MŰEGYETEM 1782

SOHAC: Efficient Storage of Tick Data That Supports Search and Analysis

Gabor I. Nagy, Krisztian Buza
Budapest University of Technology and Economics
gnagy@tmit.bme.hu, buza@cs.bme.hu

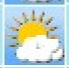
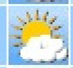
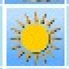
Acknowledgements:
Discussions with Dr. Ferenc Bodon and Zoltan Papp, Morgan Stanley Analytics, Budapest, Hungary are greatly appreciated.
TÁMOP - 4.2.2.B-10/1–2010-0009

Outline

- Introduction
- Problem Formulation
- Our approach
- Experiments
- Conclusions and Future Work

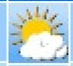
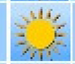
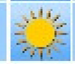
Introduction

- Real-world phenomena
  – described by several attributes
  – the dynamics of these attributes matters
- Illustrative example: weather observations
- Application domains
  – finance (tick data), seismology, medicine, sensor data…

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	20	100 200	5	SW	low	
10:22	16	20	100 200	5	SW	low	
10:38	16	30	100 100	5	SW	low	
10:40	17	30	100 100	5	SW	medium	
10:43	18	30	100 100	10	SW	medium	
10:44	18	30	100 100	15	W	medium	
10:51	18	20	100 200	15	W	medium	

Storage of tick data

- Omit rows where no attribute changes
- Challange: Find balance between two criteria
  - Storage space occupation
  - Quick access to the data → search & analysis

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	20	100 200	5	SW	low	
10:22	16	20	100 200	5	SW	low	
10:38	16	30	100 100	5	SW	low	
10:40	17	30	100 100	5	SW	medium	
10:43	18	30	100 100	10	SW	medium	
10:44	18	30	100 100	15	W	medium	
10:51	18	20	100 200	15	W	medium	

Decomposition of a tick data table (illustrative example)

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	20	100 200	5	SW	low	
10:22	16	20	100 200	5	SW	low	
10:38	16	30	100 100	5	SW	low	
10:40	17	30	100 100	5	SW	medium	
10:43	18	30	100 100	10	SW	medium	
10:44	18	30	100 100	15	W	medium	
10:51	18	20	100 200	15	W	medium	

Time	Hum. (%)	Press. (Pa)
10:21	20	100 200
10:38	30	100 100
10:51	20	100 200

Time	Temp. (°C)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	5	SW	low	
10:22	16	5	SW	low	
10:40	17	5	SW	medium	
10:43	18	10	SW	medium	
10:44	18	15	W	medium	

Decomposition of a tick data table (illustrative example)

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	20	100 200	5	SW	low	
10:22	16	~~20~~	~~100 200~~	5	SW	low	
10:38	16	30	100 100	5	SW	low	
10:40	17	~~30~~	~~100 100~~	5	SW	medium	
10:43	18	~~30~~	~~100 100~~	10	SW	medium	
10:44	18	~~30~~	~~100 100~~	15	W	medium	
10:51	18	20	100 200	15	W	medium	

Time	Hum. (%)	Press. (Pa)
10:21	20	100 200
10:38	30	100 100
10:51	20	100 200

Time	Temp. (°C)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	5	SW	low	
10:22	16	5	SW	low	
10:40	17	5	SW	medium	
10:43	18	10	SW	medium	
10:44	18	15	W	medium	

Problem Formulation

- Given a number *k*, find a decomposition into *k* tables so that the storage space is minimized
  - Usually: *k* = 2 or *k* = 3 in practice
- Clustering problem
  - Domain-specific notion of similarity

Time	Hum. (%)	Press. (Pa)
10:21	20	100 200
10:38	30	100 100
10:51	20	100 200

Time	Temp. (°C)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	5	SW	low	
10:22	16	5	SW	low	
10:40	17	5	SW	medium	
10:43	18	10	SW	medium	
10:44	18	15	W	medium	

Decomposition of a tick data table (illustrative example)

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	20	100 200	5	SW	low	
10:22	16	20	100 200	5	SW	low	
10:38	16	30	100 100	5	SW	low	
10:40	17	30	100 100	5	SW	medium	
10:43	18	30	100 100	10	SW	medium	
10:44	18	30	100 100	15	W	medium	
10:51	18	20	100 200	15	W	medium	

Time	Hum. (%)	Press. (Pa)
10:21	20	100 200
10:38	30	100 100
10:51	20	100 200

Time	Temp. (°C)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	5	SW	low	
10:22	16	5	SW	low	
10:40	17	5	SW	medium	
10:43	18	10	SW	medium	
10:44	18	15	W	medium	

Preprocessing: Construction of a binary change indicator matrix

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	15	20	100 200	5	SW	low	
10:22	16	20	100 200	5	SW	low	
10:38	16	30	100 100	5	SW	low	
10:40	17	30	100 100	5	SW	medium	
10:43	18	30	100 100	10	SW	medium	
10:44	18	30	100 100	15	W	medium	
10:51	18	20	100 200	15	W	medium	

Time	Temp. (°C)	Hum. (%)	Press. (Pa)	Wind (v) (km/h)	Wind (dir.)	Radiation	Outlook
10:21	1	1	1	1	1	1	1
10:22	1	0	0	0	0	0	0
10:38	0	1	1	0	0	0	0
10:40	1	0	0	0	0	1	1
10:43	1	0	0	1	0	0	0
10:44	0	0	0	1	1	0	0
10:51	0	1	1	0	0	0	0

Our approach

- SOHAC: Storage-Optimizing Hierarchical Agglomerative Clustering
  – Clustering algorithm in order to find an (approximately) optimal partitioning of columns
  – Basic idea:
    - cluster: set of columns
    - initially: each column is a separate cluster
    - in each iteration of HAC, we merge those clusters that lead to optimal storage

Experiments

- *Datasets*
  - Morgan Stanley Tick Data (30 columns, ≈4M rows)
  - Publicly availabe real-world datasets
    - Some of the most popular datasets from the UCI repository: Adult, Breast Cancer Wisconsin (Diagnostic), Car Evaluation, Forest Fires and Poker Hand
- *Performance measure*: compression ratio

$$CR = \frac{\text{number of cells after decomposition}}{\text{number of cells in the original matrix}}$$

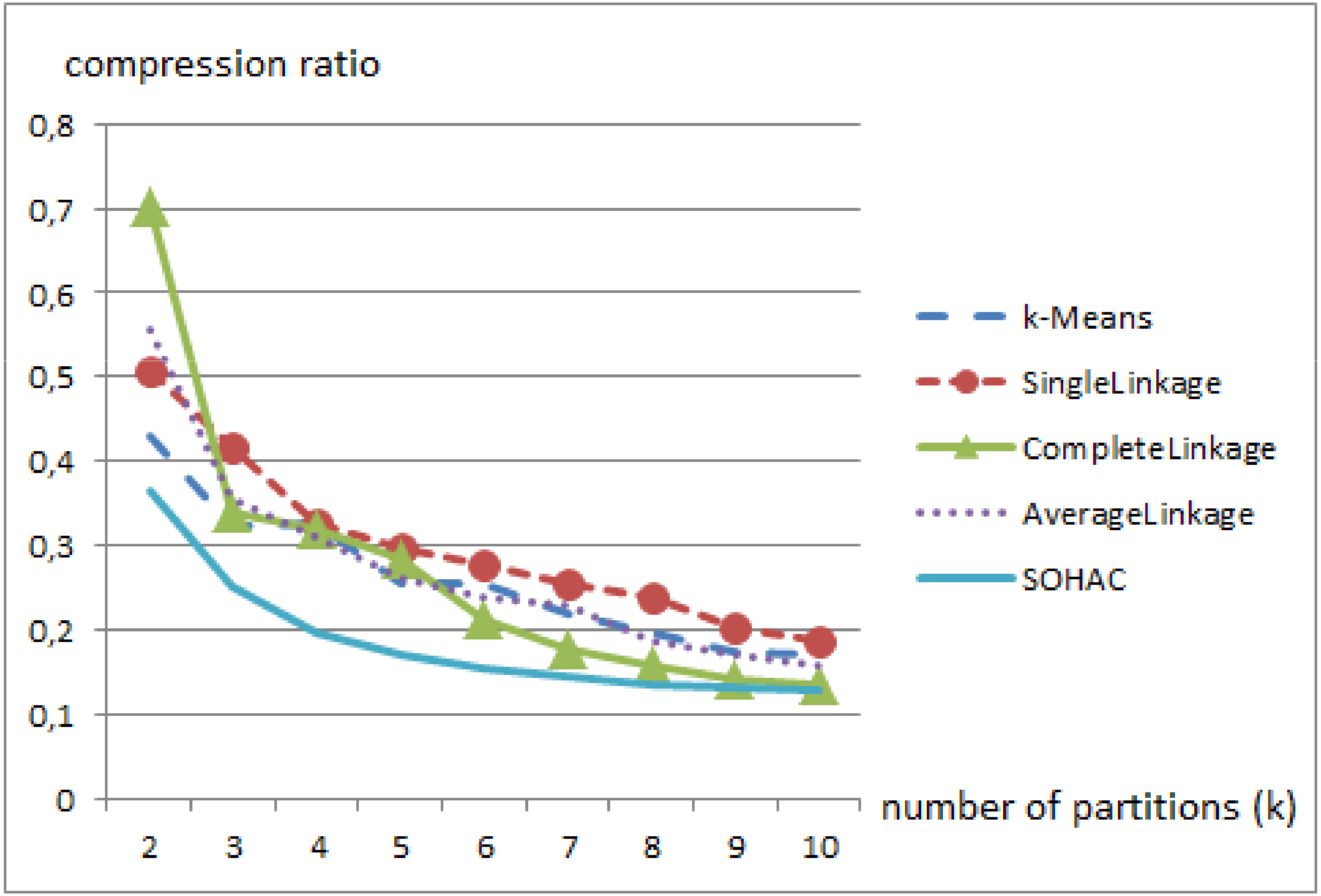
- *10 disjoint splits* → average + standard deviation
- *Baselines*:
  - Hierarchical clustering algorithms and *k*-Means with various distance measures
  - In total: 38 clustering algorithms from the literature

Results on Morgan Stanley Tick Data

Algorithm	Distance Measure	$k=2$	$k=3$	$k=4$
Average-Linkage	Dice	0.9799±0.0016	0.5805±0.0596	0.3094±0.1311
	Jaccard	0.9799±0.0016	0.5805±0.0596	0.3094±0.1311
	Kulczynski	0.9799±0.0016	0.5805±0.0596	0.3094±0.1311
	Nominal	0.7385±0.1072	0.6309±0.0731	0.5612±0.0753
	Rogers-Tanimoto	0.7320±0.1032	0.6339±0.0696	0.5312±0.1184
	RussellRao	0.9799±0.0016	0.5805±0.0596	0.3094±0.1311
	SimpleMatching	0.7320±0.1032	0.6339±0.0696	0.5312±0.1184
	Chebychev	0.9799±0.0016	0.7169±0.0412	0.6836±0.0413
	Cosine	0.5556±0.2659	0.3560±0.0852	0.3084±0.0601
	Euclidean	0.7320±0.1032	0.6339±0.0696	0.5312±0.1184
	Manhattan	0.7320±0.1032	0.6339±0.0696	0.5312±0.1184
	Overlap	0.9605±0.0172	0.8788±0.0129	0.7734±0.0222
Complete-Linkage	Dice	0.7415±0.1020	0.5805±0.0596	0.3094±0.1311
	Jaccard	0.7415±0.1020	0.5805±0.0596	0.3094±0.1311
	Kulczynski	0.7415±0.1020	0.5805±0.0596	0.3094±0.1311
	Nominal	0.7044±0.0462	0.3460±0.1328	0.3254±0.1361
	RogersTanimoto	0.7013±0.0446	0.3388±0.1273	0.3190±0.1299
	RussellRao	0.9799±0.0016	0.5805±0.0596	0.3094±0.1311
	SimpleMatching	0.7013±0.0446	0.3388±0.1273	0.3190±0.1299
	Chebychev	0.9799±0.0016	0.7169±0.0412	0.6836±0.0413
	Cosine	0.8303±0.0762	0.7306±0.1298	0.3075±0.0875
	Euclidean	0.7013±0.0446	0.3388±0.1273	0.3190±0.1299
	Manhattan	0.7013±0.0446	0.3388±0.1273	0.3190±0.1299
	Overlap	0.8696±0.0475	0.7620±0.0408	0.6970±0.0441
Single-Linkage	Dice	0.9799±0.0016	0.7301±0.1952	0.3338±0.1629
	Jaccard	0.9799±0.0016	0.7301±0.1952	0.3338±0.1629
	Kulczynski	0.9799±0.0016	0.7301±0.1952	0.3338±0.1629
	Nominal	0.7607±0.1379	0.7296±0.1511	0.5612±0.0753
	RogersTanimoto	0.7520±0.1329	0.7228±0.1441	0.5587±0.0714
	RussellRao	0.9799±0.0016	0.9055±0.0264	0.4820±0.3088
	SimpleMatching	0.7520±0.1329	0.7228±0.1441	0.5587±0.0714
	Chebychev	0.9799±0.0016	0.7169±0.0412	0.6836±0.0413
	Cosine	0.5072±0.2641	0.4150±0.1893	0.3254±0.0683
	Euclidean	0.7520±0.1329	0.7228±0.1441	0.5587±0.0714
	Manhattan	0.7520±0.1329	0.7228±0.1441	0.5587±0.0714
	Overlap	0.9799±0.0016	0.9466±0.0016	0.9134±0.0018
k-Means	Euclidean	0.4291±0.1821	0.3242±0.1216	0.3244±0.1309
	Manhattan	0.8084±0.1219	0.6029±0.1224	0.4437±0.1274
SOHAC		**0.3649±0.0772**	**0.2526±0.0587**	**0.1960±0.0499**

our approach

Varying the number of partitions



Results on publicly available real-world datasets

Dataset	SOHAC	Single Linkage	Avg. Linkage	Complete Linkage
k = 2				
Adult	**0.8051±0.0256**	0.8672±0.0473	0.8558±0.0408	0.8558±0.0408
Breast C.W.	**0.5040±0.2420**	0.5708±0.2243	0.5478±0.2181	0.5142±0.2243
Car	**0.5199±0.0291**	0.6347±0.0806	0.6108±0.0733	0.5909±0.0660
ForestFires	**0.7816±0.0208**	0.7887±0.0286	0.7834±0.0288	0.7925±0.0389
Poker Hand	**0.5490±0.0001**	0.7582±0.0572	0.7582±0.0572	0.7871±0.0018
k = 3				
Adult	**0.7101±0.0251**	0.8018±0.0515	0.7884±0.0397	0.7876±0.0388
Breast C.W.	**0.4451±0.2424**	0.5022±0.2167	0.4915±0.2189	0.4628±0.2292
Car	**0.3869±0.0190**	0.4389±0.0235	0.4391±0.0238	0.4391±0.0238
ForestFires	**0.7242±0.0202**	0.7406±0.0212	0.7402±0.0213	0.7387±0.0178
Poker Hand	**0.4477±0.0003**	0.5978±0.0011	0.5978±0.0011	0.5978±0.0011
k = 4				
Adult	**0.6491±0.0222**	0.7402±0.0125	0.7437±0.0215	0.7501±0.0272
Breast C.W.	**0.4068±0.2344**	0.4414±0.2199	0.4394±0.2215	0.4289±0.2183
Car	**0.3141±0.0206**	0.3146±0.0198	0.3146±0.0198	0.3146±0.0198
ForestFires	**0.6857±0.0191**	0.7144±0.0214	0.7113±0.0226	0.7105±0.0177
Poker Hand	**0.4016±0.0004**	0.4272±0.0005	0.4272±0.0005	0.4272±0.0005

our approach

Outlook & Future work

- Other algorithms for finding the optimal decomposition
- Study the stability of the algorithm and speed-up
- Study the presence of hubs and hub-based clustering algorithms
- New domains
  - multivariate time-series
  - sensor data
  - biomedical data

Conclusion

- Reduction of storage space while allowing quick access to the data
- Use clustering algorithms for the above problem
- SOHAC: Storage-Optimizing Hierarchical Agglomerative Clustering
- Extensive experiments: our approach outperformed other clustering algorithms