ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Optimization of Parametrized Divergences in Fuzzy c-Means

T. Geweniger[1,2], M. Kästner[2], T. Villmann[2]

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

2 - University of Applied Sciences Mittweida - Computational Intelligence Group
Dep. of Mathematics, Natural and Computer Sciences, Mittweida, Germany

Abstract. We propose the utilization of divergences as dissimilarity measure in the Fuzzy c-Means algorithm for the clustering of functional data. Further we adapt the relevance parameter to improve the data representation and therefore obtain more accurate clusterings in terms of separation and compactness. We show for two example applications that this method leads to improved performance.

1 Introduction

In machine learning the Fuzzy c-Means algorithm (FCM) plays an important role. This prototype based unsupervised clustering method has been extensively studied and applied to a great variety of problems from different research areas like medicine and biology. Commonly the Euclidean distance is used as dissimilarity measure, although any dissimilarity measure would be suited. Recently divergences are used instead [1, 2, 3, 4]. Further, relevance learning, i.e. weighting of input dimensions, was proposed for unsupervised vector quantization to improve cluster separation [5]. We transfer this idea to FCM using generalized divergences for relevance clustering of functional data. We denote this approach as Relevance FCM (R-FCM).

2 Relevance learning for the Fuzzy c-Means algorithm

2.1 Fuzzy c-Means

The standard Fuzzy c-Means algorithm was proposed in [6]. A given D-dimensional data set $V = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}, V \subseteq \mathbb{R}^D$ with N data points is partitioned into C clusters with prototypes $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C\}, W \subseteq \mathbb{R}^D$ while the objective function

$$J(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m d_{ij}^2 \tag{1}$$

is minimized. The set $U = \{u_{ij}\}$ defines a probabilistic cluster partition of V, where u_{ij} is an abbreviation for $u(\boldsymbol{v}_i, \boldsymbol{w}_j)$ and is subject to the constraints

$$\sum_{j=1}^{C} u_{ij} = 1 \quad \texttt{and} \quad u_{ij} \geq 0, \quad \forall i \in \{1, \ldots, N\}\,. \tag{2}$$

The exponent m regulates the *fuzziness.* $d_{ij} = d(\boldsymbol{v}_i, \boldsymbol{w}_j)$ is a dissimilarity function in $\mathbb{R}^D$. Note that the dissimilarity measure is fixed and most commonly the Euclidean distance is used.

The updates for u_{ij} and $\boldsymbol{w}_j$ follow an EM scheme with alternating calculation:

$$u_{ij} = \frac{d_{ij}^{\frac{-2}{m-1}}}{\sum_{l=1}^{c} d_{lj}^{\frac{-2}{m-1}}} \qquad\qquad \boldsymbol{w}_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} \boldsymbol{v}_i}{\sum_{i=1}^{n} u_{ij}^{m}} \tag{3}$$

Now we consider functional data $\boldsymbol{v}_i$, i.e. $\boldsymbol{v}$ is a discrete representation of a continous function $v(x)$. Usually these functional vectors are very high-dimensional and the vector components x are spatially correlated. Whereas for common Euclidean vectors the vector dimensions are treated independently.

If these functions are assumed to be positive with finite $\mathfrak{L}_1$-norm, the dissimilarity between such functions can be evaluated by (generalized) divergence measures taking into account the functional character [7]. If a certain divergence $D(\boldsymbol{v}||\boldsymbol{w})$ is used as dissimilarity measure, the cost function (1) changes to

$$J(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} D(\boldsymbol{v}_i||\boldsymbol{w}_j) \tag{4}$$

2.2 Relevance learning

We now explore the idea of relevance learning which consists in weighting the D input dimensions such that a better clustering is obtained. This can be realized for the Euclidean distance by weighting

$$d_{\lambda}(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{\lambda} \circ (\boldsymbol{v} - \boldsymbol{w})\|^2 = \|\boldsymbol{\lambda} \circ \boldsymbol{v} - \boldsymbol{\lambda} \circ \boldsymbol{w}\|^2 \tag{5}$$

where $\boldsymbol{\lambda} \circ \boldsymbol{v}$ and $\boldsymbol{\lambda} \circ \boldsymbol{w}$ are Hadamard products and the constraints $\lambda_i \geq 0$ and $\sum \lambda_i = 1$ are valid [8]. The λ_i are subject to optimization.

If for a cluster algorithm a divergence is used instead of the Euclidean distance we can also transfer the idea of relevance learning. For divergences we consider again the weighting of the prototypes as well as the data by the above mentioned Hadamard products.

For example consider the Kullback-Leibler divergence $D_{KL}(\boldsymbol{v}||\boldsymbol{w})$ with $\boldsymbol{v}$ and $\boldsymbol{w}$ assumed as densities, i.e. $v_i, w_j \geq 0$ and $\sum v_i = 1$ and $\sum w_j = 1$, where the positivity of $\boldsymbol{w}_j$ is assured by (3). If we now weight $\boldsymbol{v}$ by $\boldsymbol{\lambda}$, $\boldsymbol{\lambda} \circ \boldsymbol{v}$ is no longer a density but still a positive measure. Hence, for such data the generalized Kullback-Leibler divergence has to be used:

$$D_{KL}^{\lambda}(\boldsymbol{v}||\boldsymbol{w}) = \sum_{k=1}^{D} \lambda_k v_k \cdot \texttt{log}\left(\frac{v_k}{w_k}\right) - \sum_{k=1}^{D} (\lambda_k v_k - \lambda_k w_k) \tag{6}$$

Analogously the generalized Rényi divergence for relevance learning is

$$D_{R}^{\lambda}(\boldsymbol{v}||\boldsymbol{w}) = \frac{1}{\alpha - 1} \texttt{log}\left(\sum_{k=1}^{D} \left[\frac{(\lambda_k v_k)^{\alpha}}{(\lambda_k w_k)^{(\alpha-1)}} - \alpha \cdot \lambda_k v_k + (\alpha - 1) \cdot \lambda_k w_k\right] + 1\right) \tag{7}$$

A very robust divergence for positive measures is the γ-Divergence which yields for relevance learning as

$$D_{\gamma}^{\lambda}(\boldsymbol{v}||\boldsymbol{w}) = \mathtt{log}\left[\frac{(\sum_{k=1}^{D}(\lambda_k v_k)^{\gamma+1})^{\frac{1}{\gamma(\gamma+1)}} \cdot (\sum_{k=1}^{D}(\lambda_k w_k)^{\gamma+1})^{\frac{1}{\gamma+1}}}{(\sum_{k=1}^{D}(\lambda_k v_k)(\lambda_k w_k)^{\gamma})^{\frac{1}{\gamma}}}\right] \tag{8}$$

For $\gamma = 1$ the Cauchy-Schwarz divergence is obtained [9].

Generally, relevance learning is carried out by stochastic gradient descent learning (SGDL) which becomes for R-FCM

$$\frac{\partial J}{\partial \lambda_k} = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \frac{\partial D_{\lambda}(\boldsymbol{v}_i||\boldsymbol{w}_j)}{\partial \lambda_k}, \qquad \Delta\lambda_k = -\beta\frac{\partial J}{\partial \lambda_k}, 1 \leq k \leq D. \tag{9}$$

After relevance update, the prototypes have to be readjusted. Thus we get an alternating process of relevance and prototype adaptation.

For SGDL we need the derivations of the divergences with respect to λ_k which are

Generalized Kullback-Leibler divergence

$$\frac{\partial D_{KL}^{\lambda}(\boldsymbol{v}||\boldsymbol{w})}{\partial \lambda_k} = v_k \cdot \mathtt{log}\left(\frac{v_k}{w_k}\right) - v_k + w_k \tag{10}$$

Rényi divergence

$$\frac{\partial D_{R}^{\lambda}(\boldsymbol{v}||\boldsymbol{w})}{\partial \lambda_k} = \frac{w_k \cdot ((\frac{v_k}{w_k})^{\alpha} + \alpha - 1) - v_k \cdot \alpha}{(\alpha - 1)\sum_{l=1}^{N}\left[\lambda_l \cdot \left(w_l(\frac{v_l}{w_l})^{\alpha} - \alpha \cdot v_l + (\alpha - 1)\cdot w_l\right)\right] + 1} \tag{11}$$

γ-divergence

$$\frac{\partial D_{\gamma}^{\lambda}(\boldsymbol{v}||\boldsymbol{w})}{\partial \lambda_k} = \frac{v_k(\lambda_k v_k)^{\gamma}}{\gamma\sum\limits_{l=1}^{N}(\lambda_l v_l)^2} + \frac{w_k(\lambda_k w_k)^{\gamma}}{\sum\limits_{l=1}^{N}(\lambda_l w_l)^{\gamma+1}} - \frac{v_k(\gamma+1)(\lambda_k w_k)^{\gamma}}{\gamma\sum\limits_{l=1}^{N}(\lambda_l v_l)(\lambda_l w_l)^{\gamma}} \tag{12}$$

Note that during the adaptation the constraints for λ_k have to be kept and the dissimilarity measure is assumed to be fixed during the EM-like prototype learning, as above. Hence the adaptation of λ_k has to be performed in an adiabatic manner, such that the optimization process can easily follow this drift [10]. The adiabatic behavior is realized by very small learning rates β in (9).

3 Experimental results

In this section we demonstrate the effects of relevance parameter adaptation using different dissimilarity measures. For this purpose we used remote sensing data FLC1 [11] and the Wine data set [12]. The data sets were processed by the FCM algorithm using different dissimilarity measures: standard Euclidean distance, Kullback-Leibler divergence, Rényi divergence and Gamma divergence; the latter with varying values for γ. $\gamma = 1$ is equivalent to the Cauchy-Schwarz divergence and $\gamma = 0.5$ was chosen, because in [7] an improved clustering compared to the Cauchy-Schwarz divergence was obtained. The adapted relevance parameters are compared and discussed. For the analysis of the cluster solutions we applied a number of cluster validity indexes

- The *partition entropy* $PE = -\frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}\cdot\mathtt{log}_m(\mu_{ij})$ was introduced by Bezdek [13],[14] and is a measure of compactness. It is merely based on the fuzzy assigments and has no direct connection to the metric properties. The partition entropy is desired to obtain a low value.

- *Fukuyama&Seguno* [15] proposed a validity function which combines the compactness and the separation of a cluster solution $FS = \sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{m}d(v_i,w_j) - \sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{m}d(w_j,\bar{w}), \bar{w} = \sum_{j=1}^{c}w_j/c$ taking into account the metric properties of the distance measure. This index is desired to obtain a minimum.

- *Xie&Beni's* [16] validity measure $XB = \frac{\sum_{j=1}^{c}\sum_{i=1}^{n}\mu_{ij}^{m}d(v_i,w_j)}{n\cdot\mathtt{min}_{j\neq l}d(w_j,w_l)}$ is also based on the concept of compactness and separation, stating that a good clustering is obtained by minimizing the compactness and maximizing the separation. Therefore the aim is to achieve a low value for the index.

3.1 Remote sensing data

The Flightline C1 (FLC1) [11] remote sensing data was collected by an airborne scanner and consists of 11451 spectral bands in the range of 0.4 to 2.4 μm, which are assigned to 10 ground cover classes like corn, soybeans, wheat, and others.

We randomly choose 10 prototypes and after training we calculated the cluster validity index measures to show the impact of relevance parameter adaptation to the clustering. For means of comparability we always used the same prototype initializations and performed a total of 2500 relevance parameteter adaption steps with an initial learning rate of 0.001. Per adaption step we used only 10% of the data samples to achieve an adiabatic drift in the optimization process. After each relevance adaption step we executed the known FCM clustering to adapt the protoypes.

After completed clustering we calculated the before mentioned cluster validity indexes. All three measure show the expected behavior, i.e. they decrease after the adaption of the relevance parameter, which indicates an improved clustering in terms of compactness and separation. Detailed results depicting the index values with and without relevance adaption can be found in Tab. 1.

The relevance profiles in Fig. 1 indicate an emphasis on the lower spectral band, i.e. if the influence of these input dimenions during learning is increased, the results will be improved. Remarkable are the relevance profiles for the Kullback-Leibler and the Rényi divergence, since they are almost identical.

3.2 Wine data set

The Wine data set [12] contains 121 absorbing infrared spectra of wine between 4000 and 400 cm^{-1}. We used this data set ignoring the class information according to the alcohol level and clustered with 6 randomly chosen prototypes. Again we obtained lower values for the validity measures after the adaption of the relevance parameter indicating a better cluster solution, see lower part of

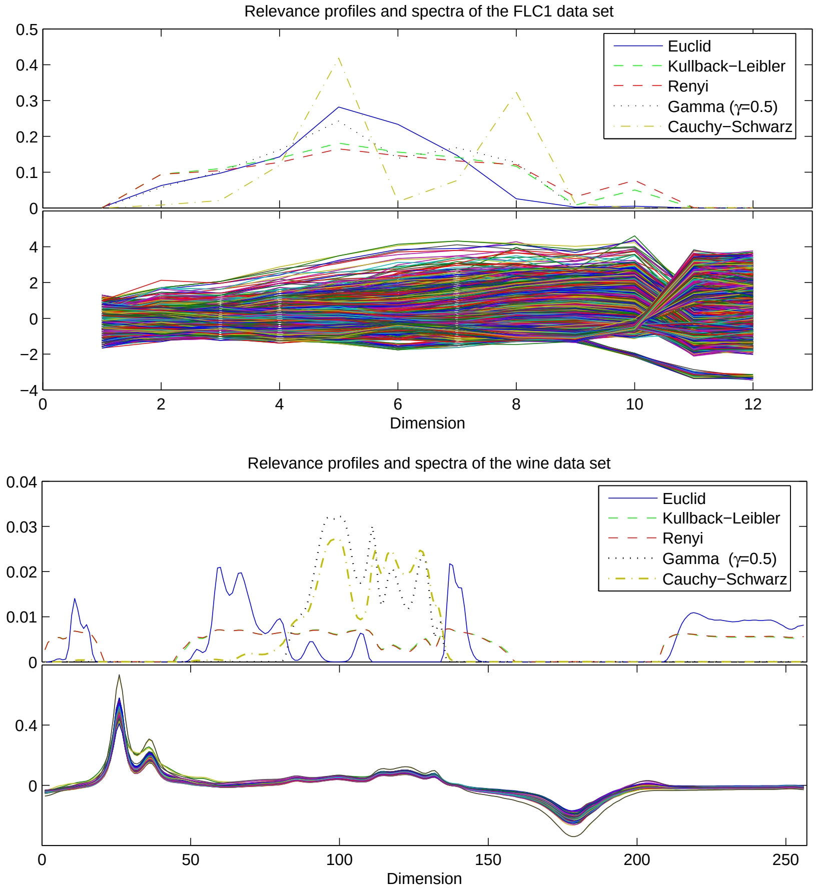

Fig. 1: Relevance profiles and spectra of the FLC1 and the Wine data set.

Tab. 1. The relevance profiles plotted in Fig. 1 show an agreement with the inverse variance of the data indicating an emphasis on those dimensions, for which the spectral bands lay close together. Remarkable are again the relevance profiles for the Kullback-Leibler and the Rényi divergence, which are again almost identical.

4 Conclusion

In this contribution we demonstrated, how different dissimilarity measures, namely Rényi, Kullback-Leibler and γ-divergences, can be incorporated in the FCM algorithm for clustering of functional data. Thereby, the derivatives of the divergences are used to adapt the relevance parameters. Hence, the metric is no longer fixed and can therefore be modulated to improve the representation of the data and to obtain a more accurate clustering in terms of compactness and separation. We demonstrated the improved performance of the R-FCM algorithm for two real life data sets.

Idx	Euclid	Rényi	Kull.-Leib.	γ-Div.(γ=0.5)	γ-Div.(γ=1.0)
					Cauchy-Schw.
PE	2.3026	2.3026	2.3026	2.3026	2.3026
	1.6516	1.4923	1.5492	1.4905	0.6474
FS	0.3392	255.0732	47.8509	464.2801	289.6427
	0.0355	4.1854	1.8236	5.7332	8.6354
XB	0.0107	0.0195	0.0126	0.0224	0.0110
	0.0054	0.0052	0.0052	0.0159	6.70e-04
PE	1.7918	1.7918	1.7918	1.7918	1.7918
	1.1592	1.3554	1.3525	1.1013	0.9657
FS	5.19e-07	0.4658	0.0433	6.8077	5.4929
	9.72e-10	2.27e-05	1.16e-05	9.60e-04	1.43e-05
XB	0.0491	0.0907	0.0528	0.0161	0.0153
	0.0070	0.0340	0.0339	0.0052	0.0057

Table 1: Validity index measures for the FLC1 (upper part) and the Wine data set (lower part) without (1st entry) and with (2nd entry) relevance parameter adaptation. A lower value indicates a better clustering in terms of separation and compactness.

References

[1] R. Inokuchi and S. Miyamoto. Fuzzy c-means algorithms using kullback-leibler divergence and helliger distance based on multinomial manifold. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 12(5):443–447, 2008.

[2] E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):158–171, February 2002.

[3] S. Rao, J. C. Sanchez, S. Han, and J. C. Principe. Spike sorting using non parametric clustering via cauchy schwartz pdf divergence. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 881–884, 14-19 May 2006.

[4] R. Jenssen, K. E. Hild, D. Erdogmus, J. C. Principe, and T. Eltoft. Clustering using Rényi's entropy. *Neural Computation*, 1:523–528, 20-24 July 2003.

[5] B. Arnonkijpanich, A. Hasenfuss, and B. Hammer. Local matrix learning in clustering and applications for manifold visualization. *Neural Networks*, 23(4):476–486, 2010.

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

[7] T. Villmann and S. Haase. Divergence based vector quantization. *Neural Computation*, page page in press, 2011.

[8] B. Hammer and T. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15:1059–1068, 2002.

[9] J. C. Principe, J.F.Ill, and D.Xu. *Unsupervised Adaptive Filtering*, chapter Information theoretic learning. Wiley, New York, NY, 2000.

[10] T. Kato. On the adiabatic theorem of quantum mechanics. *Journal of the Physical Society of Japan*, 5(6):435–439, 1950.

[11] D. Landgrebe. *Signal Theory Methods in Multispectral Remote Sensing.* Wiley, Hoboken, New Jersey, 2003.

[12] Wine data set provided by prof. marc meurens. available on http://www.ucl.ac.be/mlg/index.php?page=databases. meurens@bnut.ucl.ac.be.

[13] J. C. Bezdek. Numerical taxonomy with fuzzy sets. *Journal of Mathematics and Biology*, 1:57–71, 1974.

[14] J. C. Bezdek. Cluster validity with fuzzy sets. *Journal of Cybernetics*, 3:58–72, 1974.

[15] Y. Fukuyama and M. Sugeno. A new method of choosing the number of clusters for the fuzzy c-means method. *Proc. of the 5th Fuzzy Systems Symposium*, pages 247–250, 1989.

[16] X. L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(8):841–847, 1991.