

Jorge Cordero Hernandez and Yifeng Zeng, Department of Computer Science, Aalborg University, Selma Lagerlös Vej 300, DK-9220 Aalborg ø, Denmark

# Background

Attribute clustering is the task of selecting subsets of highly dependent variables from data.





Table 2: Performance ( $W^C$ ) of algorithms (Alg.) in five domains over different sample sizes  $\Omega$  and k=8. The k-modes algorithm is optimal when fed with the right initial modes.

#### **Motivations**

- Widest use in bioinformatics.
- Feature selection, Grouping variables.
- Tree partitioning is easy to implement and fast to execute.

## **Objective**

Given X, the objective is to find a set of disjoint clusters  $C = \{C_i | (i = 1, \dots, k) \land (\forall_{i \neq j} C_i \cap C_j = \emptyset)\}$  that maximizes:

 $W^{C} = \sum_{C_{i}} \sum_{x_{j} \in (C_{i} - \{o_{i}\})} w_{o_{i}, x_{j}}$ (1)

where  $w_{o_i,x_j}$  denote the weight (measured by association functions) from the center  $o_i$  to other variables  $x_j$  in the cluster  $C_i$ .

# Methods

#### k-modes

It identifies points in the space (variables) as centers or modes. Every variable is attached to a mode whose distance is minimal. Find a set of stars and calculate star weights



O Adj	
Leat	

Sort all star and select the stars having maximal weights

Domain      Alg.      10000      8000      6000      400        SEMST      4.13      18.41      21.61      22.9        CESMT      5.4      18.78      22.07      23.5        Alarm      ZEMST      6.11      19.10      22.85      24.6        SD      7.85      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        Barley      SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      50.97      50.32      51.49      52.3        GESMT      51.21      50.55      51.71      52.8        SD      55.57      56.98      58.34      59.5        SD      55.57      56.98      58.34      59.5 <th></th> <th colspan="4"><math>\Delta 2</math></th>		$\Delta 2$				
Alarm      SEMST      4.13      18.41      21.61      22.93        Alarm      ZEMST      5.4      18.78      22.07      23.53        ZEMST      6.11      19.10      22.85      24.6        SD      7.85      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        Amodes      8.35      21.30      23.95      25.3        Barley      SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        Barley      ZEMST      3.85      14.91      20.70      24.2        Barley      ZEMST      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.62      25.4        K-modes      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      50.57      56.	Domain	Alg.	10000	8000	6000	4000
Alarm      CESMT      5.4      18.78      22.07      23.5        ZEMST      6.11      19.10      22.85      24.6        SD      7.85      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        Barley      SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        Barley      ZEMST      3.85      14.91      20.70      24.2        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      5.61      15.39      21.02      25.4        K-modes      5.57      56.98      58.34      59.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5 <td></td> <td>SEMST</td> <td>4.13</td> <td>18.41</td> <td>21.61</td> <td>22.99</td>		SEMST	4.13	18.41	21.61	22.99
Alarm      ZEMST      6.11      19.10      22.85      24.6        SD      7.85      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        Barley      SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        Barley      ZEMST      3.85      14.91      20.70      24.2        Barley      ZEMST      3.85      14.91      20.70      24.2        Barley      ZEMST      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      50.97      50.32      51.49      52.3        SD      55.57      56.98      58.34      59.5        SD      30.26      31.33      32.42      33.6        K-modes      32.48      33.5		CESMT	5.4	18.78	22.07	23.52
SD      7.85      21.30      23.95      25.3        k-modes      8.35      21.30      23.95      25.3        Barley      SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        ZEMST      3.85      14.91      20.70      24.2        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      5.61      15.39      21.02      25.4        K-modes      51.21      50.55      51.71      52.8        CESMT      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        K-modes      30.26      31.33      32.42      33.6        GESMT      30.26      31.33      32.42      33.6        SD	Alarm	ZEMST	6.11	19.10	22.85	24.66
k-modes      8.35      21.30      23.95      25.3        SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        Barley      ZEMST      3.85      14.91      20.70      24.2        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      51.21      50.55      51.71      52.8        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        k-modes      31.02      32.00      33.01      34.1        ZEMST      30.26      31.33      32.42      33.6        SD      32.48      33.58      34.69      35.9        k-mode		SD	7.85	21.30	23.95	25.38
SEMST      2.33      14.67      19.03      22.2        CESMT      2.55      14.85      19.24      22.4        ZEMST      3.85      14.91      20.70      24.2        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      51.27      50.32      51.49      52.3        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        K-modes      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.1        ZEMST      32.41      33.28      34.69      35.9        k-modes      32		k-modes	8.35	21.30	23.95	25.38
CESMT      2.55      14.85      19.24      22.4        Barley      ZEMST      3.85      14.91      20.70      24.2        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      5.61      15.39      21.02      25.4        Leparll      SEMST      50.97      50.32      51.49      52.3        CESMT      51.21      50.55      51.71      52.8        ZEMST      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        K-modes      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.9        SD      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9		SEMST	2.33	14.67	19.03	22.23
Barley      ZEMST      3.85      14.91      20.70      24.2        SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      50.97      50.32      51.49      52.3        CESMT      51.21      50.55      51.71      52.8        CESMT      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        K-modes      31.02      32.00      33.01      34.1        SD      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      36.6 <td< td=""><td></td><td>CESMT</td><td>2.55</td><td>14.85</td><td>19.24</td><td>22.48</td></td<>		CESMT	2.55	14.85	19.24	22.48
SD      4.88      15.39      21.02      25.4        k-modes      5.61      15.39      21.02      25.4        K-modes      5.61      15.39      21.02      25.4        SEMST      50.97      50.32      51.49      52.3        CESMT      51.21      50.55      51.71      52.8        ZEMST      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        k-modes      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.1        ZEMST      32.41      33.28      33.81      34.9        SD      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        SEMST      85.98      87.53      88.75      89.6        CESMT      88.63	Barley	ZEMST	3.85	14.91	20.70	24.20
k-modes      5.61      15.39      21.02      25.4        SEMST      50.97      50.32      51.49      52.3        CESMT      51.21      50.55      51.71      52.8        ZEMST      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        k-modes      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.4        SD      32.41      33.28      34.69      35.9        K-modes      32.48      33.58      34.69      36.9        K-modes      88.315      88.75      89.64      90.6        SD		SD	4.88	15.39	21.02	25.41
SEMST      50.97      50.32      51.49      52.3        CESMT      51.21      50.55      51.71      52.8        ZEMST      51.27      51.43      52.55      53.5        SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        K-modes      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.1        ZEMST      32.41      33.28      33.81      34.9        SD      32.48      33.58      34.69      35.9        k-modes      38.315      88.75      89.64      90.6        SD      86.61      89.31      89.71      91.0        k-modes      <		k-modes	5.61	15.39	21.02	25.41
CESMT      51.21      50.55      51.71      52.8        Heparll      ZEMST      51.27      51.43      52.55      53.5        SD <b>55.57 56.98 58.34 59.5</b> k-modes <b>55.57 56.98 58.34 59.5</b> K-modes <b>55.57 56.98 58.34 59.5</b> K-modes <b>30.26</b> 31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.1        ZEMST      32.41      33.28      33.81      34.9        SD <b>32.48 33.58 34.69 35.9</b> k-modes <b>32.48 33.58 34.69 35.9</b> Resons <b>85.98 87.53 88.75 89.6</b> Pathfinder      ZEMST      88.315 <b>88.75 89.64 90.6</b> SD <b>86.61 89.31 89.71 91.02 86.61 89.41 91.32 92.7</b>		SEMST	50.97	50.32	51.49	52.32
Heparll    ZEMST    51.27    51.43    52.55    53.5      SD    55.57    56.98    58.34    59.5      k-modes    30.26    31.33    32.42    33.6      CESMT    31.02    32.00    33.01    34.1      SD    32.41    33.28    33.81    34.9      SD    32.48    33.58    34.69    35.9      k-modes    32.48    33.58    34.69    35.9      Remodes    88.63    88.22    89.40    90.4      SD    86.61    89.31    89.71    91.0      k-modes    90.33    89.41    91.32    92.7		CESMT	51.21	50.55	51.71	52.89
SD      55.57      56.98      58.34      59.5        k-modes      55.57      56.98      58.34      59.5        SEMST      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.1        ZEMST      32.41      33.28      33.81      34.9        SD      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        SD      32.48      33.58      34.69      35.9        Remodes      32.48      33.58      34.69      35.9        Remodes      32.48      33.58      34.69      35.9        SD      85.98      87.53      88.75      89.6        SD      86.61      89.31      89.71      91.0        k-modes      90.33      89.41      91.32      92.7	Heparll	ZEMST	51.27	51.43	52.55	53.54
k-modes      55.57      56.98      58.34      59.5        SEMST      30.26      31.33      32.42      33.6        CESMT      31.02      32.00      33.01      34.1        ZEMST      32.41      33.28      33.81      34.9        SD      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        k-modes      32.48      33.58      34.69      35.9        CESMT      85.98      87.53      88.75      89.8        CESMT      88.63      88.22      89.40      90.1        SD      86.61      89.31      89.71      91.0        k-modes      90.33      89.41      91.32      92.7		SD	55.57	56.98	58.34	59.56
SEMST    30.26    31.33    32.42    33.6      CESMT    31.02    32.00    33.01    34.1      ZEMST    32.41    33.28    33.81    34.9      SD    32.48    33.58    34.69    35.9      k-modes    32.48    33.58    34.69    35.9      k-modes    32.48    33.58    34.69    35.9      SEMST    85.98    87.53    88.75    89.8      CESMT    88.63    88.22    89.40    90.1      Pathfinder    ZEMST    88.315    88.75    89.64    90.6      SD    86.61    89.31    89.71    91.0    91.0		k-modes	55.57	56.98	58.34	59.56
CESMT31.0232.0033.0134.1HailfinderZEMST32.4133.2833.8134.9SD32.4833.5834.6935.9k-modes32.4833.5834.6935.9k-modes32.4833.5834.6935.9CESMT85.9887.5388.7589.6CESMT88.6388.2289.4090.1PathfinderZEMST88.31588.7589.6490.6SD86.6189.3189.7191.0k-modes90.3389.4191.3292.7		SEMST	30.26	31.33	32.42	33.65
HailfinderZEMST32.4133.2833.8134.9SD32.4833.5834.6935.9k-modes32.4833.5834.6935.9SEMST85.9887.5388.7589.8CESMT88.6388.2289.4090.1PathfinderZEMST88.31588.7589.64SD86.6189.3189.7191.0k-modes90.3389.4191.3292.7		CESMT	31.02	32.00	33.01	34.16
SD    32.48    33.58    34.69    35.9      k-modes    32.48    33.58    34.69    35.9      SEMST    85.98    87.53    88.75    89.6      CESMT    88.63    88.22    89.40    90.1      Pathfinder    ZEMST    88.315    88.75    89.64    90.6      k-modes    90.33    89.41    91.32    92.7	Hailfinder	ZEMST	32.41	33.28	33.81	34.97
k-modes32.4833.5834.6935.9SEMST85.9887.5388.7589.6CESMT88.6388.2289.4090.1PathfinderZEMST88.31588.7589.6490.6SD86.6189.3189.7191.0k-modes90.3389.4191.3292.7		SD	32.48	33.58	34.69	35.96
SEMST $85.98$ $87.53$ $88.75$ $89.8$ CESMT $88.63$ $88.22$ $89.40$ $90.1$ PathfinderZEMST $88.315$ $88.75$ $89.64$ $90.6$ SD $86.61$ $89.31$ $89.71$ $91.0$ k-modes <b>90.3389.4191.3292.7</b>		k-modes	32.48	33.58	34.69	35.96
CESMT88.6388.2289.4090.1PathfinderZEMST88.31588.7589.6490.6SD86.6189.3189.7191.0k-modes90.3389.4191.3292.7		SEMST	85.98	87.53	88.75	89.82
Pathfinder      ZEMST      88.315      88.75      89.64      90.6        SD      86.61      89.31      89.71      91.0        k-modes      90.33      89.41      91.32      92.7		CESMT	88.63	88.22	89.40	90.19
SD86.6189.3189.7191.0k-modes90.3389.4191.3292.7	Pathfinder	ZEMST	88.315	88.75	89.64	90.61
k-modes 90.33 89.41 91.32 92.7		SD	86.61	89.31	89.71	91.03
		k-modes	90.33	89.41	91.32	92.72

#### Maximum Spanning Tree (MAST) Clustering

- Standard Euclidean maximum spanning tree (SEMST):
  k 1 inconsistent edges with minimal weights are removed.
- Maximum cost spanning tree (CESMT): k 1 inconsistent edges with minimal costs are removed.
- Zahn's maximum spanning tree (ZEMST): Edges are removed if their attached weights are smaller than the average of weights in neighborhoods.

# **Star Discovery Algorithm**

- The star discovery (SD) algorithm uses both weights and topology for deleting inconsistent edges.
- The MAST is divided into Spanning Stars (we encapsulate a center *O*, adjacent variables (*Adj*) and leaf nodes (*Leaf*)).
- A Spanning Star is a sub-tree over the MAST,  $S = (V_S, E_S)$ . It has a center and a set of adjacent nodes (extended to include leaf nodes).

#### Illustration



• Every resulted star becomes a cluster



# **Experimental Results**

## *k*-modes Reliability

Feed all possible combinations of initial modes given the setting: k=2 and  $\Omega=10000$ .

Table 1: Number of local optima into which the k-modes algorithm falls.

#### Domains vs Local Optima

## **Timing Issue**

Measurements in seconds given  $\Omega$ =10000.

Table 3: Elapsed times for algorithms in all domains.

	Domain				
Alg.	Alarm	Barley	Heparll	Hailfinder	Pathfinder
SEMST	0.031	0.04	0.044	0.049	0.047
CESMT	0.04	0.042	0.056	0.05	0.062
ZEMST	0.078	0.057	0.065	0.07	0.094
SD	0.047	0.04	0.046	0.061	0.062
k-modes	0.109	0.063	0.077	0.078	0.125

# Discussions

- The SD outperforms other MAST-based clustering algorithms (quality).
- The SD is competitive (it obtains similar results as *k*-modes in a reduced time complexity).

We find the set of stars SS that maximizes:

# $W = \sum_{S_l \in SS} \left( \sum_{x_i \in Adj_l} (w_{x_i,o_l}) + \sum_{x_j \in Adj_l, x_h \in Leaf_l} (w_{x_j,x_h}) \right)$ (2)

where  $o_l$  is the star(cluster) center,  $Adj_l$  is a set of adjacent nodes to the center node  $o_l$ , and  $Leaf_l$  a set of leaf nodes that connect to either  $o_l$  or  $Adj_l$ .

Bomanio vo Local Optima				
Alarm	Heparll	Hailfinder	Pathfinder	
17	130	91	117	

## **Cluster Quality**

Initial modes of k-modes are taken from the final modes in SD.

- The SD is deterministic while the *k*-modes falls in local optima.
- The SD + k-modes produces optimal clusters.
- Complexity of learning a BN can be reduced by learning small BNs (for each cluster) and then combining them.