**Riahi, Fatemeh**; **Schulte, Oliver**
**Model-based exception mining for object-relational data.** (English) Zbl 1433.68375
Data Min. Knowl. Discov. 34, No. 3, 681-722 (2020).

Summary: This paper develops model-based exception mining and outlier detection for the case of object-relational data. Object-relational data represent a complex heterogeneous network, which comprises objects of different types, links among these objects, also of different types, and attributes of these links. We follow the well-established exceptional model mining (EMM) framework, which has been previously applied for subgroup discovery in propositional data; our novel contribution is to develop EMM for relational data. EMM leverages machine learning models for exception mining: An object is exceptional to the extent that a model learned for the object data differs from a model learned for the general population. In relational data, EMM can therefore be used for detecting single outlier or exceptional objects. We combine EMM with state-of-the-art statistical-relational model discovery methods for constructing a graphical model (Bayesian network), that compactly represents probabilistic associations in the data. We investigate several outlierness metrics, based on the learned object-relational model, that quantify the extent to which the association pattern of a potential outlier object deviates from that of the whole population. Our method is validated on synthetic data sets and on real-world data sets about soccer and hockey matches, IMDb movies and mutagenic compounds. Compared to baseline methods, the EMM approach achieved the best detection accuracy when combined with a novel outlinerness metric. An empirical evaluation on soccer and movie data shows a strong correlation between our novel outlierness metric and success metrics: Individuals that our metric marks out as unusual tend to have unusual success.

**MSC:**

68T05 Learning and adaptive systems in artificial intelligence
68T09 Computational aspects of data analysis and big data

**Keywords:**

outlier detection; exception mining; statistical-relational learning; Bayesian network; likelihood ratio; network data

**Software:**

GitHub; LOF; ROCR

**Full Text:** DOI

**References:**

[1] Achtert E, Kriegel H-P, Schubert E, Zimek A (2013) Interactive data mining with 3D-parallel coordinate trees. In: Proceedings ACM special interest group on management of data, New York, NY, USA, pp 1009-1012. doi:10.1145/2463676.2463696

[2] Aggarwal CC (2013) Outlier analysis. Springer, New York. ISBN 9781461463955. http://books.google.ca/books?id=900CkgEACAAJ · Zbl 1291.68004

[3] Akoglu L, McGlohon M, Faloutsos C (2010) Oddball: spotting anomalies in weighted graphs. In: Proceedings Pacific-Asia conference on knowledge discovery and data mining, pp 410-421. doi:10.1007/978-3-642-13672-6_40

[4] Akoglu, L.; Tong, H.; Koutra, D., Graph based anomaly detection and description: a survey, Data Min Knowl Discov, 29, 3, 626-688 (2015)

[5] Albert, J.; Glickman, ME; Swartz, TB; Koning, RH, Handbook of statistical methods and analyses in sports (2017), Boca Raton: CRC Press, Boca Raton

[6] Anderson G, Pfahringer B (2008) Exploiting propositionalization based on random relational rules for semi-supervised learning. In: Proceedings Pacific-Asia conference on knowledge discovery and data mining, pp 494-502. doi:10.1007/978-3-540-68125-0_43

[7] Angiulli, F.; Greco, G.; Palopoli, L., Outlier detection by logic programming, ACM Trans Comput Logic, 9, 1-7, 7 (2004) · Zbl 1367.68314

[8] Beirlant, J.; Györfi, L.; Lugosi, G., On the asymptotic normality of the L1-and L2-errors in histogram density estimation,

Can J Stat, 22, 3, 309-318 (1994) · Zbl 0816.62037

[9] Beirlant, J.; Devroye, L.; Györfi, L.; Vajda, I., Large deviations of divergence measures on partitions, J Stat Plan Inference, 93, 1-2, 1-16 (2001) · Zbl 0996.62052

[10] Breunig M, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings ACM special interest group on management of data, pp 93-104. doi:10.1145/342009.335388

[11] Cansado, A.; Soto, A., Unsupervised anomaly detection in large databases using Bayesian networks, Appl Artif Intell, 22, 309-330 (2008)

[12] de Campos, L., A scoring function for learning Bayesian networks based on mutual information and conditional independence tests, J Mach Learn Res, 7, 2149-2187 (2006) · Zbl 1222.62036

[13] Domingos, P.; Lowd, D., Markov logic: an interface layer for artificial intelligence (2009), San Francisco: Morgan and Claypool Publishers, San Francisco · Zbl 1202.68403

[14] Domingos, P.; Richardson, M.; Getoor, L.; Taskar, B., Markov logic: a unifying framework for statistical relational learning, Introduction to statistical relational learning (2007), Cambridge: MIT Press, Cambridge

[15] Duivesteijn, W.; Feelders, AJ; Knobbe, A., Exceptional model mining, Data Min Knowl Discov, 30, 1, 47-98 (2016) · Zbl 1411.68096

[16] Fawcett, T., An introduction to ROC analysis, Pattern Recognit Lett, 27, 8, 861-874 (2006)

[17] Fisher, RA, On the probable error of a coefficient of correlation deduced from a small sample, Metron, 1, 3-32 (1921)

[18] Gao J, Liang F, Fan W, Wang C, Sun Y, Han J (2010) On community outliers and their efficient detection in information networks. In: Proceedings ACM special interest group on knowledge discovery and data mining, New York, NY, USA, pp 813-822. ACM. ISBN 978-1-4503-0055-1. doi:10.1145/1835804.1835907

[19] Garcia-del Barrio P, Pujol F (2004) Pay and performance in the Spanish soccer league: who gets the expected monopsony rents? Faculty Working Papers 05/04, School of Economics and Business Administration, University of Navarra, March 2004. https://ideas.repec.org/p/una/unccee/wp0504.html

[20] Getoor L (2001) Learning statistical models from relational data. PhD thesis, Department of Computer Science, Stanford University

[21] Getoor, L.; Taskar, B., Introduction to statistical relational learning (2007), Cambridge: MIT Press, Cambridge · Zbl 1141.68054

[22] Hall, S.; Szymanski, S.; Zimbalist, AS, Testing causality between team performance and payroll: the cases of Major League Baseball and English Soccer, J Sports Econ, 3, 2, 149-168 (2002)

[23] Halpern, JY, An analysis of first-order logics of probability, Artif Intell, 46, 3, 311-350 (1990) · Zbl 0723.03007

[24] Heckerman, D.; Meek, C.; Koller, D.; Getoor, L.; Taskar, B., Probabilistic entity-relationship models, PRMs, and plate models, Introduction to statistical relational learning (2007), Cambridge: MIT Press, Cambridge

[25] Horváth T, Alexin Z, Gyimóthy T, Wrobel S (1999) Application of different learning methods to Hungarian part-of-speech tagging. In: Dzeroski S, Flach P (eds) Inductive logic programming: 9th international workshop. ILP-99 Bled. Springer, Berlin, pp 128-139

[26] Horváth, T.; Wrobel, S.; Bohnebeck, U., Relational instance-based learning with lists and terms, Mach Learn, 43, 1, 53-80 (2001) · Zbl 0988.68039

[27] Khosravi H, Man T, Hu J, Gao E, Mar R, Schulte O (2019) Factorbase code. https://github.com/sfu-ml-lab/FactorBase. Accessed 15 Nov 2016

[28] Khot T, Natarajan S, Shavlik JW (2014) Relational one-class classification: a non-parametric approach. In: Proceedings association for the advancement of artificial intelligence, Quebec City, Quebec, Canada, pp 2453-2459. http://www.aaai.org/ocs/index.php/AAAI/AAAI1 Accessed 10 Dec 2017

[29] Kimmig, A.; Mihalkova, L.; Getoor, L., Lifted graphical models: a survey, Mach Learn, 99, 1, 1-45 (2014) · Zbl 1320.62016

[30] Kirsten, M.; Wrobel, S.; Horváth, T.; Dzeroski, S.; Lavrac, N., Distance-based approaches to relational learning and clustering, Relational data mining, 213-232 (2001), Berlin: Springer, Berlin

[31] Knobbe, AJ, Multi-relational data mining (2006), Amsterdam: IOS Press, Amsterdam · Zbl 1138.68376

[32] Koh JLY, Lee ML, Hsu W, Ang WT (2008) Correlation-based attribute outlier detection in XML. In: Proceedings international council for open and distance education, Cancun, Mexico. IEEE, pp 1522-1524. http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=449

[33] Koller D, Pfeffer A (1997) Object-oriented Bayesian networks. In: Geiger D, Shenoy PP (eds) Proceedings uncertainty in artificial intelligence. Morgan Kaufmann, Burlington, pp 302-313. arXiv:1302.1554

[34] Kramer S, Lavrac N, Flach P (2000) Propositionalization approaches to relational data mining. In: Dzeroski S (ed) Relational data mining. Springer, Berlin, pp 262-286

[35] Kuzelka O, Zelezný F (2008) Hifi: tractable propositionalization through hierarchical feature construction. In: Late breaking papers, inductive logic programming, p 69

[36] Liu G, Schulte O (2018) Deep reinforcement learning in ice hockey for context-aware player evaluation. In: Proceedings international joint conference on artificial intelligence. International Joint Conferences on Artificial Intelligence Organization, pp 3442-3448. doi:10.24963/ijcai.2018/478

[37] Maervoet, J.; Vens, C.; Berghe, GV; Blockeel, H.; Causmaecker, PD, Outlier detection in relational data: a case study in geographical information systems, Expert Syst Appl, 39, 5, 4718-4728 (2012)

[38] Müller E, Assent I, Iglesias P, Mülle Y, Böhm K (2012) Outlier ranking via subspace analysis in multiple views of the data.

In: Proceedings international conference on data mining (ICDM), pp 529-538

[39] Nickel, M.; Murphy, K.; Tresp, V.; Gabrilovich, E., A review of relational machine learning for knowledge graphs, Proc IEEE, 104, 1, 11-33 (2016)

[40] Nielsen, F.; Nock, R., On the chi square and higher-order Chi distances for approximating f-divergences, IEEE Signal Process Lett, 21, 1, 10-13 (2014)

[41] Novak, PK; Lavrač, N.; Webb, GI, Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining, J Mach Learn Res, 10, 377-403 (2009) · Zbl 1235.68178

[42] Pearl, J., Probabilistic reasoning in intelligent systems (1988), Burlington: Morgan Kaufmann, Burlington

[43] Peralta V (2007) Extraction and integration of MovieLens and IMDb. Technical report. Alternative Project Delivery Methods

[44] Perovsek M, Vavpetic A, Cestnik B, Lavrac N (2013) A wordification approach to relational data mining. In: Proceedings DS, lecture notes in computer science, pp 141-154. Springer, Singapore. doi:10.1007/978-3-642-40897-7_10

[45] Poole D (2003) First-order probabilistic inference. In: Proceedings international joint conference on artificial intelligence, pp 985-991

[46] Ramaswamy S, Rastogi R, Shim K (2000) Efficient algorithms for mining outliers from large data sets. In: Proceedings ACM special interest group on management of data, pp 427-438. doi:10.1145/342009.335437

[47] Riahi F, Schulte O (2015a) Model-based outlier detection for object-relational data. In: Proceedings symposium series on computational intelligence. IEEE, pp 1590-1598. doi:10.1109/SSCI.2015.224

[48] Riahi F, Schulte O (2015b) Codes and datasets. ftp://ftp.fas.sfu.ca/pub/cs/oschulte/CodesAndDatasets/. Accessed 15 Nov 2016

[49] Riahi F, Schulte O (2016) Propositionalization for unsupervised outlier detection in multi-relational data. In: Proceedings international conference of the Florida artificial intelligence, Key Largo, Florida, USA, pp 448-453. http://www.aaai.org/ocs/index.php/FLAIRS/FLAIF Accessed 2 Jan 2017

[50] Riedel S, Yao L, McCallum A, Marlin BM (2013) Relation extraction with matrix factorization and universal schemas. In: Proceedings annual conference of the North American Chapter of the Association for Computational Linguistics, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA, pp 74-84. http://aclweb.org/anthology/N/N13/N13-1008.pdf

[51] Routley K, Schulte O (2015) A Markov game model for valuing player actions in ice hockey. In: Proceedings uncertainty in artificial intelligence, pp 782-791

[52] Sarawagi S, Agrawal R, Megiddo N (1998) Discovery-driven exploration of OLAP data cubes. In: Proceedings extending database technology, Valencia, Spain, pp 168-182. Springer, Berlin. doi:10.1007/BFb0100984

[53] Schulte O (2011) A tractable pseudo-likelihood function for Bayesian networks applied to relational data. In: Proceedings society for industrial and applied mathematics, pp 462-473. doi:10.1137/1.9781611972818.40

[54] Schulte O, Gholami S (2017) Locally consistent Bayesian network scores for multi-relational data. In: Proceedings international joint conference on artificial intelligence, Melbourne, Australia, pp 2693-2700. doi:10.24963/ijcai.2017/375

[55] Schulte, O.; Khosravi, H., Learning graphical models for relational data via lattice search, Mach Learn, 88, 3, 331-368 (2012)

[56] Schulte O, Routley K (2014) Aggregating predictions versus aggregating features for relational classification. In: Proceedings center for information-development management, Orlando, FL, USA, pp 121-128. IEEE. doi:10.1109/CIDM.2014.7008657

[57] Schulte, O.; Khosravi, H.; Kirkpatrick, A.; Gao, T.; Zhu, Y., Modelling relational statistics with Bayesian networks, Mach Learn, 94, 105-125 (2014) · Zbl 1319.68190

[58] Sing T, Sander O, Beerenwinkel N, Lengauer T (2012) ROCR: visualizing the performance of scoring classifiers. http://cran.r-project.org/package=ROCR. Accessed 15 Nov 2016

[59] Sun Y, Han J, Zhao P (2009) Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings extending database technology, New York, NY, USA, pp 565-576. ACM

[60] Tang G, Bailey J, Pei J, Dong G (2013) Mining multidimensional contextual outliers from categorical relational data. In: Proceedings scientific and statistical database management conference, pp 1171-1192. doi:10.3233/IDA-150764

[61] Tuffery S (2011) Data mining and statistics for decision making. Wiley series in computational statistics. http://ca.wiley.com/WileyCDA/WileyTitl 0470688297.html. Accessed 15 Nov 2016 · Zbl 1216.62005

[62] Wang DZ, Michelakis E, Garofalakis M, Hellerstein JM (2008) BayesStore: managing large, uncertain data repositories with probabilistic graphical models. In: Proceedings very large data bases. VLDB Endowment, pp 340-351. doi:10.14778/1453856.1453896. http://www.vldb.org/pvldb/1/1453896.pdf. Accessed 15 Nov 2016

[63] Xiang R, Neville J (2011) Relational learning with one network: an asymptotic analysis. In: Proceedings artificial intelligence and statistics, pp 779-788. http://proceedings.mlr.press/v15/xiang11a/xiang11a.pdf. Accessed 15 Nov 2016