∫(X)DATA

基于SVDD算法的半监督风控模型

文|索信达 杨健颖

导读

本文主要是介绍一种半监督算法——SVDD：该算法是传统SVM的一种扩展，能用于异常值（奇异值）检测和极度不平衡数据的分类问题。该算法根据样本中有标记的这类样本的特性构建一个球面作为决策边界，使用该球面来判断剩下无标记样本的属性，无标记样本能落在球内则判断为正样本，否则为负样本。在金融风控领域，我们常常遇到一类样本有标记，另一类样本无标记的数据，我们可以通过有标记样本来构建该模型，再使用该模型来识别剩下的无标记样本是否有风险。

一、引言

SVDD（Support Vector Domain Description，支持向量域描述算法）是基于传统 SVM 的一种半监督算法。SVDD 会根据样本中有标记的这类样本构建一个球面作为决策边界，球面边界上的点称为支持向量。我们可以计算那些没有标记的样本与球心的距离，若距离小于球的半径，即该无标记样本落在球内，则该样本与有标记的样本属于同一类；否则不是同一类。该算法适用于异常值（奇异值）检测和极度不平衡数据的分类问题。

严格来说，SVDD 适用于奇异值（novelty）检测，而非异常值（outlier）检测。这里使用 sklearn 官网上的定义来区分奇异值和异常值：

- 异常值检测：训练数据中含有离群点，我们希望通过算法找到训练数据最集中的区域，忽略偏差观测值。
- 奇异值检测：训练数据中没有离群点，我们的目标是用训练好的模型去检测一个新的观测值是否异常。在这种情况下，离群点也称为奇异点。

简单理解就是，训练模型时所用的数据中可以包括异常值，这属于异常值检测，适用的模型往往是无监督的，如 iForest 和 KNN；训练模型时所用的数据中不能包括异常值，这属于奇异值检测，适用的模型往往是半监督的，如 SVDD 和 OCSVM。

在金融业的风控领域，我们常常遇到一类样本有标记，另一类样本无标记的数据，即我们能确定某些样本有风险（或者没有风险），而不能确定剩余样本是否有风险。在这样的业务场景中，我们可以通过有标记样本（标记为有风险的样本或没有风险的样本）来构建 SVDD 模型，再使用该模型来识别剩下的无标记样本是否有风险，即仅通过有标记的样本数据建立 SVDD 模型来识别其它样本是否有风险。SVDD 这种能进行奇异值检测和处理极度不平衡数据分类问题的半监督模型非常适用于金融风控的领域。

二、理论知识

1. 模型推导

对于一个包含 N 个样本的数据集，SVDD 致力于找到一个能包含所有正样本的最小球面，设球面半径为 R，球心为 a。对于极少部分离球心较远的正样本而言，如果我们建立的球面想把这部分距离很远的正样本也包含进来的话，那半径势必会很大，这个半径很大的球并不能很好的代表正样本数据的的特性。所以我们这里引入 SVM 软间隔中松弛变量 ξ 的知识：即在建立分隔平面的时候，允许一些正样本分错，这些正样本到球心的距离可以超过 R。假设有 n 个样本，从而我们寻找的最小球面可以描述为：

$$F(R,a,\xi_i) = R^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

ξ_i 是松弛变量，不同样本对应的 ξ_i 的取值是不一样的，允许分错的正样本对应的 $\xi_i > 0$ 此时 $F > R^2$ 分类正确的正样本对应的 $\xi_i = 0$，此时 $F \le R^2$ 。C 为惩罚参数，当距离很远的点对应的松弛变量之和 $\sum_i \xi_i$ 固定时，C 增大表示对错分的惩罚增加，此时我们更加重视错分的损失，不愿意放弃这些距离远的点，算法会偏向于寻找包含更多样本点的球；反之，C 减小表示对错分的惩罚减少。我们的目标是要找一个半径尽可能小而错分的点也尽可能少的球面。

样本点到球心的距离将满足如下约束：

$$(x_i - a)^T(x_i - a) \le R^2 + \xi_i, \xi_i \ge 0 \tag{2}$$

SVDD要做的事就是在约束条件（2）下寻找最小的（1）。从而我们可以构造如下拉格朗日函数：

$$L(R,a,\alpha_i,\xi_i) = R^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[R^2 + \xi_i - (x_i^2 - 2ax_i + a^2)] - \sum_{i=1}^{n}\gamma_i\xi_i \tag{3}$$
$$\alpha_i \ge 0, \gamma_i \ge 0$$

其中，α_i和γ_i是拉格朗日乘子。此时我们希望最小化L，L分别对R、α_i、ξ_i求偏导，且偏导为0，于是有：

$$\sum \alpha_i = 1, \sum_i \alpha_i x_i = a, C - \alpha_i - \gamma_i = 0 \tag{4}$$

将（4）代入（3）消去γ_i，并引入SVM中核函数的知识，可以将（3）化简为：

$$L = \sum_i \alpha_i K(x_i, x_j) - \sum_{i,j}\alpha_i\alpha_j K(x_i, x_j) \tag{5}$$
$$0 \le \alpha_i \le C, \sum_i \alpha_i = 1, i \ne j$$

此时我们希望最大化L。$K(x_i \cdot x_j)$表示原本$(x_i \cdot x_j)$的核函数，SVDD的作者在论文中根据模拟实验的结果，建议使用高斯核函数，高斯核函数的公式为：

$$K_G(x_i, x_j) = e^{\frac{-(x_i - x_j)^2}{s^2}} \tag{6}$$

对于没有标记的样本（设为z）而言，它们与球心的距离为：

$$D = (z - a)^T(z - a) \tag{7}$$

将$\sum_i \alpha_i x_i = a$代入（7），并引入核函数的知识，我们有：

$$D = K(z,z) - 2\sum_i \alpha_i K(z, x_i) - \sum_{ij}\alpha_i\alpha_j K(x_i, x_j) \tag{8}$$

当$D > R^2$时，该样本落在球外，属于负样本；当$D \le R^2$时，该样本落在球内，属于正样本。

2.细节讨论（选读）

SVDD中的C是一个非常重要的参数，我们这里对C再做一些思考（SVDD的基本原理上一部分已经介绍完了，这部分是一些细节思考，不感兴趣的小伙伴可以跳过，直接看下一部分）。根据KKT条件（周志华老师的《机器学习》附录B有KKT条件的介绍和相关推导证明，这里不做KKT条件原理的细说），我们有：

$$\gamma_i \xi_i = 0 \tag{9}$$

$$\alpha_i [R^2 + \xi_i - (x_i^2 - 2ax_i + a^2)] = 0 \tag{10}$$

且根据（4）和（5），我们有：

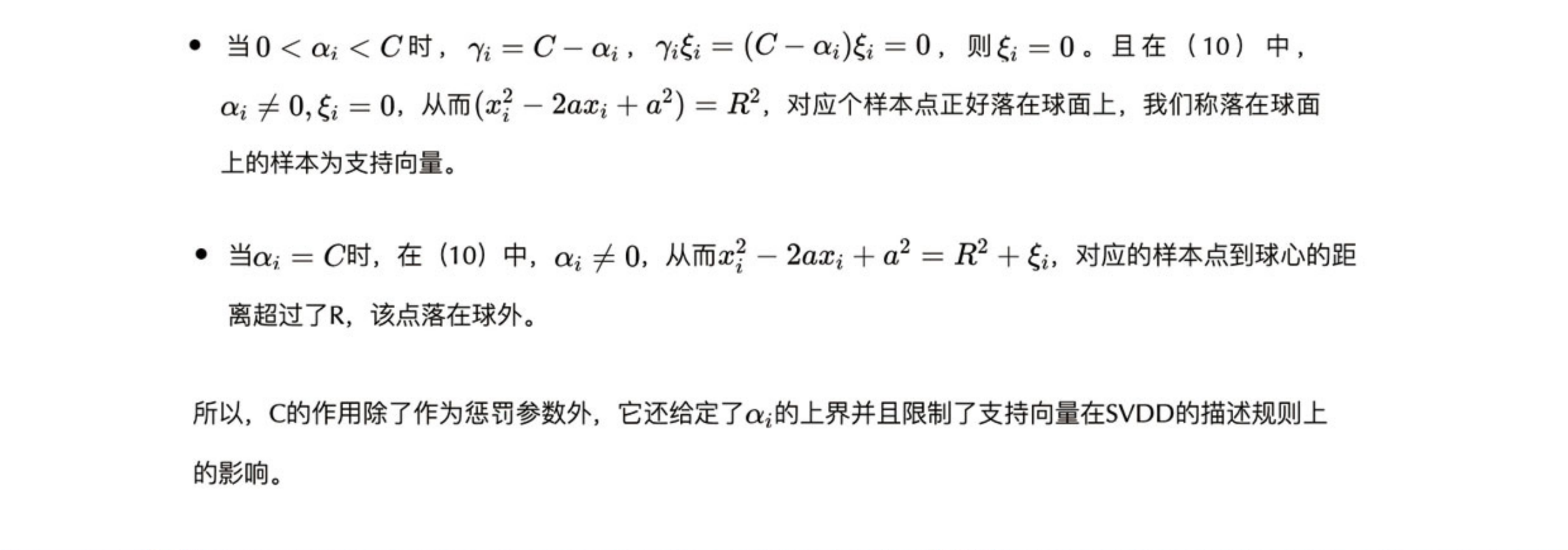
$$C - \alpha_i - \gamma_i = 0 \tag{11}$$

$$0 \leq \alpha_i \leq C \tag{12}$$

根据（9）、（10）、（11）和（12），我们有：

- 当$\alpha_i = 0$时，$C = \gamma_i$，$\gamma_i \xi_i = C\xi_i = 0$，则$\xi_i = 0$，对应的样本点落在球内。
- 当$0 < \alpha_i < C$时，$\gamma_i = C - \alpha_i$，$\gamma_i \xi_i = (C - \alpha_i)\xi_i = 0$，则$\xi_i = 0$。且在（10）中，$\alpha_i \neq 0, \xi_i = 0$，从而$(x_i^2 - 2ax_i + a^2) = R^2$，对应个样本点正好落在球面上，我们称落在球面上的样本为支持向量。
- 当$\alpha_i = C$时，在（10）中，$\alpha_i \neq 0$，从而$x_i^2 - 2ax_i + a^2 = R^2 + \xi_i$，对应的样本点到球心的距离超过了R，该点落在球外。

所以，C的作用除了作为惩罚参数外，它还给定了α_i的上界并且限制了支持向量在SVDD的描述规则上的影响。



三、应用分析

在应用 SVDD 算法时，非常重要的一点就是，SVDD 对异常点非常敏感，训练模型时用到的样本中只能包含同一类样本。因为如果训练模型时用到另一类的样本，那训练出来的 SVDD 球面也将会匹配另一类样本的特点，这样训练出来的 SVDD 球面将包含两类样本的特点，那么该球面就难以精确判断剩下的样本应落在球内还是球外。下面我们结合金融风控领域的业务场景和数据特点，对该算法在异常值检测和极度不平衡数据分类问题的处理思路上做一些介绍。

1. 异常值（奇异值 ）检测

其实，异常值检测也是金融风控领域的一类很有用的算法，由于有风险的样本一般在所有样本中占比很少，它们在一些指标上也往往不同于正常样本，我们可以将有风险的样本视作异常点，从异常值检测的角度来识别这些有风险的样本。使用 SVDD 进行异常值（奇异值）检测时，通常我们将确定没有风险的样本用于训练模型，构造球面作为决策边界，再用训练出来的模型识别剩余的点是否落在球内：落在球内的点认为不是异常点，对应的样本没有风险；落在球外的点是异常点，对应的样本有风险。

2. 极度不平衡数据的分类问题

对于极度不平衡的数据，很难直接建立一个精度高的二分类模型，这时我们可以转变思路，从异常值检测的角度入手，将极度不平衡数据中的那小部分数据视作异常值，用剩余的大部分数据训练一个 SVDD 超球面作为决策边界，从而将极度不平衡数据的分类问题转变为 SVDD 中的异常点检测问题。在金融风控领域，有风险的样本很少，大部分样本没有风险，我们可以将有风险的样本视作异常值，从异常值检测的角度建立金融风控的 SVDD 模型，从而使用 SVDD 算法解决极度不均衡数据的分类问题。

目前 python 的 libsvm 库和 slearnk 库中都已包含了该算法，可以直接调用。

参考资料

[1] Aggarwal, Charu C. Outlier Analysis, 2nd ed[M]. Berlin, Germany:Springer. 2016.

[2] Tax D M J , Duin R P W . Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11-13):1191-1199.

[3] Schölkopf B , Platt J C , Shawe-Taylor J , et al. Estimating the Support of a High-Dimensional Distribution[J]. Neural Computation, 2001, 13(7):1443-1471.

[4] 周志华 . 机器学习 [M] 北京：清华大学出版社，2016