

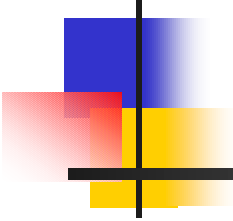
Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching

Authors:

Andrew McCallum

Kamal Nigam

Lyle H. Ungar



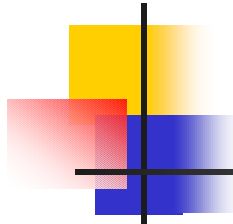
Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching

Group Members:

Surendra Byna

XiaoShan He

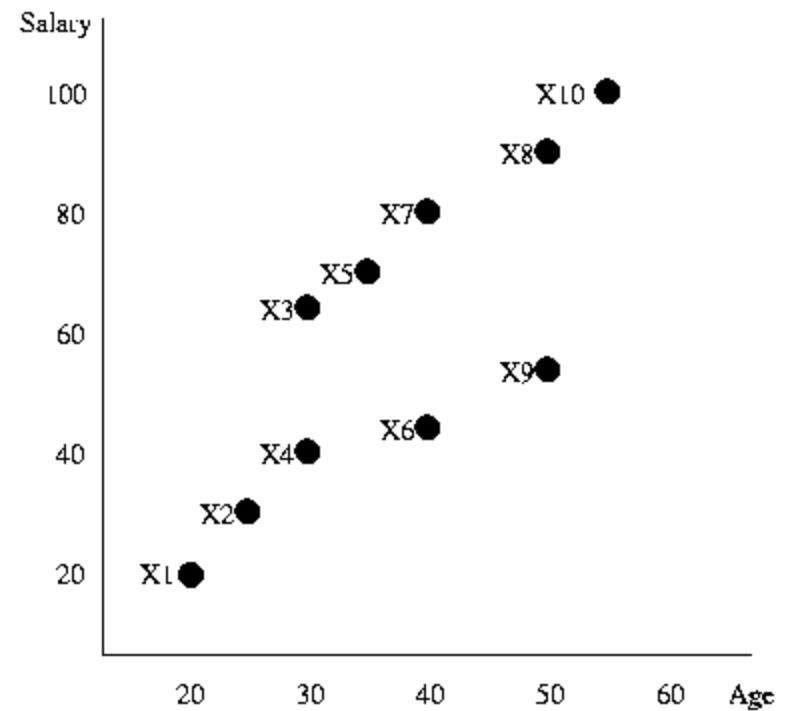
Chih-Wei Yi



Example

■ Data & Figure

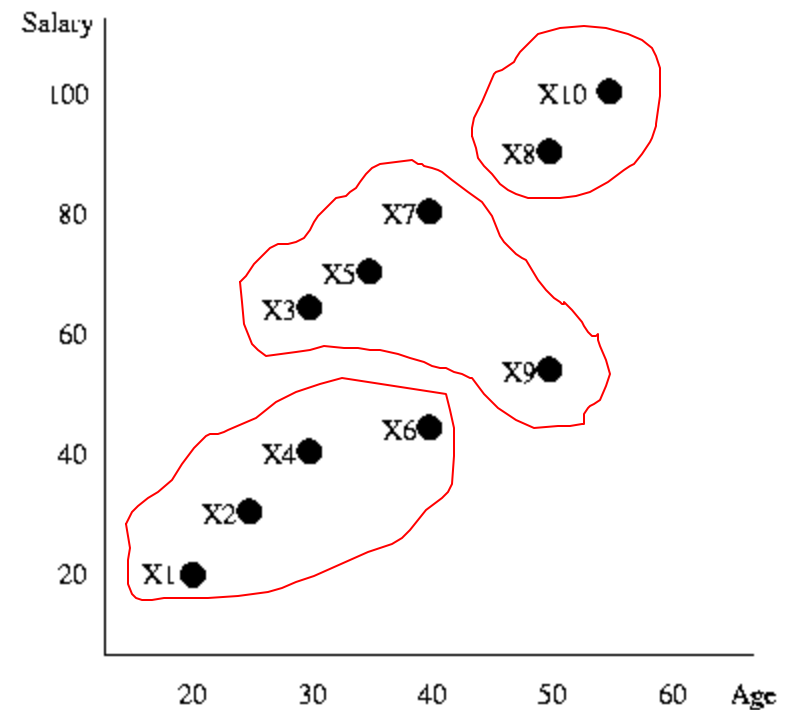
	Age	Salary
X1	20	20
X2	25	30
X3	30	65
X4	30	40
X5	35	70
X6	40	45
X7	40	80
X8	50	90
X9	50	55
X10	55	100



Example

■ K-means: 1st iteration

	K1(30,35)	K2(40,70)	K3(50,90)	
X1	25	70	100	K1
X2	10	55	85	K1
X3	30	15	45	K2
X4	5	40	70	K1
X5	40	5	35	K2
X6	20	25	55	K1
X7	55	10	20	K2
X8	75	30	0	K3
X9	40	25	35	K2
X10	90	45	15	K3
New	K1(29,34)	K2(39,68)	K3(53,95)	





Example

- K-means: 2nd iteration

	K1(29,34)	K2(39,68)	K3(53,95)	
X1	23	67	108	K1
X2	8	52	93	K1
X3	32	12	53	K2
X4	7	37	78	K1
X5	42	6	43	K2
X6	22	24	63	K1
X7	57	13	28	K2
X8	77	33	8	K3
X9	42	24	43	K2
X10	92	48	7	K3

- Additional Information

- $S(\text{Engineer}) < 60$
- $35 < S(\text{Manager}) < 85$
- $75 < S(\text{Partner})$

- What can we do?



Example

- Canopy

- $C1 = \{X1, X2, X4, X6, X9\}$
- $C2 = \{X3, X4, X5, X6, X7, X9\}$
- $C3 = \{X7, X8, X10\}$

- K-means: 1st iteration

	K1(30,35)	K2(40,70)	K3(50,90)	
X1	25	X	X	K1
X2	10	X	X	K1
X3	X	15	X	K2
X4	5	40	X	K1
X5	X	5	X	K2
X6	20	25	X	K1
X7	X	10	20	K2
X8	X	X	0	K3
X9	40	25	X	K2
X10	X	X	15	K3
New	K1(29,34)	K2(39,68)	K3(53,95)	



Example

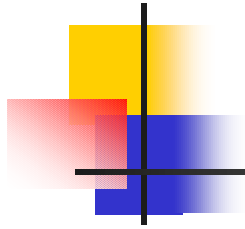
- K-means: 2nd iteration

	K1(30,35)	K2(40,70)	K3(50,90)	
X1	23	X	X	K1
X2	8	X	X	K1
X3	X	12	X	K2
X4	7	37	X	K1
X5	X	6	X	K2
X6	22	24	X	K1
X7	X	13	28	K2
X8	X	X	8	K3
X9	42	24	X	K2
X10	X	X	7	K3



Overview

- Introduction the problem
- Key contribution
- Relevant prior work
- Methodology
- Results
- Our opinion of the paper



Problem

- Some Computational Constraints in Traditional Clustering Algorithm
 - Limited number of clusters
 - Low feature dimensionality
 - Small number of data points



Problem

- Time Complexity
 - K-means
 - Each iteration
 - $O(\text{clusters} * \text{points} * \text{dimension})$
 - Greedy Agglomerative Clustering
 - All pair-wise distances
 - $O(\text{points}^2 * \text{dimension})$

(Suppose the complexity of distance calculation is proportion to dimension)



Key Contribution

- Reduce computation time
- Decrease error?
- Real-value data



Relevant Prior Work

- K-means
- Greedy Agglomerative Clustering
- KD-tree



Methodology

- Two Stage Scenario
 - Creating Canopies
 - Cheap distance metric
 - Overlapping subsets
 - Traditional Clustering Algorithm
 - Agglomerative, K-means, or Expectation-Maximization



Accuracy Guarantee

- The distance to Cluster
 - Centroid
 - For every traditional cluster, there exists a canopy such that all elements of the cluster are in the canopy.
 - Closest
 - For every cluster, there exists a set of canopies such that the elements of the cluster “connect” the canopies.



Results

■ Complexity

- K-means: 1 iteration
 - Old: $O(NK)$
 - New: $O(NKR^2 / C)$
- GAC
 - Old: $O(N^2)$
 - New: $O(N^2R^2 / C)$

■ Notation

- N: # of points
- K: # of clusters
- C: # of canopies
- R: repeated rate



Our Opinion of the Paper

- Reduce the computation time
- How to create canopies?
 - Background knowledge
 - Cheap distance metrics
 - The number of canopies
- Grade: 7