

doi: 10.12052/gdutxb.170008

基于正交投影的降维分类方法研究

滕少华, 卢东略, 霍颖翔, 张 巍

(广东工业大学 计算机学院, 广东 广州 510006)

摘要: 大数据时代要求数据挖掘算法能高效处理海量数据, 针对传统分类算法建模时间长、分类耗时久、算法难以理解等问题, 提出一种基于正交投影的降维分类方法, 通过投影方式将多维分类问题转化为多个二维投影面组合问题, 建立投影面的密度模型进行分类. 相比于常用的支持向量机(Support Vector Machine, SVM)、逻辑回归(Logistic Regression, LR)、k最近邻(K-Nearest Neighbor, KNN)等分类算法, 基于正交投影降维的分类方法能够在不丢失分类准确度的同时, 拥有较高的模型训练效率与分类效率. 其算法易于实现, 可用于实时性要求较高的应用场合, 如入侵检测、交通调度等.

关键词: 数据挖掘; 分类; 正交投影; 降维

中图分类号: TP391 文献标志码: A 文章编号: 1007-7162(2017)03-0001-07

Classification Method Based on Dimension Reduction

Teng Shao-hua, Lu Dong-lue, Huo Ying-xiang, Zhang Wei

(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Data mining algorithm in the era of big data needs to be able to efficiently deal with massive data. Traditional classification algorithms take a long time to train a model and classify the test dataset, and the algorithm is difficult to understand. To deal with the problems, a classification method based on dimension reduction is proposed in this paper. The multidimensional classification problem is transformed into multiple 2D projection surface combination by projection, and a density model of the projection surface is trained for classification. Compared with Support Vector Machines (SVM), Logistic Regression (LR), K-Nearest Neighbor (KNN) and other algorithms, the classification method based on dimension reduction has higher training efficiency and classification efficiency without loss of accuracy. The method is easy to implement, so it can be used for real-time application, such as intrusion detection and traffic scheduling.

Key words: data mining; classification; orthogonal projection; dimension reduction

数据挖掘技术如今已被广泛应用到实际生产, 然而分类算法大多复杂度很高, 模型训练与分类的时间效率低, 如支持向量机的训练计算复杂度在 $O(N^3 + L \cdot N^2 + d \cdot L \cdot N)$ 和 $O(d \cdot L^2)$ 之间 (其中 N 是支持向量的个数, L 是训练集样本的个数, d 是每个样本的维数), 该问题很大程度上限制了分类算法在时间要求较高与大数据情况下的应用, 如入侵检测问题^[1-8], 实时公车调度问题^[9-10]. 本文针对该问题提出一种基于降维的分类方法, 能够快速建模与分类, 同

时有较好的分类准确率.

1 相关工作

分类算法被广泛应用于实际生产中^[11-12], 但数据量的急速增长, 提高了对分类算法的效率的要求, 分类算法若没有很好的运行效率则无法适应生产发展的需要. 入侵检测算法的实时性要求较高, 在这个领域中已有许多方法针对分类方法的时间效率提出改进.

收稿日期: 2017-01-11

基金项目: 国家自然科学基金资助项目(61402118, 61673123); 广东省科技计划项目(2015B090901016, 2016B010108007); 广东省教育厅项目(粤教高函2015[133]号, 粤教高函〔2014〕97号); 广州市科技计划项目(201604020145, 2016201604030034, 201508010067)

作者简介: 滕少华(1962-), 男, 教授, 博士, 主要研究方向为协同计算、数据挖掘、网络安全、大数据.

通信作者: 张巍(1964-), 女, 副教授, 主要研究方向为协同计算、数据挖掘、网络安全、大数据. E-mail: weizhang@gdut.edu.cn

各种分类算法如决策树^[1-2], 贝叶斯^[3-4], KNN^[5], SVM^[6-8]已被改进并应用于入侵检测中. 在强调入侵检测准确率的同时, 为了在入侵发生的第一时间能做出识别, 降低损失, 许多学者在时间效率上做出研究, 文献[5]提出一种适应大样本集的网络入侵检测算法Cluster-KNN算法, 该算法先进行离线预处理建立大样本集的聚簇索引, 在线实时分类时利用该聚簇索引搜索得到近邻, 再采用KNN算法得出分类结果, 解决了KNN算法需要在分类过程中遍历数据集的问题, 提高了算法分类过程的速度, 让其能够应用于大数据集中. 但该算法在处理增量问题时, 必须重新进行聚簇索引的建立, 而K-means得到聚簇是一个非常耗时的过程. 文献[6]提出一种基于并行凸包分解计算和支持向量机的入侵检测分类算法(PCH-SVM), 该算法借助凸包的分解和并行计算快速提取训练样本空间几何凸包的顶点, 构建简约SVM训练样本集. 该算法能在不造成精度损失的前提下, 降低SVM训练的时空复杂度. 文献[7]提出了基于正区域快速属性约简算法, 形成SVM最优约简属性集. 但是这类算法仅是从SVM训练样本的角度进行优化, 未能根本解决SVM算法复杂度高的问题, 故时间效率

提高有限. 本文提出的基于正交投影的降维分类方法, 其分类准确率略低于SVM方法, 但其分类速度与训练速度远快于SVM方法, 能够高效地处理实时分类问题.

2 基于正交投影的降维分类方法

2.1 总体结构

本文从数据预处理, 模型训练与分类3个步骤介绍基于正交投影的降维分类方法, 预处理过程如图1(a)所示, 对于原始数据集需要对其进行归一化与区间离散映射, 这一过程是为了让数据变得规整均匀适合于投影至相同规格的矩阵中, 且适合以数组形式来将投影模型保存在计算机中. 本文将处理后的数据按一定比例随机划分为训练集与测试集, 对训练数据针对不同分类类别进行不同投影面的落点计数与模糊处理, 如图1(c)所示, 得到的是 $k \cdot C_n^2$ 个投影面密度模型, 其中 k 为分类类别数, n 表示数据维度数量, 不同维度属性组合的二维投影面数量为 C_n^2 . 得到该模型后, 用测试集对模型进行测试, 如图1(b), 得到分类的准确率与耗时, 分析分类结果, 我们可以对模型的参数进行调整, 如模糊过程中的模糊半径 σ , 映射区间范围大小range.

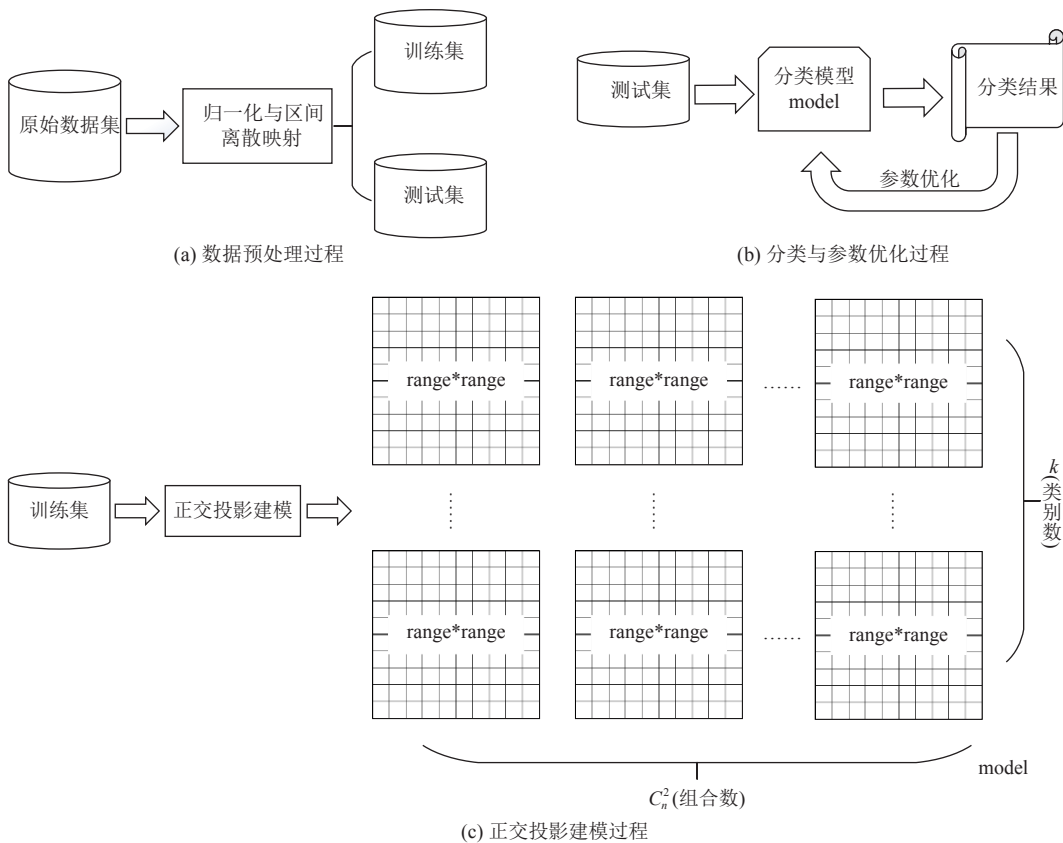


图 1 分类方法整体结构示意图
Fig.1 Schematic diagram of classification method

2.2 归一化与区间离散映射

基于正交投影的降维分类方法需要对数据进行归一化与区间离散映射处理。模型训练的时候需要将数据投影至各个二维平面保存在矩阵数组中,这一过程需要根据矩阵数组的下标进行映射,数组的下标是离散而非连续的,为了让数据能够投影至同样规格的矩阵中,在模型训练前,不论连续的实数数据还是标量类型的数据都需要进行归一化与区间的离散化处理。

对于连续属性,基于正交投影的降维分类方法需要将实数处理为正整数,通过数据规范化的方法将数据规范化至区间 $[0, 1]$,再进行映射取整。

假设 A 是数值属性,具有 n 个观测值 $v_1, v_2, \dots, v_n$,且 $\min_A$ 和 $\max_A$ 分别为属性 A 的最小值和最大值,则有

$$v'_i = \frac{v_i - \min_A}{\max_A - \min_A} \cdot \text{range}. \quad (1)$$

把 A 的值 v_i 离散化并映射为 $[0, \text{range}]$ 中的 v'_i ,其中的 $\text{range} \in \mathbb{Z}^+$ 为离散化后范围的最大值,能够反映离散化的精度,当 range 取值越大时,数据的精度越高,但需要的存储空间也越大。

为了让降维分类方法能够处理标量型的数据,同样需要对标量数据进行离散值的映射,假设 B 是标量属性,具有 n 个观测值 $b_1, b_2, \dots, b_n$,则有

$$b'_i = \frac{(i-1)}{n} \cdot \text{range}. \quad (2)$$

把 B 的值 b_i 映射为 $[0, \text{range}]$ 中的 b'_i . 需要注意的是标量数据的排列顺序应与其物理意义的相对大小对应,如当 B 表示属性为评价时, B 的排列顺序应为{差, 中, 良, 优},而非{中, 差, 优, 良}。

2.3 建立投影面密度模型

经过预处理,得到 $[0, \text{range}]$ 范围内规范的正整数数据,下面分二维数据建模与多维数据建模分别进行讨论。

2.3.1 二维数据建模

首先考虑二维一类情况,假设现有已预处理的二维数据集 $P = \{p_0, p_1, \dots, p_m\}$, $p_i = (x_i, y_i)$, $x_i \in [0, \text{range}]$, $y_i \in [0, \text{range}]$. P 表示整个数据集, p_i 表示第 i 行数据, x_i 表示 p_i 的第1维属性值, y_i 表示 p_i 的第二维属性值, m 表示数据量。建立一个二维浮点型数组 model ,其初始值均为0, model 的两个下标范围均为 $[0, \text{range}]$,即其下标范围与 P 中数据的取值范围大小一致,故可以在 model 中找到与 (x_i, y_i) 对应的下标。

遍历数据集 P ,对于每一个 $p_i = (x_i, y_i)$,找到

model 对应的下标,并让 model 对应下标的值自增。为了方便说明,本文将 model 矩阵表示为一个二维图像,如图2(a)是一个二维一类数据集生成的 model 示例。但图(a)仅能反映 model 中的某一位置有无数据,不能反映 model 中数据值的大小,即重叠的数据量,因此,本文通过灰度值来表示 model 中的数值的相对大小,如图2(b)所示。

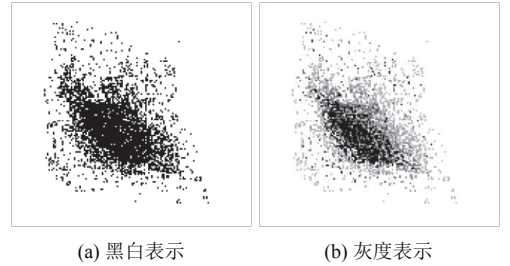
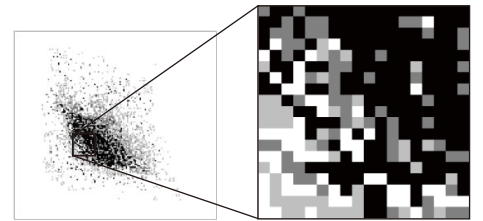


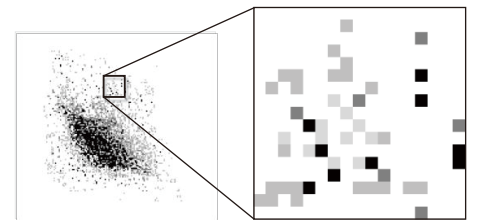
图2 二维一类model示例

Fig.2 Sample of two-dimensional one-class model

保存该模型,对于新的测试点,只需要查看 model 对应位置值的大小即可获得其分类信息。但该模型易受噪声干扰,由于训练样本是有限的,在原本高密度的区域可能会出现低密度的点,如图3(a),而在原本低密度的区域也可能出现高密度的点,如图3(b)所示。为了避免这种情况,需要对其进行高斯模糊处理。进行高斯模糊时需要选择合适的模糊半径,如图4所示,这样能够在保证保留数据分布规律的同时,降低噪声。若选择过小的模糊半径,去噪效果可能差,如图4(a)所示;若选择过度的模糊半径,则可能使数据的分布规律丢失,如图4(c)所示;图4(b)给出了合适的模糊半径。



(a) 高密度区域出现低密度点



(b) 低密度区域出现高密度点

图3 模型噪声干扰情况

Fig.3 Noise interference

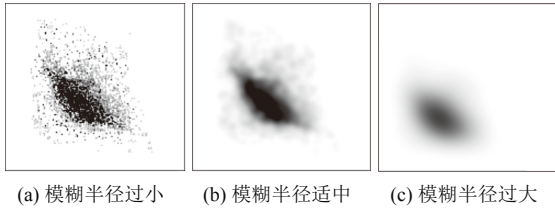


图4 模糊示意图

Fig.4 Fuzzy schematic diagram

上面讨论的情况类别数目唯一,即只能判断数据属于该类别的可能性,需要人工设置判定的阈值。

下面讨论二维多类数据的情况。先考虑类别为2的情况,假设现有已预处理的二维二类数据集 $P = \{p_0, p_1, \dots, p_m\}$, $p_i = (x_i, y_i, c_i)$, $x_i \in [0, \text{range}]$, $y_i \in [0, \text{range}]$, $c_i \in \{0, 1\}$ 。 P 表示整个数据集, p_i 表示第 i 行数据, x_i 表示 p_i 的第一维属性值, y_i 表示 p_i 的第二维属性值, c_i 表示 p_i 所属类别, m 表示数据量。即每行的数据 p_i 都增加类别属性 c_i ,此时可看作将数据集划分为两部分进行处理,即 $\alpha = \{p_i | i \in [0, m], c_i = 0\}$ 与 $\beta = \{p_i | i \in [0, m], c_i = 1\}$ 两部分,此时问题分解为求两个二维一类数据集的训练模型问题,采用前面介绍的方法,分别得到对应于不同类别的模型,如图5所示,图中以圆形表示第0类数据点,方形表示第1类数据点,则得到图1(a)和(b)两类的模型。

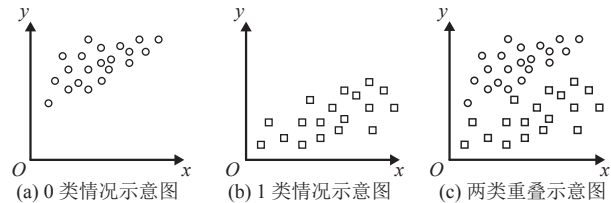


图5 二维二类情况示意图

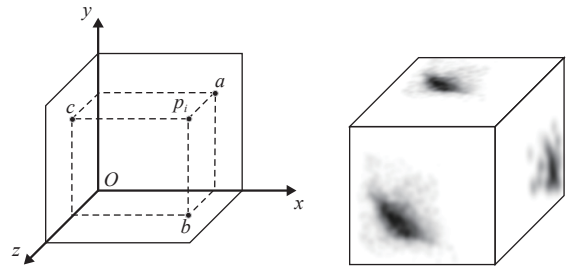
Fig.5 Sample of two-dimensional two-classes data

以相同方法进行扩展,推广到一般类别不唯一的情况,如假设此时的类别属性 c_i 取值为0至 $k-1$ 的自然数集,即此时类别数量为 k 个,那么针对这 k 个类别进行建模,即可得到 k 个表示不同类别的二维矩阵模型。

2.3.2 多维数据建模

下面将方法推广至多维的情况,先讨论类别数唯一的情况,假设现有已预处理的多维一类数据集 $P = \{p_0, p_1, \dots, p_m\}$, $p_i = (f_{i0}, f_{i1}, \dots, f_{in})$, $f_{ij} \in [0, \text{range}]$ 。 P 表示整个数据集, m 表示数据量, p_i 表示第 i 行数据, f_{ij} 表示 p_i 的第 j 维属性值,维度数量为 n 。此时无法像二维的情况一样将数据信息保存在一个二维矩阵中,本算法的做法是对多维属性进行投影,由于高维空间数据难以用图形示意,本文先以 $n=3$ 的情况为例说

明,如图6(a)所示,三维空间中的点 $p_i = (f_{i0}, f_{i1}, f_{i2})$,可以找到3个相互正交的坐标平面 xOy, xOz, yOz ,将点 p_i 向3个坐标构成的平面进行投影,可得 $a(f_{i0}, f_{i1})$, $b(f_{i0}, f_{i2})$, $c(f_{i1}, f_{i2})$ 3个投影点,如对于某三维数据点 $p_i = (30, 12, 6)$,对其向3个正交坐标轴组成的二维平面进行投影,分别得到3个坐标点 $(30, 12)$, $(30, 6)$, $(12, 6)$ 。这样就将三维空间的问题通过投影的方式转化为3个二维空间组合的问题。对于每一个投影的平面,采用与上述二维一类情况相同的做法,可以得到如图6(b)这样的模型。当 $n > 3$ 时,方法是一样的,根据上面的投影方式可以计算三维情况下投影面的数量为 $C_3^2 = 3$,那么一般情况下一个数据点需要投影的二维平面数量则为 C_n^2 ,维度数量的增加只是增加了投影的数量且相应地增加了模型占用的内存,投影后按二维一类情况处理即可。



(a) 三维数据点正交投影示意图 (b) 三维数据集正交投影建模示意图

图6 三维投影示意图

Fig.6 Projection of three-dimensional data

下面讨论多维数据类别不唯一、二维数据类别不唯一的情况,本文将问题转换为根据不同的类别建立多个二维一类模型,同理针对多维且类别不唯一的问题,本文将不同类别的数据分别建模,得到与类别对应的多维一类模型。图7以三维二类为例展示了各类别对应的各投影面中投影的数据点,图中 α 表示类别为0的数据集,数据量为 s , α_{ij} 表示数据集 α 的第 i 行数据的第 j 维属性, β 表示类别为1的数据集, β_{ij} 表示数据集 β 的第 i 行数据的第 j 维属性,数据量为 t 。投影后可得模型如图8所示。根据 α 数据集(第0类)与 β 数据集(第1类)分别进行建模,可得到 $2 \cdot C_3^2 = 6$ 个投影面,那么对于类别数量为 k ,维度数量为 n 的数据集,此时需要存储的投影数量为 $k \cdot C_n^2$ 。由于本文应用一个浮点型数组存储该模型,在通常的编程环境中,浮点型数占用4字节的内存空间,则该模型所占用的内存空间为 $4 \cdot \text{range}^2 \cdot k \cdot C_n^2$ 字节。举例说明,当要处理的数据集为13维度二类的情况,若采用 $\text{range}=100$ 处理数据集,则该模型需要占用的内存空间约为6 MB。

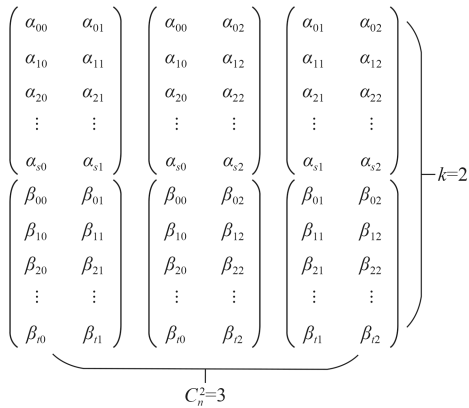


图7 三维两类各面投影数据

Fig.7 Three-dimensional two-classes projection data

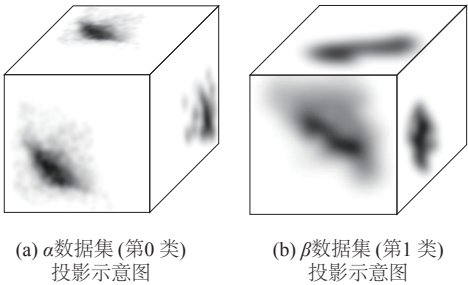


图8 三维两类投影示意图

Fig.8 Projection of three-dimensional two-classes data

2.4 快速高斯模糊算法

建模过程中,需要对投影统计后的数据进行快速高斯模糊,高斯模糊算法本质上是一种数据平滑(Data Smoothing)算法,被广泛应用于信号处理领域.

一维高斯函数为

G(x)=1/(sqrt(2*pi)*sigma)*exp(-(x-mu)^2/(2*sigma^2)). (3)

二维高斯函数为

G(x,y)=1/(2*pi*sigma^2)*exp(-(x^2+y^2)/(2*sigma^2)). (4)

其中x和y代表了相对于原始中心的偏移值,即距离中心点x轴向上的距离与y轴向上的距离. sigma表示高斯模糊的标准差.

根据二维高斯函数,可以计算出一个N×N权重矩阵,其中N取值与sigma相关. 通过将原始数据矩阵与该权重矩阵进行卷积运算,即可得到模糊后的矩阵数据. 但高斯模糊方法存在着两个问题,随着N的增加,模糊运算的时间消耗将逐级增加,其运算量如表1所示,另一个问题是,范围大于N的像素点,其权重为0,即对当前点无贡献,而实际情况是即使在3sigma的像素点也对中心点有贡献.

快速高斯模糊在一定程度上解决了以上两个问题,具体步骤分为两步,首先对图像做一次前向滤波. 再对图像做一次后向滤波,前向滤波公式为

w(n)=a0x(n)+a1x(n-1)-b1w(n-1)-b2w(n-2). (5)

后向滤波公式为

y(n)=a2x(n)+a3x(n+1)-b1y(n+1)-b2y(n+2). (6)

其中,x为输入的数据序列,y为输出的结果序列,w为中间结果序列,a0,a1,a2,a3,b1,b2均为滤波系数.

从式(5)和式(6)可以看出,每个输出点的计算与N没有关系,而6个滤波系数是sigma的函数,只需计算一次,对于不同的N,每个输出点的计算量是相同的,其计算量如表1所示,相对于原始的高斯模糊方法,其计算量大幅下降,能满足性能要求,且质量不会下降,甚至CPU也能达到实时处理要求.

表1 对于每个输出点的计算量

Tab.1 Calculation cost of each output point

	计算量	
	乘法	加法
高斯模糊	(2N+1)×模糊方向数	2N×模糊方向数
快速高斯模糊	8×模糊方向数	6×模糊方向数

2.5 分类

基于降维的分类方法,建立好模型以后,可以得到k·C_n^2个投影面,即针对每一类别均有C_n^2个投影面,如图7所示,对于新输入的测试点,分别根据不同的类别取出该测试点在各投影面的密度值进行求和(或求平方和),此时可以得到k个和值,根据和值的大小,可以判断该测试点属于k个类别中的哪一类.

这一过程运算量极低,只需进行k·C_n^2次取数据操作与k·(C_n^2-1)次加法操作即可,这决定了本算法具有很高的时间效率.

3 实验及结果分析

本文通过4组实验数据对比传统分类方法支持向量机(SVM)、逻辑回归(LR)来验证基于正交投影的降维分类方法的可行性,比较不同类别数量,不同属性数量与不同数据量的情况下它们的分类准确率,建模效率与分类效率. 4组实验数据均来源于网络公开数据集. 数据具体情况如表2所示,实验环境如表3所示.

实验1的数据来源于真钞和假钞的图像,其图像通过工业相机拍摄,大小为400×400像素. 它的4个属性均是从图像中抽取的特征,其中属性1表示图像小波变换的方差(variance),属性2表示图像小波变换的

表 2 数据集基本情况

Tab.2 Basic information of data set

编号	问题	数据量	属性数量	类别数量
实验1 ^[13]	纸币真伪	1 372	4	2
实验2 ^[14]	区分皮肤非皮肤	245 057	3	2
实验3 ^[15]	页块分类	5 473	10	5
实验4 ^[16]	字母识别	9 148	16	12

表 3 实验环境情况表

Tab.3 Experimental environment

操作系统	处理器	内存	编程环境
mirosoft windows 10, 64位操作系统	Intel(R) Core(TM)i5-4300U CPU @1.90 GHZ 2.50 GHz	4 GB	JAVA/eclipse neon

偏态(skewness), 属性3表示图像小波变换的峭度(kurtosis), 属性4表示图像的熵(entropy). 4个属性均为连续的实数, 数据集共有两个类别标识, 数据的基本情况如表4所示. 数据集的数据量并不大, 本文取其中的50%作为训练集, 另外的50%作为测试集.

表 4 实验1数据集基本情况

Tab.4 Basic information of data set

属性	最小值	最大值	平均值
属性1	-7.042 1	0.965 74	0.433 7
属性2	-13.773 1	3.192 9	1.922 3
属性3	-5.286 1	17.927 4	1.397 6
属性4	-5.874 1	-2.032 3	-1.191 6

取range=300对数据进行规范化与离散化处理, 其中range需要人为调节, 理论上其数值越大, 表示精度越高, 最后分类的准确率越高, 但同时消耗的内存也越大, 因此需要做出平衡, 在维度数量与类别数量较低的情况下, 可以选择较大的range, 而当类别数量与维度都很大的时候, 为了降低内存消耗, 可以选择较小的range.

对数据进行建模, 快速高斯模糊中取 σ 为6.02, 可以得到 $2 \cdot C_4^2 = 12$ 个投影面, 如图9所示, 图中, 本文假设4个维度属性分别对应于坐标 w, x, y, z .

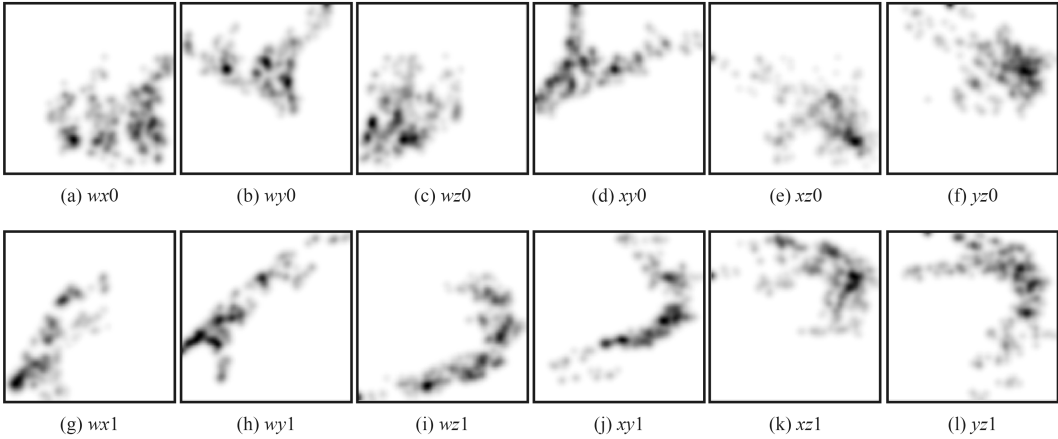


图 9 实验1投影模型示意图

Fig.9 Sketch map of projection model

应用该模型进行分类, 本文方法得到了约98.97%的分类准确率, 且训练耗时与分类耗时都很低, 只耗费了78 ms与小于1ms的时间, 这里实验得出的时间均为多组实验的平均值. 作为对比, 本文同时运用了两种应用广泛的分类方法: 支持向量机与逻辑回归进行了测试, 实验结果对比如表5中实验1所示, 可以发现对于该数据集3种方法得到的分类准确率相差不多, 且都比较高, 其中SVM>本文>LR, 但在耗时方面本文方法要远低于SVM与LR两种方法.

表 5 实验结果对比

Tab.5 Experiment results comparison

实验编号	实验方法	正确率/%	训练耗时/ms	分类耗时/ms
实验1	SVM	99.125 46	203	47
	LR	98.542 27	219	16
	本文	98.979 59	78	0.1
实验2	SVM	90.756 30	7 470	2 273
	LR	96.748 26	2 097	28
	本文	89.948 83	75	31
实验3	SVM	99.173 25	5 872 094	999 585
	LR	91.893 35	3 014	276
	本文	98.023 32	87	42
实验4	SVM	95.889 81	2 567	1 675
	LR	81.853 95	4 621	84
	本文	88.019 24	1 634	697

实验2、3、4采取与实验1相同的方法, 在相同的运行环境中进行, 这里不对其进行详细描述, 给出4组试验的最终结果, 如表5所示, 文献[14-16]中有数据集的具体描述信息.

实验2的数据量较大, 能够更清晰反映算法之间运算耗时的差距, 实验1、2针对的是二分类问题, 而实验3、4数据的维度与类别数量相对较大, 反映了对

于多分类问题算法的运行情况。4组实验得到了相似的结果,其中多数情况下SVM方法在准确率上是最优的,本文方法得到的准确率稍低,但差距并不大,而在耗时方面本文方法要远远优于SVM与LR方法。

4 结论

本文介绍了一种基于降维的分类方法,能够把高维空间的问题转化为多个二维空间问题以降低分类算法复杂度,并给出了分类模型训练方法,通过3组实验的结果表明了基于正交投影的降维分类方法与传统的分类方法相比存在着明显的速度优势,且其预测准确率也只是稍低于以准确率著称的SVM分类方法。降维分类方法算法简单,易于编程,具有实际应用价值。当然该方法也存在着一些不足,如模型的存储占用内存比一般分类方法要大,对倾斜数据敏感等问题都需要在未来的工作中研究。

参考文献:

- [1] TENG L Y, TENG S H, TANG F, *et al.* A collaborative and adaptive intrusion detection based on SVMs and decision trees[C]// IEEE International Conference on Data Mining Workshop. [S.l.]: IEEE, 2014: 898-905.
- [2] 滕少华, 严远驰, 刘冬宁, 等. 基于FCM-C4.5的双过滤入侵检测机制[J]. 计算机应用与软件, 2016, 33(1): 307-311.
TENG S H, YAN Y C, LIU D N, *et al.* A dual filtration intrusion detection mechanism based on FCM and C4.5 [J]. Computer Applications and Software, 2016, 33(1): 307-311.
- [3] VARUNA S, NATESAN P. An integration of k-means clustering and naïve bayes classifier for intrusion detection[C]// International Conference on Signal Processing, Communication and NETWORKING. [S.l.]: IEEE, 2015.
- [4] GUMUS F, SAKAR C O, ERDEM Z, *et al.* Online naïve bayes classification for network intrusion detection[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. [S.l.]: IEEE, 2014: 670-674.
- [5] 华辉有, 陈启买, 刘海, 等. 一种融合Kmeans和KNN的网络入侵检测算法[J]. 计算机科学, 2016, 43(3): 158-162.
HUA H Y, CHEN Q M, LIU H, *et al.* Hybrid K-means with KNN for network intrusion detection algorithm [J]. Computer Science, 2016, 43(3): 158-162.
- [6] 张雪芹, 顾春华, 吴吉义, 等. 基于约简支持向量机的快速入侵检测算法[J]. 华南理工大学学报(自然科学版), 2011, 39(2): 108-112.
ZHANG X Q, GU C H, WU J Y, *et al.* Fast intrusion detection algorithm based on reduced SVM [J]. Journal of South China University of Technology (Natural Science Edition), 2011, 39(2): 108-112.
- [7] 毕孝儒. 基于粗糙集属性约简和加权SVM的入侵检测方法研究[D]. 西安: 西安科技大学计算机学院, 2011.
- [8] DU H, TENG S, FU X, *et al.* A cooperative intrusion detection system based on improved parallel SVM[C]// Pervasive Computing (JCPC), 2009 Joint Conferences on. [S.l.]: IEEE, 2009: 515-518.
- [9] 刘琪. DH-SVM: 基于SVM和动态混合算法的公交车辆路段运行时间估计与预测方法的研究[D]. 济南: 山东大学微电子学院, 2015.
- [10] 柏丛, 彭仲仁. 基于动态模型的公交车行程时间预测[J]. 计算机工程与应用, 2016, 52(3): 103-107.
BAI C, PENG Z R. Bus travel time prediction based on dynamic model [J]. Computer Engineering and Applications, 2016, 52(3): 103-107.
- [11] 杨婷, 滕少华. 改进的贝叶斯分类方法在电信客户流失中的研究与应用[J]. 广东工业大学学报, 2015, 32(3): 67-72.
YANG T, TENG S H. Research and application of improved bayes algorithm for the telecommunication customer churn [J]. Journal of Guangdong University of Technology, 2015, 32(3): 67-72.
- [12] 夏琴晔, 杨宜民. 基于biSCAN和SVM的机器人目标识别新算法研究[J]. 广东工业大学学报, 2013, 30(4): 65-69.
XIA Q Y, YANG Y M. Research on a new algorithm for robots's recognition of objects based on biSCAN and SVM [J]. Journal of Guangdong University of Technology, 2013, 30(4): 65-69.
- [13] LOHWEG V. UCI Machine Learning Repository: banknote authentication Data Set[EB/OL]. (2012-03-01) [2017-02-22]. <http://archive.ics.uci.edu/ml/datasets/banknote+authentication#>.
- [14] BHATT R B, SHARMA G, DHALL A, *et al.* Efficient skin region segmentation using low complexity fuzzy decision tree model[C]// India Conference (INDICON), 2009 Annual IEEE. [S.l.]: IEEE, 2009.
- [15] MALERBA D. UCI Machine Learning Repository: Page Blocks Classification Data Set[EB/OL]. (1996-11-03) [2017-02-22]. <http://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification>.
- [16] MALERBA D. UCI Machine Learning Repository: Letter Recognition Data Set[EB/OL]. (1991-01-01) [2017-02-22]. <http://archive.ics.uci.edu/ml/datasets/Letter+Recognition>.