Fuzzy Entropy Semi-supervised Support Vector Data Description

Trung Le, Dat Tran, Tien Tran, Khanh Nguyen, and Wanli Ma

Abstract—Support Vector Data Description (SVDD) is known as one of the best kernel-based methods for one-class classification problems. SVDD requires fully labelled data sets. However, in reality, an abundant amount of data can be easily collected, while the labelling process is often expensive, time-consuming, and error-prone. Therefore, partially labelled data sets are popular and easy to obtain. In this paper, we propose a semi-supervised learning method, Fuzzy Entropy Semi-supervised SVDD (FS3VDD), to extend SVDD to cope with partially labelled data sets. The learning model employs fuzzy membership and fuzzy entropy to help the labelling of the unlabeled data.

I. INTRODUCTION

For supervised machine learning applications, training data have to be first collected and labelled to train the learning algorithms. It is relatively easy to collect a great amount of data. However, the labelling of the collected data has to be done manually by domain experts. It is well-known that manual labelling is a slow, expensive, and error-prone process [5]. As the result, very often, a data set consists of only a very small portion of labelled data, with the rest of unlabeled ones. Semi-supervised learning has the ability to deal with the situation with improved generalization performance.

Support Vector Machine (SVM) [8], [3] becomes the most popular state-of-the-art classifier. SVM has its root from Statistical Learning Theory [16]. It is proven that the optimal hyperplane with the maximal margin maximizes the generalization ability of the linear classifier [16], [17]. The original SVM requires fully labelled data sets. The idea of applying a semi-supervised learning paradigm to SVM was first introduced by Vapnik and Sterin in 1977 [18]. However, it really attracted much attention from the machine-learning community only after the work of Joachims [11]. So far, there have been many studies on semi-supervised SVM [2], [4], [6], [7], [9], [10], [13], [19].

SVM has been proven very successful for balanced data sets, where the numbers of the data samples for both normal and abnormal classes are roughly the same, and also the sample data for both classes are all properly labelled. However, SVM cannot render the same good performance for imbalanced data sets, where one of two classes is undersampled, or only data samples of one class are available [12]. In real world problems, very often, only imbalanced data sets are available. To accommodate this situation, Support Vector Data Description (SVDD) was introduced in [14], [15] as a kernel method to deal with imbalanced data sets. SVDD aims at constructing an optimal hypersphere in the feature space to include only normal data points and exclude all abnormal data points with tolerances. This optimal hypersphere is regarded as a data description, since when mapped back to the input space, it becomes a set of contours that tightly enclose the normal data samples [1]. SVDD also requires fully labelled data sets, which, as discussed before, are rarely possible in real life application domains.

In this paper, we present a novel method, Fuzzy Entropy Semi-supervised SVDD (FS3VDD), to apply a semi-supervised learning paradigm to SVDD so that it can cope with imlabelled data sets, yet with improved generalization performance. Our proposed method introduces fuzzy membership for each unlabeled data point. The membership stands for its belonging degree to the normal class. At the beginning, the fuzzy membership of each data point is randomly initialized, and the fuzzy partition is vague, and then fuzzy entropy is used to boost the purity of the fuzzy partition, when the temperature variable $T$ is approaching $0$. Eventually, the fuzzy partition converges to a crisp partition, which effectively results the labelling of these unlabeled data.

II. SUPPORT VECTOR DATA DESCRIPTION (SVDD)

SVDD [14], [15] aims at determining an optimal hypersphere to include the normal data points while exclude the abnormal data ones with tolerances. The optimization problem is as follows:

$$\min_{R,c,\xi} \left( R^2 + C \sum_{i=1}^{n} \xi_i \right)$$

subject to

$$\|\phi(x_i) - c\|^2 \leq R^2 + \xi_i, \quad i = 1, \ldots, p$$
$$\|\phi(x_i) - c\|^2 \geq R^2 - \xi_i, \quad i = p + 1, \ldots, n$$
$$\xi_i \geq 0, \quad i = 1, \ldots, n$$

where $R, c$ are the radius and the centre of the hypersphere respectively, $C$ is a constant, $\xi = [\xi_i]_{i=1,\ldots,n}$ is the vector of slack variables, $\phi(\cdot)$ is a transformation from the input space to the feature space, and $p$ is the number of normal data points.

To classify an unknown data point $x$, the following decision function is used: $f(x) = \text{sign}(R^2 - ||\phi(x) - c||^2)$. The data point $x$ is normal if $f(x) = +1$ or abnormal if $f(x) = -1$. 

Trung Le, Tien Tran, and Khanh Nguyen are with Faculty of Information Technology, HCMC University of Pedagogy, Hochiminh city, Vietnam (email: {trunling, tinntt, and khanhndk}@hcmup.edu.vn).

Dat Tran and Wanli Ma are with Faculty of Education, Science, Technology and Mathematics, University of Canberra, Australia (email: {dat.tran, wanli.ma}@canberra.edu.au).
III. Fuzzy Entropy Semi-supervised Support Vector Data Description

A. Problem Statement

Given a training set \( X = X_l \cup X_u \) where \( X_l = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\} \) and \( X_u = \{x_{l+1}, \ldots, x_n\} \), where \( n = l + u \), \( X_l \) are the labelled data, and \( X_u \) are unlabelled. The problem now is to use the unlabelled training data to enhance the generalization performance of the classifier, i.e. the optimal radius of the fully labelled data, which automatically obtained from the aforementioned labelling process through fuzzy membership, must be minimized. To avoid any bias in the label assignment, it is reasonable to assume the following constraint:

The portion of normal data in the unlabelled training set coincides with that in the labelled training set.

B. Optimization Problem

Now, we are dealing with the following optimization problem (OP):

\[
\min_{R, c, Y_u} \left( R^2 + C \sum_{i=1}^{l} V(\alpha_i, y_i) + C' \sum_{i=l+1}^{n} V(\alpha_i, y_i) \right)
\]

(3)

where \( Y_u = \{y_{l+1}, \ldots, y_n\} \) is a label assignment, the loss function \( \alpha_i = \|\phi(x_i) - c\|^2 - R^2 \) is defined as \( V(\alpha_i, y_i) = \max\{0, y_i\alpha_i\} \), and \( C \) and \( C' \) are two parameters which stand for the trade-offs between the empirical losses of the labeled and unlabeled data and the general loss, respectively.

The above OP means that we need to find out the optimal label assignment \( Y_u \) such that the optimal radius for the whole data set, i.e. \( X = X_l \cup X_u \), is minimized.

C. Solution

For each unlabeled sample \( x_i \) (\( l \leq i \leq n \)), we introduce fuzzy membership \( u_i \), which is the possibility of \( x_i \), belonging to the normal class (\( y_i = 1 \)).

Given temperature \( T > 0 \), regarding sample \( x_i \), we need to minimize the following extended loss function:

\[
EV(\alpha_i, u_i) = AV(\alpha_i, u_i) - TS(u_i)
\]

(4)

In the above extended loss function, we employ the entropy to encourage the purity of fuzzy partition. The reason is that to minimize \( EV(\alpha_i, u_i) \), when \( T \) becomes smaller, i.e. \( T \rightarrow 0 \), \( S(u_i) \) is encouraged to be smaller which also means purer fuzzy partition. The temperature variable \( T \) is regarded as a trade-off parameter, which controls the trade-off between the average loss value and the purity of the fuzzy partition.

In our experiments, the temperature \( T \) is led to approach 0.

The extended OP is of the following form (referring to Table I):

\[
\min_{R, c, U} \left( \frac{1}{u} \sum_{i=l+1}^{n} u_i \right) \sum_{i=1}^{l} V(\alpha_i, y_i) +
\sum_{i=l+1}^{n} (u_i V(\alpha_i, 1) + (1 - u_i)V(\alpha_i, 1))
\]

\[
+ C' \sum_{i=l+1}^{n} (T u_i \ln u_i + T(1 - u_i) \ln(1 - u_i))
\]

(5)

As discussed previously, the constraint regarding the ratio of normal data in \( X_u \) is the same as in \( X_l \):

\[
\frac{1}{u} \sum_{i=l+1}^{n} u_i = \frac{l}{l+u} \sum_{i=1}^{l} \max\{0, y_i\} = r
\]

(6)

To solve the above OP, the fuzzy membership array \( U \) and the optimal hypersphere are alternatively kept fixed in iterations, until the temperature variable \( T \) is driven to approach 0.

1) Keep \( U \) fixed: The OP becomes:

\[
\min_{R, c} \left( R^2 + C \sum_{i=1}^{l} V(\alpha_i, y_i) +
C' \sum_{i=l+1}^{n} (u_i V(\alpha_i, 1) + (1 - u_i)V(\alpha_i, 1)) \right)
\]

(7)

which actually is the standard SVDD. Actually, it can be transformed to yet another equivalent OP:

\[
\min_{R, c} \left( R^2 + C \sum_{i=1}^{l} \xi_i + C' \sum_{i=l+1}^{n} \xi_i + C' \sum_{i=l+1}^{n} (1 - u_i) \xi_i \right)
\]

(8)

subject to

\[
y_i \left( \|\phi(x_i) - c\|^2 - R^2 \right) \leq \xi_i; \xi_i \geq 0, i = 1, \ldots, l
\]

\[
\|\phi(x_i) - c\|^2 - R^2 \leq \xi_i; \xi_i \geq 0, i = l + 1, \ldots, n
\]

(9)

The Lagrange function is of:

\[
L \left( R, c, \xi_i, \xi_i, \alpha_i, \alpha_i', \beta_i, \beta_i', \xi_i \right) = R^2 + C \sum_{i=1}^{l} \xi_i + C' \sum_{i=l+1}^{n} \xi_i
\]

\[
+ C' \sum_{i=l+1}^{n} (1 - u_i) \xi_i + \sum_{i=1}^{l} \alpha_i \left( y_i \left( \|\phi(x_i) - c\|^2 - R^2 \right) - \xi_i \right)
\]

\[
+ \sum_{i=l+1}^{n} \alpha_i \left( \|\phi(x_i) - c\|^2 - R^2 - \xi_i \right) - \sum_{i=1}^{l} \beta_i \xi_i
\]

\[
- \sum_{i=l+1}^{n} \left( \|\phi(x_i) - c\|^2 - R^2 - \xi_i \right) - \sum_{i=l+1}^{n} \beta_i \xi_i
\]

(10)

TABLE I

The expressions of the average loss and the fuzzy entropy according to the fuzzy membership \( u_i \).

<table>
<thead>
<tr>
<th>Fuzzy membership</th>
<th>( x_i \in \text{normal class or } y_i = 1 )</th>
<th>( y_i = -1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average loss</td>
<td>( A V(\alpha_i, u_i) = u_i V(\alpha_i, 1) + (1 - u_i)V(\alpha_i, -1) )</td>
<td></td>
</tr>
<tr>
<td>Fuzzy entropy</td>
<td>( S(u_i) = -u_i ln(u_i) - (1 - u_i) ln(1 - u_i) )</td>
<td></td>
</tr>
</tbody>
</table>
Setting the partial derivatives to 0, we have:

\[
\frac{\partial L}{\partial R} = 0 \Rightarrow \sum_{i=1}^{l} y_i \alpha_i + \sum_{i=l+1}^{n} \alpha_i - \sum_{i=l+1}^{n} \alpha_i' = 1
\]

\[
\frac{\partial L}{\partial c} = 0 \Rightarrow c = \sum_{i=1}^{l} y_i \alpha_i \phi(x_i) + \sum_{i=l+1}^{n} \alpha_i \phi(x_i) - \sum_{i=l+1}^{n} \alpha_i' \phi(x_i)
\]

Furthermore, we can evaluate \( V_i - V_i' \) as follows:

\[
V_i - V_i' = V (\alpha_i, 1) - V (\alpha_i, -1) = \max \{ 0, \alpha_i \} - \max \{ 0, -\alpha_i \} = \alpha_i
\]

Substituting the above equation to Equation (17), we can evaluate \( u_i \) as follows:

\[
u_i = \frac{1}{e^{\frac{V_i - V_i'}{\lambda}} + 1} = \frac{1}{e^{\frac{\alpha_i}{\lambda}} + 1}
\]

To determine \( \lambda \), we use the constraint:

\[
\sum_{i=1}^{n} u_i = \sum_{i=1}^{n} \frac{1}{e^{\frac{\alpha_i}{\lambda}} + 1} = ur
\]

We define and investigate the following function:

\[
f (\lambda) = \sum_{i=1}^{n} \frac{1}{e^{\frac{\alpha_i}{\lambda}} + 1}
\]

The derivative of the above function is:

\[
f' (\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} e^{\frac{-\alpha_i}{\lambda}} > 0
\]

Please note that the function \( f (\lambda) \) is strictly increasing. Moreover, we have:

\[
\lim_{\lambda \to -\infty} f (\lambda) = \lim_{\lambda \to -\infty} \sum_{i=1}^{n} \frac{1}{e^{\frac{\alpha_i}{\lambda}} + 1} = u
\]

\[
\lim_{\lambda \to +\infty} f (\lambda) = \lim_{\lambda \to +\infty} \sum_{i=1}^{n} \frac{1}{e^{\frac{\alpha_i}{\lambda}} + 1} = 0
\]

which means that the equation has a unique solution \( \lambda_0 \). To find out \( \lambda_0 \), we employ the Newton-Raphson method.

The rule for updating the fuzzy membership \( u_i \) becomes:

\[
u_i = \frac{1}{e^{\frac{\alpha_i}{\lambda_0}} + 1}, \; i = l+1, \ldots, n
\]

D. The Overall Algorithm

We start with \( T = 10 \). For each \( T \), we solve the OP in (5) by alternately keeping \( R \) and \( c \) fixed and then \( U \) fixed. The KL-divergence is used as the terminating criterion for each iteration. To achieve the local minimiser attained for each \( T \) and also the global minimiser, \( T \) is led to approach 0. The detail of this algorithm is as follows:

Initialize

\[ T = 10, \; \varepsilon = 0.0001, \; U = (r, r, \ldots, r) \]

Execute

while \((T > \varepsilon)\{
\]

Keep U fixed

Calculate \( R, c, \) and \( \alpha_i = \| \phi(x_i) - c \|^2 - R^2 \)

Keep R and c fixed

\[ V = U \]

Calculate fuzzy partition array \( U \)

\[ T = \frac{T}{10} \]

\}

where \( D_{KL} (U, V) > \varepsilon \)

where \( D_{KL} (U, V) = \sum_{i=1}^{n} u_i \ln \left( \frac{u_i}{v_i} \right) \)
TABLE II

NUMBER OF SAMPLES IN 6 DATA SETS. #POSITIVE: NUMBER OF POSITIVE SAMPLES, #NEGATIVE: NUMBER OF NEGATIVE SAMPLES AND d: DIMENSION.

<table>
<thead>
<tr>
<th>Data set</th>
<th>#positive</th>
<th>#negative</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astroparticle</td>
<td>2000</td>
<td>1089</td>
<td>4</td>
</tr>
<tr>
<td>Australian</td>
<td>383</td>
<td>307</td>
<td>14</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>444</td>
<td>239</td>
<td>10</td>
</tr>
<tr>
<td>Fourclass</td>
<td>307</td>
<td>555</td>
<td>2</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>255</td>
<td>126</td>
<td>34</td>
</tr>
<tr>
<td>Liver Disorders</td>
<td>200</td>
<td>145</td>
<td>6</td>
</tr>
</tbody>
</table>

IV. EXPERIMENTS

A. Data Sets

We conducted our experiments on 6 data sets of the benchmark UCI repository. The details of the data sets are in Table II. They are fully labelled data sets. We deliberately chose these fully labelled data sets, instead of using partially labelled data, so that we could make a comparison of our proposal with the normal SVDD. To make them partially labelled, we needed to preprocess these data sets. First, we randomly removed the negative (abnormal) data points from a data set so that the ratio of the number of positive data to that of the negative data is 10 : 1 (the full training set). Second, we randomly chose the unlabeled data from the current data set such that the ratio of the unlabeled and labelled data is 4 : 1. The unlabeled data were dominant because we wanted to simulate the real world imbalanced data problems.

B. Parameter Settings and the Results

To evaluate accuracies, cross validation with 3 folds were used. The popular RBF kernel function \( K(x, x') = e^{-\gamma ||x-x'||^2} \) was used. The parameter \( \gamma \) was ranged in the grid \( \{2^k : k = 2l + 1, l = -8, -6, \ldots, 1\} \). The trade-off parameter \( C \) was searched in the grid \( \{2^k : k = 2l + 1, l = -8, -6, \ldots, 2\} \). Another trade-off parameter \( C' \) was set to \( \lambda C \), where \( \lambda \) was searched in the grid \( \{0.25, 0.5, 1, 2, 4\} \). To evaluate the classification rate, we employed the accuracy metric given by \( acc = \frac{acc^+ + acc^-}{2} \), where \( acc^+ \) and \( acc^- \) are the accuracies on positive (normal) and negative (abnormal) classes, respectively.

We made the comparison of our proposed model Fuzzy Entropy Semi-supervised SVDD (FS3VDD) with SVDD in two cases: 1) SVDD for the fully labelled data set (Full-SVDD) and 2) SVDD for the partially labelled data set (Partial-SVDD). The fold division of the cross validation process is displayed in Figure 1, where the folds in red are the unlabeled folds, and the folds in green are the labelled folds. For Partial-SVDD, the classifier is trained in the labelled folds and is tested in the remaining full fold.

The experimental results are in Table III and Figure 2. It can be seen that FS3VDD is comparable with Full-SVDD with fully labelled data sets. However, the performance of Partial-SVDD is far worse both FS3VDD and Full-SVDD. The reason is that only a small collection of the full data set is used for training and thereby cannot characterize well the fully independent testing fold.

![Figure 1. The folds in cross validation.](image)

![Figure 2. Classification results (in %) on 6 data sets for FS3VDD, Full-SVDD, and Partial-SVDD.](image)

TABLE III

CLASSIFICATION RESULTS (IN %) ON 6 DATA SETS FOR FS3VDD, FULL-SVDD, AND PARTIAL-SVDD.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Partial-SVDD</th>
<th>Full-SVDD</th>
<th>FS3VDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Astroparticle</td>
<td>60%</td>
<td>70%</td>
<td>70%</td>
</tr>
<tr>
<td>Australian</td>
<td>71%</td>
<td>82%</td>
<td>81%</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>86%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>Fourclass</td>
<td>75%</td>
<td>93%</td>
<td>93%</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>65%</td>
<td>88%</td>
<td>86%</td>
</tr>
<tr>
<td>Liver Disorders</td>
<td>51%</td>
<td>63%</td>
<td>63%</td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this paper, we have proposed a novel method on how to apply a semi-supervised learning paradigm to SVDD. The experiments conducted on the 6 data sets in the benchmark UCI repository have demonstrated that our proposed method FS3VDD completely outperforms the partial SVDD and is comparable with the full SVDD.

REFERENCES


