2016 IEEE 28th International Conference on Tools with Artificial Intelligence

Reframing in Clustering

Md Naimul Hoque Chowdhury Farhan Ahmed
Dept. of Computer Science and Engineering
University of Dhaka
Dhaka, Bangladesh
tonmoycsedu@gmail.com, farhan@du.ac.bd

Nicolas Lachiche
ICube Laboratory
University of Strasbourg
Strasbourg, France
nicolas.lachiche@unistra.fr

Carson K. Leung(✉) Hao Zhang
Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada
{kleung, zhangh15}@cs.umanitoba.ca

***Abstract*—Adaptation of the dataset shift has grown to be of great importance in machine learning problems in recent years. Reframing has emerged as a new machine learning technique that adapts the context changes between training and target domains. One of the advantages of reframing is that it can offer good performances with a limited amount of deployment data. Reframing has already been implemented in classification and regression by reusing labelled training data with the help of few labelled target data. However, *reframing in clustering* is still a challenging research problem because of its unsupervised nature. In this paper, we concentrate on building a reframing method for clustering. We also show the necessity and effectiveness of our method in contrast to retraining, which is the process of learning new model in the testing and deployment phases. Our evaluation results with extensive experiments using both synthetic and real-life datasets show that our method correctly identifies most of the shifts between datasets and builds better clustering model than retraining.**

***Keywords*—Machine learning; clustering; reframing; retraining; dataset shift**

I. Introduction and Related Works

Utilization of the existing models in different context has become a challenging task in machine learning [7]. In many real-life applications, the distribution and clustering tendency of data is often different from training to deployment. Thus, applying the learned model (i.e., *base model*) in deployment (or target domain) [2], [5], [6], [11] does not produce the expected results in many cases. *Retraining* [1] is a solution in cases where one can learn a new model in deployment. However, this costly solution may not produce good results when one do not have enough data in deployment.

In IEEE ICTAI 2014, *reframing* [1]—which uses a limited amount of data in deployment and learns the dataset shift [8] from training to deployment—was shown to produce good results in classification and regression. Reframing is highly influenced by the *unsupervised transfer learning* [9] technique, which requires some unlabeled data in the target domain to induce an objective predictive model $f_T(\cdot)$ for use in the target domain. However, to our knowledge, there has not yet been much comprehensive work detailing how to use reframing in clustering, especially on using hill-climbing. In this paper, we propose a hill-climbing method for reframing in clustering and also for improving cluster qualities in the target domain.

Consider the following real-life scenario. There are three income classes of people in Cities 1 and 2—namely, low income, middle class and high income. When applying K-means clustering on the income to the people of City 1, we get three clusters, with their cluster means being 5000, 12000 and 20000. When applying the same clustering technique to City 2, we also get three clusters. Suppose that the means of these three clusters for City 2 are 7500, 18000 and 30000. If the mean of City 1 is X and mean of City 2 is Y, then the dataset shift between the income of people in these two cities can be expressed as follows:

$$Y = \alpha X + \beta, \tag{1}$$

where α=1.5 and β=0. Now, suppose that we do not have enough manpower to collect data of City 2. Instead, we collect small number of data using our limited manpower. With this limited amount of data, K-means may not converge and thus may cluster future data incorrectly. The motivation behind our work is to (i) learn the shift between two cities using the model already built for City 1 and the small amount of data available of City 2 and (ii) ultimately build a model that correctly clusters the data for City 2. In this scenario, we can apply the model learned from City 1 to City 2 by simply multiplying the centers of the clusters of City 1 by 1.5. This scenario shows the significance of reframing in clustering. Reframing is very effective when data distribution changes from the source to the target domain and there is not enough target data to retrain the model. This type of real-life scenario gives us the motivation to build a reframing method for clustering continuous input attributes.

Our work is challenging in the sense that we are working with unsupervised learning technique. Moreover, there has not been a lot work on how to use reframing in clustering. Contributions of our work are listed below:

- We develop an efficient algorithm for reframing continuous input attributes in clustering.
- We determine the existence of dataset shifts between datasets.

2375-0197/16 $31.00 © 2016 IEEE
DOI 10.1109/ICTAI.2016.57

- We use the existing model (i.e., base model) with some target data to learn the optimum value of dataset shift.
- We compare the efficiency of the proposed method with base model and retraining.
- We show the existence of dataset shift in real-life datasets and experimentally show the results of reframing in real-life datasets.

The remainder of this paper is organized as follows. We describe our proposed reframing method for clustering in Section II. In Section III, our experimental results are presented and analyzed. Finally, conclusions are drawn in Section IV.

II. Our Proposed Method

Related to our method are the *base model* and *retraining*. Their formal definitions in the context of clustering are given below.

***Definition** 1:* Given a source domain D_S and its clustering task C_S, as well as a target domain D_T and its clustering task C_T, the use of the **base model** is the process of improving the cluster quality function $RC(\cdot)$ in D_T using only the knowledge in D_S, where $C_S \neq C_T$ and the labels of data of both the source domain and the target domain are not observable. ■

***Definition** 2:* Given a source domain D_S and clustering task C_S, as well as a target domain D_T and clustering task C_T, **retraining** is the process of improving the cluster quality function $RC(\cdot)$ in D_T using only the knowledge in D_T, where $C_S \neq C_T$ and the labels of data of both the source domain and the target domain are not observable. ■

Reframing, on the other hand, uses knowledge from both D_S and D_T. Our proposed method uses (i) K-means as clustering model and (ii) sum of squared errors [4] as clustering measurement. We describe our method—called *Reframing in Clustering using Hill-climbing* (*RCH*)—via the following illustrative example.

Consider data of a university student's mobile talking time for two days as shown in Table I. If we use the first three data points of each day as our initial values of K-means, we get three clusters as shown in Tables II and III. An observant reader may notice that each datum Y in Day 2 is similar to the corresponding datum X in Day 1 in the sense that they are shifted by α=2 and β=10, i.e., $Y = 2X + 10$. Hence, the final means (i.e., centers) of Day 2 are also shifted. The sum of squared errors for the target (Day 2) model is $(133-110)^2 + ... + (133-122)^2 + (57.43-50)^2 + ... + (57.43-56)^2 + (228.5-250)^2 + ... + (228.5-266)^2 = 6888.57$.

An option to cope with this dataset shift is to ignore the shift and apply the cluster centers learned from base model directly to the test data. When all the data of Table I(b) are used as test data, three clusters are formed as shown in Table IV(a), in which every data point except 34 is incorrectly clustered.

Table I
Mobile talking time.

PersonID	Income
1	20
2	30
3	50
4	120
5	22
6	63
7	102
8	34
9	87
10	12
11	25
12	128
13	23
14	77
15	56

(a) Day 1 (source)

PersonID	Income
1	50
2	70
3	110
4	250
5	54
6	136
7	214
8	78
9	184
10	34
11	60
12	266
13	56
14	164
15	122

(b) Day 2 (target)

Table II
Base/source model: Day 1

	Cluster1	Cluster2	Cluster3
	50	20	120
	63	12	102
	77	22	87
	56	34	128
		30	
		25	
		23	
Centers	61.5	23.71	109.25

Table III
Target model: Day 2

	Cluster1	Cluster2	Cluster3
	110	50	250
	136	34	214
	164	54	184
	122	78	266
		70	
		60	
		56	
Centers	133	57.43	228.5

An alternative option is to retrain the model for Day 2 using the limited data (e.g., first five data points of Day 2). After three iterations, the model retained from deployment data is built as shown in Table IV(b), in which the cluster centers are at 90, 52 and 250. Note that, when using all data in Table I(b) as test data and applying the centers learned from this retraining process as our initial centers, three clusters are formed as in Table IV(c), in which only 78 is incorrectly clustered into Cluster1. The final sum of squared errors for this retraining model is 8811, which is greater than 6888.57 for the original target model.

Next, we apply reframing to build a model by using only the first five data point in Table I(b). Let avg_d be the average of the available data (e.g., $avg_d = \frac{50+54+70+110+250}{5} = 106.8$), and let avg_m be the average of the means of the base model (e.g., $avg_m = \frac{61.5+23.71+109.25}{3} = 64.82$). Then, the ratio of these two averages is 1.65, which is our initial α. Multiplying the means of Day 1 by α gives us the initial means of Day 2, which are 101.47, 39.13 and 180.26. If we apply these means to all the five deployment data points of

Table IV
CLUSTERING RESULTS USING DIFFERENT MODELS.

	Cluster1	Cluster2	Cluster3
	50	34	110
	70		250
	54		136
	78		214
	60		184
	56		266
			164
			122
Centers	61.5	23.71	109.25

(a) BASE MODEL

Centers	Cluster1	Cluster2	Cluster3
	70	50	250
	110	54	
3rd iter.	90	52	250

(b) RETRAINING FROM DEPLOYMENT DATA

	Cluster1	Cluster2	Cluster3
	110	50	250
	136	70	214
	78	54	184
	122	56	266
	164	34	
		60	
Centers	122	54	228.5

(c) RETRAINING THE MEANS OF ORIGINAL DATA

Table V
REFRAMING WITH DIFFERENT α.

Centers	Cluster1	Cluster2	Cluster3
	110	50	250
		54	
		70	
1st iter.	101.47	39.13	180.26
2nd iter.	110	58	250

(a) α=1.65

Centers	Cluster1	Cluster2	Cluster3
	110	50	250
		54	
		70	
1st iter.	107.63	41.5	191.19
2nd iter.	110	58	250

(b) α=1.75

Centers	Cluster1	Cluster2	Cluster3
	110	50	250
		54	
		70	
1st iter.	126.08	48.61	223.96
2nd iter.	110	58	250

(c) α=2.05

Centers	Cluster1	Cluster2	Cluster3
	110	50	250
	70	54	
1st iter.	95.33	36.76	169.34
2nd iter.	90	52	250

(d) α=1.55

Table VI
REFRAMING WITH α=2.05 AND β=5 FOR DAY 2.

Centers	Cluster1	Cluster2	Cluster3
	110	50	250
		54	
		70	
1st iter.	131.08	53.61	228.96
2nd iter.	110	58	250

(a) AFTER FIRST TWO ITERATIONS

	Cluster1	Cluster2	Cluster3
	110	50	250
	136	70	214
	164	54	184
	122	78	266
		34	
		60	
		56	
Ctrs	131.08	53.61	228.96

(b) DEPLOYMENT OF REFRAMED MODEL FOR DAY 2

Day 2, we obtain a model as shown in Table V(a). Sum of squared errors for this model is $(110-110)^2+(58-50)^2+(58-54)^2+(58-70)^2+(250-250)^2=224$.

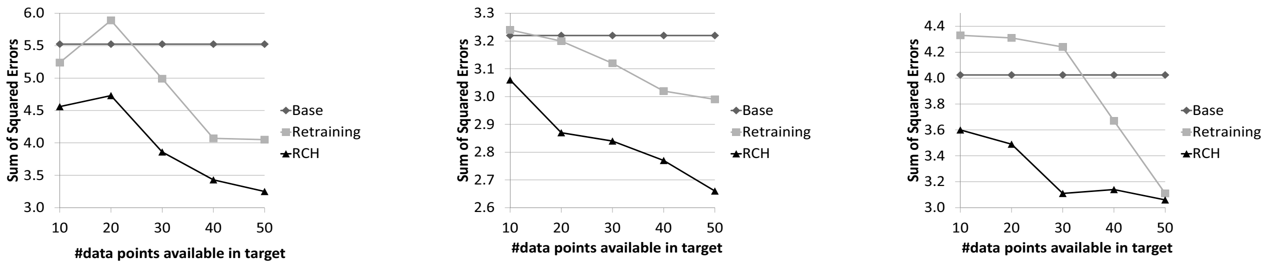
To improve the cluster quality and decrease sum of squared errors, we change α. We incrementally increase the value of α by 0.1 until the sum of squared errors is improved. With the next α=1.75, the initial cluster means are changed to 107.63, 41.5 and 191.19. The resulting clusters after the second iteration with α=1.75 (as shown in Table V(b)) are the same as those with α=1.65. We obtain similar clustering results when we increase α further until α=2.05, which is observed to be the (local) optimal α for our illustrative example.

We restart our reframing in clustering using hill-climbing process from initial α, but decrease the value of α subsequently this time. As the results with α=1.55 (as shown in Table V(d)) is not an improvement over our previous result and the sum of squared errors is greater than the previous minimum, we stop the process. Hence, our (local) optimal α is 2.05 for this reframed model.

Next, we learn β in two phases. First, we gradually decrease β from our initial β (which is 0) and run the process until there is improvement in the sum of squared errors. Then, we gradually increase β and try to learn the optimal β. Continue with our illustrative example. As increasing or decreasing the β value does not improve performance, we stop the process after running five times in both directions. As both +5 and -5 produce same results, we choose +5 as our optimal β. Result with α=2.05 and β=+5 is shown in the Table VI(a). Here, we shift the centers of our base model (Day 1) in Table II with α=2.05 and β= +5. Centers of the resulting reframed model are 131.08, 53.61 and 228.96, which almost correctly identify the shift between two datasets with only five deployment data. These centers are very close to the original centers of Day 2, which are 133.00, 57.43 and 228.5. In Table VI(b), we present our results on the whole dataset of Day 2 as test dataset. When compared with retraining (which incorrectly clusters 78 into Cluster1 as evidenced in Table IV(a)), reframing performs better than retraining because the reframing correctly clusters each data point. The resulting means of the reframing model are 133.5, 57.43 and 228.5, which are exactly the same as the original model shown in Table III. As the center and contents of each cluster are the same, sum of squared errors for both reframed and original model are the same.

III. EXPERIMENTAL RESULTS

In this section, we present the experimental results of our method when using both synthetic and real-life datasets. First, we compared the effectiveness and efficiency of our method with the base model and retrained model. Then, we show the existence of dataset shift in two real-life datasets, and analyze the performance of reframing (RCH) in comparison to the base model and retraining. We used K-means for forming the clusters and a sum of squared errors as a performance measurement.



(a) Season 2 as target

(b) Season 3 as target

(c) Season 4 as target

Figure 1. Learning curves for reframing (RCH), retraining and the base model with Season 1 as source and different seasons as targets on the real-life CBS dataset.

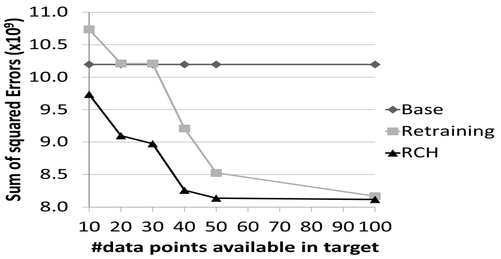


Figure 2. Learning curves for reframing (RCH), retraining and the base model with City 1 as source and City 2 as target on a synthetic dataset.

A. Performance on Synthetic Datasets

For experimental purpose, we generated two synthetic datasets using Gaussian distribution to represent income data of the people of two different cities. We used K-means for clustering the data. We generated 500 data points for City 1 and 500 data points for City 2. Out of the 500 data points of City 2, we took a small amount of data as target/deployment data and others as test data. We generated the data of City 1 using a mean of 12,000 and a standard deviation of 8,000. For City 2, the mean is 17,000 and the standard deviation is 10,000. The learning curves for reframing (RCH), retraining and the base model are shown in Figure 2.

First, we built a clustering model with the data of City 1. This model is the base/source model for our setting. We conducted our experiment with different numbers of available target data points. As shown in Figure 2, the curve for the base model is constant because the base model applies the centers learned from City 1 as the centers of K-means to the test data of City 2. As it application does not depend on the available target data in City 2, the results are static with respect to the number of available target data. Retraining learned the cluster centers of City 2 by applying K-means on the available target data and then applied these centers to the test data of City 2. Difference between them is that, the base model applied centers learned from City 1 directly to the City 2, but retraining omitted the knowledge of City 1 by simply applying the centers learned from available target data of City 2.

Recall from Section III, RCH uses both the centers from City 1 and the available target data of City 2 to learn the possible centers for City 2. As shown in Figure 2, reframing performs better than both retraining and the base model. Retraining produced a greater sum of squared errors than reframing or the base model when the number of target data is 10. The performance of retraining gets better with an increased number of target data. When the number of target data is 100, retraining catches reframing with its performance.

B. Performance on Real-life Datasets

In addition, we also used two real-life datasets from UCI Machine Learning Repository[1] to show the existence of dataset shift in real life and the effectiveness of our method. First, we show the experimental results on the bike sharing dataset [3]. The bike sharing dataset contains usage log of a bike sharing system called Capital Bike Sharing (CBS) in Washington, DC, USA. We built clustering models using four continuous attributes: (i) actual temperature (temp), (ii) apparent or feeling temperature (atemp), (iii) humidity (hum), and (iv) wind speed. Apparent temperature is “feels like” temperature perceived by humans. This dataset contains data of four seasons (i.e., summer, fall, winter and spring). From an earlier work [1], we know that there exist shifts between the data of different seasons. We took Season 1 (say, summer) as source domain and Seasons 2–4 as target domain. We used the same settings as described for the synthetic dataset.

Figure 1 shows the results of reframing in comparison to retraining and the base model. From the figure, reframing is observed to perform better than retraining and the base model. As expected, when the number of data in target increases performance of reframing and retraining becomes closer. However, reframing outperforms retraining significantly when the number of data is very small.

[1] http://archive.ics.uci.edu/ml/

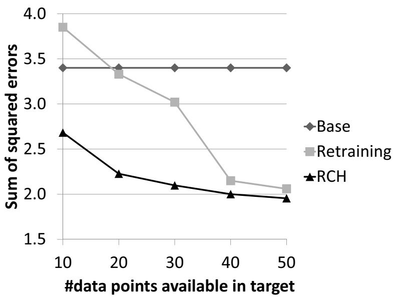


Figure 3. Learning curves for reframing (RCH), retraining and the base model with Channel 1 as source and Channel 2 as target on the real-life wholesale customer dataset.

Moreover, we also performed experiments on another real-life dataset—called wholesale customer dataset—from UCI Machine Learning Repository. This dataset, which refers to clients of a wholesale distributor, includes the annual spending in monetary units (m.u.) on diverse product categories. For our experiments, we selected six continuous input attributes: fresh products, milk products, grocery products, frozen products, detergents and paper products, as well as delicatessen products. Each attribute contains annual spending (m.u.) on the specific product. We chose the channel attribute as our shift attribute, which has two discrete values: the customer channel called Horeca (Hotel/Restaurant/Cafe) and the retail channel. In the experiment, we used data from the Horeca channel (Channel 1) as our source data and data from the Retail channel (Channel 2) as the target domain.

The result for wholesale customer dataset is presented in Figure 3, which shows that our proposed method of reframing (RCH) performed better than retraining and the base model. As expected, difference of the performances between reframing and retraining decreases as the number of target data increases.

IV. Conclusions

In this paper, we proposed a new method of reframing the values of K-means cluster centers so that they can be used in target domains. We designed an efficient algorithm—i.e., Reframing in Clustering using Hill-climbing (RCH)—to learn the optimal parameter values for the shifted input attributes. Our method is capable of tackling different changes in data distributions and decision functions, and making the existing model workable in different deployment environments. The method learns the value of the shift parameters by a popular learning technique named hill-climbing. It does so with the help of the source model and target/deployment data. Although the hill-climbing technique could inherently lead to the local maximum problem, our experimental results show than RCH leads to better results than alternatives. By extensive experiments, we showed—with the help of few unlabelled data in target—one can discover the desired optimal parameter values to handle the actual dataset shift. Our experimental results show that our method is remarkably better than retraining and base model in terms of sum of squared errors. We also showed the effectiveness of our method by giving real-life examples. The two real-life datasets used inherently dataset shifts in them.

For future work, we would like to build a unified reframing technique that will be applicable in different types of clustering method such as hierarchical clustering and grid based clustering. Moreover, we would like to build a reframing model that can handle nonlinear shift between datasets although this kind of shifts is rare in real life.

Acknowledgment

This project is partially supported by (i) the REFRAME project granted by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-NET (CHIST-ERA), as well as (ii) the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] C. F. Ahmed, N. Lachiche, C. Charnay, and A. Braud, "Reframing continuous input attributes," in *IEEE ICTAI 2014*, pp. 31–38.

[2] C. Charnay, N. Lachiche, and A. Braud, "Pairwise optimization of Bayesian classifiers for multi-class cost-sensitive learning," in *IEEE ICTAI 2013*, pp. 499–505.

[3] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, 2(2), 113–127, 2014.

[4] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*, 2011.

[5] J. Hernández-Orallo, "ROC curves for regression," *Pattern Recognition*, 46(12), pp. 3395–3411, 2013.

[6] N. Lachiche and P. Flach, "Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves," in *ICML 2003*, pp. 416–423.

[7] C. K. Leung, R. K. MacKinnon, and Y. Wang, "A machine learning approach for stock price prediction," in *IDEAS 2014*, pp. 274–277.

[8] J. G. Moreno-Torres, T. Raeder, R. Alaiz-RodríGuez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, 45(1), pp. 521–530, 2012.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE TKDE*, 22(10), pp. 1345–1359, 2010.

[10] M. G. Rahman, "An introductory survey on reframing in clustering," in *ECML-PKDD 2015 Workshop on LMCE.*

[11] H. Zhao, A. P. Sinha, and G. Bansal, "An extended tuning method for cost-sensitive regression and forecasting," *Decision Support Systems*, 51(3), pp. 372–383, 2011.