Adaptive Matching of Kernel Means

Miao Cheng[a]; Xinge You[b]
[a] Guangxi Normal University, Guilin, Guangxi, China
[b] Huazhong University of Science and Technology, Wuhan, Hubei, China

Abstract

As a promising step, the performance of data analysis and feature learning are able to be improved if certain pattern matching mechanism is available. One of the feasible solutions can refer to the importance estimation of instances, and consequently, **kernel mean matching (KMM)** has become an important method for knowledge discovery and novelty detection in **kernel machines**. Furthermore, the existing KMM methods have focused on concrete learning frameworks. In this work, a novel approach to **adaptive matching** of kernel means is proposed, and selected data with high importance are adopted to achieve calculation efficiency with optimization. In addition, **scalable learning** can be conducted in proposed method as a generalized solution to matching of appended data. The experimental results on a wide variety of real-world data sets demonstrate the proposed method is able to give outstanding performance compared with several state-of-the-art methods, while **calculation efficiency** can be preserved.

Introduction

Background

① As a promising step, the performance of pattern analysis and recognition are able to be improved if certain pattern **matching** mechanism is available.

② One of the feasible solutions can refer to the **importance estimation** of instances, and thereafter important instances hold more reference power for pattern analysis.

Kernel Mean Matching

● Derived from conception of **training (matching)** and **testing (reference)** data in pattern recognition, the importance of a given sample $W(x)$ [1] is given by the ratio of densities $p_m(x)$ and $p_r(x)$ as

$$w(x) = \frac{p_r(x)}{p_m(x)}$$

● KMM aims to minimize the discrepancy between reference distribution $p_r(x)$ and the matching distribution $p_m(x)$ in a RKHS, i.e.,

$$J_{KMM} = argmin_{\alpha} \left\| \frac{1}{n_m} \sum_{i=1}^{n_m} \alpha(x_i) \emptyset(x_i) - \frac{1}{n_r} \sum_{i=1}^{n_r} \emptyset(x_i) \right\|^2$$

By removing the constant item, the objective can be redefined as

$$J(\alpha) = argmin_{\alpha} \left[\frac{1}{2} \alpha^T K_{m,m} \alpha - \frac{n_m}{n_r} \alpha K_{m,r} e \right]$$

Fig. 1. The target groups of people are more important for certain sale businesses.

Fig. 2. Media information of matched knowledge are more attractive for corresponding persons in human society

Kernel Mean Matching with Global Importance

● To improve matching performance, a natural consideration in KMM is to select the reference instances with great importance so that calculation cost can be reduced,

$$\widetilde{w_i} = \int_r \emptyset(x_i^r) dx = \sum_{j=1}^{n_r} k(x_i^r, x_j^r)$$

Global KMM (gloKMM) algorithm

Input: Given matching instances $x_i^m (i = 1,2,\cdots,n_m)$, reference set $x_i^r (i = 1,2,\cdots,n_r)$, desired number of reference instances n_h with highest Importance.

1. Calculate the importance of each reference instance, and select the n_h instances with highest importance.

2. Calculate the kernels $K_{m,m}$ and $K_{m,h}$ with selected matching and reference instances.

3. Solve the KMM problem and obtain the optimal coefficients α.

4. Calculate estimated importance of instances by $w(x)$.

Adaptive Matching of Kernel Means

● Select a **subset** of reference data for estimation of importance, and it is verified the estimated importance results in **acceptable ranking** of reference data.

● A **refinement** stage is designed to pick up the reference instances with the highest importance associated with randomly selected instances.

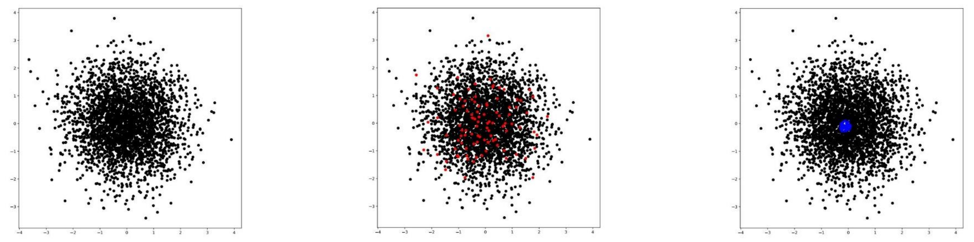

Fig. 3. A toy example of proposed method. (a) 3,000 data points of standard normal distribution. (b) Randomly selected 100 (red) points. (c) Top 50 (blue) points corresponding to random points.

● Selectively adaptive matching is repeated several times, and then a **fusion** stage is to adopted to learn the ideal matching. Finally, it aims to solve the quadratic programming (QP) problem with relaxed constraint conditions,

$$J(\beta) = argmin_{\beta_i} \sum_{i=1}^{t} \sum_{j=1}^{n_s} \left(\frac{1}{2} r_{i,j}^T K_{m,m} r_{i,j} - \frac{n_m}{n_r} r_{i,j} K_{m,r} e \right)$$
$$with \quad r_{i,j} = \alpha_{i,j} \beta_i$$
$$s.t. \quad \beta_i \geq 0$$

AKMM algorithm

Input: Given matching instances $x_i^m (i = 1,2,\cdots,n_m)$, reference set $x_i^r (i = 1,2,\cdots,n_r)$, number of repetition t, number of randomly selected instance n, desired number of important instances n_s for matching.

While *The desired repetition t has never reached* **do**

1. Randomly select n instances from x_i^r.

2. Choose the most important n_s instances from reference data associated with the previously selected n instances.

3. Follow the steps 2-3 in gloKMM algorithm.

End

4. Calculate the fusion coefficient by solving the defined QP problem.

5. Calculate estimated importance of samples $w(x)$.

Discussion

● **Differentiate AMKM from ensemble KMM:**

① Ensemble KMM relies on partition of reference set and the complete set is still absorbed, AMKM performs the selection with a separate refinement stage.

② AMKM randomly selects the subset of reference data with no explicit rule, and the volume of referred data can be changed conveniently.

● **Theoretical relationship with information theory [2],**

$$V(x^r) = \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} G(x_i^r - x_j^r, 2\sigma^2)$$

Experiments

● **KMM methods:**

standard KMM [3],
locally KMM (locKMM) [4],
ensemble KMM (ensKMM) [5],
global KMM (gloKMM)
AMKM

● **Date sets:**

Data Sets	Samples	Dimensionality
Monks	1,711	6
Ionosphere	351	34
Climate	540	18
Forest	581,012	54
Letter	20,000	16
CIFAR	10,000	255

① **KMM with different sizes of matching and reference data**

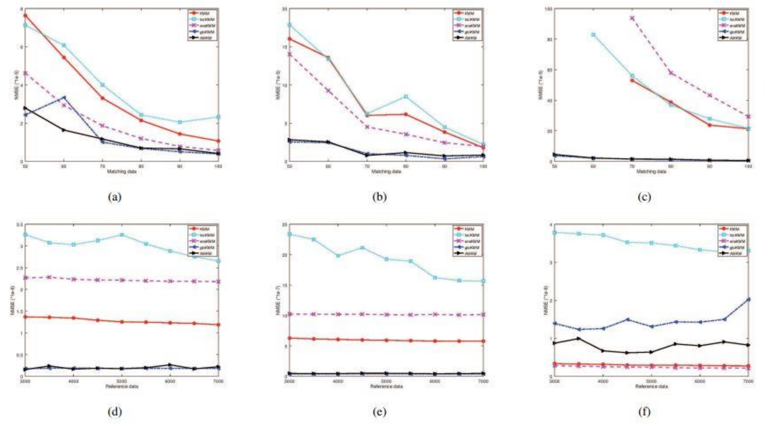

Fig. 4. The experimental results of different KMM methods on various data sets. (a)-(c): The obtained NMSE on Monks, Ionosphere, and Climate data. (d)-(f): The obtained NMSE on Forest, Letter and CIFAR data sets.

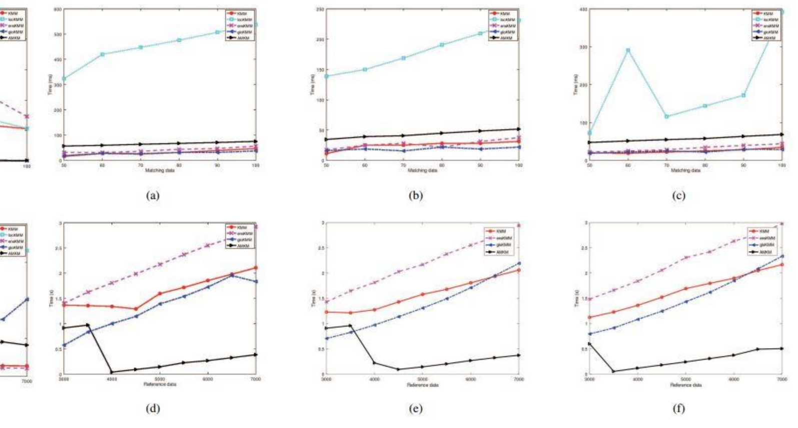

Fig. 5. The time complexities of different KMM methods on various data sets. (a)-(c): The time complexities (milliseconds) of different algorithms on Monks, Ionosphere, and Climate data sets. (d)-(f): The time complexities (seconds) of different algorithms on Forest, Letter and CIFAR data sets.

② **Scalable matching**

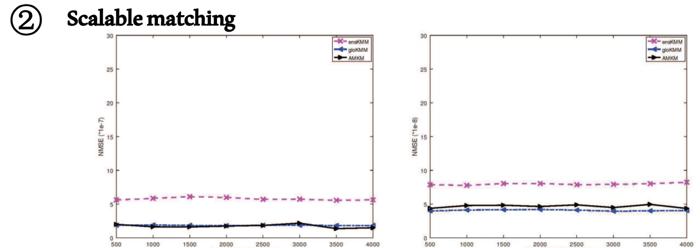

Fig. 6. The experimental results of scalable learning on Forest and Letter data sets: (a) Forest and (b) Letter

④ The obtained average NMSE ($\times 10^{-5}$ on Monks, Ionosphere, and Climate data sets. $\times 10^{-7}$ on Forest, Letter and Cifar data sets) from AMKM method with different quantities of selected top important instances n_s.

Top instances n_s		50	100	150	200
Monks	gloKMM	0.992	1.005	1.03	1.018
	AMKM	1.249	1.209	1.056	1.076
Ionosphere	gloKMM	1.034	1.052	1.03	1.025
	AMKM	0.839	0.757	0.128	0.108
Climate	gloKMM	1	1.051	1.026	1.001
	AMKM	1.531	1.475	1.468	1.453
Top instances n_s		100	200	300	400
Forest	gloKMM	1.901	1.739	1.699	1.667
	AMKM	1.782	1.49	1.353	1.381
Letter	gloKMM	0.412	0.394	0.393	0.385
	AMKM	0.469	0.413	0.43	0.419
CIFAR	gloKMM	12.56	9.645	6.715	6.388
	AMKM	9.074	7.741	5.93	5.897

③ The obtained average NMSE ($\times 10^{-5}$ on Monks, Ionosphere, and Climate data sets. $\times 10^{-7}$ on Forest, Letter and CIFAR data sets) from AMKM method with different quantities of randomly selected instances n.

Selected instances n		50	100	150	200
Data sets	Monks	1.059	1.121	1.254	1.204
	Ionosphere	0.706	0.749	0.734	0.709
	Climate	1.538	1.418	1.702	1.463
Selected instances n		100	200	300	400
Data sets	Forest	1.804	2.006	2.124	2.278
	Letter	0.502	0.509	0.535	0.528
	CIFAR	7.312	6.679	6.457	7.117

The average cost times (milliseconds) of AMKM with different quantities of randomly selected instances n.

Selected instances n		50	100	150	200
Data sets	Monks	63.031	65.618	71.402	74.994
	Ionosphere	41.284	43.677	45.672	49.268
	Climate	54.647	58.836	62.427	65.619
Selected instances n		100	200	300	400
Data sets	Forest	68.741	121.4	199.192	264.816
	Letter	77.311	117.211	191.407	258.434
	CIFAR	163.881	215.543	291.339	364.549

Conclusions

● The proposed AMKM method is able to achieve calculation efficiency with selective reference instances, and importance estimation of whole data can be avoided.

● Scalable matching of kernel means can be conducted in the proposed method.

● Experimental results on a variety of data sets demonstrate that, the proposed method is able to obtain ideal KMM performance while promising efficiency can be achieved.

References

1. M. Sugiyama, S. Nakajima, H. Kashima, P. V. BÜnau, M. Kawanabe, Direct Importance Estimation with Model Selection and Its Application to Covariate Shift Adaptation, in *Proc.* Advances in Neural Information Processing Systems, 2007.
2. D. Erdogmus and J. C. Principe, Generalized Information Potentials Criterion for Adaptive System Training, IEEE Trans. Neural Networks, vol. 13, no. 5, pp. 1035-1044, 2015.
3. T. Kanamori, S. Hido, and M. Sugiyama, A Least-squares Approach to Direct Importance Estimation, Journal of Machine Learning Research, vol. 10, pp. 1391-1445, 2009.
4. Y. Q. Miao, A. K. Farahat, M. S. Kamel, Locally Adaptive Density Ratio for Detecting Novelty in Twitter Streams, in *Proc.* International Conference on World Wide Web, 2015.
5. Y. Q. Miao, A. K. Farahat, M. S. Kamel, Ensemble Kernel Mean Matching, in *Proc.* International Conference on Data Mining, 2015.

Contact

Miao Cheng
School of Computer Science and Information Engineering
Guangxi Normal University
Guilin, Guangxi, China
Email: mcheng@mailbox.gxnu.edu.cn
miao_cheng@outlook.com