Duivesteijn, Wouter; **Feelders, Ad J.**; **Knobbe, Arno**

Exceptional model mining. Exceptional model mining, supervised descriptive local pattern mining with complex target concepts. (English) Zbl 1411.68096

Data Min. Knowl. Discov. 30, No. 1, 47-98 (2016).

Summary: Finding subsets of a dataset that somehow deviate from the norm, i.e. where something interesting is going on, is a classical Data Mining task. In traditional local pattern mining methods, such deviations are measured in terms of a relatively high occurrence (frequent itemset mining), or an unusual distribution for one designated target attribute (common use of subgroup discovery). These, however, do not encompass all forms of "interesting". To capture a more general notion of interestingness in subsets of a dataset, we develop Exceptional Model Mining (EMM). This is a supervised local pattern mining framework, where several target attributes are selected, and a model over these targets is chosen to be the target concept. Then, we strive to find subgroups: subsets of the dataset that can be described by a few conditions on single attributes. Such subgroups are deemed interesting when the model over the targets on the subgroup is substantially different from the model on the whole dataset. For instance, we can find subgroups where two target attributes have an unusual correlation, a classifier has a deviating predictive performance, or a Bayesian network fitted on several target attributes has an exceptional structure. We give an algorithmic solution for the EMM framework, and analyze its computational complexity. We also discuss some illustrative applications of EMM instances, including using the Bayesian network model to identify meteorological conditions under which food chains are displaced, and using a regression model to find the subset of households in the Chinese province of Hunan that do not follow the general economic law of demand.

MSC:

68T05 Learning and adaptive systems in artificial intelligence
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

exceptional model mining; subgroup discovery; supervised local pattern mining; regression; Bayesian networks

Full Text: DOI

References:

[1] Agresti A (1990) Categorical data analysis. Wiley, New York · Zbl 0716.62001

[2] Aidt T, Tzannatos Z (2002) Unions and collective bargaining. The World Bank, Washington, DC

[3] Anglin, PM; Gençay, R., Semiparametric estimation of a hedonic price function, J Appl Econ, 11, 633-648, (1996)

[4] Atzmüller M, Lemmerich F (2009) Fast subgroup discovery for continuous target concepts. In: Proceedings of ISMIS, pp 35-44

[5] Bay, SD; Pazzani, MJ, Detecting group differences: mining contrast sets, Data Min Knowl Discov, 5, 213-246, (2001) · Zbl 0982.68048

[6] Blockeel H, De Raedt L, Ramon J (1998) Top-down induction of clustering trees. In: Procedings of ICML, pp 55-63

[7] Boley M, Grosskreutz H (2009) Non-redundant subgroup discovery using a closure system. In: Proceedings of ECML/PKDD, vol 1, pp 179-194

[8] Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey · Zbl 0541.62042

[9] Campos, LM; Fernández-Luna, JM; Huete, JF, Bayesian networks and information retrieval: an introduction to the special issue, Inf Process Manag, 40, 727-733, (2004)

[10] Carmona, CJ; González, P.; Jesus, MJ; Herrera, F., NMEEF-SD: non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery, IEEE Trans Fuzzy Syst, 18, 958-970, (2010)

[11] Chao, C.; Velicer, C.; Slezak, JM; Jacobsen, SJ, Correlates for completion of 3-dose regimen of HPV vaccine in female members of a managed care organization, Mayo Clin Proc, 84, 864-870, (2009)

[12] Cook, RD, Detection of influential observation in linear regression, Technometrics, 19, 15-18, (1977) · Zbl 0371.62096

[13] Cook, RD; Weisberg, S., Characterizations of an empirical influence function for detecting influential cases in regression, Technometrics, 22, 495-508, (1980) · Zbl 0453.62051

[14] Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman & Hall, London · Zbl 0564.62054

[15] Costanigro, M.; Mittelhammer, RC; McCluskey, JJ, Estimating class-specific parametric models under class uncertainty: local polynomial regression clustering in an hedonic analysis of wine markets, J Appl Econ, 24, 1117-1135, (2009)

[16] Davis, GA, Bayesian reconstruction of traffic accidents, Law Probab Risk, 2, 69-89, (2003)

[17] Díez, FJ; Mira, J.; Iturralde, E.; Zubillaga, S., DIAVAL, a Bayesian expert system for echocardiography, Artif Intell Med, 10, 59-73, (1997)

[18] Dong G, Li J (1999) Efficient mining of emerging patterns: discovering trends and differences. In: Proceedings of KDD, pp 43-52

[19] Dougherty C (2011) Introduction to econometrics, 4th edn. Oxford University Press, Oxford

[20] Duivesteijn W, Feelders A, Knobbe AJ (2012) Different slopes for different folks—mining for exceptional regression models with Cook's distance. In: Proceedings of KDD, pp 868-876

[21] Duivesteijn W, Knobbe AJ, Feelders A, van Leeuwen M (2010) Subgroup discovery meets Bayesian networks—an exceptional model mining approach. In: Proceedings of ICDM, pp 158-167

[22] Duivesteijn W, Loza Mencía E, Fürnkranz J, Knobbe AJ (2012) Multi-label LeGo—enhancing multi-label classifiers with local patterns. In: Proceedings of IDA, pp 114-125

[23] Friedman, J.; Fisher, N., Bump-hunting in high-dimensional data, Stat Comput, 9, 123-143, (1999)

[24] Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D., Using Bayesian networks to analyze expression data, J Comput Biol, 7, 601-620, (2000)

[25] Galbrun, E.; Miettinen, P., From black and white to full color: extending redescription mining outside the Boolean world, Stat Anal Data Min, 5, 284-303, (2012)

[26] Garriga GC, Heikinheimo H, Seppänen JK (2007) Cross-mining binary and numerical attributes. In: Proceedings of ICDM, pp 481-486

[27] Gallo A, Miettinen P, Mannila H (2008) Finding subgroups having several descriptions: algorithms for redescription mining. In: Proceedings of SDM, pp 334-345

[28] Gentleman, JF; Wilk, MB, Detecting outliers II: supplementing the direct analysis of residuals, Biometrics, 31, 387-410, (1975) · Zbl 0322.62084

[29] Goodman, LA, The multivariate analysis of qualitative data: interaction among multiple classifications, J Am Stat Assoc, 65, 226-256, (1970)

[30] Grosskreutz, H.; Rüping, S., On subgroup discovery in numerical domains, Data Min Knowl Discov, 19, 210-226, (2009)

[31] Hand DJ, Adams NM, Bolton RJ (2002) Pattern detection and discovery, vol 2447. Lecture notes in computer science, Springer, Berlin · Zbl 1007.68737

[32] Heckerman, D.; Geiger, D.; Chickering, DM, Learning Bayesian networks: the combination of knowledge and statistical data, Mach Learn, 20, 197-243, (1995) · Zbl 0831.68096

[33] Heikinheimo, H.; Fortelius, M.; Eronen, J.; Mannila, H., Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters, J Biogeogr, 34, 1053-1064, (2007)

[34] Herrera, F.; Carmona, CJ; González, P.; Jesus, MJ, An overview on subgroup discovery: foundations and applications, Knowl Inf Syst, 29, 495-525, (2011)

[35] Hochberg Y, Tamhane A (1987) Multiple comparison procedures. Wiley, New York · Zbl 0731.62125

[36] Jensen, RT; Miller, NH, Giffen behavior and subsistence consumption, Am Econ Rev, 98, 1553-1577, (2008)

[37] Jesús, MJ; González, P.; Herrera, F.; Mesonero, M., Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing, IEEE Trans Fuzzy Syst, 15, 578-592, (2007)

[38] Jorge AM, Azevedo PJ, Pereira F (2006) Distribution rules with numeric attributes of interest. In: Proceedings of PKDD, pp 247-258

[39] Klösgen W (1996) Explora: a multipattern and multistrategy discovery assistant. In: Advances in knowledge discovery and data mining. pp 249-271

[40] Klösgen W (1998) Deviation and association patterns for subgroup mining in temporal, spatial, and textual data bases. In: Rough sets and current trends in computing. Springer, pp 1-18

[41] Klösgen W (1999) Applications and research problems of subgroup mining. In: Proceedings of ISMIS, pp 1-15

[42] Klösgen W (2002) Subgroup discovery. In: Handbook of data mining and knowledge discovery, chap. 16.3. Oxford University Press, New York

[43] Knobbe AJ, Feelders A, Leman D (2012) Exceptional model mining. In: Data mining: foundations and intelligent paradigms, intelligent systems reference library, vol 24, pp 183-198 · Zbl 1231.68206

[44] Knuth DE (1998) The art of computer programming, vol. 3: sorting and searching, 2nd edn. Addison-Wesley, Reading

[45] Kocev, D.; Vens, C.; Struyf, J.; Džeroski, S., Tree ensembles for predicting structured outputs, Pattern Recogn, 46, 817-833, (2013)

[46] Kohavi R (1995) The power of decision tables. In: Proceedings of ECML, pp 174-189

[47] van de Koppel E, Slavkov I, Astrahantseff K, Schramm A, Schulte J, Vandesompele J, de Jong E, Dzeroski S, Knobbe AJ (2007) Knowledge discovery in neuroblastoma-related biological data. In: Data mining in functional genomics and proteomics workshop at PKDD 2007, Warsaw, Poland, pp 45-56

[48] Kralj Novak, P.; 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

[49] Kriegel H-P, Kröger P, Schubert E, Zimek A (2012) Outlier detection in arbitrarily oriented subspaces. In: Proceedings of ICDM, pp 379-388

[50] Lavrač N, Flach P, Zupan B (1999) Rule evaluation measures: a unifying view. In: Proceedings of the ninth international workshop on inductive logic programming. Lecture notes in artificial intelligence, vol 1634, pp 174-185

[51] Lavrač, N.; Kavšek, B.; Flach, PA; Todorovski, L., Subgroup discovery with CN2-SD, J Mach Learn Res, 5, 153-188, (2004)

[52] Leeuwen, M., Maximal exceptions with minimal descriptions, Data Min Knowl Discov, 21, 259-276, (2010)

[53] van Leeuwen M, Knobbe AJ (2011) Non-redundant subgroup discovery in large and complex data. In: Proceedings of ECML/PKDD, vol 3, pp 459-474

[54] Leeuwen, M.; Knobbe, AJ, Diverse subgroup set discovery, Data Min Knowl Discov, 25, 208-242, (2012)

[55] Leman D, Feelders A, Knobbe AJ (2008) Exceptional model mining. In: Proceedings of ECML/PKDD, vol 2, pp 1-16 · Zbl 1231.68206

[56] Lemmerich F, Becker M, Atzmüller M (2012) Generic pattern trees for exhaustive exceptional model mining. In: Proceedings of ECML/PKDD, vol 2, pp 277-292

[57] Mampaey M, Nijssen S, Feelders A, Knobbe AJ (2012) Efficient algorithms for finding richer subgroup descriptions in numeric and nominal data. In: Proceedings of ICDM, pp 499-508

[58] Marshall A (1895) Principles of economics. MacMillan and co, New York

[59] Meeng M, Knobbe AJ (2011) Flexible enrichment with Cortana—Software Demo. In: Proceedings of Benelearn, pp 117-119

[60] Mitchell-Jones T et al (1999) The atlas of European mammals. Poyser natural history. Poyser, London

[61] Moore D, McCabe G (1993) Introduction to the practice of statistics. WH Freeman and Company, New York

[62] Morik K, Boulicaut JF, Siebes A (2005) Local pattern detection. Lecture notes in computer science, vol 3539, Springer, Heidelberg

[63] Neil, M.; Fenton, N.; Tailor, M., Using Bayesian networks to model expected and unexpected operational losses, Risk Anal, 25, 963-972, (2005)

[64] Neter J, Kutner M, Nachtsheim CJ, Wasserman W (1966) Applied linear statistical models. WCB McGraw-Hill, Boston

[65] Paine, RT, Food web complexity and species diversity, Am Nat, 100, 65-75, (1966)

[66] Ramakrishnan N, Kumar D, Mishra B, Potts M, Helm RF (1995) Turning CARTwheels: an alternating algorithm for mining redescriptions. In: Proceedings of KDD, pp 837-844

[67] Rezende, L., Econometrics of auctions by least squares, J Appl Econ, 23, 925-948, (2008)

[68] Scholz M (2005) Knowledge-based sampling for subgroup discovery. In: Morik K, Boulicaut JF, Siebes A (eds) Local pattern detection. Lecture notes in computer science, vol 3539, Springer, Heidelberg, pp 171-189

[69] Schubert E, Wolfe J, Tarnopolsky A (2004) Spectral centroid and timbre in complex, multiple instrumental textures. In: Proceedings of 8th international conference on music perception & cognition, pp 654-657

[70] Siebes A (1995) Data surveying: foundations of an inductive query language. In: Proceedings of KDD, pp 269-274

[71] Stengos, T.; Zacharias, E., Intertemporal pricing and price discrimination: a semiparametric hedonic analysis of the personal computer market, J Appl Econ, 21, 371-386, (2006)

[72] Trohidis K, Tsoumakas G, Kalliris G, Vlahavas IP (2008) Multi-label classification of music into emotions. In: Proceedings of 9th international conference on music information retrieval, pp 325-330

[73] Umek, L.; Zupan, B., Subgroup discovery in data sets with multi-dimensional responses, Intell Data Anal, 15, 533-549, (2011)

[74] Verma T, Pearl J (1990) Equivalence and synthesis of causal models. In: Proceedings of UAI, pp 255-270

[75] Whittaker J (1990) Graphical models in applied multivariate statistics. Wiley, New York · Zbl 0732.62056

[76] Wrobel S (1997) An algorithm for multi-relational discovery of subgroups. In: Proceedings of PKDD, pp 78-87

[77] Yang G, Le Cam L (2000) Asymptotics in statistics: some basic concepts. Springer, Berlin · Zbl 0952.62002

[78] Zhang B (2003) Regression clustering. In: Proceedings of ICDM, pp 451-458

[79] Zimmermann, A.; Raedt, L., Cluster-grouping: from subgroup discovery to clustering, Mach Learn, 77, 125-159, (2009)