Mining Subgroups with Exceptional Transition Behavior

KDD'16

slides by Pedro O.S. Vaz de Melo

Introduction

- Exceptional Model Mining
  - a generalization of the classic subgroup discovery task
  - framework that identifies patterns which contain unusual interactions between multiple target attributes
  - it emphasizes the detection of easy-to-understand subgroups

Objective

- Apply exceptional model mining to discover interpretable subgroups with exceptional transition behavior
- Applications
  - a transition model could show that people either move within their direct neighborhood or along main roads
  - identify subgroups of people (such as “male tourists from France") or subsegments of time (such as “10 to 11 p.m.") that exhibit unusual movement characteristics, e.g., tourists moving between points-of-interest or people walking along well-lit streets at night
  - identify subgroups of webusers with unusual navigation behavior or subgroups of companies with unusual development over time

Contributions

- a new method that enables mining subgroups with exceptional transition behavior by introducing first-order Markov chains as a novel model class for exceptional model mining
- an interestingness measure that quantifies the exceptionality of a subgroup’s transition model
- a method to find subgroups specifically matching (or contradicting) given hypotheses about transition behavior
- validation of the method using synthetic as well as real-world data

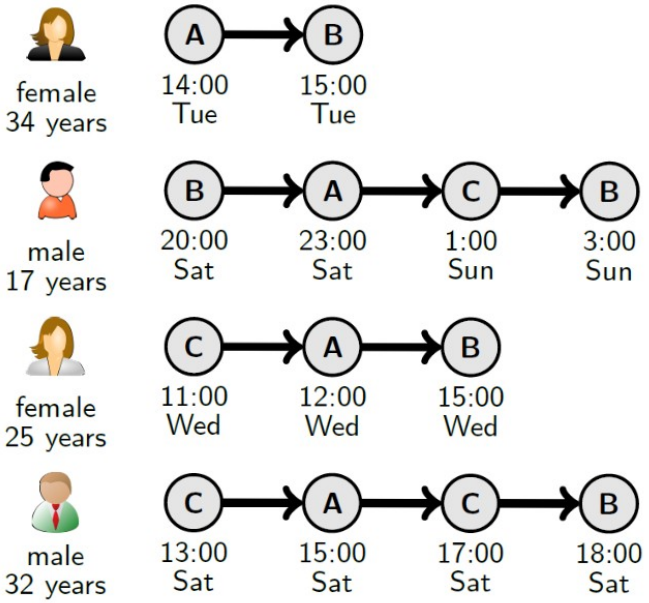
Exceptional Model Mining

- Identifies the ones that are unusual (“interesting”) with respect to a target model class and the model attributes
  - e.g.: “While overall there is a positive correlation between the exam preparation time and the score ($\rho = 0.3$), the subgroup of males that are younger than 18 years shows a negative correlation ($\rho = -0.1$)”

Exceptional Model Mining

- The goal is accomplished by using a quality measure ***q*** that maps a subgroup to a real number based on the supposed interestingness of its model parameters and performing a search for the subgroups with the highest scores

Data representation



(a) Sequence data with background knowledge

A_M		A_D				
Source State	**Target State**	Gender	Age	Hour	Weekday	# Visits of user
A	B	f	34	14	Tue	2
B	A	m	17	20	Sat	4
A	C	m	17	23	Sat	4
C	B	m	17	1	Sun	4
C	A	f	25	11	Wed	3
A	B	f	25	12	Wed	3
C	A	m	32	13	Sat	4
A	C	m	32	15	Sat	4
C	B	m	32	17	Sat	4

(b) Transition dataset

$$\begin{pmatrix} 0 & 2 & 2 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix}$$

(c) Transition matrix T_D (entire dataset)

$$\begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

(d) Transition matrix $T_{Gender=f}$

$$\begin{pmatrix} 0 & 0 & 2 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

(e) Transition matrix $T_{Weekday=Sat}$

Interestingness Measure

- Based on a comparison between the transition matrix of the subgroup and a reference transition matrix that is derived from the overall dataset

Interestingness Measure

- First, Manhattan distance between two rows:

$$\delta_{tv}(g, D, i) = \frac{1}{2} \sum_j \left| \frac{g_{ij}}{\sum_j g_{ij}} - \frac{d_{ij}}{\sum_j d_{ij}} \right|$$

- Then, distance between two matrices:

$$\omega_{tv}(g, D) = \sum_i \left(w_i \cdot \sum_j \left| \frac{g_{ij}}{\sum_j g_{ij}} - \frac{d_{ij}}{\sum_j d_{ij}} \right| \right)$$

$$w_i = \sum_j g_{ij}$$

Comparison with random samples

- Get r random samples
- Each sample has the same number of transitions from each state
- Compare the original distance with the distances generated from the random samples
- Interestingness measure:

$$q_{tv}(g,D) = \frac{\omega_{tv}(g,D) - \mu(\omega_{tv}(R_1,D),\ldots,\omega_{tv}(R_r,D))}{\sigma(\omega_{tv}(R_1,D),\ldots,\omega_{tv}(R_r,D)) + \varepsilon}$$

Subgroup search

- enumerate all candidate subgroups in the search space in order to find the ones with the highest scores

Subgroup search

- A typical problem in pattern mining is redundancy
  - eg: the subgroup male induces an exceptional transition model and thus achieves a high score, then also the subgroup males older than 18 can be expected to feature a similarly unusual model and receive a high score
- Solution: to adapt a minimum improvement constraint
  - remove a more specific subgroup with a similar (e.g., less than 10% difference) or a higher (???) score

Subgroup assessment

- Key statistics
- Exemplification
- Visualization

User-defined hypothesis

- Hypothesis for toy example (Figure 1):

$$\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Experiments

- Synthetic data
- Real data

Synthetic #1: Random transition matrices

- two random 5x5 transition matrices: M1, M2
- 2 attributes
  - A: ternary, B: binary
- 20 random binary attributes
- generation of transitions:
  - if A = A1 and B = B1, generate transition from M1
  - else, generate from M2
- goal: identify M2 in the whole data

Synthetic #1: Random transition matrices

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
A = A1 $\wedge$ B = B1	10,000	113.01 $\pm$ 2.74	5,783	1.54
A = A1	20,000	67.23 $\pm$ 1.60	4,634	0.60
B = B1	30,000	45.52 $\pm$ 0.94	3,480	0.33
B = B2	30,000	44.69 $\pm$ 1.08	3,480	0.51
A = A3	20,000	32.05 $\pm$ 0.77	2,378	0.53

Synthetic #2: random walker

- a scale free BA network
- each node has a color
- generate some random walks
  - M1: completely random walk
  - M2: higher probability to walk to nodes of the same color

Synthetic #2: random walker

(a) Comparison to the overall dataset

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	35.67 ± 0.78	51,929	125.96
Type = Random	799,085	34.34 ± 0.80	51,929	31.73
Noise9 = False	681,835	2.25 ± 0.06	51,358	36.27
Noise9 = True	318,165	2.23 ± 0.06	51,358	77.94
Noise2 = False	18,875	1.80 ± 0.05	14,844	394.51

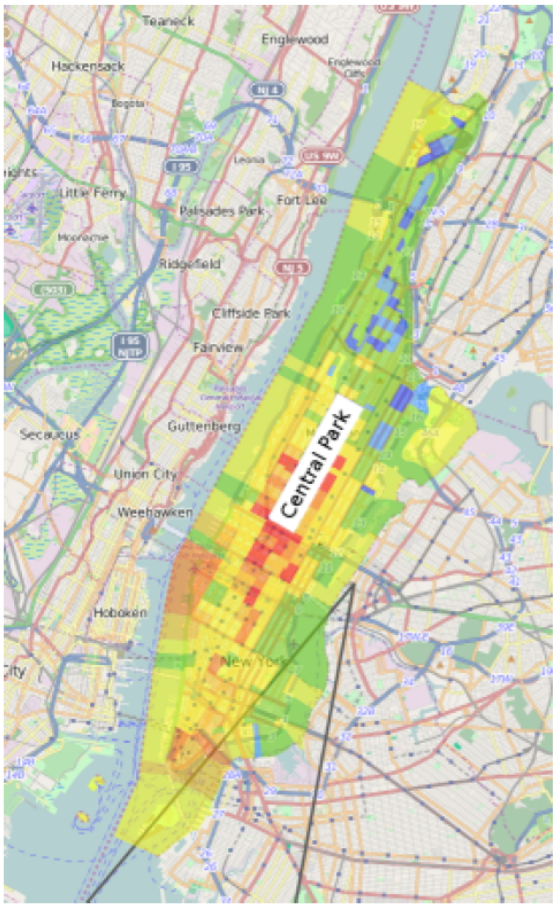
(b) Comparison to the *homophile* hypothesis, contradicting

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Random	799,085	26.88 ± 0.57	1,554,130	981.38
Noise4 = True	519,130	2.28 ± 0.06	1,008,912	981.25
Noise2 = False	18,875	2.25 ± 0.06	37,057	987.49
Noise1 = True	469,290	2.00 ± 0.05	912,032	981.26
Noise19 = True	342,765	1.93 ± 0.05	666,229	981.28

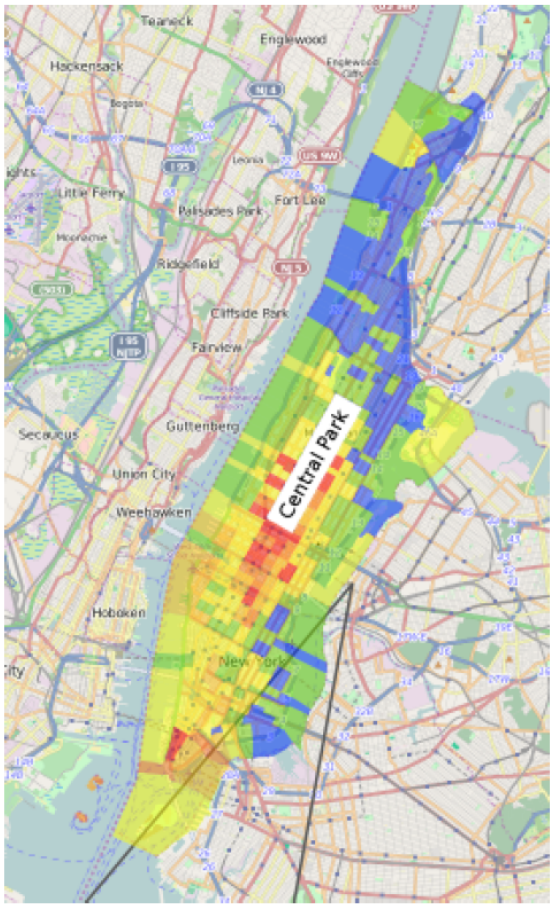
(c) Comparison to the *homophile* hypothesis, matching

Description	# Inst.	q_{tv} (score)	ω_{tv}	Δ_{tv}
Type = Homophile	200,915	12.10 ± 0.27	389,841	981.04
Noise4 = False	480,870	2.69 ± 0.07	934,190	981.20
Noise19 = False	657,235	2.27 ± 0.06	1,276,868	981.20
Noise1 = False	530,710	1.99 ± 0.05	1,031,101	981.20
Noise0 = True	523,410	1.74 ± 0.05	1,016,899	981.21

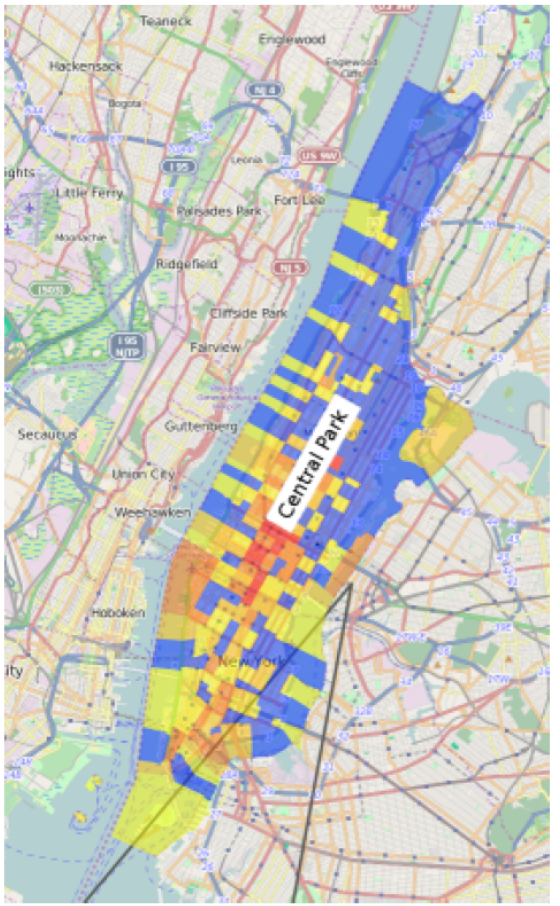
Real #1: Flickr



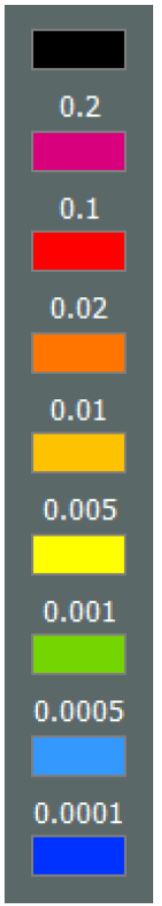
(a) All transitions



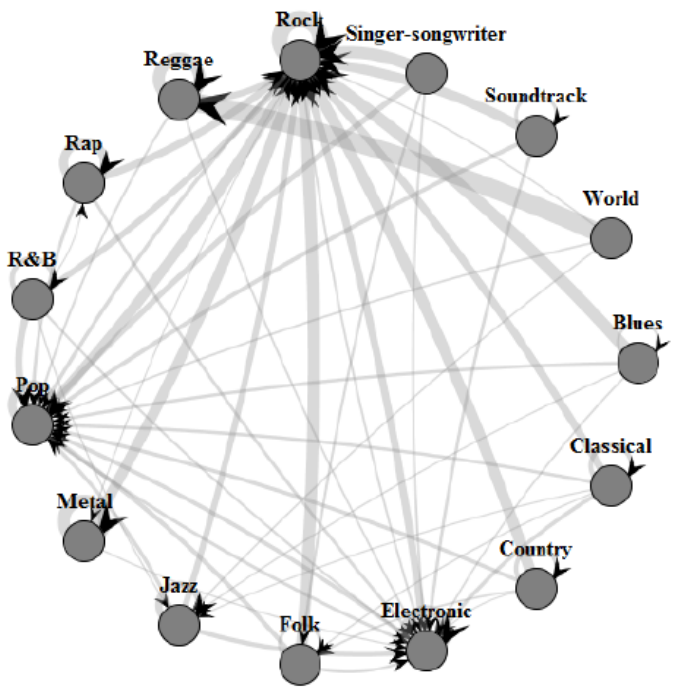
(b) Transitions of tourists



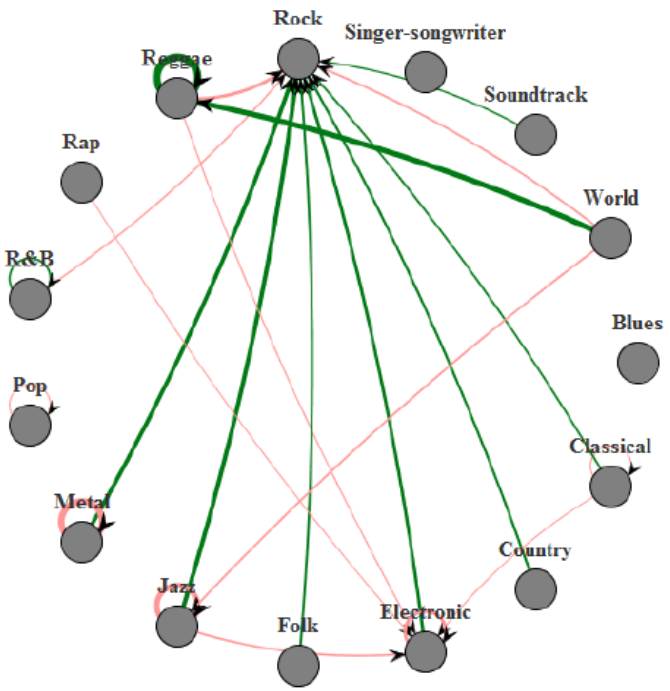
(c) Night transitions (22–23 h)



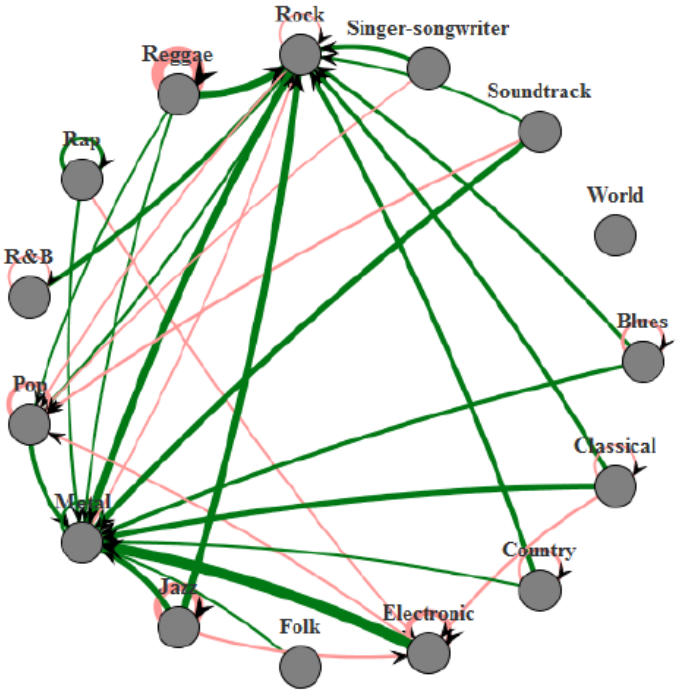
Real #2: LastFM



(a) All transitions



(b) Users from the United States



(c) Users from Finland

Conclusions