Designing GUI for Human Active Learning in Constrained Clustering

Seiji Yamada
National Institute of Informatics / SOKENDAI / Tokyo Institute of Technology
2-1-2 Hitotsubashi, Chiyoda
Tokyo, 101-8430 Japan
seiji@nii.ac.jp

Masayuki Okabe
Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku
Toyohashi, 441-8580 Japan
okabe@imc.tut.ac.jp

Junki Mizukami
Tokyo Institute of Technology
4259 Nagatsuta, Midori
Yokohama, 226-8503 Japan
mizukami@ntt.dis.titech.ac.jp

IUI 2013 Workshop on Interactive Machine Learning,
19 March, 2013, Santa Monica, CA USA.

Abstract
In this paper, a novel GUI for human active learning in constrained clustering is described. Clustering is the most popular data mining technology, and in particular, interactive constrained clustering that uses constraints from a human is promising for practical applications. We propose a GUI that can expose the effects of given constraints by emphasizing them at the clustered results and provide multiple viewpoints that a user can flexibly change to derive human active learning. We fully implemented an interactive constrained clustering system with the GUI as a web service. We also conducted an evaluation experiment on image clustering with participants, and obtained results to support the effectiveness of our approach.

Introduction
Clustering is the most popular data mining technology, and in particular, interactive constrained clustering that uses constraints from a user is promising for practical application. In interactive constrained clustering, a user plays the role of a *teacher* to a constrained clustering system [2] and gives must/cannot-links as constraints. In such a situation, a user tries to select effective constraints and this human ability is called *human active learning* [4]. Furthermore, because a user's cognitive load necessary to give constraints is well known to be significantly restricted

and the maximum number of pairwise constraints was about 50 in our experiment, we needed to prepare a mechanism to support human active learning. Thus we propose a GUI that can derive human active learning. The GUI has two characteristics, exposing the effects of the last constraint given from a user and providing multi viewpoints with which a user can easily find an effective data-pair as a constraint.

Some constraints are not effective [6] in constrained clustering, and the cognitive load necessary for a user to give constraints is high in interactive clustering, thus we need to provide a mechanism in which a user can easily select only effective constraints for an interactive constrained clustering system. Since the constraints are considered to be training data for classification learning, traditional computational active learning like uncertainty sampling [11] and query by committee [13] might be useful in non-interactive constrained clustering. For interactive constrained clustering, we should use human active learning [4].

As far as we know, popular data mining tools that include Weka [9] do not provide an environment for interactive constrained clustering and a suitable GUI for human active learning. Interactive machine learning systems are closely related to this work. Such systems provide a user interface that supports a user in giving training data to the systems. Fails et al. originally proposed an interactive machine learning framework that supports the interactive training of pixel classifiers with a user for image segmentation [7]. CueFlik [8] also provided an end-user machine learning environment in which users can easily create rules for re-ranking images according to their visual characteristics. It was implemented for web image searches. CuteT [1] was designed to use interactive machine learning that learns from triaging decisions made by operators in a dynamic environment. It also had visualizations to support operators to quickly and accurately triage alarms. These interactive machine learning systems basically dealt with training data for classification learning, not constraints for constrained clustering, and explicitly did not provide an interaction design for deriving human active learning.

Designing a GUI for human active learning

We made an interactive constrained clustering system with a GUI that can derive human active learning, and investigated its effectiveness in performing clustering. The interactive constrained clustering system was implemented on a web server by using MATLAB, Perl, and JavaScript, and could run on a web browser. Figure 1 shows a snapshot of the interactive constrained clustering system's UI. The system had the two following special functions for human active learning.

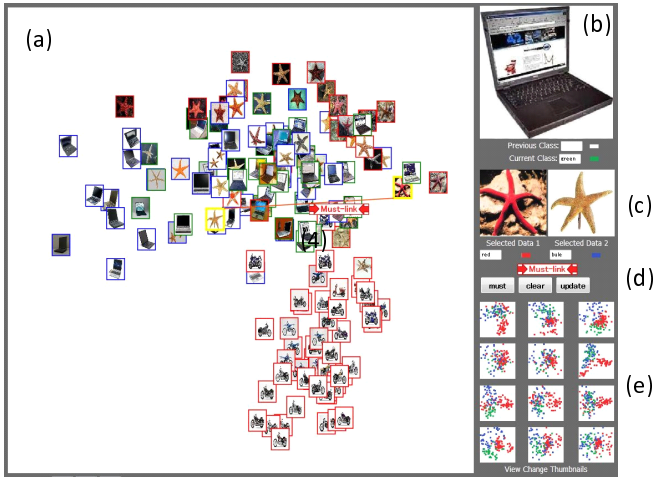

Figure 1: GUI for human active learning.

Figure 2: Exposing the effects of a given constraint.

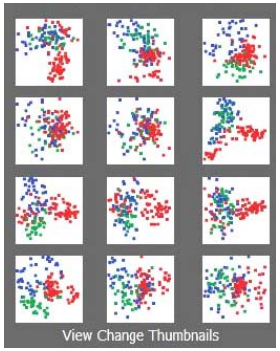

Figure 3: Multiple-views selection (Figure 1(e)).

- *Exposing the effects of given constraints*: The interactive constrained clustering system can expose data influenced by the last given constraint. We expect this function to make a user recognize the effects caused by his/her given constraint and to derive human active learning. The influenced data are identified by checking the change of a cluster to which data belonged after every clustering, and are emphasized by being circled like in Figure 2.
- *Multiple viewpoints for the results of clustering*: The interactive constrained clustering system enables a user to change viewpoints to see the distribution of data. We used multi-dimensional scaling [3] to show data in the 2D coordination, and provided multiple viewpoints by preparing pairs of eigenvalues as a 2D-coordination. In Figure 3, which shows 1(e), a user can freely select a viewpoint by mouse clicking, and the data distribution with the selected viewpoint is updated as in Figure 1(a).

COP-Kmeans [14] was used in our interactive constrained clustering system as a constrained clustering algorithm. Another approach to constrained clustering is metric/distance learning [10]. However, since it costs much because it needs complicated optimization and the response is sufficiently quick, we did not use it. The user gives only must-links. The user selects a pair of data for a must-link by using a main window (a), a magnifying window (b) and a selected data pair window (c), as shown in Figure 1. After the user determines a must-link in (d), the constrained clustering run and the data in (a) are updated. This procedure is repeated.

Experiment

We conducted a within-subject experiment with 16 participants (12 males, 4 females, ages 18∼52). Two conditions, a proposed GUI (p-GUI) and a traditional GUI without the two functions (t-GUI), were prepared, and their ordering was counterbalanced. In each trial, a participant was asked to give 50 constraints one by one. The data were image files generated from CALTECH 256 by using bag of features [5] with SIFT [12]. A data set of three clusters comprised of "motorcycles", "laptop PCs" and "starfishes" was prepared, and each cluster had about 50 pieces of image data.

We measured the number of constraints (must-links) necessary to achieve normalized mutual information (NMI) = 1.0 and the maximum NMI for 50 constraints as an evaluation. The average number of constraints (the upper limit = 50) for the p-GUI and t-GUI was 35.9 (SD = 10.6) and 46.1 (SD = 10.3), and the average maximum NMI was .97 (SD = .08) and .84 (SD = .13). These results are shown in Figures 4 and 5. We applied a t-test to them and found significant differences between the two levels in both evaluations ($p = .007, .006$). These results supported the effectiveness of our proposed GUI. We also found that this human active learning

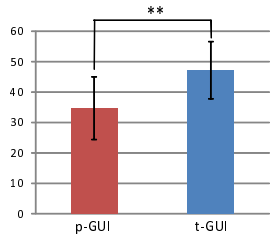

Figure 4: Number of constraints to NMI = 1.0.

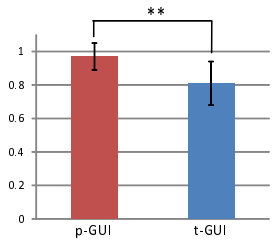

Figure 5: Maximum NMI for 50 constraints.

outperformed computational active learning with uncertain sampling-like heuristics.

Conclusion

We proposed a novel GUI for human active learning in interactive constrained clustering. The GUI can expose the effects of the last given constraint to a user so that he/she can easily recognize the causality between constraints and clustering, and it can provide multiple viewpoints so that a user can quickly find effective data pairs as constraints. We conducted experiments with image clustering by using popular data sets to evaluate our proposed GUI. We experimentally compared the proposed GUI with a conventional system without the GUI, and obtained results to support the advantages of our proposed GUI.

References

[1] S. Amershi, B. Lee, A. Kapoor, R. Mahajan, and B. Christian. Human-guided machine learning for fast and accurate network alarm triage. In *IJCAI'11*, pages 2564–2569, 2011.

[2] S. Basu, I. Davidson, and K. Wagstaff, editors. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman and Hall/CRC, 2008.

[3] I. Borg and P. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Verlag, 1997.

[4] R. M. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu. Human active learning. In *NIPS'08*, pages 241–248, 2008.

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision*, volume 1, page 22, 2004.

[6] I. Davidson. Two approaches to understanding when constraints help clustering. In *KDD'12*, pages 1312–1320, 2012.

[7] J. A. Fails and D. R. Olsen, Jr. Interactive machine learning. In *IUI'03*, pages 39–45, 2003.

[8] J. Fogarty, D. Tan, A. Kapoor, and S. Winder. Cueflik: interactive concept learning in image search. In *CHI'08*, pages 29–38, 2008.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[10] P. Jain, B. Kulis, I. Dhillon, and K. Grauman. Online metric learning and fast similarity search. In *NIPS'08*, volume 22, pages 761–768, 2008.

[11] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12, 1994.

[12] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[13] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT'92*, pages 287–294, 1992.

[14] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *ICML'01*, pages 577–584, 2001.