Learning Nonstructural Distance Metric by Minimum Cluster Distortions

Daichi Mochihashi *#,
Genichiro Kikui, and Kenji Kita *

* ATR Spoken Language Translation Research Laboratories, Japan
# Graduate School of Information Science, NAIST, Japan

EMNLP 2004

Today's Topic

- We propose a Metric distance function that can be used as an alternative/in conjunction with tf.idf.
- Agenda:
  - How to get an optimal metric?
  - Problems we met with texts.

Overview

- Motivation and Background
- Two distance functions
  - Euclid and generalized Mahalanobis
- Quadratic Minimization Problem
- Recent Related Works
- Experiments
  - Sentence retrieval, Document retrieval
  - General Machine Learning data
- Discussion & Conclusion

Background

- Comparing two linguistic expressions:
  - Structural
    - Tree kernel (Collins and Duffy 2001), HDAG (Suzuki et al. 2003), …
    - Not all NLP can be kernelized
    - Leaf comparison is still done non-hierarchically
    - Rough but fast search is needed (IR, QA, EBMT)

⇓

  - Non-structural comparison is even necessary
    - But ...

Non-structural comparison

- Comparison in vector space
- Many NLP methods still depend on naïve cosine distance function
  - Information Retrieval
  - Subtopic segmentation (ex.Hearst 94, Choi 00)
    - Method for structural text comparison itself depends on cosine distance between paragraphs!
- Feature weightings and correlations
  - like Polynomial kernel
  - But there aren’t any.

Euclidean distance

$$d(\vec{u},\vec{v}) = (\vec{u}-\vec{v})'(\vec{u}-\vec{v}) = \sum_i (u_i - v_i)^2$$

- Cosine distance is identical to Euclidean (if normalized)
- Problems:
  - Ignores correlation between the features (i.e. dimensions)
  - Ad hoc feature weighting (tf.idf)
    - No theoretical justification w.r.t. distances

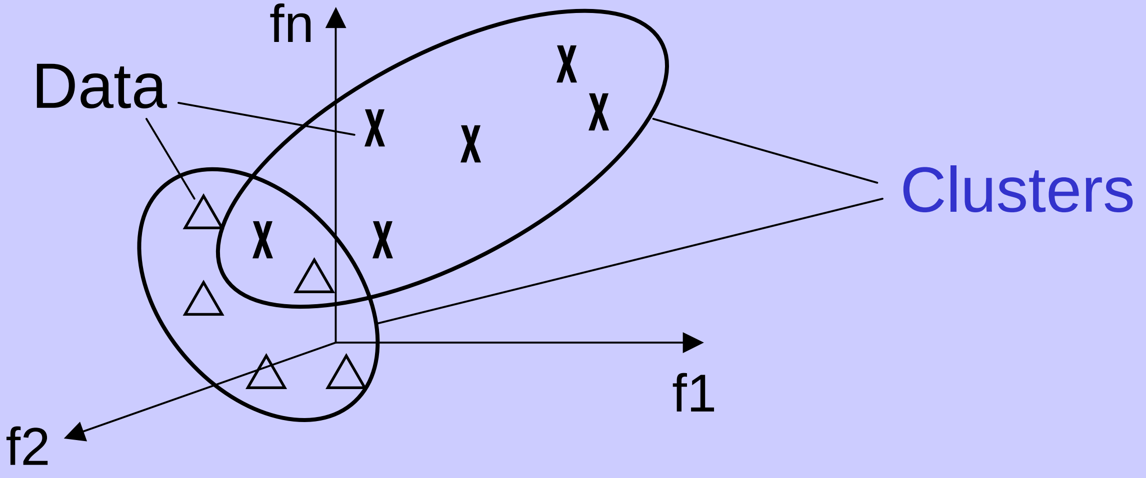
Generalized Mahalanobis distance

$$d_M(\vec{u},\vec{v}) = (\vec{u}-\vec{v})' M (\vec{u}-\vec{v})$$
$$= \sum_i \sum_j m_{ij} (u_i - v_i)(u_j - v_j)$$

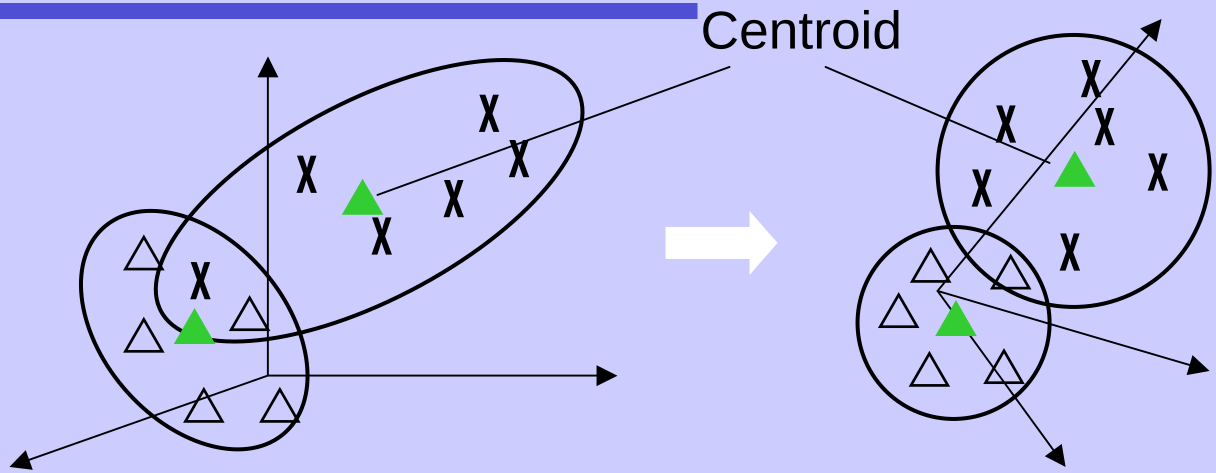
- Simultaneous feature correlation and weighting as m_{ij}
- Famous distance for pattern recognition
  - M is often a covariance matrix of some cluster
  - However, M is arbitrary in general
- What is an optimal M?

Feature space and data

- Training data is not independent in general
- Often, data have a cluster structure
  - Nested linguistic structures
- Usually, cluster doesn't form a true sphere but an ellipsoidal form (feature correlations)



Minimization Problem



- Minimization of within-cluster distances measured by $d_M(\vec{u},\vec{v})$

- Minimization of total sum of distances between cluster data and its centroid

Quadratic Optimization

- For training cluster c in C, when we write centroid of c as $\vec{c}$,

$$\min \sum_{c \in C} \sum_{\vec{x} \in c} d_M(\vec{x}, \vec{c}) = \min \sum_{c \in C} \sum_{\vec{x} \in c} (\vec{x} - \vec{c})' M (\vec{x} - \vec{c})$$

  – Constraint: $|M|=1$ (to avoid $M=0$)
- Solution ➡ Let A the sum of covariance matrix of each cluster,

$$M = |A|^{1/n} A^{-1} \propto A^{-1}$$

  – Proof: by Lagrangian.
    (Extension to Ishikawa98)

Notes

- For linguistic feature vector, *A* is often singular
  - Moore-Penrose pseudoinverse A^{+} as A^{-1}
- Interpreted as Linear projection+Euclid distance

$$d_M(\vec{u},\vec{v}) = (\vec{u}-\vec{v})' M (\vec{u}-\vec{v})$$
$$= (M^{1/2}(\vec{u}-\vec{v}))'(M^{1/2}(\vec{u}-\vec{v}))$$

  - Euclidean distance in $M^{1/2}$-mapped space (optimal geometry)

Related Works

- Xing,Ng,Jordan (NIPS ‘02)
  - Induce *M* from set *S* of “similar” pairs $(\vec{x}_i, \vec{x}_j)$
  - Optimization via Newton-Raphson
  - O(n^2) pairs are required

  → Our method can induce *M* all at once
- Fisher kernel (Jaakkola 98)
  - Same concept in kernel-based method

    $K(x, y) = U(x)' I^{-1} U(y)$ (*I*: Fisher information matrix)
  - Unit matrix approximation in reality

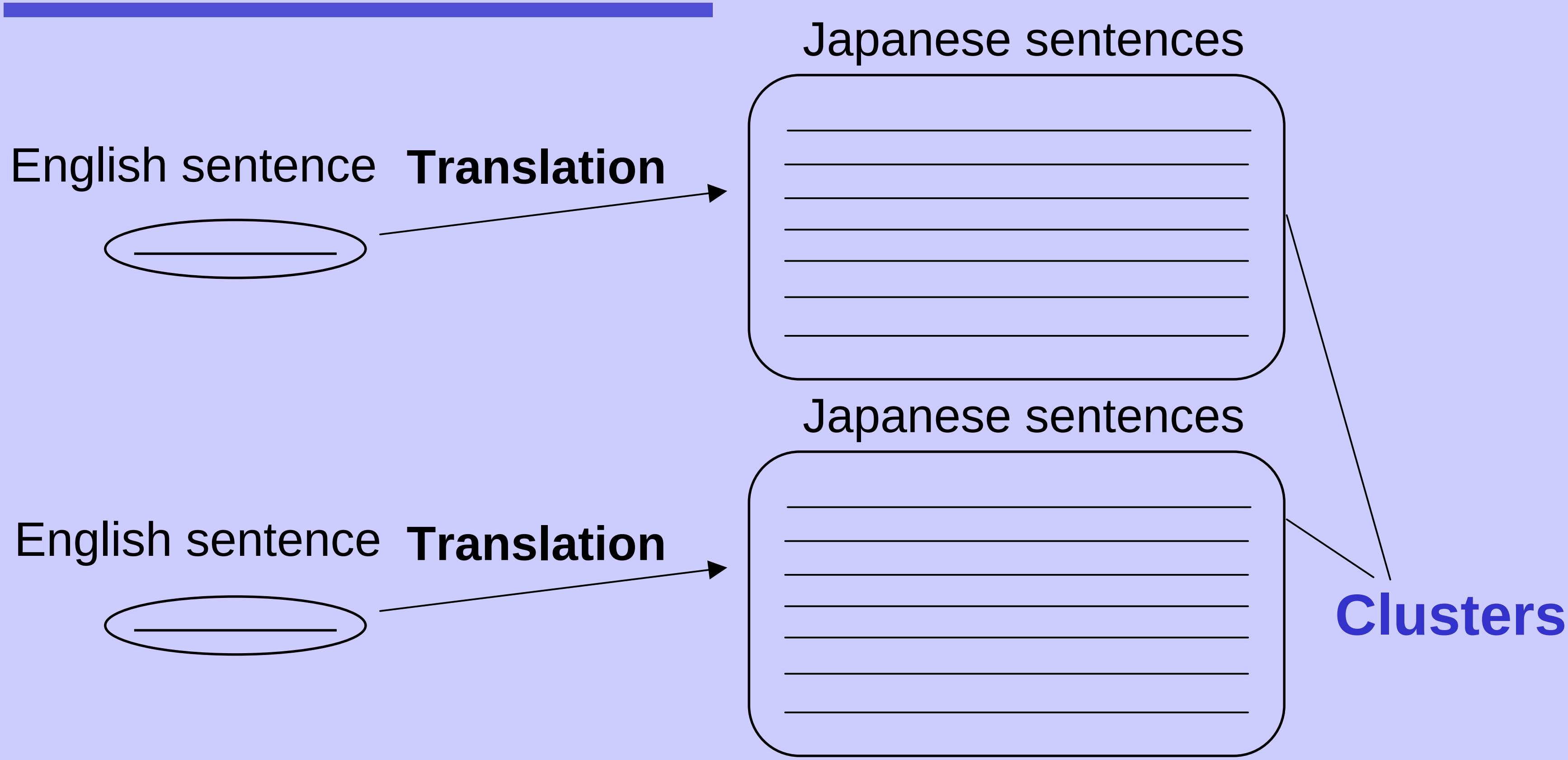
Experiments

- Synonymous sentence retrieval
- Document retrieval
- General vectorial data in Machine Learning

Synonymous sentence retrieval

- ATR paraphrasing corpus (Sugaya et al., 2002)
  - English sentence → multiple Japanese translations
    - English 10,610, Japanese 33,723,164 sentences
  - Possible Japanese translations as a cluster

Paraphrasing as a Cluster



- Possible translations can be regarded as a synonymous cluster.

Synonymous sentence retrieval

- Basic procedure:
  - Calculate a metric matrix from training clusters

    ⬇
  - How well the test data's clusters can be recovered?
- Feature・・ Unigrams, bigrams of function words
  - Large number of features
    → Dimension reduction through SVD (LSI)
  - idf feature weighting as a baseline.

Sentence retrieval result 1

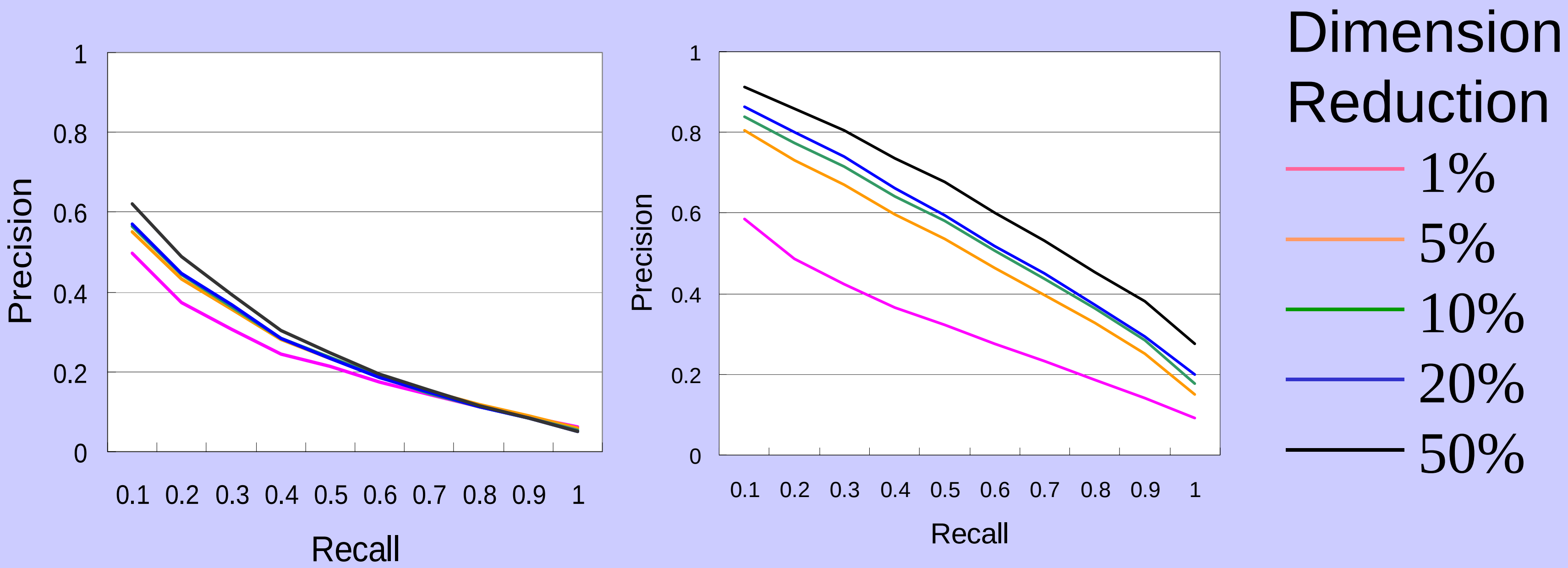
- Query: “How much is total?” (「全部でいくらですか」)

Euclid distance (~cosine)		**Metric distance**	
0.1732	全部でいくらですか	0.2712	合計でいくらでしょうか
1.781	合計でおいくらですか	0.3444	内金はいくらですか
1.902	紫外線防止ですか	0.3444	入場料はいくらですか
1.966	内金はいくらですか	0.369	手付金はいくらですか
1.966	入場料はいくらですか	0.4377	合計でいくらいたしますか
1.974	手付金はいくらですか	0.4479	合計でいくらいたしますでしょうか
1.983	全部でおいくらですか	0.4505	全部でいくらですか
2.283	どんな兆候ですか	0.4558	合計でいくらになりますか
2.505	どんな症状ですか	0.4602	合計でいくらになりますでしょうか
2.65	お一人ですか	0.4682	合計でいくらになるでしょうか
2.729	放送で呼び出してください	0.4729	合計でいくらしますか
2.749	紫外線防止ですね	0.4851	合計でいくらしますでしょうか

Blue: Correct answer
Red : Wrong answer

Result 2 (sentence retrieval)

- Precision-Recall Curve

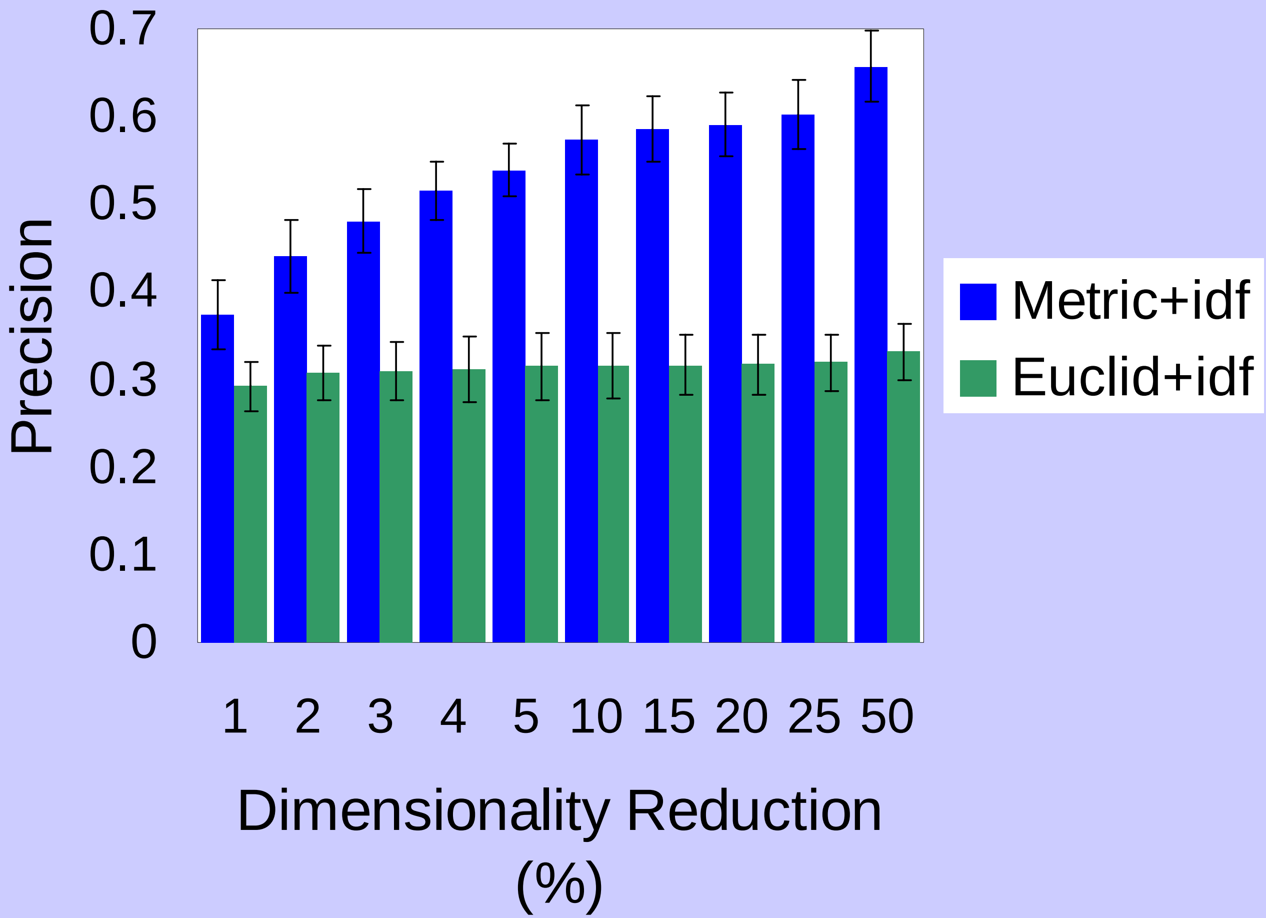


Euclidean+idf   Metric+idf

- 200/50 training/test clusters (100 sents/cluster)
  - 10-fold cross validation

Result 3 (sentence retrieval)

- 11 Point Average Precision



Document retrieval

- 20-Newsgroup dataset (@ai.mit.edu)
  - 5-fold cross validation (16/4 training/test)
  - tf.idf feature weighting as a baseline
- General text・・・・Lengths differ much (vector norm problem)
  - Cannot treat as a cluster in vector space

    ⬇
  - Sub/over sampling to median length (130 words)

    Length normalization (mapping to hypersphere) didn’t work well.

Result (Document retrieval)

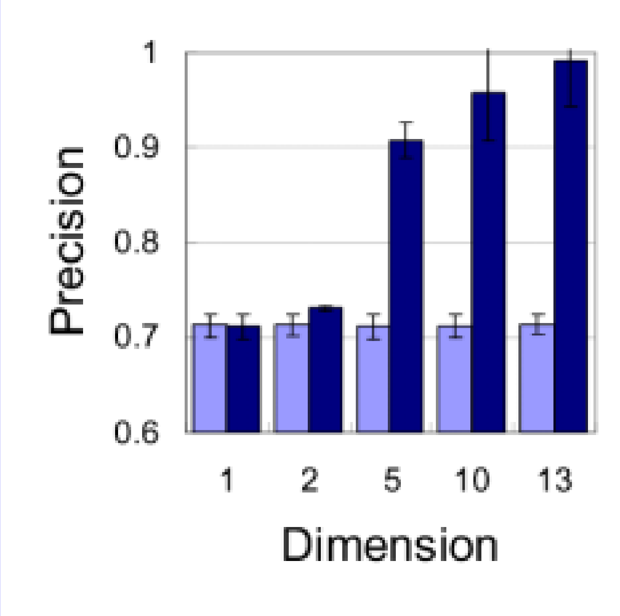
Dim.red.	R-precision		11pt Avr. Prec.	
1%	**0.388**	0.368	**0.450**	0.430
2%	**0.359**	0.343	**0.425**	0.409
5%	**0.329**	0.318	**0.397**	0.388
10%	**0.316**	0.307	**0.379**	0.376
20%	**0.343**	0.297	**0.397**	0.365

- Data from Yahoo.com web directory (http://dir.yahoo.com/*/) has the same tendency

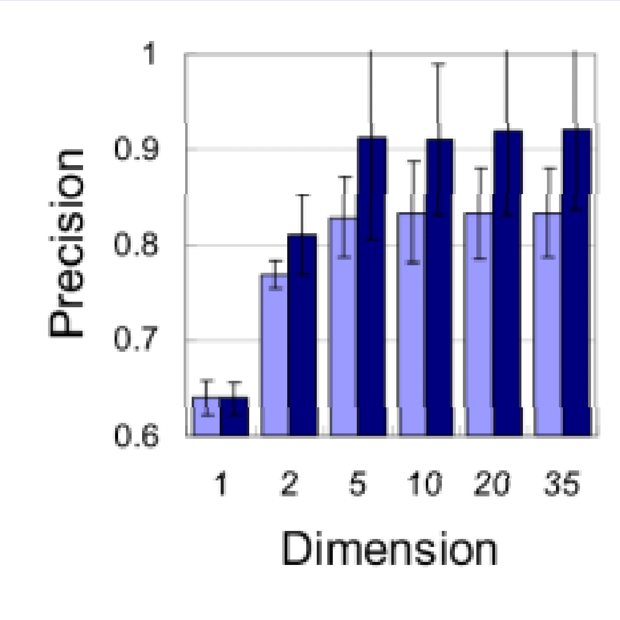
Analysis (Document retrieval)

- Only slight increase in precision
  → Dimensionality reduction
- Dim. red. by SVD: $X = USV^{-1} \rightarrow X_k = V_k X$
  $$\Rightarrow M^{1/2} X_k = M^{1/2} \cdot V_k X$$
  – Dimensionality reduction V subsumes M because of diffuse clusters
- Simultaneous metric induction and dim. red.
- Effective when clusters are tight or dim. red. is unnecessary (sentence retrieval, general data)

UCI Machine Learning datasets



“wine” dataset



“protein” dataset

- K-means clustering x 100, # of clusters known
- Precision of clustering is apparently higher.

Discussion

- Of course, our criterion is one of the possibilities
- Latest NIPS'03 saw two related works:
  - Spectral clustering setting (Bach&Jordan 04)
  - Relative comparison data with SVM (Schultz&Joachims 04)

⬇

- "Minimum distortion" concept in kernel Hilbert space?

Conclusion

- Introduced an optimal metric distance in vector space using training clusters
  - Result of quadratic minimization problem
- Simpler and faster induction than previous work and intuitive result
- Validated by sentence retrieval, document retrieval, and general vectorial data
  - Simultaneous induction of metric and dim. red. may be necessary for texts
  - Same minimization in kernel Hilbert space?