Spherical Wards clustering and generalized Voronoi diagrams

Marek Śmieja
Faculty of Mathematics and Computer Science
Jagiellonian University
Lojasiewicza 6, 30-348 Krakow, Poland
Email: marek.smieja@ii.uj.edu.pl

Jacek Tabor
Faculty of Mathematics and Computer Science
Jagiellonian University
Lojasiewicza 6, 30-348 Krakow, Poland
Email: jacek.tabor@ii.uj.edu.pl

Abstract—**Gaussian mixture model is very useful in many practical problems. Nevertheless, it cannot be directly generalized to non Euclidean spaces. To overcome this problem we present a spherical Gaussian-based clustering approach for partitioning data sets with respect to arbitrary dissimilarity measure. The proposed method is a combination of spherical Cross-Entropy Clustering with a generalized Wards approach. The algorithm finds the optimal number of clusters by automatically removing groups which carry no information. Moreover, it is scale invariant and allows for forming of spherically-shaped clusters of arbitrary sizes. In order to graphically represent and interpret the results the notion of Voronoi diagram was generalized to non Euclidean spaces and applied for introduced clustering method.**

I. INTRODUCTION

Distribution-based clustering, as Gaussian mixture model (GMM), has been proven to be very useful in many practical problems [1]. This technique has been widely applied in object detection [2], learning and modeling [3], feature selection [4] or classification [5]. The constructed groups are described by optimally fitted probability distributions. Nevertheless, this kind of methods is limited for the case of Euclidean spaces and the clustering of data with respect to Gaussian-like probability distributions in arbitrary data spaces where only distance or (dis)similarity measure is provided still remains a challenge.

In this paper we show how to partially overcome this problem and propose a spherical Wards clustering (SWARDS) which divides data sets with respect to arbitrary dissimilarity measure into groups described by spherical Gaussian-like distributions. Figure 1 shows the relationship between SWARDS and related methods. Moreover, we extend the notion of Voronoi diagram to the case of arbitrary criterion function in non Euclidean spaces and apply it for SWARDS clustering.

Introduced method permits an informal interpretation of the notion of spherical Gaussian probability distribution in non Euclidean space. The algorithm is capable of discovering spherically-shaped groups of arbitrary sizes (see Example 5.2). Moreover the clustering results are invariant on the scaling of data (see Example 5.1). In fact, data sets with unbalanced groups appear very often in practice, e.g in chemoinformatics where finding of chemical compounds acting on specific disease is rare [6], [7] or in Natural Language Processing where the numbers of documents that belong to particular domains are different [8]. Our method can be successfully applied in discovering of populations districts in biological systems modeled by a random walk procedure (see Examples
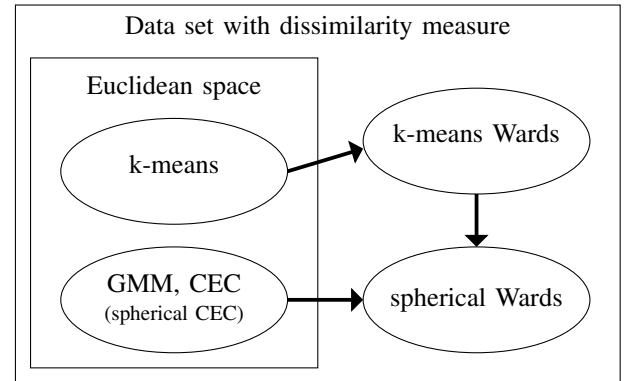

Fig. 1. **Spherical Wards clustering.** The relationship of our method with GMM, k-means and Wards clustering.

5.5, 5.6). The method is easy to implement and works in a comparable time to k-means version adapted to non Euclidean spaces [9]. Moreover, our algorithm automatically finds the resultant number of groups by reducing unnecessary clusters on-line. Voronoi diagrams for SWARDS, k-means and their kernelized versions for a mouse-like set with non Euclidean distance function are presented in Figure 2.

Proposed SWARDS method is a combination of spherical variant of Cross-Entropy Clustering (CEC) [10] with the generalized Wards approach [9], [11]. Generally, spherical CEC describes clusters by optimally fitted spherical Gaussian distributions while Wards method allows for its adaptation to non Euclidean case. Spherical CEC performs a clustering by optimizing a cross-entropy criterion function (4). Its form is very flexible since it is based on the within clusters sums of squares, the cardinalities of clusters and the dimension of space.

Applied Wards approach allows for a generalization of the notion of within cluster sum of squares for the case of any dissimilarity measure [9], [11]. The key lies in the observation that this quantity can be rewritten in Euclidean space without the use of a mean m_Y of a cluster Y in the form:

$$\sum_{y \in Y} d^2(y, m_Y) = \frac{1}{2|Y|} \sum_{y,z \in Y} d^2(y, z).$$

On the other hand, note that a dimension in arbitrary space does not have to be defined. Therefore, to adapt spherical CEC

(a) Wards k-means clustering with $k = 4$.

(b) sWARDS clustering started with 10 initial clusters.

(c) Wards k-means with RBF dissimilarity function and $k = 4$.

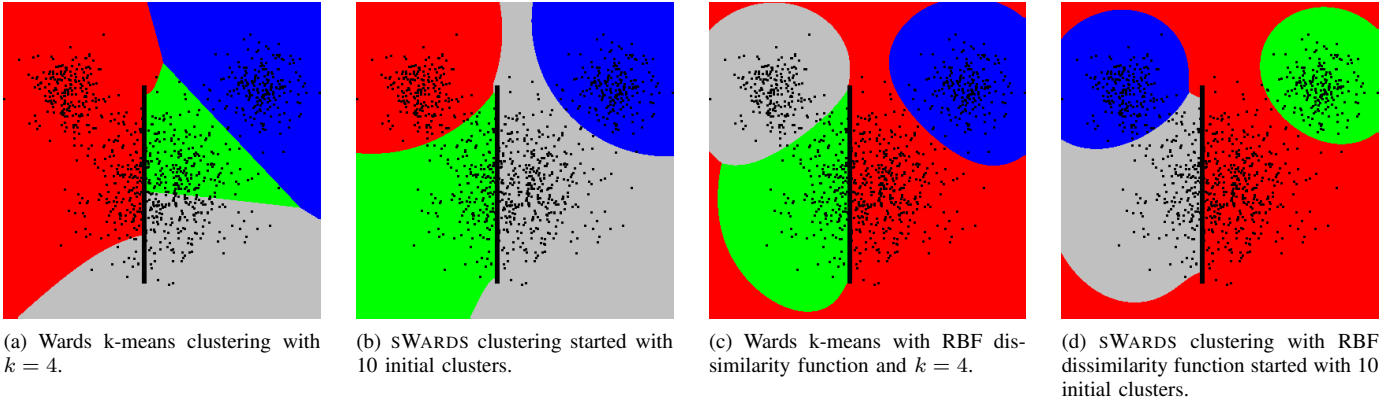(d) sWARDS clustering with RBF dissimilarity function started with 10 initial clusters.

Fig. 2. **Voronoi diagrams.** The Voronoi diagrams of introduced sWARDS compared with Wards k-means on mouse-like set with barriers with two types of similarity measures: Euclidean and RBF similarity. The barrier changes the distance between elements. The distance between elements located on the opposite sides of barrier is calculated as a shortest path which does not cross the barrier. Observe that despite the barrier sWARDS method discovered "mouse ears" as a spherical clusters while "mouse head" was divided into two smaller groups. The Wards k-means results do not have so intuitive explanation. Kernelized versions of both algorithms gave the satisfactory effects, but the main difficulty lies in finding the appropriate values of RBF parameter. An important thing is that the introduce sWARDS technique produces comparable partition without the need of parameters tuning.

criterion function to general case we recommend to estimate its value from data with use of Maximum Likelihood Estimator of intrinsic dimension [12], [13].

To graphically represent and interpret the results of clustering the notion of Voronoi diagram is widely applied. Its construction requires the answer for the question: to which cluster we should associate an arbitrary unclustered point? In the case of classical k-means the answer is simple: we assign the point to the cluster with the nearest center. In the Wards method we replace it by a generalization of distance of point x from the center of cluster Y given by [9]

$$d^2(x; Y) := \frac{1}{|Y|} \sum_{y \in Y} d^2(x, y) - \frac{1}{|Y|} ss(Y). \qquad (1)$$

In our work we calculate the analogue of above formula (1) for the case of sWARDS criterion function (5) (see (8) for precise formula and Figure 2 for sample effects).

The practical properties of proposed method are illustrated and examined on synthetic data sets and examples retrieved form the UCI repository [14]. We compare sWARDS with similar methods which can be applied for non Euclidean data as k-means, Spectral Clustering and their kernelized versions. Our tests demonstrate that introduced method can be applied for populations detection in simple biological systems.

The paper is organized as follows. Next section gives a brief description of related clustering methods. In section 3 we recall Wards approach to k-means and present its application for spherical CEC criterion function. Section 4 demonstrates the generalization of Voronoi diagrams to the case of arbitrary criterion functions in non Euclidean data paying particular attention on sWARDS method. The results of experiments and potential applications are given in section 5 while section 6 contains the conclusion.

II. RELATED WORKS

The hierarchical clustering is probably one of the most popular methods to partition data based on any kind of (dis)similarity measure [15]. The well-known k-means algorithm [16] can also be adapted to non Euclidean data by defining a medoid [17] which plays a role of a generalized notion of mean or by using the Wards method [9], [11] which reformulates the within cluster sum of squares without the notion of the cluster mean. Despite the wide use of these methods, they are sometimes unable to discover groups with complex structures and different sizes. A lot of modifications were also considered to describe clusters with arbitrary shapes [18], [19]. Spectral Clustering uses eigenvectors of similarity matrix to divide elements into groups [20].

Another issue of clustering non Euclidean data sets is the appropriate selection of dissimilarity measure. Examples showed that interesting effects can be obtained by applying Gaussian radial basis function (RBF) [21]. The difficulty is that there is no unified methodology how to choose the radius of this function for particular situation [22], [23].

In order to perform a distribution-based clustering a GMM is widely used in Euclidean space [1]. Nevertheless it cannot be directly generalized to arbitrary data sets with dissimilarity measures. On the other hand, a family of density based clustering such as DBSCAN [24] can be applied for non Euclidean data. Although the method is capable of discovering clusters of arbitrary shapes and does not require the specification of the number of groups, it does not adopt well to clusters with large differences in densities.

Proposed sWARDS method joins the simplicity and flexibility of k-means with the effects of GMM. Its can be applied in non Euclidean spaces and is based on Gaussian-like probability distributions.

III. CLUSTERING METHOD

The proposed sWARDS clustering is a combination of spherical Cross-Entropy Clustering (sCEC) [10] with a generalized Wards approach [9], [11]. In this section we first introduce a basic notation and recall the Wards version of k-means. Then, we show how sCEC can be generalized to non Euclidean data sets via Wards method.

A. Wards method

Generally, k-means method aims at producing a splitting of data set which optimizes a squared error criterion function. For a group $Y \subset \mathbb{R}^N$ the within cluster sum of squares is defined as:

$$\mathrm{ss}(Y) = \sum_{y \in Y} \|y - \mathrm{m}_Y\|^2,$$

where m_Y is a mean of Y. The k-means looks for a partition of $X \subset \mathbb{R}^N$ into k pairwise disjoint sets $Y_1, \ldots, Y_k$ such that the function

$$\sum_{j=1}^{k} \mathrm{ss}(Y_j)$$

is minimal.

Note that the above formulas cannot be used directly for non vector data since the mean is not well-defined for general data sets. There are several alternatives [25], [26], [27], [28], [29] which allow to partially overcome this difficulty as k-medoids [30] or k-clustering [31]. The technique related to k-clustering and k-means is the generalized Wards method [9], [11] which plays the basic role in our investigations. The key idea is the observation that the within cluster sum of squares in Euclidean space can be formulated equivalently without the notion of the center of cluster:

Proposition 3.1: [11] If $Y \subset \mathbb{R}^N$, then

$$\sum_{y \in Y} \|y - \mathrm{m}_Y\|^2 = \frac{1}{2|Y|} \sum_{y \in Y} \sum_{z \in Y} \|y - z\|^2,$$

where $|Y|$ is a cardinality of Y.

This allows to reasonably generalize the within cluster sum of squares to general non Euclidean data set. For this purpose let X be an arbitrary data set and let $d : X \times X \to [0, +\infty)$ be a symmetric dissimilarity measure on X, i.e,

- $d(y, y) = 0$,
- $d(y, z) = d(z, y)$,

for $y, z \in X$. Given two subsets Y, Z of X we define a function [31] connected with the average linkage function (also called average neighbor function) [32], [27] as:

$$D\langle Y, Z \rangle := \sum_{y \in Y} \sum_{z \in Z} d^2(y, z).$$

As a generalized within cluster $Y \subset X$ sum of squares we put [9]:

$$\mathrm{ss}(Y) := \frac{1}{2|Y|} D\langle Y, Y \rangle = \frac{1}{2|Y|} \sum_{y \in Y} \sum_{z \in Y} d^2(y, z). \quad (2)$$

Then, the goal of Wards method is formulated as follows:

Wards Optimization Problem [9]. Let X be a data set with a dissimilarity measure d and let $k \in \mathbb{N}$. Find a splitting of X into k pairwise disjoint sets $Y_1, \ldots, Y_k$ which minimizes the generalized squared error function:

$$\mathrm{E}_{\mathrm{Wards}}(Y_1, \ldots, Y_k) := \sum_{i=1}^{k} \mathrm{ss}(Y_i), \quad (3)$$

where $\mathrm{ss}(\cdot)$ is defined by (2).

B. Spherical Wards criterion function

The Cross-Entropy Clustering (CEC) is a kind of distribution-based clustering which divides an Euclidean data set into groups such that each group is described by optimally fitted Gaussian probability distribution [10]. The effects of the clustering are similar to those obtained by GMM, but the optimizing criterion function is different. Its value determines the statistical code length of memorization of an arbitrary element of a data set in the case when each cluster uses its own coding algorithm. In particular, the introducing of one more cluster (coding algorithm) requires an additional cost of its identification (increase of the entropy). In consequence, the maintaining of too many clusters is not optimal and it allows for the automatic reduction of unnecessary groups. Another advantage of CEC is that the clustering is performed in a comparable time to computationally efficient k-means method. For more details the reader is referred to [10], [33], [34].

Spherical Cross-Entropy Clustering (sCEC) is a variant of CEC which takes into account the family of spherical Gaussian distributions. Since for every group the optimal spherical Gaussian distribution is matched, then data set is partitioned into spherically-shaped clusters. For a splitting $Y_1, \ldots, Y_k$ of X the associated criterion function is defined by

$$\frac{N}{2} \ln\left(\frac{2\pi e}{N}\right) + \sum_{i=1}^{k} \frac{|Y_i|}{|X|} \cdot \left[-\ln \frac{|Y_i|}{|X|} + \frac{N}{2} \ln\left(\frac{|X|}{|Y_i|} \mathrm{tr}(\Sigma_{Y_i}) \right) \right],$$
$$(4)$$

where Σ_Y is a covariance matrix of group Y and $\mathrm{tr}(\Sigma_Y)$ is a trace of Σ_Y.

Let us first observe that the notion of covariance matrix can be easily removed from the expression (4).

Proposition 3.2: If $Y \subset \mathbb{R}^N$ then:

$$\mathrm{tr}(\Sigma_Y) = \mathrm{ss}(Y).$$

In consequence the application of Wards approach facilitates its interpretation in non Euclidean case for a fixed $N > 0$.

For fully explanation of the formula (4) in the context of non Euclidean space, the value of dimension N has to be specified. As the most reasonable way to set this value we recommend to use the estimation of a dimension of X. In the present study we apply the Maximum Likelihood Estimation (MLE) of intrinsic dimension of X proposed by [12] and modified by [13] (see Remark 3.1). However one can tune this value in the learning process as well as may set it to any positive number. In the experimental section we show that for high values of N the more clusters are created in the clustering while for low values of N the method prefers to reduce a number of groups.

Remark 3.1: We recall a basic procedure for an estimation of a dimension of a data set based on a maximum likelihood principle. Let $X = \{x_1, \ldots, x_n\}$ be a data set. The maximum likelihood estimator of a dimension N of X calculated for each $x \in X$ equals [12]:

$$\hat{N}_k(x) = \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{d(x, x_k)}{d(x, x_j)},$$

for $k \in \{1, \ldots, n\}$. Since the above value is dependent on the choice of k and x, then one should average the results over

$x \in X$ and $\tilde{K} \subset \{1, \ldots, n\}$ to obtain the final estimator of N [13].

All in all, the generalized Wards approach and the appropriate choice of the dimension parameter N allow for the understanding of spherical cross-entropy criterion function in arbitrary data set with a dissimilarity measure. In consequence, the informal notion of spherical Gaussian probability distribution based on any dissimilarity measure could be considered. We conclude this subsection with a formulation of spherical Wards (sWARDS) optimization problem:

Spherical Wards Optimization Problem. Let X be a data set with a dissimilarity measure d, $n \in \mathbb{N}$ be an initial number of clusters and $N > 0$ be a free parameter. Find $k \leq n$ and a partition $Y_1, \ldots, Y_k$ of X which minimizes spherical Wards criterion function

$$E_{sWards}(Y_1, \ldots, Y_k; N) :=$$
$$\frac{N}{2} \ln\left(\frac{2\pi e}{N}\right) + \sum_{i=1}^{k} \frac{|Y_i|}{|X|} \cdot \left[\frac{N}{2} \ln(ss(Y_i)) - \frac{N+2}{2} \ln\left(\frac{|Y_i|}{|X|}\right) \right],$$
$$\tag{5}$$

where $ss(\cdot)$ is defined by (2).

C. Clustering algorithm

One can show that the natural modification of the Hartigan algorithm [9], [16], [10] can be used to minimize the sWARDS criterion function (5). We will now discuss its technical aspects.

The procedure can be divided into two parts: initialization and iteration. In the initialization phase $n \in \mathbb{N}$ groups are created randomly. During iteration the algorithm reassigns elements between clusters in order to minimize the sWARDS criterion function (5).

More precisely, in the iteration part we repeatedly go over all elements of X applying the following steps:

1) Reassign $x \in X$ to this cluster for which the decrease of energy (5) is maximal,
2) If a probability of some cluster is less than a fixed number $\varepsilon > 0$, then remove this cluster and assign its elements to these groups for which the increase of energy (5) is minimal,

until no group membership has been changed.

The number ε was introduced to speed up the reduction of redundant clusters. In our experiments we always use the value $\varepsilon = 1\%$. Thus, the group is removed if it contains less than 1% of all elements of X. Clearly, the procedure is not deterministic and leads to a local minimum of (5) [25]. Therefore, to provide the satisfactory results the algorithm should be evaluated several times – the final result is that which gives the minimal value of sWARDS criterion function.

Basically, the above Hartigan procedure relies on the switching the elements between clusters and recalculating the sWARDS criterion function. This implies that to efficiently apply this procedure we have to recompute

$$ss(Y \cup \{x\}) \text{ and } ss(Y \setminus \{x\}),$$

very fast. For this purpose the following formulas are useful:

Proposition 3.3: [11] Let $Y \subset X$ and $x \in X$.

a) If $x \notin Y$, then

$$ss(Y \cup \{x\}) = \frac{|Y|}{|Y| + 1} ss(Y) + \frac{1}{|Y| + 1} D\langle\{x\}, Y\rangle.$$

b) If $x \in Y$, then

$$ss(Y \setminus \{x\}) = \frac{|Y|}{|Y| - 1} ss(Y) - \frac{1}{|Y| - 1} D\langle\{x\}, Y\rangle.$$

Given k clusters, the computational complexity of one iteration of standard Hartigan procedure requires about $k \cdot N \cdot |X|$ operations (for data sets contained in $\mathbb{R}^N$). When applying the Wards approach this complexity changes to $k \cdot |X|^2$ operations. Since the mean of cluster is not defined in general situation, one has to pay an additional cost of recalculating the within cluster sum of squares during every reassigning. However, we do not need to recalculate the distance between the reassigning elements and the mean of a cluster which decreases the computational cost N times.

IV. GENERALIZED VORONOI DIAGRAM

There arises a natural problem how to graphically present the clustering results. Clearly, we can mark the elements of each cluster with different label. However, in practice it is usually more clear to show the division of whole space. In this section we show that we can naturally obtain an equivalence of the Voronoi diagram for any criterion function in non Euclidean space. In particular we apply these results to define the Voronoi diagram for sWARDS.

A. Classical diagram

Let us recall that in the case of classical version of Voronoi diagram (k-means method) the point x is associated with this cluster whose center is the closest to x. More precisely, it is classified to this cluster Y_i which minimizes $d(x; m_i)$, where m_i is a mean of Y_i.

Remark 4.1: One can consider the alternative to the Voronoi diagrams as described in [35]. It provides the partition of data but does not induce a natural partition of the space. We refer the reader to [35] for more details where this idea in investigated.

To generalize the notion of the Voronoi diagram to non Euclidean space (Wards k-means), we need to be able to compute the distance of a point from the center of the cluster (without using it in the computations).

Proposition 4.1: Let $x \in \mathbb{R}^N$ be fixed and $Y \subset \mathbb{R}^N$ be a subset of $\mathbb{R}^N$ with mean m_Y. Then

$$\|x - m_Y\|^2 = \frac{1}{|Y|} \sum_{y \in Y} \|x - y\|^2 - \frac{1}{2|Y|^2} \sum_{y \in Y} \sum_{z \in Y} \|y - z\|^2.$$

The above allows the formulation of the analogue of the square of the "classical" distance of a point x from the center of Y. Let Y be a subset of data space X with a dissimilarity measure

d and let $x \in X$ be fixed. We define the mean square distance of x from Y by

$$d^2(x;Y) := \frac{1}{|Y|}(D\langle\{x\},Y\rangle - \mathrm{ss}(Y)). \qquad (6)$$

Applying the above formula one can draw the equivalence of the Voronoi diagram for Wards k-means, i.e. an element $x \in X$ is classified to this cluster which minimizes (6).

B. Diagram for arbitrary criterion function

We are now going to present a reasoning which allows to create a kind of Voronoi diagram for arbitrary criterion function. This will be useful for constructing a division of the space for the case of SWARDS method. Obtained results are consistent with the classical Voronoi diagram in the case of Wards k-means presented in previous section.

Let X be a space with a dissimilarity measure d and let $Y \subset X$ represent our data. We extend X by introducing a weight function

$$w : X \ni x \to \begin{cases} w(x) \in [0, +\infty) & , x \in Y, \\ 0 & , x \in X \setminus Y, \end{cases}$$

which assigns a weight to every element of X. Then we consider an extended data set

$$Y^w = \{(y, w(y)) : y \in Y\}.$$

We define the operations $D\langle\cdot,\cdot\rangle$ and $\mathrm{ss}(\cdot)$ adapted for Y^w. Given $Z, Y_1, Y_2 \subset Y$ we put:

1) $|Z^w| := \sum\limits_{z \in Z} w(z)$,

2) $D\langle Y_1^w, Y_2^w \rangle := \sum\limits_{y_1 \in Y_1} \sum\limits_{y_2 \in Y_2} d^2(y_1, y_2) w(y_1) w(y_2)$,

3) $\mathrm{ss}(Z^w) := \frac{1}{2|Z^w|} D\langle Z^w, Z^w \rangle$.

Then the analogue of k-means criterion function equals:

$$\mathrm{E_{Wards}}(Y_1^w, \ldots, Y_k^w) = \sum_{i=1}^{k} \mathrm{ss}(Y_i^w), \qquad (7)$$

where $Y_1, \ldots, Y_k$ is a splitting of Y. If $w_{|Y} \equiv 1$ then (7) coincides with (3).

In order to explain our technique assume that $Y_1, \ldots, Y_k$ is a splitting of data set Y and E is an arbitrary criterion function. For a fixed point $x \in X$ we consider a mapping

$$E^i_{x,[Y_1^w,\ldots,Y_k^w]} : h \to$$
$$E(Y_1^w, \ldots, Y_{i-1}^w, (Y_i \cup \{x\})^{w+h\delta_x}, Y_{i+1}^w, \ldots, Y_k^w),$$

where $h \geq 0$ and $i \in \{1, \ldots, k\}$. It determines the value of criterion function E when $x \in X$ is associated with i-th cluster with a weight increased by h.

We define the functions (wherever they exist)

$$\partial_i E(x, [Y_1^w, \ldots, Y_k^w]) := (E^i_{x,[Y_1^w,\ldots,Y_k^w]})'(0),$$

for $i \in \{1, \ldots, k\}$. Observe that $\partial_i E$ coincides with the infinitesimal change in energy when we add x to the i-th cluster. Thus, in Voronoi diagram the point $x \in X$ should be assigned to this cluster which minimizes $\partial_i E(x, [Y_1^w, \ldots, Y_k^w])$.

Let us show that the above reasoning is consistent with the classical results (6) for Wards k-means criterion function (7):

Theorem 4.1: Let Y be a subset of a space X with a dissimilarity measure d and let $w(y) = 1$, for all $y \in Y$, be a weight function. If E denotes the squared error function (7) and $Y_1, \ldots, Y_k$ is a fixed splitting of Y then

$$\partial_i E(x, [Y_1^w, \ldots, Y_k^w]) = d^2(x; Y_i),$$

for $x \in X$ and $i \in \{1, \ldots, k\}$.

Proof: Let $h > 0$. By Corollary 3.3, we have

$$\frac{1}{h}[E(Y_1^w, \ldots, Y_{i-1}^w, Y_i^w \cup \{(x,h)\}, Y_{i+1}^w, \ldots, Y_k^w)$$
$$- E(Y_1^w, \ldots, Y_k^w)]$$
$$= \frac{1}{h}\left[\mathrm{ss}((Y_i \cup \{x\})^{w+h\delta_x}) - \mathrm{ss}(Y_i^w)\right]$$
$$= \frac{1}{h}\left[\frac{|Y_i^w|\mathrm{ss}(Y_i^w) + D\langle\{(x,h)\},Y_i^w\rangle}{|Y_i^w|+h} - \mathrm{ss}(Y_i^w)\right]$$
$$= \frac{1}{h}\frac{|Y_i^w|\mathrm{ss}(Y_i^w) + hD\langle\{(x,1)\},Y_i^w\rangle - (|Y_i^w|+h)\mathrm{ss}(Y_i^w)}{|Y_i^w|+h}$$
$$= \frac{D\langle(x,1),Y_i^w\rangle - \mathrm{ss}(Y_i^w)}{|Y_i^w|+h}.$$

Since $w_{|Y} \equiv 1$ then

$$\frac{D\langle(x,1),Y_i^w\rangle - \mathrm{ss}(Y_i^w)}{|Y_i^w|+h} = \frac{D\langle x, Y_i\rangle - \mathrm{ss}(Y_i)}{|Y_i^w|+h} \to$$
$$\frac{1}{|Y_i|}(D\langle x, Y_i\rangle - \mathrm{ss}(Y_i)) \text{ , as } h \to 0,$$

which yields the assertion of the theorem. ∎

C. Voronoi diagram for SWARDS

The following theorem presents how to create the Voronoi diagram for SWARDS criterion function:

Theorem 4.2: Let Y be a subset of a space X with a dissimilarity measure d and let $w(y) = 1$, for all $y \in Y$, be a weight function. If E denotes the SWARDS criterion function for a data set with weights and $Y_1, \ldots, Y_k$ is a fixed splitting of Y then

$$\partial_i E(x, [Y_1^w, \ldots, Y_k^w])$$
$$= \frac{1}{|X|}\left[\frac{N}{2}\left(\ln(\mathrm{ss}(Y_i)) + |Y_i|\frac{d^2(x;Y_i)}{\mathrm{ss}(Y_i)}\right) - \frac{N+2}{2}(\ln|Y_i| + 1)\right].$$

Proof: Roughly speaking, Theorem 4.1 says that $\partial_i \mathrm{ss}(Y_i^w) = d^2(x; Y_i)$. Moreover, $\partial_i |Y_i^w| = 1$. Applying the operator ∂_i and the above to (5) we easily get the assertion of the theorem. ∎

Consequently, given a partition $Y_1, \ldots, Y_k$ to associate the point $x \in X$ to a cluster it is sufficient to find i which minimizes

$$\ln(\mathrm{ss}(Y_i)) + |Y_i|\frac{d^2(x;Y_i)}{\mathrm{ss}(Y_i)} - (1 + \frac{2}{N})\ln|Y_i|. \qquad (8)$$

V. EXPERIMENTS

In this section we discuss some fundamental properties as well as the potential applications of proposed clustering method and present a short evaluation study.
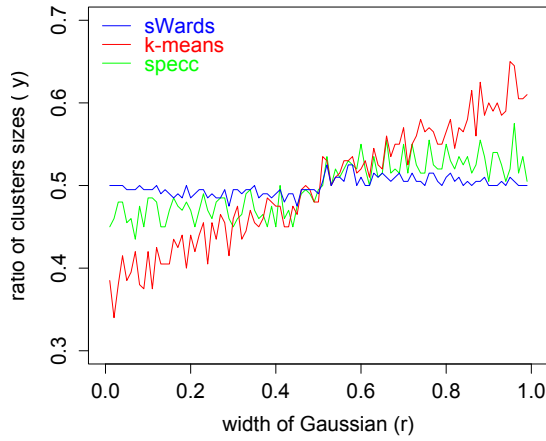
Fig. 3. **Scale invariance.** The rations of clusters sizes for a data set generated from the mixture of two spherical Gaussian distributions $\frac{1}{2}\mathcal{G}_1(r) + \frac{1}{2}\mathcal{G}_2(1-r)$ when changing the width r of Gaussians. The optimal curve should be a constant function, $y = \frac{1}{2}$.
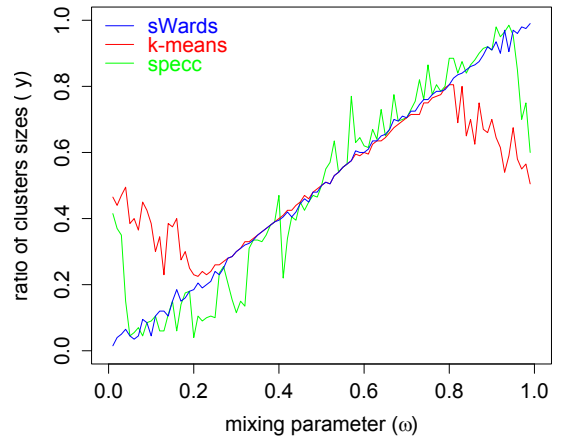

Fig. 4. **Sensitivity on the unbalanced data.** The rations of clusters sizes for a data set generated from the mixture of two spherical Gaussian distributions $\omega\mathcal{G}_1 + (1-\omega)\mathcal{G}_2$. Number $\omega \in (0,1)$ controls the number of elements produced by each Gaussian. The optimal curve should be a linear function, $y = \omega$.

A. Synthetic data sets

In order to show the capabilities of SWARDS we examined its resistance on the change of scale and its sensitivity on the unbalanced data. We compared the clustering results with the ones obtained with use of related methods which can be applied for non Euclidean spaces: Wards k-means and Spectral Clustering (kernlab R package was used for the implementations of this algorithm [36]). Since SWARDS automatically detects the resultant number of groups, then we ran it with 10 initial clusters while the other methods used the number of groups returned by SWARDS[1]. The value of parameter N (dimension of space) for SWARDS was set automatically with use of MLE method [12], [13]. To provide more stable results, each algorithm was run 10 times and the result with the lowest value of criterion function was chosen.

Example 5.1: **Scale invariance**

In the first experiment we examined the invariance of algorithms on the change of scale. A data set was generated from the mixture of two spherical Gaussian distributions,

$$\frac{1}{2}\mathcal{G}_1(r) + \frac{1}{2}\mathcal{G}_2(1-r)$$

with different covariance matrices

$$C_1 = \begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}, \; C_2 = \begin{pmatrix} 1-r & 0 \\ 0 & 1-r \end{pmatrix}, \text{ for } r \in (0,1),$$

centered at
$$m_1 = (-1,0), \; m_2 = (1,0).$$

The parameter r controls the width of Gaussians.

The Figure 3 presents the ratios of resulted clusters sizes. The SWARDS method is robust to the change of scale – the clusters remained almost equally-sized for all $r \in (0,1)$. The clustering result was the most dependent on the widths of Gaussians in the case of k-means.

Example 5.2: **Unbalanced data**

We have also tested how the number of elements generated from the individual distributions affects the clustering results. For this purpose data was generated from the mixture of two Gaussians

$$\omega \cdot \mathcal{G}_1 + (1-\omega) \cdot \mathcal{G}_2 \text{ , for } \omega \in (0,1),$$

with identical covariance matrices

$$C_1 = C_2 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix},$$

but different centers

$$m_1 = (-1,0), \; m_2 = (1,0).$$

The number of elements generated from each Gaussian is determined by the value of parameter ω.

The ratios of clusters sizes are shown in the Figure 4. One can observe that the proportions specified by ω was preserved by SWARDS method. In the Spectral Clustering the results are less stable. On the other hand Wards k-means has a tendency to build equally-sized clusters.

B. Dimension estimation

To apply the SWARDS criterion function in the case of arbitrary non Euclidean space the value of dimension parameter N needs to be specified. In the previous subsection we showed that the reasonable clustering results can be obtained calculating this value using MLE method [12], [13]. We will experimentally show how the clustering effects differ when the value of N changes.

Example 5.3: **Clusters number detection**

Let us first examine the impact of the value of parameter N on the detection of the resultant number of groups. For this purpose a mouse-like set (see Figure 2) was clustered with different values of N starting from 100 initial groups. The resultant number of groups are illustrated in the Figure 5.

The immediate observation is that the increase of the value of N results in the increase of the detected number of groups.

---

[1]Such a technique for a detection of clusters number was chosen in order to provide the correspondence between clustering results for all methods.
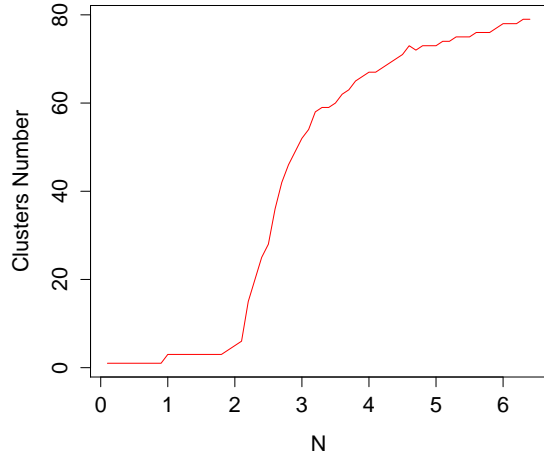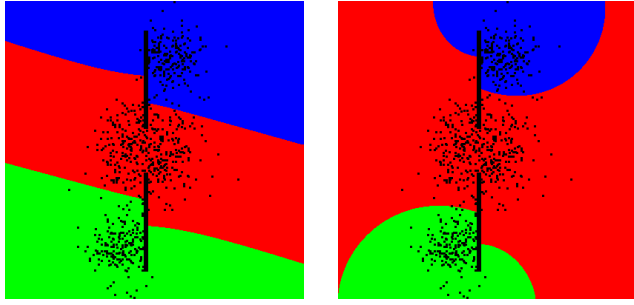
Fig. 5. **Clusters detection.** The influence of the value of parameter N on the resulted number of clusters. Maximal number of clusters was set to 100.
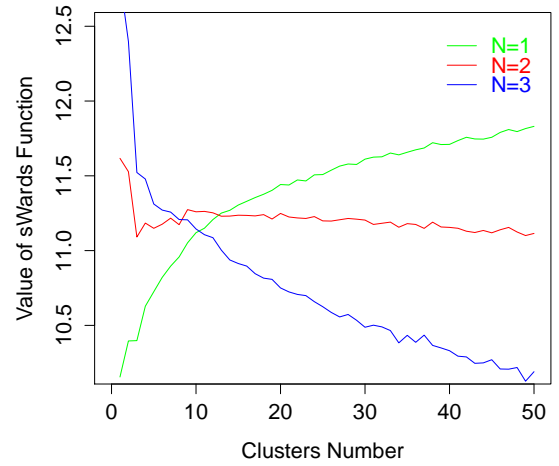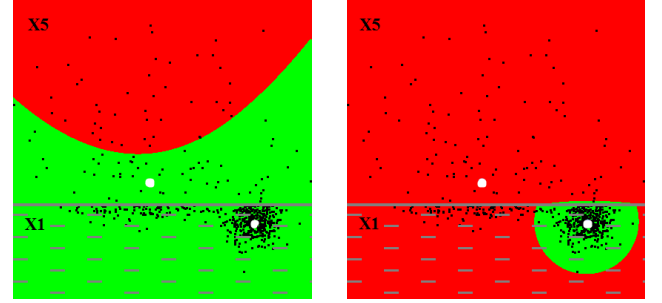

Fig. 7. **Shape of criterion function.** The influence of the clusters number on the shape of sWARDS criterion function for three exemplary values of N.

(a) Wards k-means clustering with $k = 3$.

(b) sWARDS clustering started with 10 initial clusters.

Fig. 6. **Populations districts on the space with barriers.** Voronoi diagrams constructed by Wards k-means and sWARDS on a data space with barriers containing three populations generated from random walk models.

(a) Wards k-means clustering with $k = 2$.

(b) sWARDS clustering started with 10 initial clusters.

Fig. 8. **Populations districts on the space with regions.** Voronoi diagrams constructed by Wards k-means and sWARDS on a data space divided into two regions X_1 and X_5 containing three populations generated from random walk models. The speed of movements in X_5 is 5 times higher than in X_1.

One can observe that for $N < 1$ the entire data set was recognized as one group. For $N \in (1, 2)$ the mouse-like set was partitioned into three groups which seems to be the most appropriate partitioning. For $N > 2$ the number of groups began to grow rapidly.

Example 5.4: **Shape of criterion function**

To get more insight into the influence of dimension parameter on the discovered number of clusters, we analyzed the shape of sWARDS criterion function for different values of N. Since the sWARDS automatically reduces unnecessary clusters, it is not possible to directly specify the number of groups. Therefore, a mouse-like data set (see Figure 2) was first partitioned into expected number of groups with use of k-means. Then, the sWARDS criterion function was calculated for each partition.

It is clear from Figure 7 that the criterion function yields a global minimum for 3 clusters when $N = 2$. Therefore, in most cases the algorithm ends with 3 groups. For $N = 1$ the cost of maintaining clusters increases and the algorithm generally includes all elements into one group. The function is decreasing for $N = 3$. It means that the method rarely reduces clusters. The last case can be a very useful variant of sWARDS when the resulting number of groups should not be discovered by the algorithm but specified directly by the user.

C. Applications

In this section we show that the proposed method is very useful in the analysis of biological models of populations. It is assumed that a population follows a random walk model $P(x, n, t)$ on a plane [37], where at each unit of time an instance moves randomly in one of four directions: left, right, up or down. More precisely, given a starting point (seed) $x \in X$, n-instances are generated from a random walk model assuming t time units. It is worth to mention that a probability distribution of a population converges to spherical Gaussian one [37]. Given a data set consisting of k populations we would like to discover them during a clustering process. Constructed Voronoi diagram determines the corresponding populations districts in the whole space.

Let us observe that, in practice the environment does not represent an Euclidean space. Indeed, a plane is usually crossed by rivers and barriers. Moreover, the environment can be divided into various regions, e.g. meadows, seas, forests etc., which changes the speed of movement of individuals. These modifications change the classical Euclidean metric – the distance between elements has to take into account all the aforementioned circumstances. In the experiments we analyze two cases of populations environments.

| | | | Euclidean dissimilarity | | | | RBF dissimilarity | | | |
Data	True Cl.	N	Est. Cl.	sWards	k-means	Specc	Est. Cl.	sWards	k-means	Specc
Cmc	3	2.64	4	**0.61**	0.57	0.58	2	**0.55**	0.51	0.51
Ecoli	8	3.72	9	**0.88**	0.83	0.77	10	**0.84**	0.79	0.8
Glass	7	3.07	8	**0.71**	0.7	0.68	7	0.7	**0.71**	**0.71**
Hayes-r.	3	1.85	5	**0.62**	0.58	0.59	5	**0.61**	0.5	0.6
Ionosph.	2	5.03	4	0.55	0.52	**0.61**	4	0.57	**0.61**	0.58
Iris	3	2.49	4	**0.85**	0.81	0.83	5	**0.85**	0.84	0.83
Tae	3	2.06	6	**0.61**	**0.61**	0.6	5	**0.62**	0.58	0.6
Wine	3	1.64	4	**0.75**	0.63	0.68	5	**0.58**	0.55	0.55
Yeast	10	4.81	11	0.64	**0.73**	**0.73**	10	0.63	**0.73**	**0.73**

Example 5.5: **Environment with barriers.** Let us consider three populations living in the environment showed in Figure 6 crossed by two barriers which modify Euclidean distance function. Basically, the distance between elements located on the opposite sides of barrier is calculated as a shortest path which does not cross the barrier.

Regions occupied by populations can be obtained with use of Voronoi diagram. Is is clear from Figure 6(a) that Wards k-means discovered populations districts as horizontal stripes which is not an appropriate model. More accurate partition results from sWards (see Figure 6(b)), where detected regions form circular shapes. Partitions agreement measured by Rand index [38] for Wards k-means equals 96% while for sWards is 98%.

Example 5.6: **Environment with regions.** In the second example let us assume that a data space X is divided into two regions X_1 and X_5. In X_5 the individuals moves 5 times faster than in X_1. This inducts a dissimilarity measure on X by:

$$d(x,y) := \begin{cases} d_E(x,y), & x,y \in X_5, \\ 5d_E(x,y), & x,y \in X_1, \\ \inf\{5d_E(x,z) + d_E(z,y) : \text{border point } z\}, \\ \quad x \in X_1, y \in X_5, \end{cases}$$

where $d_E(\cdot, \cdot)$ denotes the Euclidean distance. We consider two populations showed in Figure 8 with starting points marked with white dots in X_1 and X_5 respectively.

One can observe in Figure 8(b) that despite the form of the above dissimilarity measure, sWards detected the circular-like districts of populations very well. This result can be compared with k-means clustering (see Figure 8(a)) where a produced partition does not match populations distributions. The value of Rand index for sWards equals 92% while for Wards k-means is 61%.

D. Evaluation

After establishing the properties as well as demonstrating basic capabilities and potential applications of introduced method we present a short evaluation. We carried out the experiments on selected UCI data sets [14]. In all experiments the initial number of clusters for sWards was fixed two times higher than the actual number of groups. In order to provide the correspondence between the clustering results the other
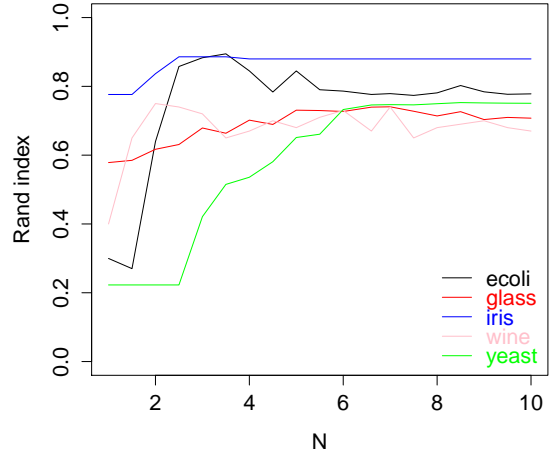

Fig. 9. **UCI evaluation.** The influence of the value of parameter N on the clustering results of UCI data sets.

examined methods assumed the number of groups returned by sWards as the input clusters number.

As a measure of agreements between partitions the Rand index (RI) was used [38]. It is defined as a ratio between pairs of true positives and false negatives, and all pairs of examples. The values close to 1 indicate that two partitions are very similar. MLE was used to calculate the optimal value of parameter N. Two kinds of dissimilarity measures were considered: the Euclidean distance and the dissimilarity determined by the Gaussian radial basis function (RBF). The value of sigma for RBF was estimated as a median of the squared distances between all pairs of data set elements [39].

The results presented in Table I show that sWards reasonably well determined the final number of groups. The advantage of our method over k-means and Spectral Clustering is the most evident for the case of Ecoli data set and Euclidean distance. The worst results were obtained for Ionosphere data set. The use of RBF similarity rarely improved the accuracy of clustering. It could be caused by the fact that it is very difficult to set the optimal value for RBF sigma parameter in particular situation.

To extend the above evaluation, in the Figure 9 we present the clustering accuracies of UCI data sets for a wide range of dimension parameter $N \in (0, 10)$. One can observe that in most cases the best results were obtained when N was

estimated as a dimension of data. The exceptions are Glass and Yeast data sets where a slight improvement was achieved for higher values of N.

VI. Conclusion

In this paper a generalization of spherical Cross-Entropy Clustering to non Euclidean spaces was presented. The proposed method uses a Wards approach to modify the cross-entropy criterion function for the case of arbitrary data sets. In consequence, obtained method allows for partitioning of non vector data into spherically-shaped clusters of arbitrary sizes. It is scale invariant technique which detects the final number of groups automatically. Our method works in comparable time to generalized Wards method while the clustering effects are similar to those produced by GMM when focusing on spherical Gaussian distributions in Euclidean spaces.

Moreover, we generalized the notion of Voronoi diagram for the case of arbitrary criterion function based on Wards approach. This leads to identical results in the case of classical methods as k-means while it allows for formal division of data space when focusing on non Euclidean methods as sWards.

References

[1] G. McLachlan and T. Krishnan, *The EM algorithm and extensions.* John Wiley & Sons, 2007, vol. 382.

[2] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381–396, 2002.

[3] J. Samuelsson, "Waveform quantization of speech using Gaussian mixture models," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004*, vol. 1. IEEE, 2004, pp. I–165.

[4] E. P. Xing, M. I. Jordan, R. M. Karp *et al.*, "Feature selection for high-dimensional genomic microarray data," in *Proceedings of the 18th International Conference on Machine Learning*, vol. 1. Citeseer, 2001, pp. 601–608.

[5] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, "Time series classification using Gaussian mixture models of reconstructed phase spaces," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 16, no. 6, pp. 779–783, 2004.

[6] J. Gasteiger, *Handbook of chemoinformatics.* Wiley Online Library, 2003, vol. 1.

[7] S. L. Dixon and H. O. Villar, "Investigation of classification methods for the prediction of activity in diverse chemical libraries," *Journal of Computer-Aided Molecular Design*, vol. 13, no. 5, pp. 533–545, 1999.

[8] O. Zamir and O. Etzioni, "Web document clustering: A feasibility demonstration," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1998, pp. 46–54.

[9] V. Batagelj, "Generalized Ward and related clustering problems," *Classification and Related Methods of Data Analysis*, pp. 67–74, 1988.

[10] J. Tabor and P. Spurek, "Cross-entropy clustering," *Pattern Recognition*, vol. 47, no. 9, pp. 3046–3059, 2014.

[11] H. Spath, "Cluster-analyse-algorithmen," *München und Wien*, 1975.

[12] E. Levina and P. Bickel, "Maximum likelihood estimation of intrinsic dimension," *Ann Arbor MI*, vol. 48109, p. 1092, 2004.

[13] "Comments on "Maximum likelihood estimation of intrinsic dimension" by E. Levina and P. Bickel (2004), author=MacKay, D. J. C. and Ghahramani, Z., year=2012, note=Available from http://www.inference.phy.cam.ac.uk/mackay/dimension/."

[14] A. Asuncion and D. Newman, "UCI machine learning repository," 2007.

[15] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[16] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied Statistics*, pp. 100–108, 1979.

[17] H. S. Park and C. H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[18] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, 2004, pp. 551–556.

[19] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl *et al.*, "Constrained k-means clustering with background knowledge," in *Proceedings of the 18th International Conference on Machine Learning*, vol. 1, 2001, pp. 577–584.

[20] D. Zhou and C. J. C. Burges, "Spectral clustering and transductive learning with multiple views," in *Proceedings of the 24th International Conference on Machine learning.* ACM, 2007, pp. 1159–1166.

[21] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *Neural Networks, IEEE Transactions on*, vol. 4, no. 4, pp. 570–590, 1993.

[22] M. J. L. Orr, "Regularization in the selection of radial basis function centers," *Neural Computation*, vol. 7, no. 3, pp. 606–623, 1995.

[23] S. Rippa, "An algorithm for selecting a good value for the parameter c in radial basis function interpolation," *Advances in Computational Mathematics*, vol. 11, no. 2-3, pp. 193–210, 1999.

[24] H. P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[25] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[26] D. Steinley, "K-means clustering: A half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[27] G. Gan, C. Ma, and J. Wu, *Data clustering.* SIAM, Society for Industrial and Applied Mathematics, 2007.

[28] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[29] R. Xu and D. Wunsch, *Clustering.* Wiley-IEEE Press, 2009.

[30] L. Kaufman and P. Rousseeuw, "Clustering by means of medoids," 1987.

[31] P. Indyk, "A sublinear time approximation scheme for clustering in metric spaces," in *Foundations of Computer Science, 1999. 40th Annual Symposium on.* IEEE, 1999, pp. 154–159.

[32] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *Univ. Kans. Sci. Bull.*, vol. 38, pp. 1409–1438, 1958.

[33] M. Śmieja and J. Tabor, "Image segmentation with use of cross-entropy clustering," in *Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013.* Springer, 2013, pp. 403–409.

[34] J. Tabor and K. Misztal, "Detection of elliptical shapes via cross-entropy clustering," in *Pattern Recognition and Image Analysis.* Springer, 2013, pp. 656–663.

[35] M. Telgarsky and A. Vattani, "Hartigan's Method: k-means Clustering without Voronoi," in *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[36] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab-an S4 package for kernel methods in r," 2004.

[37] E. A. Codling, M. J. Plank, and S. Benhamou, "Random walk models in biology," *Journal of the Royal Society Interface*, vol. 5, no. 25, pp. 813–834, 2008.

[38] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[39] B. Caputo, K. Sim, F. Furesjo, and A. Smola, "Appearance-based Object Recognition using SVMs: Which Kernel Should I Use?" in *Proceedings of NIPS workshop on Statistical Methods for Computational Experiments in Visual Processing and Computer Vision, Whistler*, vol. 2002, 2002.