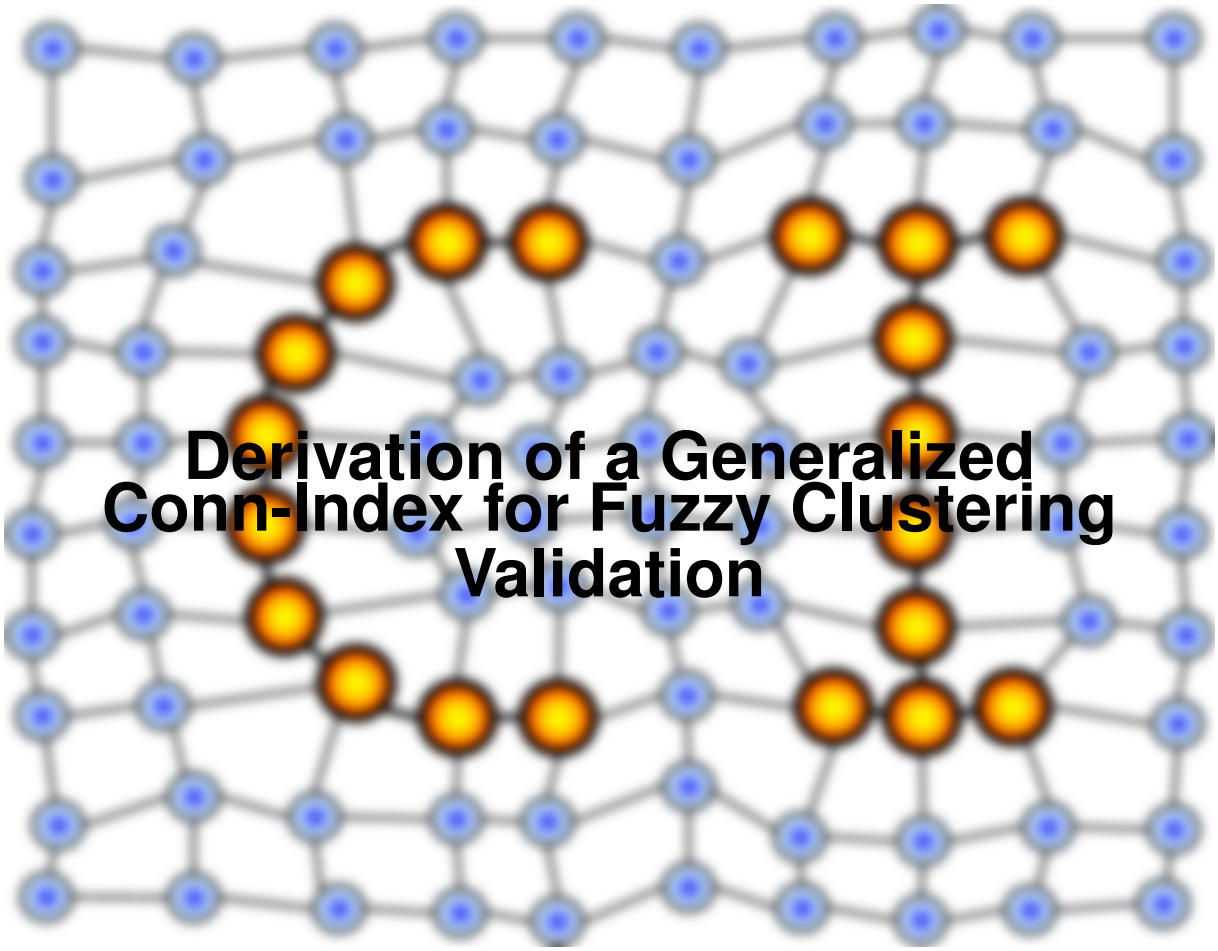
MACHINE LEARNING REPORTS



Derivation of a Generalized Conn-Index for Fuzzy Clustering Validation

Report 07/2011
Submitted: 17.10.2011
Published: 20.10.2011

T. Geweniger, M. Kästner, M. Lange, and T. Villmann*
(1) Computational Intelligence Group, University of Applied Sciences, Mittweida,
Technikumplatz 17, 09648 Mittweida, Germany
* corresponding author, *email: thomas.villmann@hs-mittweida.de*

Abstract

Clustering of very large data sets frequently requires a preprocessing such that the complexity of the clustering task is reduced. One method is to compress the information by prototype based crisp vector quantization and to cluster the prototypes subsequently. For the evaluation of such generated cluster solution the so-called Conn-index was proposed. In this paper we generalize this Conn-index such that it is also applicable to cluster solutions based on fuzzy vector quantizers, which allows a greater variability for vector quantization model selection.

Derivation of a Generalized Conn-Index for Fuzzy Clustering Validation

T. Geweniger, M. Kästner, M. Lange, and T. Villmann*

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

Abstract

Clustering of very large data sets frequently requires a preprocessing such that the complexity of the clustering task is reduced. One method is to compress the information by prototype based crisp vector quantization and to cluster the prototypes subsequently. For the evaluation of such generated cluster solution the so-called Conn-index was proposed. In this paper we generalize this Conn-index such that it is also applicable to cluster solutions based on fuzzy vector quantizers, which allows a greater variability for vector quantization model selection.

1 Introduction

Clustering of data is a challenging task. However, the evaluation of obtained cluster solutions is difficult because the task belongs to the class of ill-posed problems. Therefore, different cluster validity indices were developed reflecting different aspects [6, 7]. The Conn-index was designed for the evaluation of cluster solutions for very large data sets represented by prototypes [14]. In particular, the cluster solution generated on the basis of these prototypes is assessed using the information about the whole data set collected in the receptive fields of the prototypes.

*corresponding author, *email: thomas.villmann@hs-mittweida.de*

2 The Conn-Index

For the Conn-index C it is supposed that the data set $V = \{\mathbf{v}_l\}_{l=1}^{N} \subseteq \mathbb{R}^m$ is represented by a set $W = \{\mathbf{w}_i\}_{i=1}^{n} \subset \mathbb{R}^m$ of prototypes where N is huge and $n \ll N$. The prototypes are obtained by an arbitrary vector quantization algorithm based on a dissimilarity measure $d(\mathbf{v}_l, \mathbf{w}_i)$. Frequently, the quadratic Euclidean distance is chosen. We further assume that the prototypes itself are clustered into K clusters Ξ_k. Thereby, it is not important for the calculation of the Conn-index, how the cluster partition of the prototypes is obtained. However, the mapping information of data points to prototypes plays an essential role. This information is included using the best matching and second best matching prototype for a given data point to reflect the neighborhood relations between the prototypes are known from topology representing networks [10].

For a given data vector $\mathbf{v}_l$ the best matching prototype is determined by

$$s_0(\mathbf{v}_l) = \operatorname{argmin}_j \left(d(\mathbf{v}_l, \mathbf{w}_j) \right) \tag{1}$$

defining a crisp winner take all mapping rule (WTA). More general, a winning rank function for the ith prototype can be defined as

$$rk_i(\mathbf{v}_l, W) = \sum_{j=1}^{n} \Theta\left(d(\mathbf{v}_l, \mathbf{w}_i) - d(\mathbf{v}_l, \mathbf{w}_j) \right) \tag{2}$$

where

$$\Theta(x) = \begin{cases} 0 & if \quad x \leq 0 \\ 1 & else \end{cases} \tag{3}$$

is the Heaviside function [11]. Obviously, $rk_{s_0(\mathbf{v}_l)}(\mathbf{v}_l, W) = 0$ is the rank of the best matching prototype. Analogously, the jth winner is denoted by $s_{j-1}(\mathbf{v}_l)$ with $rk_{s_j(\mathbf{v}_l)}(\mathbf{v}_l, W) = j$. If it is clear from the context we will abbreviate $s_j = s_j(\mathbf{v}_l)$. The receptive field Ω_i of the ith prototype is then the non-empty set given by

$$\Omega_i = \{\mathbf{v}_l \in V \mid rk_i(\mathbf{v}_l, W) = 0\}. \tag{4}$$

The Conn-index weights separation and connectivity of the prototype clusters taking into account the receptive field information of the prototypes. For the estimation of the connectivity, first, the non-symmetric cumulative adjacency matrix $\mathbf{A}$

$$\mathbf{A} = \sum_{l=1}^{N} \mathbf{\Psi}(\mathbf{v}_l) \tag{5}$$

with respect to the receptive fields Ω_i and data set V is considered. Here, $\mathbf{\Psi}(\mathbf{v}_l)$ is the zero $(n \times n)$-matrix except the element

$$\Psi_{s_0,s_1} = c_1 \tag{6}$$

with a positive constant c_1, which is usually set to $c_1 = 1$. Yet, any other choice of c_1 would deliver exactly the same results for the Conn-index. The matrix $\mathbf{\Psi}(\mathbf{v}_l)$ is called the response matrix with respect to the data vector $\mathbf{v}_l$. As pointed out in [14], the row vector $\mathbf{a}_i = (a_{i,1}, \ldots, a_{i,n})$ of $\mathbf{A}$ describes the density distribution within the receptive field Ω_i with respect to the other $n-1$ prototypes.

Since the diagonal elements $a_{i,i}$ of the cumulative adjacency matrix $\mathbf{A}$ are never used for the calculation of the Conn-index, we can take these to store additional information about the winning frequency, i.e. the importance of the respective prototype. A possible choice could be $\Psi_{s_0,s_0} = c_0$ where $c_0 = \#\Omega_{s_0}$ is the size of the receptive field.

The symmetric connectivity matrix

$$\mathbf{C} = \mathbf{A} + \mathbf{A}^\top \tag{7}$$

reflects the topological relations between the prototypes based on the receptive field evaluation. Thereby, the elements $c_{i,j}$ reflect the dissimilarities between the prototypes based on the local data density.

The Conn-index C combines the intra-cluster connectivity $C_{intra} \in [0,1]$ and the inter-cluster connectivity $C_{inter} \in [0,1]$ balancing the overall cluster compactness and separation:

$$C = C_{intra} \cdot (1 - C_{inter}) \,. \tag{8}$$

The intra-cluster connectivity C_{intra} measures the compactness of the clusters. It is based on the cumulative adjacency matrix $\mathbf{A}$ from (5) and the average is taken over the local intra-cluster connectivities

$$C_{intra}(k) = \frac{\sum_{i,j|i\neq j} \{a_{i,j} \mid \mathbf{w}_i, \mathbf{w}_j \in \Xi_k\}}{\sum_{i,j|i\neq j} \{a_{i,j} \mid \mathbf{w}_i \in \Xi_k\}} \tag{9}$$

for each cluster Ξ_k, i.e. $C_{intra} = \sum_k C_{intra}(k)$. The greater the compactness of a cluster Ξ_k the closer their intra-connectivity $C_{intra}(k)$ is to one. The overall intra-cluster connectivity C_{intra} is obtained by averaging all local intra-connectivities $C_{intra}(k)$.

The inter-cluster connectivity C_{inter} evaluates the separation between the clusters. Analogously, it is the average $C_{inter} = \sum_k C_{inter}(k)$ over the local inter-cluster connectivities

$$C_{inter}(k) = \max_{1\leq l\leq K, k\neq l} C_{inter}(k,l) \tag{10}$$

of all clusters Ξ_k evaluating the separation of each cluster Ξ_k to the other cluster Ξ_l, $k \neq l$. Thereby, $C_{inter}(k,l)$ judges the separation of cluster Ξ_k to cluster Ξ_l based on the connectivity matrix $\mathbf{C}$ (7) and is defined as

$$C_{inter}(k,l) = \begin{cases} 0 & if \quad S_{k,l} = \emptyset \\ \frac{\sum_{i,j|i\neq j}\{c_{i,j}|\mathbf{w}_i \in \Xi_k, \mathbf{w}_j \in \Xi_l\}}{\sum_{i,j|i\neq j}\{c_{i,j}|\mathbf{w}_i \in S_{k,l}\}} & if \quad S_{k,l} \neq \emptyset \end{cases} \tag{11}$$

where the sets

$$S_{k,l} = \{\mathbf{w}_i \mid \mathbf{w}_i \in \Xi_k \wedge \exists \mathbf{w}_j \in \Xi_l : a_{i,j} > 0\}$$

describe the neighborhood relations between the clusters Ξ_k and Ξ_l based on the contained prototypes. Similar to C_{intra}, the value of C_{inter} increases with better separability.

The usefulness of the Conn-index C is extensively studied in [14]. For many cluster problems reflecting typical cluster situations this index is superior over most other cluster validity indices. Thus, the Conn-index can be seen as a standard for the validation of cluster solutions resulted from prototype based vector quantization.

3 Generalizations of the Conn-Index

The calculation of the Conn-index C in Sec. 2 takes only the best matching unit $s_0(\mathbf{v}_l)$ and the second best matching unit $s_1(\mathbf{v}_l)$ into account discarding any information provided by higher ranked prototypes. Yet, this information will be used in the following to achieve a more general cluster validity index.

3.1 The Generalized Conn-Index

For the generalized Conn-index C_g, which incorporates the higher winning ranks, the redefinition $\psi(\mathbf{v}_l)$ of the original response matrix $\mathbf{\Psi}(\mathbf{v}_l)$ now involves the *full* response of the whole vector quantizer model for a given input $\mathbf{v}_l$. The matrix $\psi(\mathbf{v}_l)$ is a zero matrix of the same size as $\mathbf{\Psi}(\mathbf{v}_l)$ except the row vector regarding the winner $s_0(\mathbf{v}_l)$. It is set to be

$$\psi_{s_0}(\mathbf{v}_l) = \mathbf{r}(\mathbf{v}_l) \tag{12}$$

where $\mathbf{r}(\mathbf{v}_l)$ is the so-called response vector of all prototype responses for a given input $\mathbf{v}_l$. The vector elements $r_i(\mathbf{v}_l)$ of the ith prototype are defined according to

$$r_i(\mathbf{v}_l) = \varphi_\sigma(rk_i(\mathbf{v}_l, W)) \tag{13}$$

with $\varphi_\sigma(x)$ being an arbitrary monotonically decreasing function in x and the winning ranks $rk_i(\mathbf{v}_l, W)$ are taken from (2). The parameter σ determines the range of influence and should be determined carefully. A simple choice of the function $\varphi_\sigma(x)$ would be the exponential function $\varphi_\sigma(x) = \exp\left(-\frac{1-\frac{id(x)}{n-1}}{2\sigma^2}\right)$. Yet, an alternative approach could be the direct utilization of the distances $d(\mathbf{v}_l, \mathbf{w}_i)$ instead of the winning ranks and $\varphi_\sigma(x)$. In case of vector quantization algorithms including neighborhood cooperativeness in learning like neural gas (NG, [11]) or self-organizing maps (SOM,[8]), the σ-parameter should be chosen accordingly in agreement to the neighborhood range used in the respective model.

Now, the generalized Conn-index C_g uses for the calculation of the cumulative adjacency matrix $\mathbf{A}$ in (5) the new response matrices $\psi(\mathbf{v}_l)$ instead of the original response matrices $\mathbf{\Psi}(\mathbf{v}_l)$.

The original Conn-index is obtained for the choice

$$\varphi(x) = \begin{cases} c_0 & for \quad x = 0 \\ c_1 & for \quad x = 1 \\ 0 & else \end{cases} \quad .$$

3.2 The Fuzzy Conn-Index

So far we assumed that the vector quantization model is based on a crisp mapping according to the WTA rule (1). In such models the response information of the network is collected in the response vector $\mathbf{r}(\mathbf{v}_l)$ based on the winning rank and, therefore, reflecting the topological relation between the prototypes. In fuzzy vector quantization algorithms like fuzzy c-means (FCM,[1, 5]), fuzzy variants of neural gas (FNG - [16]) or self-organizing map (FSOM, [3, 2, 4, 12, 13, 15]) this information is not longer available because data points are gradually assigned to all prototypes. This fuzzy mapping information is stored in the assignment variables $u_{l,i} \in [0, 1]$, $l = 1, \ldots, N$ and $i = 1, \ldots, n$, determining the fuzzy degree regarding the mapping of the data vector $\mathbf{v}_l$ onto the ith prototype. If the restriction $\sum_{i=1}^{n} u_{l,i} = 1$ holds, the vector quantizer realizes a probabilistic fuzzy quantization. Otherwise, the quantization is called possibilistic [9]. The crisp WTA mapping can be seen as probabilistic variant with assignments restricted to be $u_{l,i} \in \{0, 1\}$.

Now, we give an extension of the Conn-index C for such fuzzy vector quantization models. The idea is to use these fuzzy assignments instead of the response vector $\mathbf{r}(\mathbf{v}_l)$ for the determination of the response matrix $\mathbf{\Psi}(\mathbf{v}_l)$ involved in the calculation of the generalized Conn-index C_g. The resulting *fuzzy* Conn-index is denoted as C_f.

Generally, in fuzzy vector quantization schemes the information about the topographic relations between the prototypes is implicitly contained in the fuzzy assignments. Therefore, we collect all fuzzy assignments for a given data vector $\mathbf{v}_l$ in the fuzzy response vector $\mathbf{u}_l$. Obviously, the assignment vector $\mathbf{u}_l$ is comparable to the response vector $\mathbf{r}(\mathbf{v}_l)$ used for the generalized Conn-index C_g. Consequently, the best matching prototype s_0 for a given data vector can be seen as the prototype with the highest fuzzy assignment $u_{l,i}$:

$$s_0 = argmax_i \{u_{l,i}\}. \tag{14}$$

Now, the row vector $\psi_{s_0}(\mathbf{v}_l)$ of the redefined response matrix $\psi(\mathbf{v}_l)$ can simply be chosen as the fuzzy response vector $\mathbf{u}_l$:

$$\psi_{s_0}(\mathbf{v}_l) = \mathbf{u}_l. \tag{15}$$

Again, the cumulative adjacency matrix $\mathbf{A}$ is calculated as before for the original Conn-index C according to (5) and the further calculations remain unaffected.

Hence, the resulting new fuzzy Conn-index C_f is the counterpart of the generalized Conn-index C_g in case of fuzzy vector quantization models.

4 Conclusion

In this paper we propose generalizations of the Conn-index for the evaluation of cluster solution based on prototype based vector quantization. These generalizations concern, on the one hand, the incorporation of the full vector quantizer response for a given data point. On the other hand, cluster solutions based on fuzzy vector quantizers are considered. For this purpose, the assignment vectors for each data point determining the mapping probabilities to the prototypes are taken into account for the respective extension of the Conn-index. The article shows the theoretical justification of the new variants of the Conn-index.

References

[1] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York, 1981.

[2] J. C. Bezdek and N. R. Pal. A note on self-organizing semantic maps. *IEEE Transactions on Neural Networks*, 6(5):1029–1036, 1995.

[3] J. C. Bezdek and N. R. Pal. Two soft relatives of learning vector quantization. *Neural Networks*, 8(5):729–743, 1995.

[4] J. C. Bezdek, E. C. K. Tsao, and N. R. Pal. Fuzzy Kohonen clustering networks. In *Proc. IEEE International Conference on Fuzzy Systems*, pages 1035–1043, Piscataway, NJ, 1992. IEEE Service Center.

[5] T. Graepel, M. Burger, and K. Obermayer. Self-organizing maps: generalizations and new optimization techniques. *Neurocomputing*, 21(1–3):173–90, 1998.

[6] W. Hashimoto, T. Nakamura, and S. Miyamoto. Comparison and evaluation of different cluster validity measures including their kernelization. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 13(3):204–209, 2009.

[7] L. Kaufmann and P. Rousseuw. *Finding Groups in Data - A Introduction to Cluster Analysis*. John Wiley , Sons, 1990.

[8] T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[9] R. Krishnapuram and J. Keller. A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, 1(4):98–110, 1993.

[10] T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, 7(2), 1994.

[11] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[12] N. R. Pal, J. C. Bezdek, and E. C. K. Tsao. Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 4(4):549–557, 1993.

[13] N. R. Pal, J. C. Bezdek, and E. C. K. Tsao. Errata to Generalized clustering networks and Kohonen's self-organizing scheme. *IEEE Transactions on Neural Networks*, 6(2):521–521, March 1995.

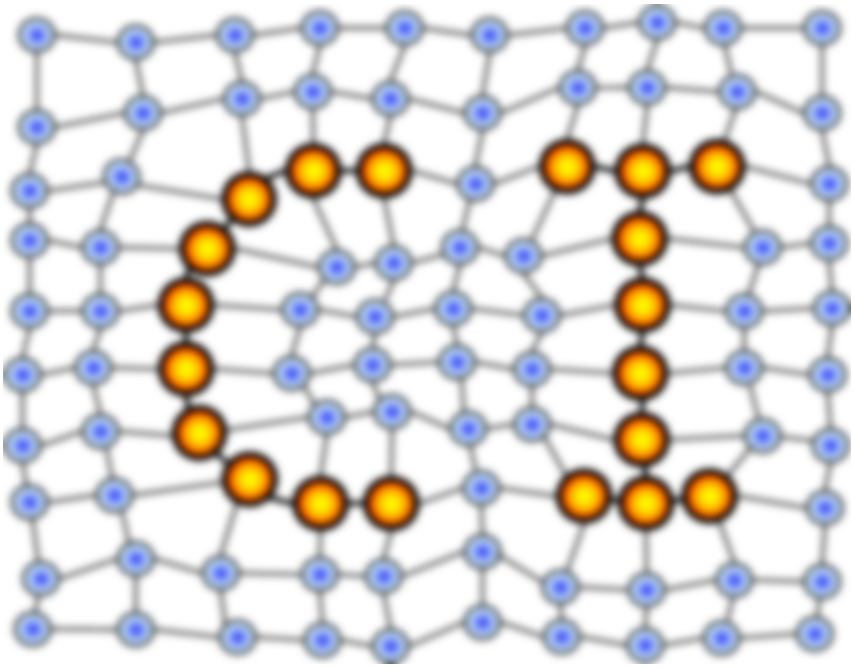
[14] K. Tasdemir and E. Merényi. A validity index for prototype-based clustering of data sets with complex cluster structures. *IEEE Transactions on Systems, Man, and Cybernetics*, 41(4):1039 – –1053, 2011.

[15] E. Tsao, J. Bezdek, and N. Pal. Fuzzy Kohonen clustering networks. *Pattern Recognition*, 27(5):757–764, 1994.

[16] T. Villmann, T. Geweniger, M. Kästner, and M. Lange. Theory of fuzzy neural gas for unsupervised vector quantization. *Machine Learning Reports*, 5(MLR-06-2011):27–46, 2011. ISSN:1865-3960, http://www.techfak.uni-bielefeld.de/~fschleif/mlr/mlr_06_2011.pdf.

MACHINE LEARNING REPORTS

Report 07/2011



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Dr. rer. nat. Frank-Michael Schleif
University of Bielefeld
Universitätsstrasse 21-23, 33615 Bielefeld, Germany
• http://www.cit-ec.de/tcs/about

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.