ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

# Modified Conn-Index for the evaluation of fuzzy clusterings

Tina Geweniger[1,2], Marika Kästner[2], Mandy Lange[2], Thomas Villmann[2]

1 - Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands

2 - Computational Intelligence Group, University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany

**Abstract**. We propose an extension of the Conn-Index to evaluate fuzzy cluster solutions obtained from fuzzy prototype vector quantization, whereas the original Conn-Index was designed for crisp vector quantization models. The fuzzy index explicitly takes the fuzzy assignments resulting from fuzzy vector quantization into account. This avoids the information loss which would occur if the original crisp index is applied to fuzzy solutions.

## 1 Motivation

Prototype based vector quantization (VQ) is an approved method to compress and cluster very large data sets. Thereby, the data are represented by a much smaller number of prototypes. If each data point is uniquely assigned to one prototype, it is called crisp clustering. Famous methods are c-Means [1], Self-Organizing Maps (SOM) [2] and Neural Gas (NG) [3]. Yet, in practical applications the clusters are often overlapping. For this kind of data fuzzy clustering methods have been developed, e. g. Fuzzy c-Means (FCM) [4] and Fuzzy SOM (FSOM) [5]. For these methods each data point is partially assigned to each prototype. The FSOM is an extension of the FCM taking the neighborhood cooperativeness into account. Thereby, the neighborhood is bound to an external topological structure like a grid as known from SOM. The new approach called Fuzzy NG (FNG) [6] introduced in Sec. 2 combines the FCM with the NG using the dynamic neighborhood cooperativeness as known from NG [3].

Clustering in general is an ill-posed problem and it is difficult to validate a cluster solution. There exist a number of validity measures based on separation and compactness, e. g. Partition Entropy, Xie-Beni-Index, and Fukuyama-Sugeno-Index [7]. They were originally developed to verify the correct number of prototypes assuming that each cluster is represented by exactly one prototype. Yet very large data sets require a higher number of prototypes to represent the data and the above mentioned measures cannot be used. Taşdemir & Merényi proposed the Conn-Index [8], which is suited to evaluate crisp clusterings, where each cluster contains more than one prototype. This Conn-Index takes the neighborhood structure between the learned prototypes into account to transfer the information of the full data set to the cluster validation process. We propose a modification of the Conn-Index for fuzzy cluster solutions in Sec. 3.

Further, we show that the new Fuzzy Conn-Index can be successfully applied to judge cluster solutions based on fuzzy VQ algorithms. We demonstrate this for an artificial and a real world data set. To obtain the cluster solutions for the real world example we apply Fuzzy Neural Gas (FNG) [6] and subsequent clustering of the prototypes by Affinity Propagation (AP) [9].

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

## 2 Fuzzy VQ and subsequent clustering of prototypes

The Fuzzy Conn-Index requires that the data set is represented by prototypes, which are obtained by a fuzzy VQ method and subsequently clustered. For the fuzzy VQ we use the FNG as a powerful fuzzy VQ method, whereas for the clustering the established AP algorithm is applied.

We now briefly introduce the recently published FNG [6], which can be derived directly by combining the well-known NG and the FCM. For that purpose the FCM cost function is equipped with a dynamic neighborhood between the prototypes as introduced for the NG. Both methods, FNG and FCM, are prototype based vector quantizers, where prototypes $W = \{\mathbf{w}_j\}_{j=1}^{n} \subset \mathbb{R}^D$ represent the data set $V = \{\mathbf{v}_i\}_{i=1}^{N} \subseteq \mathbb{R}^D$ with $n \ll N$. The cost function of the FCM is

$$E_{FCM}(\mathbf{U}, V, W) = \sum_{j=1}^{n} \sum_{i=1}^{N} (u_{ij})^m \, d(\mathbf{v}_i, \mathbf{w}_j)^2 \tag{1}$$

with $m \in (1, \infty)$ as the fuzziness parameter which is typically set to $m = 1.2, \ldots, 2$ [10]. The fuzzy assignments $u_{ij} \in \mathbf{U} \subseteq [0,1]^{N \times n}$ describe the membership of data point $\mathbf{v}_i$ to prototype $\mathbf{w}_j$. If $\sum_{j=1}^{n} u_{ij} = 1$ holds, then the assignments are probabilistic, otherwise possibilistic. The Euclidean distance is usually used as dissimilarity measure $d(\mathbf{v}_i, \mathbf{w}_j)$, but other choices are possible. In a variant of the NG the local costs for mapping the data point $\mathbf{v}_i$ to a certain prototype $\mathbf{w}_j$ can be defined as

$$lc_\sigma^{NG}(i,j) = \sum_{l=1}^{n} h_\sigma^{NG}(j,l) \cdot d(\mathbf{v}_i, \mathbf{w}_l)^2 \tag{2}$$

taking into account the dynamic neighborhood structure according to

$$h_\sigma^{NG}(j,l) = c_\sigma^{NG} \cdot e^{\left(-\frac{(rk_j(\mathbf{w}_l, W))^2}{2\sigma^2}\right)}. \tag{3}$$

The winning ranks $rk_j$ of each prototype $\mathbf{w}_j$ are calculated by

$$rk_j(\mathbf{w}_l, W) = \sum_{k=1}^{n} \Theta\left(d(\mathbf{w}_l, \mathbf{w}_j) - d(\mathbf{w}_l, \mathbf{w}_k)\right). \tag{4}$$

$\Theta(x)$ is the Heaviside function, where $\Theta(x) = 0$ iff $x \leq 0$ and 1 else [3]. The value $\sigma > 0$ is the neighborhood range and $c_\sigma^{NG}$ assures that $\sum_l h_\sigma^{NG}(j,l) = 1$ [3]. Note that the prototype neighborhood here differs from the neighborhood in original NG which is based on the winner ranks according to the data.

Now the FNG is obtained by replacing the quadratic distances $d(\mathbf{v}_i, \mathbf{w}_j)^2$ in eq. (1) with the local costs (2) from the NG. The update of the prototypes and the fuzzy assignments of the FNG now include the neighborhood function (3) [6]. Using the Euclidean distance we get

$$\mathbf{w}_j = \frac{\sum_{i=1}^{N} \sum_{l=1}^{n} (u_{il})^m \cdot h_\sigma^{NG}(j,l) \cdot \mathbf{v}_i}{\sum_{i=1}^{N} \sum_{l=1}^{n} (u_{il})^m \cdot h_\sigma^{NG}(j,l)} \tag{5}$$

$$u_{i,j} = \frac{1}{\sum_{l=1}^{n} \left(\frac{lc_\sigma^{NG}(i,j)}{lc_\sigma^{NG}(i,l)}\right)^{\frac{1}{m-1}}}. \tag{6}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

## 3 Generalization of the Conn-Index — Fuzzy Conn-Index

The Generalized Conn-Index $C$ [11], which is based on the Conn-Index proposed by Taşdemir & Merényi [8], is a validity measure to evaluate clusterings of very large data sets $V = \{\mathbf{v}_i\}_{i=1}^{N} \subseteq \mathbb{R}^D$, which were partitioned using a prototype based VQ scheme with the prototype set $W = \{\mathbf{w}_j\}_{j=1}^{n} \subset \mathbb{R}^D$. Thereby it is presumed that each of the $K$ clusters $\Omega_k$ is represented by more than one prototype. Following Taşdemir & Merényi the index balances the overall cluster compactness and separation by combining the inter-cluster connectivity $C_{inter} \in [0,1]$ and the intra-cluster connectivity $C_{intra} \in [0,1]$

$$C = C_{intra} \cdot (1 - C_{inter}) \,. \tag{7}$$

Thereby, $C_{intra}$ measures the compactness of the clusters and $C_{inter}$ evaluates the separation between them. The calculation of $C_{intra}$ is based on the cumulative adjacency matrix

$$\mathbf{A} = \sum_{l=1}^{N} \psi(\mathbf{v}_l) \tag{8}$$

with elements $a_{ij}$ and where the $\psi(\mathbf{v}_l)$ are $n \times n$ zero matrices except the row vector $\mathbf{r}_{s_0}(\mathbf{v}_l)$ corresponding to the best matching protoype $\mathbf{w}_{s_0(\mathbf{v}_l)}$ of the regarded data point $\mathbf{v}_l$ with

$$s_0(\mathbf{v}_l) = \mathrm{argmin}_j \left( d(\mathbf{v}_l, \mathbf{w}_j) \right) \tag{9}$$

where $d(\mathbf{v}_l, \mathbf{w}_j)$ is the same dissimilarity measure as used for the vector quantization. The vector $\mathbf{r}_{s_0}(\mathbf{v}_l)$ is also called response and is affiliated with the ranks of all other prototypes with respect to $\mathbf{v}_l$. In particular, each vector element $r_i(\mathbf{v}_l)$ corresponding to the winner rank of the $i$th prototype is defined as

$$r_i(\mathbf{v}_l) = \varphi\left(rk_i(\mathbf{v}_l, W)\right) \tag{10}$$

where $rk_i(y)$ is the rank function (4), but now based on the distances between a given data point and a prototype. We explicitly remark, that the rank of the winning prototype $\mathbf{w}_{s_0(\mathbf{v}_l)}$ obtained from (9) is zero. Further for the $(q+1)$th winner $\mathbf{w}_{s_q(\mathbf{v}_l)}$ the rank is $q$. The function $\varphi(x)$ is arbitrary monotonically decreasing, e. g. the exponential function. Thus, the Generalized Conn-Index (7) takes also higher winning ranks into account in contradiction to the original Conn-Index [8], where only the first and second best matching prototypes $\mathbf{w}_{s_0(\mathbf{v}_l)}$ and $\mathbf{w}_{s_1(\mathbf{v}_l)}$ are considered. Yet, the Generalized Conn-Index comprises the original Conn-Index by setting

$$\varphi(rk_i(\mathbf{v}_l, W)) = \begin{cases} 1 & \text{for } rk_i(\mathbf{v}_l, W) = 1 \\ 0 & \text{else} \end{cases} \tag{11}$$

The compactness value $C_{intra}$ is now the average of the local $C_{intra}(k)$ over all $K$ clusters $\Omega_k$, $C_{intra} = \frac{1}{K} \sum_{k=1}^{K} C_{intra}(k)$, where

$$C_{intra}(k) = \frac{\sum_{i,j|i\neq j} \{a_{ij} \mid \mathbf{w}_i, \mathbf{w}_j \in \Omega_k\}}{\sum_{i,j|i\neq j} \{a_{ij} \mid \mathbf{w}_i \in \Omega_k\}} \tag{12}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

is the ratio of the prototype connections within cluster $k$ to all connections from and between the prototypes describing cluster $k$.

For the inter-cluster connectivity $C_{inter}$ the connectivity matrix $\mathbf{C} = \mathbf{A}^T + \mathbf{A}$ is required. Their elements $c_{ij}$ can be interpreted as the dissimilarities between the prototypes, and hence, implicitly contain information about the local data density according to the magnification property [12]. The inter-cluster connectivity $C_{inter}$ is the average of the values $C_{inter}(k)$ analoguously to $C_{intra}(k)$ where

$$C_{inter}(k) = \max_{1 \leq l \leq K, k \neq l} C_{inter}(k, l) \tag{13}$$

is the maximum of the local inter-cluster connectivities $C_{inter}(k, l)$

$$C_{inter}(k, l) = \begin{cases} 0 & \text{if} \quad S_{k,l} = \emptyset \\ \frac{\sum_{i,j|i \neq j} \{c_{ij} | \mathbf{w}_i \in \Omega_k, \mathbf{w}_j \in \Omega_l\}}{\sum_{i,j|i \neq j} \{c_{ij} | \mathbf{w}_i \in S_{k,l}\}} & \text{if} \quad S_{k,l} \neq \emptyset \end{cases} \tag{14}$$

The set $S_{k,l}$ describes the neighborhood relations between the cluster $\Omega_k$ and $\Omega_l$ based on the contained prototypes $S_{k,l} = \{\mathbf{w}_i \mid \mathbf{w}_i \in \Omega_k \wedge \exists \mathbf{w}_j \in \Omega_l : a_{ij} > 0\}$, i. e. the inter-cluster connectivity $C_{inter}(k, l)$ evaluates the separation of cluster $\Omega_k$ from cluster $\Omega_l$.

Generally, high values of the (Generalized) Conn-Index $C$ (7) indicate a good clustering. That implies high values for the compactness $C_{intra}$ vs. small values for the separation $C_{inter}$ are desired.

According to the used winner rank function the Conn-Index assumes a unique crisp winner ranking. Thereby, as mentioned above, the rank function (10) used in the Generalized Conn-Index (7) reflects the topological structure of the data. In fuzzy VQ this information is implicitly contained in the fuzzy assignments $u_{ij}$ of the data to the prototypes. Thus, in the new Fuzzy Conn-Index $C_F$ the fuzzy assignments are considered instead of the prototype ranks. Particularly we redefine the response vector $\mathbf{r}_{s_0}(\mathbf{v}_l)$ by

$$r_i(\mathbf{v}_l) = u_{li} \tag{15}$$

instead of eq. (10). In this way fuzzy decisions in vector quantization can directly be used to calculate a cluster validation index. Thereby, the structural methodology of the original Conn-Index is preserved in the Fuzzy Conn-Index.

## 4 Experiments

To demonstrate the capability of the new Fuzzy Conn-Index in comparison to the crisp version we consider an artificial and a real world data set.

**Artificial dataset 'Smiley'.** This two-dimensional data set called *Smiley* consists of three Gaussian clouds with varying variances and one curved cloud. Three of them are overlapping and one is well separated from the others (see Fig. 1). We distributed 17 prototypes to describe the data. These positions are fixed for the following experiments. We varied the fuzziness ranging from crisp (equivalently to $m = 1$, [10]) to severe fuzziness ($m = 3$): For the crisp case mapping rule (9) was used for the determination of the assignments, whereas for fuzzy assignments eq. (6) was applied. In the latter case the calculation was done taking the limit $\sigma \to 0$. Subsequently, we clustered the prototypes to

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

| | Conn-Index (crisp) | Fuzzy Conn-Index 1.1 | 1.25 | 1.5 | 1.75 | 2.0 | 3.0 |
|---|---|---|---|---|---|---|---|
| C | 0.56 | 0.80 | 0.77 | 0.65 | 0.51 | 0.40 | 0.187 |
| $C_{intra}$ | 0.783 | 0.798 | 0.779 | 0.682 | 0.579 | 0.491 | 0.316 |
| $C_{inter}$ | 0.283 | 0.004 | 0.013 | 0.050 | 0.110 | 0.179 | 0.409 |

**Table 1:** Values for the original and the Fuzzy Conn-index for the artificial *Smiley* data set. The Fuzzy Conn-Index based on varying values of the fuzziness parameter $m$.

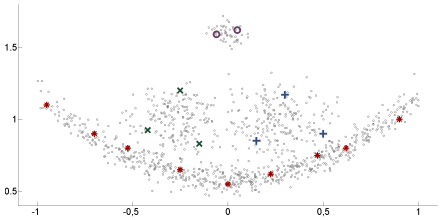

**Fig. 1:** *Smiley* with clustered prototypes (red $\star$, green $\times$, blue $+$, purple o)
Colored image can be obtained from the authors.

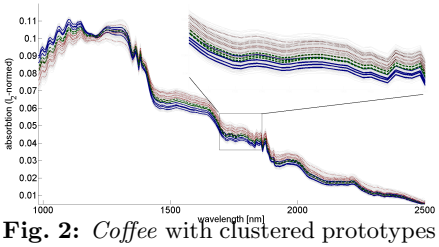


**Fig. 2:** *Coffee* with clustered prototypes (red dotted, green dashed, blue lines)
Colored image can be obtained from the authors.

approximate the data clouds. Thus, each cluster is represented by more than one prototype, i. e. the respective requirement of the Conn-Index is fulfilled. Evidently, the fuzzy cluster solutions should be more adequate because of the overlapping data clouds.

The original and the Fuzzy Conn-Index were compared according to their performance and suitability, whereby the original Conn-Index was used for crisp clusterings. The Fuzzy Conn-Index takes the fuzzy assignments $u_{ij}$ of the data points to the prototypes into account. Comparing the values in Tab. 1 we observe that the Fuzzy Conn-Index yields better results for most of the fuzzy cluster solutions reflecting the fuzziness of the data clouds. This can mainly be dedicated to a lower separation value $C_{inter}$, i. e. the incorporation of fuzziness leads to a better description of the cluster separability. However, too strong fuzziness neglects this effect. Hence, a careful choice of the fuzziness parameter $m$, which should resemble the data overlap as truly as possible, is mandatory.

**Real world data set 'Coffee'.** This data set consists of hyper spectra of ten different untreated powder coffee sorts. The spectral measurement was done using a hyper spectral camera (HySpex SWIR-320 m-e, Norsk Elektro Optikk A/S) with the short range infra-red spectral range of 970nm to 2.500nm with a resolution of 6nm yielding 256 bands per spectrum.[1]One hundred spectra were generated for each coffee sort. A representative subset of the spectra is depicted in Fig.2. We trained crisp NG and the fuzzy FNG with 15 prototypes for each. The latter one with varying fuzziness parameter $m$, see Tab.2. An AP prototype clustering was performed resulting in three clusters for each run. Comparing the values obtained by applying the original Conn-Index to fuzzy clusterings ($m > 1$) with the respective Fuzzy Conn-Index values we notice significant discrepancies. This reflects the information loss which occurs during *crispification* of fuzzy values required for the calculation of the original Conn-Index. Again it can be

[1]Special thanks to Udo Seiffert and his team from the Fraunhofer Institute IFF Magdeburg for the collection of the data.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI 28001100967420.

| | NG | | FNG | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $m = 1.25$ | | $m = 1.5$ | | $m = 2.0$ | |
| $C$ | 0.74 | $C_c = 0.905$ | 0.61 | $C_c = 0.846$ | 0.69 | $C_c = 0.858$ | 0.49 | $C_c = 0.890$ |
| | | $C_s = 0.187$ | | $C_s = 0.274$ | | $C_s = 0.201$ | | $C_s = 0.452$ |
| $C_f$ | – | – | 0.77 | $C_c = 0.781$ | 0.72 | $C_c = 0.757$ | 0.63 | $C_c = 0.738$ |
| | | – | | $C_s = 0.018$ | | $C_s = 0.045$ | | $C_s = 0.144$ |

**Table 2:** The original Conn-Index $C$ and the Fuzzy Conn-Index $C_f$ with $C_{intra} = C_s$ and $C_{inter} = C_c$ for the cluster solutions of NG and FNG for the *Coffee* data.

observed that the choice of the fuzziness should be considered carefully. Differing from the frequently applied value $m = 2$ here a lower fuzziness parameter seems to perform better, as it was also obtained for the artificial data.

## 5 Conclusion

In this paper we extended the original cluster evaluation Conn-Index [8], designed for the evaluation of cluster solutions obtained from crisp prototype based vector quantization. This extension takes fuzzy vector quantization models into account and is based on the Generalized Conn-Index proposed in [11]. The new Fuzzy Conn-Index performs better for fuzzy cluster solutions than applying the original crisp variant to them. This is due to the information loss occurring in the calculation of the original Conn-Index if applied to fuzzy clusterings, since the new index takes explicitly all the fuzzy information provided by the fuzzy assignments into account. As the original, the Fuzzy Conn-Index requires more than one prototype per cluster. Further, the experiments have shown that the choice of the fuzziness parameter in the vector quantization has to be considered very carefully.

## References

[1] Geoffrey H. Ball and David J. Hall. A clustering technique for summarizing multivariate data. *Behavioral Science*, 12(2):153–155, 1967.
[2] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.
[3] Th. M. Martinetz, S. G. Berkovich, and K. J. Schulten. "neural-gas" network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, July 1993.
[4] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.
[5] N. B. Karayiannis and J. C. Bezdek. An integrated approach to fuzzy learning vector quantization and fuzzy c-means clustering. *IEEE Transactions on Fuzzy Systems*, 5(4):622–628, November 1997.
[6] M. Kästner and Th. Villmann. Fuzzy supervised neural gas for semi-supervised vector quantization – theoretical aspects. Machine learning report, University of Bielefeld, 2011.
[7] D.-W. Kim, K. H. Lee, and D. Lee. On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition*, 37:2009–2025, 2004.
[8] K. Taşdemir and E. Merényi. A validity index for prototype-based clustering of data sets with complex structures. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 41(4):1039–1053, August 2011.
[9] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, February 2007.
[10] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
[11] T. Geweniger, M. Kaestner, M. Lange, and T. Villmann. Derivation of a generalized conn-index for fuzzy clustering validation. In T. Villmann and F.-M. Schleif, editors, *Machine Learning Reports 07/2011*, pages 1–10, 2011.
[12] T.Villmann and J. C. Claussen. Magnification control in self-organizing maps and neural gas. *Neural Computation*, 18:446–469, 2006.