META CLUSTERING

R Caruana, M Elhawary, N Nguyen, C Smith
IEEE International Conference on Data Mining, 2006

WHAT IS CLUSTERING

- Finding groups of similar objects in data
  - Clustering people with similar characteristics
    - Activities
    - Network of associations
    - Educational, socio-economic, background
    - Beliefs and behaviors
  - Clustering text/documents with similar characteristics
    - By content
    - By document type
    - By document intent
    - By intended audience
  - Clustering network events
    - By intent: attack vs. intrusion vs. denial of service vs. normal
    - By type: port scan vs. probe vs. ...

WHY CLUSTERING

- Data exploration
  - Our capacity to collect data has outstripped our capacity to understand/ interpret the data
  - Chicken and egg problem with new data
    - Don't know what you are looking for until you understand the data
    - Can't understand data until you know what you are looking for
  - Easier to find patterns in groups of objects than in single objects
  - As data grows bigger, but human brain remains fixed, must present experts with less raw, more processed data
  - Focused search and data analysis
    - soft/fuzzy/approximate/smart queries
  - Efficient transmission, presentation, summarization

STANDARD CLUSTERING IS INADEQUATE

- Disadvantages:
  - user in the loop
  - manually engineer distance metric
  - time consuming
  - requires significant expertise
  - final clustering often sub-optimal

NEW APPROACH: META CLUSTERING

- Automatically generate many different clusterings
- Cluster clusterings to organize results
- Present user with organized meta clustering
- Human out of loop: just select best clustering
- No need to manually engineer distance metric
- Faster, better final clustering for task at hand

MAIN GOALS

- Push as much work as possible required for clustering from the user to the computer
- Make clustering as automatic as possible
- More effective clustering in hands of users, not researchers
- Find better clusters/clusterings
- Find better clusters/clusterings faster
- Simultaneously provide multiple/alternate views of data
- Meta level helps users understand complex data faster
- Provide more natural user control and feedback

OVERALL APPROACH

1. Generate many good, yet qualitatively different, base-level clusterings of the same data
2. Measure the similarity between the base-level clusterings generated in the first step so that similar clusterings can be grouped together
3. Organize the base-level clusterings at a meta level and present them to the users

RESEARCH QUESTIONS

- How to generate different clusterings?
- How to measure distance between clusterings?
- How to organize clusterings for user?
- How to combine/merge clusterings?

GENERATING DIVERSE CLUSTERINGS

- Diverse clusterings from K-means minima
- Diverse clusterings from feature weightings

GENERATING DIVERSE CLUSTERINGS

- Diverse clusterings from K-means minima
  - K-means is run multiple times with different initializations, and each local minimum is recorded
  - **Finding**: the space of local minima is small compared to the space of reasonable clusterings, so an additional method for generating diverse clusterings is used...

GENERATING DIVERSE CLUSTERINGS

- Diverse clusterings from feature weightings
  - clustering many times with different random feature weights allows to find qualitatively different clusterings using the same clustering algorithm.

GENERATING DIVERSE CLUSTERINGS

- Diverse clusterings from feature weightings
  - feature weighting requires a distribution to generate the random weights
  - a **Zipf power law distribution** is used (empirical evidence shows that feature importance is Zipf-distributed in a number of real-world problems)

GENERATING DIVERSE CLUSTERINGS

- Diverse clusterings from feature weightings
  - A Zipf distribution describes a range of integer values from 1 to some maximum value K
  - The frequency of each integer is proportional to $\frac{1}{i^{\alpha}}$ where i is the integer value and α is the shape parameter

GENERATING DIVERSE CLUSTERINGS

- Diverse clusterings from feature weightings

Algorithm 3: Generate a diverse set of clusterings

Input: $\mathbf{X} = \{x_1, x_2, ..., x_n\}$ for $x_i \in R^d$, k is the number of clusters, m is the number of clusterings to be generated

Output: A set of m alternate clusterings of the data $\{C_1, C_2, ..., C_m\}$ for which $C_i : \mathbf{X} \mapsto \{1, 2, ..., k\}$ is the mapping of each point $x \in \mathbf{X}$ to its corresponding cluster

begin
  for $i = 1$ *to* m **do**
    $\alpha = rand(\text{"uniform"}, [0\ \alpha_{max}])$
    for $j = 1$ to d **do**
      $w_j = rand(\text{"zipf"}, \alpha)$
    end
    $\mathbf{X}_i = \emptyset$
    for $x \in \mathbf{X}$ **do**
      $x' = x \odot w$ where $\odot$ is pairwise multiplication
      $\mathbf{X}_i = \mathbf{X}_i + \{x'\}$
    end
    $C_i = \text{K-means}(\mathbf{X}_i, k)$
  end
end

Clustering clusterings at the meta level

- How to measure distance between clusterings?
- Measure based on Rand Index
- Given two clusterings:

$I_{ij} = 1$ if points i and j are in the same cluster in one clustering, but in different clusters in the other.

$I_{ij} = 0$ otherwise

Dissimilarity of two clusterings: $\frac{\sum_{i<j} I_{ij}}{N(N-1)/2}$

Agglomerative clustering at the meta level

- How to combine / merge clusterings?
- Meta clustering can be performed using any clustering algorithm that works with pairwise similarity data
- Agglomerative clustering is used
  - works with similarity data
  - does not require the user to specify the number of clusters
  - resulting hierarchy makes navigating the space of clusterings easier

EXPERIMENTAL RESULTS

Data Sets

Data Set	# features	# cases	# trueclasses	# clusters	# points in biggest class	# features in 95 % PCA
Australia	17	245	10	10	80	10
Bergmark	254	1000	25	25	162	130
Covertype	49	1000	7	15	476	39
Letters	617	514	7	10	126	141
Protein	ud format	639	N/A	20	N/A	N/A
Phoneme	10	990	15,11	15	N/A	9

PERFORMANCE METRICS

- **Compactness**: measures the average pairwise distance between points in the same cluster

$$\frac{\Sigma_{i=1}^{k} N_i \frac{\Sigma_{j=1}^{N_i-1} \Sigma_{k=j+1}^{N_i} d_{jk}}{N_i(N_i-1)/2}}{N}$$

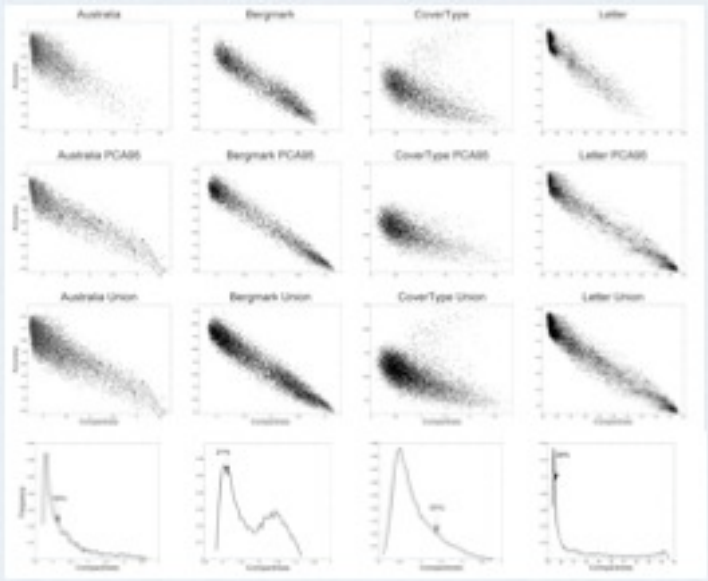
- **Accuracy** (using class labels)

Effect of Zipf Weighting

- As the α value increases, feature weighting explores a region of lower compactness
- Some of the most accurate clusterings are generated when applying feature weighting with higher α values
- A uniform distribution alone is insufficient to explore the clustering space

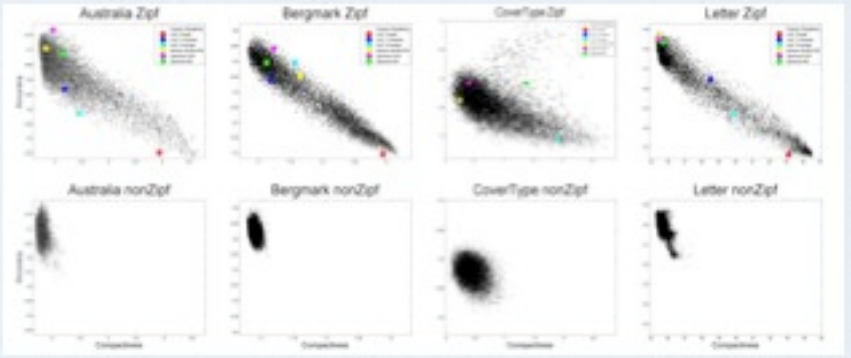
Feature Weighting Before and After PCA

- Although there is correlation between compactness and accuracy, the correlation is not perfect.
- Sometimes, the most accurate clusterings are not the most compact ones.
- PCA yields more diverse clusterings on some problems, less diverse clusterings on others.

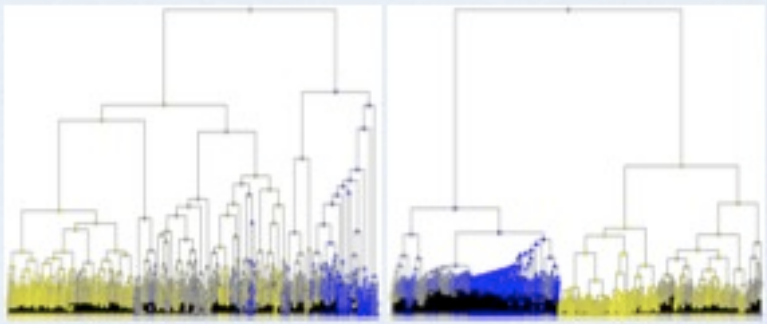


LOCAL MINIMA VS. FEATURE WEIGHTING

- For Australia, Bergmark, and Letter, weighting features yields more diverse clusterings.
- For Covertype, not applying feature weighting fails to discover the cloud of more accurate clusterings.
- For Australia, K-means finds more clusterings in the upper left corner (accurate and compact).



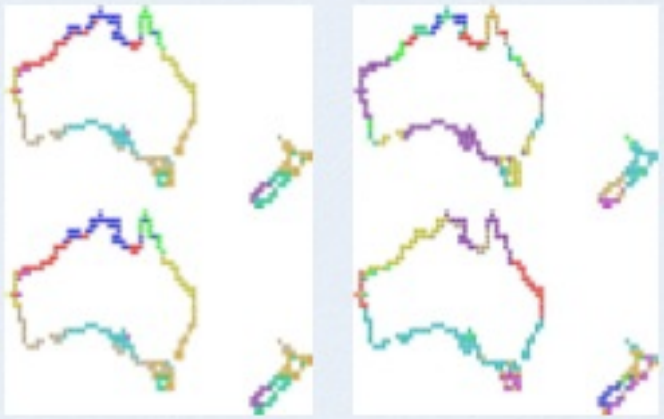
META-LEVEL AGGLOMERATIVE CLUSTERING



Australia

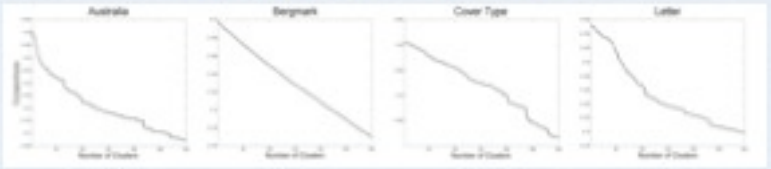
Letter

META-LEVEL AGGLOMERATIVE CLUSTERING

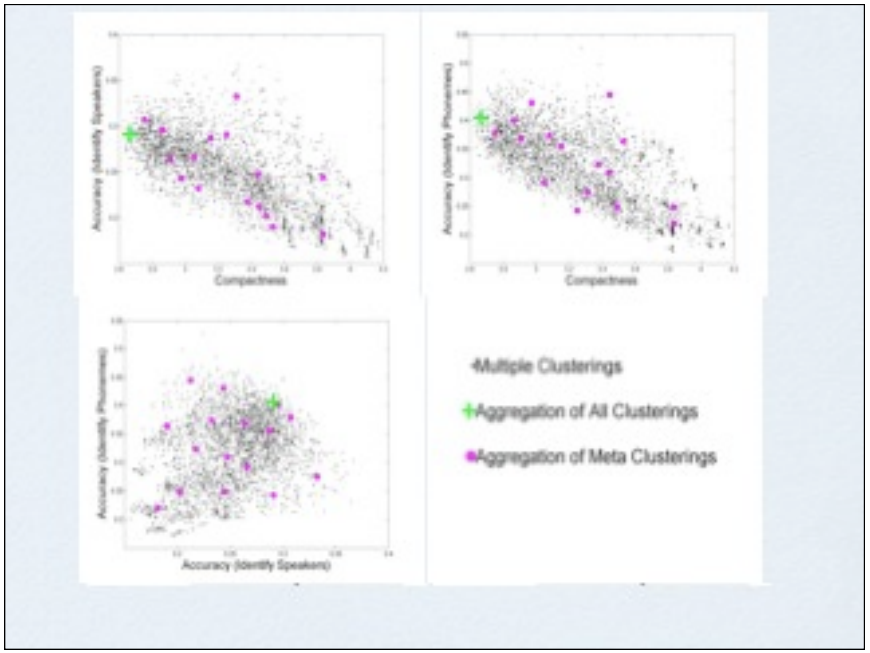


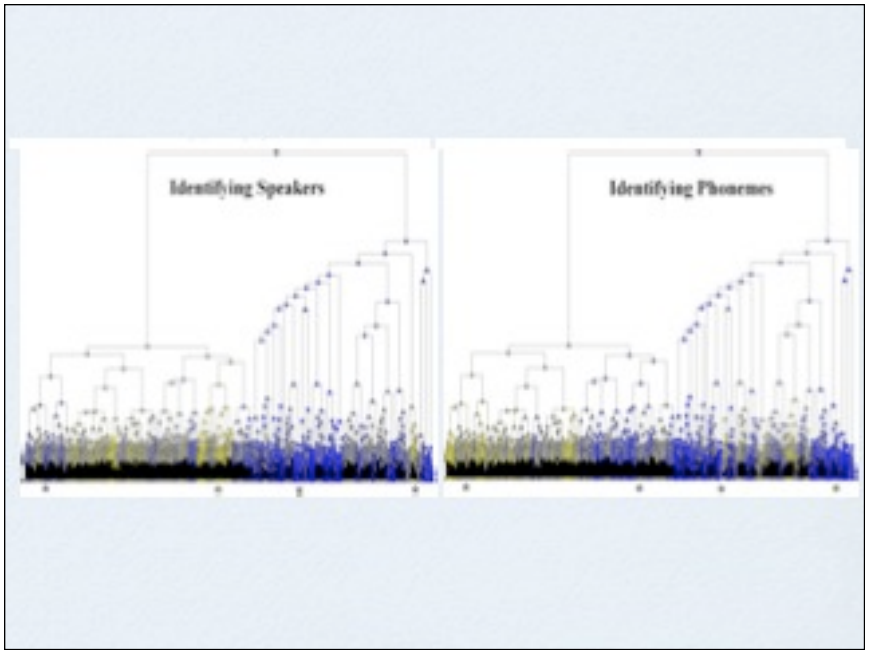
META-LEVEL AGGLOMERATIVE CLUSTERING

Compactness of the hierarchical agglomerative clustering at the meta level



CASE STUDY: PHONEME CLUSTERING





CONCLUSIONS

- Modest correlation between clustering compactness and clustering accuracy
- Searching for a single, optimal clustering may be inappropriate when correct clustering criteria cannot be specified in advance
- Clustering that is good for one criterion may be suboptimal for another criterion
- Different clustering may be needed by different users