Clusters With Core-Tail Hierarchical Structure And Their Applications To Machine Learning Classification

Dmitriy Fradkin
Dept. of Computer Science
dfradkin@paul.rutgers.edu
1-732-445-4578

Ilya B. Muchnik
DIMACS
muchnik@dimacs.rutgers.edu
1-732-445-0073

Rutgers, The State University of New Jersey
110 Frelinghuysen Rd., Piscataway, NJ 08854-8019

Abstract— *We present a method for analysis of clustering results. This method represents every cluster as a stratified hierarchy of its subsets of objects (strata) ordered along a scale of their internal similarities. The "layered structures" can be described as a tool for interpretation of individual clusters rather than for describing the model of the entire data. It can be used not only for comparisons of different clusters, but also for improving existing methods to get "good" clusters. We show that this approach can also be used for improving supervised machine learning methods, particularly "active machine learning" methods, by specific analysis and pre-processing of a training data.*

1. Introduction

Clustering is a standard technique for analysis of complex data, particularly when the data is very large and statistical hypotheses of its generation are unknown or difficult to formulate. Most data mining software packages support the results of clustering with tools that provide some statistical characteristics of clusters, and their visual representations. While these tools significantly improved general understanding of the structure of and dependencies between clusters, their frequent use in practice has shown that interpretation of clustering results is a new essential problem in need of its own methods.

In this paper we present a novel method of cluster interpretation which is related to the so-called Layered Cluster method proposed in [1] where the layered structure is used to characterize all clusters together. In contrast we apply the same idea for analysis of individual clusters.

The paper is structured as follows. We summarize some background work. We then describe in general terms how cluster analysis can be conducted. The Experiments section provides description and results of such analysis of 4 well-known supervised learning datasets.

2. Background

Layered Patterns— In [1] the authors suggested a method for extracting a set of layered patterns of a set of points. This was done by defining a tightness function on a set and finding a sequence of nested sets where this tightness function achieves local maxima. Such a sequence exists and can be found by a fast polynomial-time algorithm.

We apply these ideas to analysis of individual clusters generated by a clustering procedure, such as K-Means [2], and show how such analysis can be useful for interpreting clustering results and pre-processing training data for supervised learning problems.

K-Means Clustering— The clustering algorithms that we concentrate on are K-Means algorithms. These methods aim to minimize the criterion:

$$E = \sum_{j=1}^{K} \sum_{x_i \in S_j} d^2(x_i, s_j) \tag{1}$$

where K is the number of clusters, $s_j = \frac{\sum_{x_i \in S_j} x_i}{|S_j|}$ (i.e. s_j is the center of cluster S_j), $d(x, y)$ is Euclidean distance. This criterion was first described by Lloyd [3] and Fisher [4] and had since been used in many applications.

Current methods for minimizing (1) usually consist of taking the best solution out of a set of local minima found from a number of randomly generated initial conditions.

[0] Research reported in this paper was supported in part by funding from SROA, Rutgers University.

3. Methodology

Estimating Difficulty of a Learning Problem by K-Means Clustering of Training Data— Suppose we have a training set W of points for some pattern recognition problem. Each element x_i of the set is a vector $x_i = ({x_i}^1, \dots, {x_i}^n)$ in R^n and has a label y_i determining its class. The labels y_i of the data induce an "external" partition of the data into K non-overlapping classes $P_e = \{D_1, \dots, D_K\}$.

By performing cluster analysis (K-Means into K clusters) and comparing its results with the external partition we obtain an indication of how appropriate the feature space and training set are for the supervised learning problem.

It is clear that if the external classification corresponds to the good (low) value of (1) then the learning problem is easy - all points from one class occupy a small region around its center and are far away from other centers.

Alternatively, it can be said that the feature space for such a problem was chosen appropriately for the problem, since distances between points in this space are indicative of their class labels (that is distances between points in a class are smaller than inter-class distances).

The following criterion might be useful in estimating the difficulty of a problem. Let $E(P_e)$ be value of criterion (1) on the external partition of the data:

$$E(P_e) = \sum_{j=1}^{K} \sum_{x_i \in D_j} d^2(x_i, d_j) \tag{2}$$

where d_j is the center of class D_j.

Let $E(P_k)$ be the minimal value of (1) (achieved on partition $P_k = \{S_1, \dots, S_K\}$ of W) for the same number of clusters K.

$$E(P_k) = \sum_{j=1}^{K} \sum_{x_i \in S_j} d^2(x_i, S_j). \tag{3}$$

Let $\rho = E(P_e)/E(P_k)$ - the ratio of E-value of the external partition to that of minimal K-Means partition. If the ratio ρ is close to 1, we may consider the classification problem to be relatively easy. However, if the ratio is large, then the problem is "hard" in the chosen feature space.

We can form a better understanding of the problem if, in addition to comparing E values of the external and K-Means partitions, we consider the confusion matrix defined by these. This matrix can indicate which clusters are "naturally different" from others, and which clusters are close to each other in the feature space.

Core-Tail Cluster Structure For Estimating Difficulty of a Learning Problem— The methods described in the previous section may provide interesting interpretations but they work on the level of the whole dataset and might be too general. In order to provide better interpretation for the data we use the ideas of layered cluster structure suggested in [1].

We are going to separate each cluster (both in the training partition and in the ones produced by K-Means) into two layers (core and tail) and then analyze each layer as discussed in the previous section. Informally, the core of a cluster is a part that has a "high concentration" of data. The tail is a complementary part of the same cluster.

We can separately compute contributions of tails and cores to criterion (1) and compute separate similarity coefficients for cores and tails.

The idea is that if there are a lot of mismatches between existing classification and K-Means partition it is important to understand whether such mismatches are widespread, whether they occur only for specific clusters or only in small regions along the boundaries between the classes. Similarly, such analysis indicates whether the difference in the value of (1) between partitions is due to several boundary points, or whether the partitions are entirely different.

If the mismatches occur throughout the layers, then the classification problem is difficult. However, if the mismatches occur only in the tails then we can conclude that the problem is solvable, though not easy.

Additional benefit of such analysis is that we can use its results to weight different points according to their importance to training. For example, in a cluster where both core and tail points in external and K-Means partitions fall into the same cluster, we can give a low weight to core points in the training process of SVM [2] classifiers, since tail points can be associated with a high weight. The tail points in such cases can be seen as informal analogs of "support vectors".

The above ideas can be straightforwardly applied to situations where we partition clusters into more than 2 layers. In these situations we would be interested in points lying in the outmost correctly-clustered layers. The points outside such layers are not reliable predictors since they are mixed with points from the other classes, while the points inside these layers provide redundant information. The predictions (or training) can then be based only on the "combinatorial support vectors" and should have comparable accuracy.

Formal Model For Core and Tail of a Cluster— Let us define a similarity function on two elements i, j of a cluster S:

$$s(i,j) = e^{-\frac{d^2(i,j)}{\alpha}},\ \alpha = \sigma^2(S)/|S|, \tag{4}$$

where σ^2 is defined for any set H as

$$\sigma^2(H) = \sum_{i \in H} (x_i - \overline{x})^2 \tag{5}$$

The coefficient α is used to make similarity between two points reflect the structure of the set.

We define a monotone linkage function between an element i and a subset H of a cluster S:

$$\pi(i, H) = \sum_{j \in H} s(i, j), \ i \neq j, \ \forall i \in S, H \subseteq S \tag{6}$$

The core C_S is now defined as:

$$C_S = argmax_{H \in 2^S} \min_{i \in H} \pi(i, H) \tag{7}$$

and the tail is:

$$T_S = S - C_S \tag{8}$$

Such cores and tails can be found using a fast procedure, quadratic in number of points in S.

4. EXPERIMENTS

Data— For the experiments we used four well-known datasets from UCI [5].

1. Image Segmentation dataset has 7 classes, 19 features and 2310 observations. The points correspond to 3x3 pixel regions randomly drawn from 7 outdoor images (with the name of the image serving as a label). The features are computed based on color intensities and coordinates of the pixels in the regions. All features are numeric. Best classification results obtained on this dataset with 10-fold cross validation were around 96%-97% [6].

2. Pendigit dataset consists of 10992 points represented by 16 features and a label (a digit from 0 to 9). This dataset was created by collecting 250 samples of digits written on a pressure-sensitive tablet from each of 44 writers. The features were obtained by spacial resampling of the points in the plane and are integers in the range [0,100]. Known results of classification on a specific split of this data (7494 training and 3498 test points) using k-Nearest Neighbors methods all show above 97% accuracy [7].

3. Satellite Image (Satimage) consists of 6 classes, 36 features and 6435 (4435 train and 2000 test) observations. Points correspond to multi-spectral values of pixels in a 3x3 regions of a satellite image (4 spectral bands for each of the 9 pixels). The features are numerical, in the range 0 to 255. The labels are soil types. Best results, achieved with Nearest Neighbor methods, were 90.6% accuracy on the test data [6].

4. The fourth dataset was a **Vowel dataset**. There are 992 points, corresponding to 11 English vowel sounds, represented by 10 features and a label. The features are derived from analysis of sample windowed segments of the speech signal and are real-valued. The best classification result, obtained with Nearest Neighbor Method, achieved 56% accuracy [2].

In each of the datasets, the sizes of clusters are approximately the same. We conduct all experiments on training and test data combined, since in the relevant supervised learning problems the labels are available for all such points. The value of K is taken to be the number of classes in the data.

Results— To obtain minimal K-Means partition $P_k(W)$ we perform K-Means with 10 random initial conditions and take the best result (i.e. result with the lowest value of (1)).

We will denote core and tail of cluster S_i by C_i and T_i (and sometimes $C(S_i)$ and $T(S_i)$) (and will indicate the dataset W when necessary). Also, we will refer to the set of all core points of a dataset as $C = \cup_{i=1,\ldots,K} C_i$. Similarly for the tail points: $T = \cup_{i=1,\ldots,K} T_i$.

A *contribution* $Q(H, P)$ of set $H \subseteq W$ to $E(P(W))$ for some partition $P = \{S_1, \ldots, S_K\}$ of W is:

$$Q(H, P) = \sum_{j=1}^{K} \sum_{x_i \in S_j \cap H} d^2(x_i, s_j). \tag{9}$$

where s_j is the center of cluster S_j.

Instead of providing the raw values of E for datasets, we provide the ratio of these values to $\sigma^2(W)$. This is indicative of the quality of clustering and has an advantage of making the results comparable across datasets.

Thus, for a dataset W and its partition P we write

$$\Sigma(P) = \frac{E(P)}{\sigma^2(W)} \tag{10}$$

where E_P is computed according to (2) or (3). Σ_H is defined for contribution $Q(H, P)$:

$$\Sigma_H(P) = \frac{Q(H, P)}{\sigma^2(W)}, H \subseteq W \tag{11}$$

We will write just Σ and Σ_H when it is clear what W and P are. Clearly, $0 \leq \Sigma_H \leq \Sigma \leq 1$ for any $H \subseteq W$ and $P(W)$.

We also compute 3 confusion matrices for each dataset, to compare the best K-Means result and the external classification on the whole clusters, on the cores only and on the tails only.

For each confusion matrix we compute the optimal assignment and the ratio of number of points that are then assigned to the diagonal (of the optimal assignment matrix) to the total

TABLE 1. Evaluation of the datasets

Data	Accuracy	$\Sigma(P_e)$	$\Sigma_c(P_e)$	$\Sigma(P_k)$	$\Sigma_c(P_k)$	$M(P_e, P_k)$	$M(C_e, C_k)$	$M(T_e, T_k)$
Image	97%	0.516	0.114	0.268	0.100	0.569	0.572	0.623
Pendigit	>97%	0.465	0.168	0.302	0.099	0.667	0.820	0.577
Satimage	90.6%	0.327	0.090	0.209	0.054	0.682	0.750	0.668
Vowel	56%	0.644	0.280	0.370	0.216	0.314	0.403	0.309

TABLE 2. K-Means (P_k) clusters of Image Dataset

Clusters	N_S	$N_{C(S)}$	α	Σ_S	$\Sigma_{C(S)}$	β	γ
0	527	377	0.715	0.049	0.021	0.434	0.607
1	276	88	0.319	0.034	0.007	0.213	0.669
2	220	133	0.605	0.014	0.005	0.385	0.638
3	590	386	0.654	0.037	0.013	0.365	0.558
4	331	134	0.405	0.041	0.008	0.192	0.473
5	354	303	0.856	0.034	0.024	0.695	0.812
6	12	6	0.500	0.058	0.021	0.352	0.704

TABLE 3. K-Means (P_k) clusters of Pendigit Dataset

Clusters	N_S	$N_{C(S)}$	α	Σ_S	$\Sigma_{C(S)}$	β	γ
0	931	582	0.625	0.026	0.008	0.323	0.516
1	812	268	0.330	0.032	0.007	0.213	0.646
2	466	271	0.582	0.015	0.005	0.323	0.555
3	651	389	0.598	0.011	0.003	0.307	0.514
4	1147	691	0.602	0.032	0.011	0.356	0.592
5	740	413	0.558	0.017	0.005	0.310	0.555
6	1734	825	0.476	0.050	0.011	0.213	0.449
7	961	543	0.565	0.030	0.011	0.354	0.627
8	1140	633	0.555	0.025	0.007	0.282	0.508
9	2410	1652	0.685	0.063	0.031	0.492	0.717

TABLE 4. K-Means (P_k) clusters of Satimage Dataset

Clusters	N_S	$N_{C(S)}$	α	Σ_S	$\Sigma_{C(S)}$	β	γ
0	1164	727	0.625	0.034	0.011	0.338	0.541
1	856	516	0.603	0.047	0.013	0.267	0.443
2	1366	786	0.575	0.028	0.007	0.252	0.437
3	981	414	0.422	0.023	0.006	0.262	0.622
4	572	256	0.448	0.035	0.008	0.229	0.512
5	1496	739	0.494	0.042	0.009	0.212	0.430

TABLE 5. K-Means (P_k) clusters of Vowel Dataset

Clusters	N_S	$N_{C(S)}$	α	Σ_S	$\Sigma_{C(S)}$	β	γ
0	109	73	0.670	0.036	0.018	0.494	0.738
1	89	45	0.506	0.036	0.011	0.306	0.605
2	81	56	0.691	0.031	0.016	0.507	0.733
3	114	82	0.719	0.039	0.022	0.571	0.793
4	91	82	0.901	0.039	0.033	0.854	0.948
5	124	110	0.887	0.037	0.031	0.824	0.929
6	76	62	0.816	0.034	0.024	0.727	0.892
7	69	48	0.696	0.027	0.014	0.518	0.745
8	48	24	0.500	0.015	0.004	0.289	0.577
9	84	47	0.560	0.036	0.016	0.454	0.812
10	105	83	0.790	0.040	0.026	0.664	0.840

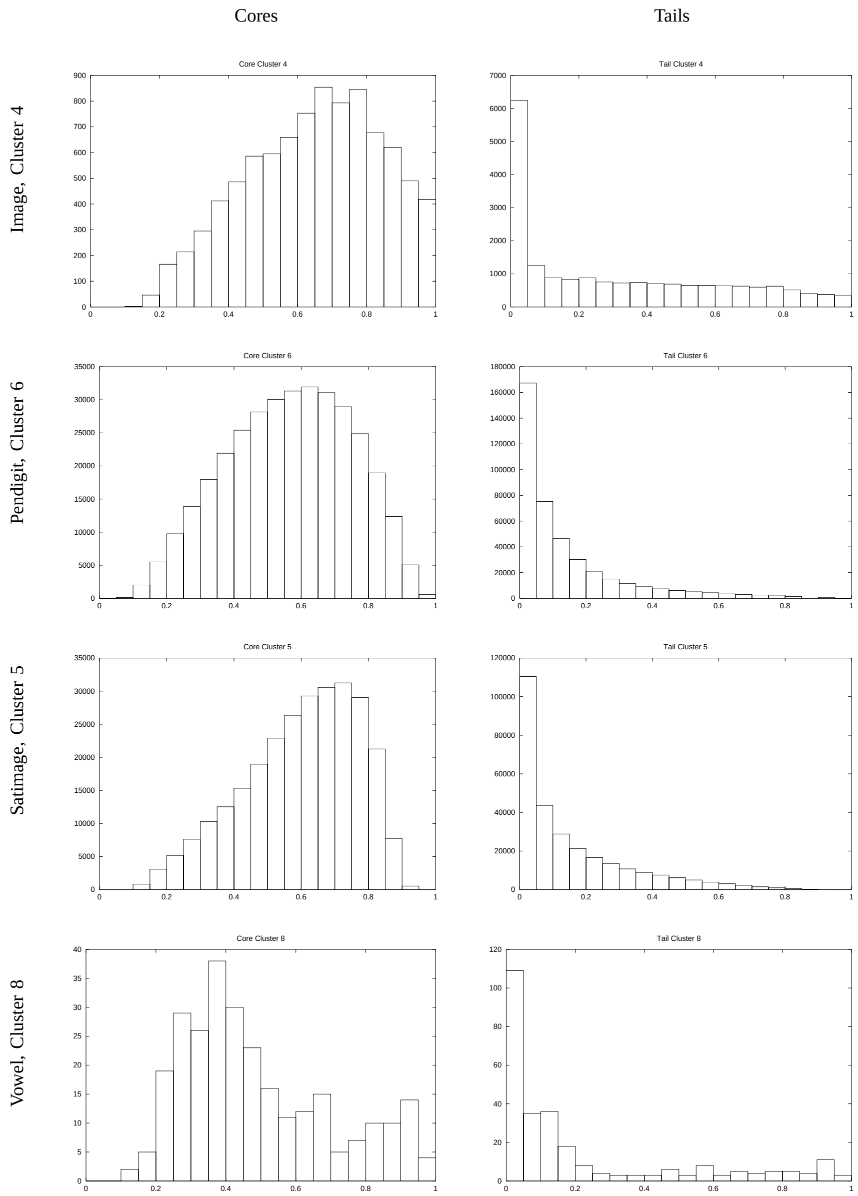

Figure 1. Histograms of inter-point similarities within cores and tails of a cluster from each of the four dataset.

number of points. In the Table 1 we denote this correspondence coefficient for external and K-Means partitions P_e and P_k by $M(P_e, P_k)$. Similarly, this ratio for cores and tails (of K-means partition) is denoted by $M(C_e, C_k)$ and $M(T_e, T_k)$. High values of M indicate that K-Means partition (or its parts) matches the external partition well, and that therefore the learning problem should be easy.

Table 1 contains the information on each of the four datasets as a whole. The first column gives the name of the dataset, the second one shows the results of supervised learning experiments taken from the literature. The next two columns give the value of Σ for the external partition P_e and the contribution Σ_C of the cores of external partition to this value. The following two columns give the same values for the best of K-Means partitions. The last three columns show the correspondence coefficients M for the whole clusters, cores and tails of K-Means partitions against the external partition.

The next four tables provided a more detailed view of each of the datasets. They show analysis of best K-Means clustering result. For each cluster S we give the number of points in the cluster N_S and its core $N_{C(S)}$ and the contribution of the cluster Σ_S and it core $\Sigma_{C(S)}$ to $\Sigma(P_K)$. In order to better illustrate the nature of the core we provide the following ratios: $\alpha = \frac{N_{C(S)}}{N_S}$, $\beta = \frac{\Sigma_{C(S)}}{\Sigma_S}$ and $\gamma = \frac{\beta}{\alpha}$. Intuitively, α indicates what fraction of points in a cluster are core points, while β shows what is the ratio of core's contribution to the cluster's Σ_S value. Then the ratio γ is indicative of the density of the core. The lower the value of γ is, the smaller is the contribution of the core points (in proportion to their number), implying that core points are all close to the center. These values for each dataset are displayed in **Tables 2-5**.

Figure 1 contains histograms of inter-point similarities in the core and tail of one of the clusters of each dataset. The similarities were computed as in (4). The cluster was chosen from each dataset to have the lowest value of γ. So in some sense we chose "the best" clusters (the ones with the densest cores) for the histograms. In all cases, there is a clear difference between histograms of cores and tails: core points are much more similar to each other than tail points are.

Image Dataset: Table 1 shows that external and K-Means partitions match equally well both in the cores and in the tails (with latter being somewhat better). However the ratio of overall to core contribution is about 2.6, indicating that the cores are quite dense. This conclusion is supported by the histograms (Figure 1).

So on the one hand the feature space induces a good partition (i.e. dense cores), but on the other hand the induced partition does not appear to match the external partition well. The results of supervised learning experiments seem difficult to predict. A more detailed analysis of confusion matrix might help discover "problem" areas or classes.

Pendigit Dataset: K-Means matches external partition significantly better for the cores than for the tails, and the correspondence coefficients M are quite high for both both cores and tails (Table 1). The cores are dense, judging by relatively low γ values in Tables 3 and by clearly unimodal core histogram with peak above 0.6 in Figure 1. These factors combined indicate that the feature space is appropriate for the problem and that good classification results can be expected. It seems likely that some preprocessing to reduce the sizes of training data would speed up learning without compromising the performance.

Satimage Dataset: The results for this dataset are presented in Table 4 and in histograms in Figure 1. They lead to conclusions similar to those for the Pendigit Dataset.

Vowel Dataset: Table 1 shows that the Vowel dataset has a higher ratio of $E(P_k)$ to $E(C_k)$ values, and the correspondence coefficients M between the external and K-Means partition are much lower, than in the other three problems. Tables 2-5 show that the value of γ is much higher on average in clusters of Vowel dataset, indicating that the cores of these clusters are not much denser than the tails. The α values (ratios of sizes of the cores to the sizes of the clusters) are also quite varied. The histograms in Figure 1 confirm this view: in the best cluster (lowest γ value), similarities between core points are much closer to those of tail points than for any other dataset. Furthermore, the histogram of core similarities is not unimodal. This implies that though the core points are more similar to each other than to tails, there is a lot of dissimilarity between them. These factors combined characterize Vowel learning problem as a difficult one.

5. Discussion

This paper suggested new ways of using clustering procedures for the analysis of training data for machine learning problems. This approach may lead to improvement of existing learning methods, especially in the field of "active machine learning".

We also provided illustrative examples of usefulness of the concepts of cores and tails of clusters. As can be seen from the results, core points are much closer to each other than the tail points and their contribution to the value of (1) is usually significantly less (relative to their number) than that of the tail points. Also, cores have much higher correspondence to the external classification than do tails.

Our results also indicate that datasets can vary greatly even in terms of how dense the cores are and how well they match external partition. Pendigit and Vowel datasets demonstrate the extremes.

References

[1] B. Mirkin and I. Muchnik, "Layered clusters of tightness set functions," *Applied Mathematics Letters*, vol. 15, pp. 147–151, 2002.http://www.datalaundering.com/download/mm012.pdf
http://www.datalaundering.com/download/monsysp.pdf

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. New York: Springer, 2001.

[3] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, pp. 128–137, 1982.

[4] W. D. Fisher, "On grouping for maximum homogeneity," *J.Am.Stat.Assoc*, vol. 53, pp. 789–798, 1958.

[5] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: www.ics.uci.edu/~mlearn/MLRepository.html

[6] D. Michie, D. Spiegelhalter, and C. Taylor, Eds., *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ: Prentice Hall, 1994.

[7] F. Alimoglu and E. Alpaydin, "Methods of combining multiple classifiers based on different representations for pen-based handwriting recognition," in *Proceedings of the TAINN 96*, 1996.