


# Selecting Indispensable Edge Patterns With Adaptive Sampling and Double Local Analysis for Data Description

Huina Li, Xuchang University, China

Yuan Ping, Xuchang University, China\*

 <https://orcid.org/0000-0001-7703-4637>

## ABSTRACT

Support vector data description (SVDD) inspires us in data analysis, adversarial training, and machine unlearning. However, collecting support vectors requires pricey computation, while the alternative boundary selection with  $O(N^2)$  is still a challenge. The authors propose an indispensable edge pattern selection method (IEPS) for data description with direct SVDD model building. IEPS suggests a double local analysis to select the global edge patterns. Edge patterns belong to a subset of the target problem of SVDD and its variants, and neighbor analysis becomes pivotal. While an excessive number of participating data result in redundant computations, an insufficient number may impede data separability or compromise the model's quality. Consequently, a data-adaptive sampling strategy has been devised to ascertain an optimal ratio of retained data for edge pattern selection. Extensive experiments indicate that IEPS keeps indispensable edge patterns for data description while reducing the interference in the norm vector generation to guarantee the effectiveness for clustering analysis.

## KEYWORDS

Adaptive Sampling, Cluster Analysis, Edge Pattern Selection, K-Means++, Support Vector Data Description

## INTRODUCTION

Inspired by support vector classifier, support vector data description (SVDD) (Tax & Duin, 1999) characterizes a data set by obtaining the spherically shaped boundary. Through a model built to describe the target data set, it benefits a wide range of applications, such as image description (Aslani & Seipel, 2021), novelty discovery (Hu et al., 2023), adversarial training (C. Chen et al., 2023), and machine unlearning (M. Chen et al., 2023). However, in collecting support vectors (SVs) for data description, the conventional solution conducts model training through solving a quadratic programming optimization problem. It poses a computational complexity of  $O(N^3)$  where  $N$  is the number of data points. Evidently, pricey computations may significantly degrade SVDD's applicability.

Let  $\mathcal{X}$  be a data set with  $N$  data points  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  where  $\mathbf{x}_i \in \mathbb{R}^d (i \in [1, N])$  in data space. The pricey model training is generally caused by solving a quadratic programming problem in terms

DOI: 10.4018/JCIT.335945

\*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

of iterative analysis on a  $N \times N$  kernel matrix. Furthermore, the number of iterative analysis is usually large and uncertain, yet a great value for the final coefficient vector  $\beta$  exacerbates the practical time-cost. Efficient solver for the quadratic programming problem is the major preference for improvement, such as the solver of dual coordinate descent (Y. Ping et al., 2017). However, the computational complexity falling in the range of  $O(N^2)$  and  $O(N^3)$  upon the specific case is still pricey (Arslan et al., 2022). Another intuitive way of improvement is to select the most representative subset of  $\mathcal{X}$ . However, few works in the literature focus on the subset's representativeness or purity strongly related to SVDD. They frequently select a subset on the basis of random sampling, data geometry analysis, and neighborhood relationships. For instance, Kim et al. (2015) define a sample rate  $\rho$  ( $0 < \rho < 1$ ) to regulate the randomly selected  $\rho N$  data points during model training, while Jung et al. (2010) and Gornitz et al. (2018) leverage data geometry information by incorporating  $k$ -means to partition  $\mathcal{X}$  into  $K$  subsets for local model training and subsequent global merge. However, these data points employed for model training, whether obtained through random sampling or cluster-based geometry analysis, may not accurately capture the true distribution of  $\mathcal{X}$ . Random sampling introduces changes in the densities of all the retained data groups that significantly impacts data separability. The circle-like pattern hypothesis employed in subsets collection for local model training may exacerbate the adverse effects of irregular cluster shapes. Despite achieving substantial efficiency improvements, these methods often result in highly unstable accuracies. As the superset of support vectors (SVs) (Y. Ping et al. 2015), boundary generally makes an equivalent contribution to the construction of demarcation hyperplanes (Chen et al., 2023). On the basis of neighborhood relationships, Aslani and Seipel (2021) introduce locality-sensitive hashing (LSH) to gather instances near decision boundaries and eliminate nonessential ones. However, it retains many inners that may be more suitable for constructing a classifier for multi-classes problems rather than describing clusters with arbitrary shapes. Furthermore, Y. Ping et al. (2015) and Y. Ping et al. (2019) utilize the boundary to directly reformulate the dual problem. Despite achieving stable performance, the boundary selection becomes computationally expensive with a large value of  $N$ .

As depicted by Figure 1, boundary consists of edge and border (Li & Maguire, 2011), which extend beyond the essential requirement for cluster discovery and description. Specifically, for unsupervised learning, boundary is effectively profiled by edge patterns alone since no shared borders should be considered. The border refers to the connection between two nearest neighboring clusters. Simultaneously, both of the kernelized SVDD (Cevikalp et al., 2020) and the convex decomposition strategy (Y. Ping et al., 2020) suggest that only a subset of samples on the decomposed convex hulls is necessary for accurate data description. Thus, an optimal edge pattern selection method should preserve cluster shapes, excluding inners and border patterns, and eliminate redundant instances even though they reside on edges, as they contribute nothing in additional contributions to cluster description.

Toward this requirement, our critical observations based on the principle of SVDD encompass three aspects: 1) Inners are exclusive to each cluster, as they can also be considered as outliers of other clusters; 2) Border patterns are solely shared by any two connected clusters or prototypes, such as convex hulls in (Y. Ping et al., 2019), and this sharing is independent of the clustering method; 3) Edge patterns represent the only type of data points shared by all clusters, regardless of the chosen clustering method. Motivated by these insights, we propose an indispensable edge pattern selection method (IEPS) with data adaptive sampling and double local analysis strategies. IEPS maximizes the utility of the shrinkable boundary selection algorithm (SBS) (Y. Ping et al., 2019),  $p$ -stable distributions based LSH (pdLSH) (Datar et al., 2004), and  $k$ -means++ (Arthur & Vassilvitskii, 2007). The main contributions lie in:

- (1) We propose an edge pattern selection strategy with double local analysis (EPSDLA) based on two rounds of data partitioning using  $k$ -means++. The inherent instability of  $k$ -means++ is leveraged

24 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: [www.igi-global.com/article/selecting-indispensable-edge-patterns-with-adaptive-sampling-and-double-local-analysis-for-data-description/335945](http://www.igi-global.com/article/selecting-indispensable-edge-patterns-with-adaptive-sampling-and-double-local-analysis-for-data-description/335945)

## Related Content

---

### Challenges of IT Adoption at Educational Institutions: Lessons From Bangladesh

Saad Hasan, Mohammad Rashedul Hoque, Shafkat Reza Chowdhury, Ashfaque A. Mohiband Md. Abdul Ahad (2020). *International Journal of Information Systems and Social Change* (pp. 66-90).

[www.irma-international.org/article/challenges-of-it-adoption-at-educational-institutions/243716](http://www.irma-international.org/article/challenges-of-it-adoption-at-educational-institutions/243716)

### Usability Evaluation of E-Learning Systems

Shirish C. Srivastava, Shalini Chandraand Hwee Ming Lam (2009). *Encyclopedia of Information Science and Technology, Second Edition* (pp. 3897-3903).

[www.irma-international.org/chapter/usability-evaluation-learning-systems/14158](http://www.irma-international.org/chapter/usability-evaluation-learning-systems/14158)

### Electronic Supply Chain Partnerships: Reconsidering Relationship Attributes in Customer-Supplier Dyads

Rebecca Angelesand Ravi Nath (2003). *Information Resources Management Journal* (pp. 59-84).

[www.irma-international.org/article/electronic-supply-chain-partnerships/1248](http://www.irma-international.org/article/electronic-supply-chain-partnerships/1248)

### Optimization of Favourable Test Path Sequences Using Bio-Inspired Natural River System Algorithm

Nisha Ratheeand Rajender Singh Chhillar (2021). *Journal of Information Technology Research* (pp. 85-105).

[www.irma-international.org/article/optimization-of-favourable-test-path-sequences-using-bio-inspired-natural-river-system-algorithm/274280](http://www.irma-international.org/article/optimization-of-favourable-test-path-sequences-using-bio-inspired-natural-river-system-algorithm/274280)

### Hierarchies in Multidimensional Databases

Elaheh Pourabbas (2005). *Encyclopedia of Information Science and Technology, First Edition* (pp. 1327-1332).

[www.irma-international.org/chapter/hierarchies-multidimensional-databases/14433](http://www.irma-international.org/chapter/hierarchies-multidimensional-databases/14433)