ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.082

Research on Local Neighbor Density Similarity Based SVDD for Data Classification

Yiwei Yuan [1,2+], Kezhuang Liu [1], Yingmin Yi [1] and Lei Bai [3]

[1] School of Automation and Information Engineering, Xi'an University of Technology, Xian, China

[2] Key Laboratory of Shaanxi Province for Complex System Control and Intelligent Information Processing, Xian, China

3 School of Electrical Engineering, Shaanxi Polytechnic Institute, Xianyang, China

Abstract. Due to the high dimension of industrial process data, there are aliasing areas when building hyper-spheres of different types of data by SVDD algorithm, which will lead to inaccurate results in condition identification. Focusing on the problem, a method based on local neighbor density similarity support vector data description (LNDS-SVDD) is proposed for data classification. Utilizing the local similar density in discriminating data similarity, the LNDS-SVDD can further judge the classification of samples distributed in the aliasing area using the density information. In this paper, a randomly generated synthetic data set and a real industrial process data set is used for simulating the proposed data classification algorithm. The results show that the algorithm has high classification accuracy, which is an effective condition identification method for the industrial process.

Keywords: SVDD, aliasing area, local neighbor density similarity, condition identification

1. Introduction

Support Vector Domain Description (SVDD) is an effective classification method. This method encloses most or all of the samples by constructing a minimum enclosed hyper-sphere[1]. This method has been in-depth research since it was proposed in 1999, and it has been widely used in problems such as gross error elimination and fault diagnosis[2-6]. However, due to the complex structure of the system and variable working conditions, some actual data will inevitably be located in the hyper-sphere area surrounded by different categories, which reduces the fault classification performance.

The most basic division of aliasing region samples is to make a separation surface at the intersection of two hyper-spheres. The samples on both sides of the separation surface belong to different categories[7]. Although this method is easy to implement, the classification error of the aliased region samples is extremely large. Some scholars believe that the separation surface should be biased in favor of the hyper-sphere with the small radius, which improves the classification accuracy[8]. Some researches construct the membership function based on the distance method as a method to further distinguish the data types of aliasing regions[9].

In this paper, a method for constructing density similarity functions based on relative distance is derived, which improves the classification accuracy of SVDD in aliasing area samples to a certain extent.

2. Basic Principle of SVDD

The basic algorithm of SVDD is as follows. Consider that the training data set is a target data set containing samples in the space. SVDD looks for the smallest hyper-sphere that can surround the target data as much as possible. The center and radius of the hyper-sphere are respectively with R, the SVDD problem can be expressed as follow:

$$\begin{aligned} &\min \quad R^2 + C\sum_{i=1}^{N}\xi_i, \\ &s.t. \quad \|x_i - \alpha\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0,\ \forall i, \end{aligned} \tag{1}$$

[+] Corresponding author.
E-mail address: yyw@xaut.edu.cn.

where ξ_i is the relaxation factor. By adjusting part of the sample points to be outside the hyper-sphere to enhance the robustness of SVDD when the data set contains gross errors, the parameter C is used to balance the relationship between the rate of misclassification of samples and the volume of the hyper-sphere. After the optimization problem of the above formula is further simplified, the dual optimization problem can be obtained as follow:

$$\begin{aligned} &\max \quad \sum_i \alpha_i (x_i \cdot x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j), \\ &s.t. \quad \sum_i \alpha_i = 1, a = \sum_i \alpha_i x_i, 0 \le \alpha_i \le C, \end{aligned} \tag{2}$$

where "•" represents the inner product, and α_i is the Lagrange factor. Different values of the Lagrange factor correspond to different positions of the sample point x_i: when x_i is located in the hyper-sphere,$\alpha_i = 0$; when x_i is located on the boundary of the hyper-sphere, $0 < \alpha_i < C$; when x_i is outside the hyper-sphere, $\alpha_i = C$. All the sample points with $\alpha_i \ne 0$ are called Support Vector (SV).

The center of the hyper-sphere can be obtained by weighting the support vector points, and the radius R of the hyper-sphere can be calculated from the distance from the BSV to the center of the sphere. If the test point z satisfies:

$$\|z-a\|^2 = (z \cdot z) - 2\sum_i \alpha_i (z \cdot x_i) + \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \le R^2 \tag{3}$$

Then the test point is considered to belong to the same class as the training sample, otherwise it is judged to be other types of data. Since the spherical description cannot guarantee a reasonable and compact boundary shape in all cases, the kernel function transformation technique can be used to obtain the original inner product operation through the kernel function calculation without knowing the specific form of the nonlinear mapping, then the above formula is transformed for:

$$\begin{aligned} &\max \quad \sum_i \alpha_i K(x_i \cdot x_i) - \sum_{i,j} \alpha_i \alpha_j K(x_i \cdot x_j), \\ &s.t. \quad \sum_i \alpha_i = 1, a = \sum_i \alpha_i x_i, 0 \le \alpha_i \le C, \\ &\|z-a\|^2 = K(z \cdot z) - 2\sum_i \alpha_i K(z \cdot x_i) + \sum_{i,j} \alpha_i \alpha_j K(x_i \cdot x_j) \le R^2. \end{aligned} \tag{4}$$

Usually,K is selected as a Gaussian kernel function, namely $K(x,y) = \exp(-\|x-y\|^2 / \sigma^2)$.

3. Sample Classification Method for Aliasing Regions

Assume that there are M types of samples through independent learning of the SVDD algorithm based on Gaussian kernel function to obtain M closed hyper-spheres $S_1, S_2, \cdots, S_M$. If the distance $D_j(x)$ between the test sample x and the center of the j-th hyper-sphere satisfies the Eq.(4), the sample x is considered to belong to the j-th category, namely:

$$I(x) = \begin{cases} 1, if & D_j(x) \le R_j \\ 0, if & D_j(x) \ge R_j \end{cases} \tag{5}$$

However, in practical applications, there may be more than one hyper-sphere that satisfies the above formula, so x may fall into two or more hyper-spheres, that is the cross-aliasing situation shown in Fig.1.

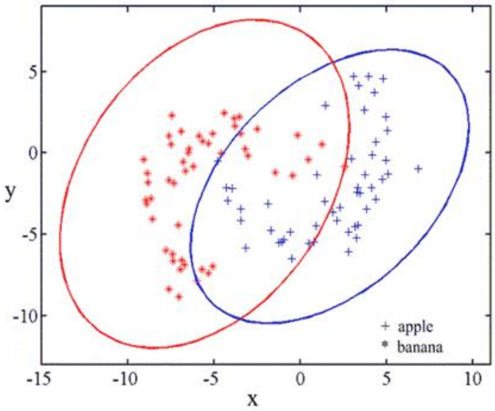


Fig. 1 Multi-class cross-aliasing of two-dimensional data

Simply classifying the aliasing area samples will reduce the classification accuracy of the SVDD algorithm. Therefore, on the basis of the SVDD algorithm, a data classification method based on the density of neighbors is proposed. This method is mainly based on the similarity between the aliasing region samples and their neighboring data densities, so as to accurately classify the aliasing region samples.

4. Data Classification Method Based on Local Neighbor Density Similarity

4.1. Nearest k-Neighbor Density of a Data

Suppose D is a data set and p is an object in D. Before giving the definition of the object k-density, the k-distance and k-distance neighborhood definitions of object p is introduced as follows[10]:

Definition 1: (k-distance of object p) For any natural number k, define the k-distance of p (k-distance(p)) as the distance ($d(p,o)$) between p and some object o, where o satisfies:

(1) There are at least k objects $o' \in D\backslash\{p\}$, such that $d(p,o') \le d(p,o)$

(2) There are at most $k-1$ objects $o' \in D\backslash\{p\}$, such that $d(p,o') < d(p,o)$.

Definition 2: (k-distance neighborhood of object p) Giving the k-distance of p, the k-distance neighborhood of p includes all objects whose distance from p does not exceed k-distance.

On the basis of the two definitions above, the k-density of p is defined as follows:

Definition 3: (k-density of p) For any natural number k, define the k-density of p as:

$$kd(p) = \frac{|N_k(p)|}{med\left\{dist(p,o) \mid o \in N_k(p)\right\}} \tag{6}$$

where $N_k(p)$ represents the k-distance neighborhood of p, $N_k(p)$ represents the number of objects in $N_k(p)$, $dist$ is the operator for finding the distance between two objects, and med is the median operator.

The data density is used to combine the two parameters of the distance between the data and the number of data in a given range to obtain the concept of "density", and then some key information of the data can be obtained according to the density

4.2. Data Classification Method Based on Local Neighbor Density Similarity (LNDS)

In order to determine the classification of the aliasing domain samples, the degree of belonging of the sample data needs to be calculated. This paper uses the density information of sample points to propose a data classification method based on neighbor density. By calculating the local density similarity between the sample point and its neighboring elements in the hyper-sphere it falls into, the sample point can be classified as the class with the greatest degree of similarity.

The local density similarity is calculated by the steps as follows:

Step 1: Determine the k neighborhood of sample p;

At this time, the elements in the k-neighborhood only contain the elements that p falls into several hyper-spheres. Assume that the set of elements belonging to the j-th category in the k-neighborhood of p is $P_j = \{d_{j1}, d_{j2}, \cdots, d_{jn}\}$.

Step 2: Calculate the spatial density of sample p in the j-th category and the k-density of each element in P_j;

At this time, the spatial density of p is only calculated based on the elements in P_j.

Step 3: According to the sample p and its k-nearest neighbor, calculate the similarity density of p;

The similar density of sample p is expressed as the ratio of the k-density of sample p to the average density of elements in P_j, expressed as:

$$lsd_j(x) = \frac{n * kd(p)}{\sum_{i=1}^{n} kd(d_{ji})} \tag{7}$$

In the formula, $kd(p)$ is the k-density of p, and $kd(d_{ji})$ is the k-density of the i-th element in P_j. The similarity density reflects the similarity between the sample p and the j-th type of data.

Definition 4: The local neighbor density similarity of sample p to the j-th category is:

$$\text{lnds}_{\text{j}}(p) = \frac{lsd_{\text{j}}(p)}{|\,lsd_{\text{j}}(p)-1\,|}. \tag{8}$$

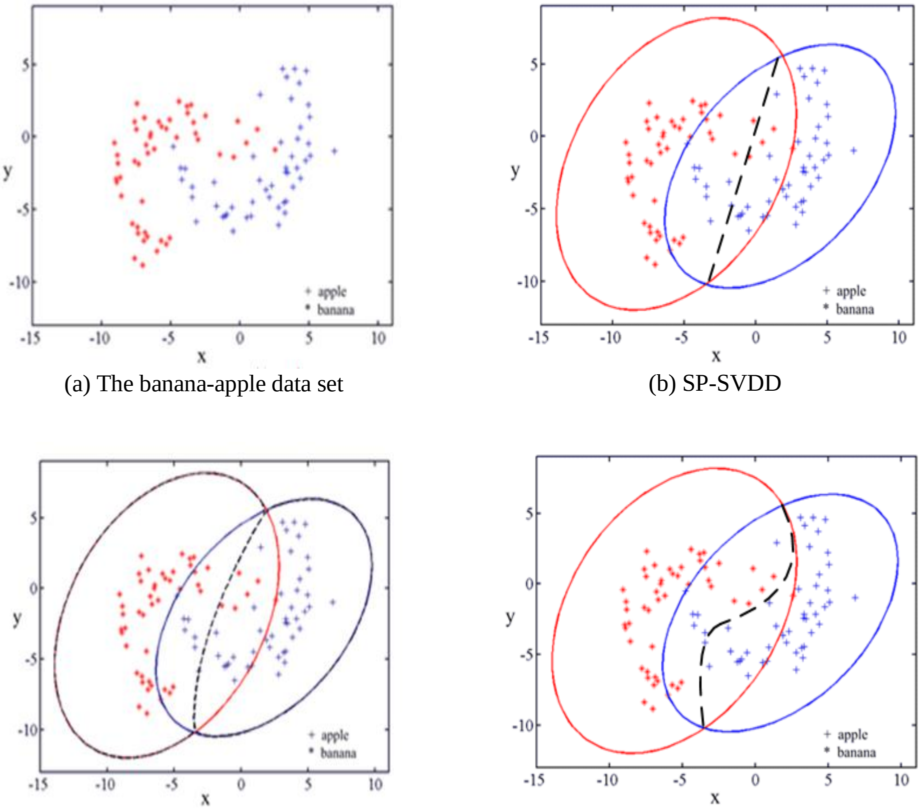
Divide the aliasing region samples into categories with high degree of belonging. The similarity of belonging of an object reflects the degree to which the aliasing region samples belong to each category. Since the local neighbor density is the core of the algorithm, this algorithm is called the SVDD classification method based on the local neighbor density similarity (LNDS-SVDD).

5. Simulation Result Analysis

In this part, we perform some experiments to evaluate the effectiveness of the proposed data classification algorithm (LNDS-SVDD) compared with the simple separation plane classification method (SP-SVDD) and the normalized radius classification method (NR-SVDD). The experiments are performed on a synthetic dataset and a real industrial dataset.

5.1. Synthetic Dataset

The simulation experiment is to use banana-apple data set to test the performance of the algorithm. The dataset contains 50 banana data and 50 apple data. The local nearest neighbor density similarity classification method needs to determine k value which is the number of neighbors. According to the determination method in [11], select 10 as the value of k in this experiment. Fig.2 shows the simulation results of three classification algorithms. It can be seen from the figure that there are 20 apple data and 12 banana data in the aliasing area. The SP-SVDD misjudges 4 banana data as apple class and at the same time misjudges 9 apple data as the banana data, the classification accuracy rate of this classification method is 88%. The NR-SVDD misjudges 6 banana data as apple class and 5 apple data as banana class. The classification accuracy of this classification method is 89%. The LNDS-SVDD misjudges 1 banana data as apple class and 6 apple data as banana class. The classification accuracy of this classification method is 93%. The experimental results show that the local nearest neighbor density classification method has a higher accuracy rate for the division of the aliasing region sample.



(a) The banana-apple data set

(b) SP-SVDD

(c) NR-SVDD

(d) LNDS-SVDD

Fig. 2 Classification results of different classification algorithms

5.2. Industrial Dataset

The industrial dataset used in the experiments was obtained from a pulverizing system. The experimental dataset includes 6 features which are ball mill load (BL), inlet-outlet pressure difference (IOD), inlet negative pressure (INP), outlet temperature (OT), coarse pulverized coal separator outlet pressure (CSOP) and fine pulverized coal separator outlet pressure (FSOP). The field dataset contains four kinds of process data in normal state, full mill fault, coal cut-off fault and blockage fault of powder return pipe, and there are 150 samples in each type of dataset. Table 1 lists some field data. Table II illustrates the classification results by different algorithms.

Table 1. Some of the field data

BL(t)	IOD(Pa)	INP(Pa)	OT(℃)	CSOP(Pa)	FSOP(Pa)
52.8	159.62	2267.41	99.1	-3676.62	-5790.87
52.81	159.62	2267.41	99	-3676.62	-5790.87
52.81	159.66	2260.26	99	-3659.6	-5778.78
52.81	159.66	2283.2	99	-3690.24	-5809.8

Table 2. The Field Data Classification Accuracy of Different Algorithms

	SP-SVDD	NR-SVDD	LNDS-SVDD
NR	89%	91%	95%
FF	90%	90%	94%
CF	86%	93%	93%
BF	92%	96%	100%

From Table2, it can be seen that the SP-SVDD has the worst performance in data classification, while the LNDS-SVDD performs best. Especially in the BF fault data classification, LNDS-SVDD can divide all BF fault data into correct categories. Obviously, the result verifies the effectiveness of proposed algorithm.

6. Conclusion

Adopting the SVDD classification algorithm based on the local neighbor density similarity can improve the classification accuracy of the SVDD algorithm for the aliasing area data. Firstly, the k-nearest neighbor method is used to obtain the nearest neighbor data that has an important influence on the aliasing area samples; then, the similarity between the density of the aliasing area sample and the density of the neighboring data is computed as the classification basis for the aliasing area samples. Through simulation verification, the LNDS-SVDD can accurately distinguish the types of data in the aliasing areas, meanwhile improve the accuracy of data classification.

7. Acknowledgement

The research is supported by the special scientific research program for local service in 2021 of Shaanxi Provincial Department of Education (21JC001), the scientific research projects of Shaanxi Provincial Department of Education (20JK0802), Natural Science Foundation of Shaanxi Province (2022JQ-609) and the scientific research projects of Shaanxi Polytechnic Institute(2020YKYB-051).

8. References

[1] TAX D M J, DUIN R P W. Support vector domain description [J]. *Pattern Recognition Letters*, 1999, 20 (11): 1191-1199.

[2] SANG J, ZHANG J, GUO T, et al. Detection of incipient faults in EMU braking system based on data domain description and variable control limit[J]. *Neurocomputing*, 2020, 383:348-358.

[3] ZHOU M , LIU Z, CAI Y, et al. Incipient fault detection based on energy efficiency and support vector data description[J]. *JOURNAL OF CHEMICAL ENGINEERING OF JAPAN*, 2019, 52(6):562-569.

[4] JEONG M K, AN S H, NAM K. SVDD-based financial fraud detection method through respective learnings of normal/abnormal behaviors[J]. *International Journal of Security and Its Applications*, 2016, 10 (3): 429-437.

[5] LI G, HU Y, CHEN H, et al. A sensor fault detection and diagnosis strategy for screw chiller system using support vector data description-based d-statistic and dv-contribution plots[J].*Energy and Buildings*, 2016, 133: 230-245.

[6] ZHANG M L, WANG T, WANG X P, et al. Online fault diagnosis algorithm based on variable dual-threshold SVDD with negative samples and its application [J]. *Journal of Vibration Engineering*, 2016, 29 (3): 555-560.

[7] CAI J Y, DU M J. A Novel Approach to Discriminate the Overlap Region of Multi-class Classification SVDD and Fault Diagnosis Application [J]. *Aerospace Control*, 2012, 30 (6): 83-88.

[8] ZHU Z B, WANG P L, SONG Z H. PCA-SVDD based fault detection and self-learning identification [J]. *JOURNAL OF ZHEJIANG UNIVERSITY (ENGINEERING SCIENCE)*, 2010, 44 (4): 652-658.

[9] LEE K Y, KIM D W, LEE K H. Density- Induced Support Vector Data Description[J]. *IEEE Trans Neural Network*, 2007, 18 (1): 284-289.

[10] CAO H , SI G Q , ZHU W Z, et al. Enhancing Effectiveness of Density-Based Outlier Mining[C]// Proceedings of the 2008 International Symposiums on Information Processing Conference. NW Washington, DC, United States: IEEE Computer Society, 2008:149-154.

[11] YUAN Y W, CAO H, ZHANG Y B, et al. Outlier mining based on neighbor-density-deviation with minimum hyper-sphere[J]. *Information Technology & Control*, 2016, 45(3):267-277.