基于核方法的一种新的模糊支持向量机

李 雷[1],鲁延玲[1],周蒙蒙[1],柏永成[2]
(1.南京邮电大学 理学院,江苏 南京 210003;
2.中国科学技术大学 软件学院,安徽 合肥 230026)

摘 要:由于支持向量机对样本中的噪声及孤立点非常敏感,因而在解决非线性、高维数、不确定问题时,使用模糊支持向量机比使用支持向量机的效果要好。在模糊支持向量机中,模糊隶属度函数的建立是关键也是难点。一般,模糊隶属度是在原始空间中根据样本点的相互距离及到类中心的距离创建的。考虑样本间的密切度,在特征空间中利用混合核函数建立一种新的模糊隶属度。通过试验比较多项式核函数、高斯径向基核函数与混合核函数,可看出新方法表现出了它的优越性。

关键词:模糊支持向量机;模糊隶属度;混合核函数
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2010)02-0009-03

A New Fuzzy Support Vector Machine Based on Kernel Method

LI Lei[1], LU Yan-ling[1], ZHOU Meng-meng[1], BAI Yong-cheng[2]
(1. School of Sciences, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Software, University of Science and Technology of China, Hefei 230026, China)

Abstract: Support vector machine is sensitive to the noises and outliers in the training samples, so fuzzy support vector machine precede support vector machine in solving the problem of non-linearity, high dimension and uncertainty. The choice of fuzzy membership is the key and difficulty for fuzzy support vector. Generally, the fuzzy membership is established according to the distance of between sample points and its cluster center. A new fuzzy membership function is established, considering the relation among samples using mixed kernel function, based on mixed kernel function. The experiments show that fuzzy support vector machine with the new fuzzy membership is superior through comparing mixed kernel function with Polynomial kernel function and Gaussian RBF kernel function.

Key words: fuzzy support vector machine; fuzzy membership; mixed kernel function

0 引 言

支持向量机是在统计学习理论基础上建立的一种新的学习机,它根据有限的样本信息,在模型的复杂性和学习能力之间寻求最佳折衷,以期获得比较好的泛化能力。

支持向量机通过结构风险最小化原理来提高泛化能力,将求解最优分类面问题转化为解凸二次规划问题,较好地解决了小样本、非线性、高维数、局部极小点等实际问题。支持向量机已被广泛应用于各个领域,例如,文本分类、光学字符识别、人脸识别[1]、入侵检测、语音识别、模式识别[2]。

在支持向量机中,样本空间中的大部分输入数据不能被线性分类或近似线性分类时,应该通过非线性映射 φ,把样本空间映射到一个高维特征空间,并在特征空间中构造最优分类面。引入核函数代替特征空间中 φ 的点积。在低维空间中不能解决的问题,通过核函数的转换,在高维空间中可以被解决。进而,核函数是支持向量机解决非线性问题的关键。对改善支持向量机的性能,基于现有核函数创建新的核函数是一种重要而且有效的方法[3~5]。

支持向量机对样本中的噪声点和孤立点是非常敏感的,针对这种情况,Lin 等学者将模糊理论与支持向量机结合,提出了模糊支持向量机(FSVM)[6~9]。在模糊支持向量机[10]中,对每个样本赋入一个模糊隶属度。对于决策面的学习,不同的隶属度呈现不同的贡献。模糊隶属度的确定是模糊支持向量机的难点,目前没有确定的建立隶属度函数的方法。

收稿日期:2009-06-09;**修回日期**:2009-09-10
基金项目:国家自然科学基金项目(10371106,10471114);江苏省高校自然科学基金项目(04KJB110097,08KJB520003);南京邮电大学攀登计划(NY207064)
作者简介:李 雷(1958-),男,教授,研究方向为智能信号处理、非线性分析与计算智能。

1 模糊支持向量机

给定模糊训练样本集

$$S = \{(x_1, y_1, s_1), (x_2, y_2, s_2), \cdots, (x_l, y_l, s_l)\} \tag{1}$$

其中 $x_i \in R^n$，为 n 维向量，$y_i \in \{-1, 1\}$ 为相应的二类划分，$0 \leqslant s_i \leqslant 1$ 为模糊隶属度。

假设 φ 是将原始空间映射到高维特征空间的非线性映射。在特征空间中利用结构风险最小化原理和分类间隔最大化思想，求解最优分类超平面问题可转化为下面的最优化问题：

$$\begin{aligned} &\text{Minimize } \varphi(w, \xi) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} s_i \xi_i \\ &\text{Subject to } y_i[(w \cdot x_i) + b] \geqslant 1 - \xi_i \\ &\xi_i \geqslant 0, i = 1, 2, \cdots, l \end{aligned} \tag{2}$$

其中 w, b 分别是分类超平面的权值和偏值，ξ_i 是非负松弛变量，$C > 0$ 是自定义的惩罚系数，保持分类最大间隔与分类误差之间的平衡。模糊隶属度 s_i 表示相应点 x_i 的价值，较小的 s_i 可以减小 ξ_i 在式中的影响，以致将相应的 x_i 看作不重要的样本。

为求解上述约束最优化问题，引入 Lagrange 函数

$$L = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{l} s_i \xi_i - \sum_{i=1}^{l} \alpha_i [y_i (x_i \cdot w + b) - 1 + \xi_i] - \sum_{i=1}^{l} \beta_i \xi_i \tag{3}$$

其中 $\alpha_i \geqslant 0, \beta_i \geqslant 0$ 是 Lagrange 系数。将 L 分别对 w、b、ξ_i 求偏微分并令其等于 0，就可以把最优化问题 (2) 转化为等价的对偶规划问题

$$\begin{aligned} &\text{Maximize } W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ &\text{Subjectto } \sum_{i=1}^{l} \alpha_i y_i = 0 \\ &0 \leqslant \alpha_i \leqslant s_i C, i = 1, 2, \cdots, l \end{aligned} \tag{4}$$

其中 $K(x_i, x_j) = (\varphi(x_i), \varphi(x_j))$ 是核函数，并且满足 Mercer 定理。

如果 $\alpha_i > 0$，相应的点 x_i 是支持向量，这里有两种支持向量：一种满足 $0 < \alpha_i < s_i C$ 的支持向量 x_i 位于分类面附近，一种满足 $\alpha_i = s_i C$ 的支持向量 x_i 为错误分类样本。

相应的最优决策函数为

$$f(x) = \text{Sgn}((w \cdot x) + b) = \text{Sgn}\left(\sum_{i=1}^{l} y_i \alpha_i K(x_i \cdot x) + b\right) \tag{5}$$

其中

设 α^* 是式 (5) 的最优解，w^*、b^* 分别为相应的权值和偏值。 (6)

2 核函数

支持向量机在处理非线性分类问题时，引入核函数，巧妙解决了高维空间中的内积运算，避免了由非线性映射造成的维数问题。目前研究最多的核函数如下：

多项式核函数 $K(x, x') = [(x \cdot x') + 1]^q$ (7)

高斯 RBF 核函数 $K(x, x') = \exp\left\{-\frac{\| x - x' \|}{2\sigma^2}\right\}$ (8)

Sigmoid 核函数 $K(x, x') = \tanh(\alpha(x \cdot x') + \beta)$ (9)

其中，Sigmoid 核函数在 $\alpha > 0, \beta < 0$ 时较适合做核函数。

核函数是特征空间中非线性映射的点积，可以直接在原始空间中解决，而不必知道非线性变换 φ 的显式形式。支持向量机的性能主要取决于核函数。不同的传统核函数在分类中有不同的优点，依据 Mercer 定理，可以从简单的核构造更复杂的核。在这篇文章，利用传统核函数创建混合核函数：

混合核函数

$$K(x, x') = \lambda[(x \cdot x') + 1]^q + (1 + \lambda)\exp\left\{-\frac{\| x - x' \|}{2\sigma^2}\right\} \quad 0 < \lambda < 1 \tag{10}$$

3 基于核方法的一种新的模糊隶属度函数

在模糊支持向量机中，选择合适的隶属度函数是非常重要的。在许多实际应用中，训练样本点的作用是不同的。因此，对不同的样本应当引入不同的隶属度函数。对孤立点和噪声点赋较小的隶属度，以便减小它们的影响。在这里，利用混合核函数创建了一种新的模糊隶属度函数。

将训练样本分为两类：正类 C^+ 和负类 C^-。

$$\begin{aligned} C^+ &= \{x_i \mid x_i \in S' \text{ 且 } y_i = +1\} \\ C^- &= \{x_i \mid x_i \in S' \text{ 且 } y_i = -1\} \end{aligned} \tag{11}$$

显然 $C = C^+ + C^-$。设两类样本中心分别是 x^+、x^-，半径分别为 r_+、r_-。

$$\varphi(x^+) = \frac{1}{n_+} \sum_{x_i \in C^+} \varphi(x_i) \tag{12}$$

$$\varphi(x^-) = \frac{1}{n_-} \sum_{x_i \in C^-} \varphi(x_i) \tag{13}$$

其中 n_+, n_- 分别为正类 C^+ 和负类 C^- 的样本数。

$$r_+^2 = \max_{x_i \in C^+} \| \varphi(x_i) - \varphi(x_+) \|^2 = \max_{x_i \in C^+} d_{i+}^2 \tag{14}$$

$$r_-^2 = \max_{x_i \in C^-} \| \varphi(x_i) - \varphi(x_-) \|^2 = \max_{x_i \in C^-} d_{i-}^2 \tag{15}$$

进而得到两类中心的距离及正(负)类样本点到正(负)类中心的距离分别为 d_{i+}^2、d_{i-}^2。

$$D = \| \varphi(x_+) - \varphi(x_-) \|^2 = \frac{1}{n_+^2} \sum_{x_l, x_m \in C^+} K(x_l, x_m) + \frac{1}{n_-^2} \sum_{x_l, x_m \in C^-} K(x_l, x_m) - \frac{2}{n_+ n_-} \sum_{x_l \in C^+} \sum_{x_m \in C^-} K(x_l, x_m) \tag{16}$$

$$d_{i+}^2 = \| \varphi(x_i) - \varphi(x_+) \|^2 = K(x_i, x_i) + \frac{1}{n_+^2} \sum_{x_l, x_m \in C^+} K(x_l, x_m) - \frac{2}{n_+} \sum_{x_m \in C^+} K(x_i, x_m) \tag{17}$$

$$d_{i-}^2 = \| \varphi(x_i) - \varphi(x_-) \|^2 = K(x_i, x_i) + \frac{1}{n_+^2} \sum_{x_l, x_m \in C^+} K(x_l, x_m) - \frac{2}{n_+} \sum_{x_m \in C^+} K(x_i, x_m) \tag{18}$$

公式(16)－(18)均采用混合核函数。

因此,模糊隶属度函数构造如下:

$$s_i^+ = \begin{cases} 0.6 * \dfrac{d_{i+}^2}{r_+^2} + 0.4 & d_{i+}^2 \leqslant D \cdot \varepsilon \\ 0.4 * \dfrac{1}{1 + [r_+^2 - d_{i+}^2]} & d_{i+}^2 > D \cdot \varepsilon \end{cases} \tag{19}$$

$$s_i^- = \begin{cases} 0.6 * \dfrac{d_{i-}^2}{r_-^2} + 0.4 & d_{i-}^2 \leqslant D \cdot \varepsilon \\ 0.4 * \dfrac{1}{1 + [r_-^2 - d_{i-}^2]} & d_{i-}^2 > D \cdot \varepsilon \end{cases} \tag{20}$$

其中 $\varepsilon > 0$ 是半径控制因子。

模糊隶属度的建立不仅可以根据样本到类中心的距离,而且可以根据样本间的密切度建立。设两个不同样本点的密切度为 d_{ij}:

$$d_{ij} = K(x_i, x_j) = \lambda[(x_i \cdot x_j) +]^q + (1 - \lambda)\exp\left\{-\frac{\| x_i - x_j \|}{2\sigma^2}\right\} \tag{21}$$

其中 $i, j \in l, i \neq j$。设 $D = \max(d_{ij})$。

考虑样本间的密切度创建的模糊隶属度为 μ_i:

$$\mu_i = \begin{cases} \mu_i^+ = \dfrac{d_{ij}}{D} & \text{where } x_i \in C^+, x_j \in C \\ \mu_i^- = \dfrac{d_{ij}}{D} & \text{where } x_i \in C^-, x_j \in C \end{cases} \tag{22}$$

最后,将以上两种隶属度结合建立一种新的模糊隶属度:

$$c_i = \begin{cases} c_i^+ = 0.5 * s_i^+ + 0.5\mu_i^+ & \text{where } x_i \in C^+ \\ c_i^- = 0.5 * s_i^- + 0.5\mu_i^- & \text{where } x_i \in C^- \end{cases} \tag{23}$$

其中 s_i^+, s_i^- 在公式(19)及公式(20)中得出,μ_i^+, μ_i^- 在公式(23)中得出。

4　实验结果

为了测试新的模糊隶属度的效果,从机器学习库的UCI知识库选取的生物医学数据,为二类分划问题,包含72个探测点和7129个特性。在此数据库中,共有1000个样本点,其中750个样本点属于正类样本,250个样本点属于负类样本。选择678个样本点作为训练集,322个样本点作为测试集。

利用Gauss RBF核函数训练并选择最优惩罚因子 C。模糊隶属度 μ_i 中采用混合核函数,在模糊隶属度 s_i 中分别采用多项式核函数、Gauss RBF核函数、混合核函数。表1为采用不同隶属度函数的FSVM的实验结果。

表1　采用不同隶属度函数的FSVM的实验结果

模糊隶属度 s_i 中采用的核	训练误差	测试误差
Poly 核函数	0.00713756	0.03031021
Gauss RBF 核函数	0.00604629	0.02835817
混合核函数	0.00348163	0.01369150

实验结果表明:混合核函数及乘积核函数的性能高于传统核函数。

5　结束语

在这篇文章中,建立的隶属度函数不仅与样本到类中心的距离有关,还与样本间的密切度有关,在隶属度函数建立过程中选择使用了混合核函数。对不同的样本如何选择合适的隶属度函数是及其重要的。提出的FSVM解决了由传统SVM引起的不能划分的区域。在理论上,混合核函数的泛化能力优越于传统核函数。通过利用人工数据及生物医学数据进行计算机仿真,确实比传统核函数的效果好。

在将来,将利用新的隶属度函数讨论多分类问题,选择其他核函数及隶属度函数进行比较。

参考文献:

[1] Zhu Shuxian, Zhang Renjie. Research for Face Recognition Base on Mixed Kernel Function[M]. China:[s. n.], 2008: 1395－1399.

[2] Hao Tang, Liang Sheng. Fuzzy Support Vector Machine With a New Fuzzy Membership Function for Pattern Classification [C]//Proceedings of the Seventh International Conference on Machine Learning and Cybernetics. Kunming:[s. n.], 2008: 768－773.

[3] Xia Hong. Feature Selection based on Fuzzy SVM[C]//Fifth International Conference on Fuzzy Systems and Knowledge Discovery. Jinan, China:[s. n.], 2008:586－589.

[4] Yang Chih－Cheng, Lee Wan－Jui, Lee Shie－Jue. Learning of Kernel Functions in Support Vector Machines[C]//2006 International Joint Conference on Neural Networks. Sheraton Vancouver Wall Centre Hotel. Vancouver, BC, Canada:[s. n.], 2006:1150－1155.

[5] Tan Ying, Wang Jun. A Support Vector Machine with a Hy-

(下转第15页)

```
public void StringSearch(String keyword, String searchDir)
{.....}
    Date endTime = new Date();
    long timeOfSearch = endTime.getTime() - beginTime.
getTime();}
```

先给出索引的路径,然后再提供要搜索的 Field 和搜索的关键字作为参数,最后生成一个查询对象即可开始查询。

通过这个实例,可以了解检索的一般流程,其它功能的复杂的检索只是对具体方法的细化和完善,但都会有上面的三个部分。

下面对实验结果做简要说明。

3.2 实验结果说明

通过对两个文档的检索,来对比应用 Lucene 进行检索和传统检索方式的效率。第一篇文档字数为 50 万左右,检索目标词汇,前者用时 120ms,后者用时 389ms,大约是前者的三倍。当文档字数增加到 2500 万字左右时,即用第二篇测试文档进行,采用 Lucene 的系统用时大约 300ms,而传统检索方法用时接近 4000ms。采用 Lucene 的检索明显优于字符串匹配方式。由此可以想象当数据量达到 TB 级别时,传统检索方式的检索速度将是无法忍受的。

以上的测试文档均为 txt 格式,其实 Lucene 并没有规定数据源的格式,它提供了一个通用的结构来接受索引的输入,因此输入的数据源可以是数据库、Word 文档、PDF 文档、HTML 文档,只要能够设计相应的解析转换器将数据源构造成 Document 对象即可进行索引。

该实例默认支持 txt 形式文档的索引和检索。如因实际应用需要,还可以利用第三方工具如 POI,XPDF 和 Jacob 包对 Excel,Word 和 PDF 进行操作,只要将数据源构造成一个 Document 对象,就可以实现对其的检索支持。

4 结束语

文中讲述了全文检索技术,描述了 Lucene 的结构和特点,提出了一种解决中文全文检索的方法,可以应用在中小企业网站站内检索,个人用户桌面搜索引擎建立,特定文档检索数据库建立等方面。Lucene 所独有的增量索引技术使其能够被大部分的项目成功应用,相比传统的检索方式在效率上有很大的改善,同时用户可以根据自己的需要构建个性化的搜索,从而摆脱对 Google,baidu 站内搜索的依赖。实现对目标文档检索和管理,降低中小网站的运营成本。个人如果有兴趣,还可以结合一些开源的爬虫开发自己的搜索引擎,这也是下一步将开展的工作。

参考文献:

[1] 孙西全,马瑞芳,李燕灵.基于 Lucene 的信息检索的研究与应用[J].情报理论与实践,2006,29(1):521-528.

[2] Gospodnetic O,Hatcher E. Lucene in action[M].[s.l.]:Manning Publications Co,2005.

[3] 林碧英,赵　锐,陈良臣.基于 Lucene 的全文检索引擎研究与应用[J].计算机技术与发展,2007,17(5):186-190.

[4] 索红光,孙　鑫.基于 Lucene 的中文全文检索系统的研究与设计[J].计算机工程与设计,2008,29(19):5083-5086.

[5] 邱　哲,符滔滔.开发自己的搜索引擎 Lucene2.0+Hertrix[M].北京:人民邮电出版社,2007.

[6] 朱学昊,王儒敬,余锋林,等.基于 Lucene 的站内搜索设计与实现[J].计算机应用与软件,2008,25(10):6-8.

[7] 郎小伟,王申康.基于 Lucene 的全文检索系统研究与开发[J].计算机工程,2006,32(4):95-99.

[8] Zhang Yuletide,Zhang Tao,Chen Shijie. Research on Lucene-based English-Chinese cross-language information retrieval[J]. Journal of Chinese Language and Computing,2005,15(1):25-32.

(上接第 11 页)

brid Kernel and Minimal Vapnik-Chervonenkis Dimension [J]. IEEE Transactions on Knowledge and Data Engineering, 2004,16(4):385-395.

[6] Czajkowska J,Rudzki M,Czajkowski Z. A New Fuzzy Support Vectors Machine for Biomedical Data Classification[C]//30th Annual International IEEE EMBS Conference. Vancouver, British Columbia, Canada:[s.n.],2008:4476-4479.

[7] TSANG E C C,YEUNG D S,CHAN P P K. Fuzzy Support Vector Machines for Solving Two-class problems[C]//Proceedings of the Second International Conference on Machine Learning and Cybernetics. Xi'an:[s.n.],2003:1080-1083.

[8] Li Xuehua,Shu Lan. Fuzzy Theory Based on Support Vector Machine Classifier[C]//Fifth International Conference on Fuzzy Systems and Knowledge Discovery. Jinan, China:[s.n.],2008:600-604.

[9] Lin C F,Wang S D. Fuzzy Support Vector Machines[J]. IEEE Transactions on Neural Networks,2002,13(12):466-471.

[10] Soria E, Martin J, Camps G, et al. A low complexity fuzzy activation function for artificial neural networks[J]. IEEE Trans Neural Networks,2003,14(6):1576-1579.