In Proc. Int'l Joint Conference on Neural Networks, 2007 (IJCNN 2007), Orlando, FL, Aug 12-17, 2007

A new cluster validity index for prototype based clustering algorithms based on inter- and intra-cluster density

Kadim Taşdemir and Erzsébet Merényi, *Senior Member, IEEE*

***Abstract*— One of the fundamental challenges of clustering is how to evaluate, without auxiliary information, to what extent the obtained clusters fit the natural partitions of the data set. A common approach for evaluation of clustering results is to use validity indices. We propose a new validity index, $Conn\_Index$, for prototype based clustering. $Conn\_Index$ is applicable to data sets with a wide variety of cluster characteristics (different shapes, sizes, densities, overlaps). We construct $Conn\_Index$ based on inter- and intra-cluster *connectivities* of prototypes, which are found through a weighted Delaunay triangulation called "connectivity matrix" [1], where the weights indicate the data distribution. We compare the performance of $Conn\_Index$ to commonly used indices on synthetic and real data sets.**

I. Introduction

Clustering means splitting a data set into groups such that the data samples within a group are more similar to each other than to the data samples in other groups. Clustering is done with many methods which can be categorized in several ways where the two major ones are partitioning and hierarchical clustering. For any method, clustering the data directly becomes computationally heavy as the size of the data set increases. In order to significantly reduce the computational cost, two-step algorithms have been proposed [2], [3], [4], [5]. Two-step algorithms (prototype based clustering) first find the quantization prototypes of data, and then cluster the prototypes. Using the prototypes instead of data can also reduce noise because the prototypes are the local averages of the data.

A widely and successfully used neural paradigm for finding prototypes is the Self-Organizing Map (SOM). The SOM is a spatially ordered quantization of a data space where the quantization prototypes are adaptively determined for optimal approximation of the (unknown) distribution of the data. The SOM also facilitates visualization of the structure of a higher-dimensional data space in one or two dimensions, which can guide semi-manual clustering. Thus, the SOM is a powerful aid in capturing clusters in high-dimensional intricate data sets [1], [2], [3], [6].

With any clustering method, whether clustering the data itself or its prototypes, the main problems are to determine the number of clusters and to evaluate the validity of the clusters. A validity measure of the clustering ideally shows how well the obtained clusters fit the natural partitions of the data set, without any a priori class information. Cluster validity approaches can use three criteria: *external criteria* (evaluate the result with respect to a pre-specified structure), *internal criteria* (evaluate the result with respect to a proximity matrix of the data vectors), and *relative criteria* (evaluate the validity by comparing it to other clustering results) [7]. Many different methods exist for investigations of the validity of crisp clustering [8], [9], [10], [11], [12] or fuzzy clustering [13], [14], [15]. Due to space constraints we refer the reader to [7], [10], [11], [12] for crisp clustering and to [7], [15] for fuzzy clustering for detailed analysis of the cluster validity indices. For crisp clustering, the Davies-Bouldin index [8] and the generalized Dunn Index [10] are some of the most commonly used. Both depend on a *separation* measure between clusters and a measure for *compactness* of clusters based on distance. Even though these two indices work satisfactorily for well-separated clusters, they may fail for complicated data structures with clusters of different shapes or sizes or with overlapping clusters. When the clusters have homogeneous density distribution, one effective approach to correctly evaluate the clustering of data sets is CDbw (composite density between and within clusters) [16]. CDbw finds prototypes for clusters instead of representing the clusters by their centroids, and calculates the validity measure based on inter- and intra-cluster densities, and cluster separation. The densities are calculated as the number of data samples within a standard deviation from the prototypes. However, it fails to represent true inter- and intra-cluster densities when the clusters have inhomogeneous density distribution.

Our objective is to define a validity index that can be used successfully for any data set with overlapping clusters, with varying cluster statistics or with clusters of different shapes or sizes. We introduce a new validity index, $Conn\_Index$, based on inter- and intra-cluster densities of the prototypes. These densities are found through a weighted Delaunay triangulation (*connectivity matrix*, [1]) where the weights indicate the data distribution between the neighbor prototypes.

In order to evaluate the effectiveness of $Conn\_Index$ we use three synthetic data sets with different properties and four real data sets: Breast Cancer Wisconsin (10-dimensional), Iris (4-dimensional), Wine (13-dimensional) and Ocean City, an 8-band remote sensing spectral image. Their prototypes are clustered with various methods: k-means, single linkage; and two semi-manual clusterings of the SOM prototypes. We describe $Conn\_Index$ in Section II and give examples of its performance on synthetic data sets in Section III. In

Authors are at Electrical and Computer Engineering Department, Rice University, Houston, TX 77005, USA, e-mail: (tasdemir,erzsebet)@rice.edu

This work was partially supported by grant NNG05GA94G from the Applied Information Research Program of NASA, Science Mission Directorate.

Many thanks to Dr. Bea Csathó, Byrd Polar Institute, Ohio State University, for the Ocean City image and ground truth.

Figures in this paper are in color. Request color reprint from (tasdemir,erzsebet)@rice.edu.

Section IV we apply $Conn\_Index$ to the real data sets. We summarize our conclusions in Section V.

II. $Conn\_Index$: A VALIDITY INDEX BASED ON INTRA- AND INTER-CLUSTER DENSITIES

A. Density representation by cumulative adjacency matrix ($CADJ$) and connectivity strength matrix ($CONN$)

The first step of prototype based clustering algorithms is to find the prototypes of the data vectors by a process such as neural networks, hierarchical clustering algorithms, etc. Each prototype is the best matching unit (BMU) for the data vectors in its receptive field RF (Voronoi polyhedron). The sizes of the receptive fields indicate how the data is distributed among the prototypes. To indicate the similarity of a prototype to others, we introduce a weighted Delaunay triangulation (*cumulative adjacency matrix* $CADJ$, *connectivity matrix* $CONN$ [1]) where the weights correspond to the data distribution among the neighbor prototypes. $CADJ$ (directed weighted Delaunay triangulation) and $CONN$ (weighted Delaunay triangulation) are defined in [1] as follows:

Definition 1: *Let $CADJ$ be an $N \times N$ matrix where N is the number of prototypes. The cumulative adjacency, $CADJ(i,j)$, of two prototypes v_i and v_j, is the number of data vectors for which v_i is the BMU and v_j is the second BMU.* By this definition, $|RF_i| = \sum_{k=1}^{N} CADJ(i,k)$.

Definition 2: *The level of connectedness (similarity) of two prototypes v_i and v_j is*

$$CONN(i,j) = CADJ(i,j) + CADJ(j,i) \qquad (1)$$

By definition, $CONN$ is symmetric and shows how similar two prototypes are by indicating the number of data vectors for which they are the BMU and the second BMU pair. $CADJ$ and $CONN$ indicate the data topology on the prototype level by showing the neighborhood relations of the prototypes. They also provide a finer resolution for data distribution than the Voronoi polyhedron level by showing how the data is distributed within the Voronoi polyhedron among the neighbor prototypes.

A visualized example of the $CONN$ matrix is shown in Figure 1 for a 2-d data set called "Lsun" created by [17]. This data set has three well-separated clusters (two rectangular and one spherical) with different variances within clusters and different inter-cluster distances. The prototypes were obtained by a 10x10 SOM. $CONN$ makes high-density regions and no-data regions (disconnected parts of the data set) visible, which outlines the boundaries of the three clusters. We show in the next sections how $CADJ$ and $CONN$ can help determine the validity of clustering for two-step clustering (prototype based clustering) algorithms.

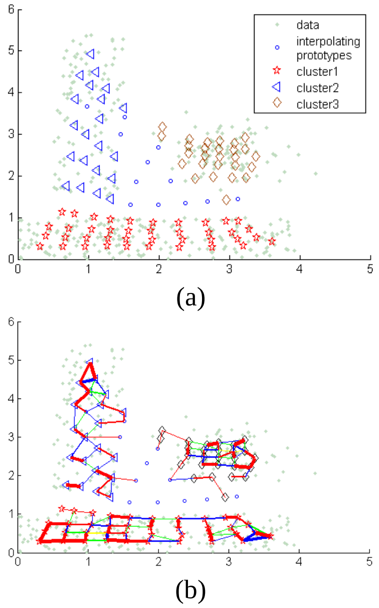


Fig. 1. Density representation by connectivity matrix $CONN$, shown on a simple 2-d dataset Lsun [17]. A 10x10 SOM is used to obtain prototypes. (a) Lsun (3 clusters) and its prototypes with true labels in data space. Interpolating prototypes (small open circles) have empty receptive fields. Variances within clusters and inter-cluster distances are different but the 3 clusters are well separated. (b) $CONN$ visualization for this dataset in data space. The width of a line indicates the number of data vectors for which the prototypes connected by this line are the BMU and the second BMU pair. The separations between clusters are indicated by unconnected prototypes.

B. Definition of $Conn\_Index$

Assume K clusters, N prototypes v in a data set and let C_k, C_l refer to two different clusters where $1 \leq k, l \leq K$. $Conn\_Index$ will be defined with the help of two quantities, *compactness* of clusters, $Intra\_Conn$, and *separation* of clusters, $1 - Inter\_Conn$, so we introduce these quantities first. The compactness of C_k, $Intra\_Conn(C_k)$, is the ratio of the number of data vectors in C_k whose second BMU is also in C_k, to the number of data vectors in C_k:

$$Intra\_Conn(C_k) = \frac{\sum_{i,j}^{N}\{CADJ(i,j):\ v_i, v_j \in C_k\}}{\sum_{i,j}^{N}\{CADJ(i,j):\ v_i \in C_k\}} \qquad (2)$$

By this definition, $Intra\_Conn(C_k) \in [0,1]$ where a greater value shows a more compact cluster. If the second BMUs of all data vectors in C_k are also in C_k, then $Intra\_Conn(C_k) = 1$. The intra-cluster connectivity of all clusters, $Intra\_Conn$, is the average compactness, *i.e.*,

$$Intra\_Conn = \sum_{k}^{K} Intra\_Conn(C_k)/K \qquad (3)$$

We define the inter-cluster connectivity between clusters C_k and C_l, $Inter\_Conn(C_k, C_l)$, as the ratio of the connectivity between C_k and C_l to total connectivity of the prototypes in C_k which have at least one connection to a prototype in C_l,

$$Inter\_Conn(C_k, C_l) = \frac{Conn(C_k, C_l)}{\sum_{i,j}^{N}\{CONN(i,j) : v_i \in V_{k,l}\}} \qquad (4)$$

with $Conn(C_k, C_l) = \sum_{i,j}^{N}\{CONN(i,j) : v_i \in C_k, v_j \in C_l\}$

and $V_{k,l} = \{v_i :\ v_i \in C_k,\ \exists\, v_j \in C_l :\ CADJ(i,j) > 0\}$.

This ratio shows how similar the prototypes at the boundary of C_k are to the ones at the boundary of C_l. If C_k and

C_l are completely separated (have no connection), then $Inter\_Conn(C_k, C_l) = 0$. A greater $Inter\_Conn(C_k, C_l)$ indicates a greater degree of similarity between C_k and C_l. $Inter\_Conn(C_k, C_l) > 0.5$ indicates that those prototypes in C_k which have connections to C_l should in fact be in C_l, or C_k and C_l should be combined. We define the interconnectivity of C_k to all other clusters, $Inter\_Conn(C_k)$, and the average similarity of clusters, $Inter\_Conn$, as

$$Inter\_Conn(C_k) = \max_{l, l \leq K} Inter\_Conn(C_k, C_l) \quad (5)$$

$$Inter\_Conn = \sum_{k}^{K} Inter\_Conn(C_k)/K \quad (6)$$

Then $1 - Inter\_Conn$ is the *separation* measure between clusters. Finally, we define $Conn\_Index$ as

$$Conn\_Index = Intra\_Conn \times (1 - Inter\_Conn) \quad (7)$$

$Conn\_Index \in [0, 1]$, increases with better clustering, where 1 means completely separated clusters. The complexity of $Conn\_Index$ is $O(N^3)$, which depends only on the number of prototypes, while the complexity of GDI [10] is $O(d \times M^2)$ and of CDbw [16] is $O(d \times M \times N^2)$ where d is the data dimensionality and M is the number of data vectors. $Conn\_Index$ is fast for large and high-dimensional data sets compared to GDI and CDbw.

To exemplify how $Conn\_Index$ is calculated, Figure 2 shows a clustering of eleven prototypes of synthetic data into three groups, A, B and C. Numbers on the connecting lines show the connectivity strengths between the respective prototypes. For the prototypes at the cluster boundaries (p_5, p_6, p_8, p_9 and p_{10}), cumulative adjacencies are also indicated. B has 10 (6+4+4) data vectors whose BMU and second BMU are in B. B also has 3 data vectors whose second BMU is in another cluster ($CADJ(p_6, p_5) + CADJ(p_6, p_9) = 2 + 1 = 3$). Thus $Intra\_Conn(B)$ is 14/17. Similarly, $Intra\_Conn(A)$ is 18/19 and $Intra\_Conn(C)$ is 11/13, which produce $Intra\_Conn = 0.87$. The total connectivity strength between B and C is $CONN(p_6, p_9) + CONN(p_8, p_{10}) = 2 + 1 = 3$. The prototypes in B that have a connection to C are p_6 and p_8. Their connections within B are $CONN(p_6, p_7)$, $CONN(p_6, p_8)$ and $CONN(p_8, p_7)$ which sum up to 10. Hence, $Inter\_Conn(B, C)$ is 3/17. Similarly, $Inter\_Conn(B, A)$ is 3/13 which results in $Inter\_Conn(B) = \max(3/13, 3/17) = 3/13$. A and C are only connected to B, so $Inter\_Conn(A) = 3/10$ and $Inter\_Conn(C) = 3/14$ which result in $Inter\_Conn = 0.25$ and $Conn\_Index = 0.87 \times 0.75 = 0.65$.

While $Inter\_Conn$ depends only on the connections of prototypes at the cluster boundaries, $Intra\_Conn$ heavily depends on the sizes of the clusters. Therefore, $Intra\_Conn$ will certainly decrease with increasing number of clusters, unless the clusters are split along natural cluster boundaries.

III. EXAMPLES FOR $Conn\_Index$ PERFORMANCE

We use three synthetic data sets created by [17]: Lsun, Wingnut, Engytime, each with different properties. We obtain the prototypes by a 10x10 SOM, and cluster them by k-means and single linkage clustering. We compare $Conn\_Index$ to the commonly used Davies-Bouldin index (DBI) [8], to the generalized Dunn index (GDI) [10] and to an index proposed for prototype based clustering (CDbw) [16]. GDI is used with centroid linkage and average distance of points to cluster centroids as the inter- and intra-cluster distance metrics, respectively. DBI decreases with the increasing cluster quality while GDI, CDbw and $Conn\_Index$ increase with better clustering. DBI, GDI and CDbw have values in $[0, \infty)$ while $Conn\_Index \in [0, 1]$.

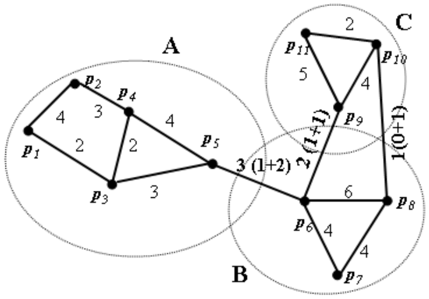


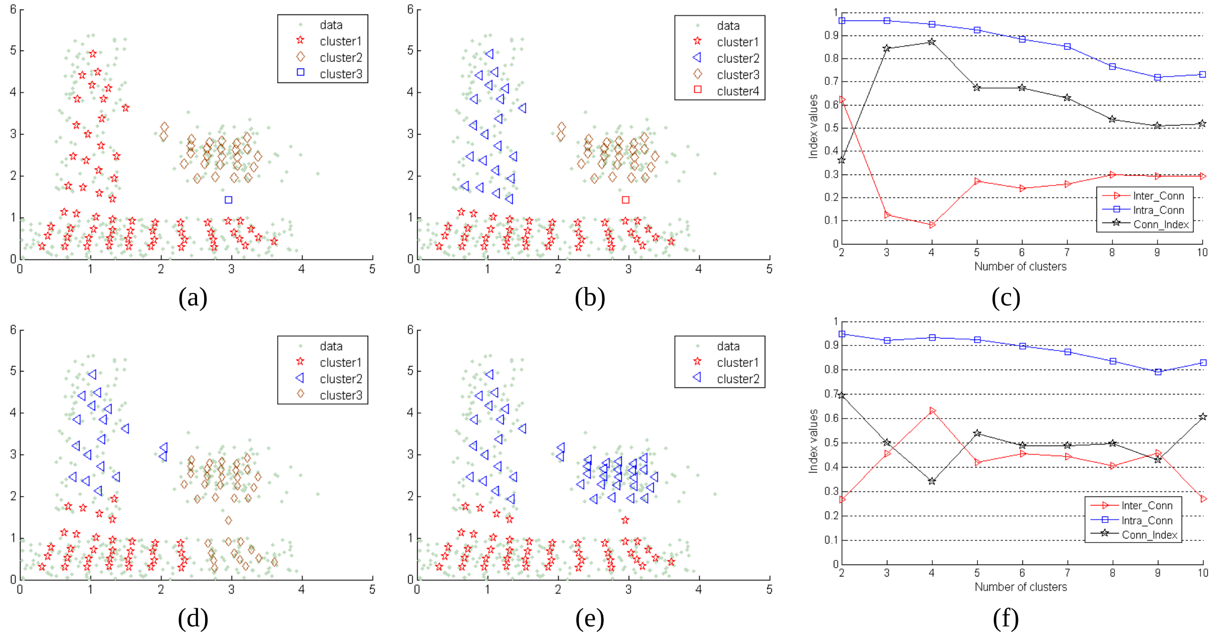
Fig. 2. An example of how $Conn\_Index$ is calculated. We have three clusters, A, B and C, and 11 prototypes indicated by p_i. The numbers on the lines connecting two prototypes are the connectivity strengths between those prototypes. We also indicate the cumulative adjacencies for the prototypes at the boundaries, $CADJ(p_5, p_6) = 1$, $CADJ(p_6, p_5) = 2$; $CADJ(p_6, p_9) = CADJ(p_9, p_6) = CADJ(p_{10}, p_8) = 1$ and $CADJ(p_8, p_{10}) = 0$.

TABLE I
VALIDITY INDICES FOR SINGLE LINKAGE CLUSTERING OF LSUN

Validity Indices	Indices for true cluster labels (k=3)	Indices for single-linkage clustering, k = # of clusters			
		k=2	k=3	k=4	k=5
DBI	0.77	1.16	1.36	**0.74**	0.78
GDI	**1.37**	0.83	0.35	0.51	0.41
CDbw	0.96	1.18	1.17	1.03	**2.05**
Conn_Index	**1.0**	0.37	0.84	0.88	0.68

A. *Lsun: 400 points, 3 well-separated clusters*

Lsun has three well separated clusters: two rectangular, one spherical, shown in Figure 1.a. Since the true clusters are completely separated, $Conn\_Index$ is 1 (Figure 1.b). Single linkage clustering with 3 clusters, merges two rectangular clusters and has an extra singleton (Figure 3.a). The closest result to the true labels is the one with 4 clusters (Figure 3.b) where the only difference is a singleton prototype. It is also favored best by $Conn\_Index$ (Figure 3.c) with a maximum index, 0.88, for k=4, although it is less than the index for the true labels, as expected. Table I provides DBI, GDI, and CDbw values for single linkage clustering. Among them, only GDI favors the true clusters. Figures 3.d-e show k-means clustering, which is unsuccessful because of the variable cluster shapes and proximity relations of clusters. Due to the same reasons, all indices except $Conn\_Index$ favor five clusters for k-means (Table II) which is the case



(a) (b) (c)

(d) (e) (f)

Fig. 3. Clustering of Lsun. Small dots are data points, symbols indicate prototypes, with labels assigned by a clustering algorithm. Top: Single linkage clustering (a) with the true number (3) of clusters, (b) k=4 (for which the $Conn\_Index$ is maximum, Table II), and (c) $Conn\_Index$ for different numbers of clusters. Bottom: k-means clustering (d) with the true number of clusters, (e) k=2 (at maximum $Conn\_Index$), and (f) $Conn\_Index$ for different numbers of clusters.

TABLE II

VALIDITY INDICES FOR K-MEANS CLUSTERING OF THE SYNTHETIC DATA SETS USED IN THIS PAPER

Data Sets	Validity indices	Indices for true clusters	Indices for k-means, k = # of clusters			
			k=2	k=3	k=4	k=5
Lsun (k=3)	DBI	0.77	0.93	0.77	0.61	**0.58**
	GDI	1.37	1.50	1.67	1.63	**1.83**
	CDbw	0.96	1.34	0.76	1.64	**1.80**
	Conn_Index	**1.0**	0.70	0.51	0.34	0.54
Wingnut (k=2)	DBI	0.97	0.95	0.88	0.85	**0.84**
	GDI	1.48	1.51	1.32	**1.55**	1.34
	CDbw	0.86	**1.02**	0.54	0.39	0.43
	Conn_Index	**0.90**	0.60	0.45	0.52	0.58
Engy time (k=2)	DBI	**0.80**	0.97	0.94	0.81	0.89
	GDI	1.36	1.35	**1.63**	1.11	1.27
	CDbw	**1.69**	1.27	0.82	0.67	0.75
	Conn_Index	**0.77**	**0.77**	0.54	0.53	0.53

where the two rectangular clusters are split into four spherical ones. The maximum $Conn\_Index = 0.70$ indicates the poorer quality of k-means compared to the true partitions.

B. Wingnut: 1048 data points, 2 non-overlapping clusters

Figure 4.a shows the data set and its prototypes with the true labels. The data set has highly varying density distribution within clusters. The clusters are not overlapping, but they are very close, which results in connections across the prototypes at the cluster boundaries. That explains $Conn\_Index < 1$ (0.90) for the true labels. Neither k-means nor single linkage clustering results in the true partitioning. We leave out single linkage clustering due to its poor performance. Figure 4.b shows the best k-means clustering as judged by $Conn\_Index$ (k=2, in Figure 4.c). $Conn\_Index$ in Table II indicates the poor performance of k-means by much smaller values than 0.90. The other indices are unsuccessful due to the data structure: they favor k-means clustering with larger k values.

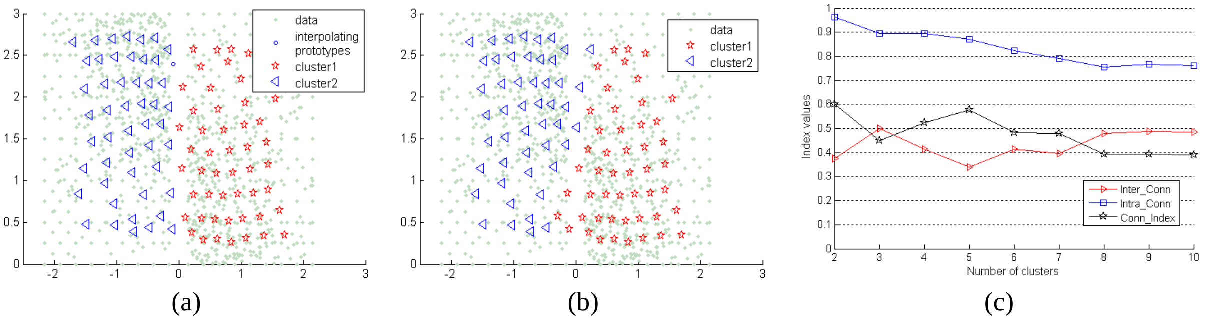
C. Engytime: 4096 data points, 2 overlapping clusters

Engytime is a mixture of two highly overlapping Gaussian distributed clusters (Figure 5.a). $Conn\_Index$ is 0.77 for the true clusters. Even though one can expect $Intra\_Conn$ to be small because of highly overlapping clusters, $Intra\_Conn$ is 0.97 due to relatively large sizes of the clusters. The best k-means clustering (Figure 5.b) according to $Conn\_Index$ is with k=2 (0.77, Table II). The index is much larger than the index for other k values and the same as the index for the true clusters. For this data set, CDbw provides the best evaluation, *i. e.*, it significantly favors k=2 for k-means, and it is by far the highest for the true clusters. This is because it is based on deviation radius of data distribution within and between the clusters.

IV. INDICES ON REAL DATA SETS

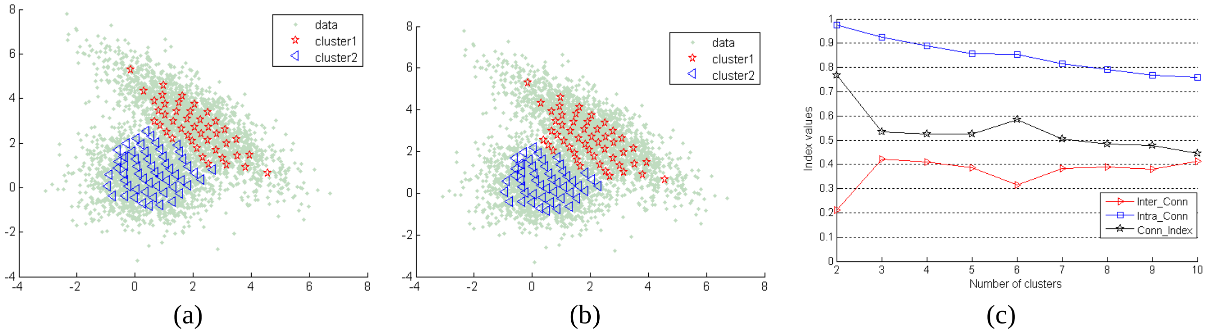
A. Data sets with small number of clusters

We used three data sets from the benchmark data sets for clustering and classification in the UCI Machine Learning Repository: Breast Cancer Wisconsin (699 samples with 2 classes, 10-d); Iris (150 samples with 3 classes, 4-d); and Wine (178 samples with 3 classes, 13-d). Table III gives the indices for the known labels, and for k-means clustering of (4x4) SOM prototypes. $Conn\_Index$ favors the known



(a) (b) (c)

Fig. 4. Wingnut: A 2-d data set with highly varying density distribution within the clusters. (a) The data vectors are shown by dots, the prototypes obtained by a 10x10 SOM are annotated with the true labels of the data in their receptive fields. (b) k-means clustering with k=2 (at maximum $Conn\_Index$). (c) $Conn\_Index$ for k-means clustering.



(a) (b) (c)

Fig. 5. Engytime: A 2-d data set which is a mixture of 2 Gaussian distributions. (a) The data vectors are shown by dots, the prototypes obtained by a 10x10 SOM are annotated with the true labels of the data in their receptive fields. (b) k-means clustering with k=2 (at maximum $Conn\_Index$). (c) $Conn\_Index$ for k-means clustering.

TABLE III

VALIDITY INDICES FOR K-MEANS CLUSTERING OF THREE REAL DATA SETS: BREAST CANCER WISCONSIN, IRIS AND WINE

Data Sets	Validity indices	Indices for true clusters	Indices for k-means, k = # of clusters			
			k=2	k=3	k=4	k=5
Breast Cancer Wisconsin (k=2)	DBI	0.80	**0.76**	1.61	1.86	1.87
	GDI	1.17	**1.27**	0.66	0.60	0.28
	CDbw	6.03	**43.7**	20.6	19.3	8.98
	Conn_Index	**0.79**	0.78	0.64	0.39	0.30
Iris (k=3)	DBI	0.81	**0.49**	0.75	0.92	0.98
	GDI	2.75	**3.61**	2.62	1.69	1.38
	CDbw	1.06	**4.77**	0.68	0.41	0.30
	Conn_Index	0.67	**1.0**	0.62	0.54	0.53
Wine (k=3)	DBI	1.36	**1.26**	1.34	1.58	2.08
	GDI	0.85	**1.12**	0.97	0.82	0.37
	CDbw	0.24	**0.67**	0.51	0.45	0.25
	Conn_Index	**0.63**	0.45	0.55	0.36	0.23

labels as the best clustering for Breast Cancer and Wine. In contrast, DBI, GDI and CDbw favor k-means clustering (k=2) for both data sets even though Wine has 3 clusters. Surprisingly, CDbw favors any k-means clustering to the known labels for these two data sets. For Iris, where two clusters are overlapping and very dissimilar to the third one, all indices including $Conn\_Index$ validate two clusters as the best, and all indices except DBI favor the known labels as the second best.

B. *Ocean City: A large real remote sensing image*

To evaluate indices on complicated data, we use a large real remote sensing spectral image of Ocean City, Maryland. It comprises 512x512 pixels and represents fairly complicated data. Each pixel is an 8-d feature vector, called spectrum. Ocean City is a long linear urban settlement on the seashore. More details of the data set are in [18]. At least 25 clusters were verified by a domain expert as meaningful physical clusters, partly by ground truthing for an earlier supervised classification of the same image [18]. Figure 6 shows the Ocean City image with 28 clusters obtained by a semi-manual clustering based on $CONN$ visualization [1]. Some major classes in the image are ocean (blue), small bays (medium blue), water canals (turquoise), lawn, trees and bushes (green and split-pea green), dry grass (orange), marshlands (brown and ocher), soil (gray), road (magenta) and concrete (red). The small rows of rectangles are buildings and their colors indicate different types of roof materials.

We have 1600 prototypes from a 40x40 SOM. When we cluster the prototypes by k-means (k≤40), $Conn\_Index$ and CDbw favor k=4 as the best result (Table IV). These four clusters appear to be superclusters of the known ones. One supercluster comprises the known vegetation classes (lawn, trees, bushes, etc.), one represents water classes (ocean, canals, pool, etc.), one represents soil (marshlands, bare soil, etc.) and one comprises roads, concrete, different roof materials, etc. Table IV presents the indices up to k=8 and Figure 7 gives $Conn\_Index$, $Intra\_Conn$ and $Inter\_Conn$

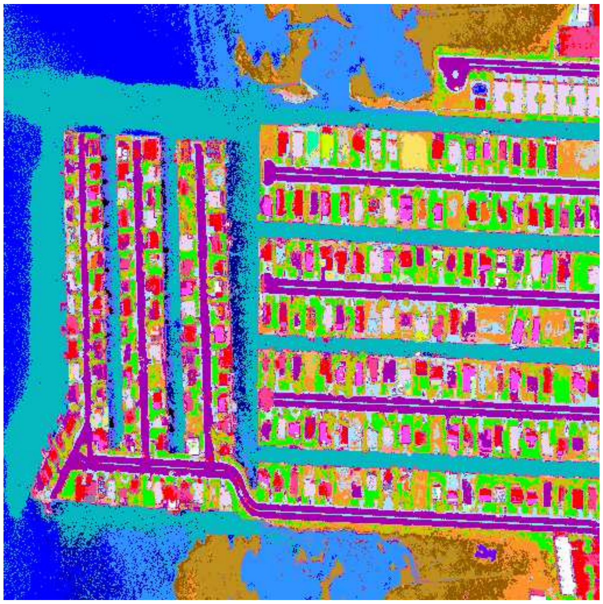


Fig. 6. Cluster map of Ocean City, an 8-band 512x512 pixel remote sensing image, obtained by semi-manual clustering based on $CONN$ visualization [1]. Each color represents a different cluster.

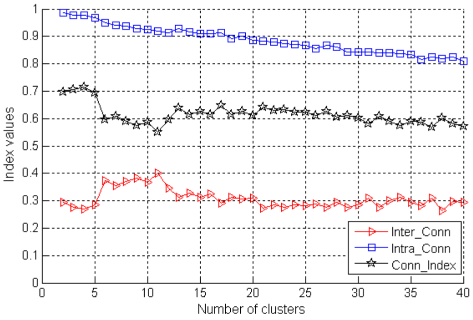
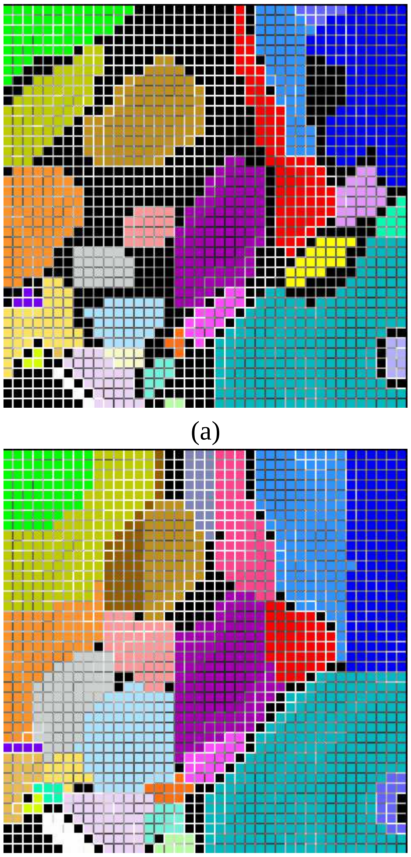


Fig. 7. $Conn\_Index$ for k-means clustering of Ocean City data.

for all k values (k≤40). DBI and GDI reflect k=2 as the best result where one cluster combines vegetation and soil, the other one contains everything else. Compared to the known clusters, k-means produces very poor partitioning for $k > 4$. The incorrect clustering for large k is indicated by all indices selecting either k=4 or k=2.

It is common practice to cluster the SOM prototypes semi-manually, based on visualizations. In many cases, different prototypes may be omitted from different clusterings of the same SOM due to different visualization schemes with different knowledge representation or processing by different users. This results in different numbers of unclustered prototypes in different cluster maps. Yet, we still need to be able to compare those different clusterings.



(a)

(b)

Fig. 8. Semi-manual clustering of the SOM prototypes of Ocean City overlain on the 40x40 SOM. Each cell is a prototype, color coded with a cluster label consistent with Fig.6. The intensities of the white fences around the cell are proportional to the distance between neighbor prototypes. Black cells are unclustered prototypes. (a) Clustering obtained from a modified U-matrix visualization [18], (b) Clustering from $CONN$ visualization [1]

TABLE IV

VALIDITY INDICES FOR K-MEANS CLUSTERING OF OCEAN CITY DATA

Validity Indices	Indices for k-means clustering, k = # of clusters						
	k=2	k=3	k=4	k=5	k=6	k=7	k=8
DBI	**0.66**	0.70	0.72	0.80	0.73	0.84	0.84
GDI	**1.93**	1.59	1.61	1.31	0.74	0.43	0.46
CDbw	0.38	1.94	**2.33**	1.56	2.17	2.27	1.91
Conn_Index	0.70	0.71	**0.72**	0.70	0.60	0.61	0.59

We compare two semi-manual clusterings of the Ocean City prototypes, based on different SOM visualizations: the first one (Figure 8.a) is obtained from a modified U-matrix [18], the second one (Figure 8.b) is obtained from $CONN$ visualization [1]. Both of these clusterings fit the data well (the clusters mapped back to the spatial image look very similar to known clusters) except Figure 8.a leaves more data vectors unclustered due to less prototypes assigned to clusters. The large number of unclustered prototypes in Figure 8.a reflects the user's conservative judgment given the uncertainty about divisions based on the modified U-matrix. Table V shows the indices for these cluster maps. We expect that the validity indices favor the semi-manual clusterings over the incorrect k-means clustering because they match the ground truth well, and because unclustered prototypes are at the cluster boundaries, thus increasing separation. DBI, GDI and CDbw favor the incorrect partitioning of k-means all the way up to k=40 even though k-means clustering for $k > 4$ is bad. DBI, GDI and CDbw are not helpful in evaluation of

TABLE V

VALIDITY INDICES FOR DIFFERENT CLUSTERINGS OF OCEAN CITY DATA

Validity Indices	U-matrix clustering [18] k=28	$CONN$ clustering [1] k=28	k-means k=2	k-means k=4
DBI	1.17	1.30	**0.66**	0.72
GDI	0.41	0.55	**1.93**	1.61
CDbw	0.18	0.21	0.38	**2.33**
Conn_Index	0.62	0.66	0.70	**0.72**
Intra_Conn	0.74	0.83	**0.99**	0.98
Inter_Conn	**0.17**	0.21	0.29	0.27

such clustering results.

Contrarily, $Conn\_Index$, $Intra\_Conn$ and $Inter\_Conn$ provide meaningful measures when the numbers of clustered prototypes resulting from different clusterings are not the same. $Intra\_Conn$ measures the effect of omitted prototypes because the smaller the clusters than their true size, the less compact they are, which produces a smaller $Intra\_Conn$. This is because the prototypes at the edges of a cluster whose fringes are trimmed have strong connections to the unclustered prototypes near edges, which probably belong to that cluster. $Intra\_Conn$ can also decrease when the data set is randomly partitioned, however, in such cases, $Inter\_Conn$ will be high. If the unclustered prototypes are at the cluster boundaries, a smaller number of clustered prototypes is expected to yield a smaller $Inter\_Conn$ value. $Conn\_Index$ indicates the combined effect of $Intra\_Conn$ and $Inter\_Conn$. In our case, $Conn\_Index$ favors the clustering based on $CONN$ visualization over the clustering based on the modified U-matrix. $Intra\_Conn$ indicates the clusters are more compact in the former while $Inter\_Conn$ indicates the clusters are more separated in the latter. It is seen in our experiments that when more prototypes remain unclustered at cluster boundaries, $Inter\_Conn$ and $Intra\_Conn$ decrease, as expected.

When clusterings with different numbers of clustered prototypes are compared to a fully clustered SOM, only $Inter\_Conn$ should be taken into account because $Intra\_Conn$ is affected heavily by the unclustered prototypes. When we compare the semi-manual clusterings to the k-means clustering, Table V shows that even the smallest $Inter\_Conn$ of k-means clustering (0.27) is larger than that of either of semi-manual clusterings (0.17 and 0.21). Hence, $Inter\_Conn$ expresses that the partitionings of semi-manual clusterings are more correct than the ones of the k-means.

V. CONCLUSIONS

In this paper, we propose a new validity index, $Conn\_Index$, for prototype based clustering algorithms. It depends on the intra-cluster ($Intra\_Conn$) and inter-cluster ($Inter\_Conn$) connectivities obtained from a weighted Delaunay triangulation (connectivity matrix, [1]). $Conn\_Index$ can be useful to evaluate cluster validity for data sets with clusters of different shapes or sizes, or with overlapping clusters. $Conn\_Index$ in our experiments measured the cluster validity in a more meaningful way, with respect to the true structure of data, compared to other indices in this study. When comparing the validity of a fully clustered SOM with one that has unclustered prototypes, $Intra\_Conn$, and consequently $Conn\_Index$ do not provide a meaningful measure. However, $Inter\_Conn$ can still be used for reliable evaluation of cluster separation. We will address this situation in future work. Although we present this discussion in the context of SOM prototypes, we stress that the construction of $Conn\_Index$ has no specifity to SOM prototypes and therefore it can be applied to prototypes produced by any other vector quantization algorithms.

REFERENCES

[1] K. Taşdemir and E. Merényi, “Data topology visualization for the Self-Organizing Maps,” in *Proc. 14th European Symposium on Artificial Neural Networks (ESANN 2006), Bruges, Belgium, D-Facto, April 26-28*, 2006, pp. 277–282.

[2] J. Vesanto and E. Alhoniemi, “Clustering of the self-organizing map,” *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586–600, May 2000.

[3] M. Cottrell and P. Rousset, “The Kohonen algorithm: A powerful tool for analyzing and representing multidimensional quantitative and qualitative data,” in *IWANN 1997 (International Work-Conference on Artificial Neural Networks)*, 1997, pp. 861–871.

[4] G. Karypis, E.-H. S. Han, and V. Kumar, “Chameleon: Hierarchical clustering using dynamic modeling,” *Computer*, vol. 32, no. 8, pp. 68–75, 1999.

[5] A. Fred and A. Jain, “Robust data clustering,” in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003, pp. II–128–133.

[6] E. Merényi, “”Precision Mining” of high-dimensional patterns with Self-Organizing Maps: Interpretation of hyperspectral images,” in *Quo Vadis Computational Intelligence: New Trends and Approaches in Computational Intelligence. Studies in Fuzziness and Soft Computing*, vol. 54. Physica-Verlag, 2000.

[7] S. Theodoridis and K. Koutroubas, *Pattern Recognition*. Academic Press, 1999.

[8] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE Trans. Pattern Analysis Machine Intelligence (PAMI)*, vol. 1, no. 2, pp. 224–227, 1979.

[9] J. C. Dunn, “Well separated clusters and optimal fuzzy partitions,” *Journal Cybernetics*, no. 4, pp. 95–104, 1974.

[10] J. C. Bezdek and N. R. Pal, “Some new indexes of cluster validity,” *IEEE Trans. on Systems, Man and Cybernetics-Part B*, vol. 28, no. 3, pp. 301–315, 1998.

[11] M. Kim and R. S. Ramakrishna, “New indices for cluster validity assessment,” *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2353–2363, 2005.

[12] G. W. Milligan and M. C. Cooper, “An examination of procedures for determining the number of clusters in a data set,” *Psychometrika*, no. 50, pp. 159–179, 1985.

[13] X. L. Xie and G. Beni, “A validity measure for fuzzy clustering,” *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 13, no. 3, pp. 841–847, 1991.

[14] N. R. Pal and J. C. Bezdek, “On cluster validity for the fuzzy c-means model,” *IEEE Trans. Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, 1995.

[15] D. Kim, K. H. Lee, and D. Lee, “On cluster validity index for estimation of the optimal number of fuzzy clusters,” *Pattern Recognition*, no. 37, pp. 2009–2025, 2004.

[16] M. Halkidi and M. Vazirgiannis, “Clustering validity assessment using multi representatives,” in *Proc. of SETN Conference, Thessaloniki, Greece, April*, 2002.

[17] A. Ultsch, “Clustering with som: U*c,” in *Proc. 5th Workshop on Self-Organizing Maps (WSOM05), Paris, France, September 5-8,*, 2005, pp. 75–82.

[18] E. Merényi and A. Jain, “Forbidden magnification? II.” in *Proc. 12th European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, D-Facto, April 28-30*, 2004, pp. 57–62.