

Supplementary Notes to KARL Library

Tsz Nam Chan Man Lung Yiu
Department of Computing
Hong Kong Polytechnic University
 {cstnchan, csmlyiu}@comp.polyu.edu.hk

Leong Hou U
Department of Computing and Information Science
University of Macau
 ryanlhu@umac.mo

I. SOTA ALGORITHM

SOTA [5] adopts the multi-step approach [7], [1], [2] for evaluating the bound functions (LB and UB), combining with existing indexing structures to support kernel density classification (query type I- τ). We extend their algorithm to support different other types of machine learning models, including approximate kernel density estimation (query type I- ϵ), 1-class SVM (query type II- τ) and 2-class SVM (query type III- τ). We name this algorithm as Multi-Step Kernel Prediction (MSKP).

Algorithm 1 Multi-Step Kernel Prediction (MSKP)

```

1: procedure MSKP(query  $q$ , weights  $\{w_1, \dots, w_n\}$ , tree  $T$ , threshold  $\tau$ )
2:   Create a max-heap  $H$ 
3:    $e \leftarrow T.R_{root}$ 
4:    $\widehat{lb} \leftarrow LB(q, e)$ ,  $\widehat{ub} \leftarrow UB(q, e)$ 
5:   enheap  $e$  to  $H$ 
6:   while  $H \neq \emptyset$  do
7:     if  $\widehat{lb} \geq \tau$  then
8:       return 1
9:     if  $\widehat{ub} < \tau$  then
10:      return -1
11:     $R \leftarrow \text{deheap an entry in } H$ 
12:     $\widehat{lb} \leftarrow \widehat{lb} - LB(q, e, R)$ ,  $\widehat{ub} \leftarrow \widehat{ub} - UB(q, e, R)$ 
13:    if  $e$  is leaf then
14:       $temp \leftarrow \sum_{p_i \in e.R} w_i \mathcal{K}(q, p_i)$ 
15:       $\widehat{lb} \leftarrow \widehat{lb} + temp$ ,  $\widehat{ub} \leftarrow \widehat{ub} + temp$ 
16:    else
17:      for each child  $e_c$  in  $e$  do
18:         $\widehat{lb} \leftarrow \widehat{lb} + LB(q, e_c, R)$ 
19:         $\widehat{ub} \leftarrow \widehat{ub} + UB(q, e_c, R)$ 
20:      enheap  $e_c$  to  $H$ 

```

Algorithm 1 is only used for the classification-based queries (types I- τ , II- τ and III- τ). For query type I- ϵ , the input threshold τ should be replaced by relative error ϵ . We also need to replace lines 7 to 10 by the following termination condition.

Algorithm 2 Termination condition for query type I- ϵ

```

1: if  $\widehat{ub} \leq (1 + \epsilon)\widehat{lb}$  then
2:   return  $\frac{\widehat{lb} + \widehat{ub}}{2}$ 

```

There is another advance implementation for the termination condition for Algorithm 3, which is based on the unpublished paper of our work [3] in Earth Mover's Distance. This paper is now under submission to TKDE.

Algorithm 3 Advanced termination condition for query type I- ϵ [3]

```

1: if  $\frac{\widehat{ub} - \widehat{lb}}{\widehat{ub} + \widehat{lb}} \leq \epsilon$  then
2:    $R = \frac{2 \times \widehat{lb} \times \widehat{ub}}{\widehat{lb} + \widehat{ub}}$ 
3:   return  $R$ 

```

II. RELATIONSHIP BETWEEN SOTA AND KARL

Both SOTA and KARL utilize the same algorithm MSKP. However, compared with SOTA, KARL utilizes tighter bound functions to boost up the efficiency performance. During the implementation of KARL, we only replace LB and UB by our bounding functions [4].

III. AUTO-TUNING (OFFLINE)

In Section III-C of our paper [4], we develop the auto-tuning method for obtaining the best index construction in the offline stage. First, we sample 1000 queries from the query dataset as the workload \mathcal{WL} . To ensure the fairness, these queries will not be used in the online phase. Then, our algorithm Auto chooses the index from either kd-tree [6] or ball-tree [8] and the most suitable leaf node capacity from the capacity list $CL = \{10, 20, 40, 80, 160, 320, 640\}$ in which the best setting bs provides the fastest running time f_{time} in the selected workload \mathcal{WL} .

Algorithm 4 shows the pseudocode of our implementation.

Algorithm 4 Auto-tuning (Offline)

```

1: procedure AUTO(Workload  $\mathcal{WL}$ , weights  $\{w_1, \dots, w_n\}$ , tree  $T$ , threshold  $\tau$ , capacity list  $CL$ , tree list  $TL = \{\text{kd}, \text{ball}\}$ )
2:    $bs \leftarrow \text{null}$ 
3:    $f_{time} \leftarrow \infty$ 
4:   for  $t \in TL$  do
5:     for  $c \in CL$  do
6:        $T \leftarrow \text{Build tree } t \text{ with capacity } c$ 
7:        $s \leftarrow \text{timer}()$ 
8:       for  $q \in \mathcal{WL}$  do
9:          $MSKP(q, weights, T, \tau)$ 
10:       $e \leftarrow \text{timer}()$ 
11:       $temp \leftarrow e - s$ 
12:      if  $temp \leq f_{time}$  then
13:         $bs \leftarrow \{t, c\}$ 
14:         $f_{time} \leftarrow temp$ 
15:      Remove tree } t
16:   return  $bs$ 

```

REFERENCES

- [1] T. N. Chan, M. L. Yiu, and K. A. Hua. A progressive approach for similarity search on matrix. In *SSTD*, pages 373–390, 2015.

- [2] T. N. Chan, M. L. Yiu, and K. A. Hua. Efficient sub-window nearest neighbor search on matrix. *IEEE Trans. Knowl. Data Eng.*, 29(4):784–797, 2017.
- [3] T. N. Chan, M. L. Yiu, and L. H. U. The Power of Bounds: Answering Approximate Earth Movers Distance with Parametric Bounds. *Submitted to IEEE Trans. Knowl. Data Eng.*
- [4] T. N. Chan, M. L. Yiu, and L. H. U. KARL: Fast kernel aggregation queries. In *ICDE*, 2019.
- [5] E. Gan and P. Bailis. Scalable kernel density classification via threshold-based pruning. In *ACM SIGMOD*, pages 945–959, 2017.
- [6] H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann, 2006.
- [7] T. Seidl and H. Kriegel. Optimal multi-step k-nearest neighbor search. In *SIGMOD*, pages 154–165, 1998.
- [8] J. K. Uhlmann. Satisfying general proximity/similarity queries with metric trees. *Inf. Process. Lett.*, 40(4):175–179, 1991.