Mining the Logging and Bookeeping Data

X. Zhang, M. Sebag, and C. Germain

October 19,, 2007

Outline

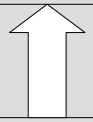
- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

Outline

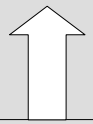
- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

Goals

self-healing (detect, diagnose and repair problems) grid system



modelling the behaviours of grid system



Mining the clusters of Logging and Bookkeeping (L&B) files

Goals

- Object: jobs submitted to grids
- Data: job traces from EGEE broker
- Short Goals:
  - characterize the jobs distribution
  - specify their failure modes

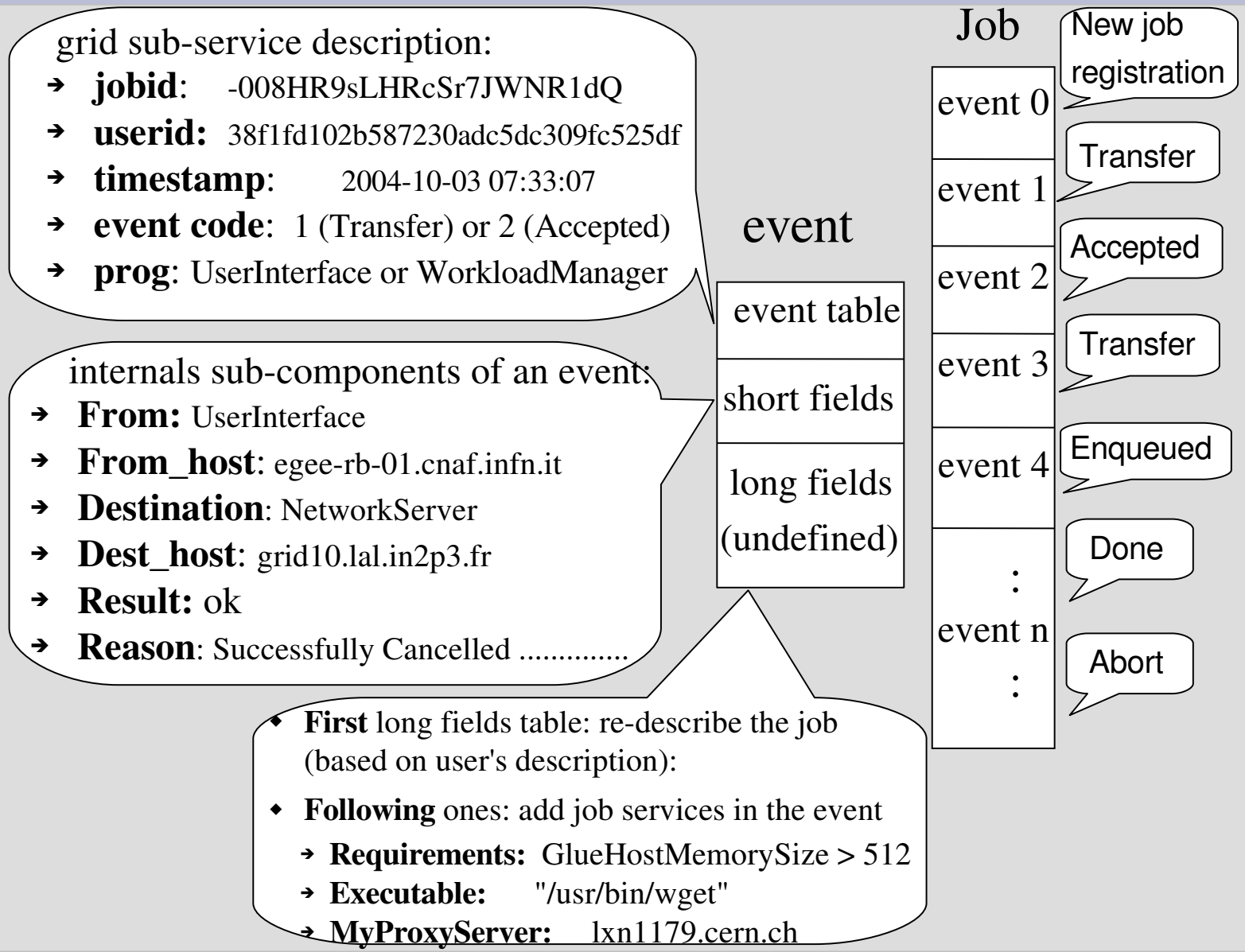
about 70% jobs failed for various reasons

Outline

- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

Data Sampling

EGEE L&B Data Structure

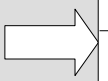


Data Sampling

Initial representation

- Job -------> numerical vector $\in \mathbb{R}^d$
  - static attributes are chosen
  - numerical attributes: normalized
  - non-numerical attributes ---> boolean attributes

Attr. / jobs	A1	A2
job 1	VA1_1	VA2_1
job 2	VA1_1	VA2_2
job 3	VA1_2	VA2_1
job 4	VA1_1	VA2_2



Attr. / jobs	VA1_1	VA1_2	VA2_1	VA2_2
job 1	1	0	1	0
job 2	1	0	0	1
job 3	0	1	1	0
job 4	1	0	0	1

Data Sampling

Initial representation

- Challenges
  - No natural distance
  - Prior knowledge
    - rough classes
      - successfully finished (good jobs)
      - failed by various reasons (bad jobs): NAR, ABU, GNG
    - heterogeneous
      - users: experience and community are different
      - weeks: load of the grid varies along time

Sampling Training Set

- Training Set (90% of all: 222, 500 jobs. 36% good and 73% bad)
  - Homogeneous subsets
    - User subsets (34)
      - all jobs submitted by a given user
    - Week subsets (45)
      - all jobs submitted during a given week
- Test Set (remaining 21512 jobs): Kept without changing

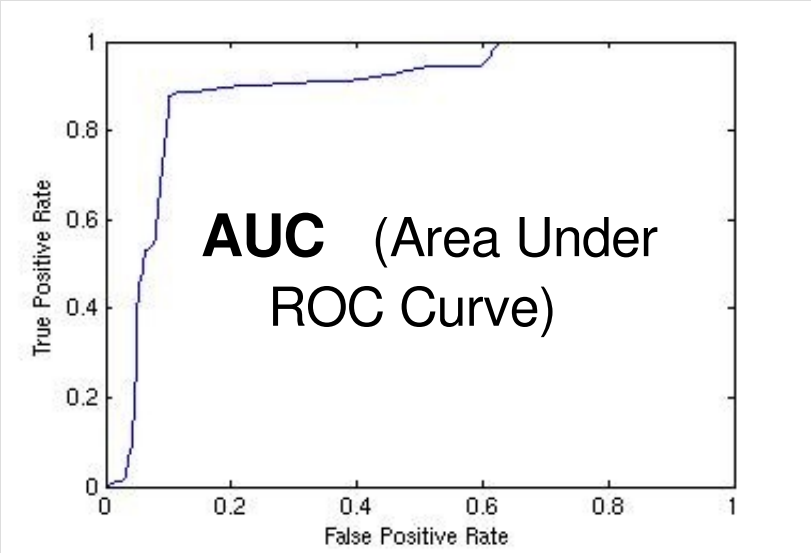
* Kearns M., Li M.: Learning in the Presence of Malicious Errors. SIAM J. Comput. 22 (1993)

Outline

- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

Feature Learning

Using ROGER (ROC-based Genetic Learner)



- Roger: Evolution Strategy algorithm which maximizes the AUC (equivalent to Wilcoxon rank test)
- hypothesis maximizing AUC can be interpreted as a probability estimation

Feature Learning

Using ROGER (ROC-based Genetic Learner)

$h \in \mathbb{R}^d \quad (job_i \rightarrow \mathbb{R} \;\; real\ value)$

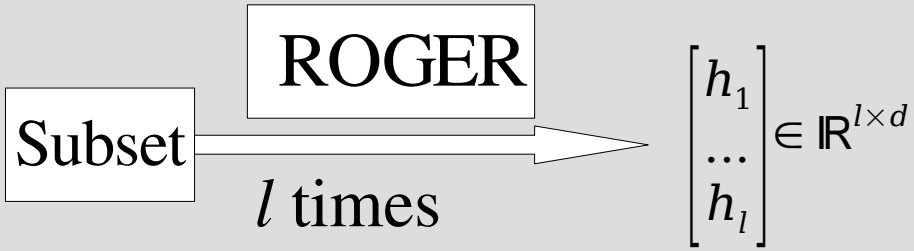
$$\begin{bmatrix} job_1 \; y_1 \\ \dots \\ job_m \; y_m \end{bmatrix} \in \mathbb{R}^{m \times (d+1)}$$

$job_i \in \mathbb{R}^d \quad y_i \in \{+1, -1\}$

- linear hypothesis h
  - provide an estimation of the classification probability
    $Pr(h(job_i) > h(job_j) \mid y_i > y_j)$
  - **as new feature**

Feature Learning

Using ROGER (ROC-based Genetic Learner)



$$\begin{bmatrix} job_1\ y_1 \\ \dots \\ job_m\ y_m \end{bmatrix} \in \mathbb{R}^{m \times (d+1)}$$

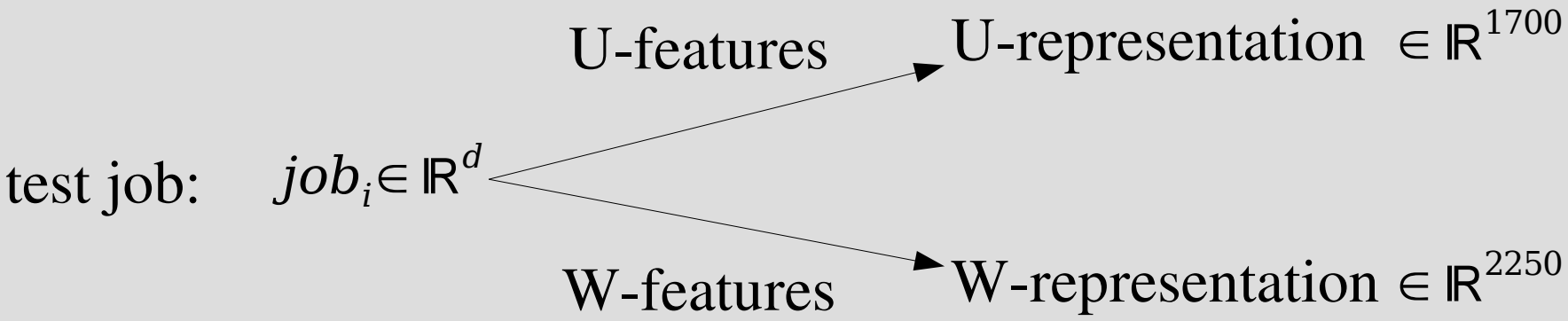
$$job_i \in \mathbb{R}^d \quad y_i \in \{+1, -1\}$$

$$l = 50$$

- Hypotheses learned from User subsets: **U-features** $\in \mathbb{R}^{(34 \times 50) \times d}$
- Hypotheses learned from Week subsets: **W-features** $\in \mathbb{R}^{(45 \times 50) \times d}$

New Representation

- Test Set New Representation



- Feature redundancy
  - from the same subset
  - redundancy of initial attributes

Outline

- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

Double Clustering

* Slonim N., Tishby N. Document clustering using word clusters via the information bottleneck method. Research and Development in Information Retrieval. (2000)

- Information bottleneck method
- double clustering
  - word clusters -----> new representations of documents
  - document clustering on *word-clusters*
- perform excellently
  - clustering by *word-clusters* is **better** than clustering by *words*

Double Clustering

feature clustering
(Dimensionality reduction)

$T << 1700$

All test jobs

U-representation

(W-representation)

$$\begin{bmatrix} job_1 \\ \dots \\ job_i \\ \dots \\ job_m \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{l,i} & \dots & x_{1,1700} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i,1} & \dots & x_{i,i} & \dots & x_{i,1700} \\ \dots & \dots & \dots & \dots & \dots \\ x_{m,1} & \dots & x_{m,i} & \dots & x_{m,1700} \end{bmatrix} \begin{bmatrix} f_{1,1} & \dots & f_{1,T} \\ \dots & \dots & \dots \\ f_{i,1} & \dots & f_{i,T} \\ \dots & \dots & \dots \\ f_{m,1} & \dots & f_{m,T} \end{bmatrix}$$

example clustering

Clusters of test jobs

$$\begin{bmatrix} C_1 \\ \dots \\ C_i \\ \dots \\ C_K \end{bmatrix}$$

$K << m$

Double Clustering

- Clustering method: K-means
- Job clustering results:
  - U-representation: U-clusters
  - W-representation: W-clusters

* Note: U-clusters are not clusters of users
W-clusters are not clusters of weeks

Double Clustering

Clustering Stability

- Clustering is an ill defined problem
  - different clustering tasks leads to different clustering paradigms
- attempts to revisit clustering *,**
- ideas
  - Compare Clustering and PCA
  - Examine the stability of clusters

* Shai Ben-David, Ulrike von Luxburg, John Shawe-Taylor and Naftali Tishby. Theoretical Foundations of Clustering. Workshop NIPS 2005.

** Meila M. The uniqueness of a good optimum for K-means. ICML 2006

Double Clustering

Clustering Stability

- Example:

Data set = {A B C D a b c d}

Case 1:

Clustering *C*: C_1 {A B C D} C_2 {a b c d}

Clustering *C'*: C'_1 {a b c d} C'_2 {A B C D}

Stable

Case 2:

Clustering *C*: C_1 {A B C D} C_2 {a b c d}

Clustering *C'*: C'_1 {A B a b} C'_2 {C D c d}

Non Stable

Clustering Stability

- A clustering C represented by matrix $\hat{C} = \{C_{1,}.., C_K\} \in \mathbb{R}^{m \times K}$

$$\hat{C}_{ik} = \begin{cases} 1/\sqrt{n_k} & \text{if } \textit{the } i^{th} \textit{ example belongs to } C_k \\ 0 & \text{otherwise} \end{cases}$$

where n_k is the size of C_k $\quad \sum_k n_k = m$

- Stability of two clustering ($\hat{c}$ and $\hat{c}'$)

$$S(\hat{C}, \hat{C}') = \|\hat{C}^T \hat{C}'\|^2_{Frobenius} = \sum_{i,j=1}^{K} n_{i,j}^2 \frac{1}{n_i n'_j}$$

where $n_{i,j}$ is the number of jobs in $C_i \cap C'_j$, n_i and n'_j are size of C_i and C'_j

Clustering Stability

- Theorem: bound of $s(\hat{C},\hat{C}')$

$$K \geq S(\hat{C},\hat{C}') \geq \frac{m}{(m-K+1)}\frac{1}{K}$$

when $K << m$, $S(\hat{C},\hat{C}') \rightarrow 1/K$

- Stability index

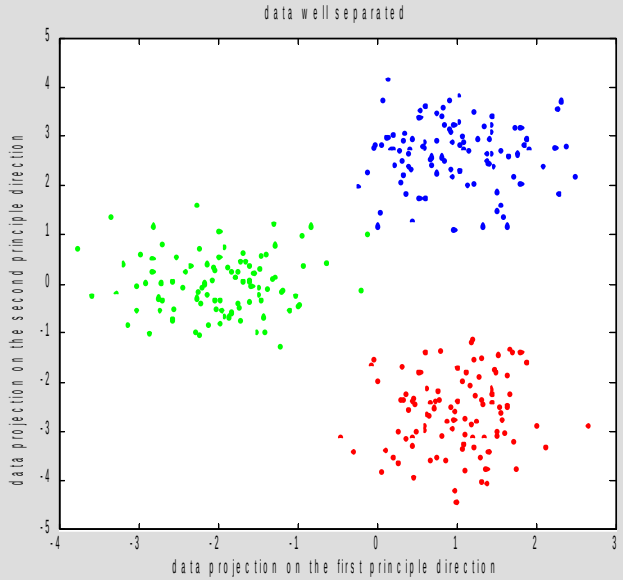
$$D(\hat{C},\hat{C}') = S(\hat{C},\hat{C}')/K$$

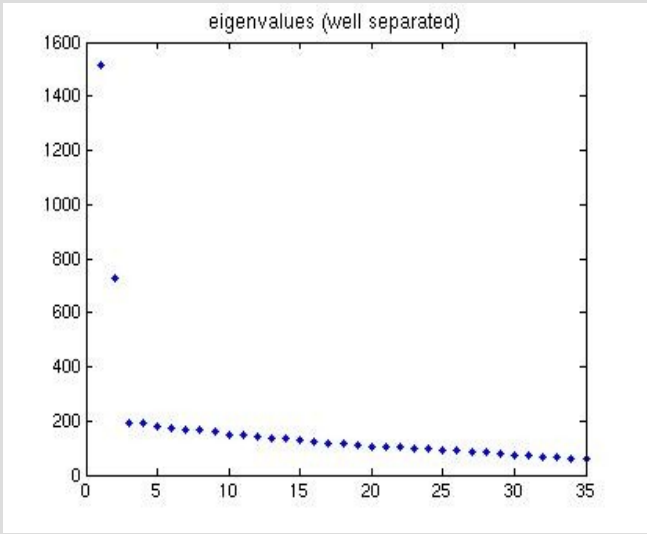
Double Clustering

Assess the quality of clustering

- well-separateness assumption:
  - data do NOT live in a manifold of dimension less than K-1

$$\sigma_{K-1} - \sigma_K >> \sigma_K - \sigma_{K+1} \qquad \sigma_K - \sigma_{K+1} > \sigma_{K+1} - \sigma_{K+2}$$

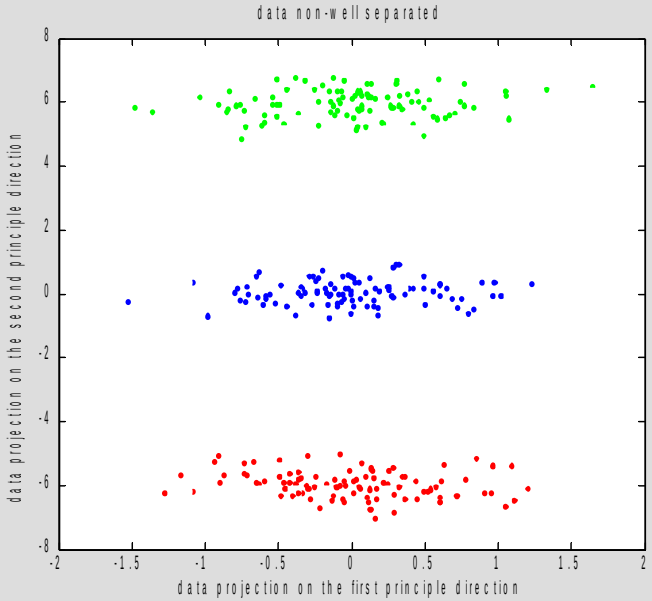


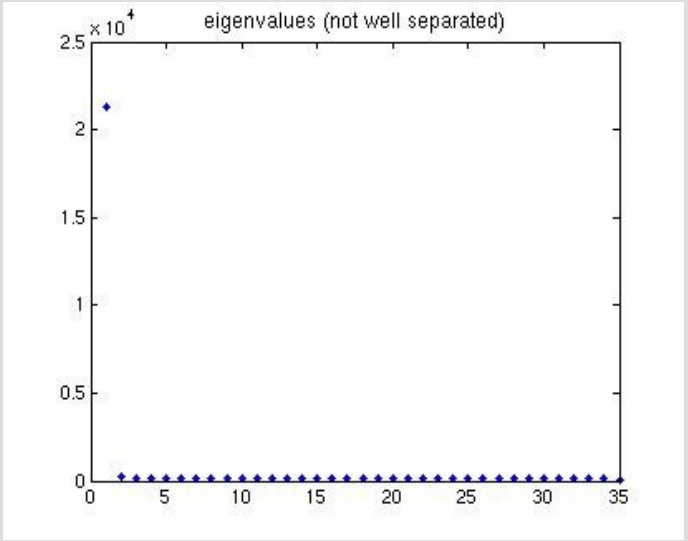


Double Clustering

Assess the quality of clustering

- Not well-separateness assumption:





- Sufficient condition
- Not necessary condition

Double Clustering

Assess the quality of clustering

- good clustering is close to principal components of the data
  - ✔ good clusterings are stable
- measure the distance between clustering and principal components

$$d(C, C^{opt}) \le 2\mathrm{p}_{max}\,\delta(1-\delta/(K-1))$$

where $p_{max} = max\{\frac{n_k}{m}\}$ and $\delta = \frac{D(C) - \sum_{k=K}^{d} \sigma_k}{\sigma_{k-1} - \sigma_k}$

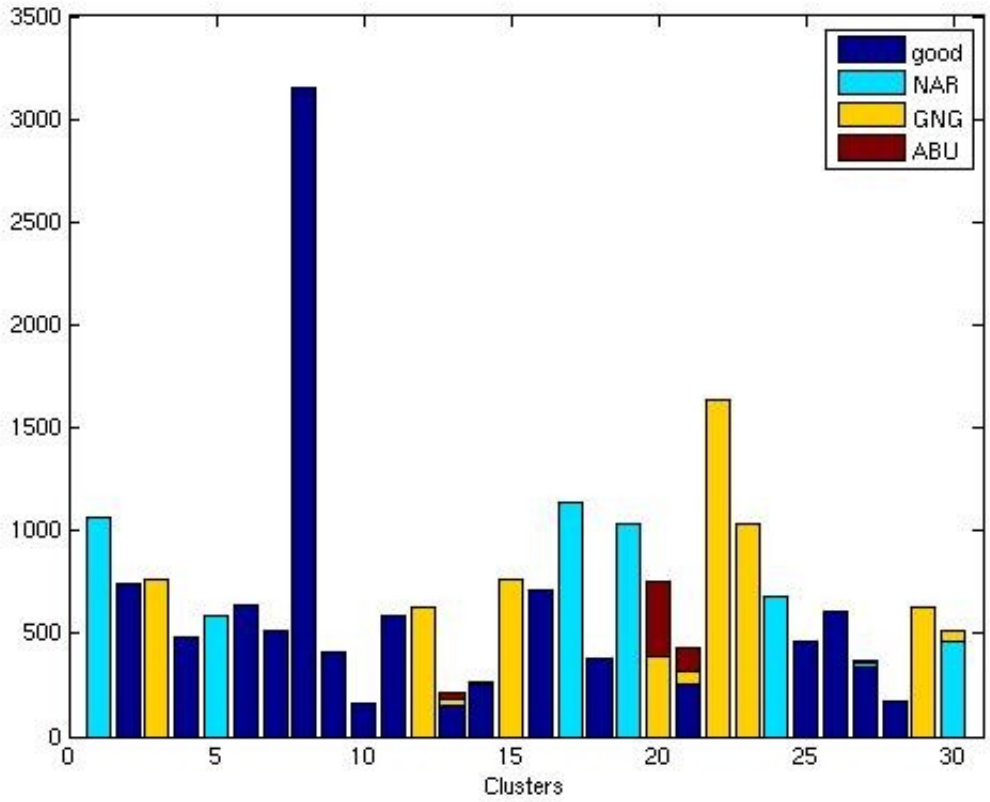
$D(C) = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|^2$ (K-means cost function)

Outline

- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

Results and Interpretations

Clustering Results



- good:

jobs terminated successfully

- NAR:

jobs failed because of
No Adequate Resource

- GNG:

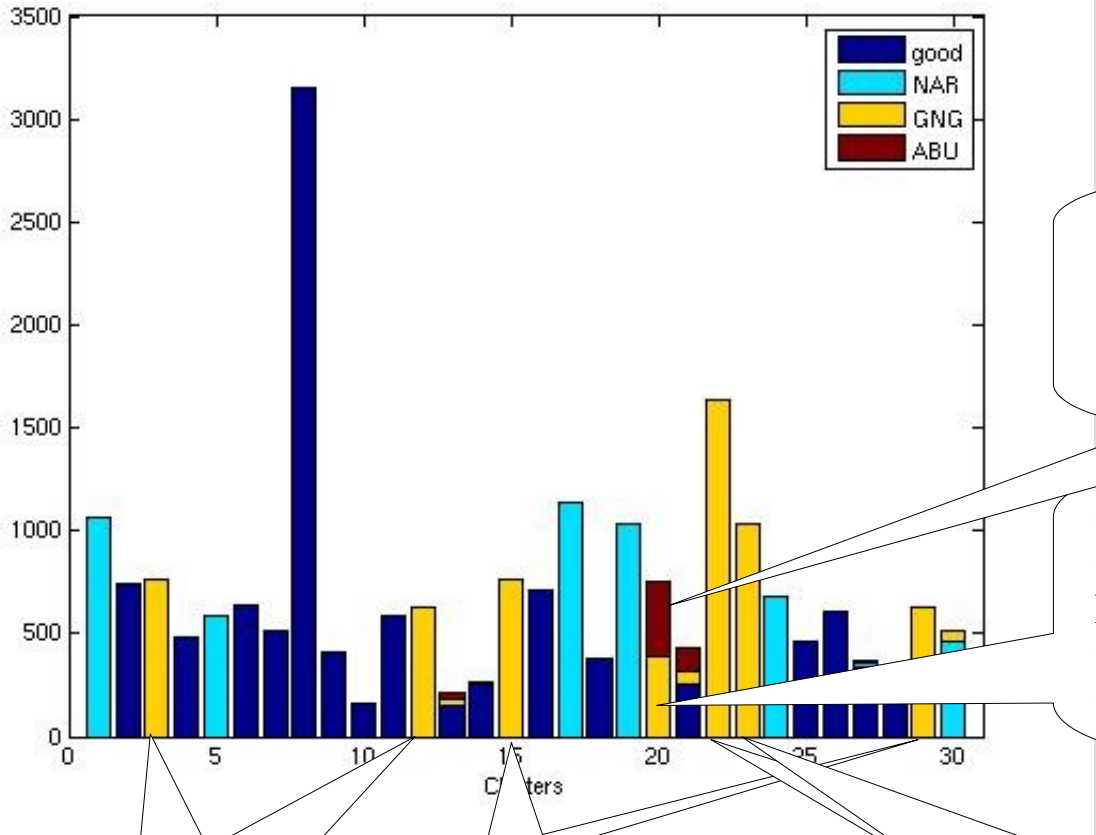
Generic and Non Generic errors

- ABU:

Aborted by Users

Results and Interpretations

Clustering Results



• Canceled by User (No specified reasons)
• unspecified error / cannot download file result in Canceling

• Job proxy is expired
• various reasons result in Job RetryCount (>=1) hit
• cannot receive/read data
• unspecified error

• various reasons result in Job RetryCount (0) hit
• Job proxy is expired

Problems during rank evaluation

• user is not authorized on any resource
• insert Data failed
• Problems during rank evaluation

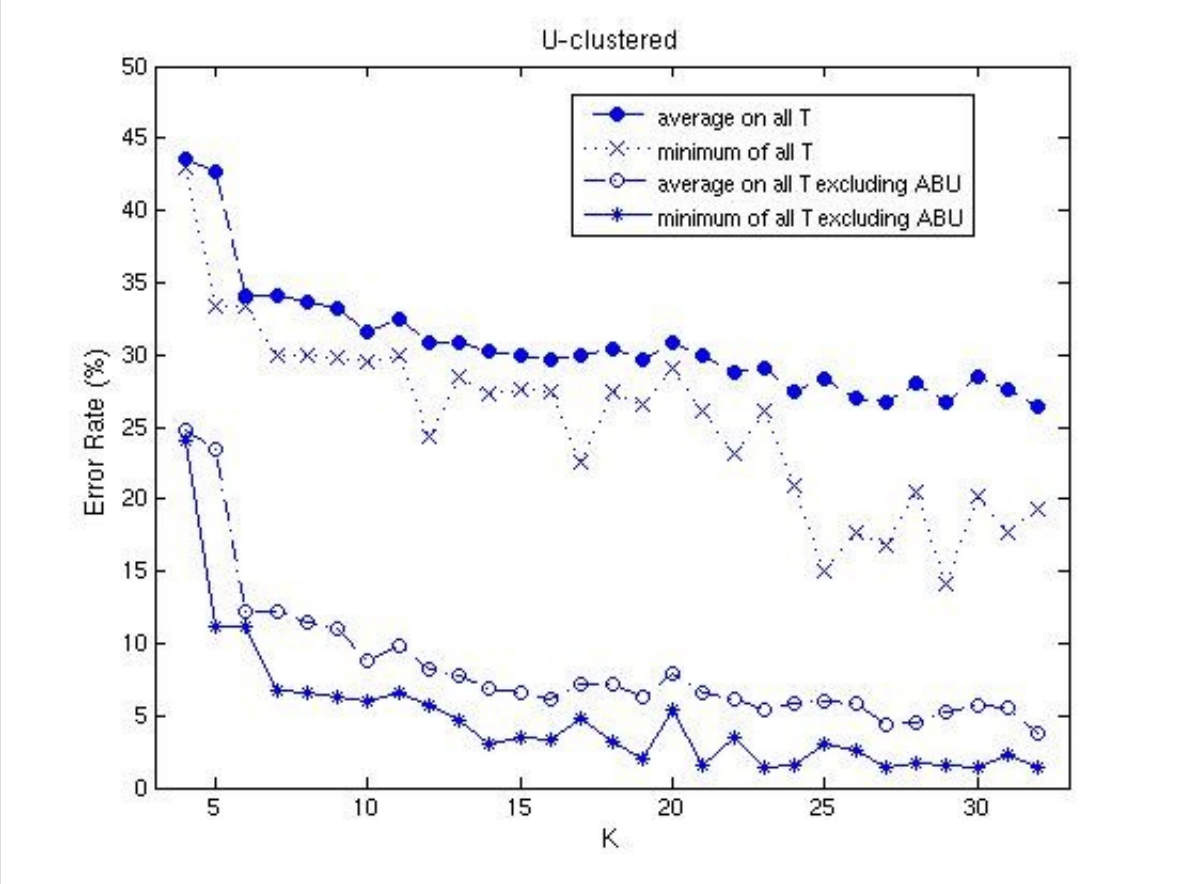
Experimental settings

- the purity of the clusters
  - errors: all jobs which do not belong to the majority class of the clusters they are in.
- Self-stability:
  - Both for W-clusters and U-clusters
  - Compute with same *K*, average on all different pairs of *T*
- Mutual-stability:
  - Between W-clusters and U-clusters
  - for given *K*, average on all pairs of W- and U-clusters with same *T*
  - for given T, average on all pairs of W- and U-clusters with same *K*

Results and Interpretations

Error Rate

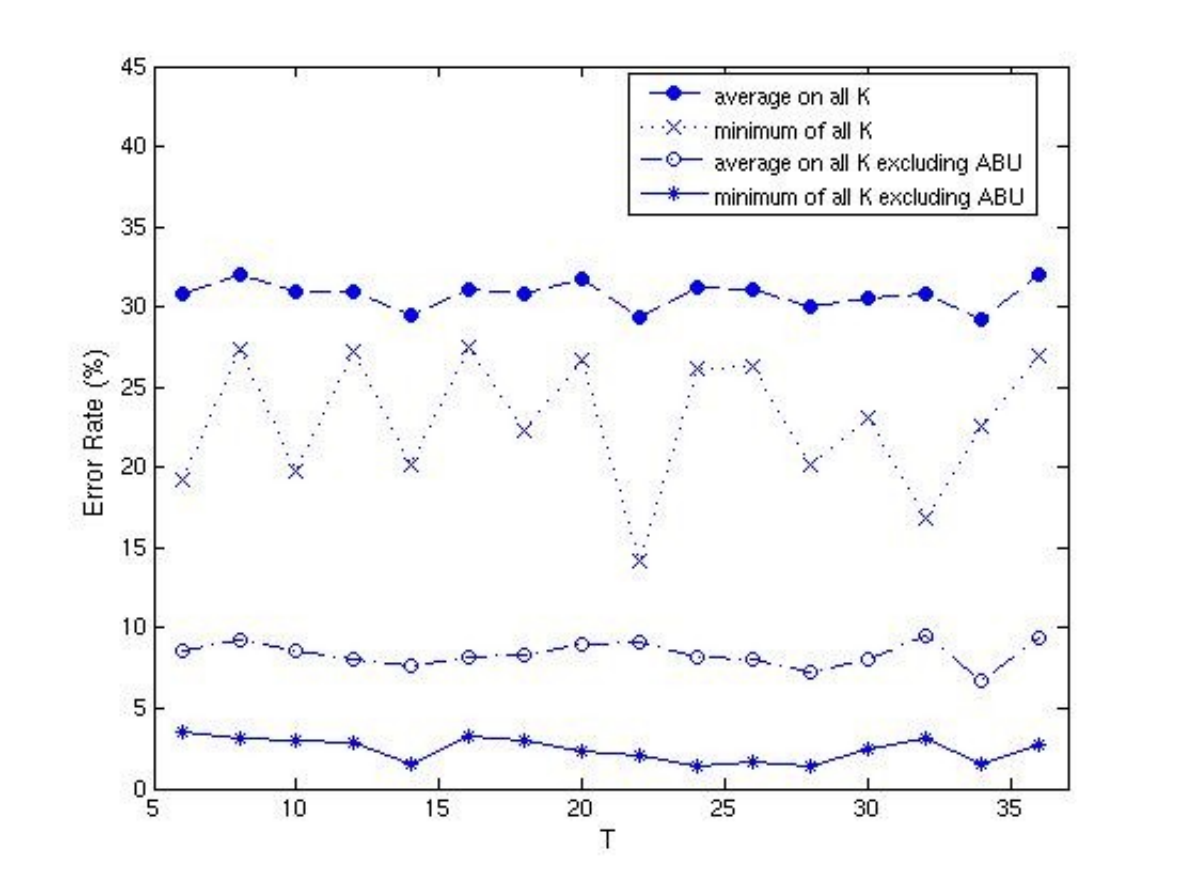
- U-clustered error rate versus *K* (the number of example clusters)



Results and Interpretations

Error Rate

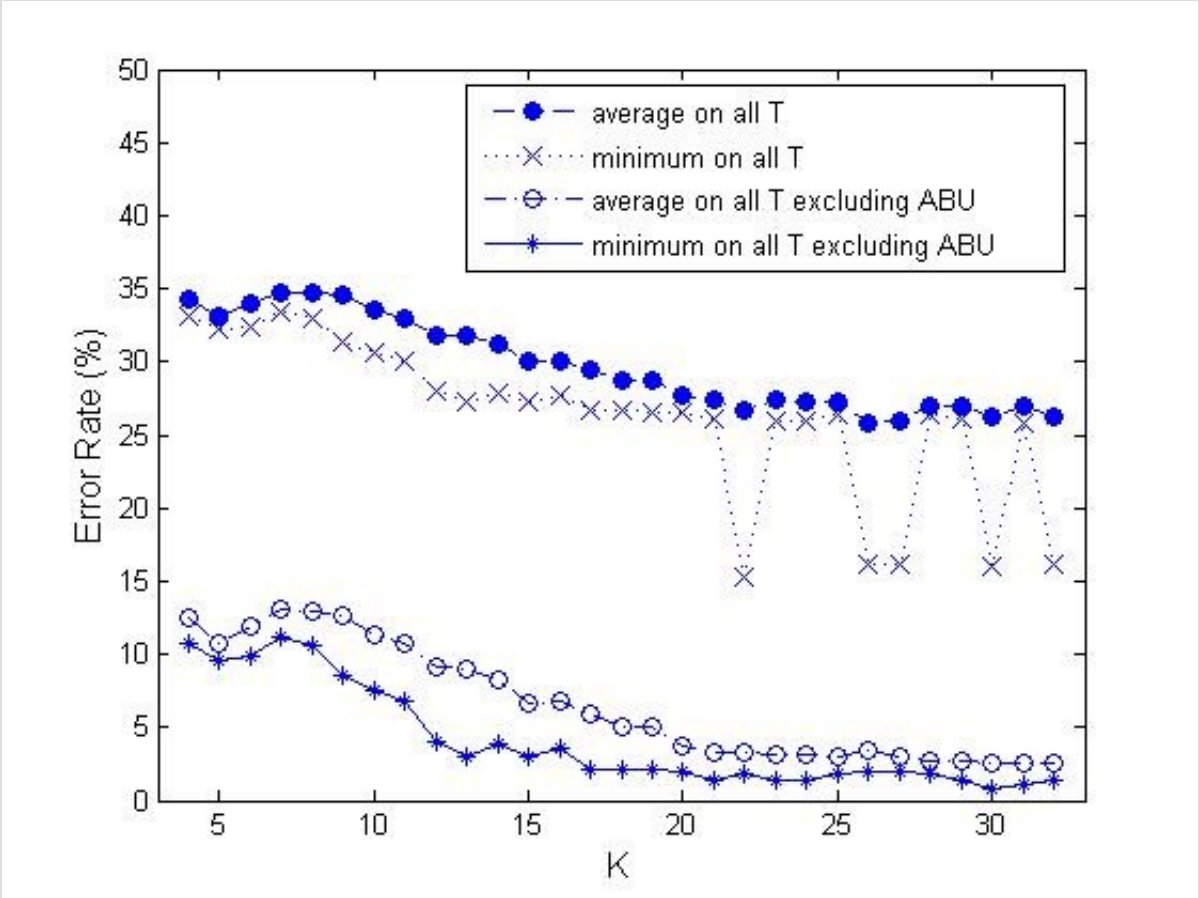
- U-clustered error rate versus *T* (the number of feature clusters)



Results and Interpretations

Error Rate

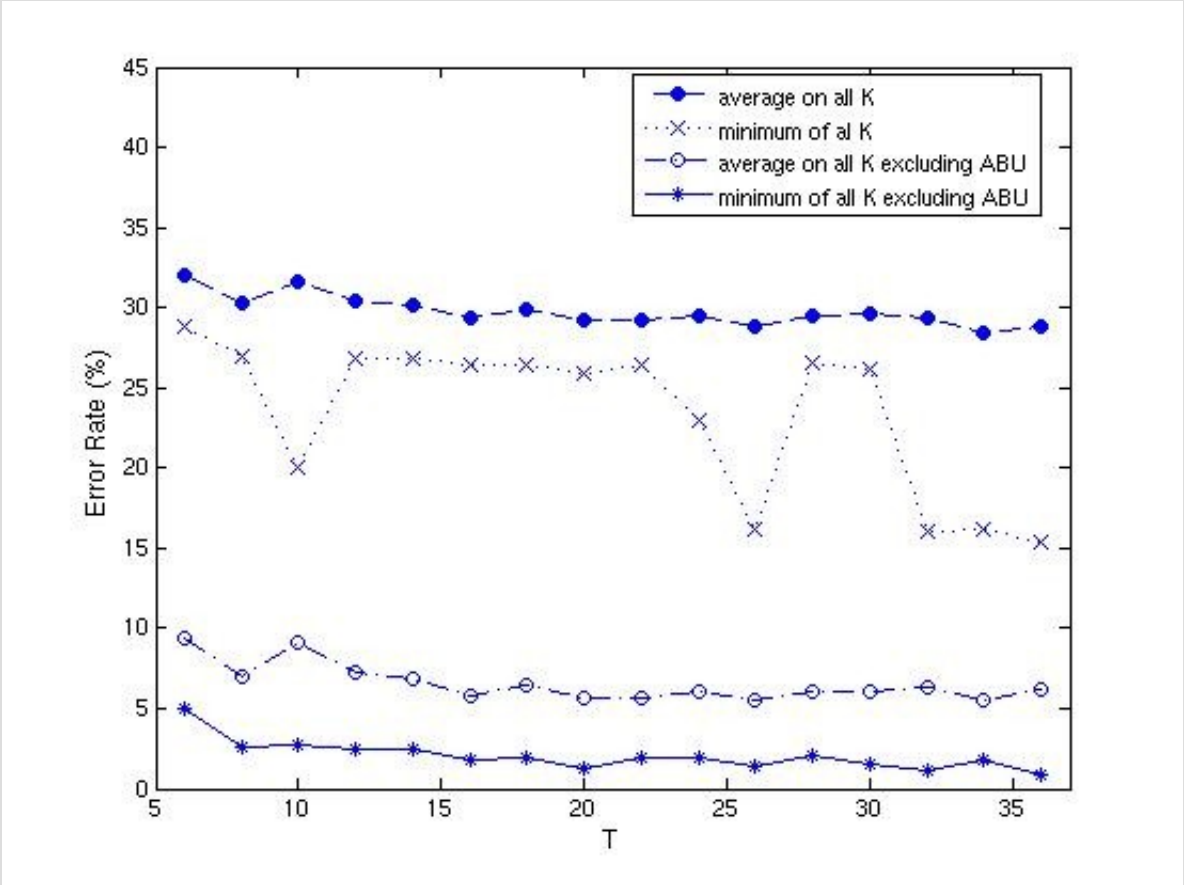
- W-clustered error rate versus *K* (the number of example clusters



Results and Interpretations

Error Rate

- W-clustered error rate versus T (the number of feature clusters)



Results and Interpretations

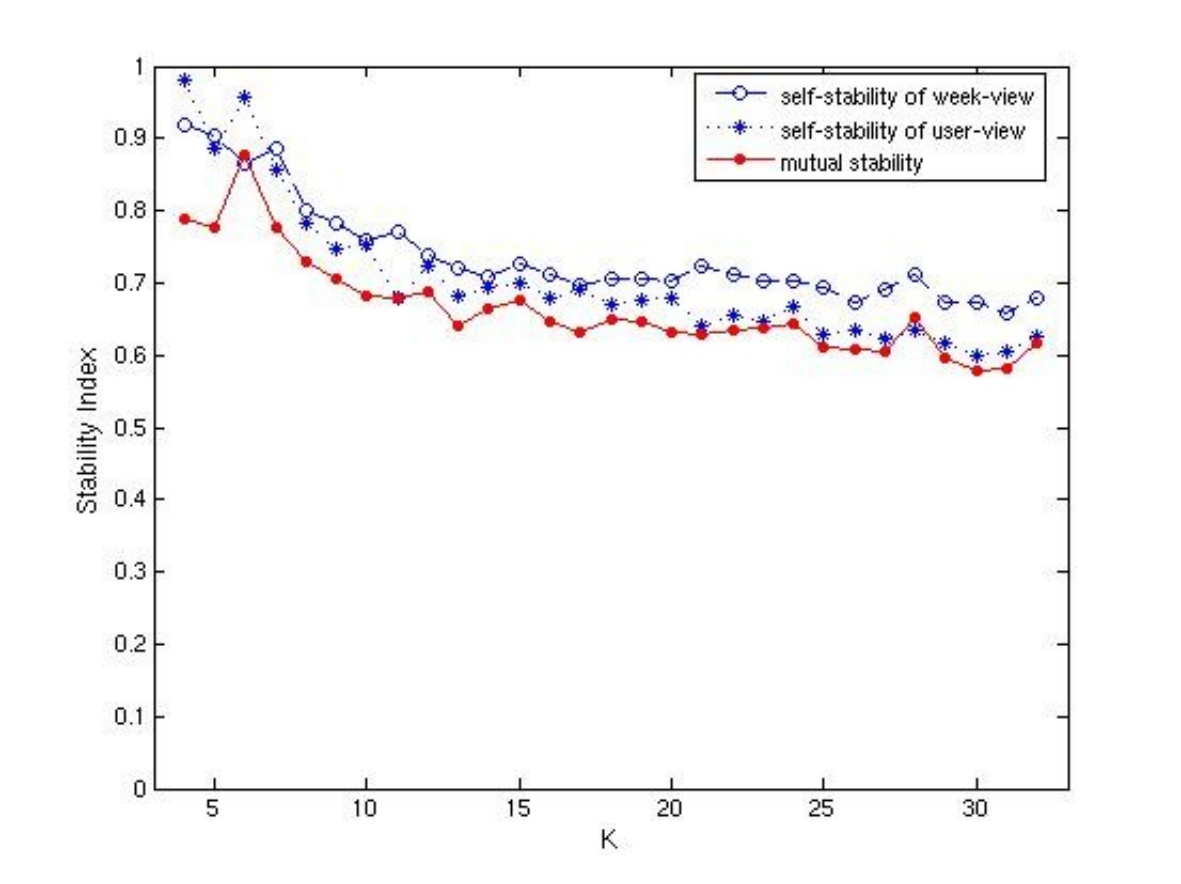
Error Rate

- Summary on Error Rate
  - decrease with *K* (K > 20)
  - ABU is difficult to classify
  - not depend much on *T*
    - feature clustering (dimensionality reduction) has no impact on clustering results
  - better performance on ABU when K and T are chosen in agreement with each other

Results and Interpretations

Clustering Stability

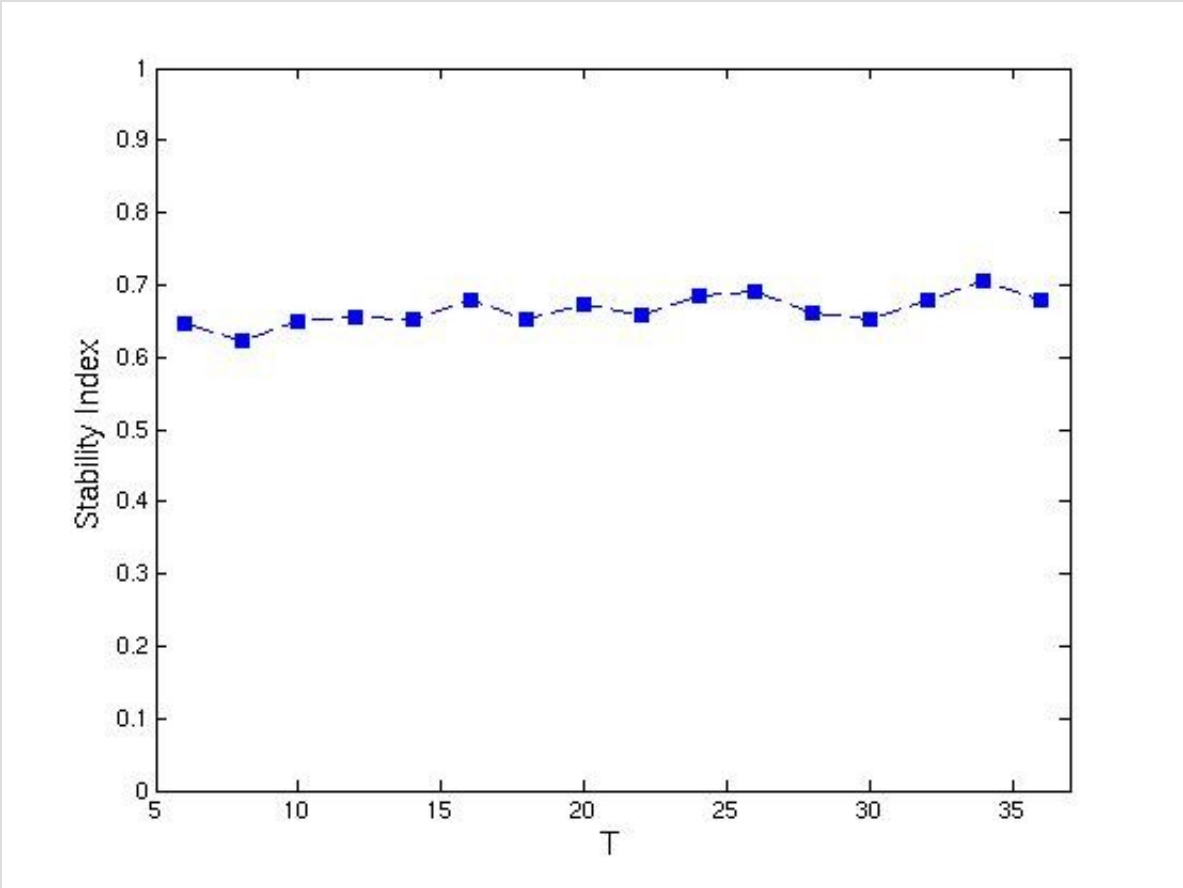
- Clustering Stability versus *K* (the number of example clusters)



Results and Interpretations

Clustering Stability

- Mutual Clustering Stability between U- and W-clustered versus *T*



Clustering Stability

- Summary on Clustering Stability
  - excellent on small *K* (K = 6)
  - quite good when error rate is low
  - slightly increase with *T*
    - feature clustering (dimensionality reduction) does not significantly affect clustering stability

Outline

- Goals
- Data Sampling
- Feature Learning
- Double Clustering
- Results and Interpretations
- Conclusion and Future work

- **Conclusion and Future work**

Conclusion

- Re-description the data
  - sampling the data by two different protocols
  - remove the heterogeneity
  - learn new features
  - two new representations

Conclusion

- Stable clustering
  - feature clustering (dimensionality reduction)
  - stable clustering on grid jobs
  - identify classes unknown to learning algorithm
    - NAR, ABU, GNG
  - find finer subclasses

Future work

- Construct user / job profiles
  - find clusters of users (physicist, biologist ...)
  - find evolution of users (beginner, mastery)
  - usages of communities
- Similar on weeks
  - work load on days

Thank you!

Question?