* ($\mathrm{M} = 35$)*. Color code: green, red and blue correspond to the membership value of the data points in the three clusters, with the corresponding color code as in (a). This representation will be used in Figs.* 2 *and* 3*.*

β which corresponds to an intermediate value of $L(\beta)$. DifFUZZY does this by finding β^* which satisfies the relation

$$L(\beta^*) = (1-\gamma_1)\left(N + \sum_{i=1}^{C} n_i\,(n_i-1)\right) + \gamma_1 N^2, \qquad (5)$$

where $\gamma_1 \in (0,1)$ is an internal parameter of the method. Its default value is 0.3. Then the auxiliary matrices are defined as follows. We put

$$W = \widehat{W}(\beta^*). \qquad (6)$$

The matrix D is defined as a diagonal matrix with diagonal elements

$$D_{i,i} = \sum_{j=1}^{N} w_{i,j}\ , \quad i = 1, 2, \ldots, N, \qquad (7)$$

where $w_{i,j}$ are the entries of matrix W. Finally, the matrix P is defined as

$$P = I + [W - D]\,\frac{\gamma_2}{\max\limits_{i=1,\ldots N} D_{i,i}}, \qquad (8)$$

where $I \in \mathbb{R}^{N\times N}$ is the identity matrix and γ_2 is an internal parameter of DifFUZZY. Its default value is 0.1. The definition of the matrices W, D and P can be intuitively understood in terms of diffusion processes on graphs, as explained in Section 2.4.

2.3 The membership values of soft data points

Let $\mathbf{X}_s$ be a soft data point. To assign its membership value $u_c(\mathbf{X}_s)$ in cluster $c \in \{1, 2, \ldots, C\}$, we first find the hard point in the c-th core which is closest (in Euclidean distance) to $\mathbf{X}_s$. This point will be denoted as $\mathbf{X}_n$ in what follows. Using the matrix W defined by Eq. (6), DifFUZZY constructs a new matrix $\overline{W}$ which is equal to the original matrix W, with the s-th row replaced by the n-th row and the s-th column replaced by the n-th column. Using $\overline{W}$ instead of W, matrices $\overline{D}$ and $\overline{P}$ are computed by (7) and (8), respectively. DifFUZZY also computes an auxiliary integer parameter α by

$$\alpha = \left\lfloor \frac{\gamma_3}{|\log \lambda_2|} \right\rfloor,$$

where λ_2 corresponds to the second (largest) eigenvalue of P and $\lfloor\cdot\rfloor$ denotes the integer part.
Next, we compute the diffusion distance between the soft point $\mathbf{X}_s$ and the c-th cluster by

$$\mathrm{dist}(\mathbf{X}_s, c) = \left|\left| P^{\alpha}\mathbf{e} - \overline{P}^{\,\alpha}\mathbf{e} \right|\right|, \qquad (9)$$

where $\mathbf{e}(j) = 1$ if $j = s$, and $\mathbf{e}(j) = 0$ otherwise. Finally the membership value of the soft point $\mathbf{X}_s$ in the c-th cluster, $u_c(\mathbf{X}_s)$, is determined with the following formula:

$$u_c(\mathbf{X}_s) = \frac{\mathrm{dist}(\mathbf{X}_s, c)^{-1}}{\sum\limits_{l=1}^{C} \mathrm{dist}(\mathbf{X}_s, l)^{-1}}. \qquad (10)$$

This procedure is applied to every soft data point $\mathbf{X}_s$ and every cluster $c \in \{1, 2, \ldots, C\}$.

2.4 Geometric and graph interpretation of DifFUZZY

In this section, we provide an intuitive geometric explanation of the ideas behind the DifFUZZY algorithm. The matrix P can be thought of as a transition matrix whose rows all sum to 1, and whose entry $P_{i,j}$ corresponds to the probability of jumping from the node (data point) i to the node j in one time step. The j-th component of the vector $P^{\alpha}\mathbf{e}$, which is used in (9), is the probability of a random walk ending up in the j-th node, $j = 1, 2, \ldots, N$, after α time steps, provided that it starts in the s-th node.
In this geometric interpretation we can give an intuitive meaning to the auxiliary parameters γ_1, γ_2 and γ_3. The parameter $\gamma_1 \in (0,1)$ is related to the time scale of this random walk. $\gamma_1 \sim 1$ corresponds to the case where all the nodes are highly connected, and therefore the diffusion will occur instantaneously, whereas for values of $\gamma_1 \sim 0$, there will be almost no diffusion between cluster cores. Therefore, we are interested in an intermediate point, where there is enough time to diffuse, but where equilibrium has not yet been reached. The parameter $\gamma_2 \in (0,1)$ ensures that none of the entries of the transition matrix P are negative, which is important, since they represent transition probabilities. It can be interpreted as the length of the time step of the random walk on the graph. For very small values of γ_2 we have $P \sim I$, for which the probabilities of transition between different data points is close to zero, therefore there will not be any diffusion during one time step.
The parameter $\gamma_3 \in (0, \infty)$ is the number of time steps the random walk is going to be run or propagated, capturing information of higher order neighborhood structure (Lafon & Lee, 2006). Small values of γ_3 give us a few time steps, whereas large values of γ_3 give us a large number of time steps. In the first situation not much diffusion has taken place, whereas in the latter case, when the random walk is propagated a very large number of times, the diffusion process is near to reaching the equilibrium.

The matrix $\overline{P}$ is used to represent a different diffusion process, an equivalent one to the first random walk, but over a new graph, where the data point $\mathbf{X}_s$ has been moved to the position of the data point $\mathbf{X}_n$. This matrix then corresponds to the transition matrix for this auxiliary graph.

3 RESULTS

In Section 3.1, we present three computer generated test data sets, designed to illustrate the strengths and weaknesses of FCM. In all three cases we show that DifFUZZY gives the desired result. Then, in Section 3.2 we apply DifFUZZY to data sets obtained from biological experiments.

3.1 Synthetic test data sets

The output of DifFUZZY is a number of clusters (C) and for each data point a set of C numbers that represent the degree of membership in each cluster. The membership value of point $\mathbf{X}_i$, $i = 1, 2, \ldots, N$, in the c-th cluster, $c = 1, 2, \ldots, C$, is denoted by $u_c(\mathbf{X}_i)$. The degree of membership is a number between 0 and 1, where the values close to 1 correspond to points that are very likely to belong to that cluster. The sum of the membership values of a data point in all clusters is always one (see Eq. (2)). In particular, for a given point, there can be only one cluster for which the membership value is close to 1, *i.e.* the point can belong to only one cluster with high certainty.

A prototypical cluster data set in two-dimensional space is shown in Fig. 1(a). Every point is described by two coordinates. We can see that the data points form three well defined clusters which are colored in green, red, and blue. Any good soft or hard clustering technique should identify these clusters. However, when we introduce intermediate data points, the clusters are less well defined, closer together, and some hard clustering techniques may have difficulty in separating the clusters. FCM can successfully handle this problem (see the Supplementary Information). The same is true for DifFUZZY. In Fig. 1(d) we present the results obtained by applying DifFUZZY to the data set in Fig. 1(a). We plot the membership values for all data points. This is a prototypical example of the type of problem for which FCM works and DifFUZZY gives comparable results. Further examples are shown in the Supplementary Information.

A classical example where the K-means algorithm fails (Filippone *et al.*, 2008) is shown in Fig. 2(a). This is a two-dimensional data set formed by three concentric rings. Using DifFUZZY we identify each ring as a separate cluster, as can be seen in Fig. 2(a)–(b). Since fuzzy clustering assigns each point to a vector of membership values, it is more challenging to visualize the results. One option is to plot the membership values as shown in Fig. 2(b). A rough idea of the behavior of the algorithm can also be obtained by making what we call a HCT-plot ("Hard Clusters by Threshold") defined as follows: a data point is assigned to a cluster only if its membership value for that cluster is higher than an arbitrary threshold z. All the other data points are unassigned, and consequently plotted in black. Such a plot is shown in Fig.2(a) for $z = 0.9$. HCT-plots do not show the complete result from applying a given fuzzy clustering method to a data set, since they contain less information than the complete result (all the membership values), and the HCT-plots depend on the threshold. However, it is illustrative to include them to clearly show how the results of different algorithms compare. The membership values obtained with FCM are plotted in Fig. 2(d). In Fig. 2(c) we present the corresponding HCT-plot with a threshold value of 0.9. Comparing Figs. 2(a)–(b) with 2(c)–(d) we clearly see that DifFUZZY gives better results than FCM.

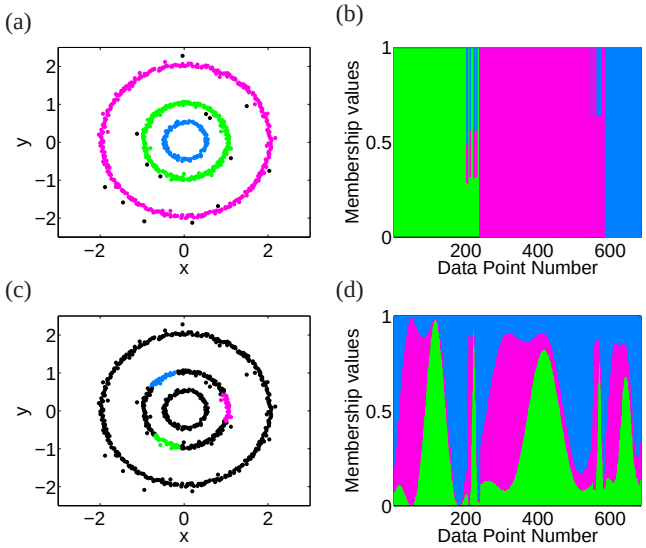

Fig. 2. *"Concentric rings" test data set.* (a) *DifFUZZY HCT-plot,* $z = 0.9$. (M $= 90$). (b) *DifFUZZY membership values.* (c) *FCM HCT-plot,* $z = 0.9$. (d) *FCM membership values. Color code for (b) and (d) as in Fig.* 1*(b).*

Another data set where K-means algorithms fail is presented in Fig. 3(a). This two-dimensional data set contains two elongated clusters, one in a diagonal orientation and the other a cross-shaped cluster. The results of DifFUZZY and FCM applied over this data set are summarized in the membership value plots in Figs. 3(b) and 3(d), respectively. DifFUZZY can separate the clusters remarkably well, as is clear from Fig. 3(a). For this data set FCM can not separate the clusters, cutting the left cluster (blue) in two parts as can be seen in the HCT-plot shown in Fig. 3(c), using the threshold value $z = 0.9$. If we compare the membership values given by FCM (Fig. 3(d)) to the one by DifFUZZY in Fig. 3(b), which basically corresponds to the desired membership values of the data points, we see the wrong identification of the data points numbered around 550–700, which in the case of FCM have been given a higher membership in the green cluster than in the cluster where they originally belong (the blue one).

3.2 Biological data sets

DifFUZZY was tested in two widely used biological data sets: Iris (Fisher, 1936) and Leukemia (Golub *et al.*, 1999). In the Supplementary Information we include results of the application of DifFUZZY to more biological data sets.

Iris dat aset: This is a benchmark data set in pattern recognition analysis, freely available at the UCI Machine Learning Repository (Asuncion & Newman, 2007). It contains three clusters (types of Iris plants: Iris Setosa, Iris Versicolor and Iris Virginica) of 50 data points each, of 4 dimensions (features): sepal length, sepal width, petal length and petal width. The class Iris Setosa is linearly

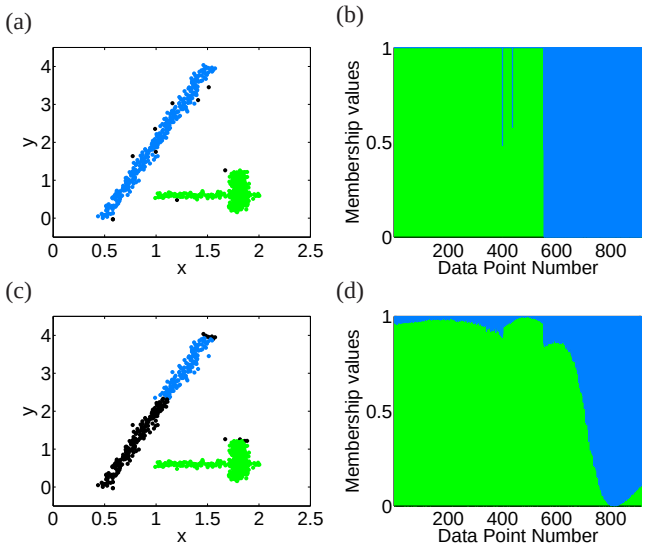

Fig. 3. *"Elongated clusters" test data set.* (a) *DifFUZZY HCT-plot,* $z = 0.9$. (M $= 150$). (b) *DifFUZZY membership values.* (M $= 150$). (c) *FCM HCT-plot,* $z = 0.9$. (d) *FCM membership values. Color code for (b) and (d) as in Fig.* 1*(b).*

separable from the other two, which are not linearly separable in their original clusters (Fisher, 1936).

We show the results of applying DifFUZZY and FCM over the Iris data set in the form of ROC (receiver operating characteristic) curves which, in machine learning, correspond to the representation of the fraction of true positive classification (TPR) vs. the rate of false positive assignments (FPR) (Fawcett, 2006). Each data point in the curve represents a pair of values (FPR, TPR) obtained for a given threshold z. The precise definitions of both TPR and FPR are given in the Supplementary Information. A perfect clustering method would give a curve that passes through the upper left corner, presenting a 100% true positive rate for a 0% false positive rate, like the one obtained for DifFUZZY (using the parameter M=15) and FCM in Fig. 4(a), whereas if true positive and false positives are equally likely, the curve would be the diagonal line TPR=FPR. In the Supplementary Information we include further information regarding how to compute a ROC curve from the membership values given by a fuzzy clustering method.

Figs. 4(a)–(c) show the ROC curves for the Iris Setosa, Iris Versicolor and Iris Virginica data, respectively. The three plots show very good clustering using DifFUZZY. For the Iris Setosa data, the DifFUZZY and FCM ROC curves correspond to perfect classifications, with both curves going through the (0,1) corner; both methods classify all those points correctly, and do not assign other points to that cluster (zero false positives), but for the Iris Versicolor and Iris Virginica data, DifFUZZY performs better than FCM, since its curves pass closer to the upper left corner. Additional material regarding the performance of DifFUZZY over the Iris data set for different values of the parameter M can be found in the Supplementary Information, such as plots of the membership values and plots of the core clusters. There we show that data sets with overlapping core clusters can be well clustered by scaling each feature by a given weight, or by choosing a smaller value of M.

Genetic expression data set: We tested DifFUZZY on the publicly available Leukemia data set (Golub *et al*., 1999), which contains genetic expression data from patients diagnosed with either of two different types of leukemia: acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) (Tan *et al*., 2004). This data set, composed of 7129 genes, was obtained from an Affimetrix high-density oligonucleotide microarray. The original data are divided into two sets, a set for training and a test set. Since our method is unsupervised we merged both sets obtaining a set with data from 72 patients: 25 with AML, 47 with ALL.

Before testing our clustering method on the Leukemia data set we pre-processed the data as done by Tan *et al*. (2004). The gene selection procedure consisted of selecting the Q genes with the highest squared correlation coefficient sums (Tan *et al*., 2004), where Q corresponds to the number of genes to be selected, which for the case of this data set was set to be 70.

Fig. 4(d) shows that DifFUZZY performs better than FCM when clustering the Leukemia data set, since for every point of the curve, at the same false positive rate DifFUZZY presents a higher or equal true positive rate than FCM. In the Supplementary Information we provide the plots of the membership values for all the data points. Through this example we are able to show that our method can also handle high dimensional microarray data and it can be successfully used for multi-class cancer classification tasks.

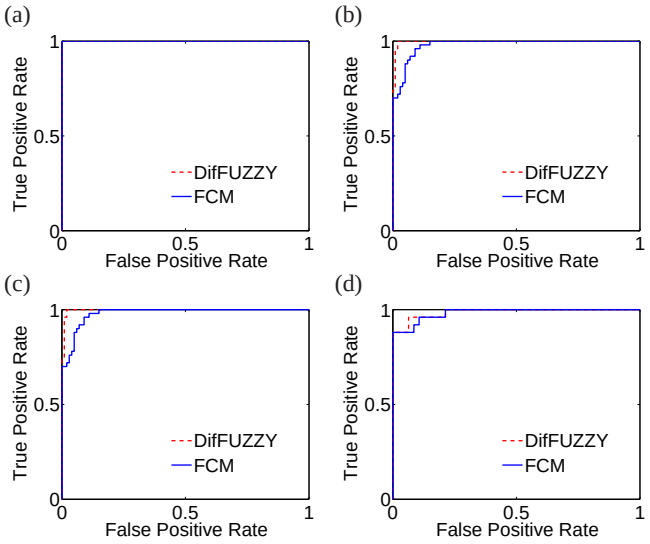

Fig. 4. *DifFUZZY and FCM ROC curves for* (a) *the Iris Setosa data* (M $=$ 15), (b) *the Iris Versicolor data* (M $= 15$), (c) *the Iris Virginica data* (M $=$ 15) and (d) *the Leukemia data set* (M $= 11$).

4 DISCUSSION

In this paper and the Supplementary Information we showed that the fuzzy spectral clustering method DifFUZZY performs well in a number of data sets, with sizes ranging from tens to hundreds of data points of dimensions as high as hundreds. This includes microarray data, where a typical size of a data set is dozens or hundreds (number of samples, conditions, or patients in medical applications) and dimension is hundreds or thousands (number of genes on the chip) (Quackenbush, 2001). It is worth noting that the dimension p of

the data points in Eq. (1). is not a bottleneck for this method, since it is only used once when computing the pairwise distances. The dimension of matrices (*i.e.* the computational intensity) is determined by the number N of data points, which is often smaller than the value p.

One of the issues that we briefly discussed was the pre-processing of data. This is crucial for some clustering applications. Noisiness of the data and the normalization used on a given data set can have a high impact on the results of clustering procedures (Kim *et al*., 2006; Karthikeyani & Thangavel, 2009). What type of normalization to use will depend on the data themselves, and when additional information on the data set is available it should be used in order to improve the quality of the data to be input in the algorithm. In the case of genetic expression data sets (such as the one analyzed in Fig. 4(d)), different steps of preprocessing commonly used are filtering, thresholding, log normalization and gene selection (Tan *et al*., 2004). The latter is done in order to reduce the dimensionality of the feature space, by discarding redundant information. Another option is to weight the different features in order to make the dimensions of the different features comparable or to augment the influence of features which carry more or better information about the data structure. The use of independent feature scaling has been described in the context of similarity matrices in Erban *et al*. (2007), where a single value of the parameter β in Eq. (3) is not necessarily appropriate for all the components (variables), given that these may vary over different orders of magnitudes. Two examples of natural weights that can be used are giving the same weight (equal importance) to the absolute values of each feature, or to rescale each variable in order for them to have the same minimum and maximum values.

A mathematical analysis of the DifFUZZY algorithm will be done in a future publication. As briefly addressed in Section 2.4, it involves an understanding of the mixing time (see, e.g., Levin *et al*. (2009)) of the random walk defined in (8) for specific types of graphs. In particular, for a given data set, the performance of the developed method relies on the parameter α determining the diffusion distance in (9). Computational experimentation with test data sets reveals that the optimal choice of α tends to be robust for a broad variety of data set geometries. In order to understand this phenomenon and the underlying mechanics of DifFUZZY, current work in progress focuses on investigating mathematically the asymptotic properties of the random walk in (8) over classes of graphs characterized by specific topologies. In this context, the transition matrix P used in (9), which can be written as $P = I + (W - D)\Delta t$, is essentially a first-order approximation to the heat kernel of the graph associated with $L = D - W$. In particular, for every $\Delta t \geq 0$, the heat kernel $H_{\Delta t}$ of a graph G with graph Laplacian L is defined to be the matrix $H_{\Delta t} = e^{-\Delta t L} = I - \Delta t L + \Delta t^2 L^2/2 - \ldots$. The importance of $H_{\Delta t}$ is that it defines an operator semigroup, describing fundamental solutions of the spatially discretized heat equation $u_t = (W - D)u$.

Heat kernels are powerful tools for defining and investigating random walks on graphs, and they provide a connection between the structure of the graph, as encoded in the graph Laplacian, and the asymptotic behavior of the corresponding random walk (Chung, 1997). Work in progress exploits these connections in order to analyze the optimal performance of DifFUZZY for data sets exhibiting specific geometries. We are also extending the applications of DifFUZZY to a variety of clustering problems emerging in bioinformatics and image analysis applications. Fuzzy clustering methods have traditionally been used for image segmentation (Bezdek *et al*., 1997; Chen & Zhang, 2004; Tziakos *et al*., 2009), especially in the field of medical imaging. Bezdek *et al*. (1997) discuss the advantages of fuzzy clustering approaches applied to the specific case of segmenting magnetic resonance images (MRIs). Several variations of the FCM method are commonly employed in this context, and recent research has been focused on images that are characterized by a non-Euclidean structure of the corresponding feature space (Chen & Zhang, 2004). The clustering methodology proposed here is specifically designed to handle non-Euclidean data sets associated with a manifold structure, as it seamlessly integrates spectral clustering approaches with the evaluation of cluster membership functions in a fuzzy clustering context.

ACKNOWLEDGEMENT

Funding: This publication is based on work (RE, AM) supported by Award No. KUK-C1-013-04, made by King Abdullah University of Science and Technology (KAUST) and the Clarendon Fund through the Systems Biology Doctoral Training Centre (O.C.). PKM was partially supported by a Royal Society-Wolfson Research Merit Award. RE would also like to thank Somerville College, University of Oxford for Fulford Junior Research Fellowship. The research of S.L. is supported in part by an Alberta Wolfe Research Fellowship from the Iowa State University Mathematics department.

REFERENCES

Asuncion, A. and Newman, D. (2007) UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html], School of Information and Computer Science, California Univ.

Belkin, M. and Niyogi, P. (2003) Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation*, **15**, 1373–1396.

Bezdek, J., Ehrlich, R. and Full, W. (1984) FCM: The fuzzy c-means clustering algorithm, *Computers & Geosciences*, **10**, 191–203.

Bezdek, J., Hall, L.O., Clark, M.C., Goldgof, D.B. and Clarke, L.P. (1997) Medical image analysis with fuzzy models, *Statistical Methods in Medical Research*, **6**, 191–214.

Chen, S. and Zhang, D. (2004) Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure, *IEEE Transactions on Systems, Man, and Cybernetics – Part B*, **34**, 1907–1916.

Chung, F. (1997) Spectral Graph Theory, *American Mathematical Society*, **92**.

Cao, F., Delon, J., Desolneux, A., Musé, P. and Sur, F. (2007) A Unified Framework for Detecting Groups and Application to Shape Recognition, *Journal of Mathematical Imaging and Vision*, **27**, 2.

Cao, F., Lisani, J., Morel, J., Musé, P. and Sur, F. (2008) A Theory of Shape Identification, *Springer*.

Coifman, R. and Lafon, S. (2006) Diffusion maps, *Applied and Computational Harmonic Analysis*, **21**, 5–30.

Dembele, D. and Kastner, P. (2003) Fuzzy C-means Method for Clustering Microarray Data, *Bioinformatics*, **19**, 973–980.

Desolneux, A., Moisan, L. and Morel, J. (2008) From Gestalt Theory to Image Analysis: A Probabilistic Approach, *Springer*.

Erban, R., Frewen, T., Wang, X., Elston, T., Coifman, R., Nadler B. and Kevrekidis, I. (2007) Variable-free exploration of stochastic models: a gene regulatory network example, *Journal of Chemical Physics*, **126**, 155103.

Fawcett, T. (2006) An introduction to ROC analysis, *Pattern Recognition Letters*, **27**, 861–874.

Filippone, M., Camastra F., Masulli F. and Rovetta, S. (2007) A survey of kernel and spectral methods for clustering, *Pattern Recognition*, **41**, 176-190.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, *Annual Eugenics*, **7**, Part II, 179–188.

French, S., Rosenberg, M. and Knuiman, M. (2008) The clustering of health risk behaviours in a Western Australian adult population. *Health Promotion Journal of Australia*, **19**, 203–209.

Gan, G., Ma, C. and Wu, J. (2007) Data Clustering: Theory, Algorithms, and Applications, *ASA-SIAM Series on Statistics and Applied Probability*, **20**.

Golub T., Slonim, D., Tamayo P, Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing J.R., Caligiuri, M., Bloomfield, C. and Lander, E. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531–537.

Jimenez, R. (2008) Fuzzy spectral clustering for identification of rock discontinuity sets, *Rock Mechanics and Rock Engineering*, **41**, 929–939.

Karthikeyani, N. and Thangavel, K. (2009) Impact of Normalization in Distributed K-Means Clusterings, *International Journal of Soft Computing and Bioinformatics*, **4** 168–172.

Kim, S., Lee, J. and Bae J. (2006) Effect of data normalization on fuzzy clustering of DNA microarray data, *BMC Bioinformatics*, **17**, 134.

Lafon, S. and Lee, A. (2006) Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning and Data Set Parameterization, *IEEE Pattern Analysis and Machine Intelligence*.

Levin, D., Peres, Y. and Wilmer, E. (2009) Markov Chains and Mixing Times, *American Mathematical Society*.

Liao, C., Lu, K., Baym, M., Singh, R., and Berger, B. (2009) IsoRankN: spectral methods for global alignment of multiple protein networks, *Bioinformatics*, **25**, i253–i258.

von Luxburg, U. (2007) A tutorial on spectral clustering, *In (MPITechnical Reports No. 149). Tubingen: Max Planck Institute for Biological Cybernetics*, **17**, 416.

Macqueen, J. (1967) Some methods for classification and analysis of multivariate observations, In *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, **1**, pages 281–297, University of California Press.

Nadler, B., Lafon, S., Coifman R. and Kevrekidis, I. (2006) Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems, *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets*, **21**.

Ng, A., Jordan, M. and Weiss, Y. (2002) On spectral clustering: Analysis and an algorithm, *Advances in neural information processing systems*, **14**.

Quackenbush, J. (2001) Computational analysis of microarray data, *Nature Reviews Genetics*, 2–418.

Shi, J. and Malik, J. (2000) Normalized cut and image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 888–905.

Singer, A., Erban, R., Kevrekidis I. and Coifman, R. (2009) Detecting the slow manifold by anisotropic diffusion maps, submitted to the *Proceedings of the National Academy of Sciences*, **106**(8), 16090–16095.

Singh, R., Xu, J. and Berger, B. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection, *Proceedings of the National Academy of Sciences*, **105**, 12763–12768.

Stuart, J., Segal, E., Koller, D. and Kim, S. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Tan, Y., Shi, L., Tong, W., Gene Hwang, G. and Wang C. (2004) Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models, *Computational Biology and Chemistry*, **28**, 235–244.

Tsao, J. and Lauterbur, P. (1998) Generalized clustering-based image registration for multi-modality images, *Engineering in Medicine and Biology Society, 1998. Proceedings of the 20th Annual International Conference of the IEEE*, **2**, 667–670.

Trivedi, M. and Bezdek, J. (1986) Low-level segmentation of aerial images with fuzzy clustering, *IEEE Transactions on Systems, Man, and Cybernetics*, **16**, 589–598.

Tziakos, I., Theoharatos, C., Laskaris, N. and Economou, G. (2009) Color image segmentation using Laplacian eigenmaps, *Journal of Electronic Imaging*, **18**, 023004.

Weiss, Y. (1999) Segmentation using eigenvectors: A unifying view, In *International Conference on Computer Vision*.

Zhang, C., Chou, K. and Maggiora, G. (1995) Predicting protein structural classes from amino acid composition: application of fuzzy clustering. *Protein Engineering, Design and Selection*, **8**, 425–435.

Zhang, A. and Cho, Y. (2008) Discovering frequent patterns of functional associations in protein interaction networks for function prediction, *2008 IEEE International Conference on Bioinformatics and Biomedicine*, 59–65.

RECENT REPORTS

23/09	Morphological instability of a nonequilibrium icecolloid interface	Peppin Majumdar Wettlaufer
24/09	The effect of polar lipids on tear film dynamics	Aydemir Breward Witelski
25/09	Preconditioning for active set and projected gradient methods as semi-smooth Newton methods for PDE-constrained optimization with control constraints	Stoll Wathen
26/09	Functional differential equations arising in cell-growth	Wake Begg
27/09	A Cell Growth Model Revisited	Derfel van Brunt Wake
28/09	Quasi-steady state reduction of molecular motor-based models of directed intermittent search	Newby Bressloff
29/09	All-at-once preconditioning in PDE-constrained optimization	Rees Stoll Wathen
30/09	An hp-Local Discontinuous Galerkin method for Parabolic Integro-Differential Equations	Pani Yadav
31/09	Stochastic neural field theory and the system-size expansion	Bressloff
32/09	A Hamiltonian Krylov-Schur-type method based on the symplectic Lanczos process	Benner Faßbender Stoll
33/09	Nematic liquid crystals : from Maier-Saupe to a continuum theory	Ball Majumdar
34/09	Tangent unit-vector fields: nonabelian homotopy invariants and the Dirichlet energy	Majumdar Robbins Zyskin
35/09	A metabolite-sensitive, thermodynamically-constrained model of cardiac cross-bridge cycling: Implications for force development during ischemia	Tran Smith Loiselle Crampin
36/09	Modelling bacterial behaviour close to a no-slip plane boundary: the influence of bacterial geometry	Shum Gaffney Smith
37/09	Optimal L2-error estimates for the semidiscrete Galerkin approximation to a second order linear parabolic initial and boundary value problem with nonsmooth initial data	Goswami Pani

38/09	Optimal L2 estimates for semidiscrete Galerkin methods for parabolic integro-differential equations with nonsmooth data	Goswami Pani Yadav
39/09	Spatially structured oscillations in a two-dimensional excitatory neuronal network with synaptic depression	Kilpatrick Bressloff
40/09	Stationary bumps in a piecewise smooth neural field model with synaptic depression	Kilpatrick Bressloff
41/09	Homogenization for advection-diffusion in a perforated domain	Haynes Hoang Norris Zygalakis
42/09	Fast stochastic simulation of biochemical reaction systems by alternative formulations of the Chemical Langevin Equation	Melykuti Burrage Zygalakis
43/09	Pseudoreplication invalidates the results of many neuroscientific studies	Lazic
44/09	Cardiac cell modelling: Observations from the heart of the cardiac physiome project	Finka *et al.*
45/09	A Hybrid Radial Basis Function - Pseudospectral Method for Thermal Convection in a 3-D Spherical Shell	Wright Flyer
46/09	Refining self-propelled particle models for collective behaviour	Yates Baker Erban Maini
47/09	Stochastic Partial Differential Equations as priors in ensemble methods for solving inverse problems	Potsepaev Farmer Aziz

Copies of these, and any other OCCAM reports can be obtained from:

Oxford Centre for Collaborative Applied Mathematics
Mathematical Institute
24 - 29 St Giles'
Oxford
OX1 3LB
England
www.maths.ox.ac.uk/occam