Theoretical and Applied Informatics
ISSN 1896–5334
Vol.24 (2012), no. 4
pp. 313–326
DOI: 10.2478/v10179-012-0019-y

Clustering in Fuzzy Subspaces

KRZYSZTOF SIMIŃSKI

Institute of Informatics
Silesian University of Technology
ul. Akademicka 16, 44-100 Gliwice, Poland
Krzysztof.Siminski@polsl.pl

Received 16 March 2012, Revised 19 October 2012, Accepted 1 November 2012.

Abstract: Some data sets contain data clusters not in all dimension, but in subspaces. Known algorithms select attributes and identify clusters in subspaces. The paper presents a novel algorithm for subspace fuzzy clustering. Each data example has fuzzy membership to the cluster. Each cluster is defined in a certain subspace, but the the membership of the descriptors of the cluster to the subspace (called descriptor weight) is fuzzy (from interval $[0, 1]$) – the descriptors of the cluster can have partial membership to a subspace the cluster is defined in. Thus the clusters are fuzzy defined in their subspaces. The clusters are defined by their centre, fuzziness and weights of descriptors. The clustering algorithm is based on minimizing of criterion function. The paper is accompanied by the experimental results of clustering. This approach can be used for partition of input domain in extraction rule base for neuro-fuzzy systems.

Keywords: subspace clustering, weighted attributes, fuzzy clustering

1. Introduction

In high dimensional data sets some of the dimensions can be of minor importance. Some of them can be treated as noise and have lower importance may (or even should) be removed. The reduction of dimensionality may be done for whole data set (global dimensionality reduction) or individually for each cluster. The global approach by feature transformation (eg. Principal Component Analysis, PCA or Singular Value Decomposition, SVD) leads to problems with interpretability of elaborated models. Dimension reduction without feature transformation can be achieved by feature selection. The global selection of dimensions may not be satisfactory because different clusters may need different dimensions. This leads to subspace clustering [8, 10, 12] where each cluster may exist in different subspace.

There are two essential ways of subspace clustering: bottom-up and top-down [12]. The first approach splits the clustering space with a grid and analyses the density of data examples in each grid cell extracting the relevant dimensions (eg. CLIQUE [3], ENCLUS [5], MAFIA [11]). The latter (top-down) approach starts with full dimensional clusters and tries to throw away the descriptors (dimensions) of minor importance (eg. PROCLUS [1], ORCLUS [2], δ-Clusters [15], FSC [10, 9]). In the algorithms mentioned above the descriptor is valid or is not valid in a certain cluster what means that the weight of the descriptor in each cluster is 0 or 1. Our paper describes the modification of the FCM (fuzzy c-means) clustering algorithm [7] where the weights of descriptors are the values from the interval $[0, 1]$, so the descriptor can have partial membership to the subspace and the subspaces are defined in a fuzzy way in the input domain. This means that the descriptors can have partial importance in the subspace. This approach creates weighted dimension subspaces – such a partition of the domain can be used to create rules for the neuro-fuzzy system where the attributes can have various weight (importance) in different rules. The described algorithm is thought to be used in neuro-fuzzy systems that apply the clustering for input domain partition.

In the paper the empty uppercase characters $(\mathbb{T})$ are used to denote the sets, upper case italics (T) – the cardinality of sets, lower case bolds $(\mathbf{T})$ – vectors, upper case bolds $(\mathbf{t})$ – matrices, lower case italics (t) – scalars and set elements. The symbols used in the paper are listed in Tab. 1.

$\mathbb{C}$	set of clusters
C	number of clusters, $C = \|\mathbb{C}\|$
c	cluster, $c \in \mathbb{C}$
$\mathbb{X}$	set of tuples, data examples
X	number of tuples, $X = \|\mathbb{X}\|$
x	tuple, data example, $\mathrm{x} \in \mathbb{X}$
x_i	i-th tuple
x	descriptor of a tuple, $\mathrm{x} = [x_1, \ldots, x_A]^{\mathrm{T}}$
$\mathbb{A}$	set of attributes
A	number of attributes in a tuple, $A = \|\mathbb{A}\|$
a	attribute, $a \in \mathbb{A}$
μ	partition matrix, membership matrix
μ_{ci}	membership value of the i-th tuple to c-th cluster
d_{cj}	distance between c-th cluster's centre and j-th tuple
v_c	centre of c-th cluster
υ_{ca}	value of a-th attribute of c-th cluster's centre
s_c	fuzziness of c-th cluster
s_{ca}	fuzziness of a-th attribute of c-th cluster
z_c	weights of descriptors in c-th cluster
z_{ca}	weight of$a$-th attribute of c-th cluster
f	the fuzzification parameter

Tab 1. Symbols used in the papers

2. Clustering with weighted attributes

Our clustering method is based on minimising the criterion function J

$$J = \sum_{c=1}^{C} \sum_{i=1}^{X} \mu_{ci}^{m} \sum_{a=1}^{A} z_{ca}^{f} (x_{ia} - v_{ca})^2. \tag{1}$$

where m and $f \neq 1$ (the case of $f = 1$ is discussed on page 6) are parameters, $\boldsymbol{\mu}$ stands for cluster membership matrix (μ_{cx} denotes the membership of the x data tuple to the cth cluster), $\mathbf{Z}$ – matrix with attributes' weights (z_{ca} denotes the weight of the ath attribute (dimension) in the cth cluster) and $\mathbf{v}$ is centre of cluster defined as

$$\mathbf{v}_c = \frac{\sum_{i=1}^{X} \mu_{ci} \mathbf{x}_i}{\sum_{i=1}^{X} \mu_{ci}}. \tag{2}$$

Two constraints are put on dimension weights $\mathbf{Z}$ and partition matrix $\boldsymbol{\mu}$. These are:

1. The sum of dimension weights z for all dimensions A in each cluster c is equal to one:
$$\forall c \in [1, C] : \sum_{a=1}^{A} z_{ca} = 1. \tag{3}$$
2. The sum of membership values to all clusters for each data tuple is one:
$$\forall i \in [1, X] : \sum_{c=1}^{C} \mu_{ci} = 1. \tag{4}$$

The criterion function J (formula 1) is minimised with Lagrange multipliers. The Lagrange function can be expressed as follows:

$$\begin{aligned} L(\boldsymbol{\mu}, \lambda_1, \lambda_2) = \sum_{c=1}^{C} \sum_{k=1}^{X} \mu_{ck}^{m} \sum_{n=1}^{A} z_{cn}^{f} (x_{kn} - {v_{cn}}^2 + \\ -\lambda_1 \left[\sum_{n=1}^{A} z_{cn} - 1 \right] - \lambda_2 \left[\sum_{c=1}^{C} \mu_{ck} - 1 \right]. \end{aligned} \tag{5}$$

The derivatives are equal to zero:

$$\frac{\partial L}{\partial \mu_{ck}} = m \mu_{ck}^{m-1} \left(\sum_{a=1}^{A} zca^{f} (x_{ka} - v_{ca})^2 \right) - \lambda_2 = 0, \tag{6}$$

$$\frac{\partial L}{\partial z_{ca}} = \sum_{i=1}^{X} \mu_{ci}^{m} \, f \, z_{ca}^{f-1} (x_{ia} - v_{ca})^2 - \lambda_1 = 0. \tag{7}$$

Transforming Eq. 6 we get:

$$\mu_{ck} = \left(\frac{\lambda_2}{m}\right)^{\frac{1}{m-1}} \left(\sum_{a=1}^{A} z_{ca}^{f}(x_{ka} - \upsilon_{ca})^2\right)^{\frac{1}{1-m}}. \tag{8}$$

Substituting 8 into 4:

$$1 = \left(\frac{\lambda_2}{m}\right)^{\frac{1}{m-1}} \sum_{c=1}^{C} \left(\sum_{a=1}^{A} z_{ca}^{f}(x_{ka} - \upsilon_{ca})^2\right)^{\frac{1}{1-m}} \tag{9}$$

and finally dividing Eq. 8 by Eq. 9:

$$\mu_{ck} = \frac{\left(\sum_{a=1}^{A} z_{ca}^{f}(x_{ka} - \upsilon_{ca})^2\right)^{\frac{1}{1-m}}}{\sum_{j=1}^{C}\left(\sum_{a=1}^{A} z_{ja}^{f}(x_{ka} - \upsilon_{ja})^2\right)^{\frac{1}{1-m}}}. \tag{10}$$

Now the dimension weights are to be calculated. Transformation of Eq. 7:

$$\sum_{k=1}^{X} \mu_{ik}^{m}\, z_{ia}^{f-1}(x_{ka} - \upsilon_{ia})^2 = \frac{\lambda_1}{f},$$

$$z_{ia}^{f-1} \sum_{k=1}^{X} \mu_{ik}^{m}\,(x_{ka} - \upsilon_{ia})^2 = \frac{\lambda_1}{f}.$$

Further:

$$z_{ia}^{f-1} = \frac{\frac{\lambda_1}{f}}{\sum_{k=1}^{X} \mu_{ik}^{m}\,(x_{ka} - \upsilon_{ia})^2},$$

$$z_{ia} = \left(\frac{\frac{\lambda_1}{f}}{\sum_{k=1}^{X} \mu_{ik}^{m}\,(x_{ka} - \upsilon_{ia})^2}\right)^{\frac{1}{f-1}},$$

$$z_{ia} = \left(\frac{\lambda_1}{f}\right)^{\frac{1}{f-1}} \left(\frac{1}{\sum_{k=1}^{X} \mu_{ik}^{m}\,(x_{ka} - \upsilon_{ia})^2}\right)^{\frac{1}{f-1}}. \tag{11}$$

Substituting Eq. 11 to Eq. 3:

$$1 = \left(\frac{\lambda_1}{f}\right)^{\frac{1}{f-1}} \sum_{n=1}^{A} \left(\frac{1}{\sum_{k=1}^{X} \mu_{ik}^{m}\,(x_{ka} - \upsilon_{ia})^2}\right)^{\frac{1}{f-1}} \tag{12}$$

and dividing Eq. 11 by Eq. 12 we get:

$$z_{ia} = \frac{\left(\frac{1}{\sum_{k=1}^{X} \mu_{ik}^{m} (x_{ka} - v_{ia})^2}\right)^{\frac{1}{f-1}}}{\sum_{n=1}^{A} \left(\frac{1}{\sum_{k=1}^{X} \mu_{ik}^{m} (x_{kn} - v_{in})^2}\right)^{\frac{1}{f-1}}} \tag{13}$$

or

$$z_{ia} = \frac{\left(\sum_{k=1}^{X} \mu_{ik}^{m} (x_{ka} - v_{ia})^2\right)^{\frac{1}{1-f}}}{\sum_{n=1}^{A} \left(\sum_{k=1}^{X} \mu_{ik}^{m} (x_{kn} - v_{in})^2\right)^{\frac{1}{1-f}}}. \tag{14}$$

These values have to be converted into premises' parameters **v** and **s**. The centre **v** of ith cluster is calculated with formula 2. The fuzzification parameter $\mathbf{s}_i$ is calculated with formula 6:

$$\mathbf{s}_i = \sqrt{\frac{\sum_{k=1}^{X} \mu_{ik}^{m} (\mathbf{x}_k - \mathbf{v}_i)^2}{\sum_{k=1}^{X} \mu_{ik}^{m}}}. \tag{15}$$

Input: $\mathbb{X}$ – array of tuples
Input: $MaxIter$ – maximal number of iterations
Input: C – number of clusters
Output: $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_C]$ – centres of clusters
Output: $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_C]$ – fuzziness of clusters
Output: $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_C]$ – weights of descriptors for each cluster
// random initialisation of . . .
initialisation $\boldsymbol{\mu}$; *// . . . membership values*
initialisation **Z**; *// . . . attribute weights*
NumberOfIter $\leftarrow 0$; *// number of iterations*
while *NumberOfIter* $<$ *MaxIter* **do**
    update $\boldsymbol{\mu}$; *// Eq. 10*
    update **Z**; *// Eq. 14*
    V $\leftarrow$ calculateCentres ; *//Eq. 2*
    NumberOfIter $\leftarrow$ *NumberOfIter* + 1;
end of while
S $\leftarrow$ calculateFuzziness ; *// Eq. 15*
return;

Fig. 1. Clustering with attribute weights

Alternating application of formulae 2, 10 and 14 leads to algorithm presented in Fig. 1.

The procedure described above cannot be used if $f = 1$. In such situation the objective function 1 becomes

$$J = \sum_{c=1}^{C} \sum_{i=1}^{X} \mu_{ci}^{m} \sum_{a=1}^{A} z_{ca}(x_{ia} - v_{ca})^2. \tag{16}$$

The attribute a of the cth rule for which the sum

$$\sum_{i=1}^{X} \mu_{ci}^{m} z_{ca}(x_{ia} - v_{ca})^2 \tag{17}$$

is minimal gets the weight $z_{ca} = 1$ and other attributes of this rule get zero weights (because of the constraint expressed by formula 3).

3. Experiments

The experiments were conducted on real-life and synthetic data sets.

3.1. Data sets

The real life data sets depict methane concentration (Sec. 3.1.1.), death rate (Sec. 3.1.2.), Wisconsin breast cancer (Sec. 3.1.3.) and concrete compressive strength (Sec. 3.1.4.). All real life data sets are normalised.

For verification of subspace identification two synthetic data sets were used (Sec. 3.1.5.). The synthetic datasets are not normalised.

3.1.1. Methane concentration

The data set contains the real life measurements of air parameters in a coal mine in Upper Silesia (Poland). The parameters (measured in 10 second intervals) are: AN31 – the flow of air in the shaft, AN32 – the flow of air in the adjacent shaft, MM32 – concentration of methane (CH_4), production of coal, the day of week. To the tuples the 10-minute sums of measurements of AN31, AN32, MM32 are added as dynamic attributes [13].

The task is to predict the concentration of the methane in 10 minutes. The data is divided into train set (499 tuples) and test set (523 tuples).

3.1.2. Death rate

The data represent the tuples containing information on various factors, the task is to estimate the death rate [14]. The first attribute (the index) is excluded from the dataset.

The precise description of the attributes is avalaible with the data set, the names of the attributes are listed in Tab. 2, so the description is omitted here[1].

3.1.3. Breast Cancer Wisconsin

The data set represents the data for breast cancer case [4]. Each data tuple contains 32 continuous attributes and one predictive attribute (the time to recur). Here again we will omit the desciption of attributes, their names are listed in Tab. 3. The symbol 'se' in attribute name stand for 'standard error' and the adjective 'worst' means the 'largest'[2].

3.1.4. Concrete Compressive Strength

The Concrete Compressive Strength set is a real life data set describing the parameters of the concrete sample and its strength [16]. The attributes are: cement ratio, amount of blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, age; the decision attribute is the concrete compressive strength.

3.1.5. Synthetic data sets

The synthetic data sets have two clusters in subspaces. The first data set $'g123'$ has 5 attributes (dimensions) and two clusters. The first cluster has points generated with Gauss distribution (mean $m = 5$ and standard deviation $\sigma = 1$) in dimensions 1, 2 and 3. The second cluster (in subspace created by dimensions 3, 4 and 5) is generated with Gauss distribution ($m = 10, \sigma = 1$). Unused attribute values are filled with uniform distribution from interval $[0, 15]$.

The second synthetic data set $'g136'$ is created in similar way with two clusters in subspaces (dimensions: 1-3-6 and 2-4-6). The clusters are generated with Gauss distribution ($m = 1, \sigma = 1$ and $m = 9, \sigma = 1$ respectively). The not used attribute values are filled with uniform distribution from interval $[0, 10]$. These data sets are not normalised.

3.2. Results

The synthetic data sets ($'g123'$ and $'g136'$) were used to verify the extraction of subspaces for clusters. The Figures 2 and 3 present the elaborated clusters. The representation is symbolical. It means that two features of the cluster: membership μ of the data tuple and weight z are shown in a combined way. The figures present the product $\mu \cdot z$ instead of separate figures of μ and z. This approach is only used for better representation in one figure. The descriptor's weight has no influence on the membership

[1] The data can be downloaded from http://orion.math.iastate.edu/burkardt/data/regression/x28.txt.

[2] The data can be downloaded from http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

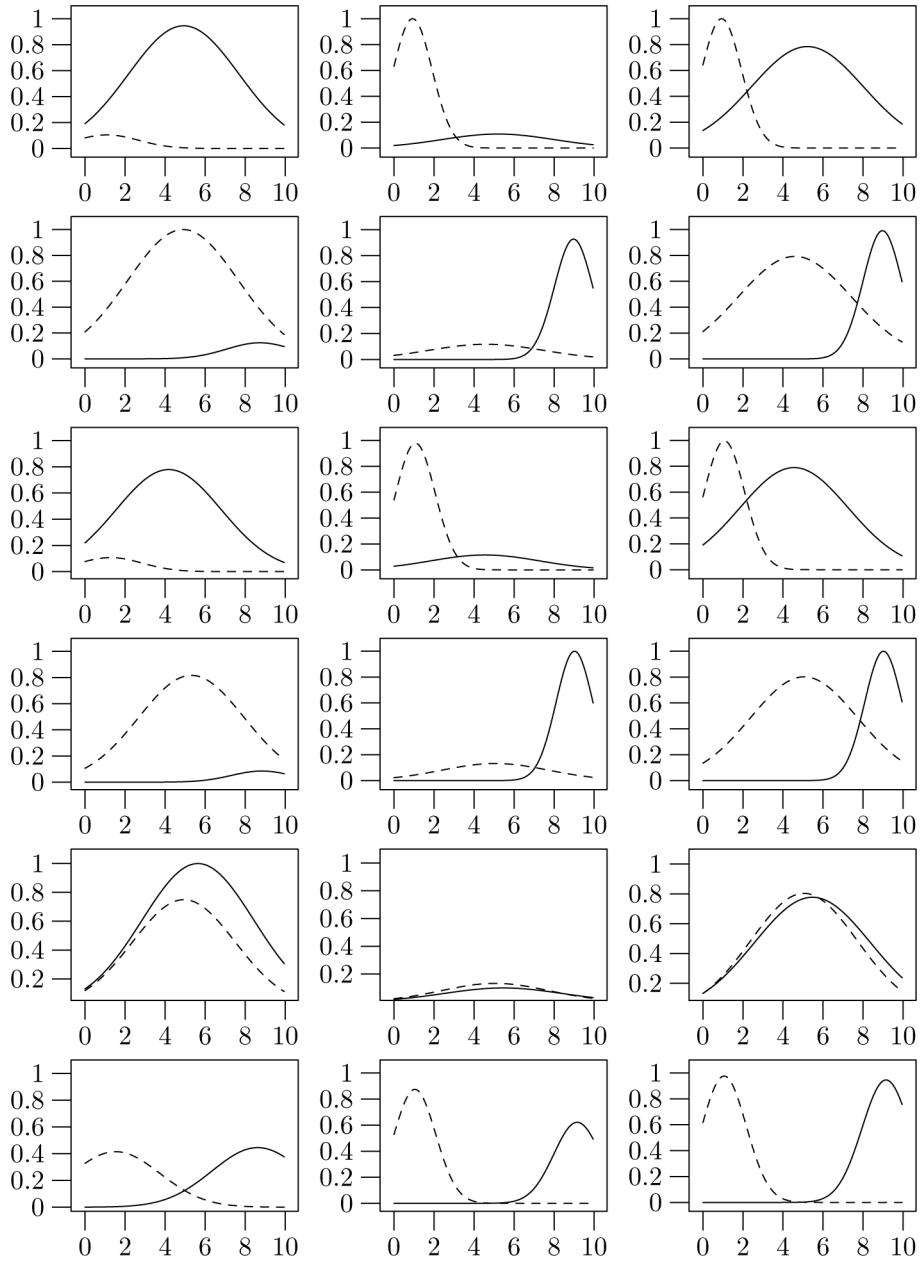

Fig. 2. The results of clustering of $'g136'$ data set with various values of f (in the first column: $f = 0.5$, second column $f = 2$, third column $f = 10$); first row – first attribute, sixth row – sixth attribute. The representation of cluster's membership functions is symbolical – the membership function is combined with the data tuple's weight (see details at the beginning of the section 3.2.). The first cluster is in 1-3-6 subspace $(m = 1, \sigma = 1)$ and the second one in 2-4-6 $(m = 9, \sigma = 1)$

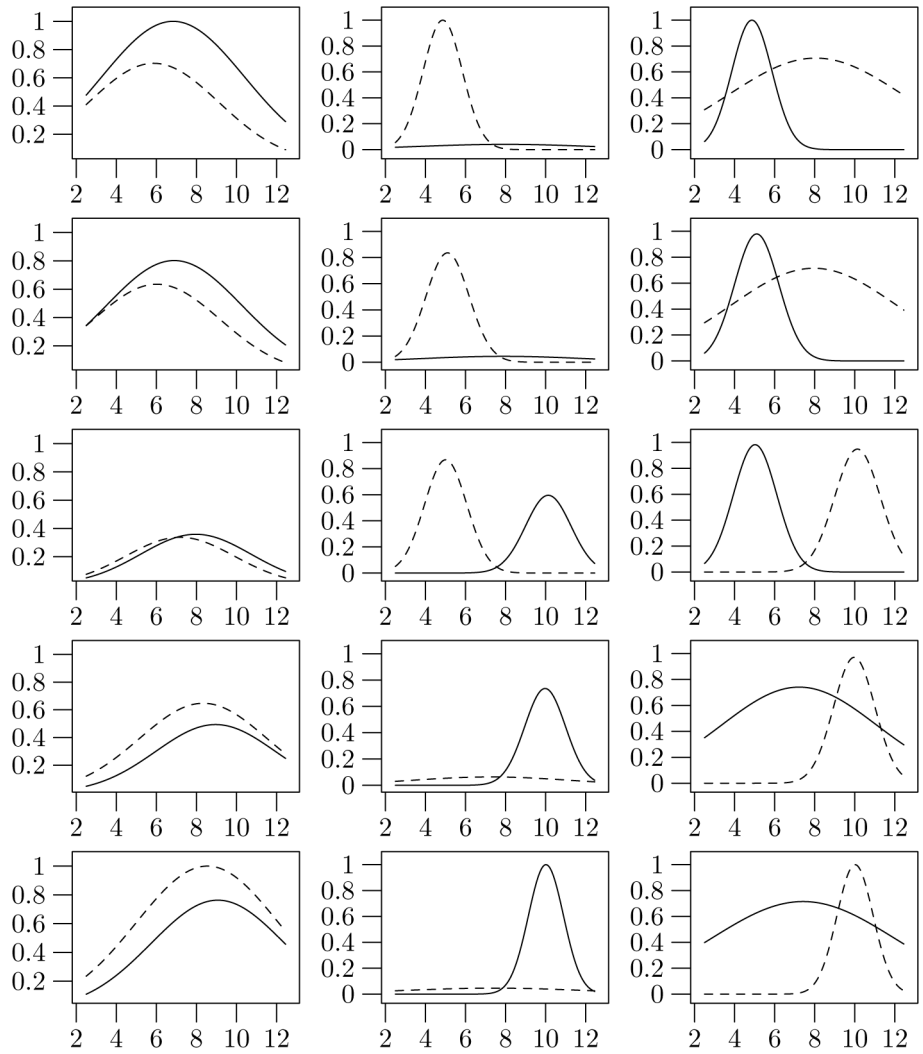

Fig. 3. The results of clustering of $'g123'$ data set with various values of f (in the first column: $f = 0.5$, second column $f = 2$, third column $f = 10$); first row – first attribute, sixth row – sixth attribute. The representation of cluster's membership functions is symbolical – the membership function is combined with the data tuple's weight (see details at the beginning of the section .2.). The first cluster is in 1-2-3 subspace $(m = 5, \sigma = 1)$ and the second one in 3-4-5 $(m = 10, \sigma = 1)$

of the descriptor. This remark should be always taken into consideration when analysing the figures mentioned above.

The Figures 2 and 3 show the influence of f (cf. the criterion function, Eq. 1) on the importance of the attributes. The values $f < 1$ lead to clusters of low reliability. The clusters are not identified correctly (the figures present only the clusters for $f = 0.5$ but similar behaviour can be observed for other values of f parameter). This can be observed in left columns of Figures 2 and 3.

If the values of f are greater than 1 the reliability of clusters is remarkably greater. The middle column in Figures 2 and 3 present the clusters for $f = 2$. For this value the clusters are correctly identified. High values of weights are assigned to the descriptors that build the identified subspace. The other attributes get low values of weights (importance). This is a desired behaviour. The greater the values of f parameter, the greater the weight (importance) assigned to the attributes that do not build the subspace. This can be noticed when the right and middle columns of Figures 2 and 3 are compared. Greater weights assigned to the attributes that do not belong to the subspace is not expected behaviour (cf. last column in Fig. 2 and 3). If $f = 1$, then only one descriptor gets weight equal to 1, all other are kept nought. Thus the values of the f parameter should be kept low but greater than one. When $f = 2$ the fifth attribute in $'g136'$ data set is assigned with very low weights (this attribute does not belong to either subspace). The experiments on clustering in subspaces lead to conclusion that the f parameter should be kept low but greater than 1.

	attributes' weights in clusters				
attribute	I	II	III	IV	V
average annual precipitation	0.002	0.000	0.002	0.090	0.001
average January temperature	0.007	0.000	0.001	0.328	0.001
average July temperature	0.003	0.000	0.001	0.022	0.001
size of the population older than 65	0.020	0.000	0.000	0.006	0.002
number of members per household	0.009	0.000	0.001	0.002	0.002
years of schooling for persons over 22	0.008	1.000	0.001	0.189	0.001
households with fully equipped kitchens	0.001	0.000	0.001	0.012	0.001
population per square mile	0.002	0.055	0.001	0.005	0.001
size of the nonwhite population	0.002	0.000	0.001	1.000	0.001
number of office workers	0.066	0.013	0.001	0.054	0.001
families with an income $< \$3000$	0.002	0.001	0.001	0.011	0.001
hydrocarbon pollution index	1.000	0.000	1.000	0.056	1.000
nitric oxide pollution index	0.269	0.000	0.440	0.023	0.110
sulphur dioxide pollution index	0.003	0.000	0.021	0.001	0.004
degree of atmospheric moisture	0.015	0.000	0.002	0.103	0.000

Tab 2. Weights of attributes elaborated for 'death rate' data set.

The Tables 2, 3, 4 and 5 present the weights of attributes in models elaborated for real life data sets. The attributes' weights for 'methane' data set gathered in Tab. 4

show that one of the most important attributes is the flow of air in the mine shaft. It is interesting that the actual concentration of methane in the mine shaft has low weights in all clusters. On the other hand the production of coal is one of the most important. It is quite reasonable because the excavation of coal causes tensions and splits in the rock that may release the methane gas. In two clusters the most important attribute is the first one, the flow of the air in the shaft in question. In the fifth cluster the interesting thing can be observed. The most important descriptor is 10-minute sum of the first attribute (flow of the air) whereas the first attribute itself has lower weight. The similar situation is to be observed in the case of second attribute (flow of air in the adjacent shaft).

	attributes' weights in clusters				
attribute	I	II	III	IV	V
lymph_node	1.000	0.008	0.030	0.061	0.044
radius_mean	0.001	0.009	0.277	0.208	0.953
texture_mean	0.001	0.015	0.037	0.147	0.039
perimeter_mean	0.001	0.009	0.293	0.194	1.000
area_mean	0.001	0.008	0.470	0.163	0.910
smoothness_mean	0.000	0.035	0.051	0.198	0.047
compactness_mean	0.000	0.016	0.041	0.241	0.039
concavity_mean	0.000	0.021	0.061	0.108	0.059
concave_points_mean	0.001	0.012	0.090	0.160	0.054
symmetry_mean	0.000	0.037	0.041	0.123	0.033
fractal_dimension_mean	0.001	0.022	0.037	0.431	0.046
radius_se	0.001	0.028	0.022	0.223	0.059
texture_se	0.000	0.052	0.028	0.431	0.034
perimeter_se	0.000	0.029	0.026	0.252	0.052
area_se	0.000	0.019	0.064	0.300	0.076
smoothness_se	0.001	0.094	0.076	0.543	0.025
compactness_se	0.000	0.040	0.069	0.356	0.064
concavity_se	0.000	0.049	0.052	0.190	0.048
concave_points_se	0.001	0.085	0.035	0.123	0.032
symmetry_se	0.000	1.000	0.051	0.175	0.066
fractal_dimension_se	0.000	0.030	0.044	0.608	0.032
radius_worst	0.001	0.009	0.544	0.196	0.379
texture_worst	0.000	0.023	0.038	0.165	0.077
perimeter_worst	0.001	0.010	0.336	0.204	0.197
area_worst	0.001	0.009	1.000	0.156	0.332
smoothness_worst	0.000	0.020	0.033	0.231	0.079
compactness_worst	0.000	0.016	0.029	0.482	0.079
concavity_worst	0.001	0.018	0.021	0.174	0.104
concave_points_worst	0.001	0.016	0.036	0.145	0.052
symmetry_worst	0.000	0.057	0.028	0.185	0.050
fractal_dimension_worst	0.001	0.014	0.021	1.000	0.106
tumor_size	0.003	0.016	0.023	0.051	0.044

Tab 3. Weights of attributes elaborated for 'wisconsin' data set.

	attributes' weights in clusters				
attribute	I	II	III	IV	V
AN31: flow of air in the shaft	1.000	0.000	0.132	1.000	0.361
AN32: flow of air in the adjacent shaft	0.009	0.000	0.087	0.002	0.102
MM32: concentration of methane	0.004	0.000	0.065	0.002	0.058
production of coal	0.028	1.000	1.000	0.011	0.885
sum of AN31	0.020	0.000	0.156	0.005	1.000
sum of AN32	0.009	0.000	0.085	0.004	0.291
sum of MM32	0.005	0.000	0.093	0.002	0.059

Tab 4. Weights of attributes elaborated for 'methane' data set.

	attributes' weights in clusters				
attribute	I	II	III	IV	V
cement ratio	0.000	0.065	0.006	0.065	0.005
blast furnace slag	0.000	0.050	0.019	0.535	1.000
fly ash	1.000	0.008	0.121	0.332	0.007
water	0.000	0.013	0.190	0.086	0.003
superplasticizer	0.000	0.057	0.057	0.112	0.004
coarse aggregate	0.000	0.028	0.006	0.090	0.003
fine aggregate	0.000	1.000	1.000	0.063	0.002
age	0.000	0.027	0.031	1.000	0.004

Tab 5. Weights of attributes elaborated for 'concrete' data set.

For 'concrete' data set in four clusters the most important attributes (all other have low weights) are the ratio of fly ash, fine aggregate, blast furnace slag and concrete age. In one cluster the weights are more varied: the most important is age, but quite high weights have concentration of blast furnace slag and fly ash.

Two clusters for the 'wisconsin' data set (Tab. 3) have among important attributes the radius (mean and standard deviation) and area (mean and standard deviation) of the lesion. In one cluster the weights are more varied. The important attributes are fractal dimension (worst and standard deviation and mean), smoothness (mean and standard deviation). In one cluster the most important are the lymph nodes – what is in concordance with medical diagnose procedures.

4. Summary

In high dimensional data sets some of the dimensions can be of minor importance. The global selection of dimensions may not be satisfactory because different clusters may need different dimensions.

The paper describes the novel clustering algorithm with weighted attributes (subspace clustering). The clustering algorithm is based on minimising of the criterion function. The clustering procedure elaborates not only the cluster centres and fuzziness but

also the weights of descriptors in each cluster. The weights of attributes are numbers from interval $[0, 1]$. This means that the attributes (dimensions) can have partial membership to the subspace.

The experiments confirm the proper subspace clustering both for synthetic and artificial data sets. The algorithm can be used for identification of rule base for fuzzy and neuro-fuzzy systems.

Acknowledgements

This work was supported by the European Union from the European Social Fund (grant agreement number: UDA-POKL.04.01.01-00-106/09).

The author is grateful to the anonymous referees for their constructive comments that have helped to improve the paper.

References

1. Ch.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, and J. S. Park: *Fast algorithms for projected clustering*, SIGMOD Rec., 28(2):61–72, 1999.
2. Ch.C. Aggarwal, and P.S. Yu: *Finding generalized projected clusters in high dimensional spaces*, In SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pages 70–81, New York, NY, USA, 2000, ACM.
3. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan: *Automatic subspace clustering of high dimensional data for data mining applications*, SIGMOD Rec., 27(2):94–105, 1998.
4. A. Asuncion, and D.J. Newman: *UCI machine learning repository*, 2007.
5. Ch.-H. Cheng, A.W. Fu, and Y. Zhang: *Entropy-based subspace clustering for mining numerical data*, In KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 84–93, New York, NY, USA, 1999. ACM.
6. E. Czogała, and J. Łęski: *Fuzzy and Neuro-Fuzzy Intelligent Systems*, Series in Fuzziness and Soft Computing. Physica-Verlag, A Springer-Verlag Company, Heidelberg, New York, 2000.
7. J.C. Dunn: *A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters*, Journal Cybernetics, 3(3):32–57, 1973.
8. J.H. Friedman, and J.J. Meulman: *Clustering objects on subsets of attributes*, J. R. Statist. Soc. B, 66:815–849, 2004.
9. G. Gan, and J. Wu: *A convergence theorem for the fuzzy subspace clustering (FSC) algorithm*, Pattern Recogn., 41(6):1939–1947, 2008.

10. G. Gan, J. Wu, and Z. Yang: *A fuzzy subspace algorithm for clustering high dimensional data*. In Advanced Data Mining and Applications, Second International Conference, ADMA 2006, Xi'an, China, August 14-16, 2006, Proceedings, volume 4093 of Lecture Notes in Computer Science, pages 271–278. Springer Berlin/Heidelberg, 2006.
11. S. Goil, S. Goil, H. Nagesh, H. Nagesh, A. Choudhary, and A. Choudhary: *Mafia: Efficient and scalable subspace clustering for very large data sets*. Technical report, 1999.
12. L. Parsons, E. Haque, and H. Liu: *Subspace clustering for high dimensional data: a review*, SIGKDD Explor. Newsl., 6(1):90–105, 2004.
13. M. Sikora, and D. Krzykawski: *Application of data exploration methods in analysis of carbon dioxide emission in hard-coal mines dewater pump stations*, Mechanization and Automation of Mining, 413(6), 2005.
14. H. Späth: *Mathematical algorithms for linear regression*, Academic Press Professional, Inc., San Diego, CA, USA, 1992.
15. J. Yang, W. Wang, H. Wang, and P. Yu: *δ-clusters: capturing subspace correlation in a large data set*, In Data Engineering, 2002. Proceedings. 18th International Conference on, pages 517–528, 2002.

Grupowanie danych w rozmytych podprzestrzeniach

Streszczenie

Niektóre dane zawierają grupy danych nie we wszystkich wymiarach, ale w pewnych podprzestrzeniach dziedziny. Artykuł przedstawia algorytm grupowania danych w rozmytych podprzestrzeniach. Każdy przykład danych ma pewną rozmytą przynależność do grupy (klastra). Każdy klaster z kolei jest rozpięty w pewnej podprzestrzeni dziedziny wejściowej. Klastry mogą być rozpięte w różnych podprzestrzeniach. Algorytm grupowania oparty jest na minimalizacji funkcji kryterialnej. W wyniku jego działania wypracowane są położenia klastrów, ich rozmycie i wagi ich deskryptorów. Przestawiono także wyniki eksperymentów grupowania danych syntetycznych i rzeczywistych.