Data as Ensembles of Records: Representation and Comparison

Nicholas R. Howe

Cornell University

Department of Computer Science

Motivation

- Many ML algorithms assume data are expressed as *feature-value* pairs:

  $$((f_1,v_1),(f_2,v_2),\ldots,(f_n,v_n))$$

- Some data aren't easily expressed in this format.

$\Rightarrow$ Need to look at other representations.

Ensembles of Records

- Some data have a collective structure:
  - Documents as collections of words.
  - Accounts as collections of transactions.
  - Episodes as collections of events.

$\Rightarrow$ *Ensembles* as collections of *records*.

- Number of records per ensemble varies.
- Records have simple description.

Roadmap

- Tools for dealing with ensemble data:
  - Uniform representation.
  - Metric for comparison.
- Application to two domains:
  - Pacific Ocean climate data.
  - Image classification and retrieval.
- Conclusions.

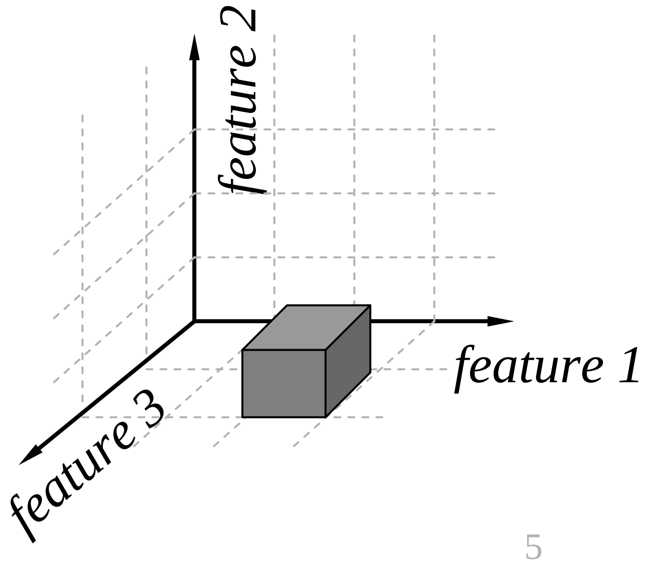
Representation

- Analogy: Document as bag of words ⇔ Ensemble as bag of records

⇒ Represent ensemble by histogram of records.

  – Discretize record features.

  – Each record maps to bin corresponding to its discretized feature values.



An Example

- Accounting domain.
- Each transaction is one record.
- Record is described by the amount and the dates of charge and payment.
- Discretize amount in $50 units, dates by month.
  - e.g., ***$50-99***, ***$100-149***, ***Jun99***, ***Sep00***, etc.

Account History		
Amount	Charge Date	Paid Date
$75.00	06/16/99	10/13/99
$20.00	09/02/99	10/13/99
$35.00	09/28/99	10/13/99

Example (cont.)

Account History		
Amount	Charge Date	Paid Date
$75.00	06/16/99	10/13/99
$20.00	09/02/99	10/13/99
$35.00	09/28/99	10/13/99

= ***($50-99,Jun99,Oct99)***

= ***($0-49,Sep99,Oct99)***

= ***($0-49,Sep99,Oct99)***

Vector representation: (...,0,0,1,0,...,0,2,0,0,...)

($50-99, Jun99, Oct99)

($0-49, Sep99, Oct99)

($50-99,Sep99, Oct99)

Comparison

- Compare ensembles using cosine metric:

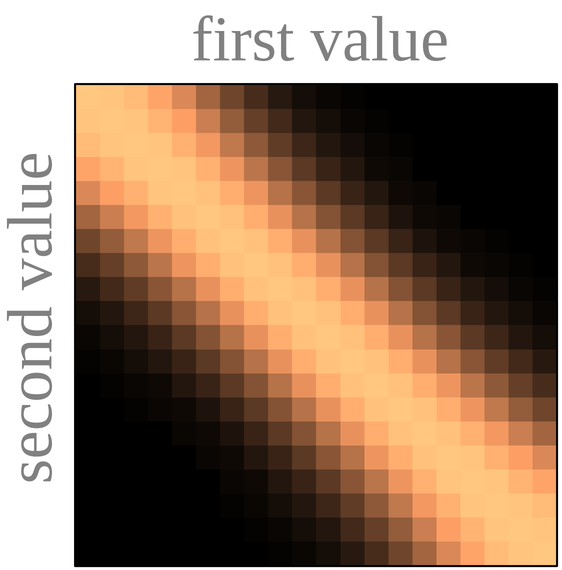
$$D(\mathbf{f_1},\mathbf{f_2}) = \cos^{-1}\left(\frac{\mathbf{f_1}^{\mathbf{T}}\mathbf{S}\mathbf{f_2}}{(\mathbf{f_1}^{\mathbf{T}}\mathbf{S}\mathbf{f}_1)(\mathbf{f_2}^{\mathbf{T}}\mathbf{S}\mathbf{f_2})}\right)$$

(Recall analogy to documents as bags of words.)

- Note generalization using **S** matrix:
  - **S** = **I** gives standard cosine metric.
  - Other values of **S** allow adjustments to metric.

Comparison: **S** Matrix

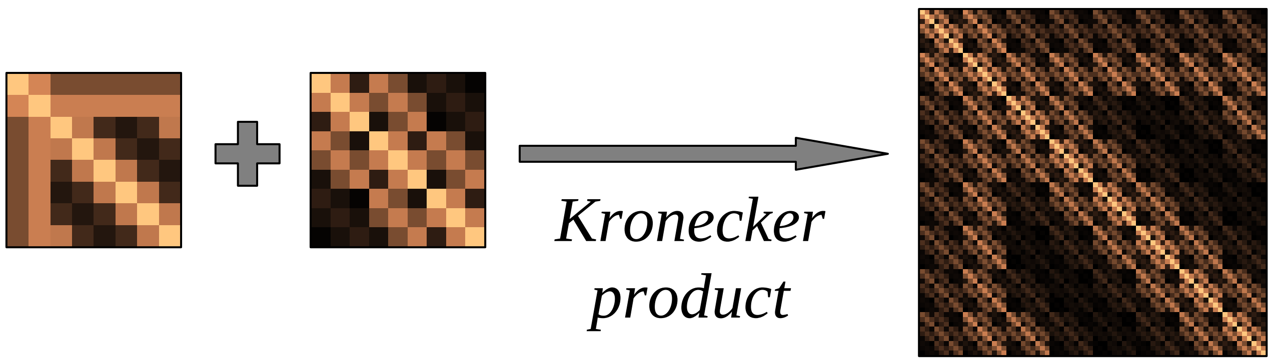
- Discretization of record features may lose order/similarity information.
  - e.g., *\$50-99* is closer to *\$0-49* than *\$950-999*.
- Such relationships may be encoded in off-diagonal terms of **S**.



(lighter ⇒ more similar)

Generating the **S** Matrix

- **S** assembled from feature matrices $\mathbf{S}_j$.

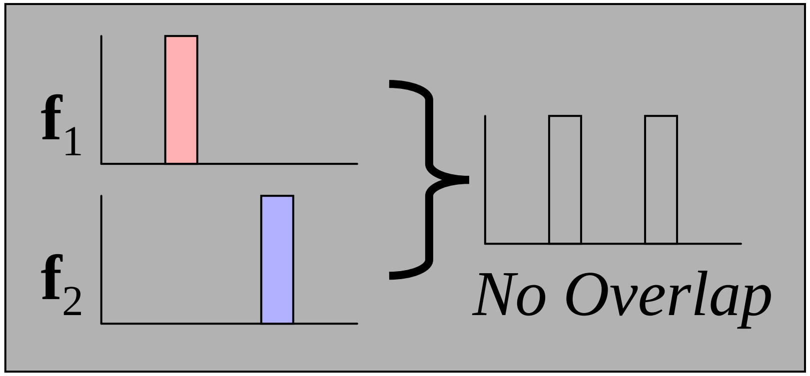


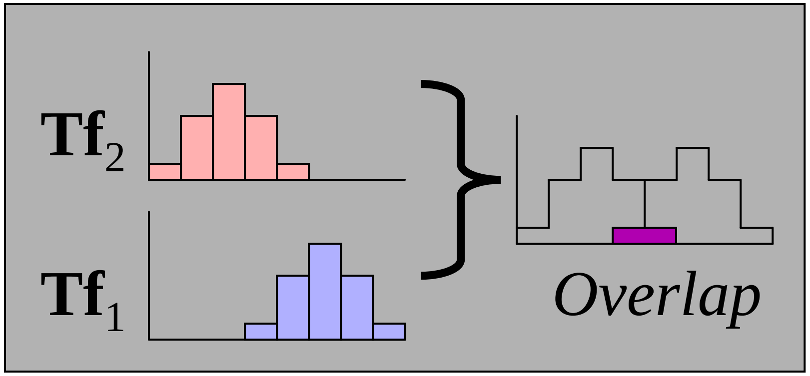
- Terms of $\mathbf{S}_j$ are a function of the distance between bin centers in feature f_j.
  - e.g., Gaussian or exponential decay.

Alternate View of **S** Matrix

- Cholesky factorization of **S**: $\mathbf{S} = \mathbf{T}^{\mathrm{T}}\mathbf{T}$
- Cosine metric of modified vectors:

$$D(\mathbf{f_1},\mathbf{f_2}) = \cos^{-1}\left(\frac{(\mathbf{Tf}_1)^{\mathrm{T}}(\mathbf{Tf}_2)}{\sqrt{\left((\mathbf{Tf}_1)^{\mathrm{T}}(\mathbf{Tf}_1)\right)\left((\mathbf{Tf}_2)^{\mathrm{T}}(\mathbf{Tf}_2)\right)}}\right)$$

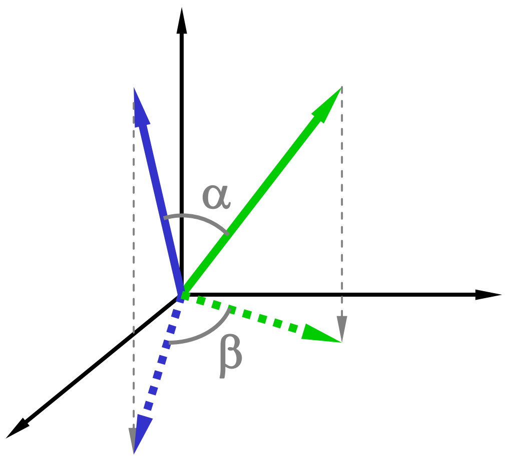




Optimizations

$$D(\mathbf{f}_1,\mathbf{f}_2) = \cos^{-1}\left(\frac{\mathbf{f}_1^{\mathrm{T}}\mathbf{S}\mathbf{f}_2}{(\mathbf{f}_1^{\mathrm{T}}\mathbf{S}\mathbf{f}_1)(\mathbf{f}_2^{\mathrm{T}}\mathbf{S}\mathbf{f}_2)}\right)$$

- Structure of **S** makes $\mathbf{f}_1^{\mathrm{T}}\mathbf{S}\mathbf{f}_2$ calculation fast.
  - (Order $n_1 n_2$, where n_1 and n_2 are the number of records that went into $\mathbf{f}_1$ and $\mathbf{f}_2$.)
  - $\mathbf{f}_1^{\mathrm{T}}\mathbf{S}\mathbf{f}_1$ and $\mathbf{f}_2^{\mathrm{T}}\mathbf{S}\mathbf{f}_2$ can be cached.
- Nearest neighbor search can be pruned by projection onto lower-dimensional spaces.



β is lower bound on α.

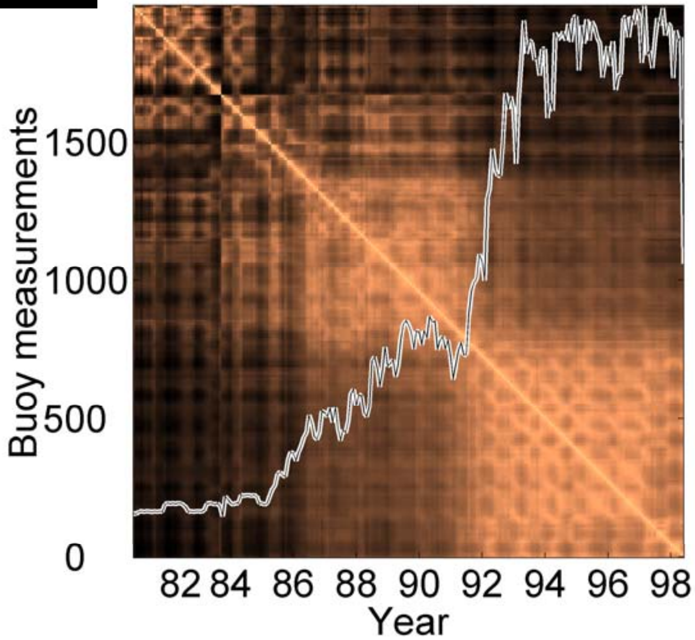
Experiments

- Pacific Ocean buoy measurements:
  - Data from NOAA meteorological buoys. (Available from UCI KDD repository.)
  - Contains four El Nino episodes.
- Image classification experiments:
  - Images from Corel stock photo collection.
  - Two sets of visually-similar categories.
- Many other data sets are proprietary.

Pacific Ocean Data

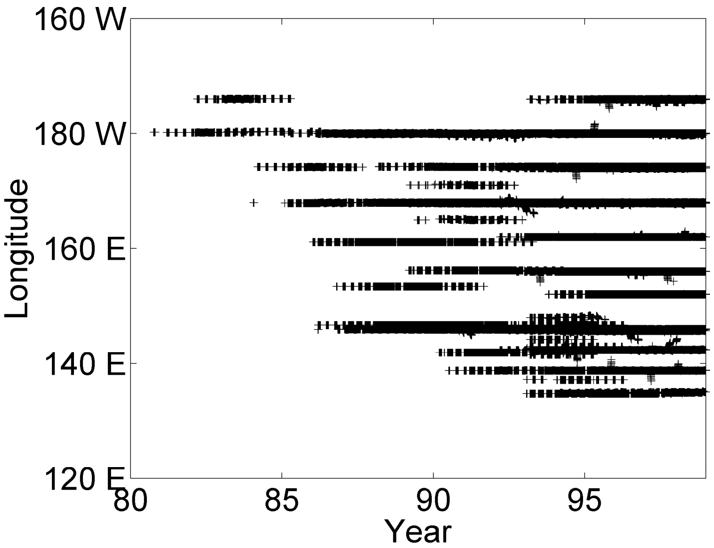
- Data from March 1980 to June 1998.
  - Some missing data.
- Features and discretization:
  - Longitude, 5 bins.
  - Zonal & meridional winds, 7 bins each.
  - Humidity, 11 bins.
  - Air & sea temperatures, 15 bins each.
  - $\Rightarrow$ Total dimensionality: 983,040 bins.
- Ensemble = aggregated measurements over one-month intervals.

First Pass Results



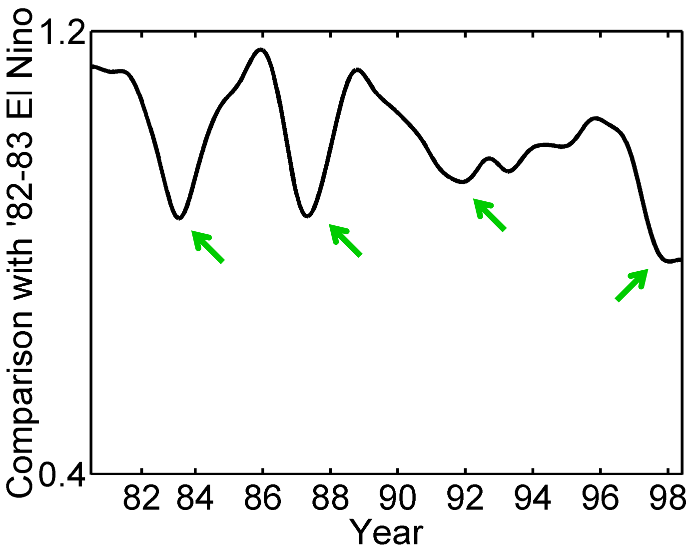
Strongest trend is a surprising dependence on measurement date.

Reflects pattern of buoy addition over time:

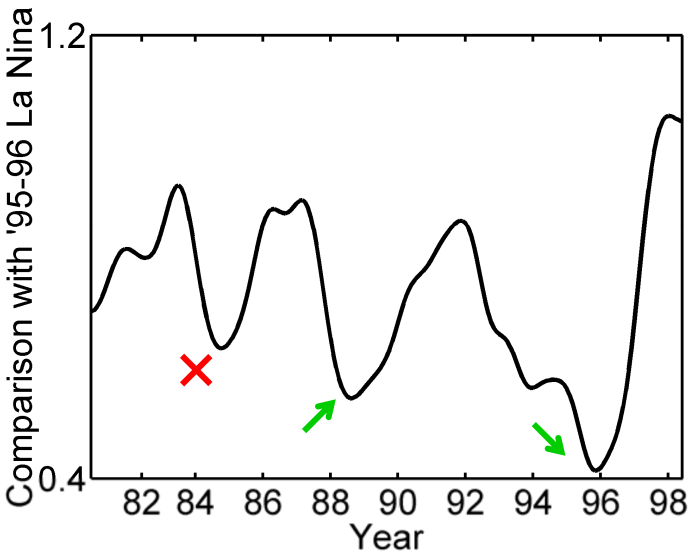


Ocean Data: Final Results

- After accounting for buoy addition, we can detect El Nino and La Nina events.



El Ninos: ’83, ’87, ’92, ’98.



La Ninas: ’89, ’96. (Not ’85).

Image Classification

- Two sets of test images:
  - 12 and 16 categories of ~100 images each.
- Features and discretization:
  - Color, 28 bins.
  - Texture (mean gradient), 3 bins.
  - Location, 25 bins.
  - Regional similarity, 4 bins.

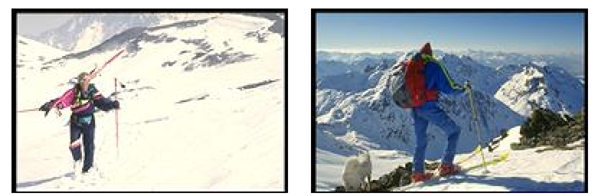
$\Rightarrow$ Total dimensionality: 8400 bins.

Sample images

Airshows

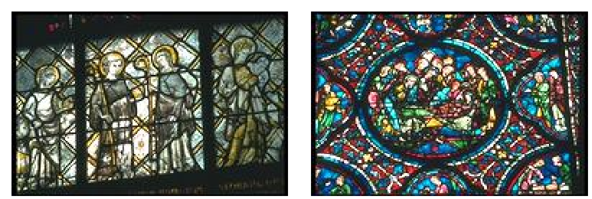
Caves

Elephants



Skiers

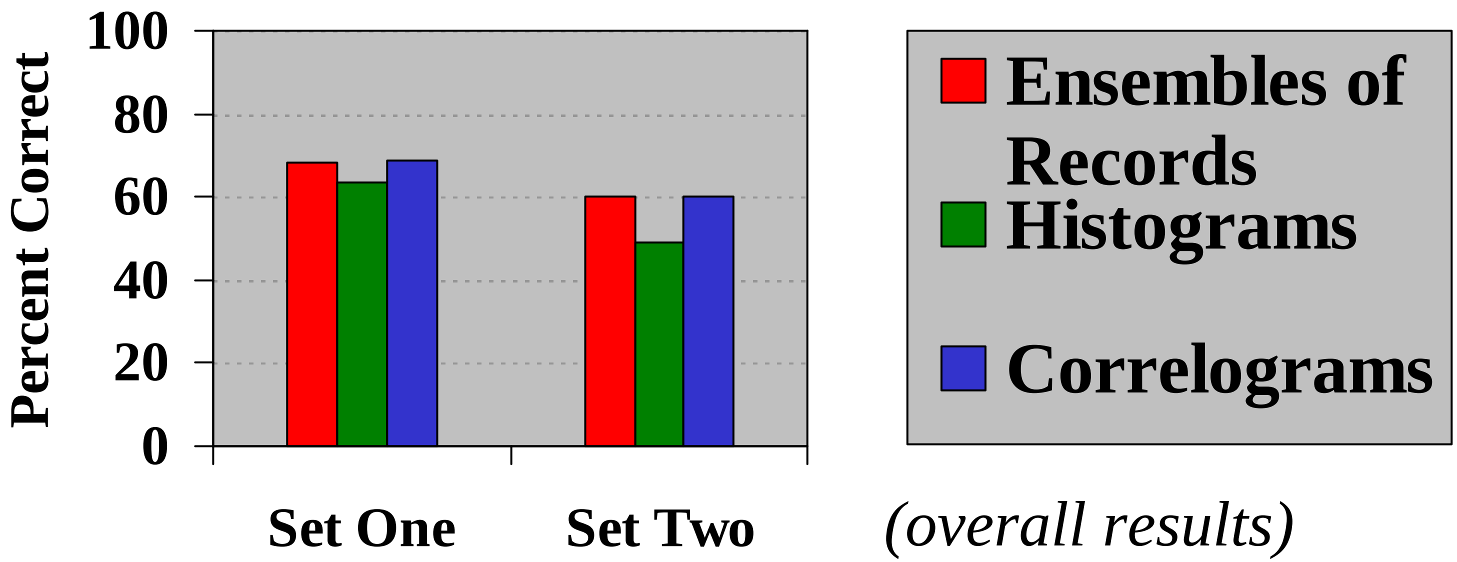
Polar Bears



Stained Glass

Classification Results

Comparison with two specialized image metrics:



(overall results)

- Outperforms baseline (green).
- Competitive with advanced image metric (blue).

Summary

- Developed representation and metric for one nonstandard data format.
- Demonstrated use of these tools on two domains.
  - Results show approach is effective.
  - Competitive with specialized tools in image domain.

Future Work

- Extend to more advanced ML techniques.
  - e.g., boosting.
- Detection of sub-patterns in ensemble data.
- Develop similar approaches for other nonstandard data.