# DATA MINING 2 Transactional Clustering

Riccardo Guidotti

a.a. 2023/2024



# Clustering

- **Clustering**: Grouping of objects into different sets, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait often proximity according to some defined distance measure
- Common distance functions:
  - Euclidean distance, Manhattan distance, ...
- This kind of distance functions are suitable for **numerical data**

## Not Only Numerical Data

	Price	Height	Width	Lenght	Acceleration	$\bigcirc$
	(€)	<b>(</b> m <b>)</b>	<b>(</b> m <b>)</b>	<b>(</b> m <b>)</b>	0-100 (s)	
	20'000	1.7	1.6	4	12	
Numerical Data	16' 000	1.65	1.5	3.7	14	
Numerical Data	12'000	1.6	1.5	3.5	15	
	<b>24'</b> 000	1.7	1.8	4.2	9.4	

**Categorical Data** 



## **Boolean and Categorical Attributes**

- A **boolean** attribute corresponding to a single item in a transaction, if that item appears, the boolean attribute is set to '1' or '0' otherwise.
- A categorical attribute may have several values, each value can be treated as an item and represented by a boolean attribute.

## Market Basket Data

- A transaction represents one customer, and each transaction contains set of items purchased by the customer.
- Clustering customers reveals customers with similar buying patterns putting them into the same cluster.
- It is useful for
  - Characterizing different customer groups
  - Targeted Marketing
  - Predict buying patterns of new customers based on profile
- A market basket database: Attributes of data points are non-numeric, transaction viewed as records with boolean attributes corresponding to a single item (TRUE if transaction contain item, FALSE otherwise).
- Boolean attributes are special case of Categorical attributes.

# **Shortcomings of Traditional Clustering**

- For categorical data we:
  - Define new criterion for *neighbors* and/or *similarity*
  - Define the ordering criterion
- Consider the following 4 market basket transactions



 using Euclidean distance to measure the closeness between all pairs of points, we find that d(P1,P2) is the smallest distance: it is equal to 1

# Shortcomings of Traditional Clustering

- If we use a hierarchical algorithm then we merge P1 and P2 and get a new cluster (P12) with (1, 1, 0.5, 1) as a centroid
- Then, using Euclidean distance again, we find:
  - d(p12,p3)= √3.25
  - d(p12,p4)= √2.25
  - d(p3,p4)= √2
- So, we should merge P3 and P4 since the distance between them is the shortest.
- However, T3 and T4 don't have even a single common item.
- So, using distance metrics as similarity measure for **categorical** data is not appropriate.

P1= (1, 1, 1, 1) P2= (1, 1, 0, 1) P3= (0, 0, 1, 0) P4= (0, 0, 0, 1)

# Algorithms for Categorical/Transactional Data

- K-Modes
- ROCK
- CLOPE
- TX-Means

#### K-Modes

Minimise 
$$P(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l} d(X_i, Q_l)$$
  
subject to  $\sum_{l=1}^{k} w_{i,l} = 1, \quad 1 \le i \le n$   
 $w_{i,l} \in \{0, 1\}, \quad 1 \le i \le n, \ 1 \le l \le k$ 

- $X = \{X_1, ..., X_n\}$  is the dataset of objects.
- $X_i = [x_1, ..., x_m]$  is an object i.e., a vector of *m* categorical attributes
- W is a matrix  $n \times k$ , with  $w_{i,l}$  equal to 1 if  $X_i$  belongs to Cluster l, 0 otherwise.
- Q = { Q<sub>1</sub>,..., Q<sub>k</sub> } is the set of representative objects (mode) for the k clusters.
- $d(X_i, Q_i)$  is a distance function for objects in the data

## K-Modes: Distance

• K-Means as distance uses Euclidean distance

$$d(X,Y) = \mathop{\text{a}}\limits_{i=1}^{m} (x_i - y_i)^2$$

 K-Modes as distance uses the number of mismatches between the attributes of two objects.

$$d_1(X, Y) = \sum_{j=1}^m \delta(x_j, y_j)$$

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

### K-Modes: Mode

- K-Modes uses the mode as representative object of a cluster
- Given the set of objects in the cluster C<sub>1</sub> the mode is get computing the max frequency for each attribute

$$f_r(A_j = c_{l,j} | X_l) = \frac{n_{c_{l,k}}}{n}$$

# K-Modes: Algorithm

- 1. Randomly select the initial objects as modes
- 2. Scan of the data to assign each object to the closer cluster identified by the mode
- 3. Re-compute the mode of each cluster
- 4. Repeat the steps 2 and 3 until no object changes the assigned cluster

# ROCK: RObust Clustering using linK

- ROCK is a **hierarchical** algorithm for clustering transactional data (market basket databases)
- ROCK uses links to cluster instead of the classical distance notion
- ROCK uses the notion of neighborhood between pair of objects to identify the number of links between two objects

# **ROCK: Clustering Algorithm**

#### Input:

A set S of data points Number of *k* clusters to be found The similarity threshold

#### **Output:**

Groups of clustered data

The ROCK algorithm is divided into three major parts:

- 1. Draw a random sample from the data set
- 2. Perform a hierarchical agglomerative clustering algorithm
- 3. Label data

#### Draw a random sample from the data set:

- Sampling is used to ensure scalability to very large data sets
- The initial sample is used to form clusters, then the remaining data on dataset is assigned to these clusters

#### Perform a hierarchical agglomerative clustering algorithm:

- ROCK performs the following steps which are common to all hierarchical agglomerative clustering algorithms, but with different definition to the similarity measures:
  - 1. Places each single data point into a separate cluster
  - 2. Compute the similarity measure for all pairs of clusters
  - 3. Merge the two clusters with the highest similarity (goodness measure)
  - 4. Verify a stop condition. If it is not met, then go to step 2.

# **ROCK: The Neighbors Concept**

- It captures a notion of **similarity** 
  - A and B are neighbors if sim(A, B) ≥ θ
- ROCK uses the Jaccard coefficient
  - sim(A, B)= |A ∩ B| / | A U B |



- A **link** defines the number of common neighbors between two objects:
- link(A, B) = |neighbor(A) ∩ neighbor(B) |
- Higher values of *link(A, B)* means higher probability that *A* and *B* belong to the same cluster
- **Similarity** is **local** while **link** is capturing **global** information
- A point is considered a neighbor of itself
- There is a link from each neighbor of the "root" point back to itself through the root
- Therefore, if a point has *n* neighbors, then *n*<sup>2</sup> links are due to it.



# ROCK: Example

- Data consisting of 6 Attributes:
  - {Book}
  - {Water, Sun, Sand, Swimming}
  - {Water, Sun, Sand, Reading}
  - {Reading, Sand}
- Resulting Jaccard Coefficient Matrix
- Set Threshold = 0.2. Neighbors:
  - N(A)={A}; N(B)={B,C,D}
  - N(C)={B,C,D}, N(D) = {B,C,D}
- Number of Links Table
  - Link (B, C) = |{B,C,D}| = 3
- Resulting Clusters after applying ROCK: {A}, {B,C,D}

{Book, Water, Sun, Sand, Swimming, Reading}

	Α	В	С	D
Α	1	0	0	0
B	0	1	0.6	0.2
С	0	0.6	1	0.5
D	0	0.2	0.5	1

	Α	В	С	D
Α	1	0	0	0
В	0	3	3	3
С	0	3	3	3
D	0	3	3	3

## **ROCK – Criterion Function**

Maximize 
$$E_l = \sum_{i=1}^k n_i * \sum_{p_q, p_r \in C_i} \frac{link(p_q, p_r)}{n_i^{1+2f(\theta)}}$$

Dividing by the number of expected links between pairs of objects in the cluster C<sub>i</sub> we avoid that objects with a low number of links are assigned all to the same cluster

 $f(\theta) = -$ 

Where  $C_i$  denotes cluster i  $n_i$  is the number of points in  $C_i$ k is the number of clusters  $\theta$  is the similarity threshold

This goodness measure helps to identify the best pair of clusters to be merged during each step of ROCK.

$$g(C_i, C_j) = \underbrace{link[C_i, C_j]}_{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)} - n_j^{1+2f(\theta)}}_{\text{Number of expected cross-links between two clusters}}$$

# ROCK: Clustering Algorithm

#### Label data

- Finally, the remaining data points are assigned to the clusters.
- This is done by selecting a random sample L<sub>i</sub> from each cluster C<sub>i</sub>, then we assign each point p to the cluster for which it has the strongest linkage with L<sub>i</sub>.

## **ROCK Summary**

Input: dataset, number of clusters.

- 1. Draw a random sample from the data set
- 2. Places each data point into a separate cluster
- 3. Compute the similarity measure for all pairs of clusters
- 4. Merge the two clusters with the highest similarity
- 5. Verify a stop condition. If it is not met, then go to step 2.
- 6. Assign not used points to clusters using linkage similarity with respect to selected samples from each cluster

# CLOPE: Clustering with sLOPE

- Transactional clustering efficient for high dimensional data
- Uses a **global criterion function** that tries to increase the intra-cluster overlapping of transaction items by increasing the height-to-width ratio of the cluster histogram.

Example: 5 transactions {a,b} {a,b,c} {a,c,d} {d,e} {d,e,f}



#### **Higher H/W means higher item overlapping**

#### **CLOPE: Criterion Function**

- For evaluating the goodness of a clustering the gradient of a cluster is
- $G(C)=H(C)/W(C)=S(C)/W(C)^{2}$



# CLOPE: Algorithm

- /\* Phrase 1 Initialization \*/
- 1: while not end of the database file
- 2: read the next transaction  $\langle t, \text{unknown} \rangle$ ;
- 3: put t in an existing cluster or a new cluster  $C_i$  that maximize profit;
- 4: write  $\langle t, i \rangle$  back to database;
  - /\* Phrase 2 Iteration \*/
- 5: repeat
- 6: rewind the database file;
- 7: moved = false;
- 8: while not end of the database file
- 9: read  $\langle t, i \rangle$ ;
- 10: move t to an existing cluster or new cluster  $C_j$  that maximize profit;
- 11: **if**  $C_i \neq C_j$  **then**
- 12: write  $\langle t, j \rangle$ ;
- 13: *moved* = **true**;
- 14: **until** not *moved*;

# **CLOPE Summary**

Input: dataset, repulsion, maximum number of clusters

- Phase 1
- 1. For each transaction, add it to a new cluster or to an existing one such that the profit is maximized
- Phase 2
- 1. For each transaction, try to move it to another cluster and do it if this maximizes the profit
- 2. Repeat 1. until all the transactions remain in the same cluster

#### **TX-MEANS**

- A parameter-free clustering algorithm able to efficiently partitioning transactional data automatically
- Suitable for the case where clustering must be applied on a massive number of different datasets
  - E.g.: when a large set of users need to be analyzed individually and each of them has generated a long history of transactions
- TX-Means automatically estimates the number of clusters
- TX-Means provides the **representative transaction** of each cluster, which summarizes the pattern captured by that cluster.





#### How It Works 2/3



#### How It Works 3/3

• Clusters

• Representative Baskets

#### **TX-Means Algorithm**



- Else
  - add B,r to the clustering result C;
- Return C;

#### **Bisecting Schema**

#### **BISECTBASKET(B: baskets):**

- SSE <-- inf;
- r1,r2 <-- select random initial baskets in B as representative;</li>
- While True:
  - C1,C2 <-- assign baskets in B with respect to r1,r2;
  - r1\_new <-- GETREPR(C1); r2\_new <-- GETREPR(C2);
  - SSE\_new <-- SSE(C1,C2,r1\_new,r2\_new);
  - If SSE\_new >= SSE Then:
    - Return C1,C2,r1,r2;
  - r1,r2 <-- r1\_new,r2\_new;</pre>

overlap-based distance function: Jaccard coefficient

#### **Get Representative Baskets**

#### GETREPR(B: baskets):

- I <-- not common items in B;
- r <-- common items in B;</li>
- While I is not empty:
  - Add to r the items with maximum frequency in I;
  - Calculate the distance between r and the baskets in B;
  - If the distance no longer decreases Then:
    - Return r;
  - Else
    - remove from I the items with maximum frequency;
- Return r;

overlap-based distance function (Jaccard coefficient)

- Clustering of a big individual transactional dataset B.
- TX-Means is scalable thanks to the following sampling strategy.
- Sampling strategy:
  - Random selection of a subset S of the baskets in B;
  - Run of TX-Means on the subset S and obtain clusters C and representative baskets R;
  - Assign the remaining baskets B/S to the clusters C using a nearest neighbor approach with respect to the representative baskets R.

# References

- Guha, S., et al. ROCK: A robust clustering algorithm for categorical attributes. 2000.
- Yang, Y., et al. CLOPE: a fast and effective clustering algorithm for transactional data. 2002.
- Guidotti, R., et al. Clustering individual transactional data for masses of users. 2017.

A and	nam Maana	DP ELL EG MO	S.CMU.EDU	
Scho	ol of Computer Science, Carnegie Mellon Universit;	, Pittsburgh, PA 15213 USA		
			and the second of the	A、15-464-44-5-737月1日年间第二
	Abstract	lutions for these problems. Speed is great	ly improved	States and a state of the
	Despite its nonularity for general clustering.	by embedding the dataset in a multiresolu	tion kd-tree	The state of the s
	K-means suffers three major shortcomings;	and storing sufficient statistics at its node analysis of the centroid locations allows	s. A careful for segmet-	
	it scales poorly computationally, the num- ber of clusters K has to be supplied by the	ric "proofs" about the Voronoi boundarie	s, and (un-	
	user, and the search is prone to local min-	like all of (Deng & Moore, 1995; Zhang Moore, 1999)) there is absolutely no an	et al., 1995; proximation	A STATE OF A
	ima. We propose solutions for the first two problems and a partial remedy for the third	anywhere in the computation. An additio	nal geomet-	A DECEMBER OF STREET
	Building on prior work for algorithmic accel-	ric computation, blacklisting, maintains a those centroids that need to be considered	list of just	
	eration that is not based on approximation, we introduce a new alcosithm that officiently	region (Pelleg & Moore, 2000). Blacklistin	g is not only	
	searches the space of cluster locations and	extremely fast but also scales very well wi ber of centroids, allowing tractable 10, 000,	th the num-	
	number of clusters to optimize the Bayesian Information Criterion (BIC) or the Akaika	rithms. This fast algorithm is used as a bu	ilding-block	
	Information Criterion (AIC) measure. The	in X-means: a new algorithm that quick K. It sees into action after each run of K-r	y estimates	
	innovations include two new ways of exploit- ing cached sufficient statistics and a new very	ing local decisions about which subset of	the current	A LEAD ALL AND SECONDER
	efficient test that in one K-means sweep se-	centroids should split themselves in order the data. The splitting decision is done by	to better fit	
	lects the most promising subset of classes for refinement. This gives rise to a fast, statis-	the new reasonable cards and	w	aci steinbach. Anuj karpathe vipin k
	tically founded algorithm that outputs both	ho su	Information Systems Vol. 25, No. 5, pp. 345	-356, 2000
	the number of classes and their parameters. Experiments show this technique reveals the	te Pergamon	© 2000 Elsevier Science Lid. All righ Printed in Gro	is reserved eat Britain
	true number of classes in the underlying dis-	sir ta	0306-4379/00 \$20	00.0 + 00.0
	tribution, and that it is much faster than re- neatedly using accelerated K means for dif	na		
	ferent values of $K$ ,	W ROCK: A ROBUST CLU	STERING ALGORITHM FOR CATEGORIC	AL CALCULATION AND AND AND AND AND AND AND AND AND AN
		tr: te	ATTRIBUTES <sup>†</sup>	The second se
1. I	ntroduction	be		A DECK OF THE OWNER WATER OF THE OWNER OF
K.m	eans (Duda & Hart, 1973; Bishop, 1995) has long	wi Sudipto Guha wi	<sup>1</sup> , RAJEEV RASTOGI <sup>2</sup> , and KYUSEOK SHIM <sup>3</sup>	A ANTON
been	the workhorse for metric data. Its attractive-	m <sup>1</sup> Stanfe	ord University, Stanford, CA 94305, USA	A STATE OF A
ness	lies in its simplicity, and in its local-minimum ergence properties. It has, however, three main	<sup>-</sup> Bell I. <sup>3</sup> Korea Advanced Institute of Science	aboratories, Murmy Hill, NJ 07974, USA and Technology and Advanced Information Technology Research	Center,
short	tcomings. One, it is slow and scales poorly with	2.	Taejon 305-701, Korea	
			31 May 2000)	
C	lustering Individual Tran	sactional Data for Mas	ses of Users	
	Riccardo Guidotti Au	nna Monreale M	irco Nanni bution patterns in the underly	ing
IS	II-CNR & University of Pisa, Italy Univ	ersity of Pisa, Italy ISTI-	CNR, Pisa, Italy	han
	riccardo.guidotti@isti.cnr.it anna.n	nonreale@di.unipi.it mirco.	nanni@isti.cnr.it that was distance hat was not	into
	Fosca Giannotti	Dino Pedreschi		
	Fosca Giannotti ISTI-CNR, Pisa, Italy	Dino Pedreschi University of Pisa, Italy	CLOPE: A Fast and Effect	ive Clustering Algorithm for
	Fosca Giannotti ISTI-CNR, Pisa, Italy fosca.giannotti@isti.cnr.it	Dino Pedreschi University of Pisa, Italy dino.pedreschi@di.unipi.it	CLOPE: A Fast and Effect Transact	ive Clustering Algorithm for tional Data
ABST	Fosca Giannotti ISTI-CNR, Pisa, Italy fosca.giannotti@isti.cnr.it	Dino Pedreschi University of Pisa, Italy dino.pedreschi@di.unipi.it from other users. This requires tha included in any data mining mathe	CLOPE: A Fast and Effect Transact Yiling Yang Xudo	ive Clustering Algorithm for tional Data ong Guan Jinyuan You
ABS1 Mining making	Fosca Ciannotti ISTI-CNR, Pisa, Italy fosca.giannotti@isti.enr.it TRACT z a large number of datasets recording human activitie genes of individual data is the key enabler of a new w	Dino Pedreschi University of Pisa, Italy dino.pedreschi@di.unji.it from other users. This requires tha included in any data mining methe rave the necessity to automatically capt	CLOPE: A Fast and Effect Transact Yiling Yang Xudo Dept. of Computer Science & Engi Schandrai 20	ive Clustering Algorithm for tional Data ng Guan Jinyuan You neering, Sharghai Jiao Tong University 2003, P.R. China
ABS1 Mining making of pers	Fosca Giannotti ISTL-CR, Pisa, Italy fosca.giamotti@isti.cnr.it TRACT z a large number of datasets recording human activitie genes of individual data is the key enabler of a new maized knowledge-based services. In this paper we fo	Dino Pedreschi University of Pisa, Italy dino.pedreschi@di.unipi.it from other users. This requires tha included in any data mining methe rave the necessity to automatically capt vidual behaviors. Due to the potent ocus of the potent of the potent	CLOPE: A Fast and Effect Transact Yiling Yang Xudo Dept. of Computer Science & Engi Shanghia, Z +85-21	ive Clustering Algorithm for tional Data org Guan Jiryuan You wening, Shimghai Jiao tong University 00030, P.R. China - 22681638
ABS1 Mining making of pers on the large n	Fosca Giannotti ISTI-CNR, Pias, Italy fosca giannotti@isti.cnr.it reACCT a large number of datasets recording human activitie gense of individual data i the key enabler of a new o nanized knowledp-based services. In this paper we for problem of clastering individual transactional data i aus of users. Transactional data is a very pervarbar ket	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.unipi.it from other users. This requires tha s for included in any data mining methe the necessity to automatically capt vidtal behaviors. Due to the potent or a (e.g. users in nowadays massive ay of generally unseable to determine in	CLOPE: A Fast and Effect Transact Viling Yang Xudo Dept. of Computer Science & Engi Shanghia, 2 (yang-yi, guan-xd.)	ive Clustering Algorithm for tional Data ng Guan Jinyuan You werking, Bhanghal Jaa Tong University 20030, P. 2 work Jack State State State work Jack State S
ABS1 Mining making of pers on the large n inform	Fosca Giannotti ISTI-CMP, Tiss, Italy foracagiannotti@isti.cn.it (Harge muber of dalasch rocording huma draiting) stamo of individual dais ibte koy mubar activitis, stamo of individual dais ibte koy mubar activitis, stamo of the stamo individual transactional data i asso of users. Transactional data is a very pervaries tek ion that is collected by several service, distensioned	Dino Pedreschi University of Pisa, Italy dino.pedreschi@di.unipi.it from other users. This requires thu included in sny data mining meth- te necessity to automatically copi visual behaviors. Due to the potent or a (e.g. users in norwadays maasive yo a dot generally unfoasible to determine in parmeter configuration for each of	CLOPE: A Fast and Effect Transaci Yiling Yang Xudo Dept. of Computer Science & Engi Shangtai, 2 (yang-yi, guan-xd, y	ive Clustering Algorithm for tional Data Org Guan Jiryuan You Deeming, Shenghai Jao Tong University 2000, P.R. Chana 2006 1030 - 2006 1030 you-Jy/@ccs.sjtu.edu.cn
ABS1 Mining makinj of pers on the large n inform huge p tering	Fosca Giannotti ISTI-CMP, Fiss, Italy fosca giannotti@isti cnr.it wa large number of datasets recording human activitie geme of individual data is the key enabler of a new problem of chattering individual transactional data is undirate laworkega-based services. In this paper we for problem of chattering individual transactional data is in that is collected by served averycies, othen invol- ced of user. We propose transact, aparameter-for- oods of user. We propose transact, aparameter-for-	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.unipi.it from other users. This requires thu- included in any data mining methe area the necessity to automatically cap for a generally uniscalito to determine in garantee configuration for each of cha- characterizies of the data tunder a characterizies of the data tunder a	CLOPE: A Fast and Effect Transact Yiling Yang Xudo Dept. of Computer Science & Engi Shanghai, 2 (yang-yi, guan-xd, y	ive Clustering Algorithm for tional Data ong Guan Jinyuan You wening, Shimghai Jiao Tong University 00030, P.R. China «25681638 you-jy}@cs.sjtu.edu.on
ABS1 Mining making of pers on the large m inform huge p tering in a co	Fosca Giannotti ISTI-CNR, Fisi, Iahy fosca giannotti@isticnr.it gange number of datasets recording human activitie genue of individual data is the key enabler of a new unilate lavorkepa-based services. In this paper ve fo problem of clustering individual transactional data is and of substrating individual transactional data is avery pervasive kit aution that is collected by several services, often invol- olo of users. Ver propose transara, a parameter-free- algorithm able to efficiently partitioning transactional patiety automatic way. Transacti si designed for the	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.wnip.it from other users. This requires that included in any data mining methe vidual behavirs. Due to the potent vidual behavirs. Due to the potent of generally unitashike to determine in grammeter configuration for each of four data in mining methods that adju- data.	CLOPE: A Fast and Effect Transact Viling Yang Xude Dept. of Computer Science & Engi Shanghia, 2 (yang-yl, guan-xd, y	tive Clustering Algorithm for tional Data ung Guan Jinyuan You werking. Simplati Jiao rang University 00030, P.R. Orina -25261638 -25261638 -000-jt/j@ccs.pitu.edu.cn
ABST Mining making of pers on the large n inform huge p tering in a co where dataset	Fosca Giannotti ISTI-CMP, Tiss, Italy foscagiannotti@bsticnrit et alorge sumber of dalandet recording human scheme steme of individual dala is the koje melhar scheme onalized knowledge-based services. In this paper we to poblem of chastering individual transactional data i asso of users. Transactional data i is a very pervasive ki alorth Mai to cellected by several services, dene invol- ools of users. We propose transma, a parameter-fore diapetitha mable of discritosy participation transactional updetely automatic way. Transacti is designed for the duriner munt be applied on a massive number of diff	Dino Pedreschi University of Fixa, Italy dino,pedreschi@di.uniyi.ti form abre user. Thir require tha included in surg data mining rethe the necessity to automatically cap visual behaviors. Due to the poter in parameter configuration for each of cast of cast of the state under a parameter configuration for each of chan- chan changes and the state under a chancetariation of the dataset under a final heat paper we focus on the pa- tional changes and the state under the state of the state of the state under the state of the state under the state of the state of the state of the state under the state of the state of the state under the state of the	CLOPE: A Fast and Effect Transact Wiling Yang Xud Dept. of Computer Science & Engr Shangha, 2 (yang-yi, guan-xd, y ABSTRACT This pager studies the professor of categorial data clustering.	tive Clustering Algorithm for tional Data ang Guan Jinyuan You sering, Shinghai Jiao Tong University 0000, P.R. China 25261638 you-Jy/@ccs.sjtu.edu.cn The Largettern [13] algorithm groups large estatories address by interver optimization of a global cetterion function. The China State State State State State State State State State State State State Stat
ABST Mining makin of pers on the large n in form huge p tering in a co where dataset individ	Fosca Giannotti ISTI-CMP, Fiss, Italy fosca giannotti@isti cnrit wa large number of datasets recording human activitie genes of individual data is the key enabler of a new formational norwelge-based services. In this paper we for problem of chattering individual transactional data is units that is collected by several services. And in paper location that is collected by several services, othen invol- cods of uner. We propose transact, aparameter-for- duerts was an entered by a several services, othen invol- cods of uner. We propose transact is designed for the advertime mathe experised on a master number of diffi- s, for instance when a large set of users need to be and will y and each of them has generated a sole plaisary of t	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.unipi.it from other users. This requires thu- included in any data mining methe the necessity to automatically cap of the necessity to automatically cap of a generally uniscalithe to determine in parameter configuration for each of chan- from the necessity of automatically cap percently uniscalithe to determine in parameter configuration for each of chan- chan- torial cap with the second second second second the second second second second the second second second second second the second second second second second the second second second second second second the second second second second second second the second second second second second second second the second second second second second second second second the second second second second second second second second second second termine second seco	CLOPE: A Fast and Effect Transact Wing Yang Xudo Dept. of Computer Science & Engi Shangha, 2 (yang-yi, guan-xd, y BESTRACT This paper studies the problem of categorical data clasteristical by his queuesionality and area volume. Statistic method	tive Clustering Algorithm for tional Data Data Jinyuan You wenny, Shimghai Jiao Tong University 00030, P.R.China -25861638 you-jy}@cs.sljtu.edu.cn The Largettem [13] algorithm groups large consported databases by literative optimization of a global criterion function. The criterion function is based on the notion of <i>large</i> than that is the criterion function is based on the notion of <i>large</i> than that is the criterion function is based on the notion of <i>large</i> than that is the criterion function is based on the notion of <i>large</i> than that is the criterion function is based on the notion of <i>large</i> than that is the criterion function is based on the notion of <i>large</i> than that is the
ABST Mining making of pers on the large m inform huge p terring in a co w here- datast individ actions	Fosca Giannotti ISTI-CMP, Tsa, Italy fosca giannotti(ejisti.cm2) ta large number of adaust: rocording human activit esme of individual dais in the koy musher activit genes of individual dais in the koy musher activit poblem of clustering individual transactional data I asso of users. Transactional data is a very pervavive ta distribute the second services, discussion data algorithm able to efficiently partitioning transactional algorithm able to efficiently partitioning transactional patiety automatics. Just 2018 and 2018 and 2018 and algorithm able to efficiently partitioning transactional data I and a large iet of users member of diffi- ta deep sceptimentations on obtin rel and on ble hand multi yan decis of them has generated a long history of to A deep sceptimentations on obtin rel and synthetic dat	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi.t from obter user. This requires the included in any data mining nether the necessity to automatically cap visual behaviors. Due to the poler energy of a generally unfeasible to determise in provide mining methods that dip data characteristics of the dataset under a personalized patterns from transact rent in bia paper we focus on the pa- tional clustering for a large number collection of transaction, transacti collection of transaction, transacti collection of transaction, transacti content issue 1920. In the data turber a common issue 1920. In the state i	CLOPE: A Fast and Effect Transact Viling Yang Xudy Dept. of Computer Science, & Evil Yang Yang Yang Yudy (yang-yi, guan-xd, y ABSTRACT This paper studies the problem of categorical data clustering operation for manascional data clustering presiding for manascional data clustering dimensionality and page volume. Starting from a beneficier method wendo a sover that and	tive Clustering Algorithm for tional Data ung Guan Juryuan You henring: Bunghal Jao Tong University 00000 the second second second second 00000 the second second second second 00000 the second second second second second 00000 the second second second second second second second 00000 the second second second second second second 00000 the second second second second second second second 00000 the second second second second second second second 00000 the second second second second second second second second 00000 the second second second second second second second second second 00000 the second se
ABS1 Mining making of pers on the large m inform huge p tering in a co w here: dataset indrivi actions shows w of diffe	Fosca Giannotti ISTI-CMP, Tiss, Italy foscagiannotti@bsticnrit eta large sumber of dalaselt recording human serio itali stano finikriola dala isthe kope number a civilia tesse of individual dala isthe kope number a civilia susso futures. Transactional data ista very pervasive ki asso of users. We propues transma, a parameter-fore olosi of users. We propues transma, a parameter-fore dimetrim units de uppled on a massive number of diffu- diaterism must be appled on a massive number of diffu- ndering must be appled on a massive number of diffu- tality and out-of them has generated a kop hythetic data the paratical effectiveness of transmission and hythetic data the paratical effectiveness of aroness for the mass charac- ter applement of the stress of the stress of the transmission.	Dino Pedreschi University of Pisa, Italy dino,pedreschi@di.uniyit. form aber user. This require the included in sup data mining rethe the necessity to automatically cap visital behaviors. Due to the poter in parameter configuration for each of class- form of the second second second second chan- chan between the second second second chan- the second second second chan- the second second second chance second chance second second second chance second chance second second chance second second countering groups of homogeneous to covering covering covering second second second covering covering second sec	CLOPE: A Fast and Effect Transact Wing Yang Xud Dept. of Computer Science & Engr Stangard (yang-yi, guan-xd, y Charge native the profession of experience of the computer memory and the second of the charge science of the dimensionality and large volume. Starting from a burvisite method increasing the heavy for the charge science of the burvisite dimensionality and large volume. Starting from a burvisite method increasing the heavy for the charge science of the charge science of the	the Clustering Algorithm for tional Data ung Guan Jinyuan You sening. Shinghai Jiao Tong University 0000, P.R. China 25061838 you-Jy/@ccs.sltu.edu.cn The Lengthem [13] algorithm groups large estapoided dathems for the senior softmariation of a global criterion function. The trainerine softmariation of a global criterion function. The trainerine softmariation of a global criterion function. The trainerine softmariation of a global criterion function function function in much faster than how how of criterion functions function function in much faster than how how for criterion functions function function in much faster than how how for criterion functions differed functions in much faster than how how for criterion functions differed functions in much faster than how how for criterion functions differed functions in much faster than how how for criterion functions differed functions in much faster than how how for criterion functions differed functions in much faster than how how for criterion functions differed functions in much faster than how how for the criterion functions differed functions in much faster than how how for the criterion functions differed functions functions differed functions differed differences for the criterion functions differed functions differed functions in much faster than how how for the criterion functions differed functions and the criterion functions differed functions differed functions and the criterion function functions differed functions differed functions and the criterion functions differed functions differed functions and the criterion functions differed functions differed functions differed functions and the criterion function function functions differed functions differe
ABST Mining making of ping of the large n inform bage p tering te	Fosca Giannotti ISTI-CMP, Fiss, Italy fosca giannotti@istir.cnr.it ga large number of datasets recording human activitie geme of individual data is the key enabler of a new formational barowledge-based services. In this paper we for protor on ext. Thereing matrixing the transmission of the protor of the transmission of the transmission of the transmission of the transmission of the spectrum abit to efficiently partitioning in priority of the spectrum abits to efficiently partitioning the particul efficient of the transmission of the transmission of the particul affectives of transmiss for diffectives. First rend personal datasets, and suggests that breneas are trent personal datasets, and suggests that breneas are trent personal datasets, and suggests that breneas are trent personal datasets, and suggests that breneas are transmissing methods in terms of gaulary and efficiency. Fit	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.unipi.it from other users. This requires thu- included in any data aming methe area the necessity to antomatically com- legation of the second second second second gate and second second second second second gate and second second second second gate and second second second second gate and second second second second collection of transactions, transacti second sec	CLOPE: A Fast and Effect Transact Wing Yang Watd Dept of Computer Science & Engi Stangals, (yang-yi, guan-xd, y Disputer the problem of categorical data clantering, repetially for transactional data clanterized by high demonstrating the Republic-world ratio of the clanter histogram, we sciable, while how guiter officies. We demonstrate here performance of our algorithm on two real world datasets, and compare CLOP with the state-off an learning.	the Clustering Algorithm for tional Data ung Guan Jinyuan You wenng, Shamphai Jiao Tong University 20030, P.R.China -25881633 you-jr)@ccs.slfu.edu.cn The Largettem [13] algorithm groups large categorieal databases by itersity oglimization of a global criterion function. The criterion function is based on the notion of <i>large</i> them tai to the scale of the scale of the scale of the scale of the scale function is much fauer than those local criterion function. The criterion function is based on the notion of <i>large</i> them tai to the scale of the scale of the scale of the scale of the scale function is much fauer than those local criterion functions function to of proview similarities. This global approach makes Largettem very unitable for clustering large categorial databases.
ABST Mining making of pers on the large n inform huge p tering in a co w here inform huge p tering in a co w here inform inform inform inform inform inform huge p tering inform inform inform inform inform inform huge p tering inform inform inform huge p tering inform inform huge p tering inform inform huge p tering inform huge p tering inform inform huge p tering inform inform huge p tering inform huge p tering huge p tering inform huge p tering inform huge p tering inform huge p tering huge p huge p tering huge p huge p	Fosca Giannotti ISTI-CMP, Tsu, Italy forca giannotti(@isti.cm2i t Jarge number of dalaust recording human sciriniti sums of individual dala in the key makes and the second sums of individual dala in the key makes and the second problem of chattering individual transactional data 1 mass of users. Transactional data is a very pervavive the olosi of mars. We propose transma, a parameter-free problem of chattering way. Threasen is designed for the dustering must be applied on a musive number of diffu- dual effect of the second services. The second is and analy and each of them has generated a long history of of a deep septementation to obta frei and synthetic dat the practical effectiveness of transmos for the mass change sent is personal dataset, and suggest but termeons not users is personal car assistant application based on term	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi.t from obre user. This requires the included in any data mining meth- the necessity o automatically capi the the necessity o automatically capi and the second second second second parameter configuration for each of classical second second second second parameters of the second second second parameters of the second second second contrast groups of homogeneous to covering groups of homogeneous to	CLOPE: A Fast and Effect Transact Willing Yang Kudo Dept. of Computer Science & Engi Shang and Willing Yang Kudo Science & Computer Science & Science (yang-yi, guan-ya, computer Science & Science Appendix for Image Volume Starting from a bencistic method proteining for Image Volume Starting from a bencistic method of Image Volume Science of the cluster biologen, we science with the state-of-ant algorithm.	the Clustering Algorithm for tional Data mg Gum _ Jnyuan You mg Kum _ Jnyuan You soot and the soot and the soot soot and the soot and the soot soot and the soot and the soot mathematical soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot soot and the soot and the soot and the soot and the soot soot and the soot and the soot and the soot and the soot soot and the soot and the soot and the soot and the soot and the soot soot and the soot and the
ABST Mining making or me of the inform inform inform there terring in a co w here: didn's activity act	Fosca Giannotti ISTI-CMP, Tiss, Italy forcagiannotti@biti.cnrit FACTOR 24 ale anumber of dalated recording human kar ohisi ante of individual dala inthe log-malkar afor dis- poletion of chalterial to cording human kar of a new onalized howe-hege-based services. In this paper were poletom of chalterial in a very pervasive kit asso of users. We propose Ermans, a parameter-for- gogenthum able of deficiently particioning irransactional dutatient must be applied on a massive number of diff algorithm able of deficiently particioning any short of using anomatic way. Ermans is designed for the participation on both real and synthetic dut the particle effectiveness of comeans for the mass chalter and a personal context, and anggests that Lermens ou cisisting methods in terms of quality and efficiency. Fit NCDDUCTION	Dino Pedreschi University of Fisa, Italy dino.pedreschi@di.uniyit form abter teasr. This requires the included in sup (ala anising erethe the necessity to automatically cap visual behaviors. Due to the poter of e.g. users in nowadays massive sy dof generally uniesable to determine in parameter configuration for each of cha- forwa data mining methods that adju a characterizic of the diaset under a configuration of transactures are configuration of transactions. transacture is covering groups of homogeneous to covering groups of homogeneous to covering groups of homogeneous to even to icade to large unrule that ai not icade to large unrule that and sen icade to large unrule the covering groups of homogeneous to covering groups of homogeneous to the state on icade to large unrule that ai not icade to large unrule the covering require eith the covering require eith the covering require eith the discont icade to large unrule the covering require eith the cristing to the state on icade to large unrule the covering require eith the cristing to a starge unrule the covering require eith the cristing to the state of the targe to covering require eith the cristing the cristing the state to covering requires the cristing the cristing the target the to the state of the st	CLOPE: A Fast and Effect Transact Wing Yang Yang Bey, d Computer Science, 8 Engr Stangar, 8 (yang-yi, guan-xd, y Character and the standard for the standard management of the standard for the standard for a source algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and every a novel algorithm or two real world datasets, and engene CLOPE with the stand-for an experime.	bit Clustering Algorithm for tional Data Uny and You see the second
ABS7 Mining of pers on the large en minors method restring in a co w here: dataset addition the of the shows in the of the shows in the show	Foreca Giannotti ISTI-CMP, They, Italy forecagiannoti/ejoiti.cm2i Partage number of dalatest recording human activitie genue of individual dais the key analytic of a new onalized toxwledge-based services. In this paper we for postbern of chartesting individual transactional data fa ass of users. Transactional data is a very pervariate the postbern of chartesting individual transactional data is a very service of the service of the service algorithm also to efficiently partitioning transactional patiety automatic way. Transans is designed for the chartering must be applied on a massive number of diffi- ing rimature when a large set of users model to be anai ually and each of them has generated a long lationy of its for practical efficiences of consons not be mass charter period data. The service of consons not be mass charter terrel personal datasets, and suggests that comeans ou setting methods in terms of quality and deficiency. Fits set a personal cart assistant application based on term <b>XTUDUCUTION</b>	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.unipi.it from other users. This requires thu included in any data mining methe wave the necessity to anisomatically copi (e.g. users: In now adays maintigation of the generally unissatible to determine in garantee configuration for each of chan- forw data mining methods that adju characterization of the data under a collection of transactions, transacti estimation of transmissions, transacti that is not totale to large user b repeatedly applying the existing pro- lisions of different datasets – which i arge population of users - is simp data, problem, e.c. the sparate indivisions	CLOPE: A Fast and Effect Curansact With the second	But Clustering Algorithms for tional Data ung Gum Jinyuan You wenng. Sumpai Jiao Tong University 20030, P.R.China - 25881633 you-jr)@ccs.sflu.edu.cm >>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>>
ABS1 Mining making of pers on the large n ministress the second second terring, a bener terring, a bener distant actions above i of diff formas w prec distant formas w prec distant formas w prec distant formas fo	Foreca Giannotti ISTI-CMP, They, Italy forecagiannotti@ehit.cm?it PL stage muther of datasets recording human activities stress of indivisional data is the key analysis of the second stress of indivisional transmission and that asso of users. Transactional data is a very pervariave ta individue to the second services, then invol- ools of uners. We propose transma, sparameter-free problem of clustering individual transmost, othen invol- ools of uners. We propose transma, sparameter-free approxima labe to efficiently partitioning in transactional dustering mut be applied on a musive number of diffu- to primatic effective second services of the smark ululy and each of them has generated a long history of of a deep experimentation on obtin rel and synthetic data the present datasets, and suggests that renseas on existing methods in iterms of quality and efficiency. Fir and approximation consider application based on their NTEDUCTION	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi.ti form obre user. This requires the included in say data mining methe the necessity to automatically capi visual behaviors. Due to be poter ing parameter configuration for each of deal- bear state mining methods that algo persently uniesable to determine in parameter configuration for each of deal- personalized patterns from transpor- tional clustering for a large number rent in this paper we focus on the p- tional clustering for a large number rent and the state of the state of the covering groups of homogeneous to rent and the state of the state of the the state of the state of the state of the the state of the state of the state of the the state of the state of the state of the the state of the state of the state of the the state of the state of the state of the the state of the state of the state of the transformed dataset, as mere char in the state of the state of the state of the state of the transformed dataset is mere char in the state of the state of the state of the state of the intermed dataset is mere the state of th	CLOPE: A Fast and Effect Transact Willing Yang	bit Clustering Algorithms for tional Data Use of the second s
ABST Mining making or the here are inform per print terring; in a co w here: dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset instant dataset dataset instant dataset d	Fosca Giannotti ISTI-CMP, Tiss, Italy foscagiannotti@biti.cm?it Part and the second program of the second second program of the second program of the second second program of the second second second program problem of chaltering individal transactional data i sass of users. We propose Ermans, a parameter-fore- problem of chaltering individal transactional data i sub of the second second second second second and the second second second second second second application while of the second second second second application when a large set of users need to be and duriner must be applied on a masser ender to difficult the practical effectiveness of cranes need to be and a door perpendention on both red and synthesis data the practical effectiveness of cranes need to be made in a personal coart assistant application based on terms CHCDDUCTON With distantional transaction of the transaction of the brackwardshiel bette for our energy digital benchospies. Thanks to these data, human acti- ding observable meaning, busing thirdle add, predicti- ding observable meaning, busing thirdle add, predicti- ding observable meaning, busing thirdle add, predicti- ding observable meaning down and the second	Dino Pedreschi University of Fixa, Italy dino pedreschi@di.unipi.it form diber torsen. This requires the host of the torsen. This requires the the included in may data requires the the included in may data requires the the included in may data requires the torsen the torsen. This requires the torsen the torsen the torsen to the second second second second second data characteristics of the dataset under a case personalized patterns from transat characteristics of the dataset under a construction of transactions. Its mascul- correcting groups of homogeneous to collection of transactions, remanci- tic constructions, dataset, or and the transactional clustering require eith that is not actor to large user b remote the existing the existing transactional dataset, as mass char transactional dataset, as mass char the been addressed in the context of nos	CLOPE: A Fast and Effect Transact Wing Yang Yang Budded Computer Science & Engr Science (yang-yi, guan-xd, y Assert and the problem of categorical data clastering specially for transactional data characterized by Maj for creating the heightbo-widt fination of the special data characterized by Maj for creating the heightbo-widt fination of the special data characterized by Maj for creating the heightbo-widt fination of the special data characterized by Maj for creating the heightbo-widt fination of the special data characterized by Maj for creating the heightbo-widt fination of the special data characterized by Maj for advantage that the special data characterized by Maj for advantage that the special data characterized by the special data characterized by Maj for advantage that the special data characterized by the special data characterized by Maj for advantage that the special data characterized by the special data characterized by Maj for advantage that the special data characterized by the special data characterized by Maj for advantage that the special data characterized by the special data characterized by Maj for advantage that the special data characterized by the special data characterized by Maj for advantage that the special data special data special for advantage that the special data special data special data special for advantage that the special data special data special data special for advantage that the special data special data special data special for advantage that the special data special da	bit Clustering Algorithms for tional Data Uny and You hereings, Shinghai Jao Tong University 00030, P.R. Chana 2008103, B. Chana 200810, Chana 20
ABST Mining making of pers on pers on pers on pers in a con- mission of the mining of	Foreca Giannotti ISTI-CMP, They, Italy forecasiannotti(e)isti.cm2i Harris and the second second second second sense of individual datis in the key mainteent of a new omained toxwledge-based services. In this paper we for poblem of clustering individual transactional data i ass of users. Transactional data is a very pervariave ta distribute the second services, the second services, discuing and the second services, the second services, discuing we algorithm also to efficiently partitioning transactional algorithm also to efficiently partitioning transactional data is a second second service of the second second second second second second be and algorithm also to allog second second be and algorithm also generated a long histight of of A deer segmentation on both real and histight of of a deer second second second second second be and and second second second second second transaction of the second second second be and algorithm effection the segments and algorithm of the second second second second second second transaction second second second second second second transaction second second second second second second transaction second second second second second transaction second second second second second second transaction second second second second second transactio	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi.t from obter user. This requires the included in any data mining nether the necessity to automatically capi visual behaviors. Due to the poler energy of a generally unfeasible to determine in provide noise methods on the state provide the included in the state of the provide the state of the state of the provide the state of the state of the provide the state of the state of the provide patterns from transact energy of the state of the state of the provide patterns from transact content state patterns from transact content state patterns from the state of the state of the state of the state of the provide patterns from the state of the provide states of the state of the state common states [19]. In the state is the state of the state of the state the state of the state of the state of the problem to design parameter able. The problem to design parameter able.	CLOPE: A Fast and Effect Lansacian Willing Yang Yung Yung Yang Yung Yang Yang Yang Yang Yang Yang Managara Ang Yang Yang Ang Yang Yang Ang Yang Yang Ang Yang Yang Managara Ang Yang Ang Yang Managara Ang Yang Managara Ma	Biotechnic Control (1) (1) (1) (1) (1) (1) (1) (1) (1) (1)
ABS1 Mining of pers on per- son per- ning of pers of person with the second sec	Foreca Giannotti ISTI-CMP, They, Italy forecagiannotti@ebit.cm?it exact_innotti@ebit.cm?it exact_innotti@ebit.cm?it exact_innotti@ebit.cm?it issues of individual is the key means activities asso of users. Transactional data is the key means observed by several services, the this paper we per poblem of clustering individual Transactional data in that is collected by several services, often invol- obol of users. We propose transma, a parameter-free problem of clustering individual Transactional data is a service of the several services, often invol- obol of users. We propose transma, a parameter-free disputing must be applied on a musice cambre of diffu- tion transce when large set of users more to be and multy and each of them has generated a key history of of the parameter of the severation of the mass channel on the parameter and section of the mass channel on the exact a personid or are satisfue application based on them <b>CHCDUCTION</b> of disrupter effect of our abways-connected nocicity in influence. The history bases that the parameter of a severation of the parameter of the seven set of the mass channel on the provide of the parameter of the seven shall be appreciated or and sever disputs the observables. Thanke users that has a post of the parameter of the severation set of the seven shall be appreciated or and sever disputs the observables. Thanke users that has a post of the parameter observables one take the parameter of the severation of the parameter observables one take the parameter of the severation of the parameter observables one take the parameter of the severation of the parameter observables one take the parameter ob	Dino Pedreschi University of Pisa, Italy dino, pedreschi (jedi, uniyi, it- form ofter user. This requires the included in say data mining methe- the necessity to automatically capi the second second second second second equation of the second second second parameter configuration for each of chan- form in nowadays massive sy of equation of the second second second parameter configuration for each of chan- form of the second second second second parameter configuration for each of chan- form of the second second second second parameter configuration for each of chan- form of the second second second second parameter of the second second parameter of the second second second content groups of homogeneous to covering groups of homogeneous to covering groups of homogeneous to reposited second chartering require eff lange oppolation of users - is simp problem, i.e., the segmate individual transpectual dataset, a mean char line in the second second second parameter information of the chartering problems. University in the second second second second second parameter [22], which are perfec- lated second second second second second parameter [23] which are perfec- lated second seco	CLOPE: A Fast and Effect Transact Willing Yang	But Clustering Algorithms for tional Data Using Gum Juryuan You wayang Algorithms and algorithms (2003) PL China 2004) State (2004) State
ABST Mining making or the her the the her the inform minor the the the the the the the the the the	Fosca Giannotti ISTI-CMP, Tiss, Italy forcagiannotti@biti.cnrit Participation and the second second second second and the second second second second second second second second second second second second second problem of chaitering individal transactional data i saso of users. Vare propose transma, a parameter-fore- problem of chaitering individal transactional data i sato of users. We propose transma, a parameter-fore- diaterism must be efficiently particioning it ransactional adultering must be applied on a masser semimer of diffi- s, for indiance when a large set of users need to be and duriner must be applied on a masser semimer of the spectra second second second second second second second second second second second second second during second second second second second second sector proposed transmatic application based on terms CHCDDUCTION will show effect of our always-connected society in the bracknubs in behavior and second second second sector of the second second second second second mign observable, meannable, quantificate ada, product visual generation to the second second second second mign observable, meannable, quantificate ada, product visual second second second second second second second mign observable meannable, quantificate ada, product visual second second second second second second second markance of information thub, for the tem paratheres and for the second second second second second second second sectors (to thanket), i.e., a special kind of calegorical models and second second second second second second sectors (to thanket), i.e., a special kind of calegorical second second second second second second second second second sectors (to thanket), i.e., a special kind of calegorical second second second second second second second second second second second second second sectors (to thanket), i.e., a special kind of calegorical second second second second second second second second second second sec	Dino Pedreschi University of Fixa, Italy dino pedreschi@di.unipi.it from other users. This requires the home of the intervent of the second research of the second research of the intervent of the second research of the form other users. This requires the intervent of the second research of the form other users. This requires the intervent of the second research of the form other users in now adays making or the parameter configuration for each of chan- chance-training of the second research of the second research of the second research chance-training of the second research chance-training of the second research chance-training of the second research chance-training of the second research correcting groups of homogeneous to correcting groups of homogeneous to correcting groups of homogeneous to research of the second chartering require eff that is not automatic, or an extrem that does not icule to large users here the second chartering requires the here and descend a means chart improbem i.e., the segments . Unlet of the chartering problems. Unlet of the chartering problems. Unlet of the chartering problems. Unlet on correct and groups and the training the correcting groups of homogeneous to one of the chartering problems. Unlet on correct and groups and the the training of the correcting groups of the the training of the the training of the chartering problems. Unlet on correct and groups and the the training the the the the correction group of the the training of the the training of the the training of the other chartering problems. Unlet on correct and the training problems. Unlet on correction groups and the the training of the	CLOPE: A Fast and Effect Transact Wing Yang Yang Gud, of Computer Science & Engr Yang Yang (yang-yi, guan-xd, y Assertion of the problem of categorical data clustering, specially for transactional data clustering by high problem of the problem of categorical data clustering, specially for transactional data clustering by high of increasing the heighbo-widt finates of the cluster histogram, we develop a noval algorithm on two real world datasets, and coppare CLOPE with the state-fort algorithm. <b>Exercise</b> Management of the state-fort algorithm. <b>Exercise</b> Management of the state-fort algorithm. <b>Exercise</b> Management of the state-fort algorithm. <b>Exercise</b> Management of the state-fort algorithm.	Bet Clustering Algorithms for the Clustering Subject to the second transformer of the second second second second transformer of the second second second second transformer of the second second second second second second second second second second transformer of the second se
ABST Mining in a second making of pers- on pers- on pers- ning of pers- ning of pers- ning of pers- ning of pers- ning of pers- ning of pers- distant actions a drow in a second actions a drow in a second actions a drow in a second drom in a second action	Fisca Giannotti ISTI-CRP, Tas, Italy foreaciannotti(e)isti.cn?t ENCL 1 stage number of adausts recording human activities y stage number of adausts recording human activities y stage number of adausts recording human activities asso of users. Transactional data is the year parent poblem of chartering individal transactional data I has of users. Transactional data is a very pervasive to obol of unret. We propose transact, a parameter free poblem of chartering users are services, diene invol- obol of unret. We propose transact, a parameter free human the applied on annaster number of diffi- tion status of when a generated a long history of to for instance when a large set of users and yntheic dat the parketia effectiveness of transact and the discinstry. Fin sense is devised and efficiency. Fin sense a personal cord assistant application based on terms <b>CTUDCUTCION</b> of dispublic reflective from theory-connected so tochy is in a value of a dispublic reflective of a relevance of the sense of the ana side efficient of the sense of the sense of the taber provide of an advance-connected so tochy is in the provende of another op-connected so tochy is in the provende of a dispublic reflective of a relevan- tioning observable, measurable, quantitable and, predict in available well, advergence operation of the ana S(d) addia a valued of a dispublic reflective of a coregories of another of the sense of the anatom and S(d) of data a value of a dispublic reflective of a coregories of another of the sense of the anatom and the S(d) of data a value of a dispublic reflective of a coregories of the sense of the sense of the sense of the sense of the anatom (to the back), i.e., a special time and S(d) of data a value of a dispublic reflective of the sense of the anatom (to the back), i.e., a special time and S(d) of data a value of a dispublic reflective of the sense of the sense of the sense of the sense of the sense of the sense of the anatom (to the back), i.e., a special time and S(d) of data a value of the sense of the sense	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi. timo bella selectional and the selection form offer user. This requires the included in any data mining meth- the necessity to automatically capi visual behaviors. Due to the poten- tion of the selection of the selection of parameter configuration for each of the selection of the selection of the period selection of the selection of the constraint of the selection of the selection of the selection of the selection of the constraint of the selection of the selection of the selection of the selection of the selection of the selection of the selection of the selection of the constraint of the selection of the selection of the constraint of the selection of the selection of the constraint of the selection of the selection of the transactional cludering require of the proputed typping the selection of the selection of the transactional datasets – which it happed problem to design parameter blee madessed in the context of magnitude to transaction data. To phole to transactional data. To applicate to the selection of the clustering problems. United applicate to transactional data. To applicate to transactional data. To applicate to the selection of the clustering the other applicate to the selection of the selectio	CLOPE: A Fast and Effect Lynamics. Willing Yang Micho Digit of Computer Science & Engl Willing Yang Micho Digit of Computer Science & Engl Willing Yang Micho Wang Yang Yang Micho Hang Yang Michowald Rate & Chancetrized And Science Proteining for Immaschical data Antonization for science in the Immaschical data Chancetrize And Bernemang the Angelia-owidh Rate & Chanceterized And Science Michowald And Science and Andrew Andrew Andrew Science and Andrew Andrew Andrew Andrew Science and Andrew Andrew Andrew Science and Andrew Andrew Andrew Science and Andrew Andrew Andrew Science and Andrew Andrew Science and Andrew Andrew Andrew Andrew Andrew Science and Andrew	Description of the second
ABST Minning of pre- son pre- ning of pre- ning of pre- ning of pre- son pre- son pre- son pre- son pre- son pre- son pre- son pre- ation of pre- son pre- ation of pre- ation of pre- ation of pre- son pre- ation of pre- son pre- bistory - bistory - - 	Fosca Giannotti ISTI-CMP, Tsu, Italy foracagiannotti@ehti.cmit et alage muther of dalanset recording human activities tesme of individual dais the key muther of a new omained knowledge-based services. In this paper we per pollem of clustering individual transmissional data i hass of users. Transactional data is a very pervaries the individual data is the key muther of a new pollem of clustering individual transmissional data in that is collected by servell arrives, often invol- oolo of users. We propose transmis, a parameter-free appender of the server arrives of the same individual action of the server and the server and the server arrives of the same data and the server arrives of the same data and a server and the server arrives of the same data and a server and the server arrives of the same data and a server and the server arrives of the same data and a server and the server arrives of the same data and a server and the server arrives of the same data and a server and the server arrives of the same data and a server and the server arrives of the server and and the server and arrives and angulation based on tene <b>XTEDUCIEND</b> Of dissiptive effect of our always-connected society is a server (or banker), c., a special idm on Sul of data and arrives (or banker), c., a special idm on Sul of data and arrives (or banker), c., a special idm on Sul of data and and the data of the same data and the server and the server form of sets of event data, and as the itemp partnase arrives (or banker), c., a special idm or and working measures (or an even arrives (or banker), c., a special idm or and working measures (or and server) and and the server and the server and the server of a server and the server of	Dino Pedreschi University of Pisa, Italy dino, pedreschi (jedi, uniyi) tiforo, aber user. This require it his included in surg data mining rether the necessity to automatically capi the second second second second second parameter configuration for each of classical second second second second control of the second second second transfer of the second second second second transfer of the second second second second second second second second second second transfer of the second second second second second transfer of the second second second second second transfer of the second second second second second second second second second second second second transfer of the second second second second second transfer of the second second second second second transfer of the second second second second transfer of the second second second second second transfer of the second second second second second transfer of the second second second second second second second transfer of the second second second second second second transfer of the second	CLOPE: A Fast and Effect Lycansaci Willing Yang	But Clustering Algorithms for the clustering Simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify and the simplify
ABS7 Mining of pers of the large or inform being in a co where dataset additional to the shower of diffi- formas we pers of the shower of the	Fisca Giannotii ISTI-CRP. They, laby foreaciannotii(qibiti.cm2i Parta) and the sember of dalatest recording human activities genus of individual tais the key analysis genus of individual tais the key analysis problem of charlengin (individual transactional data fa ass of users, Transactional data is a very pervariave tai ben and the sember of the sember of the sem- al tais of the sember of the sember of the sem- al tais of the sember of the sember of the sem- al tais of the sember of the sember of the sember algorithm and to efficiently partitioning transactional tails and the officiently partitioning transactional tails and the officiently partitioning transactional tails and the officiently applied to a massive number of diffi- ing transactive sember of the sem of the sema share there personal datasets, and suggests that <i>transacts</i> to the particul affectives of transact not be made with the particul affectives of transact not deficiency. Fit ial breakcourses of transact not be data to more <b>XTUDUCTUON</b> this displatic behindual to a a side effect of our ener- vision devices of the same to not part, con surfaces, each person generates more than SiGi of data avaanche of information that, for the more part on surfaces of the subscript, i.e., a special kind of calogorian in of soci of overlapsic versication in a particular to individual devices, the sema tain the subscript of the surface of the web pages visited in a horowarge session page tains during the string the subscript is a large of subscript, the web pages visited in a large of a distribution of soci of overlapsic section in a page of subscript. The web pages visited in a horowarge session page based vervices, and of more viscentified biccovertics.	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi.t form ofber user. This requires the included in any data mining metho the necessity o automatically capi visual behaviors. Due to the poten- ic (e.g. users in nowadays maxies ay of at generally uniteasible to determine in program and an implementation of the angle program of the second second second program of the second second second program of the second second second fragmentation of the second second program of the second second second fragmentation of the second second second fragmentation of the second second second covering groups of homogeneous to covering groups of homogeneous to the second second second second second fragmentation of the second second second second fragmentation of the second second second second fragmentation of the second second second fragmentation of the second second second second second fragmentation second s	CLOPE: A Fast and Effected Cursues           Compared and and and and and and and and and an	Bet Clustering Algorithms for the Clustering Subject to the second transformer and the second second second transformer and the second second second second second transformer and the second second second second second transformer and the second second second second second transformer and second second second second second transformer and seco
ABST Mining in making of pers- on per- on per- ning or in a co- min or in a co- in a	Fisca Gannotti ISTI-CRP, Tas, Italy foreaginnotti@ehit.crr.ht ISTI-CRP, Tas, Italy foreaginnotti@ehit.crr.ht Istage muther of datasets recording human activities y large number of datasets recording human activities asso of users. Transactional data is the year part of poblem of chattering individal transactional data I has of users. Transactional data is a very pervasive to obol of unret. We propose transact, a parameter free poblem of chattering users and activity of the chattering mut be applied on massive number of data the practical effectiveness of transaction of the trans and the second second second second second patients and the second second second second the practical effectiveness of transact and the first activity of the second second second transactivity effects of the second second second second second second second second transactivity effects and the second second transactivity effects of the second second transactivity effects of the second second transactivity effects and the second second transactivity of the second second second transactivity effects and the second second second transactivity effects and the second second second transactivity effects and the second second second transactivity of the second second second second transactivity of the second second second second second transactivity of t	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi. timo bella dina pedreschi@di.uniyi. timo bella dina sug data haning methe the necessity o automatically capi visual behaviors. Due to the poten- tical distribution of the second second second parameter configuration for each of a parameter configuration for each of the second second second second second parameter of the second second second control of transactions, transact correct groups of homogeneous to correct groups of homogeneous to transactional chartering for a large number to automatic or a extrem that does not scale to large user b reposited applying the existing pr likes a means [22], which are perfec- tional classical and providens. Unitor applicables to transaction data. To applicables to transaction data. To applicables to transaction data. To applicables to transaction data. To applicable to transaction data. To applicable to transaction data. To applicables to transaction data. To applicable to providing a video solitike soliti	CLOPE: A Fast and Effected Lycansacia           Willing Yang         Wilding Yang           Willing Yang         Willing Yang     <	<text><text><text><text><text></text></text></text></text></text>
ABST Minning of per- or per- or per- ning of the second se	Foreca Giannotti ISTI-CAP, Fas, Italy forecagiannotti@ehti.cmit et alage muther of dalanset recording human activities to the sense of individual data is the key muther of a new omained knowledge-based services. In this paper we per polehom of clustering individual transactional data i hass of users. Transacticnal data is a very pervaries the individual data is the key muther devices, often invol- obol of users. We propose transma, a parameter-free polehom of clustering individual transactional data is a very pervasive the appender of the service and the services of the services in the service of the service and the services of the services of the services of the services of the services and any services and services that the services of a distribute effect of our always-connected society is in the services and any service of our every digital technologies. Thanks to these data, human activ- dia distribute effect of our always-connected society is in the services. The services of the services of the services area (services), a special data of capterial distribute effect of our always-connected society is in a services of the services of the services of the services of the services. The services of the services area (services), a special data of capterial form of sets of event data, and any service in a Societ of an avained or distribution as a and setting of the services in the services of the services of the services of a services area (services), a special data of the services in a pairs interes of the services. The services of the services of a service area of the services of the services of the services of a service of a services of the service of the services of a service of the services, cervices the services of the services of a service area of the services of the services of the services of a service area of the services of the services of the services of a service area of the services of the services of the services of the services of a service area of the services of the services of the services of th	Dino Pedreschi University of Pisa, Italy dino, pedreschi (jedi, uniyi) time pedreschi (jedi, uniyi) time bare mest. This require it his included in surg data mining rether the necessity to automatically capi the second second second second second parameter configuration for each of classical second second second second characteristics of the dataset under a construction of the second second second characteristics of the dataset under a characteristic of the dataset under a characteristic of the dataset under a construction of transactions, transacti covering groups of homogeneous to covering groups of homogeneous to covering groups of homogeneous to reposited applying the existing pr problem, i.e., the separate individue transactioni datasets, as mas chara transactioni dataset, as mas chara transactioni transactioni data. To appleabilito of transactioni data. To appleabilito providing variabilito silvelito soluti a masativo number of different data transativo number of different data. transativo number	CLOPE: A Fast and Effect Lynamics Willing Yang	<text><text><text><text><text></text></text></text></text></text>
ABST Mining of pers of pers of the large in inform being in a co where dataset indrift forms we pre- the the person of diffi- forms we per- the the person of the the person of the the person of the person of the person of the person of the person of the person of the person of the person of the person of the	Fisca Giannotii ISTI-CMP. They, laby forecasiannotii(qibiti.cm2) Text. (A large number of dalatest: recording human activities genus of individual is the key analysis of a new onalized toxwledge-based services. In this paper we for poblem of chartering individual Transactional data 1 ass of users. Transactional data is in key any pervariave ku and the service of the service arrow, deni mod algorithm also to efficiently partitioning transactional algorithm also to efficiently partitioning transactional transaction and the service of arrows of the analy algorithm also to efficient aggest of users and the chartering must be applied on a massive number of align and particular differences of transacts for the same shaft are personal darkers and suggests that transacts on transactional datasets, and suggests that transacts on the displatic methodogist. That the the data human acti- noming observable, measurable, quantitable and, predi- tion displatic denicology. That has the data human acti- oming observable, measurable, quantitable and, predi- tion of sect of oreal parts-connected society is in arrawner of information that, for the most part con- sactors of the human parts of the same has a data inform of sect of oreal parts-connected society is a starken of aligned transactione in the parts in a patia- tion of stark of oreal parts of the same parts of a new way a special context involve the analysis of a larger in application context involve the analysis of a larger patient and into previde, the data and a patient in a pati- diabate, human be that that read to be and analysis of a larger in application context involve the analysis of a larger patient and into previde. In a moreal patient of an and parts of the distribute transactiones, det, when the evel baselines, recent caracterized by different properties in advestore of individual transactions	Dino Pedreschi University of Pisa, Italy dino, pedreschi@di.uniyi. throu obter user. This requires the included in any data mining nether the necessity to automatically capt visual behaviors. Due to the poten- ic (e.g. users in nowadys maxive sy of all generally unfatable to determine in provide an inning nether the to determine provide inning nether the trans- transactional clustering for a large number to all clustering for a large number of clustering for a large number to all clustering for a large number common item [50]. In the state ( common item [50]. In the state ( provide and commission and cluster) that that does not cache to large user b reposition and commission or an extrem and, the state ( common item [50]. In the state ( common item [50]. In the state ( the in a datomatic, or an extrem and the constraint of a large outer) that is not anomatics, as man cluster the notes (22), which are perfor like remose [25]. Nuch are perfor like to remose [25]. Nuch are perfor like to remose [26]. Nuch are perfor any existing manufer for futtherent data a massive number of different data a trategy similar to remoran [27]. Nuch for a large to remove the futtherent data is nuch and there the source to remove the futthere data is nuch and there the source to remove the futtherent data is nuch and there the source to remove the futtherent data is nuch and there the source to remove the futtherent data is nuch and there the source to remove the futtherent data trategy similar to remove the source to remove the futtherent data trategy similar to remove there the source to remove the futtherent data trategy similar to	CLOPE: A Fast and Effective Linearized           Willing the problem of categorial data categorial (yang-y, quan-xd, yang)           Appendix the problem of categorial data categorial (yang-y, quan-xd, yang)           Appendix the problem of categorial data categorial data categorial data categorial data (18), 8, 6, 7, 1)           Monte data data data data data data data categorial data categorial data categorial data data data data data data categorial data categorial data (18), 8, 6, 7, 1)           Monte data data data data data data data da	<text><text><text><text><text></text></text></text></text></text>
ABST Mining in making of pers- on the large n in the second second second the second second second the second seco	Fisca Gannotti ISTI-CRP, Tas, Italy foreaciannotti@ehit.crrit et alarge number of adamset recording human activities are sub-finited table and the transmer activities are sub-finited table and in the key and table of a new omained knowledge-based services. In this paper we of poblem of chattering individal Transmericand data 1 has of users. Transactional data is a very pervavive the obsolo for the service services, the service and the service services. The service services are problem of chattering individal Transmericand data 1 and the service services. The service services are problem of chattering was a service service service and and the service services. The service service services are problem of chattering was a service service service service and and the service services are the service service services and service services and services that the near one on service percentilized effectiveness of transment of the mass chatter that percentilized effectiveness of transment of the service that percentilized effectiveness of transment of the service that percentilized effectiveness of transment of the service of the displant chatteriologic services and the service service of the service line service of the service service service service services are service to a service service service service service services are service to a service service service service services and the service services are service to a service service service service service services are service to a service service service service service service services	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi. timo bella in sug data in the period in the period of the period of the period the period of the period of the period of a parameter in nowadays massive sy of of period of the period of the period parameter configuration for each of Amateria in the period of the period of the period of the period of the period of the period of the period of the period of the period of the period of the period parameter configuration for each of Amateria in the period of the period of the period of the period of the period of the context of the period of the period period of the period of the period of the period of the context of the period of the period period of the period of the period of the period of the context of the period of the period period the transaction data. To of the context of the period of the period of the period period the transaction data. To of the context of the period of the period of the period of the chartering period the transaction data. To of the period of the period of the period of the period of the chartering the the transact of the the transaction data. To of the chartering the the transact of the the transaction data. To of the period of the period of the transaction of the transaction data. To of the period of the period of the period the transaction of the transaction data. To of the period of the period of the transact of the transaction data. To of the period of the period of the period the transact of the transaction data. To of the period of the period the transact of the transaction data. To of the period of the period the transact of the transaction data. To of the period of the period the transact of the transaction data. To of the period of the period the period the transact of the the transact of the the transact	CLOPE: A Fast and Effective Lynamics           Willing Yang, Yang	<text><text><text><text><text></text></text></text></text></text>
ABST Mining of person of person of the information of the information of difference of	Fisca Gannotti ISTI-CAP, Fas, Italy forcaginnotti@bit.cmit. ISTI-CAP, Fas, Italy forcaginnotti@bit.cmit. Italy and the second se	Dino Pedreschi University of Pisa, Italy dino, pedreschi (jedi, unija) timo bed in sug data insing renthe included in sug data insing renthe the necessity to automatically cap visual behaviors. Due to the poter in garantee configuration for each of ceg. users in nowadays massive sy def garantee configuration for each of chan- the configuration of the state under construction of the state under the state of the state of the construction of the state under construction of the state under the state of the state under the state of the state of the state construction of the state under the state of the state of the state construction of the state of the state construction of the state of the state construction of the state of the state trans- tice of the state of the state of the state trans- ting common items [9]. In the state of the state of the state of the state of the the state of the state of the state of the trans-the state of the state of the state trans-the state of the state of the state of the state of the state of the state of the state state of the transcience of the state of the state of the state of the state of the state of the state of the state of the transcience of the state of the state of the state of the transcience of the state of the state of the state of the transcience of the state of the stat	CLOPE: A Fast and Effect: Lanasaci         Wing Yang Yang Yang Yang Yang Yang Yang Ya	<section-header><text><text><text><text><text></text></text></text></text></text></section-header>
ABST Mining in the second seco	Fisca Gammotii ISTI-CRP, They, Italy forecaginantii(qibiti.crit): ISTI-CRP, They, Italy forecaginantii(qibiti.crit): They are an ember of adatest: recording burnas activities stress of individual tas the key analysis of a new onalized toxwledge-based services. In this paper we by obselom of clustering individual transactional data I ass of users. Transactional data is the key analysis and that is collected by ascertal services, due in ited algorithm also to efficiently partitioning transactional algorithm also to efficiently partitioning transactional data is a very a service as a service of a data and the service of the service as a service of a data and the service of the service as a service of the tand and and the officiently partitioning transactional data of the service as a service as a service of the service rent personal data service and agorets that connexes to the data personal data service as a service as a service of the service rent personal data service as a service as a service of the service rent personal data service as a size of the service of the service integration service. The service of the service of the service of digital ternshologies. That is to these data, human acti- oming observable, measurable, quantitable and, predi- digital chendrologies. That is to these data is don't of a dis- part of the service of the service of a service of a new we paper card in due to a proving service now the service in a paties as starts and of the service the service in a paties as starts and of the service the service in a paties as starts and of the service the service in a paties as starts and on the service of the service in a paties as starts and one conclusterize the service in a paties as starts and one conclusterize the service in a paties as starts and one conclusterize the service of a new we have been as integrated than the rent or bear of the service in a paties as the service as the service of the service of a new we have been as the service of the service of a service we have been as the	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi. throw other user. This requires the included in any data mining nether the necessity to automatically capital equation of the second second second second equation of the second second second second parameters in nowadays maximum second equation of the second second second second parameters in our adays making and equation of the second second second second experimental second second second second construction of the second second second construction of the second second second second constructions of the second second second construction of the second second second second constructions of the second second second second constructions (1910). The state is the second second second second second second the second second second second second second the second second second second second second terms and the second second second second second terms and the second second second second second terms and the second second second second terms and the second second second second second terms and the second second second second second terms and the second second second second terms and the second second second second second terms and the second second second second terms and the second second second second second terms and the second second second second second second terms and the second second second second second second terms and the second second second second second second terms and	<section-header><text><text><section-header><text><text><text></text></text></text></section-header></text></text></section-header>	<section-header><text><text><text><text><text></text></text></text></text></text></section-header>
ABST Mining in or pers- on the large ne higher the tering, is where, whether what is a second of different forms or the we pre- are been at least the advertised of the forms	Fisca Gamnotii ISTI-CRP, They, Italy foracajannotii(@isti.crwit. ISTI-CRP, They, Italy foracajannotii(@isti.crwit. The analysis of the analysi	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi. timo bella di suga dana di suga di foru obter user. This require it his necessi y outionatically capi tella della di suga dana di suga di e (e.g. users in novadays massive sy generally unfassible to determise in parameter configuration for each of class- della di suga di suga di suga di personali della di suga di personali della di suga di personali della di suga di personali della di suga di personali di suga di suga di personali di suga di personali di suga di personali di suga di suga di personali di suga di personali di suga di suga di personali di suga di suga di personali di suga di personali di suga di personali di suga di suga di personali di suga di personali di suga di di di di di suga di personali di suga di suga di persona di suga di personali di suga di suga di personali di suga di suga di di suga di personali di suga di suga di di suga di personali di suga di di suga di suga di suga di suga di suga di di suga di suga di suga di suga di suga di di suga di suga di suga di suga di suga di di suga di suga di suga di suga di suga di di suga di suga di suga di suga di suga di suga di di suga di suga di suga di suga di suga di di suga di suga di suga di suga di suga di suga di di suga di suga di suga di s	<section-header><text><text><text><section-header><text><text><text><text></text></text></text></text></section-header></text></text></text></section-header>	<text><text><text><text><text></text></text></text></text></text>
ABS7 Mining of pers of the large in a co where, dataset individual dataset individual dataset above i dataset	The second method of the second secon	Dino Pedreschi University of Pisa, Italy dino pedreschi@di.uniyi. the pedreschi@di.uniyi. the pedreschi@di.uniyi. the network. This requires the included in any data mining metho the necessity to automatically capt (e.g. uners) to automatically capt (e.g. uners) to automatically capt percental unitage method the data percental environment of the second percental capter of the second second percental capter of the second second capter of the second second second second percental second second second second percental second second second second capter of the second second second second percental second second second second second covering groups of homogeneous to covering groups of homogeneous to covering groups of homogeneous to covering droug data second second second the second second second second second the second second second second second the second second second second second from the second second second second second of the chartering percent individual a massive number of different datasets - which i generative of different data coverestim a massive conversion generate the first data second second to the generative data a massive covere on the deferent dataset a massive covere second be deferent dataset a massive covere data de dataset and tratagey similar to a massive cover dataset dataset a massive covere second be deferent dataset a massive covere dataset dataset to deferent dataset a massive covere dataset dataset and tratagey similar to a massive covere dataset dataset a massive covere dataset dataset to dataset and tratagey similar to a massive covere dataset dataset dataset and tratagey similar to a massive covere dataset dataset to a dataset and tratagey similar to a massive covere dataset dataset to a dataset dataset to a dataset	<section-header><text><text><section-header><section-header><section-header><section-header><section-header><text></text></section-header></section-header></section-header></section-header></section-header></text></text></section-header>	<text><text><text><text><text></text></text></text></text></text>

IGKDD '02, July 23-26, 2002, Edmonton, Alber

X-means: Extending K-means with Efficient Estimation of the Number of Clusters

#### analyzing the beights and widths of the clusters. Leaving out the two identical histograms for cluster (*de*, *deff* and cluster (*ab*, *abc*) the other two histograms are of different quality. The histogram for cluster (*ab*, *abc*, *acd*) has only 4 distinct idents for 8 blocks $H^{2-2}$ . $H^{H^{2-2}}$ . So, but the one for cluster (*acd*, *de*, *deff*) has 5, for

# **Exercises Transactional Clustering**

- Suppose we have four verses contains some subjects, as follows:
- P1={ judgment, faith, prayer, fair}
- P2={ fasting, faith, prayer}
- P3={ fair, fasting, faith}
- P4={ fasting, prayer, pilgrimage}
- the similarity threshold = 0.3, and number of required cluster is 2.

Using Jaccard coefficient as a similarity measure, we obtain the following similarity table

	P1	P2	P3	P4
P1	1	0.4	0.4	0.17
P2		1	0.5	0.5
<b>P</b> 3			1	0.2
P4				1

- Since we have a similarity threshold equal to 0.3, then we derive the adjacency table: →
- By multiplying the adjacency table with itself, we derive the following table which shows the number of links (or common neighbors): →

	P1	P2	P3	P4
P1	1	0.4	0.4	0.17
P2		1	0.5	0.5
P3			1	0.2
P4				1

	P1	P2	P3	P4
P1	1	1	1	0
P2		1	1	1
P3			1	0
P4				1

	P1	P2	P3	P4
P1	-	3	3	1
P2		-	3	2
P3			-	1
P4				-

• we compute the goodness measure for all adjacent points ,assuming that

• 
$$f(\theta) = 1 - \theta / 1 + \theta = 1 - 0.3 / 1 + 0.3 = 0.54$$

- we obtain the following table  $\rightarrow$
- we have an equal goodness measure for merging ((P1,P2), (P2,P3), (P3,P1))

$$g(P_i, P_j) = \frac{link[P_i, P_j]}{(n+m)^{1+2f(\theta)} - n^{1+2f(\theta)} - m^{1+2f(\theta)}}$$

Pair	Goodness measure
P1,P2	1.35
P1,P3	1.35
P1,P4	0.45
P2,P3	1.35
P2,P4	0.90
P3,P4	0.45

- Now, we start the hierarchical algorithm by merging, say P1 and P2.
- A new cluster (let's call it C(P1,P2)) is formed.
- It should be noted that for some other hierarchical clustering techniques, we will not start the clustering process by merging P1 and P2, since Sim(P1,P2) = 0.4, which is not the highest. But, ROCK uses the number of links as the similarity measure rather than distance.

- Now, after merging P1 and P2, we have only three clusters. The following table shows the number of common neighbors for these clusters:→
- Then we can obtain the following goodness measures for all adjacent clusters:→

	C(P1,P2)	P3	P4
C(P1,P2)	-	3+3	2+1
P3		-	1
P4			-

Pair	Goodness measure			
C(P1,P2),P3	1.31			
C(P1,P2),P4	0.66			
P3,P4	0.45			

• Since the number of required clusters is 2, then we finish the clustering algorithm by merging C(P1,P2) and P3, obtaining a new cluster C(P1,P2,P3) which contains {P1,P2,P3} leaving P4 alone in a separate cluster.

 Given the following similarity matrix find the clustering result knowing that the similarity threshold = 0.4, and number of required cluster is 2.

	p1	p2	р3	р4	р5
p1	1	0.7	0.2	0.5	0.5
p2		1	0.6	0.8	0.1
р3			1	0.5	0.4
p4				1	0.3
р5					1

	p1	p2	р3	р4	р5
p1	1	0.7	0.2	0.5	0.5
p2		1	0.6	0.8	0.1
р3			1	0.5	0.4
p4				1	0.3
р5					1

	p1	p2	р3	р4	р5
p1	1	1	0	1	1
p2	1	1	1	1	0
р3	0	1	1	1	1
p4	1	1	1	1	0
р5	1	0	1	0	1

	p1	p2	р3	р4	р5
p1	1	1	0	1	1
p2	1	1	1	1	0
р3	0	1	1	1	1
p4	1	1	1	1	0
р5	1	0	1	0	1

	p1	p2	р3	p4	р5
p1	-	3	3	3	2
p2		-	3	4	2
р3			-	3	2
p4				-	2
р5					-

- $f(\theta) = 1 \theta / 1 + \theta = 1 0.4 / 1 + 0.4 = 0.43$
- $1 + 2 f(\theta) = 1.86$

$$g(P_i, P_j) = \frac{link[P_i, P_j]}{(n+m)^{1+2f(\theta)} - n^{1+2f(\theta)} - m^{1+2f(\theta)}}$$

	p1	p2	р3	р4	р5
p1	-	3	3	3	2
p2		-	3	4	2
р3			-	3	2
p4				-	2
p5					-

	p1	p2	р3	р4	р5
p1	-	1.84	1.84	1.84	1.22
p2		-	1.84	2.45	1.22
р3			-	1.84	1.22
p4				-	1.84
р5					-

• 
$$f(\theta) = 1 - \theta / 1 + \theta = 1 - 0.4 / 1 + 0.4 = 0.43$$

•  $1 + 2 f(\theta) = 1.86$ 

$$g(P_i, P_j) = \frac{link[P_i, P_j]}{(n+m)^{1+2f(\theta)} - n^{1+2f(\theta)} - m^{1+2f(\theta)}}$$

	p1	p2	р3	p4	р5		p1	p2p4	р3	р5		p1	p2p4	р3	р5
p1	-	3	3	3	2	p1	_	6	3	2	p1	_	1.94	1.84	1.22
p2		-	3	4	2	•					•				
-				-		p2p4		-	6	4	p2p4		-	1.94	1.29
p3			-	3	2										
p4				-	2	p3			-	2	p3			-	1.22
						nE					ηĘ				
p5					-	h2				-	ps				-

• Final Clusters: p1234 p5

# **Clope Exercise 1**

Transactions: abc, abc, ab, ad, def, ade, ade

Split1:

- 4 transactions: abc, abc, ab, ad
  - a:4, b:3, c:2, d:1 -> S=10; W=4; H=10/4=2,5; H/W=2,5/4=0,625
- 3 transactions: def, ade, ade
  - a:2, d:3, e:3, f:1 -> S=9; W=4; H=9/3=3; H/W=3/4=0,75

Split2:

- 2 transactions: abc, abc, ab
  - a:3, b:3, c:2 -> S=8; W=3; H=8/3=3,6; H/W=0,88
- 2 transactions: ad, def, ade, ade
  - a:3, d:4, e:3, f:1 -> S=11; W=4; H=11/4=2,75; H/W=2,75/4=0,68

Split1 is the best clustering considering r=2 Profit(Split1) =  $(10/4^2 * 4 + 9/4^2 * 3)/7 = 0.59$ Profit(Split2) =  $(8/3^2 * 3 + 11/4^2 * 4)/7 = 0.77$ 



# **Clope Exercise 2**

Split1:

- 4 transactions: abc, abc, ab, a
  - a: 4, b:3, c: 2 -> sol: S=9; W=3; H=9/3=3; H/W=1
- 3 transactions: def, de, de
  - d: 3, e:3, f: 1 -> sol: S=7; W=3; H=7/3=2.33; H/W=0.77

Split2:

- 2 transactions: abcd, ab
  - a: 2, b:2, c: 1, d:1 -> sol: S=6; W=4; H=6/4=1.5; H/W=0.37
- 2 transactions: ec, ec
  - e:2, c: 2 -> sol: S=4; W=2; H=4/2=2; H/W=1

Split1 is the best clustering considering r=2 Profit(Split1) =  $(9/3^2 * 4 + 7/3^2 * 3)/7 = 0.90$ Profit(Split2) =  $(6/4^2 * 2 + 4/2^2 * 2)/4 = 0.16$   $Profit_{r}(\mathbf{C}) = \frac{\sum_{i=1}^{k} \frac{S(C_{i})}{W(C_{i})^{r}} \times |C_{i}|}{\sum_{i=1}^{k} |C_{i}|}$