基于最大最小模块化的支持向量机多分类问题

金巧巧[1]　李研[2]

摘要

本文探讨了解决三级分类问题两种方法的性能比较，一是使用支持向量机分类器，二是解决三级分类问题使用 $Min-Max\quad Module\quad SVM$ 和 $one-vs-rest$ 任务分解方法。

关键字：任务分解策略, 最小最大模块化

Abstract

This paper discusses the method of dividing training samples by using random task decomposition and task decomposition strategy with prior knowledge.

Keywords: min-max-module, task decomposition method

1　前言

解决三级分类问题，我们讨论了两种方法，一是在给定的数据集使用支持向量机分类器和 $one-vs-rest$ 策略。支持向量机分类器在很多机器上都有学习库，我们将使用库来解决这个问题。二是解决三级分类问题使用 $Min-Max\quad Module\quad SVM$ 和 $one-vs-rest$ 任务分解方法。利用 $one-vs-rest$ 方法将三类问题转化为两类问题，再利用随机任务分解和有先验知识的任务分解策略，作为训练样本的划分方法，得到相对平衡的划分子集，将这些不平衡的两类问题分解为平衡的两类问题。最后把两种方法所得到的支持向量机的性能进行比较。

集成学习是一种有效的海量数据学习方法。其中集成学习方法之一就是在训练阶段先对大规模训练集按照一定规则进行适度的划分，并对每个划分好的训练集进行学习得到各个子分类器。最后通过一定的集成规则将各个子分类器结果集成，得到原问题的结果。最小最大模块化网络方法就是一种非常有效的集成学习方法。该方法的基本思想就是在训练阶段对训练集进行划分，然后通过基于涌现理论的 MIN 规则和 MAX 规则对子分类器的结果进行集成得出原问题的结果[1]。

1.1　数据集

在实例运用中，我们使用了两套训练集，一套是课题提供的数据集和标签集，另一套是关于垃圾分类的图像识别，其中包含金属，塑料和纸箱三个类别。我们均使用了上述两套方法进行性能比较，针对第一套数据我们分别使用随机任务分解和基于先验知识分解，对于第二套数据集我们仅使用随机任务分解。

1.1.1　数据集介绍

垃圾分类是当下全国都在计划并实施的环保政策，我们收集到了包括硬纸板 $(cardboard)$、金属 $(metal)$、塑料 $(plastic)$ 三种可回收垃圾的 1295 张

[1]解晓敏, 李云，最小最大模块化网络中基于聚类的数据划分方法研究, 南京大学学报 (自然科学),2012-3

[2]引自和鲸社区. 草莓小救星的“另一个垃圾分类数据集”数据集，有删改

图片[2]，其中包括硬纸板 403 张，金属 410 张和塑料 482 张，物品都是放在白板上在日光/室内光源下拍摄的，压缩后尺寸为 512*384。我们对数据集进行了人工标记，使三类样本的标签分别为 0，1，2。

1.1.2 数据集预处理

我们使用 PIL 对图像进行了预处理，将原始的 RGB 图像转化为灰度图；考虑到图像中的检测目标大多位于图片中央，图片周围区域包含信息较少，我们对图片进行了再裁剪，使其尺寸统一为 256*256。

1.1.3 图片的特征提取

由于 SVM 处理图像信息有一定的局限性，我们首先对图像进行 hog 特征提取。$hog(Histogram\ \ of\ \ Oriented\ \ Gradient)$ 特征提取，也称方向梯度直方图，用来描述图像局部纹理特征。其具体步骤为：

(1) 首先将图片分割成 $cell$ 和 $block$，其中每个 $cell$ 的尺寸为（16，16）个像素，$cell$ 是计算方向梯度的最小单元，求出 $cell$ 中各像素的方向梯度之和后相加，得到每个 $cell$ 的方向梯度直方图；每个 $block$ 的尺寸为 2*2 个 $cell$，作为单元格来规范化每个 $cell$ 的柱状图，规范函数为参数中“$block-norm$”，本程序选择 $L2-Hys$ 作为规范函数，其中 $Orientation=9$ 的含义是将方向分组为特定的 9 组方向，相当于把周角九等分便于统计画图。（图源网络）

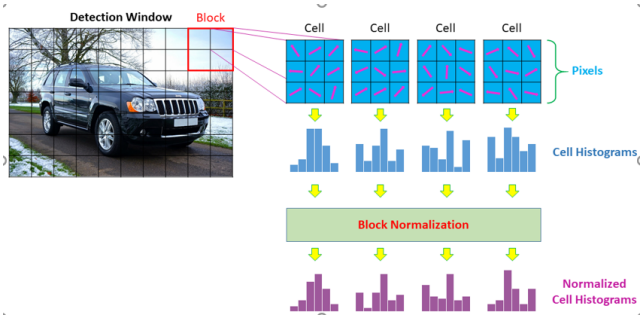

图 1: 图片的特征提取

(2) 之后以 $block$ 为单位，$block/2$ 为步长在图片内进行滑动，最后将所有 $block$ 中的标准化柱状图收集到一个 HOG 的特征向量中，得到 HOG 特征提取的结果。

2 求解过程

2.1 任务分解

在最小最大模块化网络中，对于一个的 K 类问题通过“部分对部分”的策略分解成 $K(K-1)/2$ 个二类问题。由于每个二类问题都含有较大数量的样本，所以需要进一步划分成一系列小的相对平衡的二类子问题。我们首先根据训练集的标签集，把数据集划分为三类，并利用 $one-vs-rest$, 把数据集划分为不平衡的三个两类问题。

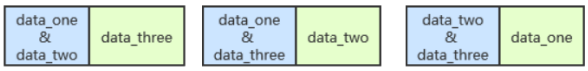

图 2: 划分结果

对于每一个二类问题，进一步划分为正训练集和负分类集[1]，即 $T=X^+\cup X^-$

$$X^+=(x_t^+,+1)_{t=1}^{l^+}$$

$$X^-=(x_t^-,-1)_{t=1}^{l^-}$$

l^+ 和 l^- 为正训练集和负训练集的样本个数。我们利用任务分解策略把正训练集和负训练集划分为 N^+ 和 N^- 个样本相关性较大的且互不相交的子训练集：

$$X_j^+=(x_t^{+,j},+1)_{t=1}^{l_j^+}$$

$$X^+=\cup_{j=1}^{N^+},\cap_{j=1}^{N^+}=\Phi,j=1,...,N^+$$

$$X_j^-=(x_t^{-,j},+1)_{t=1}^{l_j^-}$$

$$X^-=\cup_{j=1}^{N^-},\cap_{j=1}^{N^-}=\Phi,j=1,...,N^-$$

X_j^+ 和 X_j^- 分别表示划分出的第 j 个正训练子集和负训练子集，于是每个二类问题组合成为 $N^+\times N^-$ 个子问题：

$$C_{i,j}\cup X_j^-,i=1,2,...,N^+,j=1,2,...,N^-$$

所以一个三类问题分解为 $\sum_{i}^{3}\sum_{j=i+1}^{3}N_i^+\times N_j^-$

[1]解晓敏, 李云，最小最大模块化网络中基于聚类的数据划分方法研究, 南京大学学报 (自然科学),2012-3

2.2 任务分解策略

我们需要把正训练集和负训练集划分为 N^+ 和 N^- 个样本相关性较大的且互不相交的子训练集。

2.2.1 随机分解

随机规则: 随机规则是最基本、最简单的分解规则。它首先随机打乱训练样本，并将其切成互不相关的片段。当训练样本足够多时，每个片段仍然代表原类别。在这种情况下，随机规则可以正常工作[3]。

2.2.2 基于先验知识分解

先验知识规则: 如果在训练过程中可以获得先验知识，则可以指导任务分解 [3]。

基于现有的数据集特征和规模，我们求得第一维数据的期望值，取负样本集中第一维数据在期望值附近的数据，将负样本数量控制在与正样本相近的区间内。从而实现由不平衡样本到平衡样本之间的转换。

2.3 svm 训练

因为各个子问题之间是相互独立的，可以分别对 $N^+ \times N^-$ 个二类子问题 $C_{i,j}$ 运用支持向量机 svm 进行训练得到 N^+ 和 N^- 个子分类器 [2]。采用线性核函数 SVM 模型，将数据映射到线性区间进行分类，调节正则化系数以调整模型的拟合程度。

而后在测试阶段，将测试样本提交给训练得到的所有子分类器，就可以得到 $N^+ \times N^-$ 个分类结果. 假设每个子分类器的输出记为 $C_{i,j}(X), (i = 1, 2, ..., N^+; j = 1, 2, ..., N^-)$。

2.4 模块集成——最大最小模块化支持向量机

在成功地对每个个体进行两类支持向量机的训练后，将所有训练好的支持向量机按照最小和最大的组合原则和最大化原则集成到一个 $M^3 - SVM$ 中。最小单位的作用是从它的多个输入中找出一个最小值，而最小单位的作用是从它的多个输入中找出一个最小值[4]。MIN 单元是对拥有相同正类训练样本和不同负类训练样本的问题集成得到：

$$C_i(X) = \min_{1<j<N^-} C_{i,j}(X), i = 1, 2, ..., N^+$$

然后通过 MAX 单元，MAX 单元是对拥有相同负类训练样本和不同正类训练样本的问题集成，得到：

$$C(X) = \max_{1<i<N^+} C_i(X)$$

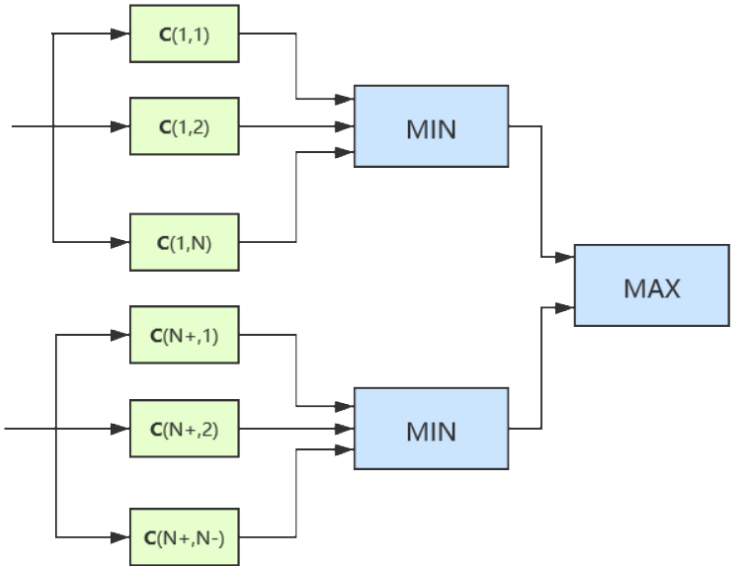

图 3: 二类问题的最小最大模块化网络集成图

3 求解结果

3.1 课题数据集

通过实验我们得到的分类结果如表所示：

[3]Zhi-Fei Ye and Bao-Liang Lu,Learning Imbalanced Data Sets with a Min-Max Modular Support Vector Machine,Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

[4]Chao Ma1,Bao-Liang Lu,Masao Utiyama,Incorporating Prior Knowledge into Task Decomposition for Large-Scale Patent Classification

		训练精度	测试精度
传统 SVM		1.0	0.480
基于随机任务分解的 Min-max	二分类器(一)	1.0	0.716
	二分类器(二)	1.0	0.697
	二分类器(三)	1.0	0.766
	OvR-SVM	/	0.528
基于先验知识分解的 Min -max	二分类器(一)	1.0	0.757
	二分类器(二)	1.0	0.713
	二分类器(三)	1.0	0.700
	OvR-SVM	/	0.598

图 4: 课题数据集的训练结果

从表中结果我们可以看出使用随机任务分解策略的 $MIN-MAX$ 分类器的分类结果与 $One-vs-Rest$ 传统 SVM 分类器的分类结果相近，而采用随机任务分解和基于先验知识任务分解策略的 $MIN-MAX$ 分类器的分类结果比传统 SVM 分类结果提高近 10% 的正确率，且基于先验知识任务分解策略正确率更高。说明基于先验知识的 $MIN-MAX$ 分类器分类效果比 SVM 更好。下面基于 $MIN-MAX$ 的原理对实验结果进行解释分析。

直观的来说，MIN 过程可以看作是将样本空间设定为正类中某部分样本与所有负类样本，只有当这个样本无可争议的被分类到正类子部分，才认为样本在该正类部分空间中；MAX 过程则表示当正类中所有子部分都经过上述过程后，如果存在某一部分，样本被确切的分到正类，那就可以认为整个系统的输出即为正，否则输出为负[5]。因此使用 $MIN-MAX$ 策略得到的分类结果，相当于 SVM 分类结果集成并提高模型置信度的结果，从而准确度高于传统的 SVM 模型。

3.2 垃圾图片分类结果

通过实验我们得到的分类结果如表所示：

		训练精度	测试精度
传统 SVM		0.989	0.575
Min-Max 策略分类结果	二分类器(一)	1.0	0.792
	二分类器(二)	1.0	0.645
	二分类器(三)	1.0	0.645
	OvR-SVM	/	0.602

图 5: 垃圾图片数据集的训练结果

从表中我们可以看出最终结果差不多，之所以正确率提高减少的原因是我们进行了 hog 特征提取，提取出特征后提高了 SVM 的分类结果的同时，也降低了 $MIN-MAX$ 在模型训练中提高正确率所起到的贡献。

4 总结

从实验可以发现，最小最大模块化方法是在训练阶段通过对训练集进行划分，后利用 MIN 规则和 MAX 规则对子分类器的结果进行集成得出原问题的结果。可以扩大数据处理范围，尽管在实验过程中，由于任务分解策略不够优略，数据集的数量不够等原因，体现出来的优势不够明显，但是通过查阅资料表明，最小最大模块化方法是一种非常有效的机器学习方法。

[5]Bao-Liang Lu and Masami Ito,Task Decomposition and Module Combination Based on Class Relations: A Modular Neural Network for Pattern Classification