Model-based Exception Mining for Relational Data

Fatemeh Riahi and Oliver Schulte

School of Computing Science
Simon Fraser University, Vancouver, Canada

Exception Mining Task

- Identify *exceptional individuals* whose statistical patterns deviate from the general population
- Can also be used for outlier/anomaly detection
- Our approach: apply the *Exceptional Model Mining* framework (EMM) to multi-relational data (Duivesteijn, W.; Feelders, A. J. & Knobbe, A. 2016.)

Highlights

- *Leverage*: Framework applies to any relational learning method
- *Ranking:* Provides single score for individual entities
- *Interpretability:* Scores can be explained by statistical differences in local feature distributions

Other Approaches

- Association Rules, e.g.
Maervoet, J.; Vens, C.; Vanden Berghe, G.; Blockeel, H. & De Causmaecker, P. (2012), 'Outlier Detection in Relational Data: A Case Study in Geographical Information Systems', *Expert Systems With Applications* **39**(5), 4718--4728.
- Clustering, e.g.
Sun, Y.; Han, J.; Zhao, P.; Yin, Z.; Cheng, H. & Wu, T. (2013), Community Distribution Outlier Detection in Heterogeneous Information Networks., *in* 'ECML/PKDD', pp. 557-573.
- Extracting network features, e.g. ODDBALL
Akoglu, L.; Mcglohon, M. & Faloutsos, C. (2010), OddBall: Spotting Anomalies in Weighted Graphs, *in 'PAKDD', pp. 410-421.*
- Propositionalization, *e.g.*
Riahi, F. & Schulte, O. (2016), Propositionalization for Unsupervised Outlier Detection in Multi-Relational Data, in 'FLAIRS', 448-453

EMM for I.I.D. Data

Entire Population Data

attribute 1	attribute 2	attribute 3

learning → population model

Subgroup Data

attribute 1	attribute 2	attribute 3

learning → subgroup model

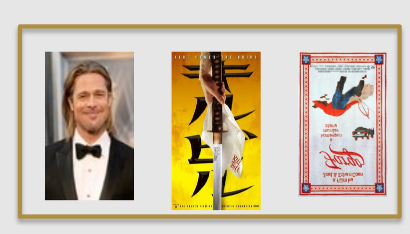
Quality measure = Measure of dissimilarity between population and subgroup models

EMM for Relational Data

Entire Observed Network

learning → statistical-relational population model

Individual Subnetwork aka "egnoet"/"interpretation"



learning → individual SRL model

Outlierness Metric (quality measure) = Measure of dissimilarity between population and individual SRL models

We investigate several metrics based on *log-linear likelihood functions*

Example

population model B_{P0} for random actor A

gender(A) → ActsIn(A,M) ← Drama(M)

Gender (A)	Drama (M)	CP ActsIn (A,M)
M	T	1/2
M	F	0
W	T	0
W	F	1

individual model B_o for A = Brad Pitt

gender(brad_p) → ActsIn(brad_p,M) ← Drama(M)

Gender (B.P.)	Drama (M)	CP ActsIn (B.P.,M)
M	T	0
M	F	0

Outlierness Metrics

- Starting point is KLD between population and individual model
- Promising novel variant ELD= mutual information decomposition + absolute values to avoid cancellations

$$KLD(B_o \| B_P) = \underbrace{\sum_{\text{nodes } i} \sum_{\text{values k}} \sum_{\text{parent-state } j}}_{\text{summation over local features}} P_{B_o}(X_i = x_{ik}, Pa(X_i) = pa_j) \times \ln\Bigg(\underbrace{\frac{P_{B_o}(X_i = x_{ik} \mid Pa(X_i) = pa_j)}{P_{B_P}(X_i = x_{ik} \mid Pa(X_i) = pa_j)}}_{\text{log-difference in empirical conditional probabilities (confidences)}}\Bigg)$$

B_p models the population network B_o models the individual network

Evaluation

AUC for detecting ground-truth outliers
(e.g. Goalies injected into set of Strikers)

Dataset	ELD	KLD
PL: Strikers	**0.89**	0.65
PL: Midfielders	**0.66**	0.55
IMDb: Drama	**0.70**	0.66

Case Studies

For each individual, drill down on the aggregate outlierness score to find

1. most unusual feature
2. most unusual feature value.

Individual	Group	Rank	Max Node	Max Value	Individual Probability	Group Probability
Edin Dzeko	Striker	1	Dribble Efficiency	DE = Low	0.16	0.5
Paul Robinson	Goalie	2	Saves Made	SM = Medium	0.3	0.04
Brave Heart	Drama	1	Actor Quality	a_quality=4	0.93	0.42
Austin Powers	Comedy	2	Cast position	cast_num =3	0.78	0.49

Conclusion and Future Work

- Exceptional Model Mining: New approach for applying SRL models to relational exception mining
- New log-linear outlierness metric
- New Model and new metric showed promising results on Soccer and IMDB datasets.
- Future work:
1) explore other SRL models (e.g. Markov Logic Networks)
2) incorporate difference in model structure as well as parameters

Datasets

- Soccer Data: The Opta dataset released by Manchester City.
- IMDB Data: From The Internet Movie Database.
- For synthetic data please see paper

References

Duivesteijn, W.; Feelders, A. J. & Knobbe, A. (2016), 'Exceptional model mining', *Data Mining and Knowledge Discovery 30(1), 47--98.*
Tutorial on Learning Bayesian Networks for Complex Relational Data, Schulte and Kirkpatrick 2017, https://oschulte.github.io/srl-tutorial-slides/