Appendix A: Bibliography with Machine Learning Literature and Remaining References

Only literature referenced from the "Applications in behavioral sciences" subsections was included into the main bibliography. This appendix contains the remaining references.

Agrawal, Mayank, Joshua C Peterson, and Thomas L Griffiths (2019). "Using Machine Learning to Guide Cognitive Modeling: A Case Study in Moral Reasoning". In: *arXiv preprint arXiv:1902.06744*.

Agrawal, Rakesh, Tomasz Imielinski, and Arun Swami (1993). "Mining association rules between sets of items in large databases". In: *ACM SIGMOD record*. Vol. 22. ACM, pp. 207–216.

Agrawal, Rakesh and Ramakrishnan Srikant (1995). "Mining sequential patterns". In: *Data Engineering, 1995. Proceedings of the Eleventh International Conference on*. IEEE, pp. 3–14.

Alcala-Fdez, Jesús, Rafael Alcala, and Francisco Herrera (2011). "A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning". In: *IEEE Transactions on Fuzzy systems* 19.5, pp. 857–872.

Andrews, Robert, Joachim Diederich, and Alan B. Tickle (1995). "Survey and critique of techniques for extracting rules from trained artificial neural networks". In: *Knowledge-Based Systems* 8.6, pp. 373–389.

Arnulf, Jan Ketil, Kai Rune Larsen, Øyvind Lund Martinsen, and Chih How Bong (2014). "Predicting survey responses: How and why semantics shape survey statistics on organizational behaviour". In: *PloS one* 9.9, e106361.

Atzmüller, Martin (2015). "Subgroup discovery". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5.1, pp. 35–49.

Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani (2010). "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." In: *Lrec*. Vol. 10, pp. 2200–2204.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 238–247.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3. Jan, pp. 993–1022.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992). "A training algorithm for optimal margin classifiers". In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, pp. 144–152.

Breese, John S., David Heckerman, and Carl Kadie (1998). "Empirical Analysis of Predictive Algorithms for Collaborative Filtering". In: *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI-98)*. Ed. by G.F. Cooper and S. Moral. Madison, WI: Morgan Kaufmann, pp. 43–52.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32.

Breiman, Leo, Jerome Friedman, Charles J. Stone, R. A. Olshen (1984). *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Brooks.

Buchanan, E (2010). "Access into memory: Differences in judgments and priming for semantic and associative memory". In: *Journal of Scientific Psychology* 1, pp. 1–8.

Cimiano, Philipp, Antje Schultz, Sergej Sizov, Philipp Sorg, Steffen Staab (2009). "Explicit Versus Latent Concept Models for Cross-Language Information Retrieval". In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09)*, pp. 1513–1518.

Cohen, William W. (1995). "Fast Effective Rule Induction". In: *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*. ICML'95. Tahoe City, California, USA: Morgan Kaufmann Publishers Inc., pp. 115–123.

Dash, Sanjeeb, Oktay Günlük, and Dennis Wei (2018). "Boolean Decision Rules via Column Generation". In: *Advances in Neural Information Processing Systems 31 (NeurIPS-18)*. Ed. by Samy Bengio et al. Montréal, Canada, pp. 4660–4670.

Dojchinovski, Milan, Dinesh Reddy, Tomás Kliegr, Tomas Vitvar, Harald Sack (2016). "Crowdsourced Corpus with Entity Salience Annotations." In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC-16)*.

Duchi, John C., Elad Hazan, and Yoram Singer (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic

Optimization". In: *Journal of Machine Learning Research* 12, pp. 2121–2159.

Dumais, Susan and Hao Chen (2000). "Hierarchical classification of Web content". In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pp. 256– 263.

Fellbaum, Christiane (2010). "WordNet". In: *Theory and applications of ontology: computer applications*. Springer, pp. 231–243.

Fernandez-Delgado, Manuel, Eva Cernadas, Senén Barro, and Dinani Amorim (2014). "Do we need hundreds of classifiers to solve real world classification problems?" In: *The Journal of Machine Learning Research* 15.1, pp. 3133–3181.

Friedman, Jerome H. and Nicholas I. Fisher (1999). "Bump Hunting in HighDimensional Data". In: *Statistics and Computing* 9.2, pp. 123–143.

Frosst, Nicholas and Geoffrey E. Hinton (2017). "Distilling a Neural Network Into a Soft Decision Tree". In: *Proceedings of the 1st AI*AI International Workshop on Comprehensibility and Explanation in AI and ML*. Ed. by Tarek R. Besold and Oliver Kutz. Vol. 2071. CEUR Workshop Proceedings. Bari, Italy: CEUR-WS.org.

Fürnkranz, Johannes (1997). "Pruning Algorithms for Rule Learning". In: *Machine Learning* 27.2, pp. 139–171.

Fürnkranz, Johannes, Dragan Gamberger, and Nada Lavrač (2012). *Foundations of Rule Learning*. Springer-Verlag.

Fürnkranz, Johannes and Eyke Hüllermeier, eds. (2010). *Preference Learning*. Springer-Verlag.

Fürnkranz, Johannes, Tomás Kliegr, and Heiko Paulheim (2018). "On Cognitive Preferences and the Interpretability of Rule-based Models". In: *arXiv preprint arXiv:1803.01316*.

Gabrilovich, Evgeniy and Shaul Markovitch (2007). "Computing semantic relatedness using Wikipedia-based explicit semantic analysis". In: *Proceedings of the 20th international joint conference on Artifical intelligence*. IJCAI'07. Hyderabad, India: Morgan Kaufmann Publishers Inc., pp. 1606–1611.

Gamon, Michael, Tae Yano, Xinying Song, Johnson Apacible, and Patrick Pantel (2013). "Identifying salient entities in web pages". In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM, pp. 2375–2380.

Gandomi, Amir and Murtaza Haider (2015). “Beyond the hype: Big data concepts, methods, and analytics”. In: *International Journal of Information Management* 35.2, pp. 137–144.

García, David, Antonio González, and Raúl Pérez (2014). “Overview of the SLAVE learning algorithm: A review of its evolution and prospects”. In: *International Journal of Computational Intelligence Systems* 7.6, pp. 1194– 1221.

Gemmis, Marco de, Leo Iaquinta, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Giovanni Semeraro: (2010). “Learning Preference Models in Recommender Systems”. In: *Preference Learning*. Ed. by Johannes Fürnkranz and Eyke Hüllermeier. Springer-Verlag, pp. 387–407. isbn: 978-3642141249.

González, Camila, Eneldo Loza Mencía, and Johannes Fürnkranz (2017). “Retraining Deep Neural Networks to Facilitate Boolean Concept Extraction”. In: *Proceedings of the 20th International Conference on Discovery Science (DS-17)*. Vol. 10558. Lecture Notes in Computer Science. Springer-Verlag, pp. 127–143.

Goodfellow, Ian J., Yoshua Bengio, and Aaron C. Courville (2016). *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press. isbn: 978-0262-03561-3.

Hájek, Petr, Martin Holeňa, and Jan Rauch (2010). “The GUHA method and its meaning for data mining”. In: *Journal of Computer and System Sciences* 76.1, pp. 34–48.

Han, Eui-Hong Sam and George Karypis (2000). “Centroid-based document classification: Analysis and experimental results”. In: *European conference on principles of data mining and knowledge discovery*. Springer, pp. 424–431.

Harris, Steve, Andy Seaborne, and Eric Prud’hommeaux (2013). “SPARQL 1.1 query language”. In: *W3C recommendation* 21.10.

Hashem, Ibrahim A. T., Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan (2015). “The rise of “big data” on cloud computing: Review and open research issues”. In: *Information systems* 47, pp. 98–115.

Herrera, Francisco, Cristóbal J. Carmona, Pedro González, and María José del Jesús (2011). “An overview on subgroup discovery: foundations and applications”. In: *Knowledge and information systems* 29.3, pp. 495–525.

Hinton, G. E., James L. McClelland, and David E. Rumelhart (1986). “Distributed Representations”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*.

Ed. by David E. Rumelhart, James L. McClelland, and CORPORATE PDP Research Group. Cambridge, MA, USA: MIT Press, pp. 77–109.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Hornik, Kurt (1991). "Approximation capabilities of multilayer feedforward networks". In: *Neural Networks* 4.2, pp. 251–257.

Hühn, Jens and Eyke Hüllermeier (2009). "FURIA: an algorithm for unordered fuzzy rule induction". In: *Data Mining and Knowledge Discovery* 19.3, pp. 293– 319.

Jannach, Dietmar Markus Zanker, Alexander Felfernig, and Gerhard Friedrich (2010). *Recommender Systems: An Introduction*. Cambridge, UK: Cambridge University Press.

Johns, Brendan T and Michael N Jones (2012). "Perceptual inference through global lexical similarity". In: *Topics in Cognitive Science* 4.1, pp. 103–120.

Kamishima, Toshihiro, Hideto Kazawa, and Shotaro Akaho (2010). "A Survey and Empirical Comparison of Object Ranking Methods". In: *Preference Learning*. Ed. by Johannes Fürnkranz and Eyke Hüllermeier. Springer-Verlag, pp. 181–201. isbn: 978-3642141249.

Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data". In: *Applied Statistics* 29, pp. 119–127.

Kitchin, Rob (2017). "Big data-Hype or revolution". In: *The SAGE handbook of social media research methods*, pp. 27–39.

Kralj Novak, Petra, Nada Lavrac, Geoffrey I. Webb (2009). "Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining ". *Journal of Machine Learning Research* 10, pp. 377–403.

Landauer, Thomas K. and Susan T. Dumais (1997). "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." In: *Psychological review* 104.2, p. 211.

Landauer, Thomas K, Danielle S McNamara, Simon Dennis, and Walter Kintsch (2013). *Handbook of latent semantic analysis*. Psychology Press.

Lantz, Brett (2015). *Machine learning with R*. Packt Publishing Ltd.

Lecun, Yann, Yoshua Bengio, and Geoffrey E. Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer: (2015). "DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia". In: *Semantic Web* 6.2, pp. 167–195.

Lin, Dekang (1998). "An Information-Theoretic Definition of Similarity". In: *Proceedings of the Fifteenth International Conference on Machine Learning*. ICML '98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 296–304.

Liu, Bing (2011). *Web data mining: exploring hyperlinks, contents, and usage data, 2nd ed.* Springer Science & Business Media.

Liu, Bing, Wynne Hsu, and Yiming Ma (1998). "Integrating Classification and Association Rule Mining". In: *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. KDD'98. New York, NY: AAAI Press, pp. 80–86.

Malioutov, Dmitry and Kuldeep S. Meel (2018). "MLIC: A MaxSAT-Based Framework for Learning Interpretable Classification Rules". In: *Proceedings of the 24th International Conference on Principles and Practice of Constraint Programming (CP-18)*. Ed. by John N. Hooker. Vol. 11008. Lecture Notes in Computer Science. Lille, France: Springer, pp. 312–327.

McKay, Dean, Jonathan S Abramowitz, and Eric A Storch (2017). *Treatments for Psychological Problems and Syndromes*. John Wiley & Sons.

Mihalcea, Rada and Andras Csomai (2007). "Wikify!: linking documents to encyclopedic knowledge". In: *Proceedings of the 16th ACM Conference on Information and Knowledge Management*. ACM, pp. 233–242.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*, pp. 3111–3119.

Minsky, Marvin and Seymour A. Papert (1969). *Perceptrons : Introduction to Computational Geometry*. Expanded Edition 1990. MIT Press.

Mitchell, Tom (1997). *Machine Learning*. McGraw-Hill Education.

Navigli, Roberto (2009). "Word sense disambiguation: A survey". In: *ACM computing surveys (CSUR)* 41.2, p. 10.

O'Dea, Bridianne, Mark E. Larson, Philip J. Batterham, Alison L. Calear, and Helen Christensen (2017). "A linguistic analysis of suicide-related Twitter posts". In: *Crisis 38,* pp. 319–329.

Pang, Guansong, Huidong Jin, and Shengyi Jiang (2015). "CenKNN: a scalable and effective text classifier". In: *Data Mining and Knowledge Discovery* 29.3, pp. 593–625.

Paulheim, Heiko (2018). "Machine learning with and for semantic web knowledge graphs". In: *Reasoning Web International Summer School*. Springer, pp. 110–141.

Peharz, Robert, Robert Gens, Franz Pernkopf, and Pedro M. Domingos (2017). "On the Latent Variable Interpretation in Sum-Product Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.10, pp. 2030–2044.

Pennebaker, James W., Ryan L. Boyd, Kayla Jordan, and Kate Blackburn (2015). *The development and psychometric properties of LIWC2015*. Tech. rep. LIWC.net, Austin, Texas.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532– 1543.

Pirro, Giuseppe and Nuno Seco (2008). "Design, Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content". In: *Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. Part II on On the Move to Meaningful Internet Systems*. OTM '08. Monterrey, Mexico: Springer-Verlag, pp. 1271–1288.

Quinlan, John Ross (1986). "Induction of Decision Trees". In: *Machine Learning* 1, pp. 81–106.

Rauch, Jan and Milan Simunek (2017). "Apriori and GUHA–Comparing two approaches to data mining with association rules". In: *Intelligent Data Analysis* 21.4, pp. 981–1013.

Resnik, Philip (1995). "Using Information Content to Evaluate Semantic Similarity in a Taxonomy". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453.

Ribeiro, Marco Túlio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings*

of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-16). Ed. by Balaji Krishnapuram et al. San Francisco, CA, USA: ACM, pp. 1135–1144.

Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.

Rumelhart, David E., Geoffrey E. Hinton, and R. Williams (1986). "Learning Internal Representations by Error Propagation". In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Ed. by D.E. Rumelhart and J. McClelland. Vol. 1: Foundations. Cambridge, MA: MIT Press, pp. 318–363.

Sammut, Claude (1996). "Automatic Construction of Reactive Control Systems Using Symbolic Machine Learning". In: *Knowledge Engineering Review* 11.1, pp. 27–42.

Schäfer, Dirk and Eyke Hüllermeier (2018). "Dyad Ranking Using PlackettLuce Models Based on Joint Feature Representations". In: *Machine Learning* 107.5, pp. 903–941.

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview". In: *Neural Networks* 61, pp. 85–117.

Serrano-Guerrero, Jesus, José Angel Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. (2015). "Sentiment analysis: A review and comparative analysis of web services". In: *Information Sciences* 311, pp. 18–38.

Siddharthan, Advaith, Nicolas Cherbuin, Paul J. Eslinger, Kasia Kozlowska, Nora A. Murphy, Leroy Lowe (2018). "WordNet-feelings: A linguistic categorisation of human feelings". In: *arXiv preprint arXiv:1811.02435*.

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15.1, pp. 1929–1958.

Stecher, Julius, Frederik Janssen, and Johannes Fürnkranz (2016). "Shorter Rules Are Better, Aren't They?" In: *Proceedings of the 19th International Conference on Discovery Science (DS-16)*. Ed. by Toon Calders, Michelangelo Ceci, and Donato Malerba. Springer-Verlag, pp. 279–294.

Tjong Kim Sang, Erik F and Fien De Meulder (2003). "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". In: *Proceedings of the seventh conference on*

Natural language learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics, pp. 142–147.

Tonidandel, Scott, Eden B King, and Jose M Cortina (2018). "Big data methods: Leveraging modern data analytic techniques to build organizational science". In: *Organizational Research Methods* 21.3, pp. 525–547.

Torgo, Luís (2010). *Data Mining with R: Learning with Case Studies*. Chapman and Hall/CRC Press.

Turney, Peter D and Patrick Pantel (2010). "From frequency to meaning: Vector space models of semantics". In: *Journal of artificial intelligence research* 37, pp. 141–188.

Tversky, Amos (1977). "Features of similarity". In: *Psychological Review* 84, pp. 327–352.

Varian, Hal R (2014). "Big data: New tricks for econometrics". In: *Journal of Economic Perspectives* 28.2, pp. 3–28.

Vembu, Shankar and Thomas Gärtner (2010). "Label Ranking Algorithms: A Survey". In: *Preference Learning*. Ed. by Johannes Fürnkranz and Eyke Hüllermeier. Springer-Verlag, pp. 45–64. isbn: 978-3642141249.

Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol (2010). "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *Journal of Machine Learning Research* 11, pp. 3371–3408.

Vrandečic, Denny and Markus Krötzsch (2014). "Wikidata: a free collaborative knowledgebase". In: *Communications of the ACM* 57.10, pp. 78–85.

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, Perry MacNeille (2017). "A Bayesian Framework for Learning Rule Sets for Interpretable Classification". In: *Journal of Machine Learning Research* 18, 70:1–70:37.

Widrow, Bernard, David E. Rumelhart, and Michael A. Lehr (1994). "Neural Networks: Applications in Industry, Business and Science". In: *Communications of the ACM* 37.3, pp. 93–105.

Wrobel, Stefan (1997). "An algorithm for multi-relational discovery of subgroups". In: *European Symposium on Principles of Data Mining and Knowledge Discovery*. Springer, pp. 78–87.

Appendix B: Software

This appendix contains description of software packages implementing methods described in article *Recent Advances in Machine Learning for Behavioral Sciences*. The appendix has the same structure including section and subsection headings as the original article. The appendix focuses on R packages, but includes also other types of software, such as Python libraries and web-based systems. Names of R packages are typeset in a `monospace` font.

2 Tabular Data

2.1 Induction of Decision Trees

The R ecosystem has support for various decision tree induction algorithms scattered across multiple packages. The caret package provides a uniform interface to many classification and regression functions in R, including CHAID (package `CHAID`), CART (package `rpart`), random forests (package `party`), and also to the C5.0 family of algorithms, which includes C5.0 decision trees and C5.0 boosted trees ensemble (package `C50`). BigML is an easy-to-use (yet commercial) Machine Learning as a Service system with good support for visualizing decision trees (cf. Figure 5).

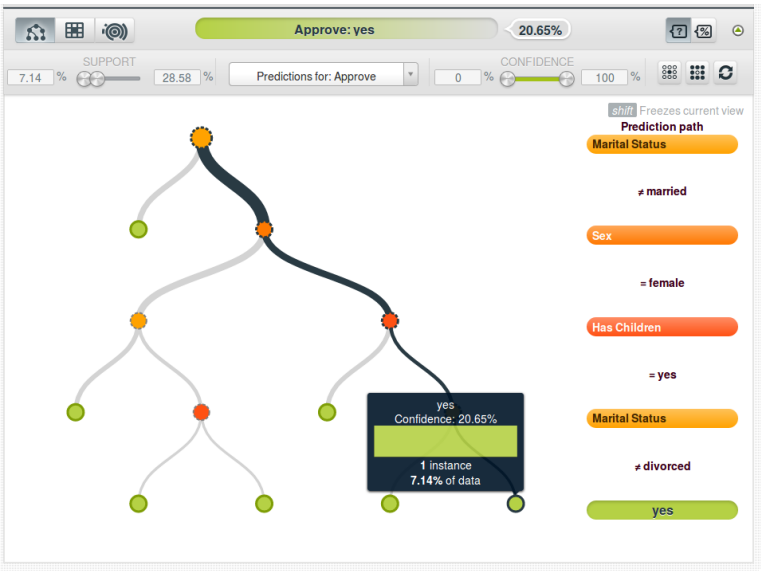

Figure 5: Decision tree induced from data included in Figure 1 by BigMLcom.

2.2 Induction of Predictive Rule Sets

A freely accessible re-implementation of RIPPER can be found in the Weka machine learning library under the name of JRip. It is also made available to R users via the caret package. The `caret` package also provides several versions of various fuzzy rule induction algorithms, including SLAVE. Implementations of the CBA algorithm are available via the `arc`, `arulesCBA` and `rCBA` packages. Web-based graphical interface to CBA is provided by EasyMiner (Vojíř et al., 2018), see also Figure 6. Additional rule learning algorithms can be used from R via package RWeka from desktop-oriented systems, such as Orange, RapidMiner, and Weka.

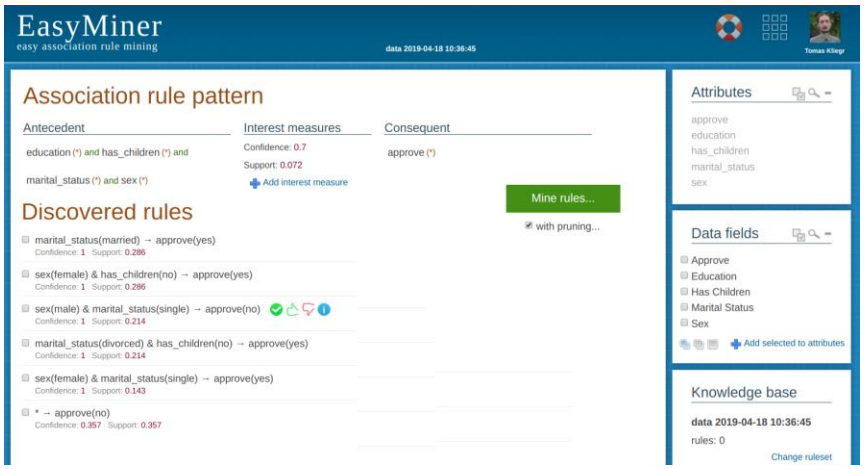

Figure 6: Predictive rule list induced from data included in Figure 1 with EasyMiner.eu. The "with pruning" option activated postprocessing of the discovered association rules into a predictive model by the CBA algorithm.

2.3 Discovering interesting patterns

The `arules` package is in the center of R ecosystem for association rule mining. Supplementary packages include

arulesViz for visualisation, arulesExplain for generating human readable explanations, and several packages with rule pruning capabilities (arc, arulesCBA and rCBA).

Especially for descriptive pattern mining, interactive graphical systems may provide an advantage over command line access offered by R. Interactive systems include EasyMiner (Vojíř et al., 2018), see also Figure 6, and LISp-Miner (Simunek, 2003), a desktop-based implementation of the GUHA method. Software for subgroup discovery includes R package rsubgroup and graphical system Vikamine (Atzmueller and Lemmerich, 2012).

2.4 Neural networks and deep learning

R package tensorFlow provides interface to the TensorFlow project (Abadi et al., 2016), a popular open source machine learning framework used for deep learning. Package keras provides interface to the comprehensive Keras project (Chollet et al., 2018), which can work with several machine learning systems (TensorFlow, CNTK (Seide and Agarwal, 2016), and Theano (Bergstra et al., 2011) focused on neural networks and deep learning.

3 Behavioral Data

3.1 Web Log and Mobile Usage Mining

Keyes, Rudis, and Jacobs (2016) present several R packages that can help with handling web server logs. For tracking users with Javascript-based systems, R package googleAnalyticsR can be used to retrieve data from the Google Analytics platform. For researchers that do not want to rely on Google Analytics, for example for privacy reasons (Chandler and Wallace, 2016), there are several open source systems with versatile tracking capabilities. These include Matomo (https://matomo.org/), which can perform both web and mobile app tracking, and Inbeat (https://www.inbeat.eu), which is a generic system focusing on collecting streaming user interaction data from sensor-based devices such as Microsoft Kinect.

Semantic description of data being interacted with can be obtained with web crawlers and scrapers, such as R package

`Rcrawler` (Khalil and Fakir, 2017). The `rgeolocate` R package serves for mapping IP addresses to regions.

There is an R package `arulesSequences` (Buchta, Hahsler, and Daniel Diaz, 2018) for extracting sequential association rules from data in the sequential format. Multiple R packages for analyzing time series data could also be applicable.

3.2 Preference learning

In some cases, preferences can be processed with algorithms for ordinal classification available in the caret R package. Specialized packages for preference models include `PlackettLuce` (Plackett-Luce model) and `BradleyTerry2` for Bradley-Terry models. Another implementation of Plackett-Luce is available in the `PLMIX` package (Mollica and Tardella, 2016).

4 Textual data

4.1 Word vectors and word embeddings

We have not found an ESA algorithm R package in CRAN. However, an easy-to-use ESA implementation that can be operated from the command line is available at https://github.com/ticcky.

For LSA, CRAN contains the package `lsa`, and a complementary package `LSAfun`. For LDA, `topicmodels` is a maintained package.

Word2vec and related predictive models are supported in general deep learning R packages `tensorFlow` and `keras` introduced earlier. Package `fastTextR` provides interface to the „Library for fast text representation and classification", which can be viewed as an evolution of word2vec. `text2vec` is a standalone package, which provides implementation of GloVe in addition to several related functionalities.

For many tasks, pretrained word embeddings might suffice, which are available from various sources, such as http://vectors.nlpl.eu/repository/.

4.2 Text annotation

R package supporting NER analysis is `cleanNLP`. This package offers two analysis backends, a Java-based Stanford CoreNLP and Python-based Spacy. The CoreNLP system is with over 4.000 citations to date a favourite choice across research disciplines. The newer Spacy system is claimed to have the fastest syntactic parser (https://spacy.io/usage/facts-figures). Both backends offer a range of other natural language processing functions beyond NER. Alternatives to `cleanNLP` include the `openNLP` R package, which provides access to functions in Apache Open NLP tools.

To the best of our knowledge, there are no R packages available in the official CRAN repository, which would provide entity linking to knowledge graphs, fine-grained entity classification ("wikification"), or computation of entity salience. However, there are several open source web-based systems, such as DBpedia Spotlight (Mendes et al., 2011) or EntityClassifier (Dojchinovski et al, 2017), which provide application programming interfaces (APIs) for which R-based wrappers are available. For entity salience, there are several Python packages as well as web applications providing this functionality. One example is the SWAT system (Ponza, Ferragina, and Piccinno, 2018), which also provides a convenient web interface (https://swat.d4science.org/).

For sentiment analysis, there are several packages in CRAN inluding: `syuzhet`, `SentimentAnalysis`, and `RSentiment`. Applicable is also the general purpose `coreNLP` package, which provides interface to Java implementation in Stanford CoreNLP. Outside the R ecosystem, LIWC (Pennebaker et al., 2015) provides a notable implementation of a sentiment analysis system used in many studies in behavioral sciences. SentiStrength (Thelwall, 2017) is another sentiment analysis system, considered as state-of-the-art by Saif et al. (2016). In last several years, web services for sentiment analysis have also gained on popularity, their review is provided by Serrano-Guerrero et al. (2015).

Finally, GATE NLP framework (https://gate.ac.uk/) is a representative of an integrated system providing coverage of most of the tasks described above. The system is written in Java but provides easy-to-use graphical user interface.

4.3 Document classification

An implementation of nearest centroid classifier is provided by R package `lolR`. Implementation of SVM is included in the `e1071` R package. Handling multilabel classification problems in R is described in (Probst et al., 2017).

4.4 Knowledge graphs

There is an official R package providing access to Wikidata API: `WikidataQueryServiceR`. For DBpedia, there is R package `datamart`, but it does not seem to be maintained for several years. Both packages allow to issue queries in the SPARQL language. Fundamentals of SPARQL can be learned at various online tutorials (e.g. https://www.w3.org/2009/Talks/0615-qbe/).

4.5 WordNet and related lexical resources

Access to WordNet is available in R via the CRAN package `wordnet`. Note that the setup of this package is somewhat more involved, since it requires setup of R-Java bindings and also manual installation of WordNet. WordNet can also be used online at the web site of Princeton University at https://wordnet.princeton.edu/.

To our knowledge, there is no package in CRAN which can directly compute word similarity with WordNet. There are, however, multiple implementations for Python. For example, the Pyhon sematch library (https://github.com/gsi-upm/sematch) also provides a freely accessible web interface (http://sematch.cluster.gsi.dit.upm.es/).

Software references

Abadi, Martín et al. (2016). "Tensorflow: A system for large-scale machine learning". In: *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283.

Atzmueller, Martin and Florian Lemmerich (2012). "VIKAMINE–open-source subgroup discovery, pattern mining, and analytics". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 842–845.

Bergstra, James et al. (2011). “Theano: Deep learning on GPUs with python”.
In: *NIPS 2011, BigLearning Workshop, Granada, Spain*. Vol. 3, pp. 1–48.

Buchta, Christian, Michael Hahsler, and with contributions from Daniel Diaz (2018). *arulesSequences: Mining Frequent Sequences*. R package version 0.2-20. url: https://CRAN.R-project.org/package=arulesSequences.

Dojchinovski, Milan, and Tomáš Kliegr. "Entityclassifier. eu: real-time classification of entities in text with Wikipedia." Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2013.

Chandler, Adam and Melissa Wallace (2016). “Using Piwik Instead of Google Analytics at the Cornell University Library”. In: *The Serials Librarian* 71.34, pp. 173–179.

Chollet, Francois et al. (2018). “Keras: The python deep learning library”. In: *Astrophysics Source Code Library*.

Keyes, Oliver, Bob Rudis, and Jay Jacobs (2016). “R Packages to Aid in Handling Web Access Logs.” In: *R Journal* 8.1.

Khalil, Salim and Mohamed Fakir (2017). “RCrawler: An R package for parallel web crawling and scraping”. In: *SoftwareX* 6, pp. 98–106.

Mendes, Pablo N et al. (2011). “DBpedia spotlight: shedding light on the web of documents”. In: *Proceedings of the 7th international conference on semantic systems*. ACM, pp. 1–8.

Mollica, Cristina and Luca Tardella (2016). “PLMIX: An R package for modeling and clustering partially ranked data”. In: *arXiv preprint arXiv:1612.08141*.

Pennebaker, James W et al. (2015). *The development and psychometric properties of LIWC2015*. Tech. rep. LIWC.net, Austin, Texas.

Ponza, Marco, Paolo Ferragina, and Francesco Piccinno (2018). “SWAT: A System for Detecting Salient Wikipedia Entities in Texts”. In: *arXiv preprint arXiv:1804.03580*.

Probst, Philipp et al. (2017). “Multilabel classification with R package mlr”. In: *arXiv preprint arXiv:1703.08991*.

Saif, Hassan et al. (2016). “Contextual Semantics for Sentiment Analysis of Twitter”. In: *Inf. Process. Manage.* 52.1, pp. 5–19.

Seide, Frank and Amit Agarwal (2016). “CNTK: Microsoft’s open-source deeplearning toolkit”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 2135–2135.

Serrano-Guerrero, Jesus et al. (2015). “Sentiment analysis: A review and comparative analysis of web services”. In: *Information Sciences* 311, pp. 18–38.

Simunek, Milan (2003). “Academic KDD project LISp-miner”. In: *Intelligent Systems Design and Applications*. Springer, pp. 263–272.

Thelwall, Mike (2017). “The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength”. In: *Cyberemotions: Collective Emotions in Cyberspace*. Cham: Springer International Publishing, pp. 119– 134.

Vojíř, Stanislav et al. (2018). “EasyMiner.eu: Web framework for interpretable machine learning based on rules and frequent itemsets”. In: *Knowledge-Based Systems* 150, pp. 111–115.