

一种孤立点挖掘的混合核方法

田 江 顾 宏

(大连理工大学电子与信息工程学院 大连 116023)

摘 要 孤立点是不具备数据一般特性的数据对象。支持向量机(SVM)将数据点映射到高维特征空间,通过划分最大间隔的超平面来分离孤立点和正常点。利用支持向量机在处理小样本、高维数及泛化性能强等方面的优势,提出了一种新的基于高斯过程潜变量模型(GPLVM)和支持向量分类的检测模型算法。利用 GPLVM 提供潜变量到数据空间的平滑概率映射实现数据降维,然后通过 SVM 交叉验证进行孤立点检测。在 KDD99 数据集上进行了仿真实验,数值结果表明该算法在保证低误报率的前提下能有效地提高检测率,证明了方法的有效性。

关键词 孤立点检测,支持向量机,数据降维,高斯过程潜变量模型

中图法分类号 TP181 **文献标识码** A

Hybrid Method for Outliers Detection Using GPLVM and SVM

TIAN Jiang GU Hong

(School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China)

Abstract Outliers are objects that do not comply with the general behavior of the data. SVM(support vector machine) finds the maximal margin hyperplane in feature space for the purpose of distinguishing the outliers from normal samples. Based on the high performance of SVMs in tackling small sample size, high dimension and its good generalization, we proposed a new method for outlier detection, which combines a novel unsupervised algorithm GPLVM(Gaussian process latent variable model) with standard SVM. GPLVM provides a smooth probabilistic mapping from latent to data space, embeds the dataset in a low-dimensional space which is used for cross validation of SVM. The proposed approach was applied to KDD99 benchmark problems, and the simulation results show its validity.

Keywords Outlier detection, Support vector machine, Dimensionality reduction, GPLVM

1 引言

孤立点检测用于发现不具备一般数据特性的数据,进而发现潜在的有用信息。它可以应用到很多领域,如信用卡欺诈检测、安全系统、医学诊断、网络入侵检测和检索等,是数据挖掘领域的一个重要研究方向^[1]。

孤立点检测和分类问题具有一定的相似性,前者的一个显著特征是数据中的绝大部分都不是孤立点,而且在很多情况下并不具备识别孤立点的先验知识。传统的孤立点检测方法主要包括基于统计分布的方法和基于距离的方法。统计学方法的一个主要不足在于绝大多数检验是针对单个属性的,而许多数据挖掘问题需要在多维空间中发现孤立点。另外统计学方法还要求关于数据集的参数知识,而许多情况下这些参数知识都是未知的^[1]。基于距离的方法要面临的一个困难是对非均匀结构化数据如何定义距离^[2],而基于核的方法通过将原始数据映射到高维特征空间进行处理^[3],能有效地克服这些不足。

由于其诸多的优良特性,近年来支持向量机^[4]引起了广泛的关注。利用支持向量机分类算法检测孤立点,对于线性

不可分的数据,支持向量机通过非线性映射将原始输入数据变换到高维空间,在新的空间内找到最大边缘超平面分离两类数据。但在实际应用中,原始数据的特征选择会影响到分类性能,而很多应用数据包含大量的特征,过高的特征维数使得训练及分类的速度变慢,同时会使分类的正确率下降,影响到算法的应用性能^[5,6]。高斯过程潜变量模型(GPLVM)^[7,8]由 Lawrence 于 2005 年提出,是一种新的无监督非线性数据降维方法,实现概率非线性的主成分分析,近年来被成功地应用于故障诊断、步态识别等领域^[9-11]。本文结合 SVM 和 GPLVM 这两种核方法,提出了一种新的孤立点检测模型算法。在 KDD99 标杆数据集上进行了相关的仿真实验,结果表明本文提出的方法能够有效地提高孤立点检测性能,在保证低误报率的前提下能够有效地提高检测率,证明了方法的有效性。

2 理论基础

高斯过程潜变量模型(GPLVM)^[7,8]是一个完全概率非线性的实现主成分分析的潜变量模型。对主成分分析进行概率解释,写成一组具有线性协方差方程的高斯过程的积形式。

到稿日期:2009-04-28 返修日期:2009-07-07 本文受国家自然科学基金(60605022)资助。

田 江(1970—),男,博士生,主要研究方向为数据挖掘等,E-mail:tianjiang@gmail.com;顾 宏(1961—),男,博士,教授,主要研究方向为数据挖掘等。

设 d 维观测数据集为:

$$Y = [y_{1,1}, \dots, y_{n,1}]^T = [y_{1,d}, \dots, y_{n,d}]^T \in \mathbb{R}^{n \times d}$$

q 维潜变量数据集为:

$$X = [x_{1,1}, \dots, x_{n,1}]^T = [x_{1,q}, \dots, x_{n,q}]^T \in \mathbb{R}^{n \times q}$$

概率 PCA^[12] 是一个简单的潜变量模型,给出了观测数据 y 和潜变量 x 之间的关系,它利用低维的潜变量来表示高维观测数据,具体描述的是线性关系:

$$y_{i,1} = Wx_{i,1} + \eta_{i,1} \quad (1)$$

其中, $W \in \mathbb{R}^{d \times q}$ 是映射矩阵,噪声向量 $\eta_{i,1}$ 服从于均值为 0、方差为 $\sigma^2 I$ 的正态分布,记为 $\eta_{i,1} \sim N(0, \sigma^2 I)$ 。将潜变量同观测数据联系起来,假定各数据相互独立,令条件概率为:

$$p(Y|X, W) = \prod_{i=1}^n N(y_{i,1} | Wx_{i,1}, \sigma^2 I) \quad (2)$$

定义潜变量的高斯先验分布为:

$$p(X) = \prod_{i=1}^n N(x_{i,1} | 0, I) \quad (3)$$

$$p(Y|W) = \prod_{i=1}^n N(y_{i,1} | 0, C) \quad (4)$$

其中, $C = WW^T + \sigma^2 I$ 为协方差,进而可推导出潜变量 x 关于观测变量 y 的后验概率分布为:

$$p(X|Y, W) = \frac{p(Y|X, W)p(X)}{p(Y|W)}$$

在该模型下的观测数据对数似然函数为:

$$\log p(Y|W) = -\frac{n}{2} \log |C| - \frac{1}{2} \text{tr}(C^{-1} Y^T Y) + c \quad (5)$$

其中, c 为常数,则将式(5)最大化可得:

$$W = U_q L V^T, L = (\Lambda_q - \sigma^2 I)^{\frac{1}{2}} \quad (6)$$

其中, U_q 是协方差矩阵 $n^{-1} Y^T Y$ 的前 q 个主特征向量,对应的特征值矩阵为 Λ_q , V 为任意 $q \times q$ 正交旋转矩阵。这样用最大似然参数估计代替了通常的样本协方差矩阵特征值分解, W 和 σ^2 可以用 EM 迭代算法高效地求解。在 GPLVM 中,对概率 PCA^[12] 进行了对偶分析,并将均值矩阵和协方差矩阵转换为高斯过程中定义的均值函数和协方差函数。将式(3)转换成对映射参数 W 的先验分布,定义映射参数 W 的高斯先验分布为 $p(W) = \prod_{i=1}^n N(w_{i,1} | 0, I)$, 并且有:

$$p(Y|X) = \prod_{i=1}^n N(y_{i,1} | 0, K)$$

其中, $K = XX^T + \sigma^2 I$, 进而可推导出映射参数 w 关于观测变量 y 的后验概率分布。在该模型下的观测数据对数似然函数为:

$$\log p(Y|X) = -\frac{d}{2} \log |K| - \frac{1}{2} \text{tr}(K^{-1} Y Y^T) + c \quad (7)$$

其中, c 为常数,则将式(7)最大化可得:

$$X = U'_q L V^T, L = (\Lambda_q - \sigma^2 I)^{\frac{1}{2}}$$

其中, U'_q 是协方差 $d^{-1} Y Y^T$ 的前 q 个主特征向量,对应的特征值为 Λ_q 。在推导过程中,用非线性核代替线性核 K ,即引入非线性协方差方程。如使用 RBF 核:

$$k(x_{i,1}, x_{j,1}) = \exp\left(-\frac{(x_{i,1} - x_{j,1})^T (x_{i,1} - x_{j,1})}{2\sigma^2}\right)$$

3 孤立点检测模型

3.1 检测模型

GPLVM 提供从潜变量到数据空间的平滑概率映射,与大多数非线性降维方法专注于保持数据空间原有局部距离不

同,GPLVM 侧重于在潜变量空间内分离原数据空间中距离较远的点。通过加以后向约束,GPLVM 同样实现了对原空间局部距离的保持。在反向约束 GPLVM 中,似然函数的优化需要加上对局部距离进行保持的约束,这个约束通过学习数据空间到潜变量空间的映射来实现。因此,在训练过程中同时会有两个模型并行工作:一个用于相异保持,从潜变量空间到数据空间的概率 GPLVM 映射;另外一个模型用于局部距离保持,从数据空间到潜变量空间的映射。

本文提出的检测算法分为两个步骤:首先使用 GPLVM 作为数据样本的降维方法,然后使用支持向量机(SVM)交叉验证对数据进行分类,进而检测孤立点。两个过程都使用了核方法来处理数据,在高维空间进行特征选取和分类操作,具体的检测模型如图 1 所示。

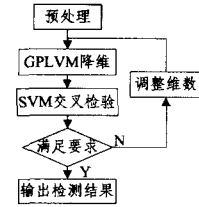


图 1 基于混合核方法的孤立点挖掘过程

3.2 算法详细描述

孤立点检测算法分为 5 个步骤,具体如下。

步骤 1 数据预处理,转换为实验需要的数据格式。

步骤 2 GPLVM 降维,设定初始目标维数为 2,进行数据降维操作。目标维数根据后面步骤的计算结果进行动态更新,以保证获得最佳检测结果。

步骤 3 使用 SVM 进行交叉验证,得到分类正确率等结果,作为下一步的评判依据。

步骤 4 判断分类结果是否满足预定要求。这里设定为是否超过原始数据的 SVM 校验检验的正确率。如果满足,则输出检测结果;如果不满足,则递增目标维数,进行降维操作,反馈循环。

步骤 5 输出孤立点检测结果。

4 实验与分析

4.1 仿真实验描述

对孤立点进行检测可以看作是一种机器学习的过程。机器学习中常用的算法评估标准有分类正确率(classification accuracy)、检测率(true positive rate,简称 TPR)和误报率(false positive rate,简称 FPR)。具体定义如下:

分类正确率 = 正确分类的样本数量 / 总的样本数量

检测率 = 正确检测的孤立点样本数量 / 总的孤立点样本数量

误报率 = 被错误判断为孤立点的正常样本数量 / 总的正常样本数量

实验使用了 FGPLVM^[13] 的 MATLAB 工具箱,在 WEKA^[14] 平台上通过广泛应用的 SVM 实现 LIBSVM^[15],完成 10 折交叉验证分类。为了验证本文方法的有效性,FGPLVM 使用了默认的参数设置:FITC 逼近,SCG 优化器;LIBSVM 也没有对参数进行优化配置,而是直接使用缺省的参数配置,具体如下:C-SVC;RBF 核函数;C=1;gamma=1/k(k 为维数数量)。

4.2 入侵检测实验结果

提取 KDDCUP99 数据集中属性 logged-in 为 1 的 703066 条记录构成测试集,其中攻击记录 3377 条。对测试集进行预处理,将多类问题的数据转换为正常记录和入侵记录的两类问题测试集。从测试集中随机抽取 10 组样本集,每组含有正常样本 1900 条、入侵攻击记录 100 条,两类数据比例为 19:1,构成本文实验的样本数据集。每条数据包含 41 个特征,其中 1—9 是数据包的基本特征、10—22 是与数据包内容相关的特征、23—31 属于网络交通特征、32—41 是主机相关的特征。经过初步实验,在表 1 中给出了分类正确率的对比情况。

表 1 分类正确率比较

编号	SVM		本文方法	
	特征	正确率	特征	特征
1	41	0.98	5	0.994
2	41	0.9955	11	0.9995
3	41	0.966	9	0.9985
4	41	0.9755	7	1
5	41	0.978	4	1
6	41	0.9925	3	1
7	41	0.9775	3	0.9995
8	41	0.983	4	1
9	41	0.9655	2	1
10	41	0.989	4	1

从表 1 中可以看出,利用本文方法对入侵数据进行检测,分类的正确率在 10 个数据集上均有显著提高,特征数也明显减少。第 4,5,6,8,9 和 10 组数据集的正确率提高到 100%,孤立点数据全部被检测处理,并且特征数降低为原来的 17% 到 5%。

在孤立点检测中,往往一类数据比另外一类更加重要,表 2 给出了在 TPR 和 FPR 两个评价标准上的对比。从表中可以看出本文方法在保证误报率为 0 的同时,能有效提高检测率。数据集 2 的孤立点检测率提高得最少,为 8 个百分点;数据集 9 的检测性能改进最为显著,提高了 69 个百分点;其他几个数据集的 TPR 也均有不同程度的提高。而传统 SVM 在数据集上的 TPR 比较低,并且具有很大的不稳定性;在 3 和 9 两组数据集上降低到 32% 和 31%;而本文方法不存在这个问题,在所有的数据上均表现良好。

表 2 TPR 和 FPR 的检测结果比较

编号	SVM		本文方法	
	TPR	FPR	TPR	FPR
1	0.6	0	0.88	0
2	0.91	0	0.99	0
3	0.32	0	0.97	0
4	0.51	0	1	0
5	0.56	0	1	0
6	0.85	0	1	0
7	0.55	0	0.99	0
8	0.66	0	1	0
9	0.31	0	1	0
10	0.78	0	1	0

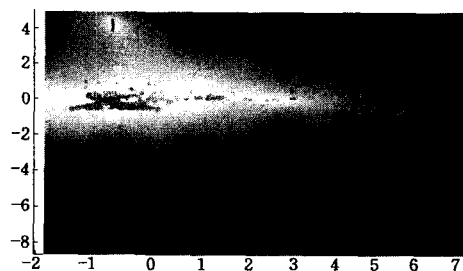


图 2 潜变量示意图

为了进一步说明降维效果,图 2 给出了数据集 9 的潜变量示意图,绿色样本(图中中间部分的数据,用○表示)为正常数据,红色的代表孤立点样本数据(图中上方数据,用×表示)。从中可以看出,经过高斯过程潜变量模型的有效降维,孤立点和正常样本在潜变量空间内显著分开,提高了下一步的孤立点检测性能。

结束语 利用支持向量机分类算法实现孤立点检测,数据过高的维度会对实际的检测性能造成影响。针对这个问题,本文将支持向量机同高斯过程潜变量模型相结合,提出了一种新的基于核方法的孤立点检测模型算法。在利用 GPLVM 进行特征选取的前提下,通过 SVM 交叉验证进行孤立点检测操作。实现后的仿真软件在 KDD99 数据集的实验结果表明,同标准的支持向量机算法相比,本文算法能够在整体上提高分类正确率。在检测率和误报率两个评价指标上,本文方法也表现出优良的性能。在保证低误报率的前提下,有效地提高了检测率。同时本文方法的检测效果更加稳定可靠,具有更好的应用性。

参考文献

- [1] Han J, Kamber M. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann, 2006
- [2] Petrovskiy M I. Outlier Detection Algorithms in Data Mining Systems[J]. Programming and Computer Software, 2003, 29 (4): 228
- [3] Smola A J. Learning with Kernels [M]. CambridgeMass: MIT Press, 2002
- [4] Vapnik V N. The Nature of Statistical Learning Theory[M]. Springer, 2000
- [5] Blanchard G, Zwald L. Finite-dimensional projection for classification and statistical learning[J]. IEEE Transactions on Information Theory, 2008, 54(9): 4169-4182
- [6] Bouchaffra D, Amira A. Structural hidden Markov models for biometrics; Fusion of face and fingerprint[J]. Pattern Recognition, 2008, 41(3): 852-867
- [7] Lawrence N D. Gaussian process latent variable models for visualization of high dimensional data[M]. Advances in Neural Information Processing Systems (NIPS) 16, Cambridge, MA: MIT Press, 2004
- [8] Lawrence N D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models[J]. Journal of Machine Learning Research, 2005, 6: 1783-1816
- [9] Eciolaza L, Alkarouri M, Lawrence N D, et al. Gaussian Process Latent Variable Models for Fault Detection[C]//Computational Intelligence and Data Mining. CIDM 2007. IEEE Symposium on. 2007: 287-292
- [10] Cheng M H, Ho M F, Huang C L. Gait analysis for human identification through manifold learning and HMM. [J]. Pattern Recognition, 2008, 41(8): 2541-2553
- [11] Quirion S, et al. Comparing GPLVM Approaches for Dimensionality Reduction in Character Animation[J]. Journal of Wscg, 2008, 16(1-3): 41-48
- [12] Tipping M E, Bishop C M. Probabilistic Principal Component Analysis[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 1999, 61(3): 611-622
- [13] www.cs.man.ac.uk/~neill/fgplvm/
- [14] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques[M]. Morgan Kaufmann, 2005
- [15] Chang C C, Lin C J. LIBSVM: a library for support vector machines[EB/OL]. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>