Outlier Robust Geodesic K-means Algorithm for High Dimensional Data

Aidin Hassanzadeh[1(✉)], Arto Kaarna[1], and Tuomo Kauranne[2]

[1] Machine Vision and Pattern Recognition Laboratory,
School of Engineering Science, Lappeenranta University of Technology,
Lappeenranta, Finland
{aidin.hassanzadeh,arto.kaarna}@lut.fi

[2] Mathematics Laboratory, School of Engineering Science,
Lappeenranta University of Technology, Lappeenranta, Finland
tuomo.kauranne@lut.fi

Abstract. This paper proposes an outlier robust geodesic K-mean algorithm for high dimensional data. The proposed algorithm features three novel contributions. First, it employs a shared nearest neighbour (SNN) based distance metric to construct the nearest neighbour data model. Second, it combines the notion of geodesic distance to the well-known local outlier factor (LOF) model to distinguish outliers from inlier data. Third, it introduces a new ad-hoc strategy to integrate outlier scores into geodesic distances. Numerical experiments with synthetic and real world remote sensing spectral data show the efficiency of the proposed algorithm in clustering of high-dimensional data in terms of the overall clustering accuracy and the average precision.

Keywords: Clustering · K-means · High-dimensional data · Geodesic distance · Shared nearest neighbour · Local outlier factor

1 Introduction

The K-means algorithm is one of the widely used clustering algorithms in the area of cluster analysis and it has been integrated into various image processing, data mining and pattern recognition applications. K-means is basically an objective function based optimization scheme that iteratively assigns data to K clusters while attempting to minimize intra-cluster variation. The K-means algorithm is simple and scalable to a wide range of data types. K-means assumes Euclidean distance as the dissimilarity metric and thus tends to produce clusters of spherical shape. Although this assumption has been shown to be reasonable for many applications [10,11], it is not universally true with data clusters of non-spherical and complex shapes, such as spatial data and hyperspectral remote sensing imaging. Moreover, the classic K-means algorithm can adversely be affected by outliers and thus is not able to achieve realistic results if the clusters are contaminated by outlying data.

A. Robles-Kelly et al. (Eds.): S+SSPR 2016, LNCS 10029, pp. 252–262, 2016.
DOI: 10.1007/978-3-319-49055-7_23

Several variants of K-means algorithm have been introduced to address these shortcomings [4,7,11]. The density sensitive geodesic K-means algorithm, henceforth DGK-means, proposed by [1] is an approach that tries to address the issues of non-spherical clusters and outlying data. The DGK-means algorithm replaces the Euclidean distance with a manifold based geodesic distance metric which is resistant to outliers. This algorithm suffers from two main difficulties: first, it can easily be affected by the curse of dimensionality, and second it may fail if the data clusters come from different density patterns.

This paper investigates the DGK-means and proposes an outlier robust geodesic distance based K-means algorithm, called ORGK-means. The proposed algorithm is similar to the DGK-means but utilizes a new geodesic distance metric. By this means, the ORGK-means algorithm attempts to address the issues of high-dimensionality, data of varying cluster densities and data with outliers.

The proposed ORGK-means algorithm includes three main contributions. First, an alternative distance measure based on the notion of shared nearest neigbor (SNN) is proposed for pairwise distance estimation. SNN, originally introduced as a similarity measure based on nearest neighbours [6], is considered an efficient method for problems involving high-dimensional data, clustering of data of varying size and distribution and data contaminated with outliers [5]. Its utilization has been reported in several applications with high-dimensional data [3,8,9,12] and outlier-scoring algorithms [13]. Second, the well-known local outlier factor (LOF) based on the notion of geodesic distance is used to rank outlierness of data. By using geodesic distance based LOF, the ORGK-means algorithm is expected to be more robust to density fluctuations. Third, to provide more flexibility in modelling and improved usability, a double sigmoid function with adaptive parameter estimation is proposed to integrate outlier scores into the distances.

The remainder of the paper is organized as follows. Section 2 briefly reviews the DGK-means algorithm, presenting the main steps involved. Section 3 introduces the new elements proposed to address the shortcomings in the DGK-means algorithm. Section 4 presents experimental results and evaluations. Section 5 concludes the paper.

2 Density Sensitive Geodesic K-means Algorithm

There are three main features in the DGK-means algorithm: general distance K-means, density sensitive geodesic distance, and geodesic K-means.

2.1 General Distance K-means

The DGK-means reformulates the whole update process in the classic K-means to a generative procedure, called general distance K-means, that can utilize any distance metric. Let $X = \{\mathbf{x}_i\}_{i=1}^{n}$ be the set of n real-valued data points of dimension p to be clustered onto m data clusters $C = \{C_l\}_{l=1}^{m}$. Given the distance metric $d : X \times X \rightarrow \mathbb{R}^+$, the general K-means algorithm aims to minimize the objective loss function:

$$W_{GD}(X) = \sum_{l=1}^{m} \sum_{\mathbf{x}_i \in C_l} \sum_{\mathbf{x}_j \in C_l} d^2(\mathbf{x}_i, \mathbf{x}_j). \tag{1}$$

Let (X, d) be the metric space. Provided that X can be mapped onto Euclidean space, the minimization of the objective function in Eq. 1 can be performed iteratively without explicitly calculating the cluster centroids. Denoting $\gamma_{t+1}(\mathbf{x}_i) : X \rightarrow \{l\}_{l=1}^{m}$ as the cluster membership function at iteration $t+1$, the update cluster assignment for every data instance $\mathbf{x}_i$ is given as follows:

$$\gamma_{t+1}(\mathbf{x}_i) = \arg\min_{1 \leq l \leq m} \left(\frac{2}{n_l(t)} \sum_{\mathbf{x}_r \in C_l(t)} d^2(\mathbf{x}_i, \mathbf{x}_r) - \frac{1}{n_l^2(t)} \sum_{\mathbf{x}_r, \mathbf{x}_{r'} \in C_l(t)} d^2(\mathbf{x}_r, \mathbf{x}_{r'}) \right). \tag{2}$$

2.2 Density Sensitive Geodesic Distance

The geodesic distance is the fundamental element in the DGK-means algorithm imposing the global structure of the data. Through the sparse k-neighbourhood graph representation of the data, denoting the data points as the graph nodes and the corresponding pairwise distances as the edge-weights, the geodesic distance between two data points is given by the sum of the edge weights of the shortest path connecting them. By this means, the geodesic distance of points residing on different geometrical structures is of higher values compared to those from similar geometrical structures.

Geodesic distance based on pure Euclidean distance may be inaccurate in the presence of outliers. In order to reduce the effects of outliers, the DGK-means algorithm incorporates the outlierness profile of data into geodesic distance estimation. It achieves this by defining the graph edge weights through combining the pairwise distances of the data points with their local densities. In particular, DGK-means uses an exponential transfer model to compute graph edge weights ω_{ij} that adjusts the pairwise Euclidean distances based on the local densities of their end points:

$$\omega_{ij} = \exp\big(\frac{1}{\sigma^2} \max\big(\hat{f}(\mathbf{x}_i), \hat{f}(\mathbf{x}_j)\big)\big) \|\mathbf{x}_i - \mathbf{x}_j\|, \tag{3}$$

Here $\hat{f}(\mathbf{x}_i)$ is the local density estimate of point $\mathbf{x}_i$ with respect to its local neigborhood and is computed using the multivariate k-NN density estimator:

$$\hat{f}(\mathbf{x}_i) = \frac{k-1}{n \; vol(\mathbf{x}_i)} \tag{4}$$

This formulation produces robust geodesic distances in low dimensions, but it does not perform well in high dimensions. First, the geodesic distances on the

k-nearest neighbour graph based on Euclidean distance suffer from concentration problem in high dimensions and can not capture thoroughly the similarity of data points [4]. Second, the k-NN density estimator in the DGK-means algorithm can easily be affected by the curse of dimensionality. k-NN density estimator relies on the computation of the volume of the sphere to represent the far-end local neighbourhood of data which is numerically intractable with a low number of data samples in high dimensions.

2.3 Geodesic K-means

The DGK-means as per the general distance approach requires all intra-cluster distances to be computed and is computationally expensive. The DGK-means algorithm may be reformulated within a randomized baseline that eliminates the multiple invocation of intra-cluster pairwise distance computations. The DGK-means achieves this through a randomized process in which virtual cluster centroids are estimated over a random sample set of each data cluster at each iteration.

3 Outlier Robust Geodesic K-means Algorithm

The ORGK-means follows the same outline as that of DGK-means algorithm but introduces three particular improvements in the formulations of the pairwise distances, the density estimation and the weighting transform model.

3.1 Distance Metric Based on SNN

The proposed ORGK-means algorithm adopts an alternative strategy based on the concept of SNN similarity to compute pairwise distances. Distance measure based on SNN, also referred to as the secondary SNN-distance measure, have been shown efficient in high dimensional data space [5,9] that can perform well with data of different size, shape and varying distributions [3]. SNN-based similarity of two data points is the degree by which their underlying patterns overlap with one another. In terms of the sparse k-neighbourhood graph as described in Sect. 2.2, the SNN-similarity of two data points is seen as the number of points shared by the k-nearest neighbour lists of those points.

Given the set of k-nearest neighbours $\mathcal{N}_k(\mathbf{x}_i)$ and $\mathcal{N}_k(\mathbf{x}_j)$ of the points $\mathbf{x}_i$ and $\mathbf{x}_j$, the SNN similarity is given by the number of their common neighbours:

$$sim_{\mathcal{SSN}_k}(\mathbf{x}_i, \mathbf{x}_j) = |\mathcal{N}_k(\mathbf{x}_i) \cap \mathcal{N}_k(\mathbf{x}_j)|. \tag{5}$$

The normalized SNN similarity measure $simcos_k$ is defined as follows:

$$simcos_k(\mathbf{x}_i, \mathbf{x}_j) = \frac{sim_{\mathcal{SSN}_k}(\mathbf{x}_i, \mathbf{x}_j)}{k}. \tag{6}$$

Several SSN-based distance measures can be constructed based on the $simcos_k$ metric [5]. In this work, the SNN based inverse distance $dinv_k$ is utilized.

$$dinv_k(\mathbf{x}_i, \mathbf{x}_j) = 1 - simcos_k(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

3.2 Geodesic Based Local Outlier Factor

The ORGK-means algorithm utilizes a geodesic based local outlier factor ($gLOF$) to rank the outlierness of data. LOF, originally introduced by Breunig et al. [2], is an outlier scoring algorithm that relies on the k-nearest neighbour model and the notion of local reachability density. $gLOF$ enhances LOF with the geodesic distance.

As the same way in LOF, $gLOF$ benefits from several advantages. It is a non-parametric and model-free approach that does not make any assumption regarding the distribution of data. It is resistant to changes in density of data distribution patterns. In addition, $gLOF$ compared to LOF incorporates both local and global structure of the data and it provides richer outlier scoring scheme.

To compute the outlier score of an individual point, $gLOF$ compares its local density with the points in its neighbourhood. It takes the ratio of the local reachability density of the data point to the median local reachability density of its surrounding neighbours. This is different from the original LOF where the arithmetic mean is utilized to approximate the local neighbourhood reachability density. The idea in utilizing the median is to make the density estimator more robust to the outlying points. When an outlying point or a point belonging to other clusters is located in the neighbourhood, the mean is likely to misrepresent (underestimate) the dispersion of neighbouring local densities as they are dominated by the local density of that outlying point. In such cases, the median is considered a reasonable choice that is more robust to outliers.

Formally, the $gLOF$ score of a point $\mathbf{x}_i$ is given by:

$$gLOF_k(\mathbf{x}_i) = \frac{median\ \{lrd_k(\mathbf{x}_j)\}_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)}}{lrd_k(\mathbf{x}_i)}, \tag{8}$$

where $lrd(\mathbf{x}_i)$ describes the local reachability density of the point $\mathbf{x}_i$ over its local neighbourhood.

The local reachability density is loosely estimated as the inverse of the median of the reachability geodeisc distances to the point $\mathbf{x}_i$ from its neighbours. The local reachability density at point $\mathbf{x}_i$ is defined as follows:

$$lrd_k(\mathbf{x}_i) = \left(median\ \left\{ rd_k(\mathbf{x}_i, \mathbf{x}_j) \right\}_{\mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i)} \right)^{-1}, \tag{9}$$

where $rd_k(\mathbf{x}_i, \mathbf{x}_j)$ refers to the reachability geodesic distance from $\mathbf{x}_i$ to $\mathbf{x}_j$ given by:

$$rd_k(\mathbf{x}_i, \mathbf{x}_j) = max(R(\mathcal{N}_k(\mathbf{x}_i)), d_{G,W}(\mathbf{x}_i, \mathbf{x}_j)). \tag{10}$$

Here, $R(\mathcal{N}_k(\mathbf{x}_i))$ is the geodesic distance from $\mathbf{x}_i$ to its k-th nearest neighbour and $d_{G,W}(\mathbf{x}_i, \mathbf{x}_j)$ is the geodesic distance from $\mathbf{x}_i$ to $\mathbf{x}_j$. The geodesic distance is defined by the shortest path over the k shared nearest neighbourhood graph representation with the SNN based inverse distance metric $dinv_k$. The reachability distance $rd_k(\mathbf{x}_i, \mathbf{x}_j)$ is asymmetric and its role is to enhance the stability of results. It is defined to smooth out the statistical fluctuations in $d_{G,W}(.,.)$ when it is small compared to the distance of k-th neighbouring point. The larger the value of k, the higher smoothing is applied.

The $gLOF$ value of a point $\mathbf{x}_i$ located in a region of homogeneous density (inlier) is approximately 1, but it is of higher value if the density of the local neighbourhood of its neighbours is higher that the density of the local neighbourhood of the point itself (outlier).

3.3 Weighting Transform Model

The first-order exponential weighting model used in the DGK-means algorithm is designed to map density values within [0 1) interval onto [1 ∞). Such weighting model does not suit to the ORGK-mean algorithm where the outlierness of the data is ranked by $gLOF$ scores not limited to the range [0 1], and either the normalization is not straightforward. In addition, the model used in DGK-means strongly depends on the scaling parameter σ whose selection is not well defined [1].

The ORGK-means algorithm relies on a sigmoid function model to transform the outlier scores to the geodesic distances where the extreme outlier scores are eliminated. Specifically, the weighting transform model is built upon the double sigmoid function whose parameters can be adaptively tuned by the statistics of outlier score distribution in an ad hoc manner.

Denote the outlier scores of the points $\mathbf{x}_i$ and $\mathbf{x}_j$, obtained from $gLOF$ model, by s_i and s_j respectively. The proposed weighting function is given by:

$$\omega_{ij} = \begin{cases} \left(1 + \beta \; exp \left[-2 \; \frac{max(s_i, s_j) - \tau}{\alpha_1} \right] \right)^{-1} & \text{if } s < \tau, \\ \left(1 + \beta \; exp \left[-2 \; \frac{max(s_i, s_j) - \tau}{\alpha_2} \right] \right)^{-1} & otherwise \end{cases} \tag{11}$$

where β is a scaling parameter and τ is the threshold parameter beyond which a data point is considered as an outlier. α_1 and α_2 are edge parameters that define the region $[\tau - \alpha_1 \; \tau + \alpha_2]$ at which the weighing function is approximately linear.

Choosing a proper value for the threshold parameter τ is highly dependent on the data as $gLOF$ produces a varying range of scores relative to the underlying local and global structures of data. However, since the scores obtained by $gLOF$ are typically positively skewed, the value of parameter τ can be set to the mean of the score distribution. In positively skewed distribution, the mean pulls toward the direction of skew (the direction of the outliers) and therefore can provide an approximate basis for the decision about the outlierness of data.

The mean of the score distribution can reasonably approximate the maximal inlier score value but it can be of too large values if there are erratic deviations in score values or the distribution is extremely skewed. To address this issue, the mean is estimated over truncated data such that a certain percentage of data, o_p, corresponding to the largest scores is discarded. The mean obtained in this manner resembles the truncated mean estimator that is less sensitive to extreme outliers.

Inspired by the definition of skewness as the measure of asymmetry about the mean, the value of α_1 can be set to $truncmean_{o_p}(s) - mode(s)$ and the value of α_2, not as critical as τ and α_1, can be set to any value lower than $max(s) - truncmean_{o_p}(s)$.

4 Experimental Results

Evaluation of the proposed ORGK-means algorithm for clustering in high dimensions was carried using synthetic and real remote sensing hyperspectral data.

4.1 Test Data

Synthetic Data: To investigate the ability of the proposed ORGK-means algorithm several datasets of different dimensionality were generated. The datasets were constructed to generate unbiased data sampled from the Gaussian distribution over an increasing range of dimensions that share certain attributes, [2, 4, 8, 16, 32, 64, 128, 256, 512]. Outliers were uniformly scattered to the space with the ratio of 5 % and 10 % of the total number of samples. Inspired by [5], data series were generated from three classes of *8-Relevant*, *Half-Relevant* and *All-Relevant* each differing in the portion of informative variables that are relevant to clusters. Each dataset contained 500 samples that are uniformly divided into 7 clusters whose mean and variance were uniformly randomized ensuring data comes from well separated clusters of various distributions and that every pair of cluster has 10 % overlap at most.

Real Data: Two different benchmark hyperspectral image datasets were used for experiments: *Botswana* and *SalinasA*. The *Botswana* dataset was acquired by NASA EO-1 satellite over the Okavango Delta, and the corrected version of the data includes 145 bands. 7 out of the 14 classes with an equal number of samples were chosen in the experiment. 1575 data samples of dimension 145 are distributed to 7 classes. The *SalinasA* dataset was collected by AVIRIS over Salinas Valley in southern California, USA. The test hypercube consists of 7138 samples of dimension 204 comprising of 6 different classes[1].

4.2 Results and Discussion

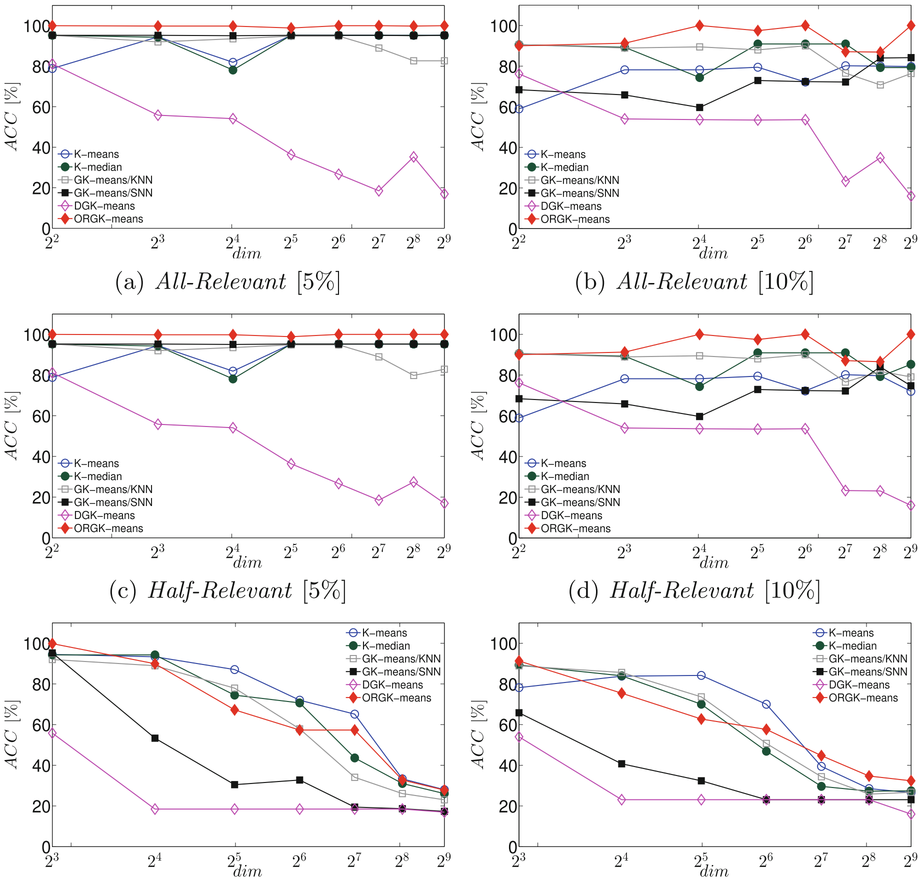
The performance of the proposed ORGK-means was compared to a number of methods including classic K-means, k-NN based geodesic K-means, SNN based geodesic K-means and density based geodesic K-means, respectively denoted as K-mean, GK-means/kNN, GK-means/SNN and DGK-means. To ensure fair evaluation, all the algorithms were experimented with the pre-known number of clusters and identical randomly initialized data-to-cluster assignments. Given the number of clusters as well as the initial data-to-cluster assignments, the proposed ORGK-means algorithm requires the parameters the top percentage of outliers o_p and the number of nearest neighbours k to be specified.

[1] These datasets are available at http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

The top outlier percentage o_p defines the percentage of data points that are considered as the extreme outliers to be excluded. It can be set empirically by a domain expert or can be approximated by the analysis of the outlier score distribution. Here, in the case of the synthetic datasets, the value for o_p was chosen based on the original percentage of added outliers and in the case of the real datasets, it was chosen empirically by searching the range from 1 % to 10 %.

The number of nearest neighbours k defines the neighbourhood size in the NN-model and significantly affects the performance of computing geodesic distances and outlier scores. The number of nearest neighbour versus overall accuracy was searched within the range form 5 to 150 and the best cases were only reported for the methods used.

Performance comparisons of the proposed ORGK-means algorithm and the other methods applied to synthetic data are shown in Fig. 1, giving the



(a) *All-Relevant* [5%]

(b) *All-Relevant* [10%]

(c) *Half-Relevant* [5%]

(d) *Half-Relevant* [10%]

(e) *8-Relevant* [5%]

(f) *8-Relevant* [10%]

Fig. 1. Clustering performance comparison on the synthetic dataset over increasing dimensionality; (a–b) *All-Relevant (%)*, (c–d) *Half-Relevant*, (e–f) *8-Relevant*.

overall accuracy versus the scaling dimension. Simulation results on *All-Relevant* and *Half-Relevant* data series show the improvement introduced by ORGK-means over the other competing methods, specifically GK-means/kNN, GK-means/SNN and DGK-means. However, there are a few exceptions in which the K-median stands as the out-performer. This observation can be explained by the synthetic data having been generated from an isotropic Gaussian mixture model residing on linear space where K-means equipped with geodesic distance does not necessarily yield higher separability power.

The results on *8-Relevant* as expected show that the performance of all compared clustering algorithms affected as dimension grows. This is to confirm the hampering effects of irrelevant attributes on the distinguishability of data clusters present in high dimensions. This observation also indicates that the ORGK-means algorithm, similarly to the other compared methods, is not able to handle high dimensional data when the feature space is dominated by irrelevant variables.

Figures 2 and 3 show the clustering maps obtained by the competing methods on real hyperspectral test data, *Botswana* and *SalinasA*, respectively. Overall, in both cases, but notably in *SalinasA*, the proposed ORGK-means gave the best results among the competitors, reaching a higher separation rate, though some data points are assigned as outliers.

Table 1 summarises the clustering accuracies, in terms of overall accuracy (ACC) and macro average Positive Predictive Value (PPV_m), achieved by the proposed ORGK-means algorithm and the other methods. The results confirm the superior performance of ORGK-means over other the methods in terms of both ACC and PPV_m.

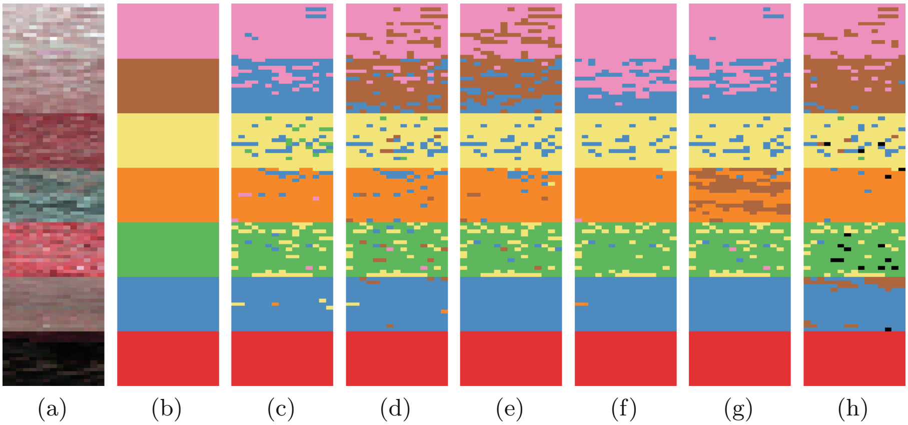


Fig. 2. Clustering performance comparison on *Botswana* test data; (a) RGB rending, (b) ground truth (c) K-means, (d) K-median, (e) GK-means/kNN, (f) GK-means/SNN (g) DGK-means and (h) ORGK-means.

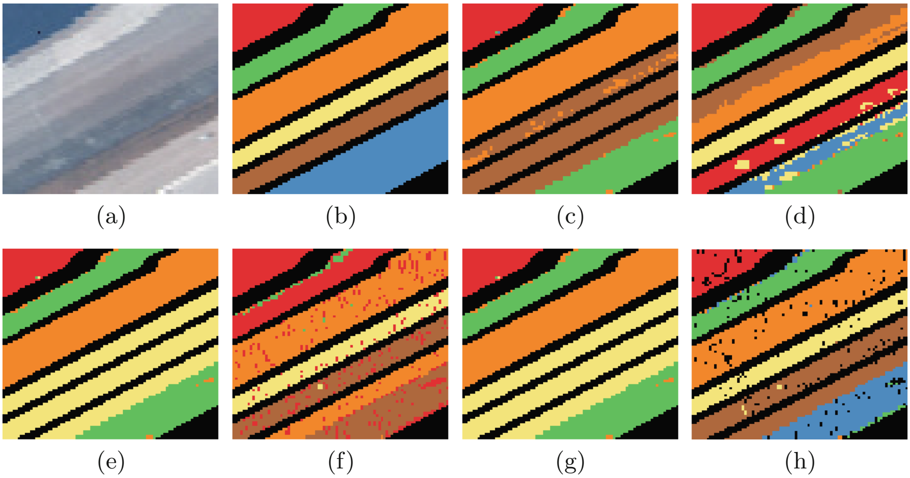


Fig. 3. Clustering performance comparison on *SalinasA* test data; (a) RGB rending, (b) ground truth (c) K-means, (d) K-median, (e) GK-means/kNN, (f) GK-means/SNN (g) DGK-means and (h) ORGK-means.

Table 1. Highest overall clustering ACC and macro averaged PPV (in percentage) obtained by the considered clustering algorithms.

	Botswana			SalinasA		
Methods	k	ACC [%]	PPV_m [%]	k	ACC [%]	PPV_m [%]
K-means	-	77.8	71.5	-	61.8	46.0
K-median	-	86.2	87.7	-	52.7	58.7
GK-means/kNN	75	84.2	87.1	10	59.4	45.5
GK-means/SNN	125	81.5	73.2	10	59.0	57.5
DGK-means	15	74.2	73.1	10	32.4	45.5
ORGK-means	125	89.3	90.9	50	83.8	91.3

5 Conclusion

In this paper, an outlier robust geodesic K-means (ORGK-means) algorithm is proposed for clustering of high dimensional data. The proposed ORGK-means extends the standard K-means algorithm by using an outlier-adjusted geodesic distance. In the proposed ORGK-means algorithm, SNN based distance metric is utilized as the pairwise dissimilarity measure. Geodesic based LOF, exploiting both local and global structural information, is introduced to rank the degree of outlierness, and an adaptive weighting transform model based on the double sigmoid function is proposed to adjust geodesic distances. The efficiency of the proposed ORGK-means algorithm in clustering high-dimensional data was evaluated using synthetic and real world remote sensing spectral data. The numerical results on the overall clustering accuracy and the average precision showed the utility of the proposed algorithm.

References

1. Asgharbeygi, N., Maleki, A.: Geodesic k-means clustering. In: 19th International Conference on Pattern Recognition 2008, pp. 1–4. IEEE, Tampa, December 2008
2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD 2000, pp. 93–104. ACM, New York (2000)
3. Ertöz, L., Steinbach, M., Kumar, V.: Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Proceedings of Second SIAM International Conference on Data Mining. SIAM (2003)
4. Francois, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Trans. Knowl. Data Eng. **19**(7), 873–886 (2007)
5. Houle, M.E., Kriegel, H.-P., Kröger, P., Schubert, E., Zimek, A.: Can shared-neighbor distances defeat the curse of dimensionality? In: Gertz, M., Ludäscher, B. (eds.) SSDBM 2010. LNCS, vol. 6187, pp. 482–500. Springer, Heidelberg (2010). doi:10.1007/978-3-642-13818-8_34
6. Jarvis, R., Patrick, E.A.: Clustering using a similarity measure based on shared near neighbors. IEEE Trans. Comput. **C–22**(11), 1025–1034 (1973)
7. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, chap. 2, pp. 68-125. Wiley (2008)
8. Moëllic, P.A., Haugeard, J.E., Pitel, G.: Image clustering based on a shared nearest neighbors approach for tagged collections. In: Proceedings of the 2008 International Conference on Content-Based Image and Video Retrieval, CIVR 2008, pp. 269–278. ACM, New York (2008)
9. Tomasev, N., Mladeni, D.: Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification. Knowl. Inf. Syst. **39**(1), 89–122 (2014)
10. Wang, D., Ding, C., Li, T.: K-subspace clustering. In: Buntine, W., Grobelnik, M., Mladenić, D., Shawe-Taylor, J. (eds.) ECML PKDD 2009. LNCS (LNAI), vol. 5782, pp. 506–521. Springer, Heidelberg (2009). doi:10.1007/978-3-642-04174-7_33
11. Wu, J.: Cluster analysis and k-means clustering: an introduction. In: Wu, J. (ed.) Advances in K-means Clustering. Springer Theses, pp. 1–16. Springer, Heidelberg (2012)
12. Yin, J., Fan, X., Chen, Y., Ren, J.: High-dimensional shared nearest neighbor clustering algorithm. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 494–502. Springer, Heidelberg (2005). doi:10.1007/11540007_60
13. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Stat. Anal. Data Min. **5**(5), 363–387 (2012)