

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

Dmitriy Fradkin and Ilya Muchnik

DIMACS, Rutgers University, NJ

October 3, 2003

Overview

We present a method for interpreting clustering results. It is based on partitioning each cluster into core and tail.

- Using this idea we can analyze not only clustering result, but any predefined partition of data
- From supervised learning perspective it is interesting to compare the partition given by supervisor on a training data with the partition of the same data produced by a clustering procedure.
- The concept of cores and tails of clusters makes this comparison more interpretive.

This talk describes results of such comparison conducted on 4 machine learning datasets.

K-Means Clustering

We consider only one clustering procedure, K-Means clustering.

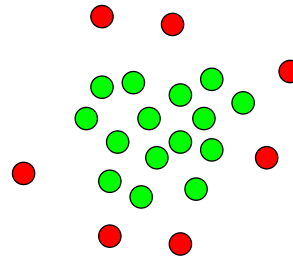
This procedure tries to find a **partition** $P(W) = \{S_1, \dots, S_K\}$ of a set W , $W = \bigcup_{i=1}^K S_i$, into K non-overlapping groups of objects that minimizes the following criterion of quality of the partition $P(W)$:

$$E(P(W)) = \sum_{i=1}^K \sum_{x \in S_i} d^2(x, s_i)$$

where $s_i = \frac{\sum_{x \in S_i} x}{|S_i|}$ (i.e. s_i is the center of cluster S_i), $d(x, y)$ is Euclidean distance.

Core-Tail Structure of a Cluster

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.



Formal Model for Core and Tail of a Cluster

The similarity function on two elements k, l of a cluster S_i :

$$a(k, l) = e^{-\frac{d^2(k, l)}{\sigma^2(S_i)/|S_i|}},$$

where

$$\sigma^2(S_i) = \sum_{x \in S_i} (x - s_i)^2.$$

Formal Model (Continued)

We define a linkage function $\pi(k, H)$ between an element k and a subset H of a cluster S_i :

$$\pi(k, H) = \sum_{l \in H} a(k, l), \quad \forall k \in S, H \subseteq S$$

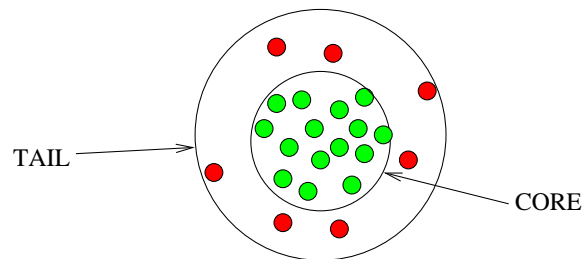
The core C_{S_i} is now defined as:

$$C_{S_i} = \operatorname{argmax}_{H \subseteq S_i} \left(\min_{k \in H} \pi(k, H) \right)$$

and the tail is:

$$T_{S_i} = S_i - C_{S_i}$$

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



Cores and Tails of Clusters in Training Set Analysis

Let us denote core and tail of cluster S_i by C_i and T_i .

Definition: A *contribution* $Q(H, P(W))$ of set $H \subseteq W$ to the quality $E(P(W))$ of the partition $P(W)$, $P(W) = \{S_1, \dots, S_K\}$, of W is:

$$Q(H, P(W)) = \sum_{i=1}^K \sum_{x \in S_i \cap H} d^2(x, s_i).$$

where s_i is the center of cluster S_i .

Using this definition we can calculate the quality values $Q(C_i, P(W))$ and $Q(T_i, P(W))$ for cores and tails of all clusters S_i in the partition $P(W)$.

Normalized Criteria

To be able to compare quality of partitions over different set, we consider normalized criterion:

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

$\Sigma_{C_i}(P(W))$ is the normalized criterion related to contribution of the core C_i of cluster S_i :

$$\Sigma_{C_i}(P(W)) = \frac{Q(C_i, P(W))}{\sigma^2(W)}.$$

Similarly we define criterion $\Sigma_{T_i}(P(W))$ for tails of clusters.

Relation between Classification and Clustering

Both supervised classification and clustering of training data can be represented as partitions $P_e(W)$ and $P(W)$ respectively. To simplify description we consider situations where the number of class in $P_e(W)$ and the number of clusters in $P(W)$ are the same, equal to K .

Two Statistics for Partitions

For a given class/cluster S of partition P , consider the following two statistics:

- $\frac{N_{C(S)}}{N_S}$ - fraction of points in a cluster that are core points
- $\frac{\Sigma_{C(S)}(P)}{\Sigma_S(P)}$ - ratio of core's contribution to the value Σ_S of the cluster

Statistics for the Pendigit Dataset

$P(W)$ clusters:

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

$P_e(W)$ clusters:

Clusters	N_S	$N_{C(S)}$	$\frac{N_{C(S)}}{N_S}$	Σ_S	$\Sigma_{C(S)}$	$\frac{\Sigma_{C(S)}(P)}{\Sigma_S(P)}$
0	1143	632	0.553	0.048	0.009	0.197
1	1143	501	0.438	0.059	0.017	0.279
2	1144	569	0.497	0.020	0.005	0.265
3	1055	534	0.506	0.014	0.004	0.294
4	1144	691	0.604	0.032	0.011	0.348
5	1055	567	0.537	0.100	0.037	0.370
6	1056	604	0.572	0.021	0.006	0.312
7	1142	654	0.573	0.039	0.011	0.271
8	1055	844	0.800	0.081	0.055	0.680
9	1055	470	0.445	0.051	0.013	0.248

Correlations between $P_e(W)$ and $P(W)$

Consider the matrix $A = a_{ij}$, where a_{ij} is the number of elements that belong to S_i in $P(W)$, that also belong to class j in $P_e(W)$.

Using this matrix A we calculate a coefficient $\alpha(A)$ that is indicative of the optimal correspondence $M(P_e, P)$ between $P(W)$ and $P_e(W)$:

$$\alpha(A) = \frac{1}{|W|} \sum_{j=1}^K a_{jj}^M$$

We use this coefficient to estimate correlation between $P_e(W)$ and $P(W)$. We can similarly compute correspondences between cores and tails of clusters in partitions $P_e(W)$ and $P(W)$: $\alpha(A_C)$ and $\alpha(A_T)$ respectively.

Estimates of Correlations between $P_e(W)$ and $P(W)$

Dataset W	K	$ W $	Dim.	$\alpha(A)$	$\alpha(A_C)$	$ A_C $	$\alpha(A_T)$	$ A_T $
Image	7	2310	19	0.569	0.674	1023	0.682	538
Pendigit	10	10992	16	0.677	0.874	5043	0.509	3702
Satimage	6	6435	36	0.682	0.998	2475	0.665	2282
Vowel	11	992	10	0.314	0.498	522	0.411	175

Another Correlation Measure

Consider a contingency matrix $B_C = b_{ij}$, where b_{ij} is the number of elements that belong to C_i in $P(W)$, that also belong to class j in $P_e(W)$. We similarly define B_T for tails of $P(W)$.

We can compute $\alpha(B_C)$ and $\alpha(B_T)$ as before.

More Estimates of Correlations

Data	$\Sigma(P_e)$	$\Sigma_c(P_e)$	$\Sigma(P_k)$	$\Sigma_c(P_k)$	$\alpha(A)$	$\alpha(B_C)$	$\alpha(B_T)$
Image	0.516	0.114	0.268	0.100	0.569	0.572	0.623
Pendigit	0.465	0.168	0.302	0.099	0.677	0.820	0.577
Satimage	0.327	0.090	0.209	0.054	0.682	0.750	0.668
Vowel	0.644	0.280	0.370	0.216	0.314	0.403	0.309