

基于相似度量的模糊支持向量机算法研究

王宇凡, 梁工谦, 张淑娟

(西北工业大学 管理学院, 陕西 西安 710025)

摘 要: 基于统计学习理论,核被看做是一种相似度测量模型.核函数是支持向量机算法的核心,利用核函数可以将低维不可分数据映射到高维空间,并进行最优分类研究.但孤立点或噪声数据都会影响最优分类平面和最优分类函数,所以提出利用相似度测量构建模糊核函数.相比高斯核函数和模糊 sigmoid 核函数的分类支持向量模型,本文提出的模糊相似核函数在支持向量机运算中计算成本最低,可以提供更高的准确率,同时可以避免传统模糊核函数的限制.

关键词: 相似测量;核函数;模糊集;支持向量机

中图分类号: TN911.7

文献标识码: A

文章编号: 1000-7180(2014)04-0112-05

Support Vector Machine with Fuzzy Kernel Function Based on Similarity Measure

WANG Yu-fan, LIANG Gong-qian, ZHANG Shu-juan

(College of Management, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Based on the statistical learning theory, kernels are often presented as measures of similarity measure. Kernels play a very important role in support vector machine (SVM) algorithms, which can mapper the unclassification data to high dimensional space, to get the optimal classification result. Unfortunately, SVM lacks ability to deal with the system or data with ambiguous characters. So this paper proposes to SVM with fuzzy kernel, which is obtained by similarity measure method. Compared to Gaussian RBF kernel function and fuzzy sigmoid kernel function, SVM with fuzzy similarity kernel gets more accuracy results with less computation requirements, and alleviates current limitation of the fuzzy kernel function.

Key words: Similarity measures; Kernel function; Fuzzy set; Support vector machine

1 引言

支持向量机 (Support Vector Machine, SVM) 是 Cortes 和 Vapnik 于 1995 年首先提出的机器学习方法^[1]. 该方法是建立在统计学习理论的 VC 维理论上,利用核函数理论,将输入不可分向量映射到一个高维特征空间,并在高维空间内寻求结构风险最小化的最优分类超平面.该方法在解决小样本、非线性及高维模式识别中表现出许多特有的优势,因此被广泛的应用,如图像识别,非线性样本分类,语音识别和空间数据分析等研究中^[2-4].

尽管支持向量机具有良好的小样本分析能力能力,但对于数据集中的孤立点或噪声数据非常敏感,其分类最优超平面或决策函数将受到这些噪声数据和异常数据的影响较为明显.模糊集理论可以用于处理不确定性、非线性和不适定的问题^[5]. Lin 等学者^[6-8],提出了模糊支持向量机 (FSVM) 理论,将模糊集理论应用于 SVM 中,利用核函数和惩罚系数对噪声或异常数据进行处理,从而减小或消除其对分类最优超平面或决策函数的影响.现有研究多集中于 FSVM 惩罚系数 (隶属度函数) 的研究,对于模糊核函数构建研究相对较少^[9-11].因此,有效地解决

收稿日期: 2013-05-28; 修回日期: 2013-07-20

基金项目: 国家自然科学基金 (71001085)

模糊核函数的构建的问题是本文研究的出发点和目标。

根据泛函相关理论,当某函数满足 Mercer 条件,并对应低维空间中的向量内积时,其高维特征空间映射函数可以称为核函数^[12]。高维特征空间的映射反应输入样本之间的相似度(即距离),利用相似度测量可以很好的构建模糊核函数,有效的减弱孤立点或噪声数据对分类精度的影响,同时减少了支持向量的数目,在处理分类问题中起到了较好的效果。

本文利用相似测量理论,研究模糊集相似度,并构建适用于模糊支持向量机的模糊相似核函数,将其用于分类支持向量机研究。对比运用传统高斯核函数和模糊 sigmoid 核函数的分类支持向量研究,其数值验证结果表明该方法对于模糊集数据处理方面具有更好的精准率和综合表现。

2 模糊集相似性度量

定义 1 设 $F_0(x)$ 为有限论域 x 上的模糊集 $F(x)$ 的子集,对于任意的两个集 $A, B \in F_0(x)$,有数 $S(A, B)$ 对应,且满足条件:

$$0 \leq S(A, B) \leq 1, S(A, A) = 1;$$

$$S(A, B) = S(B, A);$$

$$A \in B \in C \text{ 时}, S(A, C) \leq S(A, B)$$

则称 S 为 $F_0(x)$ 有限论域的一个相似度^[13]。满足上述基本条件的相似度计算方法很多^[14-15]。假设 A 和 B 为有限论域 U 上的模糊子集, $x_i \in U$, $\mu_A(x_i)$ 和 $\mu_B(x_i)$ 分别为 A 和 B 的隶属函数,常见的模糊相似测量方法如下所示。

欧几里得相似度:

$$S(A, B) = 1 - \sqrt{\frac{1}{n} \sum_{i=1}^n (\mu_A(x_i) - \mu_B(x_i))^2}$$

Hsieh 模糊相似度:

$$S(A, B) = 1 - \frac{1}{1 + \frac{1}{2} \|\mu_A(x_i) - \mu_B(x_i)\|^2};$$

最大最小值相似度:

$$S(A, B) = \left| \frac{\min(A, B)}{\max(A, B)} \right| = \frac{A \cap B}{A \cup B}.$$

模糊集合的相似度计算公式很多,本文利用最大最小值相似度讨论和构建模糊核函数。

定义 2 假设 A 为一个广义三角模糊数,三角模糊数用三元数组 (L, M, R) 表示, M 为三角模糊数的中值, L 和 R 分别为三角模糊数的左,右边界值。

$\mu(A)$ 为模糊数隶属函数。

定理 1 给定三角模糊集 A , 存在 p 使得三角模糊数左右边界点分别表示为 $((1-p)m/p, (1+p)m/p)$, 其中 m 为模糊集中值。图 1 显示了三角模糊数,其三角形隶属度函数如下:

$$\mu(A) =$$

$$\begin{cases} \frac{p}{m} \left| x - \frac{(p-1)}{p} m \right|, & \frac{(p-1)}{p} m \leq x \leq \frac{(p+1)}{p} m \\ 0, & \text{otherwise} \end{cases}$$

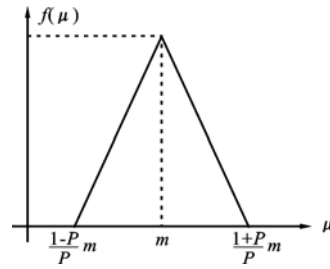


图 1 三角模糊数

采用三角模糊数描述本体映射相似度的过程中,本文对相关三角模糊数的左值(L),中值(M)和右值(R),分别进行了限定,并提出了适用于模糊相似度描述的三角模糊相似度。

定义 3 三角模糊相似度同样采用一个三元组 (L, M, R) 来表示, M 为三角模糊数的中值, L 和 R 分别为三角模糊数的左值和右值。 L, M 和 R 的取值范围为 $0 \leq L \leq M \leq R \leq 1$, $\mu(L)$ 和 $\mu(R)$ 模糊数隶属函数, $\mu(L) = \mu(L) = 0, \mu(M) = f$ 。给定任意实数 $x, \mu_L(x) = \mu_R(x) = f$, 表明其三角模糊数相似结果可能性最大。当 $L \leq x \leq M$ 时, $\mu_L(x)$ 较大, 表明其三角模糊数左侧相似可能性越大; 当 $L \leq x \leq R$ 时, $\mu_R(x)$ 较大, 表明其三角模糊数右侧相似可能性越大。

假设,三角模糊集 A 和 B , 其模糊集中值分别为 $M(a)$ 和 $M(b)$, 不妨设 $A \leq B$ 。存在 p 使得三角模糊数左右边界点分别表示为 $((1-p)a/p, (1+p)a/p)$ 和 $((1-p)b/p, (1+p)b/p)$ 。则三角模糊集 A 和 B 的存在四种条件下的相似度测量,如图 2 所示。

根据图 2 所示,三角模糊集 A 和 B 存在交点 $(2ab/a+b, 1-(b-a)p/(a+b))$ 。前三种相交情况,其对应的模糊条件分别为 $(p-1/p+1 \leq a/b \leq p-1/p), (p-1/p \leq a/b \leq p/p+1)$ 和 $(p/p+1 \leq a/b \leq 1)$, 其三角模糊集 A 和 B 相似度公式可以表示为

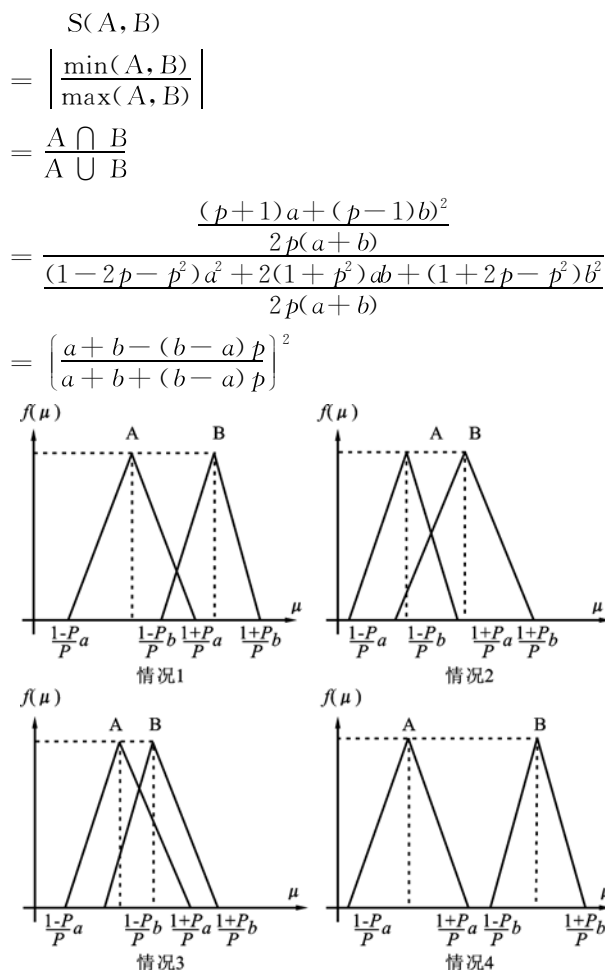


图2 不同情况下三角模糊集相似度

第四种情况 ($a/p-1 \leq b/p+1$), A 和 B 不存在相交和相并, 所以其三角模糊集 A 和 B 相似度公式可以表示为

$$S(A, B) = \left| \frac{\min(A, B)}{\max(A, B)} \right| = \frac{A \cap B}{A \cup B} = 0.$$

根据上述讨论, 三角模糊集 A 和 B 的相似度量函数可以表示为

$$S(A, B) = \begin{cases} \left(\frac{a+b-(b-a)p}{a+b+(b-a)p} \right)^2, & \frac{p-1}{p+1} \leq \frac{a}{b} \leq 1 \\ 0, & \frac{a}{b} \leq \frac{p-1}{p+1} \end{cases}$$

根据定义 1, 可以证明公式符合模糊相似性的定义.

3 模糊相似支持向量构建

支持向量机是基于统计学习理论构建, 解决小样本, 非线性及高维模式问题的机器学习方法. 当研究样本中存在孤立点或噪声数据时, 其研究准确性将受到一定的影响. 因此, 本文在模糊集相似度测量

的基础上, 构建模糊相似核函数, 利用该核函数, 对存在孤立点或噪声数据的支持向量机分析得到最优结果.

3.1 支持向量机概述

给定训练样本集 $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, 其中 $x_n \in R^N$ 是输入指标向量, y_n 为样本标签, 为输出指标.

假设 ϕ 是从低维向量空间中映射到高维特征空间的非线性映射. 在高维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, 最优分类超平面的优化问题的二次优化问题可归结为

$$f(w, b, \epsilon) = \min \left\{ \frac{1}{2} \|w\|^2 + c \sum_{i=1}^l \epsilon_i \right\}$$

$$\text{s. t. } y_i \{ (w \cdot \phi(x_i)) + b \} \geq 1 - \epsilon_i$$

$$\epsilon_i \geq 0, i = 1, 2, \dots, l$$

式中, w, b 表示是分类超平面的权值和偏值, ϵ_i 为非负松弛变量, C 为模型的惩罚参数.

运用拉格朗日函数, 二次优化问题可以用对偶形式计算. 根据泛函的有关理论, 只要一种核函数满足 Mercer 条件, 其对应着某一变换空间中的内积^[16]. 公式转换为

$$\max - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j K(x_i, x_j) + \sum_{i=1}^n a_i$$

$$\text{s. t. } \sum_{i=1}^n a_i y_i$$

$$0 \leq a_i \leq C \quad i = 1, 2, \dots, l.$$

式中, $K(x_i, x_j)$ 为满足 Mercer 条件的核函数. 于是可得到最优分类面的决策函数为

$$f(x) = \text{sign} \left(\sum_{i=1}^n a_i y_i K(x, x_i) + b \right).$$

上述分析可以看出, 支持向量机是利用核函数将低维空间映射到高维空间, 可以得到最优的分类平面和决策函数的影响.

3.2 模糊相似核函数的构建

核函数的选取是支持向量机最优超平面或决策函数计算的关键. 对于存在孤立点或噪声数据的研究样本, 根据模糊集理论, 利用模糊相似度测量, 构建模糊相似核函数, 是模糊相似支持向量机的关键. 现有模糊支持向量机多采用传统高斯和 sigmoid 等核函数进行计算, 但其最优超平面或决策函数计算结果仍存在一些限制. 因此, 本文构建使用模糊三角集的相似核函数.

给定模糊向量:

$\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$, 其中 $\tilde{x}_i = (L, M, R)$ 为模

糊三角集, M 为三角模糊数的中值, L 和 R 分别为三角模糊数的左值和右值。

假设三角模糊集 $\tilde{T} = (\tilde{L}, \tilde{M}, \tilde{R})$, 存在 P 使得模糊向量转化为 $\tilde{x} = (\tilde{x}_i, \tilde{x}_i/p)$, 其中 $i = 1, 2, \dots, n$. 对于两组三角模糊集 $\tilde{T} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n)$ 和 $\tilde{T}^j = (\tilde{x}^1, \tilde{x}^2, \dots, \tilde{x}^j)$, 其相似度测量表示为

$$S(\tilde{T}, \tilde{T}_i) = \frac{1}{n} S(\tilde{x}_i, \tilde{x}^j).$$

根据三角模糊集相似度量公式, $\tilde{x}_i$ 和 $\tilde{x}^j$ 相似度计算表现为如下:

$$S(\tilde{x}_i, \tilde{x}^j) = \begin{cases} \left[\frac{\tilde{x}_i + \tilde{x}^j - (\tilde{x}^j - \tilde{x}_i)p}{\tilde{x}_i + \tilde{x}^j + (\tilde{x}^j - \tilde{x}_i)p} \right]^2, & \text{if } \frac{p-1}{p+1} < \frac{\tilde{x}_i}{\tilde{x}^j} \leq 1 \\ \left[\frac{\tilde{x}_i + \tilde{x}^j - (\tilde{x}_i - \tilde{x}^j)p}{\tilde{x}_i + \tilde{x}^j + (\tilde{x}_i - \tilde{x}^j)p} \right]^2, & \text{if } \frac{p-1}{p+1} < \frac{\tilde{x}^j}{\tilde{x}_i} \leq 1. \\ 0, & \text{if } \frac{\tilde{x}_i}{\tilde{x}^j} \leq \frac{p-1}{p+1} \text{ or } \frac{\tilde{x}^j}{\tilde{x}_i} \leq \frac{p-1}{p+1} \end{cases}$$

从上述公式可知 $0 \leq S(\tilde{x}_i, \tilde{x}^j) \leq 1$, 进而 $0 \leq S(\tilde{T}, \tilde{T}_i) \leq 1$.

构建模糊相似核函数为

$$K(x, x_i) = \frac{1}{n} \left| \frac{x + x_i - (x_i - x)p}{x + x_i + (x_i - x)p} \right|^2.$$

根据统计学习原理, 满足 Mercer 条件, 则可作为变换空间的内积使用, 即可作为核函数. 由定理可以很好的得出如下公式, 于是 $k(x, x_i)$ 可以作为一个核函数:

$$\iint_{\forall j \in I_2} K(x, x_i) f(x) f(z) dx dz \geq 0.$$

4 数值例子

为了验证本文算法的性能和有效性, 利用 UCI 机器学习数据库中的数据验证提出的方法^[17]. 五个数据集被利用验证算法, 包括: Wine、Iris、Ionosphere、Liver-disorders、Pima-diabetes, 数据描述如表 1 所示. 选择前 20% 的数据集作为训练样本, 剩余的样本集作为测试样本。

同时, 利用高斯核函数(高斯核)和模糊 sigmoid 核函数(模糊核)比较分析本文算法的优越性. 模糊 sigmoid 核函数支持向量机研究结果来自文献[18]. 本文构建模糊相似核函数(相似核)支持向量机和高斯核函数支持向量机算法是用 Matlab2011B 语言实现的, 实验所采用的计算机配置为: Intel I5 CPU, 主频 2.66 GHz 2.26 GHz, 内存 4 GB. 高斯核函数和模糊 sigmoid 核函数分别表示为如下公式:

高斯核函数:

$$K(x, x_i) = \exp(-\|x - x_i\|^2 / 2\sigma^2).$$

模糊 sigmoid 核函数:

$$K(x, x_i) = \tanh(\gamma x_i^T x + \alpha).$$

参数选取决定支持向量机, 及其核函数运算优劣的关键. 参数的选择采用 8 折交叉验证法. 初始参数选择如表 2 所示, 其参数包括模糊惩罚参数 C , 高斯核函数参数 σ , 相似性核函数参数 p .

表 1 数据集描述

数据集	样本数	属性	类别
Wine	178	13	3
Iris	150	4	3
Ionosphere	351	34	2
Liver-disorders	345	6	2
Pima-diabetes	768	8	2

表 2 模型初始参数

	C	σ	p
高斯核	$10^{-2}, \dots, 10^3$	$10^{-3}, \dots, 10^2$	/
模糊核	$10^{-2}, \dots, 10^3$	/	1, 2, $\dots$, 9

为了更好地对比算法结果, 采用平均绝对误差 (MAP) 和模型学习运算时间 (CPU) 来衡量各个算法的误差大小, 平均绝对误差即:

$$\text{MAP} = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right|,$$

式中, $\hat{y}_i$ 为算法预测值, y_i 为模型实际值, n 为样本个数。

实验结果如表 3 所示. 从实验结果可以看出, 本文算法在所有数据集上的分类精度均为最高, 优于普通的高斯核函数和模糊 sigmoid 核函数. 与传统的 SVM 相比, 本文算法可以大大缩短训练时间, 表 4 反映了其运算速度. 可见其对三角模糊集分析研究中, 本文构建的算法具有一定的优势。

表 3 UCI 数据集研究结果比较

数据集	相似核		模糊核		高斯核	
	MAP/ %	CPU/ s	MAP/ %	CPU/ s	MAP/ %	CPU/ s
Wine	3.51	2.29	3.73	2.89	3.29	4.28
Iris	3.17	5.04	3.31	4.90	2.83	6.47
Iono	2.74	26.32	2.83	29.41	2.48	33.83
Liver	2.28	9.64	2.30	10.47	1.99	12.64
Pima	1.02	16.70	1.04	18.66	0.97	24.35

注: Iono、Liver 和 Pima 分别表示 Ionosphere Liver-disorders

和 Pima-diabetes 数据集.

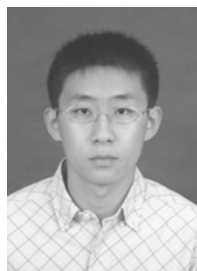
5 结束语

本文在模糊相似度测量基础上,构建适用于支持向量机算法的模糊相似核函数的.通过实验数据证明此方法,相比利用高斯核函数和模糊 sigmoid 核函数的支持向量机算法,该方法提供更高的准确率,其运算速度和运算成本更低.这一方法突破了传统核函数寻找最优分析的瓶颈.将此模糊相似支持向量机应用与工程研究领域是未来笔者研究的方向.

参考文献:

- [1] Vapnik V, Levin E, Lecun Y. Measuring the Vc-Dimension of a Learning-Machine [J]. Neural Computation, 1994,6(5):851-876.
- [2] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995,20(3):273-297.
- [3] Dibike Y B, Velickov S, Solomatine D, et al., Model induction with support vector machines: Introduction and applications [J]. Journal of Computing in Civil Engineering, 2001,15(3):208-216.
- [4] Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines [J]. Machine Learning, 2002,46(1-3):131-159.
- [5] Luo Q, Yang W Q, Yi D Y. Kernel shapes of fuzzy sets in fuzzy systems for function approximation [J]. Information Sciences, 2008,178(3):836-857.
- [6] Lin C J, Chen C H. A self-constructing compensatory neural fuzzy system and its applications [J]. Mathematical and Computer Modelling, 2005, 42 (3-4): 339-351.
- [7] Lin C J, Ho W H. An asymmetry similarity measure based neural fuzzy inference system [J]. Fuzzy Sets and Systems, 2005,152(3):535-551.
- [8] Lin C F, Wang S D. Fuzzy support vector machines [J]. Ieee Transactions on Neural Networks, 2002,13 (2):464-471.
- [9] Demirci M. Fuzzy functions and their fundamental properties [J]. Fuzzy Sets and Systems, 1999,106 (2):239-246.
- [10] Celikyilmaz A, Turksen I B. Fuzzy functions with support vector machines [J]. Information Sciences, 2007,177(23):5163-5177.
- [11] 唐浩, 廖与禾, 孙峰, 等, 具有模糊隶属度的模糊支持向量机算法 [J]. 西安交通大学学报, 2009,43(07): 40-43.
- [12] Vert J P, Tsuda K, Scholkopf B. Kernel Methods in Computational Biology [M]. Cambridge, The MIT Press: MA, 2004.
- [13] Wang W J. New similarity measures on fuzzy sets and on elements [J]. Fuzzy Sets and Systems, 1997,85 (3):305-309.
- [14] Beg I, Ashraf S. Similarity measures for fuzzy sets [J]. Applied and Computational Mathematics, 2009,8 (2):192-202.
- [15] Lee S H, Park W J, Jung D Y. Similarity measure design and similarity computation for discrete fuzzy data [J]. Journal of Central South University of Technology, 2011,18(5):1602-1608.
- [16] (美)瓦普尼克. 统计学习理论的本质 [M]. 张学工, 译. 北京:清华大学出版社, 2000.
- [17] UCI [EB/OL]. [2013-04-20]. <http://mllearn.ics.uci.edu/>.
- [18] Camps-Valls G, Martin J D, Soria-olivas, et al. Fuzzy sigmoid kernel for support vector classifiers [J]. Neurocomputing, 2004(62):501-506.

作者简介:



王宇凡 男,(1983—),博士研究生. 研究方向为质量管理、设备管理.