面向分组类别概率问题的模糊 SVM 分类算法

陈自洁[1,2],陆小兵[3],杨晓伟[4]
(1. 广东药学院 医药商学院,广东 中山 528485;
2. 华南理工大学 计算机科学与工程学院,广东 广州 510640;
3. 华为技术有限公司,广东 深圳 518129;
4. 华南理工大学 数学科学学院,广东 广州 510640)

摘 要:分组类别概率问题(Q-GP)给定样本的群组统计信息或类别概率分布,寻求每个个体样本的实际类标签,有着广泛的实际应用,但目前相应的研究仍较少。Q-GP 问题求解的关键是如何利用已知的样本群组信息来获取单个样本的分类信息。文中通过比较二分类 Q-GP 问题与有监督及半监督二分类问题的异同,提出利用模糊分类的思想,根据已知的各群组类别概率分布近似获取个体样本的类隶属度,以此构造有监督样本进行学习。具体方法是:首先使用 fuzzy 层次分类构造各群组的等价类,并利用等价类将二分类 Q-GP 问题变换成多个带模糊隶属度的有监督二分类子问题;然后实施 fuzzy SVM 训练子分类器;最后整合多个子分类器的结果即得到每个样本的类标签估计。

关键词:分组类别概率;fuzzy 层次分类;fuzzy SVM
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2013)11-0046-04
doi:10.3969/j.issn.1673-629X.2013.11.012

A Fuzzy SVM Classification Algorithm from Group Probabilities Problem

CHEN Zi-jie[1,2], LU Xiao-bing[3], YANG Xiao-wei[4]
(1. School of Medicine Business, Guangdong Pharmaceutical College, Zhongshan 528485, China;
2. School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China;
3. Huawei Technology Co., Ltd, Shenzhen 518129, China;
4. School of Mathematical Sciences, South China University of Technology, Guangzhou 510640, China)

Abstract: The problem of estimation from group probabilities (Q-GP) is to find the actual labels of the individual samples given the label proportions in each group, which has a wide application, but lacking of the existing study. The Q-GP solution is critical to use the known information of group probabilities to obtain the classification information for single sample. In this paper, present a fuzzy classification method based on fuzzy support vector machine (SVM) to solve this problem by comparing the binary Q-GP with the supervised and semi-supervised binary classification in difference. Firstly, introduce the fuzzy hierarchical classification to find the relationships between objects in a group, so as to decompose the binary Q-GP into supervised sub-problems with fuzzy memberships. Then A fuzzy SVM is trained for each sub-problem. At last combine multiple sub-classifiers to get the final labels of all individual samples.

Key words: group probabilities; fuzzy hierarchical classification; fuzzy SVM

0 引 言

根据训练样本中所包含的标识样本的多少,机器学习主要可分为有监督学习(supervised learning)、半监督学习(semi-supervised learning)和无监督学习(unsupervised learning)三大类。支持向量机(support vector machine,SVM)[1]能较有效地处理以上三类机器学习问题。最近有学者提出一种基于类别比例的分类问题(estimating labels from label proportions),也叫

收稿日期:2013-01-22 **修回日期**:2013-04-27 **网络出版时间**:2013-08-28
基金项目:国家自然科学基金资助项目(61273295)
作者简介:陈自洁(1980-),女,讲师,博士生,研究方向为模式识别、人工智能;导 师:郝志峰,教授,研究方向为智能计算与仿生算法、人工智能。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130828.0829.012.html

做基于分组类别概率的分类问题(estimation from group probabilities)[2-4],这类问题介于有监督学习与无监督学习之间,而又有别于传统定义上的半监督学习。该类问题给定若干个独立同分布的无标签样本集,每个集合中已知各类别样本的比例,其目标是根据这些信息来估计样本的类别标签(如图 1 所示)。问题描述如下:假设有独立同分布样本集为 $X = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in R^m, i = 1, \cdots, n\}$,有监督分类问题的类别为 $y \in Y$。该问题给定 l 个 X 的子集 $S_k = \{\boldsymbol{x}_i^k \mid \boldsymbol{x}_i^k \in X, i = 1, \cdots, m_k\}$,以及 S_k 中属于类别 y 的样本所占的比例 $p_{ky} = |\{\boldsymbol{x}_i^k \in S_k : y_i = y\}| / m_k$,测试样本集 $X' = \{\boldsymbol{x}_1', \boldsymbol{x}_2', \cdots, \boldsymbol{x}_t'\}$。

目标是根据已知的多个训练样本集 (S_k, p_{ky}),以最小的错误寻找分类函数 $f: X \to Y$,来估计各个样本的类别标签。其中

$$S_k \subset X, \bigcup_{k=1}^{l} S_k = X, m_k = |S_k|, k = 1, \cdots, l \qquad (1)$$

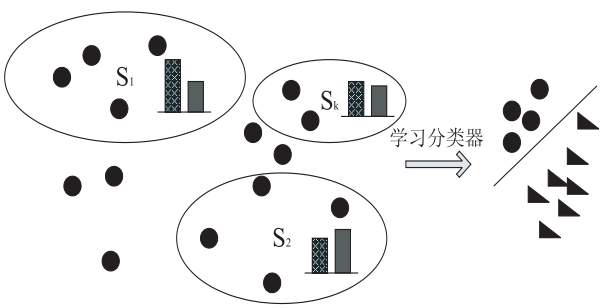

图 1　基于分组类别概率的分类估计问题

文献[5]指出,产生这类不同于标准有监督分类问题的主要原因是:

(1)获取每一个训练样本的类标签是比较困难的,费时费力成本高,而获取一群样本的类别分布或比例相对容易;

(2)由于通信和存储空间的限制,海量的数据只能以聚集的方式进行存储,保存数据的特征及统计信息;

(3)出于隐私保护和信息隐藏的考虑,不允许揭露个体的全部信息,数据只能以统计信息出现。

实际应用中又往往需要在这样的样本条件下,建立模型来对个体样本进行估计。基于分组类别概率的分类估计问题其实就是根据样本的统计信息学习得到个体样本的分类估计,其在电子商务、垃圾邮件过滤、图像内容识别、隐私保留数据挖掘、欺诈检测、行为决策等领域都有重要的应用。

文中尝试结合 fuzzy 分类和有监督模糊支持向量机来求解该分类估计问题。首先通过 fuzzy 层次分类,将基于类别概率的分类问题转换成有监督分类问题,然后用模糊支持向量机(FSVM)算法寻找分类估计函数。

1　相关研究成果

1.1　基于分组类别概率的分类估计

目前关于基于分组类别概率的分类估计问题(下文简称 Q-GP)的研究还不多。可以将 Q-GP 问题看作是一种特殊的半监督学习问题,只是其有指导的信息不再是少数单个样本的类别标签,而是样本的统计信息。

文献[2]针对二分类的 Q-GP 问题(即 Q-GP 中的每一个样本组中给定正类样本的比例的估计)设计了一个层次概率模型来求解不确定参数以及个体样本的未知标签,并通过 MCMC 抽样算法进行优化。该二分类 Q-GP 也可以看作是标准多实例学习(multi-instance learning[6])的一种变形。文献[5,7]也使用与文献[2]相似的方法进行处理。文献[3]提出的 Mean Map 方法则使用指数条件模型 $p(y \mid x, \theta) = \exp((\Phi(x,y)\theta) - g(\theta \mid x))$ 来对类条件概率 $p(y \mid x, \theta)$ 进行建模,并归结为求解一个凸优化问题。文献[3]所定义的学习问题同时假设测试样本集已知类别分布。学者 Platt 于文献[8]中提出使用 scaling 函数将 SVM 分类的数值输出转换为概率输出,文献[4]通过分析 Q-GP 问题与 Platt scaling 的联系,结合适用于 Q-GP 问题的逆 scaling 与支持向量机回归,建立二次优化模型进行有效求解,并通过大量对比实验表明,文献[4]的方法最有效,其次为 Mean Map 方法。对于可扩展性而言,文献[2]的二分类问题可容易地推广到多分类问题,文献[3]的方法原本就假设任意数量的类别,文献[4]的方法要如何应用到多分类问题还有待研究。

1.2　模糊支持向量机(FSVM)

为了解决 SVM 对噪声点和异常点敏感的问题,文献[9-10]等对每个样本引入一个模糊隶属度,将输入样本进行模糊化后变为:

$$S = \{(\boldsymbol{x}_i, y_i, s_i) \mid \boldsymbol{x}_i \in R^d, y_i \in \{+1, -1\}, s_i \in (0,1]\} \qquad (2)$$

模糊隶属度 s_i 表示相应样本 $\boldsymbol{x}_i$ 属于 y_i 的程度。于是求解样本(2)的 FSVM 模型为:

$$\min \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + C\sum_{i=1}^{n} s_i\xi_i$$
$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b) \geqslant 1 - \xi_i, i = 1, \cdots, n; \xi_i \geqslant 0, i = 1, \cdots, n \qquad (3)$$

其中,ξ_i 为松弛变量;C 为预定给定的正则化参数。

引入拉格朗日乘子 α_i,相应的对偶问题为:

$$\max W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{s.t.} \sum_{i=1}^{n} y_i \alpha_i = 0, 0 \leqslant \alpha_i \leqslant s_i C, i = 1, \cdots, n \tag{4}$$

对于模糊隶属度 s_i 的确定有很多方法,也是很多学者研究的方向,如文献[9-12]。

1.3 fuzzy 分类

Fuzzy 等价关系在模式识别、聚类识别、fuzzy 决策、fuzzy 控制、知识发现与约简等方面有着重要的应用。而在实际问题中更多遇到的是不具备传递性的 fuzzy 相似关系,为了得到协调的分类,需要先将 fuzzy 相似关系改造成一个 fuzzy 等价关系,再构造分类[11]。设 $X = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in R^m, i = 1, \cdots, n\}$ 为待分类对象的集合,实现 fuzzy 分类的一般步骤如下:

(1)使用某种度量方法计算任意两个对象 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 的相似系数,以此描述两者的相似度;

(2)确定 X 的 fuzzy 相似矩阵 $\boldsymbol{R}$;

(3)计算 $\boldsymbol{R}$ 的传递闭包,由此得到 X 的 fuzzy 等价关系 $\boldsymbol{R}'$;

(4)对于 $\lambda \in [0,1]$,由小到大动态变化 λ 的值,得到 $\boldsymbol{R}'$ 的不同的 λ 截关系 $\boldsymbol{R}'_\lambda$,从而构成一个由粗到细的动态的多层次 fuzzy 分类。

目前对于 fuzzy 集与 SVM 的研究大多集中在利用模糊隶属度来解决 SVM 的样本选择问题[13-14]、不平衡数据问题、噪声数据问题[9]、多分类的不可分区域等方面,着重于研究新的模糊隶属函数的设计与确定[9-10,12]、fuzzy 聚类的应用[14-15]、fuzzy SVM 的设计与改进等问题[10,12-15]。文献[15-16]则提出一种新的基于 SVM 的 fuzzy 层次分类方法来处理多分类问题,首先对训练数据进行 k-means 聚类,然后动态地建立一个 fuzzy 层次结构,将多分类问题转化成一个多层次的二分类问题,并在每个层次节点上运用 SVM 进行二分类,层层递进最终得到一个层次的多分类结果,应用到文本分类中取得了不错的效果。

2 面向 Q-GP 问题的模糊 SVM 分类方法

Q-GP 与半监督问题的不同之处在于分组给出各类别样本所占比例的统计信息,而不知道任何个体样本的标签信息。这使得不能直接应用常规的分类方法来处理 Q-GP 问题。一个直观的想法是,是否可以将 Q-GP 变换成有监督问题或者半监督问题来求解呢?受文献[11,16-17]启发,fuzzy 分类可以发现对象之间的关系,动态地实现由粗到细的等价分类,可作为 SVM 训练的先验信息。下文沿着这个思路,尝试使用 fuzzy 分类将 Q-GP 转换成有监督问题,然后再进行模糊 SVM 训练。

简化起见,这里先考虑二分类的 Q-GP 问题,即问题(1)中,类别 $y \in Y = \{+1, -1\}$,每个样本子集 S_k 中给定

$$p_{k+} = |\{\boldsymbol{x}_i^k \in S_k : y_i = +1\}| / m_k \text{ or}$$

$$p_{k-} = |\{\boldsymbol{x}_i^k \in S_k : y_i = -1\}| / m_k \tag{5}$$

作为正类样本或负类样本的比例(其中 $k = 1, \cdots, l$)。

于是,可以将问题(5)分解为 l 个二分类子问题,分别求解每一个问题,然后再将结果进行整合。当 l 较大时,每一个子问题的规模一般不大。

2.1 子问题的 fuzzy 层次分类

为了找出无标签样本的聚集特征,对子数据集 S_k 进行 fuzzy 等价分类,位于同一等价类中的样本更有可能属于同一给定类别。在 S_k 中使用 1.3 节的步骤进行操作:

(1)将样本数据进行标准化,压缩到[0,1]闭区间上。

(2)选用 RBF 函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/\sigma^2)$ 作为描述两对象 $\boldsymbol{x}_i, \boldsymbol{x}_j$ 的相似性的度量函数,计算 S_k 中样本之间的相似性。可以在后面的 SVM 训练中使用相同的 RBF 函数,这样可以通过存储 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 的值来避免重复运算。

(3)记 $r_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$,则矩阵 $\boldsymbol{R} = (r_{ij})_{m_k \times m_k}$ 是一个 fuzzy 矩阵,同时容易证明 $\boldsymbol{R}$ 是自反的、对称的,因此 $\boldsymbol{R}$ 是一个 fuzzy 相似矩阵,但通常 $\boldsymbol{R}$ 不直接满足传递性。

(4)采用平方法求 $\boldsymbol{R}$ 的传递闭包 $\boldsymbol{R}^*$,则 $\boldsymbol{R}^*$ 是一个 fuzzy 等价关系。其中 fuzzy 矩阵的合成运算“ ∘ ”中使用 Zadeh 算子 (∧, ∨),即一般矩阵乘法中的“ • ”改为“ ∧ ”,“+”改为“ ∨ ”。

(5)根据 fuzzy 集的理论[11],对于 $\lambda \in [0,1]$,$\boldsymbol{R}^*$ 的 λ 截关系 $\boldsymbol{R}^*_\lambda$ 是通常等价关系。动态变化 λ,再按 $\boldsymbol{R}^*_\lambda$ 分类,则得到 S_k 的一簇等价类 $S_k/\boldsymbol{R}^*_\lambda$。

2.2 建立近似的有监督二分类问题

容易知道,当 $\lambda = 0$ 时,$S_k/\boldsymbol{R}^*_\lambda$ 是最粗的分类,所有的对象都属于同一类;当 $\lambda = 1$ 时,$S_k/\boldsymbol{R}^*_\lambda$ 是最细的分类,每一个对象都各自为一类。也就是说,在 $\lambda = 0 \to 1$ 的过程中,$S_k/\boldsymbol{R}^*_\lambda$ 不断分裂调整原等价类簇,等价类的数量逐渐增多,同一等价类中的对象的相似性也随着 λ 的增大而增加。而在二分类的 Q-GP 问题中,希望通过调整 λ 得到只包含两个等价类的商集 $S_k/\boldsymbol{R}^*_\lambda$。显然,由 $\lambda = 0$ 开始,经过少数的几步调整就有可能出现希望的等价类。

```
设 γ 为 λ 的增长步长;
初始化 λ: = 0;flag: = 1;a: = ∞;
while(flag = = 1)
```

}

$\lambda := \lambda + \gamma$;

计算 $S_k/\boldsymbol{R}_\lambda^*$;

if($|S_k/\boldsymbol{R}_\lambda^*| < 2$) continue;

else if($|S_k/\boldsymbol{R}_\lambda^*| = 2$){

计算 $a = \min(a, \min(\left|\frac{|A_1|}{m_k} - p_{k+}\right|, \left|\frac{A_2|}{m_k} - p_{k+}\right|)$, 其中 $A_1, A_2 \subset S_k/\boldsymbol{R}_\lambda^*$;

continue;

}

else{

flag = 0;

记此时与 a 值对应的商集 $S_k/\boldsymbol{R}_{\lambda_a}^* = \{A_1^a, A_2^a\}$ 为最接近二分类 Q - GP 给定比例的等价类簇;

if($\left|\frac{|A_1^a|}{m_k} - p_{k+}\right| < \left|\frac{|A_2^a|}{m_k} - p_{k+}\right|$) 则赋予 A_1^a 中的样本 + 1 标签,A_2^a 样本 - 1 标签;

else 赋予 A_2^a 中的样本 + 1 标签,A_1^a 样本 - 1 标签;

}

}

由此得到一个二分类问题的有监督样本

$$(\boldsymbol{x}_{1i}^a, +1), (\boldsymbol{x}_{2j}^a, -1), i = 1, \cdots, |A_1^a|; j = 1, \cdots, |A_2^a|; \boldsymbol{x}_{1i}^a \in A_1^a, \boldsymbol{x}_{2j}^a \in A_2^a \quad (6)$$

2.3 fuzzy SVM 训练及子分类器整合

由(2.2)所得到的有监督样本(6)的临时标签是近似的,给予每一个样本对于其临时标签的隶属度,得到样本(7):

$$(\boldsymbol{x}_{1i}^a, +1, s_{1i}), (\boldsymbol{x}_{2j}^a, -1, s_{2j}), i = 1, \cdots, |A_1^a|; j = 1, \cdots, |A_2^a|; \boldsymbol{x}_{1i}^a \in A_1^a, \boldsymbol{x}_{2j}^a \in A_2^a, s_{1i}, s_{2j} \in (0,1] \quad (7)$$

由此可用(4)的 FSVM 来对(7)进行训练,得到子数据集 S_k 的一个模糊分类函数 $f_k(\boldsymbol{x})$,则 $y^{(k)} = \text{sign}(f_k(\boldsymbol{x}))$ 为第 k 个分类器对样本 $\boldsymbol{x}$ 的分类结果。问题(5)可以得到 l 个模糊分类函数 $f_1(\boldsymbol{x}), \cdots, f_l(\boldsymbol{x})$,及对应的 $y^{(1)}, \cdots, y^{(l)}$,任一样本 $\boldsymbol{x}$ 的最终类别标签通过投票表决的方式来决定。

3 结束语

基于分组类别概率的分类问题(Q-GP)不同于已被广泛研究的有监督、半监督和无监督问题,其关于样本的监督信息是群组的统计信息,而非个体标签,难以直接利用现有的分类方法求解。

文中在分析了 Q-GP 问题特点的基础上,讨论了使用 fuzzy 层次分类将 Q-GP 问题转化成带模糊隶属度的有监督问题的可行性和具体方法,并且通过整合多个子问题的 fuzzy SVM 子分类器最终求得原问题中各个样本的类别标签。在相关的 Q-GP 数据集上进行数值实验来验证所提出算法的有效性是下一步工作的方向。

参考文献:

[1] Vapnik V. Statistical learning theory[M]. Chichester: Wiley, 1998.

[2] Kueck H, de Freitas N. Learning about individuals from group statistics[C]//Proc of UAI. Arlington, Virginia: AUAI Press, 2005: 332-339.

[3] Quadrianto N, Smola A J, Caetano T S, et al. Estimating labels from label proportions[J]. Journal of machine learning research, 2009, 10: 2349-2374.

[4] Stefan R. SVM classifier estimation from group probabilities [C]//Proceedings of the 27th international conference on machine learning. Haifa, Israel: [s. n.], 2010.

[5] Chen B, Chen L, Ramakrishnan R, et al. Learning from aggregate views[C]//Proceedings of the 22nd international conference on data engineering. [s. l.]: [s. n.], 2006: 3-12.

[6] Dietterich T G, Lathrop R H, Lozano-Perez T. Solving the multiple instance learning with axis-parallel rectangles[J]. Artificial intelligence, 1997, 89(1/2): 31-71.

[7] Musicant D, Christensen J, Olson J. Supervised learning by training on aggregate outputs[C]//Proc of 7th IEEE international conference on data mining. Omaha, NE: [s. n.], 2007: 252-261.

[8] Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods [M]//Advances in large margin classifiers. [s. l.]: MIT Press, 1999.

[9] Lin C F, Wang S D. Training algorithms for fuzzy support vector machines with noisy data[J]. Pattern recognition letters, 2004, 25(14): 1647-1656.

[10] 杨晓伟,闫　丽. 基于模糊分割的支持向量机分类器[J]. 计算机工程与应用,2007,43(28):187-189.

[11] 徐宗本,张讲社,郑亚林. 计算智能中的仿生学:理论与算法[M]. 北京:科学出版社,2003.

[12] 赵克楠,李　雷,邓　楠. 一种构造模糊隶属度的新方法[J]. 计算机技术与发展,2012,22(8):75-77.

[13] Chen Zijie, Liu Bo, He Xupeng. A SVC iterative learning algorithm based on sample selection for large samples[C]//Proc of international conference on machine learning and cybernetics. Hong Kong: [s. n.], 2007: 3308-3313.

[14] 陈自洁,夏成锋. 基于模糊 c-均值聚类的 SVC 迭代训练算法[J]. 仲恺农业工程学院学报,2011,24(1):39-43.

[15] Guernine T, Zeroual K. SVM fuzzy hierarchical classification method for multi-class problems[C]//Proc of international conference on advanced information networking and applications. Bradford: [s. n.], 2009: 691-696.

[16] Guernine T, Zeroual K. A new fuzzy hierarchical classification based on SVM for text categorization[C]//Lecture notes in computer science. [s. l.]: [s. n.], 2009: 865-874.

面向分组类别概率问题的模糊SVM分类算法

作者： 陈自洁， 陆小兵， 杨晓伟， CHEN Zi-jie， LU Xiao-bing， YANG Xiao-wei

作者单位： 陈自洁,CHEN Zi-jie(广东药学院 医药商学院，广东 中山 528485; 华南理工大学 计算机科学与工程学院，广东 广州 510640)， 陆小兵,LU Xiao-bing(华为技术有限公司,广东 深圳,518129)， 杨晓伟,YANG Xiao-wei(华南理工大学 数学科学学院,广东 广州,510640)

刊名： 计算机技术与发展

ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2013(11)