Śmieja, Marek; **Wiercioch, Magdalena**

Constrained clustering with a complex cluster structure. (English) Zbl 1414.62273

Adv. Data Anal. Classif., ADAC 11, No. 3, 493-518 (2017).

Summary: In this contribution we present a novel constrained clustering method, Constrained clustering with a complex cluster structure (C4s), which incorporates equivalence constraints, both positive and negative, as the background information. C4s is capable of discovering groups of arbitrary structure, e.g. with multi-modal distribution, since at the initial stage the equivalence classes of elements generated by the positive constraints are split into smaller parts. This provides a detailed description of elements, which are in positive equivalence relation. In order to enable an automatic detection of the number of groups, the cross-entropy clustering is applied for each partitioning process. Experiments show that the proposed method achieves significantly better results than previous constrained clustering approaches. The advantage of our algorithm increases when we are focusing on finding partitions with complex structure of clusters.

MSC:

62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:

constrained clustering; model-based clustering; mixture of models; pairwise equivalence constraints; semi-supervised learning; cross-entropy clustering

Software:

ilastik; AS 136; UCI-ml

Full Text: DOI

References:

[1] Arbelaez, P.; Maire, M.; Fowlkes, C.; Malik, J., Contour detection and hierarchical image segmentation, IEEE Trans Pattern Anal Mach Intell, 33, 898-916, (2011)· doi:10.1109/TPAMI.2010.161

[2] Bar-Hillel A, Hertz T, Shental N, Weinshall D (2003) Learning distance functions using equivalence relations. In: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003. DC, USA, AAAI Press, Washington, pp 11-18 · Zbl 1222.68140

[3] Basu S, Banerjee A, Mooney RJ (2002) Semi-supervised clustering by seeding. In: Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), July 8-12, 2002. Australia, Morgan Kaufmann, Sydney, pp 27-34

[4] Baudry, JP; Cardoso, M.; Celeux, G.; Amorim, M.; Ferreira, A., Enhancing the selection of a model-based clustering with external categorical variables, Adv Data Anal Classif, 9, 177-196, (2015)· doi:10.1007/s11634-014-0177-3

[5] Bellas, A.; Bouveyron, C.; Cottrell, M.; Lacaille, J., Model-based clustering of high-dimensional data streams with online mixture of probabilistic PCA, Adv Data Anal Classif, 7, 281-300, (2013) · Zbl 1273.62137 · doi:10.1007/s11634-013-0133-7

[6] Bennett KP, Demiriz A (1998) Semi-supervised support vector machines. In: Advances in Neural Information Processing Systems, MIT Press, Cambridge, pp 368-374

[7] Bilenko M, Basu S, Mooney RJ (2004) Integrating constraints and metric learning in semi-supervised clustering. In: Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), July 4-8, 2004, Banff, Alberta, Canada, ACM, New York, NY, USA, p 11. doi:10.1145/1015330.1015360

[8] Cayton L (2005) Algorithms for manifold learning. University of California at San Diego Tech Rep, pp 1-17

[9] Collingwood EF, Lohwater AJ (2004) The theory of cluster sets. Cambridge University Press, Cambridge

[10] Ding JJ, Wang YH, Hu LL, Chao WL, Shau YW (2011) Muscle injury determination by image segmentation. In: Visual Communications and Image Processing (VCIP), 2011 IEEE, pp 1-4. doi:10.1109/VCIP.2011.6115925

[11] Hartigan, JA; Wong, MA, Algorithm AS 136: a k-means clustering algorithm, J R Stat Soc Ser C (Appl Stat), 28, 100-108, (1979) · Zbl 0447.62062

[12] Hennig, C., Methods for merging Gaussian mixture components, Adv Data Anal Classif, 4, 3-34, (2010) · Zbl 1306.62141 · doi:10.1007/s11634-010-0058-3

[13] Hruschka ER, Campello RJGB, Freitas AA, De Carvalho ACPLF (2009) A survey of evolutionary algorithms for clustering.

IEEE Trans Syst Man Cybern C Cybern 39(2):133-155

[14] Hubert, L.; Arabie, P., Comparing partitions, J Classif, 2, 193-218, (1985) · Zbl 0587.62128 · doi:10.1007/BF01908075

[15] Jain, AK, Data clustering: 50 years beyond k-means, Pattern Recogn Lett, 31, 651-666, (2010)· doi:10.1016/j.patrec.2009.09.011

[16] Jain, AK; Murty, NM; Flynn, PJ, Data clustering: a review, ACM Comput Surv, 31, 264-323, (1999)· doi:10.1145/331499.331504

[17] Klein D, Kamvar SD, Manning CD (2002) From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), July 8-12, 2002. Australia, Morgan Kaufmann, Sydney, pp 307-314

[18] Klekota, J.; Roth, FP, Chemical substructures that enrich for biological activity, Bioinformatics, 24, 2518-2525, (2008)· doi:10.1093/bioinformatics/btn479

[19] Lee, S.; McLachlan, G., On mixtures of skew normal and skew t-distributions, Adv Data Anal Classif, 7, 241-266, (2013) · Zbl 1273.62115 · doi:10.1007/s11634-013-0132-8

[20] Li Z, Liu J, Tang X (2009) Constrained clustering via spectral regularization. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp 421-428. doi:10.1109/CVPR.2009.5206852

[21] Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml

[22] Lu Z, Leen TK (2004) Semi-supervised learning with penalized probabilistic clustering. In: NIPS

[23] McLachlan G, Krishnan T (2008) The EM algorithm and extensions, Wiley series in probability and statistics, 2nd edn. Wiley, Hoboken · Zbl 1165.62019

[24] McNicholas, PD; Murphy, TB, Model-based clustering of microarray expression data via latent Gaussian mixture models, Bioinformatics, 26, 2705-2712, (2010)· doi:10.1093/bioinformatics/btq498

[25] Melnykov V, Melnykov I, Michael S (2015) Semi-supervised model-based clustering with positive and negative constraints. Adv Data Anal Classif 1-23. doi:10.1007/s11634-015-0200-3

[26] Morlini, I., A latent variables approach for clustering mixed binary and continuous variables within a Gaussian mixture model, Adv Data Anal Classif, 6, 5-28, (2012) · Zbl 1284.62384 · doi:10.1007/s11634-011-0101-z

[27] Morris, K.; McNicholas, P.; Scrucca, L., Dimension reduction for model-based clustering via mixtures of multivariate t-distributions, Adv Data Anal Classif, 7, 321-338, (2013) · Zbl 1273.62141 · doi:10.1007/s11634-013-0137-3

[28] Narayanan H, Mitter S (2010) Sample complexity of testing the manifold hypothesis. In: Advances in Neural Information Processing Systems, pp 1786-1794

[29] Nguyen, HD; McLachlan, GJ, Maximum likelihood estimation of Gaussian mixture models without matrix operations, Adv Data Anal Classif, 9, 371-394, (2015)· doi:10.1007/s11634-015-0209-7

[30] Olivier B, Soudijn W, van Wijngaarden I (1999) The 5-ht$_{1A}$ receptor and its ligands: structure and function. In: Jucker E (ed) Progress in Drug Research, Progress in Drug Research, vol 52, pp 103-165

[31] Pavel B (2002) Survey of clustering data mining techniques. Technical report, Accrue Software

[32] Rubinstein RY, Kroese DP (2004) The Cross Entropy Method: A Unified Approach To Combinatorial Optimization, Monte-Carlo Simulation (Information Science and Statistics). Springer-Verlag New York Inc, Secaucus, NJ, USA

[33] Samuelsson J (2004) Waveform quantization of speech using Gaussian mixture models. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, vol 1, pp I-165-8, vol 1. doi:10.1109/ICASSP.2004.1325948

[34] Scrucca, L.; Raftery, AE, Improved initialisation of model-based clustering using Gaussian hierarchical partitions, Adv Data Anal Classif, 9, 447-460, (2015)· doi:10.1007/s11634-015-0220-z

[35] Shental, N.; Bar-Hillel, A.; Hertz, T.; Weinshall, D., Computing Gaussian mixture models with EM using equivalence constraints, Adv Neural Inf Process Syst, 16, 465-472, (2004) · Zbl 1161.68775

[36] Śmieja M, Tabor J (2013) Image segmentation with use of cross-entropy clustering. In: Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013. Springer, Advances in Intelligent Systems and Computing, pp 403-409

[37] Śmieja, M.; Tabor, J., Entropy approximation in lossy source coding problem, Entropy, 17, 3400-3418, (2015) · Zbl 1338.94053 · doi:10.3390/e17053400

[38] Śmieja M, Tabor J (2015b) Spherical Wards clustering and generalized Voronoi diagrams. In: Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on, IEEE, pp 1-10

[39] Śmieja M, Warszycki D (2016) Average information content maximization—a new approach for fingerprint hybridization and reduction. PLoS One 11(1):e0146,666

[40] Sommer C, Strähle C, Köthe U, Hamprecht FA (2011) ilastik: Interactive Learning and Segmentation Toolkit. In: Eighth IEEE International Symposium on Biomedical Imaging (ISBI). Proceedings, pp 230-233. doi:10.1109/ISBI.2011.5872394

[41] Spurek P, Tabor J, Zając E (2013) Detection of disk-like particles in electron microscopy images. In: Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013, Springer, pp 411-417

[42] Subedi, S.; McNicholas, P., Variational Bayes approximations for clustering via mixtures of normal inverse Gaussian distributions, Adv Data Anal Classif, 8, 167-193, (2014)· doi:10.1007/s11634-014-0165-7

[43] Tabor J, Misztal K (2013) Detection of elliptical shapes via cross-entropy clustering. In: Pattern Recognition and Image Analysis, Springer, Berlin 7887:656-663

[44] Tabor, J.; Spurek, P., Cross-entropy clustering, Pattern Recogn, 47, 3046-3059, (2014) · Zbl 1342.68279 · doi:10.1016/j.patcog.2014.03.006

[45] Telgarsky M, Vattani A (2010) Hartigan's method: k-means clustering without Voronoi. In: Teh YW, Titterington DM (eds)

AISTATS, JMLR.org, JMLR Proceedings, vol 9, pp 820-827

[46] Vyas R, Gao J, Cheng L, Du P (2014) An image-based model of the interstitial cells of cajal network in the gastrointestinal tract. In: Goh J (ed) The 15th International Conference on Biomedical Engineering, IFMBE Proceedings, vol 43, Springer International Publishing, pp 5-8

[47] Wagstaff K, Cardie C, Rogers S, Schrödl S (2001) Constrained k-means clustering with background knowledge. In: Machine Learning, Proceedings of the Eighteenth International Conference (ICML 2001), June 28-July 1, 2001. Williams College, Williamstown, MA, USA, Morgan Kaufmann, pp 577-584

[48] Wang X, Davidson I (2010) Flexible constrained spectral clustering. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, KDD '10, pp 563-572. doi:10.1145/1835804.1835877

[49] Warszycki D, Mordalski S, Kristiansen K, Kafel R, Sylte I, Chilmonczyk Z, Bojarski AJ (2013) A linear combination of pharmacophore hypotheses as a new tool in search of new active compounds—an application for 5-ht$_{1A}$ receptor ligands. PLoS One 8(12):e84,510. doi:10.1371/journal.pone.0084510

[50] Willett, P., Searching techniques for databases of two- and three-dimensional chemical structures, J Med Chem, 48, 4183-4199, (2005)· doi:10.1021/jm0582165

[51] Wolfe J (1963) Object cluster analysis of social areas. University of California

[52] Wu Q, Merchant FA, Castleman KR (2008) Microscope image processing. Elsevier/Academic Press, Amsterdam

[53] Xiong Z, Chen Y, Wang R, Huang T (2002) Improved information maximization based face and facial feature detection from real-time video and application in a multi-modal person identification system. In: Multimodal Interfaces, 2002. Proceedings. Fourth IEEE International Conference on, pp 511-516. doi:10.1109/ICMI.2002.1167048

[54] Xu R, Wunsch D (2009) Clustering. Wiley-IEEE Press, Hoboken

[55] Xu, R.; Wunsch, I., Survey of clustering algorithms, IEEE Trans Neural Netw, 16, 645-678, (2005)· doi:10.1109/TNN.2005.845141