
Principal Direction Partitioning in Text Data Mining

Daniel Boley
Computer Science and Engineering
University of Minnesota

<http://www.cs.umn.edu/~boley/PDDP.html>
Supported in part by NSF

1

Outline

- Divisive Partitioning for Unsupervised Clustering
- Related Methods
- Algorithmic Issues – Fast Lanczos Solver
- Experimental Results
 - Speed
 - Quality
 - Most Distinguishing Words
- Conclusions and Future Work

2

Divisive Partitioning for Unsupervised Clustering

- Unsupervised, as opposed to Supervised:
 - no predefined categories;
 - no previously classified training data;
 - no a-priori assumptions on the number of clusters.
- Top-down Hierarchical:
 - imposes a tree hierarchy on unstructured data;
 - tree is source for some taxonomic information for dataset;
 - tree is generated from the root down.
- Principal Direction Divisive Partitioning
 - operates on real-valued data, even with missing data;
 - embedded in high dimensional Euclidean space;
 - fast & scalable by using efficient Lanczos solver.
 - ranks attributes by ability to distinguish between clusters.

3

Divisive Partitioning

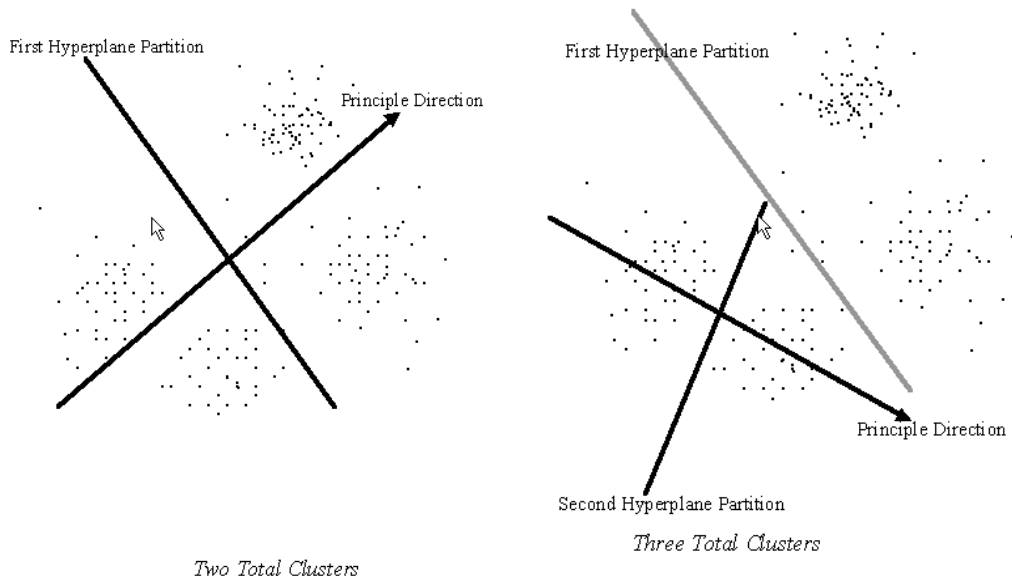
- Each document represented by n -vector \mathbf{d} of word counts.
- Each \mathbf{d} scaled to $\|\mathbf{d}\| = 1$ to make independent of document length.
- Vectors assembled into Term Frequency Matrix $\mathbf{M} = (\mathbf{d}_1 \ \dots \ \mathbf{d}_m)$.

Splitting Process:

- Get leading principal direction \mathbf{u} of $\mathbf{M} - \mathbf{w}\mathbf{e}^T$ with SVD, where $\mathbf{w} \triangleq \frac{1}{m}\mathbf{M}\mathbf{e} = \text{centroid}$, $\mathbf{e} \triangleq (1 \ \dots \ 1)^T$.
- Split documents by value of projection $\mathbf{u}^T(\mathbf{d}_j - \mathbf{w})$, $j = 1, 2, \dots$.
- Repeat recursively on each set of documents.
- Large entries in \mathbf{u} vector \iff most “distinguishing” words.
Large entries in \mathbf{w} vector \iff most “frequent” words.

4

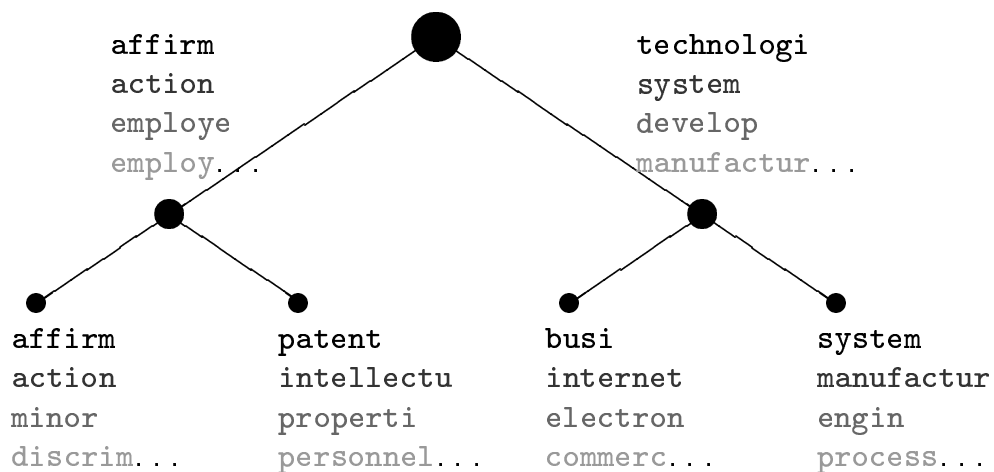
Divisive Partitioning - Splitting Step



5

Principal Direction Divisive Partitioning

- Start with root cluster representing all the documents.
- Split the root cluster into two children clusters.
- Recursively split each leaf cluster into two children
- Stop when stopping test satisfied.



6-X

Related Methods

Principal Component Analysis

- Select best rank k approximation to $\mathbf{M} - \mathbf{w}\mathbf{e}^T$.
- Result: original data represented with fewer degrees of freedom.

Latent Semantic Indexing

- Select best rank k approximation to \mathbf{M} .
- Get vectors giving inter-word relationships.
- Usual use: preprocessor for Info Retrieval (IR) tasks.

In contrast, PDDP is different

- Get just one eigenvector, but repeats on subparts.
- Unsupervised clustering method, not intended for IR (but you could use it as such).
- Getting one eigenvector much simpler computationally.

7

Algorithmic Issues – Fast Lanczos Solver

- Total cost dominated by cost of finding principal direction.
- Use efficient sparse matrix eigensolver “Lanczos”.
- Matrix used only to form matrix-vector products.
- Convergence depends on distribution of eigenvalues.
- On matrices of word counts from document sets, convergence appears to be fast (~ 20 iterations).
- Cost to find first principal direction:

Lanczos iters	·	mat-vec products per iter	·	cost of mat-vec product
~ 20	·	2	·	fill fraction $\cdot m \cdot n$
- Subsequent principal directions are cheaper [fewer documents].

8

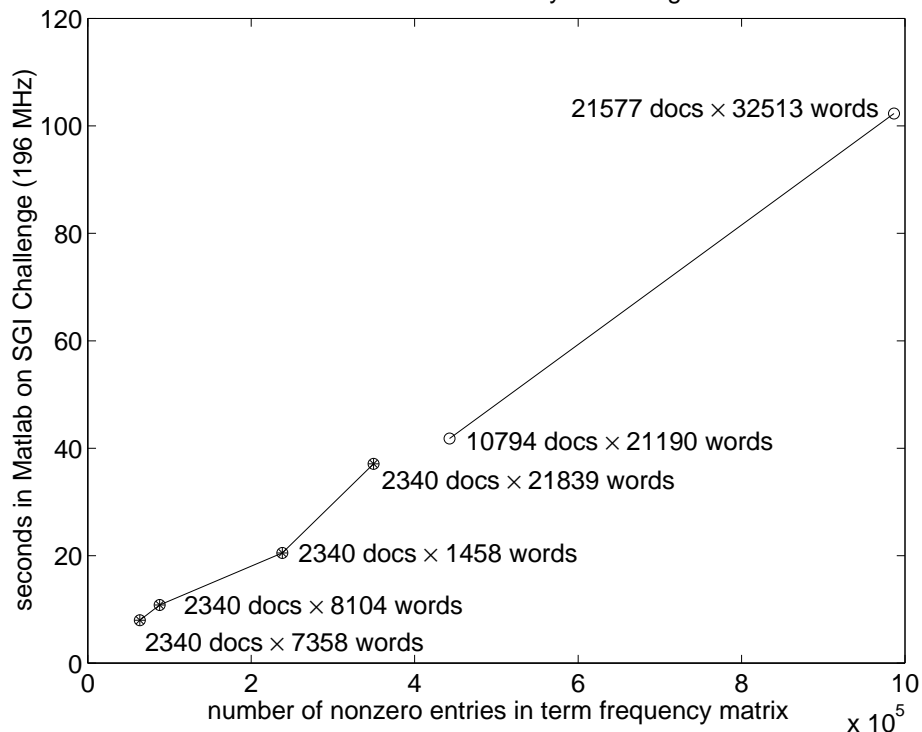
Experimental Results: Document Test Sets

Exp #	Term Frequency Matrix Size			Selection Criteria
	F-series	J-series	K-series	
1	98 × 5623	185 × 10536	2340 × 21839	all words
2	98 × 619	185 × 946	2340 × 7358	quantile filtering
3	98 × 1239	185 × 1763	2340 × 8104	top 20+ words
4	98 × 1432	185 × 2951		top 5+ words plus emphasized words
5	98 × 399	185 × 449	2340 × 1458	frequent item sets
6	98 × 2641	185 × 5106		all with TF > 1
7	98 × 1004	185 × 1328		top 20+ & TF > 1
8	98 × 827	185 × 1105		top 15+ & TF > 1
9	98 × 622	185 × 805		top 10+ & TF > 1
10	98 × 332	185 × 474		top 5+ & TF > 1
Reuters-21578			21577 × 32513	all documents
Reuters-21578			10794 × 21190	docs w/ topic labels

9

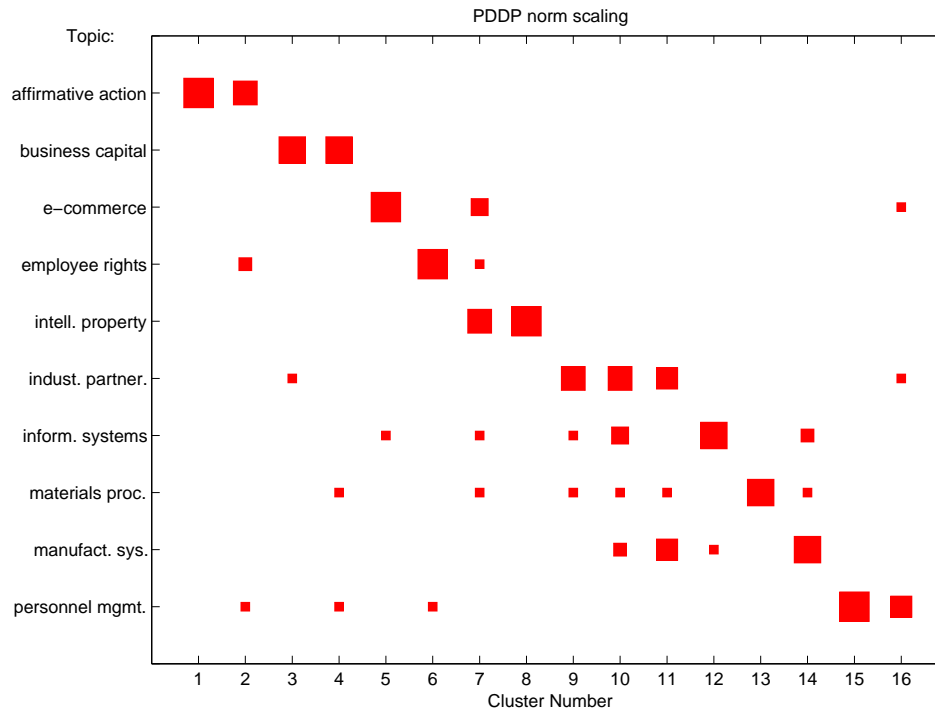
Speed on Text Documents

time to obtain 16 clusters by PDDP algorithm



10

Cluster Distribution



c

11

Cluster Contents

<i>cluster:</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
business	90	0	0	0	7	0	5	12	0	6	0	1	18	3	0	0
health	0	150	166	171	3	0	1	1	0	0	0	0	0	2	0	0
politics	2	0	0	0	100	1	2	0	0	1	0	2	1	5	0	0
sports	0	0	0	0	1	62	35	0	0	1	0	0	0	42	0	0
techno.	8	0	0	0	0	1	14	24	0	8	0	1	4	0	0	0
entertain.	24	0	0	4	11	4	22	61	135	131	148	159	143	137	204	206



topic



number of documents of each topic in each cluster

Confusion Matrix

Quality on Text Documents

- Need to “quantify” cluster quality.

- Our choice: use an Entropy measure:

- [Entropy in cluster i] $\triangleq \sum_{l \in \text{labels}} -\frac{n_{li}}{n_i} \log \frac{n_{li}}{n_i}$

where n_{li} = number of instances of label l in cluster i
 n_i = total number of instances in cluster i .

- [Total Entropy] $\triangleq \sum_i \frac{n_i}{n} \cdot [\text{Entropy in cluster } i]$,

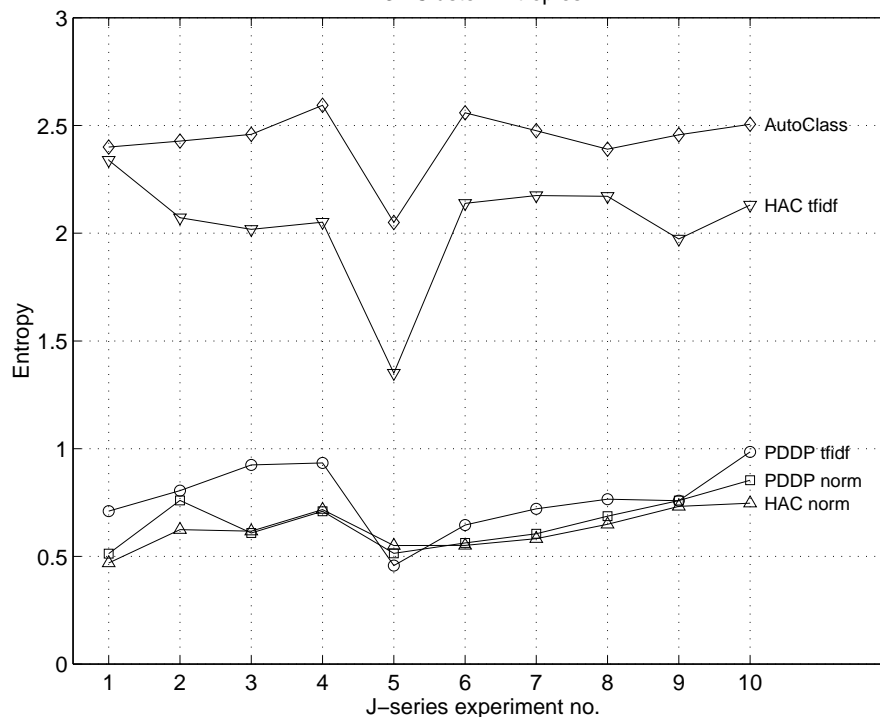
where n = total number of instances in entire collection.

- Entropy is 0 when each cluster has only one label.
Entropy is higher when clusters have mixed labels.
Entropy depends on the number of clusters.

13

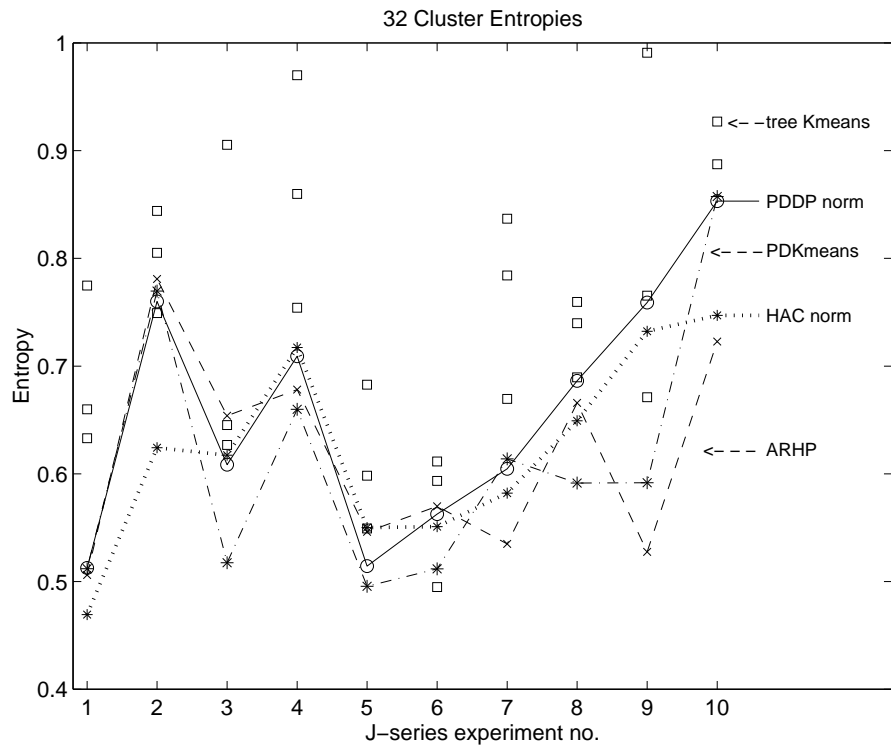
Quality on Text Documents

32 Cluster Entropies



14

Quality on Text Documents



15

Taxonomic Guide – Significant Words

Leading words for this cluster, from PDDP vectors

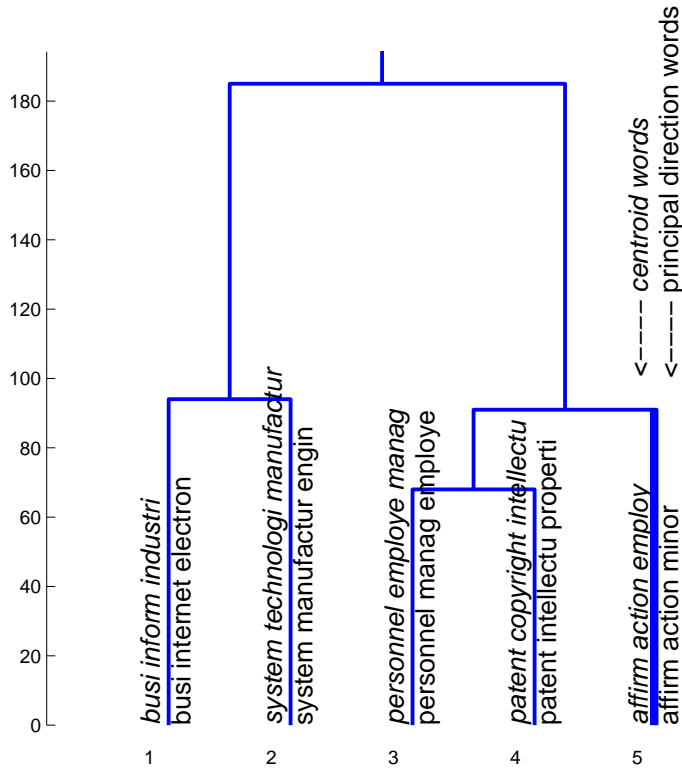
centroid help entertai health new polit sport tech bize index
principal dir. dengue fever dna gene gold probe diseases gubler guttman

The headlines for each of the 12 documents in the cluster.

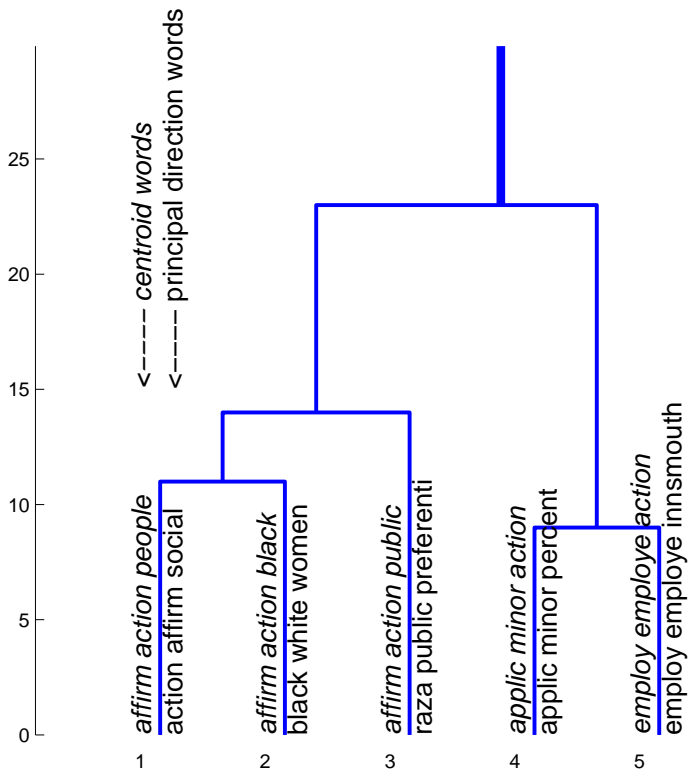
- + Enzyme Triggers Emphysema in Smokers
- + Drug May Offer New Way to Beat Colds
- + Lens Coat Cuts Cataract Cost
- + Gold Probe Detects DNA
- + "Dry Brushing" Beats Dental Plaque
- + Dengue Fever Strikes "Dr. Quinn"
- + Pregnancy Factor Protects Fetus
- + New Doping Agent for Athletes Reported
- + Dengue Fever Strikes "Dr. Quinn"
- + Gold Probe Detects DNA
- + FDA Moves to Ban Laxative Ingredient
- + Muscles Adapt to Exercise Lifestyle

16

Taxonomy from Hierarchical Structure



Taxonomy from Hierarchical Structure



Experiment: Search for MBTE on Altavista

Found 222 documents, clustered as follows:

CENTROID WORDS

62:found.serv.request.url.alt.html.fram.pleas.http.fil.de.ww
44:inc.servi.corp.ttm.compan.com.sit.stock.fre.pri.mrq.syste
38:fuel.car.gasolin.vehic.rav.re.com.gas.engin.pri.messag.su
78:wat.mtb.environmental.air.health.gasolin.program.sit.cali

PRINCIPAL DIRECTION WORDS

62:found.serv.url.request.html.pleas.apach.fil.port.htm.http
44:inc.corp.ttm.ltd.servi.corpor.international.mrq.stock.fin
38:rav.car.fuel.tir.subject.toyota.vehic.driv.wd.engin.com.h
78:wat.mtb.environmental.health.california.air.gasolin.clean

19-X

Ad hoc trial: search for mbte on Altavista

62:found.serv.url.request.html.pleas.apach.fil
23:found.serv.url.request.html.pleas.http
5:alt.tw.edu.descript.fre.net.sinica.new
5:fram.track.content.ih.returnto.tick.cl
29:transtec.inlin.de.fuelon.mb.pmn.ag.pr
44:inc.corp.ttm.ltd.servi.corpor.international
10:inc.corp.ltd.corpor.stock.international
6:ttm.mrq.mil.pri.compan.nm.ratio.inc.se
13:sit.ap.fre.inlin.research.sex.reut.liv
15:system.reg.board.micro.gms.provid.proc
38:rav.car.fuel.tir.subject.toyota.vehic.driv
8:rav.car.toyota.tir.wd.com.subject.driv
14:fuel.gasolin.vehic.car.gas.pri.engin.o
8:re.messag.previou.thread.mbt.farm.fuel
8:mcs.chemical.caus.chem.symptom.drug.ex
78:wat.mtb.environmental.health.california.ai
27:mtb.gasolin.wat.air.california.oxygena
13:environmental.health.conferen.top.wat.
17:wat.stat.sit.ypf.oil.environmental.com
21:cit.program.work.fund.school.issu.gove

20

Conclusions

- Unsupervised Clustering: get structure on large unstructured datasets.
- PDDP exhibits good scalability properties.
- PDDP generates clusters of high quality, comparable to other methods.
- PDDP identifies the distinctive attributes of the individual clusters [to get a more refined analysis of text docs, you'd need a parser].
- PDDP can be applied to non-text data.
- PDDP needs a self-contained, portable implementation.

21

Future Work

- Applications:
 - Organize Alcohol Laws for Minn. Health Dept. study
 - Classify speech recognition errors left over after all other processing.
 - Image data: classification or anomaly detection
 - Minnesota Automated Plate Scan [Sky Survey]
 - Find Distinctions among Gene Expression Data.
 - Predict toxic levels of pesticides from physical characteristics.
- Method Development
 - Two principal directions at a time (4-way split?).
 - Adjust hyperplanes during course of partitioning.
 - Adjust clusters at end to repair crude hyperplane cuts.
 - Study statistical significance of separation based on direction of maximal variance.
 - Handle datasets too big to fit in memory.

22