基于图的组合半监督 SVM 聚类核算法研究

郑文静,李 雷
(南京邮电大学 理学院,江苏 南京 210023)

摘 要:为了在聚类假设的基础上,进一步提高支持向量机的分类精度,文中通过引入线性分段转换函数,将加权无向图上的相似矩阵重新表示,改变该图上的距离度量,使得在同一群集中两点间的距离更小,从而建立基于图的聚类核,与多项式核函数线性组合后,构造出基于图的组合半监督聚类核,并将其用于支持向量机的训练和分类。实验表明,与标准 SVM 算法相比,该算法分类精度较高,且高于组合前的单个核函数。随着标记样本比例的增加,该算法的分类精度也在增加,有效利用了未标记样本蕴含的信息。

关键词:半监督支持向量机;聚类核;图;分类
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2014)05-0109-04
doi:10.3969/j.issn.1673-629X.2014.05.026

Research on Combined Semi-supervised SVM Cluster Kernel Algorithm Based on Graph

ZHENG Wen-jing, LI Lei
(School of Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: In order to further improve the classification accuracy of SVM based on the cluster assumption, represent the similarity matrix of the weighted undirected graph by linear-step transfer function to establish a cluster kernel based on graph, which alters the map distance metric, so the distance between two points in the same cluster is smaller. Combining it linearly to the polynomial kernel function, a combined cluster kernel for semi-supervised SVM based on graph is constructed. Then train support vector machine with it and obtain the classification accuracy. Experiments show that, compared with the standard SVM algorithm, the classification accuracy of the proposed algorithm is higher, and better than the individual ones. With the increase in the proportion of labeled samples, the classification accuracy of this algorithm is also increasing, using the information of unlabeled samples effectively.

Key words: S^3VM; cluster kernel; graph; classification

0 引 言

随着计算机科学技术尤其是近些年网络通信技术的发展,海量的数据充斥在人们的日常生活之中。大量的未标记样本可以自动廉价地获取,而获取大量有标记的样本则相对较为困难。且对样本进行类别标记往往需要耗费大量的人力物力。而在模式识别、机器学习及相关领域中,传统的监督学习需要大量的类别标记数据来保证算法的泛化能力[1]。相对于传统方法,半监督学习既考虑了一定的标记数据信息[2],又结合了大量的未标记数据信息,进而建立更好的分类器,将半监督学习的思想引入到支持向量机学习算法就可以弥补标准 SVM 带来的缺陷,可以提高分类性能,获得更好的分类效果。

最初,Joachims 将半监督学习引入到支持向量机中,应用于文本分类,形成了直推式支持向量机(TSVM)[3],TSVM 在通常情况下也被认为是 Semi-Supervised SVM (S^3VM)。由于 S^3VM 同时利用已标记和未标记样本去最大化分类间隔,从而使得其目标函数是非凸的[4]。目前,S^3VM 的研究主要集中在非凸目标函数的优化上,主要有:梯度下降法(Gradient Descent)[5]、凹凸法(Convex-Concave Procedure)[6]、确定性退火方法(Deterministic Annealing)[7]、连续优化方法(Continuation Techniques)[8]、半定规划方法

收稿日期:2013-07-02 **修回日期**:2013-10-14 **网络出版时间**:2014-02-11
基金项目:国家自然科学基金资助项目(61070234,61071167)
作者简介:郑文静(1990-),女,江苏徐州人,研究方向为数据挖掘与智能计算;李 雷,博士,教授,研究方向为智能信号处理、非线性分析与计算智能。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20140211.1448.003.html

(Semi-definite Programming)[9] 等。这些算法主要解决的是 S^3VM 目标函数的非凸优化问题,需要反复迭代运算,计算复杂度较高,难以应用于大规模数据的分类[10]。且一般仅利用高斯核函数对样本进行表述,不能很好地满足半监督学习必须遵循的聚类假设。

而基于聚类假设,Chapelle 等提出聚类核概念[11],使用核函数,而不是明确的特征向量,重新表示给定的数据,以反映未标记数据透露的结构。文中通过引入线性转换函数,将加权无向图上的相似矩阵重新表示,使得在同一群集中两点间的距离更小,从而建立基于图的聚类核,与多项式核函数(Polynomial Kernels)线性组合后,构造出基于图的组合半监督聚类核,并将其用于支持向量机的训练和分类。实验表明,与标准 SVM 算法相比,该算法分类精度较高,且高于单独的核函数。随着标记样本比例的增加,该算法的分类精度也在增加,有效利用了未标记样本蕴含的信息。

1 聚类核

聚类假设[12]是指同一聚类中的样本点很可能具有相同的类别标签。也就是说,如果高密度区域某两个点可以通过区域内某条路径相连接,那么这两个样本点拥有相同标签的可能性就比较大。聚类核,使用的是核函数,而不是明确的特征向量,重新表示给定的数据,以反映未标记数据透露的结构。其主要思想是改变距离度量,使在同一群集中两点间的距离更小[13]。

Chapelle 提出了构造聚类核的整体框架,即通过修改核矩阵的能量本征谱构建聚类核,其中两个主要方法是随机游走核和谱聚类核。随机游走核是一个标准化和对称化随机游动过程的 t 步转移矩阵。根据 Szummer 和 Jaakkola[14] 的描述,可以通过把 RBF 核看作是在以 x_i 为顶点的图上的随机游动过程的转移矩阵,来定义一个随机游走表示,那么就可以使用一步转移矩阵计算 t 步随机游走核。谱聚类核的依据是谱聚类的思想,即通过对样本之间的相似度矩阵进行谱分解,对样本点在特征空间中进行重新表示,使得位于同一聚类或高密度区域中的样本点在新的空间中分布得更紧凑。不同的转换函数会导出不同的聚类核,主要有线性转换函数、分段转换函数、线性分段转换函数和多项式转换函数[11]。

2 基于图的复合半监督 SVM 聚类核算法

为了更好地满足聚类假设,文中提出基于图的复合半监督 SVM 聚类核算法(GPS^{3}VM),使用线性分段转换函数将图上的相似矩阵重新表示,建立基于图的聚类核,将其与多项式核函数进行线性组合,构成了满足 Mercer 条件的基于图的复合半监督聚类核。

2.1 算法思想

假设给定一个来自某未知分布的数据集 $X = \{x_1, x_2, \cdots, x_u\}$ $(u = m + n)$,其中 $X_m = \{x_1, x_2, \cdots, x_m\}$ 为给定的标记样本集合,即样本点 $x_i (i = 1, 2, \cdots, m)$ 的类别标签 y_i 是已知的,X_n 为未标记样本集合,即样本点 $x_i (i = m + 1, m + 2, \cdots, m + n)$ 的类别标签是未知的,$X = X_m \cup X_n$。假定所有的标记样本都是训练样本,所有的未标记样本都是测试样本。

与传统的 k-均值、EM 等聚类算法相比,建立在谱图理论上的谱聚类算法具有能在任意形状的样本空间上聚类且收敛于全局最优解的优点[15]。文中基于样本相似度,在所有标记和未标记样本间建立一个加权无向图 $G = (V, E, \boldsymbol{W})$,其中 V 表示所有标记和未标记的样本点 $x_i (i = 1, \cdots, n, n + 1, \cdots, n + m)$,$E$ 表示用以连接两节点的边,边的权重为 $\boldsymbol{W}$,通常用相似度或距离来度量。相似矩阵 $\boldsymbol{W}$ 为

$$W_{ij} = \begin{cases} e^{-q \times d(x_i, x_j)} & (i \neq j) \\ 1 & (i = j) \end{cases} \tag{1}$$

其中,x_i 和 x_j 表示两个样本点,$d(x_i, x_j)$ 取 $\| x_i - x_j \|^2$;q 为给定的控制参数。

根据聚类核一般算法,计算对角矩阵 $\boldsymbol{D}$,其元素是 $\boldsymbol{W}$ 的行和。即

$$D_{ii} = \sum_{j=1}^{n+m} W_{ij} \tag{2}$$

计算 $\boldsymbol{L} = \boldsymbol{D}^{-1/2} \boldsymbol{W} \boldsymbol{D}^{-1/2}$ 和其特征分解 $\boldsymbol{L} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^{\mathrm{T}}$,其中 $\boldsymbol{\Lambda}$ 的对角线元素为 $\boldsymbol{L}$ 的特征值 $\lambda_1, \cdots, \lambda_{n+m}$,$\boldsymbol{U}$ 为特征值对应的特征向量矩阵。给出转换函数 φ,使得 $\tilde{\lambda}_i = \varphi(\lambda_i)$,文中取的转换函数为

$$\varphi(\lambda_i) = \begin{cases} \lambda_i, \lambda_i \geqslant \lambda_{\mathrm{cut}} \\ 0, \lambda_i < \lambda_{\mathrm{cut}} \end{cases} \tag{3}$$

这样就可以构建矩阵 $\tilde{\boldsymbol{L}} = \boldsymbol{U} \tilde{\boldsymbol{\Lambda}} \boldsymbol{U}^{\mathrm{T}}$,其中 $\tilde{\boldsymbol{\Lambda}}$ 的对角线元素为 $\tilde{\lambda}_1, \cdots, \tilde{\lambda}_{n+m}$,$\boldsymbol{U}$ 为 $\boldsymbol{L}$ 的特征值对应的特征向量矩阵。进一步令 $\tilde{\boldsymbol{D}}$ 为 $\tilde{D}_{ii} = 1/\tilde{L}_{ii}$ 的对角矩阵,可以取得基于图的聚类核 $\tilde{\boldsymbol{K}} = \tilde{\boldsymbol{D}}^{1/2} \tilde{\boldsymbol{L}} \tilde{\boldsymbol{D}}^{1/2}$。

在处理测试点时,把测试点 x 近似地看做训练点和未标记点的一个线性组合,在特征空间里,用这一近似结果表达测试点和其他点之间所需的内积[11],即

$$\alpha^0 = \arg \min_{\alpha} \| \Phi(x) - \sum_{i=1}^{n+n_u} \alpha_i \Phi(x_i) \| = \boldsymbol{K}^{-1} \boldsymbol{v} \tag{4}$$

其中,$v_i = K_{\mathrm{rbf}}(x, x_i)$;$\Phi$ 是 $\boldsymbol{K}$ 的特征映射,也就是说 $K(x, x') = (\Phi(x) \cdot \Phi(x'))$。

则测试点 x 和其他点的新内积,可以表示为 $\tilde{\boldsymbol{K}}$ 线

性组合,即

$$\tilde{K}(x,x_i) \equiv (\tilde{\boldsymbol{K}}\alpha^0)_i = (\tilde{\boldsymbol{K}}\boldsymbol{K}^{-1}\boldsymbol{v})_i \quad (5)$$

为了进一步提高分类性能,文中提出将上述聚类核 $\tilde{\boldsymbol{K}}$,与目前使用广泛的多项式核函数线性组合,构成基于图的组合聚类核。取 $\beta \in [0,1]$ 为权重因子,则

$$K_G(x_i,x_j) = \beta\tilde{K}(x_i,x_j) + (1-\beta)K_P(x_i,x_j) \quad (6)$$

其中,$K_P(x_i,x_j) = [(x_i,x_j)+1]^p$。

由两个核函数的和仍为核函数,易知 $K_G(x_i,x_j)$ 为满足 Mercer 条件的核函数。在仿真实验中,线性分段转换函数 cut 的值选为 10,多项式核函数 $p=2$。

2.2　算法表述

从上述描述,可以得到基于图的组合聚类核半监督 SVM 算法 GPS^{3}VM,算法的表述如下:

Step1 根据式(1)计算相似矩阵 $\boldsymbol{W}$。

Step2 计算 $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 和其特征分解 $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}}$。

Step3 给出转换函数 φ 即式(3),$\tilde{\lambda}_i = \varphi(\lambda_i)$,其中 λ_i 是 $\boldsymbol{L}$ 的特征值,并构建 $\tilde{\boldsymbol{L}} = \boldsymbol{U}\tilde{\boldsymbol{\Lambda}}\boldsymbol{U}^{\mathrm{T}}$。

Step4 令 $\tilde{\boldsymbol{D}}$ 为 $\tilde{D}_{ii} = 1/\tilde{L}_{ii}$ 的对角矩阵,计算 $\tilde{\boldsymbol{K}} = \tilde{\boldsymbol{D}}^{1/2}\tilde{\boldsymbol{L}}\tilde{\boldsymbol{D}}^{1/2}$。

Step5 计算 $v_i = K_{\mathrm{rbf}}(x,x_i)$,带入 $\tilde{K}(x,x_i) \equiv (\tilde{\boldsymbol{K}}\alpha^0)_i = (\tilde{\boldsymbol{K}}\boldsymbol{W}^{-1}\boldsymbol{v})_i$ 得到 $\tilde{\boldsymbol{K}}$。

Step6 计算标记样本集的核矩阵 $\boldsymbol{K}_{\mathrm{Ptrain}}(1 \leqslant i \leqslant m, 1 \leqslant j \leqslant m)$,计算标记样本和无标记样本所组成的测试核矩阵 $\boldsymbol{K}_{\mathrm{rbftest}}(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m)$。

Step7 根据矩阵 $\tilde{\boldsymbol{K}}$ 的前 m 列得到 $\tilde{\boldsymbol{K}}_{\mathrm{train}}$(维数 $m \times m$),$\tilde{\boldsymbol{K}}$ 的后 n 列得到 $\tilde{\boldsymbol{K}}_{\mathrm{test}}$(维数 $m \times n$)。

Step8　取权重因子 $\beta \in [0,1]$,将 $\tilde{\boldsymbol{K}}_{\mathrm{train}}$ 和 $\boldsymbol{K}_{\mathrm{Ptrain}}$ 带入式(6)进行求和,得到训练矩阵 $\boldsymbol{K}_{\mathrm{train}}(x_i,x_j)(1 \leqslant i \leqslant m, 1 \leqslant j \leqslant m)$;将 $\tilde{\boldsymbol{K}}_{\mathrm{test}}$ 和 $\boldsymbol{K}_{\mathrm{Ptest}}$ 带入式(6)进行求和,得测试矩阵 $\boldsymbol{K}_{\mathrm{test}}(x_i,x_j)(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m)$。

Step9 用核矩阵 $\boldsymbol{K}(x_i,x_j)$ 去训练支持向量机,用得到的 SVM 分类器对测试样本分类。

3　实验与结果分析

为验证文中提出的 GPS^{3}VM 算法的有效性,给出以下数值实验。首先,在随机生成的数据集 S_1 上分析标记样本数量对该算法分类精度的影响,在随机生成的数据集 S_2 上分析权重因子 β 的取值对于该算法分类精度的影响;然后,在数据集 S_1、S_2 和 UCI 的 3 个数据集上,比较分析文中算法与标准 SVM 算法的分类精度和分类时间。

算法均在 MatlabR2009a 和 libsvm-3.1-[FarutoUltimate]工具包的基础上实现。关于参数的选择,使用启发式算法 GA(遗传算法)来进行参数寻优,用网格划分(grid search)来寻找最佳的参数 C 和 σ。由于标记样本随机生成,因此算法的分类时间和分类精度都为多次实验的平均值。

3.1　数据集

表 1 给出了实验数据集的特征描述。其中 S_1 与 S_2 为 Matlab7.0 的 mvnrnd()函数生成的正态随机分布的样本。S_1 为线性可分二维离散点集,$\boldsymbol{\mu}_1 = (0,0)^{\mathrm{T}}$,$\boldsymbol{\Sigma}_1 = \begin{bmatrix}4 & 0\\ 0 & 4\end{bmatrix}$;$\boldsymbol{\mu}_2 = (10,10)^{\mathrm{T}}$,$\boldsymbol{\Sigma}_2 = \begin{bmatrix}9 & 0\\ 0 & 9\end{bmatrix}$。$S_2$ 为非线性可分二维离散点集,$\boldsymbol{\mu}_1 = (2.1,1.5)^{\mathrm{T}}$,$\boldsymbol{\Sigma}_1 = \begin{bmatrix}1.3 & 0\\ 0 & 1.3\end{bmatrix}$;$\boldsymbol{\mu}_2 = (-0.5,-0.8)^{\mathrm{T}}$,$\boldsymbol{\Sigma}_1 = \begin{bmatrix}1.1 & 0\\ 0 & 1.1\end{bmatrix}$。UCI 数据集是数据挖掘中的公共测试数据集,已知该数据集中数据的属性和类别,使用者可以用自己的方法将数据进行分类,然后将分类结果与数据说明进行对比,以此说明算法的性能。文中选用 UCI 数据集中的 Liver、Sonar 和 Heart 这 3 个数据集进行实验。

表 1　实验数据集描述

数据集	属性	样本总数	类别数
S_1	2	200	2
S_2	2	200	2
Liver	6	345	2
Sonar	60	208	2
Heart	13	270	2

3.2　实验结果分析

在数据集 S_1 上,标记样本比例分别为 1%,2%,5%,10%,15%,20%,25%时,算法分类精度如图 1 所示。在数据集 S_2 上,选取 10 个标记样本,当权重因子 $\beta = 0, 0.1, 0.2, \cdots, 0.9, 1$ 时,算法分类精度如图 2 所示。

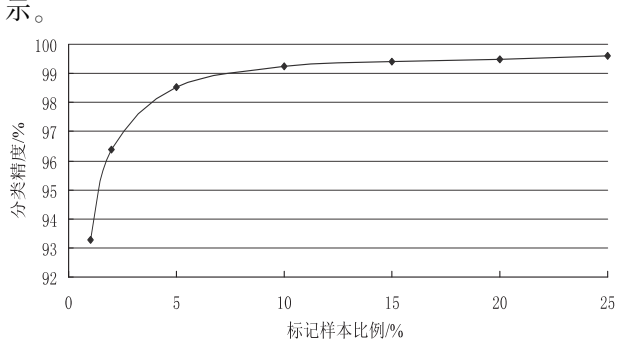

图 1　S_1 上标记样本比例对分类精度的影响

由图 1 可知,随着标记样本比例的增加,GPS^{3}VM

算法的分类精度提高,且保持在较高水平,最终趋向于100%。这说明,标记样本越多,该算法对于未标记样本中携带的信息利用越充分,分类精度就越高。由图2可以看出,当$\beta=0$时,即只是基于图的聚类核在起作用时,GPS^{3}VM算法的分类精度为86.648 7%;当$\beta=1$时,即只是多项式核函数在起作用时,GPS^{3}VM算法的分类精度为78.066 7%;而$\beta=0.8$时,GPS^{3}VM算法的分类精度达到最大为90.610 3%。可见基于图组合半监督SVM聚类核,有效提高了基于图的聚类核与多项式核的分类效果。

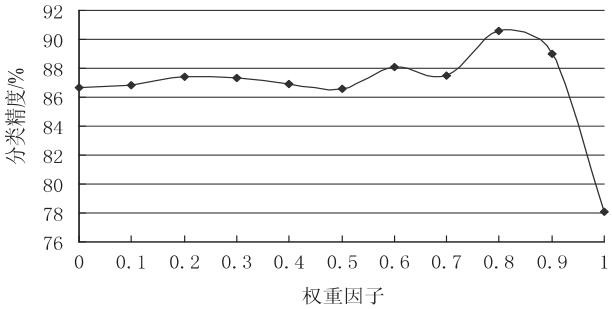

图2 S_2上权重因子对分类精度的影响

下面在各个数据集上,比较GPS^{3}VM算法与标准SVM算法的分类精度和分类时间。在S_1、S_2数据集上选取6个标记样本,标记样本比例为3%;在Liver、Sonar、Heart数据集上选取10个标记样本,标记样本比例分别为2.898 6%、4.807 7%和3.703 7%。实验结果如表2和图3所示。

表2 实验结果比较

数据集	分类方法	平均分类精度/%	最大分类精度/%	平均运行时间/s
S_1	标准SVM	77.167 3	99.746 2	0.001 9
	GPS^{3}VM	97.649 7	100.00	2.956 7
S_2	标准SVM	86.113 2	95.897 4	0.007 8
	GPS^{3}VM	87.544 2	97.058 8	0.934 9
Liver	标准SVM	53.792 2	66.268 7	0.012 6
	GPS^{3}VM	57.960 6	71.940 3	2.349 3
Heart	标准SVM	52.134 6	76.767 7	0.012 5
	GPS^{3}VM	62.612 2	73.461 5	49.438 5
Sonar	标准SVM	60.242 4	76.767 7	0.005 1
	GPS^{3}VM	64.912 52	75.757 6	96.016 3

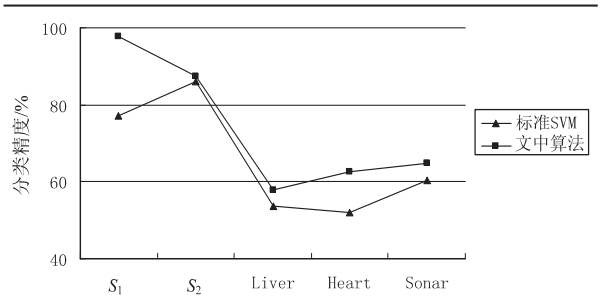

图3 实验结果比较图

可以看出在相同条件下,在训练精度方面,各个数据集上GPS^{3}VM算法的分类精度明显高于标准SVM算法,这主要是由于GPS^{3}VM算法利用了"无标签"样本中携带的大量分类信息,提高了SVM分类器的分类精度,体现了半监督式学习的优势。与其他算法一样,该算法分类精度的提高也牺牲了一定的分类时间。

4 结束语

文中根据Chapelle提出的聚类核概念,选取线性分段转换函数$\tilde{\lambda}_i=\varphi(\lambda_i)$,将加权无向图上的相似矩阵$\boldsymbol{W}$,用聚类核$\tilde{\boldsymbol{K}}$重新表述,并将$\tilde{\boldsymbol{K}}$与多项式核函数线性组合,构成了基于图的复合半监督SVM聚类核算法(GPS^{3}VM)。实验证明,随着标记样本比例的增加,该算法的分类精度也在增加;与标准SVM算法相比,该算法分类精度较高,且高于组合前单独的核函数。针对不同数据集,维数、特性等方面的不同,线性分段函数cut值、多项式核函数的p值和权重因子β值等参数的选取都会有所不同。并且,随着数据维数的增大,运行时间也会随之增加。因此,如何更有效的进行参数选择,降低时间复杂度仍然有待进一步研究。

参考文献:

[1] 李昆仑,曹 铮,曹丽萍,等. 半监督聚类的若干新进展[J]. 模式识别与人工智能,2009,22(5):735-742.

[2] 徐庆伶,汪西莉. 一种基于支持向量机的半监督分类方法[J]. 计算机技术与发展,2010,20(10):115-117.

[3] Joachims T. Transductive inference for text classification using support vector machines[C]//Proceedings of the 16th international conference on machine learning. San Francisco:Morgan Kaufmann Publishers,1999:200-209.

[4] Chapelle O,Sindhwani V,Keerthi S. Optimization techniques for semi-supervised support vector machines[J]. Journal of Machine Learning Research,2008,9(2):203-233.

[5] Chapelle O,Zien A,Cowell R,et al. Semi-supervised classification by low density separation[J]. Encyclopedia of Biostatistics,2005,34:57-64.

[6] Collobert R,Sinz F,Weston J,et al. Large scale transductive SVMs[J]. Journal of Machine Learning Research,2006,7:1687-1712.

[7] Sindhwani V,Keerthi S,Chapelle O. Deterministic annealing for semi-supervised kernel machines[C]//Proceedings of international conference on machine learning. Pittsburgh:[s. n.],2006:108-116.

[8] Chapelle O,Chi M,Zien A. A continuation method for semi-supervised SVMs[C]//Proceedings of international conference on machine learning. Pittsburgh:[s. n.],2006:184-192.

[9] de Bie T,Cristianini N. Semi-supervised learning using semi-

(下转第 116 页)

PTS 方法好,当 β 大于 0.15 时,叠加训练序列降低 PAPR 的能力比原始 PTS 好。为确保发送信号所占的功率分配因子最优化,在保证系统性能的前提下,应尽可能降低所融合的叠加训练序列所占的功率分配比值。

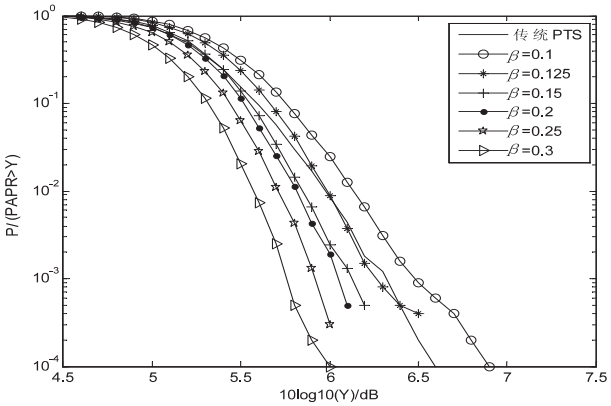

图 6 功率分配因子的值在 0.2 附近的 PAPR 曲线

4 结束语

与传统的 PTS 算法相比,基于叠加训练序列的 PTS 算法,不仅可以提高系统的带宽利用率,而且可以较好地降低系统的 PAPR。对整个 OFDM 系统来说,训练序列 M 并行叠加在信号上,它还可以用于信道估计和系统同步;训练序列 M 作为相位旋转因子时,边带信息的传输就不需要占用新的频谱资源,于是提高了频谱利用率。文中比较了不同自相关性的训练序列作叠加序列、旋转因子以及功率分配因子的大小对降低 PAPR 性能的影响。仿真结果表明:叠加训练序列的自相关性越好,降低 PAPR 的能力就越好;当训练序列只作相位旋转因子时,降低 PAPR 的效果并没有传统 PTS 方法好;当 Hadama 码序列做叠加及相位旋转因子、功率分配因子大于 0.15 时,该方法能够更有效地降低系统的 PAPR。

参考文献:

[1] Weinstein S B. The history of orthogonal frequency-division multiplexing[J]. Communications Magazine, 2009, 47(11): 26-35.

[2] Wu Yiyan. Orthogonal frequency division multiplexing: A multi-carrier modulation scheme[J]. IEEE Transactions on Consumer Electronics, 1995, 41(3): 392-399.

[3] Jiang Tao, Wu Yiyan. An overview: Peak-to-average power ratio reduction techniques for OFDM signals[J]. IEEE Transactions on Broadcasting, 2008, 54(2): 257-268.

[4] Wang Yingming, Zhang Guodong, Wang Xiaodong. Polyphase codes for uplink OFDM-CDMA systems[J]. IEEE Transactions on Communications, 2008, 56(3): 435-444.

[5] Lee B M, de Figueiredo R J P, Kim Y. A computationally efficient tree-PTS technique for PAPR reduction of OFDM signals[J]. Wireless Personal Communications, 2012, 62(2): 431-442.

[6] 罗仁泽. 新一代无线移动通信系统关键技术[M]. 北京:北京邮电大学出版社, 2007.

[7] Tellambura C. Upper bound on peak factor of N-multiple carriers[J]. IEEE Electronics Letters, 1997, 33(19): 1608-1609.

[8] 黄润林, 周 克. 利用伪随机序列降低 OFDM 系统 PAPR 方法[J]. 电子科技大学学报, 2007, 36(6): 1515-1518.

[9] He Shuangchi, Tugnait J K. On Doubly selective channel estimation using superimposed training and discrete prolate spheroidal sequences[J]. IEEE Transactions on Signal Process, 2008, 56(7): 3214-3228.

[10] Nair J P, Kumar R V R. An iterative channel estimation method using superimposed training in OFDM systems[C]//Proc of IEEE international conference on vehicular technology conference. Calgary, BC: IEEE, 2008: 1-5.

[11] Ma Yangjun, Hu Yaonu. An improved training sequence-based OFDM synchronization algorithm[J]. Study on Optical Communications, 2009, 35(6): 65-67.

[12] Pan Zongshan, Li Xiaomin. A novel synchronization algorithm for OFDM system based on training sequence added scramble code[C]//Proceedings of IEEE international conference on communications technology and applications. Beijing: IEEE, 2009: 527-531.

(上接第 112 页)

definite programming[M]//Semi-supervised Learning. Massachusetts: MIT Press, 2006: 119-135.

[10] Zhu Xiaojin. Semi-supervised learning literature survey[R]. Wisconsin: University of Wisconsin Madison, 2008.

[11] Chapelle O, Weston J, Scholkopf B. Cluster kernels for semi-supervised learning[C]//Proceedings of the 16th annual conference on neural information processing systems. Massachusetts: MIT Press, 2003: 321-328.

[12] Zhou Z H. Co-training paradigm in semi-supervised learning [C]//Proceeding of the Chinese workshop on machine learning and applications. Nanjing, China: [s. n.], 2007.

[13] Weston J, Leslie C, Ie E, et al. Semi-supervised protein classification using cluster kernels [J]. Bioinformatics, 2005, 21(15): 3241-3247.

[14] Szummer M, Jaakkola T. Partially labeled classification with Markov random walks[M]//Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002: 945-952.

[15] Bach F R, Jordan M I. Learning spectral clustering[C]//Proceedings of 17th annual conference on neural information processing systems. Massachusetts: MIT Press, 2003: 305-312.

基于图的组合半监督SVM聚类核算法研究

作者： 郑文静， 李雷， ZHENG Wen-jing， LI Lei
作者单位： 南京邮电大学 理学院,江苏 南京,210023
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2014(5)