Adaptive Matching of Kernel Means

Miao Cheng

School of Computer Science and Information Engineering
Guangxi Normal University
Guilin, Guangxi, China

Email: mcheng@mailbox.gxnu.edu.cn

Overview

Pattern Matching

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.

Importance Estimation

- For instance, the target groups of people are more **important** for certain sale businesses, as professional market investigations disclosed.

(a)

(b)

Figure: Importance estimation in business market.

Importance Estimation

1. Media information of **matched knowledge** are more attractive for corresponding persons in human society, associated with common characteristics, e.g., ages, locations, favorites, and so on.

(a)

(b)

Figure: Importance estimation in social media.

Kernel Mean Matching

1. As a standard approach, **kernel mean matching (KMM)** brings broad attentions for importance estimation, and knowledge discovery as well.
2. Derived from conception of **training (matching)** and **testing (reference) data** in pattern recognition, the importance of a given sample $w(x)$ [Sugiyama07] is given by the ratio of densities $p_r(x)$ and $p_m(x)$ as

$$w(x) = \frac{p_r(x)}{p_m(x)}. \tag{1}$$

Kernel Mean Matching

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

$$J_{KMM} = \arg\min_{\alpha} \left\| \frac{1}{n_m} \sum_{i=1}^{n_m} \alpha(x_i)\phi(x_i) - \frac{1}{n_r} \sum_{i=1}^{n_r} \phi(x_i) \right\|^2$$
$$= \arg\min_{\alpha} \left[\frac{1}{n_m^2} \sum_{i,j=1}^{n_m} \alpha_i k(x_i, x_j) \alpha_j - \right.$$
$$\left. \frac{2}{n_m n_r} \sum_{i=1}^{n_m} \sum_{j=1}^{n_r} \alpha_i k(x_i, x_j) + \frac{1}{n_r^2} \sum_{i,j=1}^{n_r} k(x_i, x_j) \right] \quad (2)$$

.

Kernel Mean Matching

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

$$J(\alpha) = \arg\min_{\alpha} \left[\frac{1}{2} \alpha^T K_{m,m} \alpha - \frac{n_m}{n_r} \alpha K_{m,r} e \right], \qquad (3)$$

.

2. As a result, the ideal α can be analytically obtained with a penalty item, e.g.,

$$\alpha = \frac{n_m}{n_r} (K_{m,m} + \lambda I)^{-1} K_{m,r} e \qquad (4)$$

.

3. After obtaining α, the importance of instances with *Gaussian* model is calculated as

$$\widehat{w}(x) = \sum_{i=1}^{n_m} \alpha_i k_{ga}(x, x_i^m) \qquad (5)$$

.

Global Importance

1. 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,

$$\widetilde{w_i} = \int_r \phi\left(x_i^r\right) dx = \sum_{j=1}^{n_r} k\left(x_i^r, x_j^r\right) \tag{6}$$

or equivalently,

$$\omega(x_i^r) = \frac{\int_r \phi\left(x_i^r\right) dx}{\sum_{j=1}^{n_r} \widetilde{w_j}} = \frac{\sum_{j=1}^{n_r} k\left(x_i^r, x_j^r\right)}{\sum_{j=1}^{n_r} \widetilde{w_j}}. \tag{7}$$

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.
- **Output:** The estimated importance $w(x)$.

1. Calculate the importance of each reference instance as done in (7), 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 in (3) and obtain the optimal coefficients α.
4. Calculate estimated importance of instances by $w(x)$.

Adaptive Matching

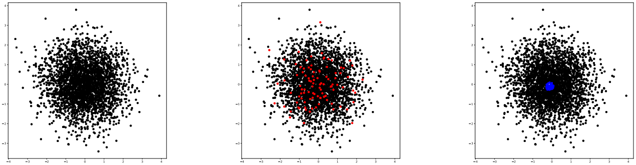
1. Select a **subset** of reference data for estimation of importance, and it is verified the estimated importance results in **acceptable ranking** of reference data.
2. As a consequence, the modified estimation of instance importance is defined as

$$\omega(x_i^r) = \frac{\int_{n_s} \phi\left(x_i^r\right) dx}{\sum_{j=1}^{n_s} \widetilde{w_j}} = \frac{\sum_{j=1}^{n_s} k\left(x_i^r, x_j^r\right)}{\sum_{j=1}^{n_s} \widetilde{w_j}} \tag{8}$$

.

Adaptive Matching

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

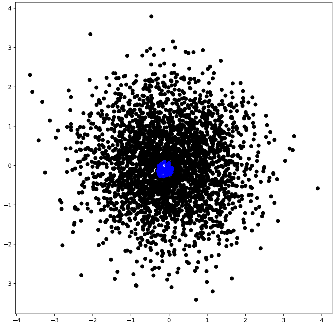


(a) (b) (c)

Figure: 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.

Adaptive Matching of Kernel Means

Nevertheless, the obtained matching results rely on *unaccurate* means, and further calibration may be necessary.



(a) Top important points

Adaptive Matching of Kernel Means

1. Selectively adaptive matching is repeated several times, and then a **fusion** stage is to adopted to learn the ideal matching.
2. Suppose that, there are t approximately matching results, i.g., $\mathrm{M}_i = [\alpha_{i,1}, \alpha_{i,2}, \cdots, \alpha_{i,n_s}], \quad i = 1, 2, \cdots, t$, then KMM can be defined as a combination of different matching coefficients,

$$\begin{aligned} J(\beta) \quad &= \arg\min_{\beta_i} \sum_{i=1}^{t} \sum_{j=1}^{n_s} \left(\frac{1}{2} \gamma_{i,j}^T K_{m,m} \gamma_{i,j} - \frac{n_m}{n_r} \gamma_{i,j} K_{m,r} e \right) \\ \textit{with} \quad & \gamma_{i,j} = \alpha_{i,j} \beta_i \end{aligned} \tag{9}$$

Adaptive Matching of Kernel Means

1. As a traditional consideration, the constraints of such **quadratic programming (QP)** can be referred to certain equivalent conditions of β_i as well as the *lower* or *upper* bounding.
2. The relaxed constraint conditions are adopted to restrict β_i to be values **larger** than *zero* only,

$$\begin{aligned} J(\beta) &= \arg\min_{\beta_i} \sum_{i=1}^{t} \sum_{j=1}^{n_s} \left(\frac{1}{2} \gamma_{i,j}^T K_{m,m} \gamma_{i,j} - \frac{n_m}{n_r} \gamma_{i,j} K_{m,r} e \right) \\ \textit{with} \quad & \gamma_{i,j} = \alpha_{i,j} \beta_i \\ s.t. \quad & \beta_i \geq 0 \end{aligned} \tag{10}$$

.

AMKM 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 instances n, desired number of important instances n_s for matching.
- **Output:** The estimated importance $w(x)$.

1. **While:** *The desired repetition t has never reached*
   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 Algorithm 1.
2. Calculate the fusion coefficients by solving the QP defined in (10).
3. 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.

Discussion

Theoretical bases

1. The measure of selection of instances is identical with information potentials [Erdogmus02],

$$V\left(x^{r}\right)=\frac{1}{n_{s}^{2}}\sum_{i=1}^{n_{s}}\sum_{j=1}^{n_{s}}G\left(x_{i}^{r}-x_{j}^{r},2\sigma^{2}\right)\tag{11}$$

.

2. The *Renyi* quadratic entropy can be succinctly written as,

$$H\left(x\right)=-\int_{x^{r}}\log p^{2}\left(x^{r}\right)dx=-\log V\left(x^{r}\right)\tag{12}$$

.

Discussion

Theoretical bases

1. The selected important instances can be explained as the ones corresponding to the **maximum** information potentials of the pre-selected random instances, and the **minimum** disorder of data as well.

Experiments

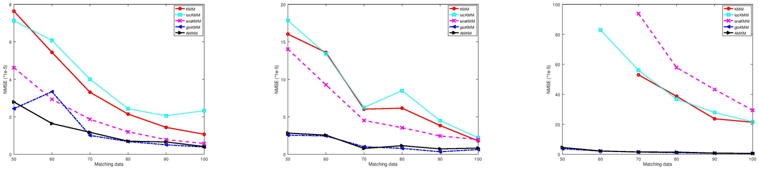
1. The efficiency of proposed AMKM are evaluated with several state-of-the-art methods, i.g., standard KMM [Kanamori09], locally KMM (locKMM) [Miao15], ensemble KMM (ensKMM) [Miao15], global KMM (gloKMM).
2. The details of different data sets

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

Experiment 1

1. A fixed size of reference data is set to be 500, 250, 400 with random selection for Monks, Ionosphere, and Climate data sets. And different sizes of instances are selected to be the matching data, which are changed in range from 50 to 100.
2. Among instances of Forest, Letter and CIFAR data sets, respective 500 instances are randomly selected to be matching data, while instances in the range from 3,000 to 7,000 are selected to be the reference data during each execution.

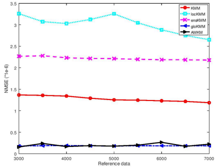
Experiment 1



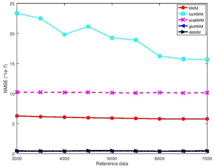
(a) Monks (b) Ionosphere (c) Climate

Figure: The obtained NMSE on Monks, Ionosphere, and Climate data sets with different sizes of matching data.

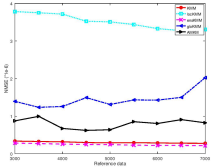
Experiment 1



(a) Forest



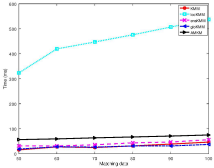
(b) Letter



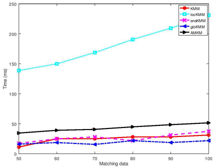
(c) Cifar

Figure: The obtained NMSE on Forest, Letter and CIFAR data sets with different sizes of reference data.

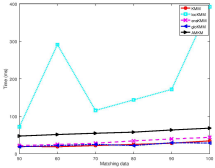
Experiment 1



(a) Monks



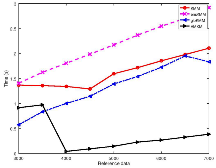
(b) Ionosphere



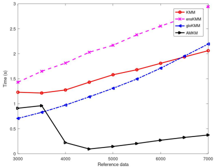
(c) Climate

Figure: The cost time complexities (milliseconds) on Monks, Ionosphere, and Climate data sets with different sizes of matching data.

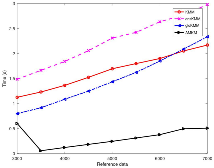
Experiment 1



(a) Forest



(b) Letter

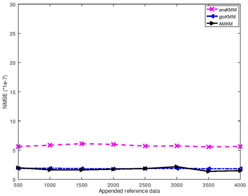


(c) Cifar

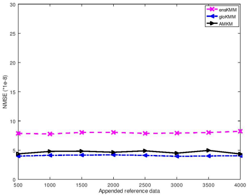
Figure: The cost time complexities (seconds) on Forest, Letter and CIFAR data sets with different sizes of reference data.

Experiment 2

1. For Forest and Letter data sets, 500 and 3,000 instances are respectively selected to be matching data and reference data.
2. Then, reference data are appended with another 500 instances each time for batch matching.



(a) Forest



(b) Letter

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

Expriment 3

Monks, Ionosphere, and Climate data sets

1. 70 instances are randomly selected to be matching data and another 500, 250, 400 instances are respectively selected to be the reference data.
2. The randomly selected instances of AMKM are set to be in range from 50 to 200, while top 100 important instances are used for final matching.

Forest, Letter, and CIFAR data sets

1. 500 instances are selected from respective three data sets to be matching data, while 4,000 instances are selected to be reference data.
2. The randomly selected instances of AMKM are set to be in range from 100 to 400, and the top 100 instances are adopted.

Expriment 3

Table: 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

Expriment 3

Table: 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

Experiment 4

Fixed 50 instances are randomly selected, and different quantities of important instances are selected for matching of gloKMM and AMKM.

Table: The obtained average NMSE ($\times 10^{-5}$ on Monks, Ionosphere, and Climate 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

Experiment 4

Table: The obtained average NMSE ($\times 10^{-7}$ on Forest, Letter and CIFAR data sets) from AMKM method with different quantities of selected top important instances n_s.

Top instances n_s		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

Conclusion

1. In this work, a novel KMM method is proposed to adaptive learning of KMM.
2. The proposed AMKM method is able to achieve calculation efficiency with selective reference instances, and importance estimation of whole data can be avoided.
3. Scalable matching of kernel means can be conducted in the proposed method.
4. 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.

Thank You for Your Attentions