A Distance Metric that Combines Linkage, Connectivity and Density Information for Clustering in Image Processing

Priyanka Khandelwal
M.Tech,Dept. of Computer Science and Engineering
Rajasthan College Of Engineering For Women, Jaipur, Rajasthan, India

Sonal Saxena
Asst.Prof., Dept. of Computer Science and Engineering
Rajasthan College Of Engineering For Women, Jaipur, Rajasthan, India

ABSTRACT

Rapid development in Internet technology has generated data at high velocity in large volume and variety. It needs newer methods of analysis. Combining traditional and popular methods with specialized techniques give interesting clustering outputs and are of much use in some real life applications. This paper suggests a new dissimilarity metric to handle complex data. It combines the linkage and density information of data together. Multi-dimensional scaling summarizes the data model based on the proposed distance metric to use it for image processing. The low dimensional model obtained after dimensionality reduction can be easily clustered using standard algorithms.

Keywords

Clustering, Distance Metrics, Multi dimensional Scaling, Ensembling, density-based clustering, linkage, image processing

1. INTRODUCTION

Clustering aims to group similar data points together in clusters with no similarity between data points of different clusters and leaves behind outliers or points not belonging to any of the clusters. In conventional clustering output of clusters depends completely on distance metrics and detects regular shaped clusters like Euclidean distance metric detects only spherical shaped clusters; Manhattan detects rhombus shaped clusters; Mahalanobis detects elliptical clusters and Chebychev detects square shaped clusters. In order to identify arbitrary shapes clusters density metrics are used as proposed in DENCLUE [1], DDC [2], Chen and He's work [3]. Or some graph-theoretic approach is used to convert dataset into graphs and then discovering connected components within it , like in PKNNG[4].

Sometimes due to randomization in the clustering algorithm, it is not fit enough to produce a good quality output by itself. So in order to minimize the bad effect of randomization ensembling is used. Ensembling runs an algorithm multiple times on a cluster and group the results in order to significantly improve the quality of the clustering results. Ensembling comprises of two steps, namely generation and consensus, the ensemble approach first generates a set of partitions by the different clustering algorithms and then combines the results into a single resultant partition. Ensembling technique is used by Strehl and Ghosh [5], Vega-Pons and Ruiz-Shulcloper [6] , Fred and Jain [7].

Clustering result also gets effected by the dimensionality of data. As some attributes of data are very important for clustering and some are not applicable for the desired application. So in order to get the desired cluster output we need to use techniques that reduce the dimensionality of the data for clustering. Like in Principal Components Analysis (PCA) [8] constructs a low-dimensional mapping of the high-dimensional input space by preserving the variances of the data points and Multi-Dimensional Scaling (MDS) [9] does the same but uses the inter-point distances, instead of the variances for the mapping.

This paper proposes how dissimilarity metric can combine the information from graph-theory and density based techniques. Such dissimilarity can then be used with any conventional clustering algorithms. If the clustering algorithm at the user end requires a dissimilarity equivalent to a geometric distance, as needed by k-means [10], then dimension reduction is used to convert the proposed dissimilarity into a distance metric.

The paper is organized as follows. Section II discusses the background techniques that form a part of proposed work flow. Section III describes the PKD [11] and proposed protocols. Section IV illustrates through experiments on real life datasets how the proposed algorithm can be used for image processing. Last two sections conclude the paper and list some future directions.

2. BACKGROUND

2.1 Linkage Clustering

Linkage clustering is a hierarchical clustering method that creates a hierarchical tree structure from observations in the input dataset. The pairwise distances between data points are computed internally by one of the metrics accepted by the approach. Linkage clustering methods differ from each other in terms of the measure they use for finding distances. Considering the following notations, the linkages for use by the above discussed methods are then described. Cluster r is formed from cluster p and q, n_r is the number of objects in cluster r and x_{ri} is the i^{th} object in cluster r.

2.1.1 Unweighted Average Distance (UPGMA) Approach: Referred to as Average linkage method. The approach considers the average distance between all object pairs in any two clusters. The equivalent distance is

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \qquad (1)$$

2.1.2 Centroid Distance (UPGMC) Approach: Referred to as Centroid linkage method. The method considers Euclidean distance between the centroids of the two clusters. The equivalent distance is

$$d(r,s) = \|\overline{x_r} - \overline{x_s}\|_2 \quad (2)$$

Where, $x_r = \frac{1}{n_r}\sum_{i=1}^{n_r} x_{ri}$

2.1.3 Furthest Distance Approach: Referred to as Complete linkage method the method considers the largest distance between objects in the two clusters. The equivalent distance is

$$d(r,s) = \max\left(dist(x\_ri, x\_sj\ )\right), \quad (3)$$

For, $i \in (i, \dots\dots, n_r), j \in (1, \dots\dots\dots, n_s)$

2.1.4 Weighted Center of Mass Distance (WPGMC) Approach: Referred to as Median linkage method, it considers Euclidean distance between weighted centroids of the two clusters. The equivalent distance is

$$d(r,s) = \|\widetilde{x_r} - \widetilde{x_s}\|_2 \quad (4)$$

For weighted centroids $\widetilde{x_r}, \widetilde{x_s}$ of clusters r and s respectively. Combining clusters p and q for cluster r,

$\widetilde{x_r}$ is defined as $\widetilde{x_r} = \frac{1}{2}\left(\widetilde{x_p} + \widetilde{x_q}\right)$

2.1.5 Shortest Distance Approach: Referred to as Single linkage method it considers the smallest distance between the objects in the two clusters. The equivalent distance is

$$d(r,s) = \min\left(dist\left(x_{ri}, x_{sj}\right)\right), \quad (5)$$

For, $i \in (i, \dots\dots, n_r), j \in (1, \dots\dots\dots, n_s)$

2.1.6 Inner Squared Distance (Minimum variance) Approach: Referred to as Ward linkage method, it uses the incremental sum of squares, that is, the increase in the total within-cluster sum of squares as a result of joining two clusters. The within cluster sum of squares is defined as the sum of squares of the distances between all objects belonging to the cluster and the centroid of the cluster. The equivalent distance is

$$d^2(r,s) = n_r n_s \frac{\|\widetilde{x_r} - \widetilde{x_s}\|_2{}^2}{n_r + n_s} \quad (6)$$

Where, $\|.\|_2$ is Euclidean distance, $\widetilde{x_r}, \widetilde{x_s}$ are weighted centroids of clusters r and s respectively.

2.2 Penalized k-nearest Neighbor Graph

Often dataset can be represented as a graph where nodes are the objects and edges between them are weighted with the labels equal to distance/dissimilarity between them. Since distances occur among every pair, it is a complete graph. A k-nearest neighbor (knn) graph is constructed from this by retaining edges at every node to only the k-nearest neighbors and deleting the rest. If the complete graph is considered to be directed, then for any edge (x,y), if the reciprocal (y,x) gets removed then (x,y) is also removed. Thus, only edges of type (x,y) are retained where object x is k-nearest of object y and vice versa. So we get a disconnected graph of strongly connected components. Authors in [4] suggest that the edges which have outstandingly high values should also be removed to curtail the outliers. Particularly, edges with weight more than the 3rd quartile plus 1.5 times interquartile distance should be removed. Now, some new edges are introduced with penalized weights. The penalized weight is

$$w = de^{d/\mu}$$

Where d is original distance and μ is the mean of distances. The penalized edges are added to make the graph either minimally connected or completely connected.

3. PROPOSED CLUSTERING METHOD

3.1 PKD Approach for Clustering

The method suggested by Baya et al [11] has been named as PKD and is actually a connectivity and density measure. To use this metric for clustering purposes, it is combined with MDS and then input to any standard clustering algorithm. The flow of process is shown in Fig. 1. The steps of this process are described below.

Step 1: Evidence Accumulation

The method suggested by Fred and Jain [7] is used in this step. The idea is to consolidate information gained from the output of different clustering process over same data. Instead of clustering processes, k-means [10] with random initialization is used. Due to random initialization, the output may vary. It is run M number of times over entire dataset X. Suppose, in the final output, the desired number of clusters is ac, then c>ac is specified as the number of output clusters in this first step. Such large number of clusters ensures detection of arbitrary shapes. Output of this step is a single clustering decision which combines all M clustering decisions using a simple frequency counting method. But, in the PKD process flow, a similarity matrix is created from the information gathered in M clustering decisions. Let this density information be denoted as T, with each element t_{ij} computed as

$$t_{ij} = \frac{N_{ij}}{M}$$

where N_{ij} is the number of times objects x_i and x_j belong to the same cluster as per the M outputs.

Step 2: Constructing a distance metric

The similarity matrix obtained in the previous step is combined with the original Euclidean distances among the data objects to prepare a new distance metric. Let this dissimilarity matrix be denoted as D. Then each element d_{ij} is computed as simple division

$$d_{ij} = \frac{d_{ij}^{euc}}{t_{ij}} \quad (6)$$

Where, d_{ij}^{euc} is the Euclidean distance between objects x_i and x_j in the original dataset. It is easily observed that D→∞ when T→0. Hence, all values of T are bound by a small constant α by replacing all the values lower than α by α.

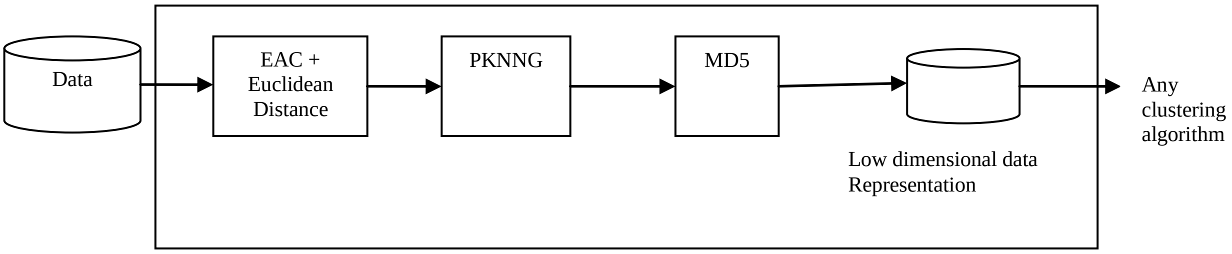

Fig.1 Working of Baya et al's work

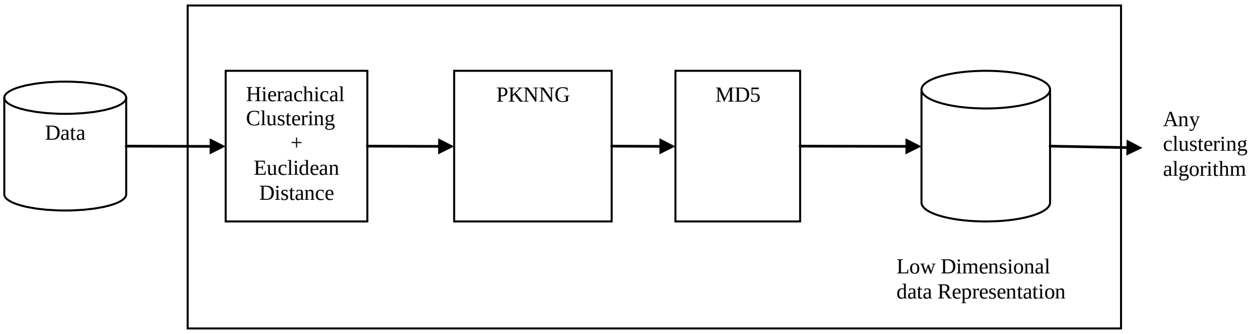

Fig.2 Working of the proposed algorithm

Step 3: Computing PKNNG metric

Output of the previous step satisfies all the conditions of geometric distances and hence is directly used as an input to the PKNNG module. The output is again a transformed dissimilarity matrix D'. This has the connectivity information also in addition to the previous density information.

Step 4: Multi-Dimensional Scaling

The final decision metric D' is treated like a dataset in itself and input to the MDS module to obtain a low dimensional representation of the data. Let this data be denoted as X'.

Step 5: Clustering

Any standard clustering algorithm which works on the concept of geometric distances or linkages is used in this step for final clustering decision. The inputs are the transformed data X' and desired number of clusters ac.

3.2 Proposed Dissimilarity Metric

Existing work by Baya et al [11] combines two different concepts of dissimilarities and uses concept of ensembles to produce good quality clusters in high dimensional data of non-Gaussian distribution. But the method uses k-means in two of its phases which may make the method limited by the performance of k-means algorithm in itself. Hence, we aim to experiment other kind of approaches for clustering in the PKD approach. The proposal is to substitute a basic partition based clustering by a hierarchical clustering process. The simplest linkage based hierarchical clustering is considered in the first step of the PKD flow structure in place of k-means. Rest all the process is kept same. The proposed work flow is given in Fig 2.

4. APPLICATION OF PROPOSAL FOR IMAGE DATA PROCESSING

To illustrate how the proposed clustering method can be used with image data, we have taken datasets of MNIST Dataset [12] consists of handwritten digits. A set of 60000 training and 10000 testing images are contained in the dataset. The number of image objects is too large, hence for experimental purpose, we pick 200 images of each digit and merge them into a single dataset of 2000 images. Olivetti Face Dataset [13] consists of ten different images of each of the 40 distinct subjects. We use both the forms of this dataset separately as two individual datasets. The first dataset represents the 32X32 converted images and the second dataset consists of 64X64 images. The Pen-Based Recognition of Handwritten Digits Dataset [14] consists of 250 digit samples from 44 writers. The total instances in the dataset are 10992 and total classes are 10 of unequal distribution of instances for each of the training and testing part. For practical purposes, 10% random samples of each class are picked such that the original distribution of each class is maintained.

The Performance of the proposed algorithm and the PKD method of Baya et al [11] are compared as MATLAB programs. The parameters used in the experiments for the different datasets along with the other quick details are listed in Table 1, as number of instances (n), number of features (d), k of PKNNG step, c is number of clusters formed in EAC step of PKD, c' as number of maximum clusters formed in linkage step of proposed algorithm, ac is desired number of clusters, M is number of times clustering is repeated for accumulation, and the thresholding value (α**)**.

Table 1. Parameters of different datasets used

Dataset	n	d	k	c	c'	ac	M	α
Olivetti (32x32)	400	1024	17	100	40	40	30	0.1
Olivetti (64x64)	400	2048	17	100	40	40	30	0.1
MNIST	2000	784	17	10	10	10	10	0.1
Pendigits	1105	17	17	10	10	10	20	0.1

The output dissimilarity matrix of both PKD and the proposed methods are inputs to same clustering algorithm for appropriate comparison. For ease of experiments and evaluation of output, we use k-means with random initialization as the final clustering algorithm.. The obtained clustering labels are compared with the ground-truth clustering labels using the Adjusted Rand Index (ARI).

The implemented methods are represented through numbers instead of names in the figures of boxplots. The notations are as listed in Table 2.

Table 2 .Methods used and their notations

Method	Description
1	PKD Clustering
2	Proposal with Complete Linkage
3	Proposal with Average Linkage
4	Proposal with Ward Linkage
5	Proposal with Single Linkage
6	Proposal with Median Linkage
7	Proposal with Centroid Linkage
8	Proposal with Weighted Linkage

4.1 Results on Olivetti Face Dataset

The boxplots in Fig. 3 and 4 list the results for Olivetti Face Dataset of 400 images of 32x32 and 64x64 pixel resolutions respectively using different clustering algorithms. X-axis representing the method used and y-axis denote the ARI values.. The boxplot shows a superior performance for proposal with Ward Linkage method. Superior performance implies that the clusters can match the actual classes more accurately. Variation in results is very less for this and the red line marking the average values is well above than red lines of other methods. The performance of Ward method is similar for Olivetti Face Dataset of both pixel resolutions.

In addition, proposal with Average, Complete and Weighted Linkage methods give enhanced performances though the variability factor is high, especially for Average Linkage method.

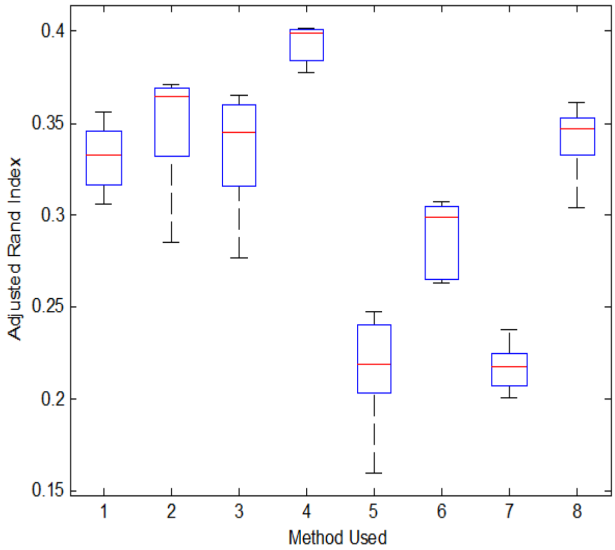

Fig.3 Variation in Adjusted Rand Index values of cluster outputs for Olivetti Face Dataset (32x32)

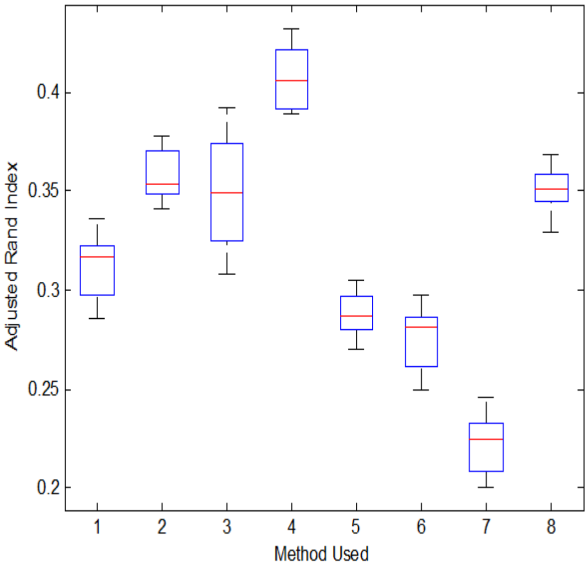

Fig.4 Variation in Adjusted Rand Index values of cluster outputs for Olivetti Face Dataset (64x64)

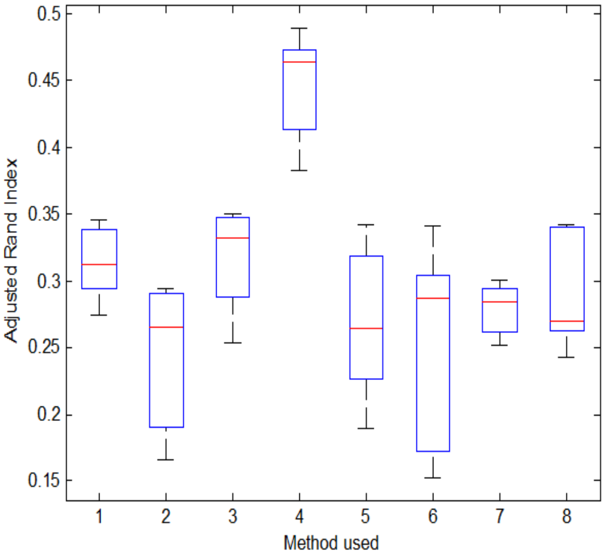

Fig.5 Variation in Adjusted Rand Index values of cluster outputs for MNIST Digits Dataset

4.2 Results on MNIST digits dataset

The boxplots in Fig. 5 list the results for MNIST Digits Dataset of 2000 images using different clustering algorithms. The best performance is portrayed in case of Ward Linkage method combined with proposal. The performance however is not consistent as implied through high variation (box plot is large). The performances of the remaining methods are more or less same and considerably less than that of proposal with Ward Linkage Method.

4.3 Results on PenDigits Dataset

The boxplots in Fig. 6 list the results for the processed PenDigits Dataset. Superior performance is shown in case of proposal with Average Linkage method. The variability factor is high indicating that the performance is inconsistent. Yet, the average line (red) of methods 2,3 and 4 are nearer to maximum value of ARI achieved. This indicates that average performance is better in these.

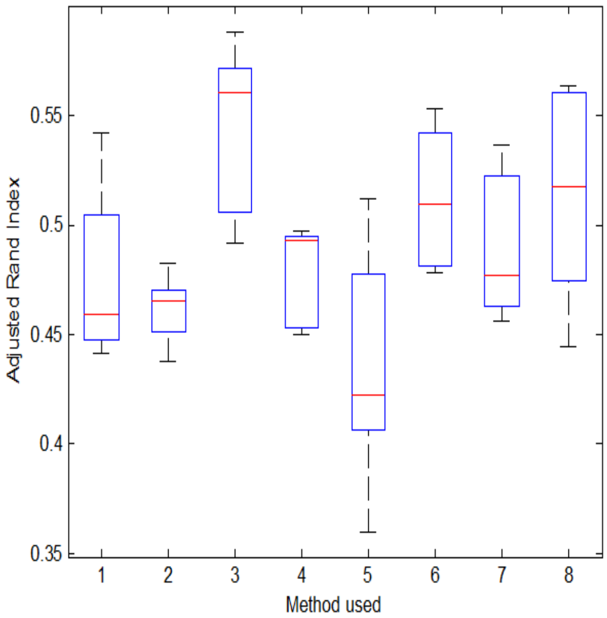

Fig.6 Variation in Adjusted Rand Index values of cluster outputs for Pendigits Dataset

5. CONCLUSIONS

This paper proposes an approach similar to PKD by Baya et al [11]. But they used k-means at both ends of the process. Since k-means has a performance bias towards Gaussian distribution, we have replaced it with many variants of hierarchical clustering, viz Ward, Centroid, Median, Average Linkage methods and more. A combined distance metric is obtained from ensembles and k-nearest neighbour weighted graphs. This metric is used to construct a dissimilarity matrix which is summarized using Multi-Dimensional Scaling to obtain a dissimilarity that follows distance properties. Hence, proposed distance metric can be used in any conventional clustering method like k-means. Experiments over few real life image datasets are conducted to verify how effective the proposal is in identifying images belonging to same group or bearing similarity to each other. The results are compared to those obtained through PKD. Proposed technique is better than PKD for image processing purpose. The results also compare the performance of proposal on the various used linkage based clustering techniques, out of which Ward's method proves out to be the most effective in almost all cases.

6. FUTURE SCOPE

As an extension to the current work, the effectiveness of the proposal can be checked for data other than image data. A comprehensive study of effect of using the proposed distance metric in popular clustering algorithms can be undertook. The current proposal uses ensembling as accumulation of results of several runs of same algorithm. Instead ensembles can be constructed by using output from more than one algorithm. Thus, the proposal provides many phases and steps which can be individually analyzed and improved.

7. REFERENCES

[1] A. Hinneburg. and D. Keim, "An efficient approach to clustering large multimedia databases with noise", *Proceedings of the 4th ACM SIGKDD Conference*, 1998, pp. 58-65.

[2] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks", Science, Vol. 344, 2014, pp. 1492–1496.

[3] J. Y. Chen and H. H. He, "A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data", Information Sciences, Vol. 345, 2016, pp. 271–293

[4] A. E. Bayá and P. M. Granitto, "Clustering gene expression data with a penalized graph-based metric", *BMC Bioinformatics*, vol. 12, no. 1, 2011.

[5] A. Strehl and J. Ghosh, "Cluster ensembles: A knowledge reuse framework for combining multiple partitions", Journal of Machine Learning Research, vol. 3, pp. 583-617, 2002.

[6] S. Vega-Pons and J. Ruiz-Shulcloper, "Clustering ensemble method for heterogeneous partitions", eds. E. Bayro-Corrochano and J.-O. Eklundh, Proceedings CIARP 2009, Vol. 5856 of Lecture Notes in Computer Science, 2009, pp. 481-488.

[7] A. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, 2005, pp. 835–850.

[8] S. Mika, B. Schölkopf, A. Smola, K-R Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces", Proceedings Of International Conference On Advances In Neural Information Processing Systems, 1999, pp. 536–542.

[9] T. F. Cox and M. A. A Cox, Multidimensional scaling (2nd ed.), Chapman & Hall/CRC, 2000.

[10] E. W. Forgy, "Cluster analysis of multivariate data: efficiency v/s interpretability of classifications", Biometrics, Vol. 21, 1965, pp. 768–769.

[11] A. E. Bayá, M. G. Larese and P. M. Granitto, "Clustering using PK-D: A connectivity and density dissimilarity", Expert Systems with Applications, Vol. 51, pp. 151–160, 2016.

[12] Y. Lecun and C. Cortes, "The MNIST database of handwritten digits", 1998. http://yann.lecun.com/exdb/mnist. Accessed June 2015.

[13] AT&T Labs, "The Olivetti faces dataset", Cambridge: AT&T Lab, 1992. http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. Accessed June 2015.

[14] F. Alimoglu, "Combining Multiple Classifiers for Pen-Based Handwritten Digit Recognition", MSc Thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University, 1996.