Effect of Term Distribution Weighting for K-means Clustering

Uraiwan Buatoom, Waree Kongprawechnon, and Thanaruk Theeramunkong

***Abstract*— While unsupervised learning is non-guided, supervised learning is full power to guide, and semi-supervised learning is a small number of some sufficient labeled instances to guide a large number of unlabeled instances. The key idea for traditional classification task is just focused on a set of limited predefined classes that aims to increase the performance of supervised learning (i.e., classification) using unlabeled data. This work proposes another type of semi-supervised learning to handle, where semantic-labels are not directly applied, statistics extracted from predefined classes with objects are used to guide the clustering process. In contrast with most previous works on constrained clustering where pairwise constraints (MUST-LINK vs CANNOT-LINK) are specified, however our work proposes how weighting obtained from a training set effects the clustering results. Two groups of distribution statistics are used in an experiments to improve the k-means clustering that is deviation- and entropy-based schemes on term weightings. Both schemes utilized by class distribution; in-collection, intra-class, and inter-class distribution as constraints to guide clustering towards user intention. The experiment is performed on text datasets to evaluate the performance. Finally, the experimental result shows that term weighting scheme has a potential to control/guide clustering process.**

***Index Terms*—semi-unsupervised; term weighting; multi-dimension; and constrained clustering**

ACKNOWLEDGMENT

This research work is partially supported by the Sirindhorn International Institute of Technology (SIIT), Thammasat University, Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Intelligent Informatics and Service Innovation (IISI) Research Center, National Electronics and Computer Technology Center (NECTEC) of Thailand for the STEM Workforce, as well as the Thailand Research Fund (TRF) under grant number RTA-6080013. The corresponding author is also thankful to Burapha University (Chanthaburi Campus) who provided financial assistance every month.

REFERENCES

[1] K. Wagstaff, C. Cardie, S. Rogers, S. Schrdl, et al., Constrained k-means clustering with background knowledge, in Proceedings of the Eighteenth International Conference on Machine Learning, pp. 577584, 2001.

[2] V. Lertnattee and T. Theeramunkong, Effects of term distributions on binary classification, IEICE TRANSACTIONS on Information and Systems, vol. 90, no. 10, pp. 15921600, 2007.

[3] C. Largeron, C. Moulin, and M. Gry, Entropy based feature selection for text categorization, in Proceedings The Symposium on Applied Computing, pp. 924928, ACM, 2011.

[4] A. Guo and T. Yang, Research and improvement of feature words weight based on tfidf algorithm, in IEEE International Conference on Information Technology, Networking, Electronic and Automation Control, pp. 415419, 2016.

[5] U. Buatoom, W. Kongprawechnon, and T. Theeramunkong, Constrained clustering with seeds and term weighting scheme, in IEEE Knowledge, Information and Creativity Support Systems, pp. 116 121, 2018.

Authors are with the School of Information, Computer, and Communication Technology (ICT), Sirindhorn International Institute of Technology, Thammasat University, PathumThani, Thailand.

Author3 is with Associate Fellow, The Royal Society of Thailand, Bangkok, Thailand.

..

Uraiwan Buatoom received a B.Sc. in Computer Science and M.Sc. in Information Technology from Burapha University, in 2003 and 2008, respectively. Currently, she is a doctoral candidate in the Information Technology program at Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand. Her research interests include data mining, clustering, and knowledge discovery

Waree Kongprawechnon received a B.Eng. degree in Electrical Engineering from Chulalongkorn University, Thailand, in 1992; and an M.Eng. in Control Engineering from Osaka University, Japan, in 1995 and Ph.D. in Mathematical Engineering and Information Physics from the University of Tokyo, Japan, in 1998. She is an Associate Professor at SIIT, Thammasat University, Thailand. Her research interests include H^{∞} control, control theory, robust control, system identification, modeling, adaptive control, learning control, neural networks and fuzzy control.

Thanaruk Theeramunkong received a B.Eng. in Electric and Electronics Engineering, and an M.Eng. and Ph.D. in Computer Science from the Tokyo Institute of Technology in 1990, 1992 and 1995, respectively. Now, he is serving as a Professor at SIIT, Thammasart University, and Associate Fellow at the Royal Society of Thailand. His research interests include data mining, machine learning, natural language processing, and knowledge engineering.