An Introductory Survey on Reframing in Clustering

REFRAME

http://www.reframe-d2k.org/

Rethinking the Essence, Flexibility and Reusability of Advanced Model Exploitation

Md Geaur Rahman
Postdoctoral research fellow
University of Strasbourg
Email: grahman.au@gmail.com
11 September 2015, LMCE, Porto,

Presented by:
Dr. Nicolas Lachiche
Associate Professor
Head of the BFO Team
University of Strasbourg, France.

Outline

- Introduction and significance of reframing in clustering
- State-of-the-art reframing in clustering approaches
- Potential future issues
- Conclusion

Introduction to Reframing in Clustering

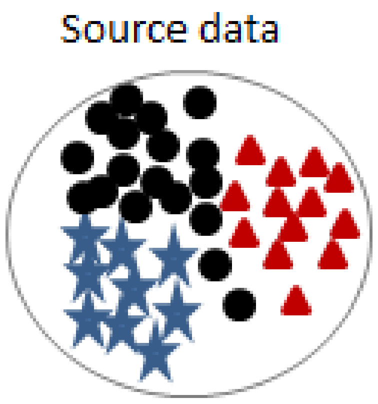
- Dataset shift between source and target data exists in many real-life applications
- Performances of existing approaches suffer from a large amount of errors due to dataset shift

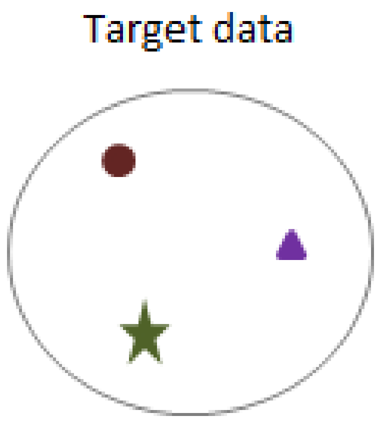
Significance of Reframing in Clustering

- Retraining of a model on the target data may improve the performance of the model

- However, most cases retraining may not be feasible due to insufficient target data and time
- Reframing between the source and target domains can be useful

Reframing in Clustering





Existing Reframing in Clustering Methods

- Self taught clustering
- Incremental clustering
- Online clustering
- Mean shift clustering

Self Taught Clustering

Objective function:

$$R(\tilde{X}_T, \tilde{X}_S, \tilde{Z}) = I(X_T, Z) - I(\tilde{X}_T, \tilde{Z}) + \lambda \left[I(X_S, Z) - I(\tilde{X}_S, \tilde{Z}) \right]$$

Where,

X_S=source data, X_T=target data

Z=common feature space between X_S and X_T

I(.,.) is the mutual information between two random variables

$$I(\underset{\sim}{X}; \underset{\sim}{Z}) = \sum_{x \in X} \sum_{z \in Z} p(x, z) \log \frac{p(x,z)}{p(x)p(z)}.$$

λ is a user-defined parameter, balances the influences between source and target data

Incremental Clustering: COBWEB

- Assigns the first record into a cluster
- Considers the next record and assigns it either to one of the existing clusters or to a new cluster.
- Repeats the second step till all the records are clustered

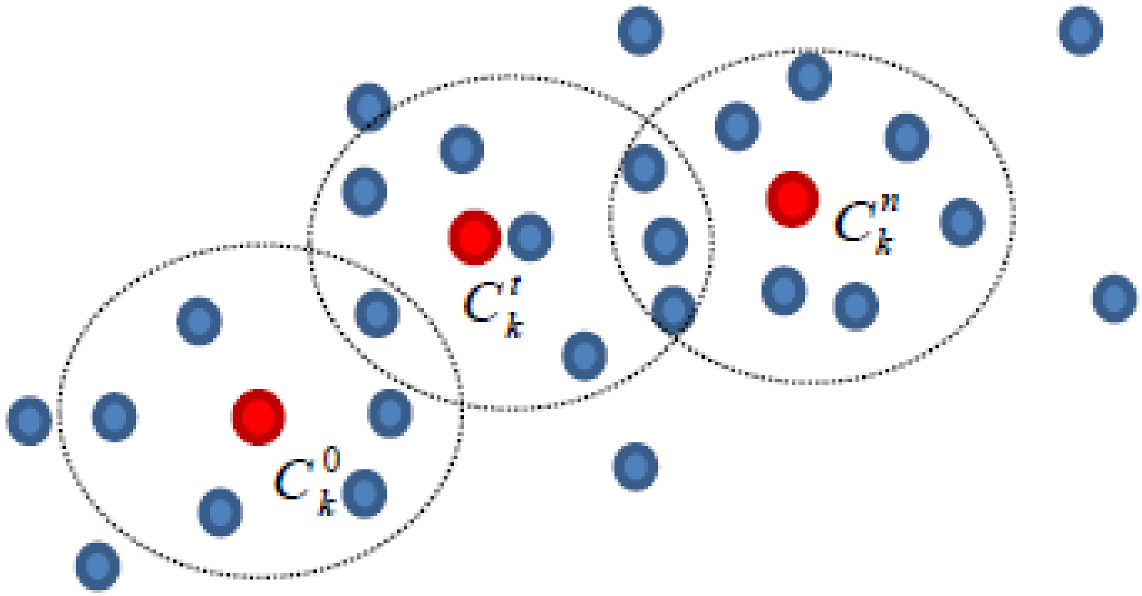
Online clustering

- Finds prototypes based on source data
- Uses the prototypes to groups the target data.
- Allows the prototypes to learn online.
- Iteratively updates the prototypes as follows.

$$V_k^{new} = V_k + \zeta(x_i - V_k)$$

REFRAME

Mean shift clustering



REFRAME

Potential Future Issues

- Use of real-life datasets to evaluate the techniques could be a better motivation for improvements
- Automatic tuning of parameters
- Better feature representation between the source and target data

REFRAME

Conclusion

- We present an introductory survey on the state-of-the-art clustering algorithms that can reframe prototypes from a source environment to a target environment
- Existing techniques have limitations
- In the future, we aim to develop a new algorithm by addressing issues and evaluate the technique on a real world application.

Some References

[1] Ahmed, C.F., Lachiche, N., Charnay, C., Braud, A.: Reframing continuous input attributes. In: Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence. pp. 31–38. IEEE (2014)

[2] Barbakh, W., Fyfe, C.: Online clustering algorithms. International Journal of Neural Systems18(03), 185–194 (2008)

[3] Cheng, Y.: Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(8), 790–799 (1995)

[4] Dai, W., Yang, Q., Xue, G.R., Yu, Y.: Self-taught clustering. In: Proceedings of the 25th international conference on Machine learning. pp. 200–207. ACM (2008)

[5] Pan, S.J., Yang, Q.: A survey on transfer learning. Knowledge and Data Engineering, IEEE Transactions on 22(10), 1345–1359 (2010)

[6] Jose, H.O., Ricardo, B.P., Kull, M., Flach, P., Chowdhury, F.A., Lachiche, N., Martiynez-Uso, A.: Reframing in context: A methodology for model reuse in machine learning. AI Communications (submitted) (2015)

Questions and Suggestions

Please send your valuable questions and suggestions to grahman.au@gmail.com

REFRAME

Thank you