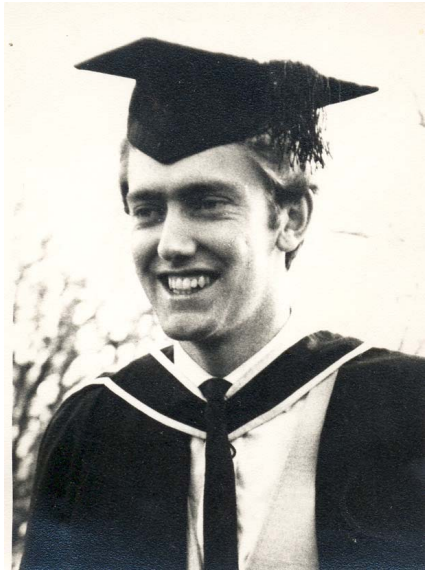
Cluster Analysis

- Homogeneous-objects or individuals in the same cluster are ‘alike’
- Separate-objects or individuals from different clusters are not ‘alike’

Good-1965

- For mental clarification and communication
- For discovering new fields of research,
- For planning an organizational structure,
- As a check list,
- For fun



Who Belongs in a Family

Robert L. Thorndike

Psychometrika, 1953, 18, 267-276

The arrival of Dr. Idnozs Tenib

‘I was sitting before my TV set, a while back, watching Captain Video and pondering the organizational problems of psychologists, psychometricians, psychodiagnosticians, psycho-somatists, psychosomnabulists, and psycho-ceramics (crack-pots to you). Wondering what I might do, in my small way to help, I decided to enlist Captain Video’s help to bring me from the Black Planet that super-galactian hypermetrician, Dr. Idnozs Tenib, cosmos-famous discoverer of Serutan.’

The Doctor Begins

‘Bring sample of population; I measure’

We set out to design a sample

‘The problem presented some interesting theoretical aspects, but the final solution was relatively simple. We stationed representatives at each of the three state beverage stores and followed every third badge-wearing individual who came out of a store. We selected only out-going patrons for obvious reasons. After assisting each respondent to unburden himself, we brought him to Dr. Idnozs (as we came to call him among ourselves) for study.’

Dr. Idnozs gives tests

- First is **Draw-a-Psychiatrist Test**
- We score this by if it gives horns.
- Next the **physiological test battery**
- We draw off saliva drop by drop and see does he drool when we bring in Skinner Box
- Later we come to the **Peculiar Preference Blank**
- Forced-choice; ‘Would you rather make mud pies or kiss gorgeous blond?’

The Doctor's Test Battery

Needless to say the tests were all orthogonal, completely diagnostic, of highest reliability, and representative of the fundamental dimensions of psycho-personality (the personality of psychologists and psychopaths).

Dr. Idnozs deals with the problems of scale and metric

'Is simple, take a number from one to ten. Is a score. Single digit. Standardized. When I say one equals one, one equals one.'

Dr. Idnozs replies

'No good. Have no a priori groups. Multiple discriminant only perpetuates sins of fathers. Tells which divisions to put man in. Not tell what divisions should be.'

Dr. Idnozs recommends

'We run cluster analysis. Find distances between sheep and goats. Assign to clusters so that average of distances within cluster is a minimum, when summed over all clusters. Define families, boundaries, and family membership like so.'

Dr. Idnozs deals with the optimization problem

‘Is easy, Finite number of combinations. Only 563 billion billion billion. Try all keep best.’

Dr. Idnozs takes his leave

‘Is dinner time. Don’t bother me.’

And the good doctor vanished rapidly into the stardust of outer space.

Data: Average ratings of 12 air force specialities on 19 attributes

Specialities:

Radio mechanic, aircraft mechanic, cook, supply technician, petroleum supply technician, clerk, career guidance specialist, personnel specialist, general instructor, budget and fiscal clerk, medical corpsman, air policeman

Attributes

Attributes

Strength, tools, fluency of expression, accuracy, manipulative ability, speed, spatial judgement etc etc

Three Group Solution

- **Group 1**

Radio mechanic, aircraft mechanic, petroleum technician

- **Group 2**

Cook, supply technician, medical corpsman, air policeman

- **Group 3**

Clerk, career guidance specialist, personnel specialist, general instructor, budget and fiscal clerk

Thorndike-The End

'At this point I can sense the bubbling up of doubts and questions: But what about your units?...How can you decide what dimensions to use?..What about the error variance in the location of a single specimen?...What has all this got to do with the organization of psychological associations?

I can do no better than emulate the good Dr. Tenib. Is time to go home. Sleep on question. Maybe tomorrow you give me answers.'

Friedman and Rubin-On Some Invariant Criteria for Grouping Data

'The objective is to analyze multivariate heterogeneous data and to present the results in such a way as to lend insight into the structure of the data so as to suggest more formal models for further analysis as well as to provide guidelines for the collection of other data.'

Friedman and Rubin-Methods

'The methods to be described apply to data consisting of *p* measurements on each of *n* objects where there is some reason to believe that these *n* objects are a heterogeneous collection. Further, the data should be such that the spatial distribution of the objects represented as points, can be meaningfully summarized by the location of the centre of gravity of each cluster and by the sample scatter matrix of each cluster.'

Friedman and Rubin-Hopes

‘Hopefully this type of analysis will be a step forward in helping to define clinically relevant subcategories of poorly defined illnesses such as schizophrenia, in isolating different disease syndromes or in defining useful categories in such fields as biological taxonomy.’

Friedman and Rubin-Begin With

T=**W**+**B**

T: Total scatter of the *n* points

W: Pooled within groups scatter

B: Between groups scatter

- For *p*=1 equation is a statement about scalars and leads to minimizing W as a natural criterion
- For *p*>1 equation involves matrices and the question of suitable criteria for grouping is more complex.

Friedman and Rubin-Suggested Criteria

- Minimization of trace(W)
- Minimization of |W|
- Maximization of trace $(W^{-1}B)$

Pearson’s Model for the Crab Data

$$f(x) = p\phi(\mu_1, \sigma_1) + (1-p)\phi(\mu_2, \sigma_2)$$

Crab data

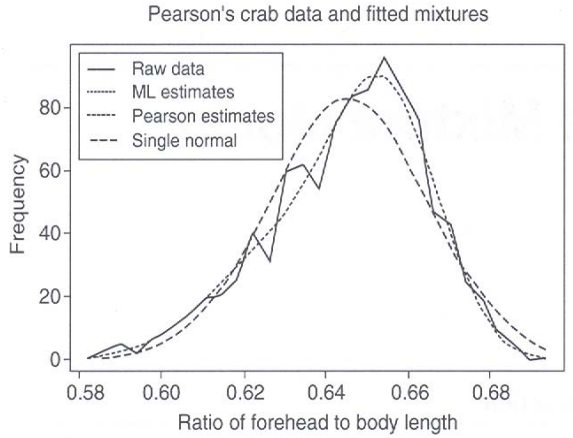


Figure 10.1 Frequency polygon of ratio of forehead to body length in 1000 crabs, fitted single normal density and two component mixtures fitted by moments and maximum likelihood.

Cluster Analysis Books-1970s

- *Cluster Analysis for Applications*-Anderberg, 1973
- *Cluster Analysis*-Everitt, 1974
- *Clustering Algorithms*-Hartigan, 1975

Finite Mixture Monographs

- *Finite Mixture Distributions*, Everitt and Hand, 1981
- *Statistical Analysis of Finite Mixture Distributions*, Titterington, Smith and Makov, 1985.
- *Finite Mixture Models*, McLachlan and Peel, 2000.

Data Mining

The nontrivial extraction of implicit, previously unknown and potentially useful information from data, or a process for discovering and presenting knowledge in a form that is easily comprehensible to humans

Classification

FLAME

Fisher's **L**inear **A**llocation **Me**thod

The Need for Data Mining?

The value of data is no longer in how much of it you have. In the new regime, the value is in how quickly and how effectively can the data be reduced, explored, manipulated and managed.

Usama Fayyad-President & CEO of digiMine Inc.

References-Clustering of gene expression data

- Eisen, Spellman, Brown and Bostein (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95, 14863-14868
- Wen et al (1998) Large-scale temporal gene expression mapping of central nervous system development, *Proc. Natl. Acad. Sci. USA*, 95, 334-339.
- Sherlock (2000) Analysis of large-scale gene expression data, *Curr. Opin. Immunol*., 12, 201-205.

References-genes' splice sites

- Ying et al (1996) GRAIL: A multi-agent neural network system for gene identification, *Proc. IEEE*, 84, 1544-1552
- Kulp et al (1996) A generalized hidden Markov model for the recognition of human genes in DNA, *Proc. Int. Conf. Intell. Syst.Mol. Biol*., 4, 134-142.

References

- Mechelen, Bock and DeBoeck (2004) Two-mode clustering methods: a structured overview, *Statistical Methods in Medical Research*, 13, 363-394.
- Friedman and Meulman (2004) Clustering objects on subsets of attributes, *J.R. Statist. Soc. B*, 66, 815-849.

Gap Statistic

- Computationally intensive method suggested by Tibshirani et al (2001)

Tibshirani et al (2001) Estimating the number of clusters in a data set via the gap statistic. *JRSS, B*, 63, 411-423.

Transaction Data

- Transaction data consists of collections of items.
- Typical example is market basket data where each transaction is the collection of items purchased by a customer in a single transaction.
- For example;

Bananas
Plums,lettuce,tomatoes,
Celery,confectionary,
Confectionary,
Apples,carrots,tomatoes,potatoes

Transaction Data

- Clustering of transaction data has an important role in the recent development of web technologies and data mining.

Wang et al (1999) Clustering transactions using large items. *Proc 8th International Conference on Information and Knowledge Management*, ACM Press

Time Series

- Clustering time series to find natural groupings of time series in a database under some similarity or disimilarity measure.
- Most classical clustering algorithms do not work well for time series data.
- Ramoni et al (2002) introduced a Bayesian method for grouping time series into clusters so that the elements of each cluster have similar dynamics.

Time Series

Ramoni et al (2002) Bayesian clustering for dynamics, *Machine Learning*, 47, 91-121

Keogh and Kasetty (2003) On the need for time series data mining benchmarks:A survey and empirical semonstration, *Data Mining and Knowledge Discovery*, 7, 349-371.

Caveats

- While techniques are important…knowing when to use them and why to use them is more important.
- In the long run it does not pay a statistician to fool either himself or his clients.
- But how in practice does one tailor statistical methods to the real need of the users, when the real need of the user is to be forced to sit and think?

THE END

Come back

Dr. Idnozs Tenib

We Need you!