Classes and clusters in data analysis

A.M. Rubinov, N.V. Sukhorukova and J. Ugon *

CIAO and School of Information Technology and Mathematical Sciences
University of Ballarat
P.O. Box 663
Ballarat, Victoria 3353 Australia

Abstract

We discuss the relation between classes and clusters in datasets with given classes. We examine the distribution of classes within obtained clusters, using different clustering methods which are based on different techniques. We also study the structure of the obtained clusters. One of the main conclusions, obtained in this research is that the notion purity cannot be always used for evaluation of accuracy of clustering techniques.

Key words: data mining, dataset with classes, point-based clustering, optimization clustering model, comparison of classes and clusters.

1 Introduction

The goal of this paper is to study relations between classes and clusters in real-world datasets. This research topic is important in the area of data classification. There are two types of data classification: supervised and unsupervised.

In supervised classification the algorithms are provided with the data points (observations) and the labels to indicate the class each observation belongs to. The task of supervised classification is to develop decision rules for prediction of the class for any new observation from its attribute values. In other words the algorithms should predict the label for each new observation.

In unsupervised classification (clustering) the algorithms work just with data points, there is no division into classes. The task of clustering algo-

* Corresponding author: jugon@ballarat.edu.au

rithms is to divide data items into several homogeneous groups (clusters): similar observations are grouped in the same cluster. The division is based on the attribute values of the data points. Unsupervised classification allows us to divide the dataset into several groups without knowing how the records "should" relate to each other. Sometimes the ways of grouping the records are very surprising and such study can bring many fresh ideas about the problems appeared in the area where data have been collected, and how to solve these problems.

The notion of a cluster is rather informal and we can give a formal definition of a cluster only in the framework of a formal mathematical model. Different formalizations are possible. In this paper we consider one of the most popular models, which we call the *point-based clustering model*.

An important characteristic of supervised and unsupervised classification methods is their *accuracy*. There are different approaches to *supervised classification accuracy* that can be used for comparing different classifiers (see for example, [10]). We mention here only n-fold cross-validation. This approach is very popular, however it cannot always give a good comparison of classifiers (see [3] for discussion). Indeed, the classification accuracy obtained for the same dataset in the case of n_1-fold cross-validation and n_2-fold cross-validation are not necessarily the same if $n_1 \neq n_2$. The estimation of accuracy of clustering methods is much more difficult than supervised classification methods.

One of the approaches for evaluation of accuracy of clustering methods is based on the following idea: let us consider the dataset with given classes and apply a technique for clustering this dataset assuming that classes are unknown. Then the clusters obtained by this technique can be compared with classes and the accuracy of the technique can be estimated by the degree of coincidences of classes and clusters. This approach has been discussed in [7] where the notion of cluster "purity" was introduced (see also [6]). In this paper we show that comparison of classes and clusters is not always appropriate, hence this comparison cannot be always used for assessment of a clustering technique. Indeed, it is possible that the points have been grouped by this technique according to some other characteristic rather than the classes. A simple and interesting example of such a case can be found in [8]).

We now present another example. Assume that we have records that describe one of the two letters, either A or B typed in different fonts made of several typefaces (Arial, Times New Romans, Tahoma etc), each written in two different styles: bold and italics. Different decomposition of this dataset onto two classes can be considered. For example, we can consider classes that contains fonts of A and fonts of B, respectively. We can also consider classes that contains both letters printed in bold and both letters printed in italics,

respectively. Different classes also can be considered. For example, one of them contains both letters typed in one collection of typeface and the other contains both letters typed in collection of different typefaces. Applying a clustering technique to this dataset we can divide it for two clusters based on a certain characteristic. For the sake of definiteness assume that one of these clusters contains letters printed in bold and the other contains letters printed in italics. On the other hand we can consider this dataset as a dataset with two classes, one of them consists of prints of the letter A and the other consists of prints of the letter B. In such a case the clusters and classes are pretty different, however we could not say the clustering technique is not good.

In such a situation we can suggest that some characteristics link points more strongly than their belonging to classes.

In this paper we show that a similar situation can appear in real-world datasets. We consider two real-world datasets with classes ("Pendigits" and "Letters") and compare classes and clusters for these datasets. We also give a very short description of results obtained in [13] for the Australian credit dataset that indicate another reason for non-coincidence between clusters and classes.

As was mentioned, we examine clusters in the framework of the point-based clustering model. In this model we can use different optimization techniques for the search for clusters. We also apply the notions of cluster function and the notion of structure of clusters (see [6]) in order to check the quality of clusters obtained by this technique. The goal of this investigation is two-fold. First, we show that the relation between clusters and classes are very different in these datasets. Then we also compare different optimization techniques for real world datasets.

The paper is organized as follows. In section 2 we present a point-based clustering model and give a short description of non-smooth optimisation methods used in the implementation of this model. In section 3 we discuss the relations between clusters and classes for the Pendigits dataset where clusters are close to classes. In section 4 we discuss the relations between clusters and classes for the Letters dataset where classes and clusters are different. Section 5 summarizes the obtained results and draws some guidelines for future research.

2 Preliminaries

2.1 Clusters

The following definition can be found in [8]: *Cluster Analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity. Intuitively, patterns within a valid cluster are more similar to each other that they are to a pattern belonging to a different cluster.*

Generally speaking, it is impossible to say whether given groups of points form clusters based on this definition (indeed, we need to explain what "more similar to each other" and "intuitively" mean). In order to find clusters we should formalize this definition. It can be done in many ways. We use one of the most popular models that we call a *point-based clustering model*. In the framework of this model clusters represent groups of points, centred in their centres, which are also points.

Suppose that a dataset contains N points $a_i \in \mathbb{R}^n, i = 1, \ldots, N$ and suppose that we need to find k clusters in this dataset. In order to determine the clusters we have to find k points which are the centres of these clusters and then we assign each point to the cluster with the nearest centre. The search for the centres can be reduced to the minimization of the so-called *cluster function*:

$$f(x_1, \ldots, x_k) = \sum_{i=1}^{N} \min_{j=1,\ldots,k} \|a_i - x_j\|_p, \quad p \geq 1. \tag{1}$$

Thus we need to solve the following mathematical programming problem

$$\text{minimize} \quad \sum_{i=1}^{N} \min_{j=1,\ldots,k} \|a_i - x_j\|_p, \quad p \geq 1, \tag{2}$$

$$\text{subject to } (x_1, \ldots, x_k) \in \mathbb{R}^{nk} \tag{3}$$

The dimension of this problem is nk (the number features multiplied by the number of clusters). If $k > 1$, the cluster function is nonsmooth and nonconvex. This function has many local minima, especially if the number of records in the dataset is large.

Remark 1 *There are some techniques which allow one to reduce the number of components in the sum (see [6]).*

After finding centres, we construct clusters: each point is assigned to the cluster with the nearest centre.

The detailed explanation of the approach to clustering based on cluster function can be found in [6]. In the following we assume that $p = 2$.

Remark 2 *Models that are different from the point-based one also can be used. For example, clusters that are centred by some hyperplanes ("hyperplane-based clusters" or "skeletons") can be considered (see [14] and references therein).*

Often we have different candidate vectors $X = (x_1, \ldots, x_k)$ which contend to be considered as centres of clusters. In the framework of the point-based clustering model, the cluster function can be used as a tool for comparison of candidates. Namely, the vector $X^* = (x_1^*, \ldots, x_k^*)$ is better suited for the role of centres of clusters then $X = (x_1, \ldots, x_k)$ if $f(x_1^*, \ldots, x_k^*) < f(x_1, \ldots, x_k)$. See [6] for details.

2.2 Step by step clustering and simultaneous clustering

Finding an initial point for minimizing the cluster function is a problem whose difficulty grows with the number of clusters sought. For one cluster, the optimization problem is convex, however for $k \geq 2$, it is not anymore convex. A heuristic, step by step method was proposed in [4] for minimizing the cluster function.

This algorithm, which constructs the clusters one by one and refines the solutions stepwise, allows one to avoid the difficult task of finding a suitable initial point for the local method. This algorithm has been introduced and discussed in [4], where numerical results were also presented.

We can also try to find simultaneously all clusters if we have some initial guess. For example, comparing clusters and classes we can use the collection of centres of classes as an initial point for the optimization.

The k-means method is one of the most popular clustering methods (see, for example, [9], [11]). This clustering method is very fast, however, very often it produces clustering systems with higher dissimilarity than methods based on nonsmooth optimization technique that we use (see subsection 2.3). It is known that the results obtained by k-means highly depends on the initial points. It is efficient to use the k-means method to find a good starting point for nonsmooth optimisation (see [6] for further discussions).

2.3 Optimisation techniques

In our numerical experiments we use the discrete gradient method (briefly DG) for minimizing (2). DG has been developed by Adil Bagirov (see [1], [2] for details).

We also use a combination of DG with the cutting angle method (briefly CAM, see [5], [12] for details). We call this combination DG+CAM. The results obtained by DG and especially by DG + CAM do not depend so strongly on the initial point as those obtained by k-means.

DG and DG+CAM are the main optimisation techniques, provided by CIAO-GO (optimisation software developed at the University of Ballarat). This software has been developed for constrained and unconstrained optimisation. It allows researchers to choose different optimisation methods (DG, DG+CAM), different types of penalty functions (for constrained optimisation) and different types of initial points (initial points can be appointer by researchers or generated randomly in a feasible region). For more details about CIAO-GO software, refer to [15]

2.4 Structure of clusters

As was mentioned in subsection 2.1, the cluster function can be used for evaluation of the quality of centers of clusters and through this for evaluation of quality of clusters. Another way to evaluate the quality of unsupervised classification is to check the distribution of the points within the clusters (structure of clusters). The goal is to check how "deep" the points are in the clusters.

Suppose that we work with a dataset which contains N observations. A clustering method has been applied to this dataset and k centres of clusters have been obtained. Consider a point a from the dataset which belongs to the l-th cluster (with the centre x^l). For this point we determine a value $c(a)$ which can be found as follows:

$$c(a) = \frac{||a - x^l||}{\min\limits_{j=1,\ldots,k, j \neq l} ||a - x^j||}. \tag{4}$$

From the definition of clusters it is easy to conclude that $c(a) \in [0, 1]$. Very often in our study we use a grid for $c(a)$ such as $0.1, 0.2, \ldots, 1$. We do not

work with the exact value for each $c(a)$, but use intervals. For example,

$$c(a) \in [(i-1) * 10^{-1}, i * 10^{-1}), \quad i = 1, \ldots, 9 \quad \text{and} \quad c(a) \in [0.9, 1] \tag{5}$$

or

$$c(a) \in [0, i * 10^{-1}) \quad i = 1, \ldots, 9 \quad \text{and} \quad c(a) \in [0, 1]. \tag{6}$$

We use the value $c(a)$ for each point to describe how "deep" this point is inside the cluster. We should underline that different values of $c(a)$ do not represent the radiuses for some spheres centred at the centres of the corresponding clusters. They rather represent some levels of confidence that the chosen point belongs to this cluster but not another one. It is possible that some points which are not "deep" enough inside the corresponding cluster move to another cluster (change their membership). It could happen, for example,

- if we change the norm in the definition of the cluster function;
- if we change the location of points (another accuracy to represent numbers in the computer);
- if we change the value for some internal parameters for the optimization methods etc.

If $c(a) = 1$ or close to 1 there are two centres such that the distances between the point and these two centres are (almost) the same. In this case some changes within the data may change the membership of the point (the point is "unstable" inside the cluster). If $c(a)$ is close to 0 the level of confidence for the point to keep its membership is high (the point is "stable").

Suppose that we obtain two different clustering results. First, we can compare the value of cluster function in the centres of corresponding clusters. The systems of clusters with the lowest value of cluster function is better from the point of view of cluster function. Second, we check the structure of corresponding clusters by means of (4). If for the first of the clustering result the values $c(a)$ are smaller than for the second clustering result for most of the points a, we assume that the first collection of the centres is preferable to the second one in the sense of the structure of the clusters (by means of (4)). The most important is to investigate the points a with the values $c(a) \in [0.9, 1]$. It is possible that the two approaches: evaluation of quality of clusters by means of clusters function and the structure of clusters, respectively, contradict each other (multi criteria problem).

2.5 Classes and clusters: purity

The notion of *purity* (see for example [7]) is used in the literature for evaluation of accuracy of clustering methods. Assume that we have a dataset A composed of the classes $\{D_1, \ldots, D_l\}$ and we apply a clustering procedure for finding clusters $\{C_1, \ldots, C_k\}$ in this dataset.

The purity of a set of clusters $\{C_1, \ldots, C_k\}$ is calculated as follows:

$$p(\{C_1, \ldots, C_k\}) = \frac{\sum_{i=1}^{k} \max_{j=1,\ldots,l} |C_i \bigcap D_j|}{|A|},$$

where $|B|$ is the cardinality of a finite set B.

We illustrate this notion in the simplest case where a dataset A contains only 2 classes $\{D_1, D_2\}$. Suppose that 4 clusters $\{C_1, C_2, C_3, C_4\}$ have been found in this dataset.

If the majority of points from the $j-$th cluster $(j = 1, \ldots, k)$ is in the $i-$th class $(i = 1, 2)$, we assign the whole $j-$th cluster to the $i-$th class. When all the clusters are assigned to one of the classes, the percentage of the correctly classified points for the test set is considered as the classification accuracy.

$$p(\{C_1, \ldots, C_4\}) = \frac{\sum_{i=1}^{4} \max(|C_i \bigcap D_1|, |C_i \bigcap D_2|)}{|A|}$$

2.6 Classes and clusters: preliminary example

There are many methods for supervised classification (see [10] and references within) which are based on quite different techniques. It is suggested, that if there is a group of points in a dataset which are misclassified by several methods then this group of points should be studied separately.

In this subsection we present some of the results obtained in [13] for comparing the division into classes and into clusters and for discovering possible reasons why sometimes these divisions are not the same.

The following procedure has been suggested in [13] for examination of quality of a dataset. First, different classifiers that are based on completely different ideas should be chosen. Then these methods should be applied to the dataset and all misclassified points for each method should be identified. The points that are misclassified by all methods are called *questionable*. Since

the used methods are based on completely different approaches we can accept that misclassification of a questionable point does not depend on classifier and depends on the quality of the dataset.

It was indicated in [13] that the Australian credit dataset consists of 5 well-defined clusters. One of the clusters contains 56 points from the first class and 77 points from the second class. The majority of questionable points belongs to this cluster and they are "deep" inside the cluster.

The fact that questionable points are mainly placed in the same cluster is of great interest. One of the possible interpretation of this fact is as the following: questionable points appear as the result of a certain systematic error and they have no real links with both classes.

3 Pendigits

This dataset was introduced by E. Alpaydin and Fevzi Alimoglu (see [10]). It contains 10 classes, 10992 observations, 16 attributes. All input attributes are integers $0 \ldots 9$.

The dataset has been created by collecting 250 samples from 44 writers. The samples written by 30 writers are used for the training set and the digits written by the other 14 are used for writer independent testing.

3.1 Classes and centres

The experiment are first carried out on the Pendigits dataset. Although this dataset is quite large, it is usually well handled by most classification methods. The first experiments is to carry out a step by step clustering without any knowledge of the classes and compare these clusters with the classes. Table 1 shows the amount of points from each class in each cluster.

As a result it is noticeable that a great majority of points in each cluster belongs to the same class. This means that the clusters seem to coincide with classes. The purity is 78.58%.

The second experiment carried out is to find for each class one centre. The points obtained are then considered as cluster centres and the same analysis as previously is applied. Table 2 presents the results.

A similar - but more previsible - observation is made about this table: the correspondence between clusters and classes is very strong. The purity is

	1	2	3	4	5	6	7	8	9	10	size	% class
1	470	0	0	0	0	107	2	1	1	0	581	80.89
2	1	1122	334	5	1	1	0	0	6	1	1471	76.27
3	0	15	634	26	0	13	0	123	146	31	988	64.17
4	0	0	2	1051	5	28	0	79	0	1	1166	90.13
5	11	0	1	29	1049	3	1	0	1	0	1095	95.79
6	247	0	0	0	0	989	0	12	0	0	1248	79.24
7	183	0	0	0	1	0	625	0	3	0	812	76.97
8	8	0	82	32	0	2	191	715	0	2	1032	69.28
9	69	2	1	0	0	0	0	0	962	0	1034	93.03
10	66	5	89	1	0	0	236	125	23	1020	1565	65.17
size	1055	1144	1143	1144	1056	1143	1055	1055	1142	1055		
	44.54	98.07	55.46	91.87	99.33	86.52	59.24	67.77	84.23	96.68		

Table 1
Pendigits: repartition of the classes in the clusters obtained by step by step clustering

	1	2	3	4	5	6	7	8	9	10	size	% class
1	561	0	0	0	0	142	14	1	0	0	718	78.13
2	0	1101	325	9	2	1	0	0	7	1	1446	76.14
3	0	37	651	18	0	12	1	126	148	31	1024	63.57
4	0	0	3	1060	9	79	0	34	0	1	1186	89.37
5	42	0	0	11	1043	25	3	0	0	0	1124	92.79
6	157	0	0	0	0	876	0	8	0	0	1041	84.14
7	67	0	0	0	1	0	613	0	0	0	681	90.01
8	6	0	75	45	0	8	180	770	0	2	1086	70.9
9	154	1	0	0	0	0	0	0	968	0	1123	86.19
10	68	5	89	1	1	0	244	116	19	1020	1563	65.25
size	1055	1144	1143	1144	1056	1143	1055	1055	1142	1055		
	53.17	96.24	56.95	92.65	98.76	76.64	58.1	72.98	84.76	96.68		

Table 2
Pendigitits: repartition of the classes in the clusters defined by class centres

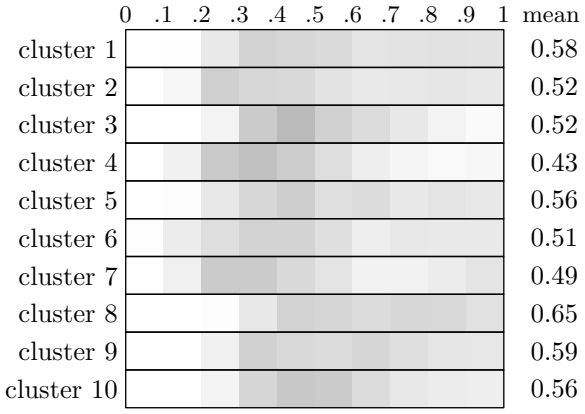

Fig. 1. Structures of the clusters of Pendigits

78.81%.

The class centres are used as an initial point to the local optimisation method to find a local minimum to the cluster function. In the case of Pendigits the solution obtained is the same as the one reached by the step by step method.

Because the same result was reached by two different methods, it can be expected that it is a very good local minimum. For comparison, the objective function value for a solution reached by the k-means method with a randomly chosen initial point was more than 2% larger.

The conclusion of this is that even by applying a clustering method without any knowledge of classes, the solution obtained is the one which corresponds at best to the classes. The classification accuracy is relatively high. In the Pendigits dataset the classes and cluster seem to be perfectly equivalent.

Indeed when the number of clusters is increased to 20 (twice the number of classes) for the step by step clustering, the purity becomes 87%. This result, obtained by a method which does not use any knowledge of classes, is comparable to most specifically designed methods presented in [10].

3.2 Classes and cluster structures

Figure 1 represents the structure of the clusters, The darker the square the larger amount of points are present in the layer. All the clusters present a similar structure: the points are deep, showing a good clusterization.

The Pendigits dataset has been shown to present a strong correlation between clusters and classes. It may be interesting to consider a deeper relation, by finding the layers of the clusters.

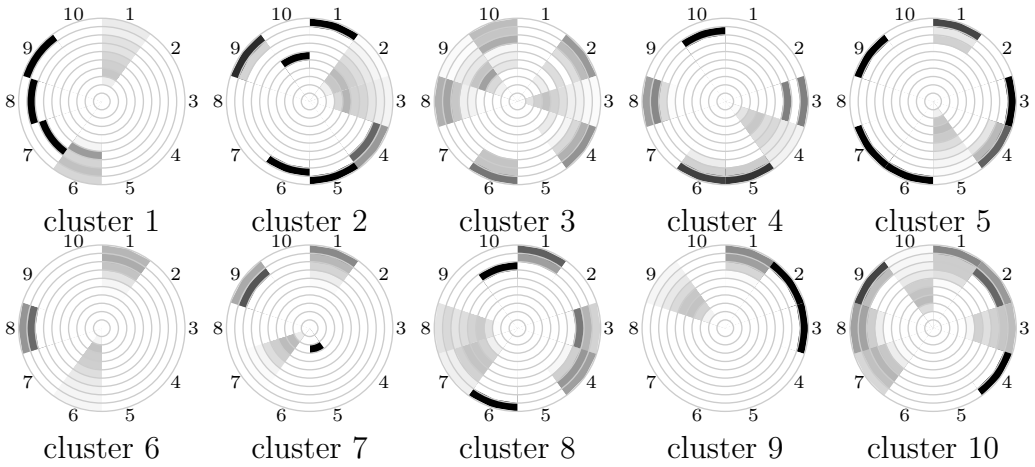

Fig. 2. Structures of the clusters in Pendigits

Figure 2 shows the depth of the classes in each cluster. Each disk represents one cluster, and each slice of this disk represents one class. The darker one circle is, the larger the proportion of points from this class is deeper inside the cluster.

Table 3 shows the repartition of the main class and the other points in each cluster. All the clusters present the same characteristic: although the points of each clusters are quite deep for all the clusters the main class is deeper than the others. This means that not only the clusters centres represent the classes well, but also the "misclassified points" do not belong so strongly to the cluster.

Table 4 shows the average depth of each class inside each cluster. The results presented in this table confirm that in each cluster the majoritary class is the deepest.

3.3 Summary

- The relationships between classes and clusters are very explicit
- The points are "deep" inside the clusters, and therefore the clustering is considered of high quality
- The correctly classified points are deeper.

4 Letters

The dataset was introduced by David Stale (see [10]). It is based on various fonts representation of the letters of the Latin alphabet. The dataset consists of 20000 observations, 26 classes, 16 numerical attributes. This is

Cluster		[0,0.1]	[0.1,0.2]	[0.2,0.3]	[0.3,0.4]	[0.4,0.5]	[0.5,0.6]	[0.6,0.7]	[0.7,0.8]	[0.8,0.9]	[0.9,1]
1	Main class	0	0	6.8	22.12	18.72	16.17	11.06	11.27	6.17	0
	Other classes	0	0	0	0	0	3.6	37.83	18.01	24.32	0
2	Main class	0	0.08	5.97	22.01	26.73	18.36	13.72	8.55	3.83	0
	Other classes	0	0	1.43	14.61	27.79	16.61	15.47	9.74	7.73	0
3	Main class	0.31	11.35	18.61	23.34	17.5	12.77	6.46	5.36	2.52	0.31
	Other classes	0	0	2.25	6.49	16.94	12.42	6.49	15.53	19.2	0
4	Main class	0	1.23	10.56	17.6	21.4	13.32	14.93	8.65	7.61	0
	Other classes	0	0	0	0	0	0	0.86	12.17	38.26	0
5	Main class	0	5.71	22.11	25.64	20.59	13.34	7.14	3.62	1.42	0
	Other classes	0	0	0	0	0	0	0	10.86	19.56	0
6	Main class	0	4.24	23.55	19.61	18.7	12.84	8.08	5.66	3.74	0
	Other classes	0	0	0	0	0.77	3.47	10.42	22.39	33.59	0
7	Main class	0	7.68	26.08	26.24	19.36	13.44	4.32	1.76	0.96	0
	Other classes	0	0	0.53	0	0	1.6	6.95	16.57	28.34	0
8	Main class	0	0	0.97	11.74	20.27	17.34	13.28	10.48	14.82	0
	Other classes	0	0	0.31	3.47	10.72	13.56	14.51	26.81	16.71	0
9	Main class	0	0.1	4.67	17.15	23.38	22.03	14.55	9.04	5.5	0
	Other classes	0	0	0	0	0	0	1.38	15.27	36.11	0
10	Main class	0	0.49	13.23	26.17	23.62	18.82	8.43	4.7	3.13	0
	Other classes	0	0	0	0.18	1.28	6.42	14.67	22.2	26.42	0

Table 3
Pendigits: repartition of the classes by cluster layers

samples of 26 capital letters, printed in different fonts. 20 fonts have been considered and the location of the corresponding samples has been distributed randomly within the dataset.

The experiments described and applied on the Pendigits dataset in the previous section are here applied on the Letters dataset. The results are presented and discussed.

	1	2	3	4	5	6	7	8	9	10
1	0.55	-	-	-	-	0.76	0.77	0.87	0.91	-
2	0.99	0.51	0.56	0.91	0.98	0.88	-	-	0.96	0.56
3	-	0.81	0.41	0.84	-	0.88	-	0.83	0.5	0.78
4	-	-	0.85	0.53	0.93	0.93	-	0.87	-	0.85
5	0.92	-	0.91	0.91	0.41	0.99	0.97	-	0.99	-
6	0.82	-	-	-	-	0.45	-	0.89	-	-
7	0.86	-	-	-	0.26	-	0.38	-	0.88	-
8	0.93	-	0.81	0.86	-	0.97	0.62	0.63	-	0.76
9	0.87	0.96	0.99	-	-	-	-	-	0.54	-
10	0.86	0.9	0.77	0.98	-	-	0.76	0.86	0.94	0.47

Table 4
Average depth for each class in each cluster

4.1 Classes and centres

The Letters dataset contains 26 classes. Therefore, 26 clusters are found using the step by step algorithm. The size of the intersection of each cluster with each class is then evaluated.

Table 5 presents the results.

Unlike for the Pendigits dataset, where each cluster is strongly associated with a class (and where each class is represented by one cluster), the repartition of the classes in the clusters of the Letters dataset is much more scattered. The purity is very low: 28.25%

Table 6 presents the repartition of the classes when the cluster centres considered are one centre per class. It is noticeable that here again the repartition of the points is very disseminate. The purity is 23.94%, the value of the cluster function is much higher than in the case of the step by step clustering. We suggest that the classes are not "point-based".

Finally, these class centres are used as an initial point for the local algorithm to minimize the cluster function. While in the case of the Pendigits dataset the result of this experiment was similar to the solution found by step by step clustering, this time the centres are quite far apart and another solution is reached. Once again, the repartition, shown in Table 7 is very diffuse.

A possible interpretation of these results could be that the clustering

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26		
1	242	0	140	0	0	0	0	0	0	0	0	0	187	0	0	0	1	0	0	0	0	0	104	34	0	0	708	34
2	123	0	90	0	0	0	7	0	0	0	0	35	107	0	0	0	81	0	0	0	0	0	97	41	95	0	676	18
3	16	84	32	69	27	40	62	2	67	34	36	71	15	53	32	25	39	44	15	0	59	10	21	40	47	9	949	8
4	0	276	0	0	17	0	0	0	0	2	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	3	301	91
5	115	50	125	63	61	11	103	0	111	0	28	74	79	16	43	73	111	1	74	0	0	15	0	5	83	57	1298	9
6	0	21	7	3	274	2	0	0	0	91	0	8	7	0	0	0	11	125	1	0	105	157	68	0	0	6	886	30
7	0	0	0	155	0	171	0	159	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	485	35
8	0	28	6	24	3	27	31	116	4	24	1	30	6	29	18	5	0	16	0	76	35	2	5	36	33	0	555	20
9	0	0	0	31	0	40	0	44	0	0	0	14	0	0	0	0	0	0	0	25	37	0	0	0	0	0	191	23
10	0	0	0	0	0	0	0	0	0	291	0	0	0	0	0	0	0	212	0	0	2	198	0	0	0	0	703	41
11	3	0	15	3	0	1	0	8	3	0	567	1	0	0	4	5	0	0	1	0	0	0	0	3	0	2	616	92
12	0	56	0	43	0	36	88	0	93	0	0	32	0	153	134	0	45	0	0	0	66	0	21	99	14	0	880	17
13	0	28	0	40	0	27	85	0	169	1	0	35	23	266	138	13	8	7	0	0	45	14	70	150	3	15	1137	23
14	23	66	52	69	52	27	80	13	65	45	44	76	66	11	120	34	55	76	16	18	20	50	34	66	111	26	1315	9
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	324	0	0	0	0	0	0	0	0	0	0	324	100
16	104	55	98	72	33	1	89	0	114	0	1	93	116	0	1	5	122	0	30	0	0	2	0	1	98	12	1047	11
17	0	7	0	0	4	0	1	0	4	0	59	0	1	0	0	60	0	0	350	0	0	0	0	0	4	190	680	51
18	0	5	0	0	5	0	0	0	0	0	0	0	0	0	0	158	1	0	229	0	0	0	0	0	0	190	588	38
19	0	0	0	2	8	127	0	239	18	26	13	1	0	6	0	1	0	1	0	327	26	0	8	1	0	1	805	40
20	0	10	0	34	10	251	11	203	9	27	21	20	1	7	2	1	0	1	1	304	41	3	17	8	0	2	984	30
21	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	169	0	0	0	0	0	174	97
22	15	26	41	40	79	7	0	2	0	84	1	92	44	0	0	2	69	119	2	0	126	127	87	3	0	12	978	12
23	0	68	0	3	163	5	0	2	0	137	0	9	3	0	0	0	4	146	0	2	28	174	40	0	0	0	784	22
24	0	7	0	50	0	9	148	0	123	0	1	119	16	210	187	9	128	2	4	0	54	5	123	249	23	2	1469	16
25	93	16	142	28	39	1	53	4	25	2	17	29	97	2	104	46	112	36	29	0	0	39	41	37	255	52	1299	19
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	168	168	100
size	734	803	748	734	775	783	758	792	805	764	789	739	768	753	783	761	787	786	755	752	813	796	736	773	766	747		
	32	34	18	21	35	32	19	30	20	38	71	16	24	35	23	42	16	26	46	43	20	24	16	32	33	25		

Table 5. Letters: repartition of the classes in the clusters obtained by step by step clustering

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26		
1	113	1	50	0	9	0	0	0	1	0	0	0	36	0	16	15	0	0	12	0	0	1	6	3	4	16	283	39
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	242	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	242	100
4	220	2	273	7	8	0	84	0	5	0	0	59	77	0	19	17	165	1	11	0	0	8	8	14	179	22	1179	23
5	17	112	13	0	320	1	0	0	0	77	0	0	18	0	0	0	16	127	14	0	0	194	0	0	0	16	925	34
6	5	0	7	1	0	0	8	0	0	0	0	0	0	0	0	7	1	0	0	0	0	0	0	0	0	1	30	26
7	33	5	54	27	0	0	78	0	38	0	0	5	36	0	4	2	34	0	5	0	0	0	0	0	54	0	375	20
8	3	39	5	11	10	17	28	0	123	1	16	2	1	9	11	57	30	5	23	0	11	0	0	7	13	19	441	27
9	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	100
10	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	50
11	0	5	3	0	0	6	0	0	0	28	0	0	0	0	0	0	1	2	0	0	0	1	0	0	0	0	46	60
12	0	3	1	6	19	0	0	0	0	346	0	44	5	0	0	0	51	177	6	0	22	86	2	0	0	0	768	45
13	15	0	5	4	0	0	0	0	0	44	0	3	37	0	0	0	2	28	35	0	0	0	0	0	5	0	178	24
14	15	0	31	6	0	1	18	0	42	1	455	0	0	2	3	92	131	0	37	0	2	0	0	7	1	64	908	50
15	0	0	0	0	0	4	0	0	0	0	0	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	92
16	0	0	0	0	0	0	0	0	0	5	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	77
17	0	21	0	26	0	28	40	15	2	1	0	1	0	44	43	0	0	0	0	0	23	0	1	30	2	4	281	15
18	42	76	78	299	20	216	219	188	312	28	28	128	134	457	473	53	27	83	3	32	399	82	215	372	247	54	4265	11
19	13	72	10	39	56	25	46	0	111	0	136	3	1	96	41	262	43	0	138	3	0	0	0	8	25	77	1205	21
20	1	0	0	4	0	1	0	0	0	0	0	50	10	0	0	16	54	0	15	0	0	0	0	0	1	31	183	29
21	140	165	194	157	116	7	96	1	139	40	39	142	205	76	115	74	207	118	54	0	13	54	63	110	198	89	2612	7
22	0	44	0	120	41	471	135	586	32	110	115	181	0	48	11	18	9	6	3	717	126	7	42	41	37	9	2909	24
23	0	0	2	24	5	6	0	2	0	4	0	19	8	2	0	0	5	1	0	0	190	0	46	0	0	44	358	53
24	9	15	10	0	171	0	0	0	0	15	0	0	5	0	0	147	6	201	347	0	2	358	66	0	0	225	1577	22
25	108	1	12	0	0	0	5	0	0	0	0	0	195	0	0	1	0	0	4	0	0	0	282	83	0	76	767	36
26	0	0	0	3	0	0	1	0	0	60	0	32	0	19	47	0	5	37	48	0	25	5	5	98	0	0	385	25
size	734	803	748	734	775	783	758	792	805	764	789	739	768	753	783	761	787	786	755	752	813	796	736	773	766	747		
	29	30	36	40	41	60	28	73	38	45	57	24	26	60	60	34	26	25	45	95	49	44	38	48	32	30		

Table 6. Letters: repartition of the classes in the clusters defined by class centres

algorithm does not reach meaningful clusters. The number of clusters is fairly large, and the cluster function possess a very large number of local minima. Its minimization is a very difficult task, even for a powerful method like the step by step method.

However neither the improved class centres nor the class centres themselves seem to represent the classes well. Moreover the value of the function at the point reached by the step by step method is sensibly lower than the one at the other points.

The distribution of the points of the dataset in clusters independent from their class may be due to the fact that some more information is contained in the dataset (for example the font of the letters), and this information may bring some noise for an eventual classification.

The latter interpretation is confirmed by the fact that when the number of clusters is decreased to 4, thus making the problem easier to solve, the repartition of the classes remains scattered.

4.2 Classes and cluster structures

Figure 3 shows the structure of the clusters. These cluster present generally a similar structure: for most of the clusters the majority of points is deep inside the cluster. This shows that the clustering created clusters of good quality.

Experiments show that the distribution of the classes inside the clusters is very diffuse for the letters dataset. This means that the classes do not represent the most obvious clusterization of this dataset. Let us consider the structures of the clusters and examine the distributions of the classes among the layers.

Figure 4 and table 8 show the distribution of the classes in each cluster. Clearly there are several types of clusters. Some of them contain only points from a few classes. The majority of the clusters, however, contains points belonging to many classes. In general, it can be noticed that the repartition in layers for all the classes present in the cluster is quite similar. This result emphasizes the conclusion that a cluster cannot be associated with a particular class in the Letters dataset.

In this dataset 20 different fonts have been considered and the location of the corresponding samples has been distributed randomly within the dataset. The hypothesis is that the meaning of the clusters is font recognition rather than letter recognition. However, it is impossible to check this hypothesis,

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26		
1	114	0	97	0	0	0	15	0	0	0	0	48	107	0	0	3	92	0	2	0	0	2	108	63	109	0	760	15
2	242	0	183	0	0	0	0	0	0	0	0	0	192	0	3	0	7	0	0	0	0	0	85	42	94	0	848	28
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	19	90	37	54	36	36	61	1	72	41	25	68	15	59	45	36	44	46	26	0	63	23	31	43	45	16	1032	8
7	32	41	24	110	35	10	71	1	84	0	21	89	57	11	20	8	51	9	12	0	9	25	1	5	78	13	817	13
8	0	223	1	1	344	0	0	0	0	26	0	2	1	0	0	0	2	89	2	0	58	131	71	0	0	2	953	36
9	1	51	26	39	21	41	65	101	23	42	11	60	26	32	49	23	13	51	1	79	49	12	18	67	74	2	977	10
10	0	0	0	160	0	178	0	181	0	0	0	0	0	0	0	0	0	0	0	0	169	0	0	0	0	0	688	26
11	79	29	76	65	18	3	73	0	100	0	1	72	100	0	0	1	80	0	9	0	4	1	1	2	84	6	804	12
12	0	0	0	0	0	1	0	0	0	385	0	0	0	0	0	0	0	307	0	0	6	300	0	0	0	0	999	38
13	3	0	25	3	0	13	0	12	7	0	364	2	0	2	4	13	0	0	7	0	0	0	0	5	0	15	475	76
14	0	0	0	0	0	0	1	0	0	0	299	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	301	99
15	0	53	0	70	22	72	69	20	79	0	0	37	0	133	125	0	40	0	45	15	107	0	22	56	10	0	975	13
16	0	90	0	30	0	25	99	0	175	0	0	45	19	283	130	15	13	6	0	0	38	16	77	151	23	14	1249	22
17	94	57	111	55	66	11	73	9	55	25	38	60	108	7	176	55	101	79	38	0	5	53	58	72	134	67	1607	10
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	326	0	0	0	0	0	0	0	0	0	0	326	100
19	146	54	135	32	49	0	75	0	70	0	0	64	103	11	36	101	154	3	88	0	0	19	4	3	62	63	1272	12
20	0	3	0	0	5	0	0	0	0	0	0	0	0	0	0	168	0	0	374	0	0	0	0	0	0	285	835	44
21	0	8	0	31	10	238	12	196	9	22	20	18	1	6	2	0	0	1	1	313	40	2	17	8	0	1	956	32
22	0	1	0	5	6	137	0	270	15	32	9	2	0	6	1	1	0	1	0	343	26	0	11	2	0	2	870	39
23	3	49	27	23	162	10	0	1	0	190	0	61	22	0	0	0	61	192	6	2	171	205	103	0	0	41	1329	15
24	1	54	6	56	1	8	144	0	116	1	1	111	17	203	192	10	129	2	3	0	68	7	129	254	53	4	1570	16
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	100
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	141	0	0	0	0	0	0	214	355	60
size	734	803	748	734	775	783	758	792	805	764	789	739	768	753	783	761	787	786	755	752	813	796	736	773	766	747		
	32	27	24	21	44	30	18	34	21	50	46	15	25	37	24	42	19	39	49	45	21	37	17	32	17	38		

Table 7. Letters: repartition of the classes in the clusters defined by improving the class centres

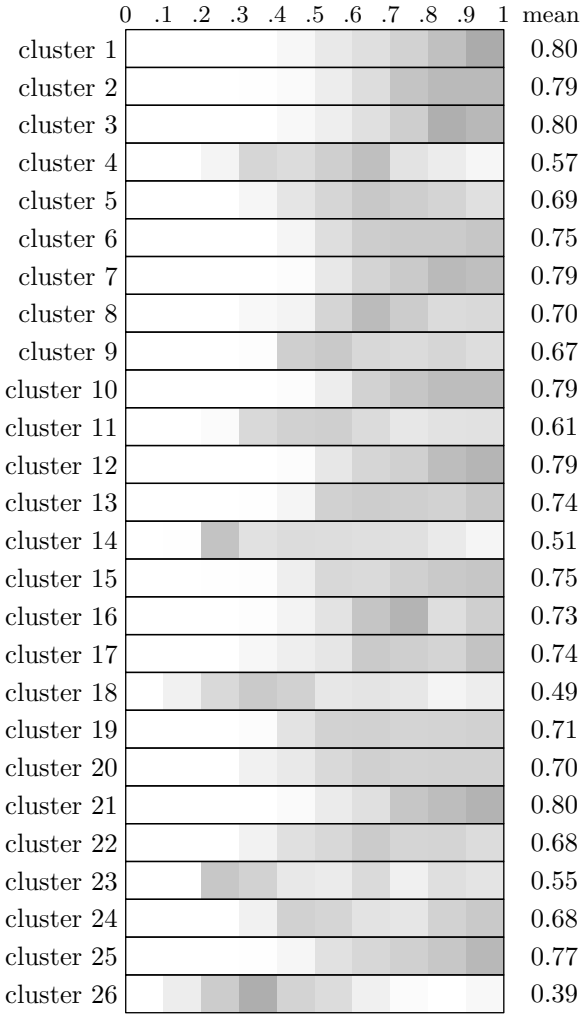

Fig. 3. Structures of the clusters for the Letters dataset

because the dataset does not keep the information about the corresponding font for each particular observation.

Cluster		[0,0.1]	[0.1,0.2]	[0.2,0.3]	[0.3,0.4]	[0.4,0.5]	[0.5,0.6]	[0.6,0.7]	[0.7,0.8]	[0.8,0.9]	[0.9,1]
1	Main class	0	0	0	2.06	25.61	29.33	13.63	9.91	5.78	0
	Other classes	0	0	0	0.21	15.02	16.95	16.3	16.52	21.67	0
2	Main class	0	0	0.81	16.26	24.39	6.5	28.45	10.56	4.06	0
	Other classes	0	0	0	0.36	2.71	10.3	19.16	20.79	19.71	0
3	Main class	0	0	0	0	1.19	15.47	32.14	20.23	17.85	0
	Other classes	0	0	0	0.11	1.15	8.67	14.68	18.26	26.47	0
4	Main class	0	0	0	6.15	19.56	18.11	11.95	10.5	17.02	0
	Other classes	0	0	0	0	0	0	0	4	24	0

5	Main class	0	0	0	0	6.4	21.6	13.6	16.8	18.4	0
	Other classes	0	0	0	0.17	2.89	7.33	12.36	17.81	25.31	0
6	Main class	0	0	0	0	2.55	6.2	14.23	27	25.91	0
	Other classes	0	0	0	0.16	1.63	8.49	11.27	20.09	25.81	0
7	Main class	0	0	11.69	45.61	26.9	12.28	2.33	0.58	0	0
	Other classes	0	0	0	0	6.36	22.61	37.26	16.87	11.78	0
8	Main class	0	0	0	0.86	5.17	19.82	25.86	18.1	16.37	0
	Other classes	0	0	0	1.13	12.3	16.85	16.17	16.62	17.53	0
9	Main class	0	0	25	40.9	13.63	4.54	9.09	6.81	0	0
	Other classes	0	0	21.08	10.88	7.48	8.84	16.32	5.44	16.32	0
10	Main class	0	0	0	1.37	7.9	23.02	27.49	13.74	14.43	0
	Other classes	0	0	0	4.85	12.13	11.4	17.71	23.05	18.68	0
11	Main class	0	0	1.76	16.22	19.57	20.28	14.99	8.81	10.58	0
	Other classes	0	0	0	0	0	0	6.12	14.28	22.44	0
12	Main class	0	0	0	21.56	26.79	26.14	9.8	3.26	4.57	0
	Other classes	0	0	0	2.33	4.67	12.37	20.08	20.22	20.63	0
13	Main class	0	0	0.37	1.5	9.02	22.55	24.06	14.28	12.4	0
	Other classes	0	0	0	0.11	1.95	8.49	13.31	19.51	25.02	0
14	Main class	0	0	0	0	0	0.83	16.66	24.16	26.66	0
	Other classes	0	0	0	0	1.08	7.78	17.65	22.25	25.77	0
15	Main class	0	0.3	23.45	11.72	13.88	13.27	12.65	12.34	8.33	0
	Other classes	0	0	0	0	0	0	0	0	0	0
16	Main class	0	0	0	0	4.91	18.85	17.21	15.57	20.49	0
	Other classes	0	0	0.32	0.75	6.81	14.91	14.27	19.02	21.4	0
17	Main class	0	0	0	0	0	18.85	38.28	25.42	10.28	0
	Other classes	0	0	0	5.75	10	14.24	14.24	14.24	18.18	0
18	Main class	0	0	0	12.66	19.65	16.15	25.32	11.35	6.98	0
	Other classes	0	0	0.27	0.27	7.52	14.76	17.27	19.49	22.84	0
19	Main class	0	0	0	0	4.58	22.32	26.6	20.48	15.59	0
	Other classes	0	0	0	0.2	3.34	6.69	15.06	20.08	23.84	0
20	Main class	0	0	0	0.32	2.96	20.39	22.03	20.72	16.77	0

	Other classes	0	0	0	0.14	4.41	17.05	18.67	17.79	17.94	0
21	Main class	0	7.1	20.71	32.54	17.75	14.2	5.91	1.77	0	0
	Other classes	0	0	0	0	0	0	0	0	0	0
22	Main class	0	0	0	0	0	2.36	17.32	22.04	31.49	0
	Other classes	0	0	0	0	1.52	9.87	16.68	20.79	26.43	0
23	Main class	0	0	0	1.72	12.06	13.21	22.41	25.28	12.64	0
	Other classes	0	0	0	0.49	2.45	10.49	22.62	30.49	12.95	0
24	Main class	0	0	0	0	3.61	10.04	21.28	21.68	23.69	0
	Other classes	0	0	0	0.24	1.55	6.31	11.88	23.52	27.95	0
25	Main class	0	0	0	0	0.39	3.92	7.45	15.68	43.92	0
	Other classes	0	0	0	0.19	3.16	7.27	13.4	20.3	28.06	0
26	Main class	0	5.35	14.88	20.83	17.85	9.52	10.71	9.52	4.16	0
	Other classes	0	0	0	0	0	0	0	0	0	0

Table 8: Letters: repartition of the classes by cluster layers

Table 9 presents the average depth of each class in each cluster. All the classes are at the same depth inside the clusters. Moreover this depth - generally between 0.5 and 0.9 - shows that the points are quite far from the boundary of each cluster. An interesting case is the 14-th cluster: most classes of the dataset are very deep inside the cluster. This can be seen as a strong belonging to the clusters, and it can be concluded that although the cluster is strongly constituted, it does not permit to discriminate between classes. A similar situation appeared for the Australian credit dataset (subsection 2.6).

4.3 Summary

- The relationships between classes and clusters are not very explicit
- The points are not as "deep" inside the clusters as they were for the Pendigits dataset. However, they are "deep" enough for most of the clusters and therefore clustering is considered of good quality.
- For most of the clusters the correctly classified points are "deeper" than misclassified points.
- The clusters do not discriminate the classes. A similar situation appeared for the Australian credit dataset. A possible hypothesis is that the clusters grouped the records according to their font recognition rather than letter recognition.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	0.63	-	0.68	-	-	-	-	-	-	-	-	-	0.6	-	-	-	0.97	-	-	-	-	-	0.86	0.84	-	-
2	0.59	-	0.74	-	-	-	0.91	-	-	-	-	0.87	0.65	-	-	-	0.73	-	-	-	-	-	0.87	0.82	0.83	-
3	0.87	0.73	0.8	0.77	0.82	0.9	0.72	0.91	0.75	0.89	0.84	0.78	0.85	0.71	0.87	0.81	0.83	0.85	0.9	-	0.85	0.9	0.88	0.76	0.73	0.92
4	-	0.67	-	-	0.92	-	-	-	-	0.89	-	-	-	-	-	-	-	-	0.93	-	-	-	-	-	-	0.96
5	0.83	0.82	0.74	0.79	0.86	0.95	0.74	-	0.77	-	0.88	0.8	0.72	0.89	0.88	0.88	0.8	0.98	0.86	-	-	0.89	-	0.89	0.71	0.9
6	-	0.88	0.86	0.89	0.79	0.93	-	-	-	0.87	-	0.91	0.92	-	-	-	0.9	0.72	0.95	-	0.85	0.77	0.85	-	-	0.94
7	-	-	-	0.74	-	0.41	-	0.61	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
8	-	0.83	0.93	0.76	0.88	0.67	0.7	0.71	0.79	0.68	0.94	0.72	0.79	0.62	0.75	0.92	-	0.79	-	0.68	0.6	0.73	0.83	0.71	0.77	-
9	-	-	-	0.45	-	0.29	-	0.43	-	-	-	0.91	-	-	-	-	-	-	-	0.89	0.71	-	-	-	-	-
10	-	-	-	-	-	-	-	-	-	0.68	-	-	-	-	-	-	-	0.61	-	-	0.91	0.8	-	-	-	-
11	0.92	-	0.85	0.92	-	0.93	-	0.88	0.84	-	0.59	0.96	-	-	0.83	0.95	-	-	0.94	-	-	-	-	0.85	-	0.97
12	-	0.84	-	0.6	-	0.81	0.71	-	0.7	-	-	0.89	-	0.55	0.62	-	0.74	-	-	-	0.9	-	0.86	0.75	0.89	-
13	-	0.9	-	0.77	-	0.7	0.85	-	0.77	0.99	-	0.88	0.92	0.69	0.78	0.83	0.94	0.92	-	-	0.79	0.85	0.83	0.79	0.91	0.74
14	0.88	0.81	0.83	0.83	0.74	0.81	0.77	0.76	0.82	0.8	0.84	0.81	0.69	0.8	0.83	0.75	0.78	0.81	0.83	0.91	0.86	0.76	0.8	0.78	0.77	0.78
15	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.52	-	-	-	-	-	-	-	-	-	-
16	0.83	0.87	0.72	0.68	0.86	0.95	0.77	-	0.74	-	0.95	0.78	0.65	-	0.97	0.93	0.75	-	0.88	-	-	0.96	-	0.93	0.65	0.91
17	-	0.95	-	-	0.9	-	0.92	-	0.95	-	0.9	-	0.9	-	-	0.7	-	-	0.7	-	-	-	-	-	0.75	0.65
18	-	0.87	-	-	0.93	-	-	-	-	-	-	-	-	-	-	0.78	0.95	-	0.6	-	-	-	-	-	-	0.69
19	-	-	-	0.98	0.86	0.79	-	0.78	0.8	0.9	0.8	0.99	-	0.93	-	0.78	-	0.91	-	0.7	0.77	-	0.89	0.97	-	0.89
20	-	0.93	-	0.82	0.83	0.73	0.84	0.69	0.83	0.88	0.77	0.89	0.95	0.81	0.95	0.97	-	0.97	0.98	0.73	0.82	0.85	0.74	0.78	-	0.91
21	-	-	-	0.98	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.38	-	-	-	-	-
22	0.9	0.9	0.84	0.77	0.82	0.95	-	0.91	-	0.85	0.99	0.74	0.76	-	-	0.97	0.69	0.8	0.99	-	0.73	0.82	0.8	0.96	-	0.87
23	-	0.84	-	0.92	0.74	0.88	-	0.97	-	0.71	-	0.82	0.89	-	-	-	0.83	0.69	-	0.97	0.9	0.7	0.83	-	-	-
24	-	0.93	-	0.73	-	0.78	0.83	-	0.85	-	1	0.84	0.91	0.73	0.77	0.73	0.81	0.95	0.92	-	0.89	0.71	0.86	0.76	0.91	0.8
25	0.81	0.77	0.78	0.82	0.72	0.91	0.86	0.92	0.83	0.93	0.79	0.84	0.8	0.9	0.88	0.68	0.82	0.84	0.66	-	-	0.83	0.77	0.77	0.83	0.74
26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.5

Table 9. average depth for each class in each cluster

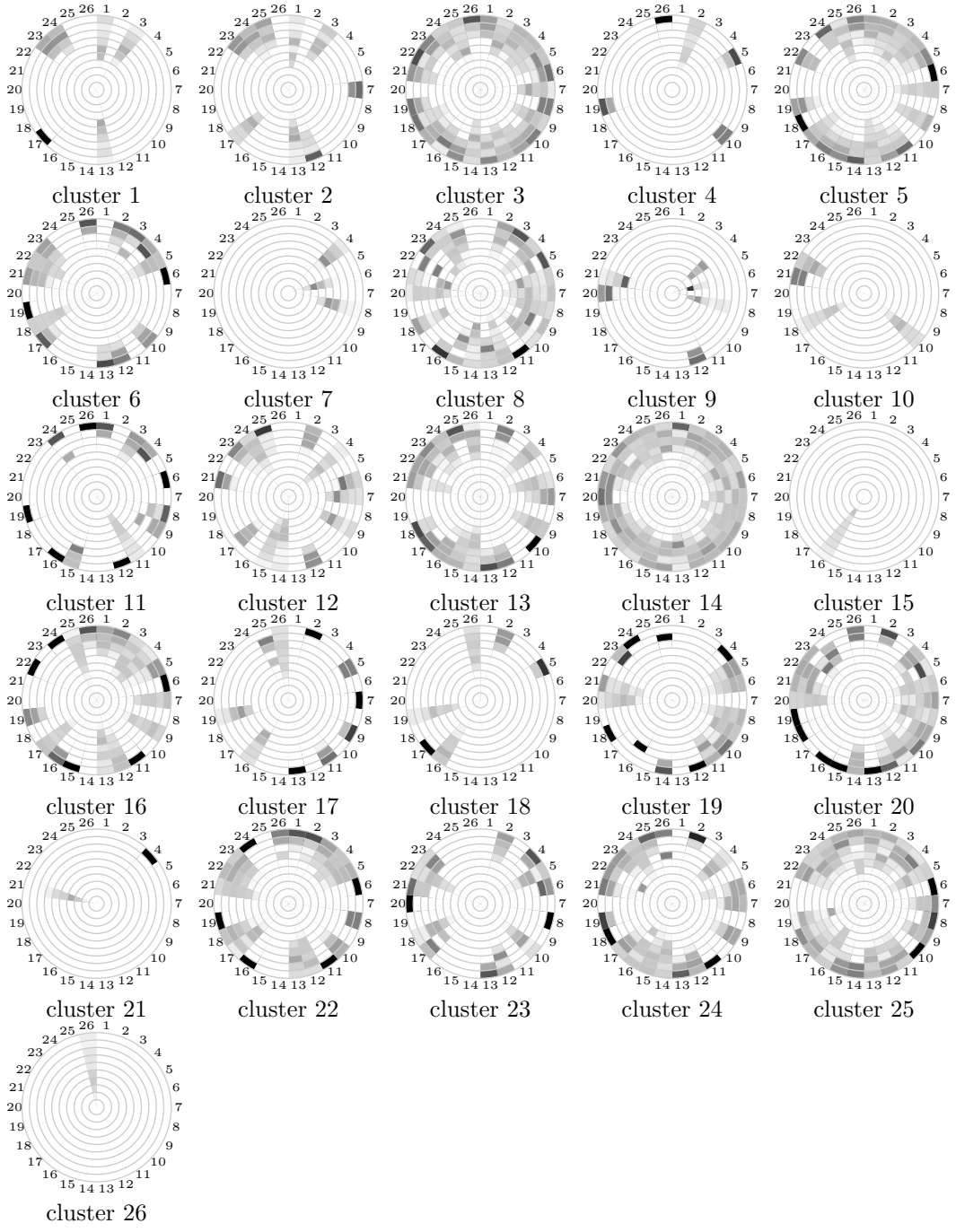

Fig. 4. The structures of the clusters

- The lowest dissimilarity function value has been reached by the step by step clustering method. The k-means method produced a higher value for the dissimilarity function.

5 Conclusions

1. The main conclusion to this research is that clusters do not necessarily coincide with classes. Several reasons can cause such results:

- the chosen clustering model does not match the classification structure. (For example, it is possible that some classes in Letters dataset are not point based; at the same time we use the point based model for clustering.)
- the dataset contains a high proportion of noisy records and/or possible mistakes, appeared, for example, on the stage of data collection;
- some characteristics link points more strongly than their belonging to classes.

2. The notion of purity cannot be always used for evaluation of accuracy of clustering methods. If the purity is high enough, we can conclude that the chosen clustering method is efficient for the dataset under consideration. However, we can not make any conclusion regarding the efficiency of clustering methods, if purity is low. In such a case the classes and clusters do not coincide. There can be different reasons for this: either the applied clustering technique works not very well, or there might be some other "hidden" characteristics, which link the records together and which are independent from the classes.

Some other approaches (different from comparing classes and clusters) can be used for evaluating the quality of clustering:

- the value of a cost function (called *dissimilarity function*, and in our case *cluster function*);
- the structure of the clusters (how "deep" the points are inside the clusters).

References

[1] A. M. Bagirov, *Derivative-free methods for unconstrained nonsmooth optimization and its numerical analysis*, Investigacao Operacional, Vol. 19, 1999, pp 75-93.

[2] A. M. Bagirov, *Numerical methods for minimizing quasidifferentiable functions: a survey and comparison*, In: V.F. Demyanov and A.M. Rubinov (eds.), *Quasidifferentiability and Related Topics*, Kluwer Academics Publisher, pp 33-71, 2000.

[3] A.M. Bagirov, A.M. Rubinov and J. Yearwood, A global optimization approach to classification, *Optimization and Engineering*, 3, 2002, 129-155.

[4] Adil M. Bagirov and Julien Ugon, *An algorithm for minimizing clustering functions*, Optimization, accepted for publication.

[5] A. M. Bagirov and A. M. Rubinov, *Global minimization of increasing positively homogeneous function over unit simplex. Annals of Operations Research*, Vol. 98, 2000, pp 171-187.

[6] A. Bagirov, A. Rubinov, N. Soukhoroukova, J. Yearwood, *Unsupervised and Supervised Data Classification Via Nonsmooth and Global Optimization*, Top, Vol. 11, Number 1, 1-93, Sociedad de Estadistica Operativa, Madrid, Spain, June 2003.

[7] I.S. Dhillon, J. Fan and Y. Guan, Efficient clustering of very large document collections, In: *Data Mining for Scientific and Engineering Applications*, Kluwer Academic Publishers, Dordrecht, 2001.

[8] A. K. Jain, M. N. Murty and p. J. Flynn *Data Clustering: A Review* ACM Computing Surveys, Vol. 31, No 3, September 1999.

[9] J. B. MacQueen, *Some Methods for Classification and Analysis of Multivariate observations*. In: L. M. LeCam and J. Neyman (eds.), Proceedings of the Firth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Berkeley: University of California Press, 1967.

[10] D. Michie, D. J. Spiegehalter and C. C. Taylor (eds.), *Machine learning, neural and statistical classification*. Ellis horwood series in artificial intelligence, London, 1994.

[11] B. Mirkin, *Mathematical Classification and Clustering*, Kluwer Academic Publishers, 1996.

[12] A. Rubinov, *Abstract Convexity and Global Optimization*, Kluwer Academic Publishers, 2000, ISBN 0-7923-6323-X.

[13] A. Rubinov, N. Soukhoroukova, J. Yearwood, *Clustering for studying the structure and quality of datasets*, L. Caccetta, V. Rehbock, Industrial Optimisation, Vol 1, Proceedings of Symposium in Industrial Optimisation, Western Australian Centre of Excellence in Industrial Optimisation (WACEIO), Curtin University, Perth, Australia, 2003.

[14] A. Rubinov, N. Soukhoroukova, J. Ugon, *Minimization of the sum of minima of convex functions and its application to clustering*, to appear in Continuous optimization: current trends and applications, Springer.

[15] http://www.ciao-go.com.au/index.php