Fuzzy Subspace Clustering

Christian Borgelt

Abstract In clustering we often face the situation that only a subset of the available attributes is relevant for forming clusters, even though this may not be known beforehand. In such cases it is desirable to have a clustering algorithm that automatically weights attributes or even selects a proper subset. In this paper I study such an approach for fuzzy clustering, which is based on the idea to transfer an alternative to the fuzzifier [15] to attribute weighting fuzzy clustering [14]. In addition, by reformulating Gustafson–Kessel fuzzy clustering, a scheme for weighting and selecting principal axes can be obtained. While in [5] I already presented such an approach for a global selection of attributes and principal axes, this paper extends it to a cluster-specific selection, thus arriving at a *fuzzy subspace clustering* algorithm.

Key words: fuzzy clustering, fuzzifier alternative, feature weighting, feature selection, subspace clustering

1 Introduction

A serious problem in distance-based clustering is that the more dimensions (attributes) a datasets has, the more the distances between data points—and thus also the distances between data points and constructed cluster centers—tend to become uniform. This, of course, impedes the effectiveness of clustering, as distance-based clustering exploits that these distances *differ*. In addition, in practice often only a subset of the available attributes is relevant for forming clusters, even though this may not be known beforehand. In such cases it is desirable to have a clustering algorithm that automatically weights the attributes or even selects a proper subset.

European Center for Soft Computing, c/ Gonzalo Gutiérrez Quirós s/n, 33600 Mieres, Spain, christian.borgelt@softcomputing.es

In general, there are three principles to do feature selection for clustering. The first is a *filter* approach (e.g. [8, 13]), which tries to assess and select features without any explicit reference to the clustering algorithm to be employed. The second is a *wrapper* approach (e.g. [7, 9, 6]), which uses a clustering algorithm as an evaluator for chosen feature subsets and may employ different search strategies for choosing the subsets to evaluate. The final approach tries to combine clustering and feature selection by pushing the feature selection method into the clustering algorithm (e.g. [19, 17]). It should also be noted that any feature weighting scheme (which may, in itself, employ any of these three principles) can be turned into a feature selection method by simply applying a weight threshold to the computed feature weights.

In this paper I study weighting and selecting features in fuzzy clustering [1, 2, 12, 4]. The core principle is to transfer the idea of an alternative to the fuzzifier [15] to attribute weighting fuzzy clustering [14]. By reformulating Gustafson–Kessel fuzzy clustering [11] this can even be extended to a scheme for weighting and selecting principal axes. While the basics of this approach were already introduced in [5] for global attribute weighting and selection, this paper extends this approach to a cluster-specific operation. By carrying out experiments on artificial as well as real-world data, I show that this approach works fairly well and may actually be very useful in practice, even though the fact that it needs a normal fuzzy clustering run for initialization (otherwise it is not sufficiently robust) still leaves room for improvement.

2 Preliminaries and Notation

Throughout this paper I assume that as input we are given an m-dimensional data set $\mathbf{X}$ that consists of n data points $\mathbf{x}_j = (x_{j1}, \ldots, x_{jm})$, $1 \leq j \leq n$. This data set may also be seen as a data matrix $\mathbf{X} = (x_{jk})_{1 \leq j \leq n, 1 \leq k \leq m}$, the rows of which are the data points. The objective is to group the data points into c clusters, which are described by m-dimensional cluster centers $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})$, $1 \leq i \leq c$. These cluster centers as well as the feature weights that will be derived (as they can be interpreted as cluster shape and size parameters) are jointly denoted by the parameter set $\mathbf{C}$. The (fuzzy) assignment of the data points to the cluster centers is described by a (fuzzy) membership matrix $\mathbf{U} = (u_{ij})_{1 \leq i \leq c, 1 \leq j \leq n}$.

3 Attribute Weighting

This section reviews two basic methods to compute attribute weights in fuzzy clustering. Its main purpose is to contrast these closely related methods and to set the stage for the attribute selection approach developed in this paper.

3.1 Axes-parallel Gustafson–Kessel Fuzzy Clustering

A very direct way to determine attribute weights is to apply axes-parallel Gustafson–Kessel fuzzy clustering [16]. In this case we have to minimize the objective function

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) \sum_{k=1}^{m} \sigma_{i,k}^{-2} \, (x_{jk} - \mu_{ik})^2$$

subject to $\forall i, 1 \leq i \leq c : \prod_{k=1}^{m} \sigma_{i,k}^{-2} = 1$ and the standard constraints $\forall j, 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$ and $\forall i, 1 \leq i \leq c : \sum_{j=1}^{n} u_{ij} > 0$. The inverse variances $\sigma_{i,k}^{-2}$ are the desired cluster-specific attribute weights, which have to be found by optimizing the objective function. The membership transformation function h is a convex function on the unit interval. Usually $h(u_{ij}) = u_{ij}^{\alpha}$ with a user-specified *fuzzifier* α (most often $\alpha = 2$) is chosen, but there are also other suggestions (for example [15]). As the methods discussed in this paper work with any choice of the function h, its exact form will be left unspecified in the following. The resulting update rules are

$$\forall i; 1 \leq i \leq c : \forall j; 1 \leq j \leq n : \qquad u_{ij} = \frac{d_{ij}^{\frac{2}{1-\alpha}}}{\sum_{k=1}^{c} d_{kj}^{\frac{2}{1-\alpha}}} \qquad \text{if } h(u_{ij}) = u_{ij}^{\alpha},$$

$$\text{where} \qquad d_{ij}^2 = \sum_{k=1}^{m} \sigma_{i,k}^{-2} (x_{jk} - \mu_{ik})^2$$

$$\forall i; 1 \leq i \leq c : \qquad \boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} h(u_{ij}) \, \mathbf{x}_j}{\sum_{j=1}^{n} h(u_{ij})} \qquad \text{and}$$

$$\forall i; 1 \leq i \leq c : \forall k; 1 \leq k \leq m : \qquad \sigma_{i,k}^2 = s_{i,k}^2 \Big(\prod_{r=1}^{m} s_{i,r}^2 \Big)^{-\frac{1}{m}},$$

$$\text{where} \qquad s_{i,k}^2 = \sum_{j=1}^{n} h(u_{ij}) (x_{jk} - \mu_{ik})^2.$$

3.2 Attribute Weighting Fuzzy Clustering

An alternative, but equally simple scheme to obtain attribute weights was suggested in [14]. The objective function to minimize is

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) \sum_{k=1}^{m} w_{ik}^{v} (x_{jk} - \mu_{ik})^2.$$

The constraints on the membership degrees u_{ij} are the same as for Gustafson–Kessel fuzzy clustering, but the attribute weight constraint now reads $\forall i; 1 \leq i \leq c : \sum_{k=1}^{m} w_{ik} = 1$. The additional parameter v controls the influence of the attribute weights in a similar way as the fuzzifier α (as in $h(u_{ij}) = u_{ij}^{\alpha}$) controls the influence of the membership degrees. The update rules for membership degrees and cluster centers coincide with those of Gustafson–Kessel fuzzy clustering. The weights are updated according to

$$\forall i; 1 \leq i \leq c : \forall k; 1 \leq k \leq m : \qquad w_{ik} = \frac{s_{i,k}^{\frac{2}{1-v}}}{\sum_{r=1}^{m} s_{i,r}^{\frac{2}{1-v}}}$$

with $s_{i,k}^2$ defined as in the preceding section. By rewriting the update rule of Gustafson–Kessel fuzzy clustering (see Section 3.1) as

$$\forall i; 1 \leq i \leq c : \forall k; 1 \leq k \leq m : \qquad \sigma_{i,k}^{-2} = \frac{s_{i,k}^{-2}}{(\prod_{r=1}^{m} s_{i,r}^{-2})^{-\frac{1}{m}}},$$

the similarities and differences become very obvious: they consist in a different normalization (sum instead of product) and the additional parameter v.

4 Attribute Selection

The methods reviewed above yield attribute weights either as inverse variances $\sigma_{i,k}^{-2}$ or directly as weights w_{ik}, $1 \leq i \leq c$, $1 \leq k \leq m$. It is important to note that in both cases it is impossible that any attribute weight vanishes. Therefore a modification of the approach is necessary in order to select attributes (which may be achieved by allowing attribute weights to become 0).

The core idea of the proposed attribute selection method is to transfer the analysis of the effect of the fuzzifier α (as in $h(u_{ij}) = u_{ij}^{\alpha}$) and its possible alternatives, as it was carried out in [15], to attribute weights. As [15] showed, it is necessary to apply a convex function $h(\cdot)$ to the membership degrees in order to achieve a fuzzy assignment. Raising the membership degrees u_{ij} to a user-specified power (namely the fuzzifier α) is, of course, such a convex function, but has the disadvantage that it forces all assignments to be fuzzy (that is, to differ from 0 and 1). The reason is that the derivative of this function vanishes at 0. If we want to maintain the possibility of crisp assignments, we rather have to choose a function h with $h'(0) > 0$.

With the approach of attribute weighting fuzzy clustering it becomes possible to transfer this idea to the transformation of the attribute weights. That is, instead of raising them to the power v as in [14], we may transform them by

$$g(x) = \alpha x^2 + (1 - \alpha)x \qquad \text{with } \alpha \in (0, 1].$$

The same function was suggested as an alternative transformation of the membership degrees in [15], and a fuzzy clustering algorithm was derived that allowed for crisp (and thus in particular: vanishing) memberships in case the distances of a data point to different clusters differed considerably. Here the idea is that the same method applied to attribute weights should allow us to derive a fuzzy clustering algorithm that assigns zero weights to some attributes, thus effectively selecting attributes during the clustering process.

However, as was also discussed in [15], the above function has the disadvantage that its parameter α is difficult to interpret and thus difficult to choose adequately. Fortunately, [15] also provided a better parameterization:

$$g(x) = \frac{1-\beta}{1+\beta}\,x^2 + \frac{2\beta}{1+\beta}\,x \qquad \text{with } \beta \in [0,1).$$

Generally, we now have to minimize the objective function

$$J(\mathbf{X},\mathbf{C},\mathbf{U}) = \sum_{i=1}^{c}\sum_{j=1}^{n} h(u_{ij}) \sum_{k=1}^{m} g(w_{ik})(x_{jk}-\mu_{ik})^2$$

subject to $\forall i : \sum_{k=1}^{m} w_{ik} = 1$ with $g(x) = \frac{1-\beta}{1+\beta}\,x^2 + \frac{2\beta}{1+\beta}\,x$ where $\beta \in [0,1)$. The constraints on the membership degrees (see Section 3.1), of course, also apply. This leads to the update rule

$$\forall i; 1 \le i \le c : \forall k; 1 \le k \le m : \qquad w_{ik} = \frac{1}{1-\beta}\left(\frac{1+\beta(m_{i\oplus}-1)}{\sum_{r=1;w_{ir}>0}^{m} s_{i,r}^{-2}} s_{i,k}^{-2} - \beta\right)$$

$$\text{where} \qquad m_{i\oplus} = \max\left\{ k \;\middle|\; s_{i,\varsigma(k)}^{-2} > \frac{\beta}{1+\beta(k-1)} \sum_{r=1}^{k} s_{i,\varsigma(r)}^{-2} \right\}.$$

Here $\varsigma(\cdot)$ is a function that describes the permutation of the indices that sorts the $s_{i,k}^{-2}$ into descending order (that is, $s_{i,\varsigma(1)}^{-2} \ge s_{i,\varsigma(2)}^{-2} \ge \ldots$).

5 Principal Axes Weighting

A standard problem of attribute weighting and selection approaches is that correlated attributes will receive very similar weights or will both be selected, even though they are obviously redundant. In order to cope with this problem, an approach in the spirit of principal component analysis may be used: instead of weighting and selecting attributes, one may try to find (and weight) linear combinations of the attributes, and thus (principal) axes of the data set. This section shows how the methods of Section 3 can be extended to principal axes weighting by reformulating Gustafson–Kessel fuzzy clustering so that the specification of the (principal) axes and their weights is separated.

5.1 Gustafson–Kessel Fuzzy Clustering

Standard Gustafson-Kessel fuzzy clustering uses a (cluster-specific) Mahalanobis distance, which is based on cluster-specific covariance matrices $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, c$. The objective function is

$$J(\mathbf{X}, \mathbf{C}, \mathbf{U}) = \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) (\mathbf{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_i),$$

which is to be minimized subject to the constraints $\forall i; 1 \leq i \leq c : |\boldsymbol{\Sigma}_i^{-1}| = 1$ (intuitive interpretation: fixed cluster volume) and the standard constraints $\forall j, 1 \leq j \leq n : \sum_{i=1}^{c} u_{ij} = 1$ and $\forall i, 1 \leq i \leq c : \sum_{j=1}^{n} u_{ij} > 0$. The resulting update rule for the covariance matrices $\boldsymbol{\Sigma}_i$ is

$$\begin{aligned} &\forall i; 1 \leq i \leq c : && \boldsymbol{\Sigma}_i = \mathbf{S}_i |\mathbf{S}_i|^{-\frac{1}{m}} \\ &\text{where} && \mathbf{S} = \sum_{j=1}^{n} h(u_{ij}) (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^\top. \end{aligned}$$

In order to obtain explicit weights for (principal) axes, we observe that, since the $\boldsymbol{\Sigma}_i$ are symmetric and positive definite matrices, they possess an eigenvalue decomposition $\boldsymbol{\Sigma}_i = \mathbf{R}_i \mathbf{D}_i^2 \mathbf{R}_i^\top$ with $\mathbf{D}_i = \mathrm{diag}(\sigma_{i,1}, \ldots, \sigma_{i,m})$ (i.e., eigenvalues $\sigma_{i,1}^2$ to $\sigma_{i,m}^2$) and orthogonal matrices $\mathbf{R}_i$, the columns of which are the corresponding eigenvectors.[1] This enables us to write the inverse of a covariance matrix $\boldsymbol{\Sigma}_i$ as $\boldsymbol{\Sigma}_i^{-1} = \mathbf{T}_i \mathbf{T}_i^\top$ with $\mathbf{T}_i = \mathbf{R}_i \mathbf{D}_i^{-1}$. As a consequence, we can rewrite the objective function as

$$\begin{aligned} J(\mathbf{X}, \mathbf{C}, \mathbf{U}) &= \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) (\mathbf{x}_j - \boldsymbol{\mu}_i)^\top \mathbf{T}_i \mathbf{T}_i^\top (\mathbf{x}_j - \boldsymbol{\mu}_i) \\ &= \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) ((\mathbf{x}_j - \boldsymbol{\mu}_i)^\top \mathbf{R}_i \mathbf{D}_i^{-1}) ((\mathbf{x}_j - \boldsymbol{\mu}_i)^\top \mathbf{R}_i \mathbf{D}_i^{-1})^\top \\ &= \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) \sum_{k=1}^{m} \sigma_{i,k}^{-2} \Big(\sum_{l=1}^{m} (x_{jl} - \mu_{il}) r_{i,lk} \Big)^2. \end{aligned}$$

In this form the scaling and the rotation of the data space that are encoded in the covariance matrices $\boldsymbol{\Sigma}_i$ are nicely separated: the former is represented by the variances $\sigma_{i,k}^2$, $k = 1, \ldots, m$ (or their inverses $\sigma_{i,k}^{-2}$), the latter by the orthogonal matrices $\mathbf{R}_i$. In other words: the inverse variances $\sigma_{i,k}^{-2}$ (the eigenvalues of $\boldsymbol{\Sigma}_i^{-1}$) provide the desired axes weights, while the corresponding eigenvectors (the columns of $\mathbf{R}_i$) indicate the (principal) axes.

[1] Note that the eigenvalues of a symmetric and positive definite matrix are all positive and thus it is possible to write them as squares.

5.2 Reformulation of Gustafson–Kessel Fuzzy Clustering

In order to transfer the approach of [14] and the one developed in Section 4, we start from the rewritten objective function, in which the scaling and the rotation of the data space are separated and thus can be treated independently. Deriving the update rule for the scaling factors $\sigma_{i,k}^{-2}$ is trivial, since basically the same result is obtained as for axes-parallel Gustafson–Kessel fuzzy clustering (see Section 3.1), namely

$$\sigma_{i,k}^2 = s_{i,k}^2 \Big(\prod_{r=1}^{m} s_{i,r}^2 \Big)^{-\frac{1}{m}},$$

with the only difference that now we have

$$s_{i,k}^2 = \sum_{j=1}^{n} h(u_{ij}) \Big(\sum_{l=1}^{m} (x_{jl} - \mu_{il}) r_{i,lk} \Big)^2.$$

Note that this update rule reduces to the update rule for axes-parallel Gustafson–Kessel clustering derived in Section 3.1 if $\mathbf{R}_i = \mathbf{1}$ (where $\mathbf{1}$ is an $m \times m$ unit matrix), which provides a simple sanity check of this rule.

In order to derive an update rule for the orthogonal matrix $\mathbf{R}_i$, we have to take into account that in contrast to how the covariance matrix $\mathbf{\Sigma}_i$ is treated in normal Gustafson–Kessel fuzzy clustering, there is an additional constraint, namely that $\mathbf{R}_i$ must be orthogonal, that is, $\mathbf{R}_i^\top = \mathbf{R}_i^{-1}$. This constraint can conveniently be expressed by requiring $\mathbf{R}_i \mathbf{R}_i^\top = \mathbf{1}$. Incorporating this constraint[2] into the objective function yields the Lagrange functional

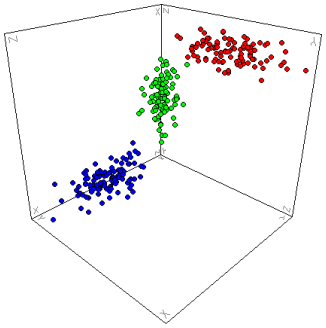
$$\begin{aligned} \mathcal{L}(\mathbf{X}, \mathbf{C}, \mathbf{U}, \mathbf{L}) = & \sum_{i=1}^{c} \sum_{j=1}^{n} h(u_{ij}) ((\mathbf{x}_j - \boldsymbol{\mu}_i)^\top \mathbf{R} \mathbf{D}^{-1}) ((\mathbf{x}_j - \boldsymbol{\mu}_i)^\top \mathbf{R} \mathbf{D}^{-1})^\top \\ & + \sum_{i=1}^{c} \operatorname{trace}\big(\mathbf{\Lambda}_i (\mathbf{1} - \mathbf{R}_i \mathbf{R}_i^\top) \big), \end{aligned}$$

where $\mathbf{L} = \{\mathbf{\Lambda}_1, \ldots, \mathbf{\Lambda}_c\}$ is a set of symmetric $m \times m$ matrices of Lagrange multipliers and $\operatorname{trace}(\cdot)$ is the trace operator, which for an $m \times m$ matrix $\mathbf{M}$ is defined as $\operatorname{trace}(\mathbf{M}) = \sum_{k=1}^{m} m_{kk}$. The resulting update rule[3] for the rotation matrices is $\mathbf{R}_i = \mathbf{O}_i$, where $\mathbf{O}_i$ is derived from the eigenvalue decomposition of $\mathbf{S}_i = \sum_{j=1}^{n} h(u_{ij}) (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^\top$, that is, from $\mathbf{S}_i = \mathbf{O}_i \mathbf{E}_i^2 \mathbf{O}_i^\top$ where $\mathbf{E}_i = \operatorname{diag}(e_{i,1}, \ldots, e_{i,m})$ is a diagonal matrix containing the eigenvalues.

---

[2] Note that, in principle, the orthogonality constraint alone is not enough as it is compatible with $|\mathbf{R}_i| = -1$, while we need $|\mathbf{R}_i| = 1$. However, the unit determinant constraint is automatically satisfied by the solution and thus we can avoid incorporating it.

[3] Note that this rule satisfies $|\mathbf{R}| = 1$ as claimed in the preceding footnote.



$\beta = 0.3$	X	Y	Z	U_1	U_2	U_3
$w_{1,k}$	0.63	0.37	—	—	—	—
$w_{2,k}$	0.13	—	0.87	—	—	—
$w_{3,k}$	—	0.29	0.71	—	—	—

$\beta = 0.4$	X	Y	Z	U_1	U_2	U_3
$w_{1,k}$	1.00	—	—	—	—	—
$w_{2,k}$	—	—	1.00	—	—	—
$w_{3,k}$	—	0.28	0.72	—	—	—

Fig. 1 Artificial data set with three Gaussian clusters and 300 data points.

6 Principal Axes Selection

In analogy to the transition from attribute weighting (Section 3) to attribute selection (Section 4), it is possible to make the transition from (principal) axes weighting (Section 5) to (principal) axes selection (this section): we simply replace the update rule for the weights (which are now separate from the axes) with the one obtained in Section 4. This leads to the update rule

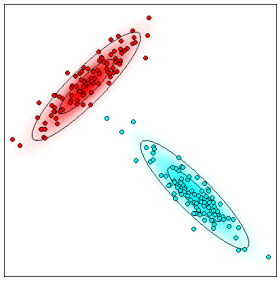
$$\forall i; 1 \leq i \leq c : \forall k; 1 \leq k \leq m : \quad w_{ik} = \frac{1}{1-\beta} \left(\frac{1 + \beta(m_{i\oplus} - 1)}{\sum_{r=1; w_{ir}>0}^{m} s_{i,r}^{-2}} s_{i,k}^{-2} - \beta \right).$$

with $s_{i,k}$ defined as in Section 5.2 and $m_{i\oplus}$ as defined in Section 4.

7 Experiments

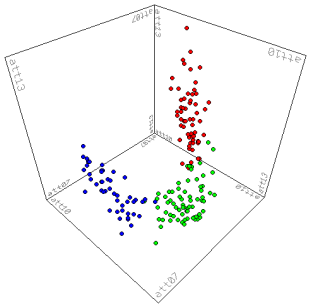
Of all experiments I conducted with the described method on various data sets, I report only a few here, due to limitations of space. Since experimental results for a global weighting and selection of attributes can be found in [5], I confine myself here to cluster-specific attribute weighting and selection.

Figures 1, 2 and 3 show two artificial and one real-world data set and the clustering results obtained on them. In all three cases the algorithm was initialized by axes-parallel Gustafson–Kessel fuzzy clustering (see Section 3.1), which was run until convergence. (Without such an initialization the results were not quite stable.) As can be seen from these results, the method is promising and may actually be very useful in practice. In all three cases uninformative attributes were nicely removed (received weights of zero or coefficients close to zero), while the informative attributes received high weights, which nicely reflect the structure of the data set.



$\beta = 0.5$	X	Y	U_1	U_2	U_3
$\mathbf{r}_1$	$\mathbf{-0.66}$	$\mathbf{0.76}$	0.02	-0.02	0.00
$\mathbf{r}_2$	$\mathbf{0.67}$	$\mathbf{0.74}$	0.00	-0.02	0.01

Fig. 2 Artificial data set with two Gaussian clusters and 200 data points.



- $\beta = 0.5$ selects attributes 2, 10, and 13 (one attribute per cluster).
- $\beta = 0.3$ selects the attribute sets $\{7, 10, 12\}$, $\{6, 7, 12, 13\}$, and $\{2\}$.
- Clustering the subspace spanned by attributes 7, 10 and 13 yields:

$\beta = 0.3$	att7	att10	att13
$w_{1,k}$	—	—	1.00
$w_{2,k}$	—	1.00	—
$w_{3,k}$	0.77	0.23	—

Fig. 3 The wine data set, a real-world data set with three classes and 178 data points. The diagram shows attributes 7, 10 and 13.

8 Summary

In this paper I introduced a method for selecting attributes in fuzzy clustering that is based on the idea to transfer an alternative to the fuzzifier, which controls the influence of the membership degrees, to attribute weights. This allows the attribute weights to vanish and thus effectively selects and weights attributes at the same time. In addition, a reformulation of Gustafson–Kessel fuzzy clustering separates the weights and the directions of the principal axes, thus making it possible to extend the scheme to a weighting and selection of principal axes, which helps in dealing with correlated attributes. Using this scheme in a cluster-specific fashion yields a *fuzzy subspace clustering* approach, in which each cluster is formed in its own particular subspace.

Software

The program used for the experiments as well as its source code can be retrieved free of charge under the GNU Lesser (Library) Public License at

`http://www.borgelt.net/cluster.html`

References

1. J.C. Bezdek and N. Pal. *Fuzzy Models for Pattern Recognition.* IEEE Press, New York, NY, USA 1992
2. J.C. Bezdek, J. Keller, R. Krishnapuram, and N. Pal. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.* Kluwer, Dordrecht, Netherlands 1999
3. C.L. Blake and C.J. Merz. *UCI Repository of Machine Learning Databases.* University of California, Irvine, CA, USA 1998
4. C. Borgelt. *Prototype-based Classification and Clustering.* Habilitation thesis, University of Magdeburg, Germany 2005
5. C. Borgelt. Feature Weighting and Feature Selection in Fuzzy Clustering. *Proc. 17th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE 2008, Hongkong, China).* IEEE Press, Piscataway, NJ, USA 2008
6. R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici. On Feature Selection Through Clustering. *Proc. 5th IEEE Int. Conf. on Data Mining (ICDM 2005, Houston, TX)*, 581–584. IEEE Press, Piscataway, NJ, USA 2005
7. M. Dash and H. Liu. Feature Selection for Clustering. *Proc. 4th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD 2000, Kyoto, Japan)*, 110–121. Springer, London, United Kingdom 2000
8. M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature Selection for Clustering: A Filter Solution. *Proc. 2nd IEEE Int. Conf. on Data Mining (ICDM 2002, Maebashi, Japan)*, 51–58. IEEE Press, Piscataway, NJ, USA 2002
9. J.G. Dy and C.E. Brodley. Visualization and Interactive Feature Selection for Unsupervised Data. *Proc. 6th ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD 2000, Boston, MA)*, 360–364. ACM Press, Ney York, NY, USA 2000
10. G.H. Golub and C.F. Van Loan. *Matrix Computations*, 3rd edition. The Johns Hopkins University Press, Baltimore, MD, USA 1996
11. E.E. Gustafson and W.C. Kessel. Fuzzy Clustering with a Fuzzy Covariance Matrix. *Proc. of the IEEE Conf. on Decision and Control (CDC 1979, San Diego, CA)*, 761–766. IEEE Press, Piscataway, NJ, USA 1979. Reprinted in [1], 117–122
12. F. Höppner, F. Klawonn, R. Kruse, and T. Runkler. *Fuzzy Cluster Analysis.* J. Wiley & Sons, Chichester, England 1999
13. P.-E. Jouve and N. Nicoloyannis. A Filter Feature Selection Method for Clustering. *Proc. 15th Int. Symp. on Foundations of Intelligent Systems (ISMIS 2005, Saratoga Springs, NY)*, 583–593. Springer, Heidelberg, Germany 2005
14. A. Keller and F. Klawonn. Fuzzy Clustering with Weighting of Data Variables. *Int. J. of Uncertainty, Fuzziness and Knowledge-based Systems* 8:735-746. World Scientific, Hackensack, NJ, USA 2000
15. F. Klawonn and F. Höppner. What is Fuzzy about Fuzzy Clustering? Understanding and Improving the Concept of the Fuzzifier. *Proc. 5th Int. Symp. on Intelligent Data Analysis (IDA 2003, Berlin, Germany)*, 254–264. Springer, Berlin, Germany 2003
16. F. Klawonn and R. Kruse. Constructing a Fuzzy Controller from Data. *Fuzzy Sets and Systems* 85:177–193. North-Holland, Amsterdam, Netherlands 1997
17. M.H.C. Law, M.A.T. Figueiredo, and A.K. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 26(9):1154–1166. IEEE Press, Piscataway, NJ, USA 2004
18. L. Parsons, E. Haque, and H. Liu. Subspace Clustering for High-Dimensional Data: A Review. *ACM SIGKDD Explorations Newsletter* 6(1):90-105. ACM Press, New York, NY, USA 2004
19. V. Roth and T. Lange. Feature Selection in Clustering Problems. *Advances in Neural Information Processing 16: Proc. 17th Ann. Conf. (NIPS 2003, Vancouver, Canada).* MIT Press, Cambridge, MA, USA 2004
20. X. Wang, Y. Wang, and L. Wang. Improving Fuzzy c-Means Clustering based on Feature-Weight Learning. *Pattern Recognition Letters* 25(10):1123–1132. Elsevier, New York, NY, USA 2004