基于K均值聚类和多核SVM的微钙化簇检测

常甜甜[1] 刘红卫[1] 冯 筠[2]

(西安电子科技大学理学院 西安 710071)[1] (西北大学信息技术学院 西安 710069)[2]

摘 要 考虑到乳腺微钙化簇样本分布不平衡以及特征的多样性,提出了基于K均值聚类的多核支持向量机。即首先将训练样本聚合成K类,对每类样本加不同的惩罚因子,以平衡样本分布不平衡。其次针对样本特征多样性,将核函数做组合,得到多核支持向量分类器。使用主动反馈学习的方法来得到稳定的训练样本。实验结果表明,本方法与单核SVM及多核SVM相比,检对率至少可以提高两个百分点。

关键词 K均值聚类,多核支持向量机,微钙化簇,主动反馈学习

中图法分类号 TP181 **文献标识码** A

Microcalcification Detection Based on K-means Cluster and Multiple Kernel Support Vector Machine

CHANG Tian-tian[1] LIU Hong-wei[1] FENG Jun[2]

(School of Science, Xidian Univ., Xi'an 710071, China)[1]

(School of Information Science and Technology, Northwest University, Xi'an 710069, China)[2]

Abstract Considering the unbalanced distribution of the training samples and the multiformity of the features. A multiple kernel SVM based on K-means cluster algorithm was proposed. Firstly, training samples was clustered into K classes, different penalty factors were used for each class in order to balance the contributions of each class. Secondly, the multiple kernel support vector machine was proposed for diversity of the features. The stabilized training sample was obtained via active feedback learning. The result show that the detection rate can be improved at least 2 percent by the proposed method, compared with the single kernel SVM and the multiple kernel SVM.

Keywords K-means cluster, Multiple kernel SVM, Microcalcification, Active feedback learning

乳腺钼靶X光片中微钙化簇检测是医学影像领域的难点。该技术是指抽取有诊断价值的可能含微钙化簇的感兴趣区域(Region of Interest ROI)特征,并进行特征优化和分类。

很多学者将不同分类器用于乳腺微钙化点检测,得出SVM比其他分类器具有更好分类性能的结论[1-5]。Bazzani得出对于小数目训练样本来说SVM具有比多层感知器更好的分类性能[2]。Papadopoulos将SVM与神经网络作对比,得出SVM的分类精度更高[3]。在将SVM、相关向量机、核fisher判别法、前馈神经网络、committee machines等分类器作对比后,Wei得出基于核的分类器,可以比SVM、相关向量机、核fisher判别法得到更好的分类性能[4]。然而,由于微钙化簇区域的特征空间与正常组织很接近,传统的SVM并不能得到很满意的精度。为提高SVM分类器的性能,Issam El-Naqa提出了一种增强学习的方法,将有限图像窗作为输入,用于SVM分类[5]。Li Ying提出多核SVM[6]来检测微钙化簇,得出线性核函数与多项式核函数的组合SVM(MKSVM)在保持真阳性率(TPR)不变的情况下降低了假阳性率(FPR)。

考虑到疑似钙化簇区域多样化的影像学特点,即(1)病灶区域特征的多样性;(2)正常乳腺组织的多变性;(3)对于在乳腺图像中被脂肪、腺体、高密度乳腺组织覆盖的微钙化簇病灶的低对比度,本文提出基于K均值聚类的多核SVM。从样本与特征两方面考虑,首先,将训练样本聚合成类,由每类样本对分类器的贡献大小是不同的,且特征特别明显的样本往往又是小数目的。对每类样本加不同的惩罚因子,以加大小样本的贡献,这是从样本方面考虑的。其次,对于样本特征的多样性,提出将核函数做组合的多核支持向量机,这是从特征方面考虑的。使用主动反馈学习的方法来得到稳定的训练样本。实验结果表明,本方法与单核SVM以及多核SVM相比,检对率至少可以提高两个百分点。

1 多核支持向量机(MKSVM)简介

1.1 支持向量机(SVM)[7]

支持向量机是Vapnik等人在1995年提出的以有限样本统计学习理论为基础的一种机器学习方法,其原始问题为:

$$\min_{w\in H,b\in R,\xi\in R^l} \frac{1}{2}\|w\|^2+\sum C\xi_i$$
$$s.t.\ y_i((w\cdot x_i)+b)\geqslant 1-\xi_i, i=1,\cdots,l \quad (1)$$

到稿日期:2008-10-14 返修日期:2008-12-04 本文受国家自然科学基金(60603098),陕西省教育厅科学研究计划项目(07JK381)资助。

常甜甜(1981-),女,博士研究生,主要研究方向为机器学习及其最优化方法研究,E-mail:changtiantian@gmail.com;刘红卫(1967-),男,博士生导师,主要研究方向为半定规划及其应用、机器学习及其最优化方法研究;冯 筠(1972-),女,硕士生导师,主要研究方向为医学图像处理、三维重建、模式识别。

$\xi_i \geqslant 0, i=1,\cdots,l$

其对偶问题为

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i,x_j) - \sum_{j=1}^{l}\alpha_j$$

$$s.t. \sum_{i=1}^{l} y_i \alpha_i = 0 \tag{2}$$

$$0 \leqslant \alpha_i \leqslant C$$

1.2 多核支持向量机(MKSVM)

多核 SVM 是 2004 年由 Lanckriet[12] 提出的，主要是针对特征的多源性问题。主要思想就是将核矩阵经过组合变成新的核矩阵：

$$K = \sum_{d=1}^{D} \mu_d K_d \tag{3}$$

其中，μ_d 是 K_d 的系数。每个核矩阵归一化为：

$$K_d(x_i,x_j) = K_d(x_i,x_j)/trace(K_d) \tag{4}$$

本文考虑 1 范数软间隔 SVM，有

$$\begin{cases} \max \sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j K(x_i \cdot x_j) y_i y_j \\ s.t. \sum_i \alpha_i y_i = 0, 0 \leqslant \alpha_i \leqslant C \end{cases} \tag{5}$$

其中，α 是拉格朗日乘子，b 是常数项，y 是样本标记，C 是错分样本的惩罚因子。最优分类超平面为：

$$y_i\left(\sum_{j=1}^{N}\alpha_j^* y_j K(x_i,x_j) + b^*\right) = 1, i=1,\cdots,N \tag{6}$$

将式(4)，(5)结合起来考虑，可以得到：

$$\begin{cases} \max \frac{1}{2}\sum_{i=1}^{l}\alpha_i - \sum_{i,j=1}^{l}\alpha_i\alpha_j(\sum\mu_d K_d) y_i y_j \\ s.t. \sum_{i=1}^{l}\alpha_i y_i = 0, 0 \leqslant \alpha_i \leqslant C; \sum\mu_d = 1 \end{cases} \tag{7}$$

2 K 均值聚类多核 SVM(KM-MKSVM)

2.1 K 均值聚类 SVM(KSVM)

在训练集中的负类点和正类点的个数有较大差距时，如果对正类点集和负类点集应用相同的惩罚参数 C，则意味着哪一类点的个数多，就更看重哪类点。然而，我们希望对正类点和负类点的惩罚是不相同的。为此，对适当选定的参数 C，令 $C_+ = \frac{l_-}{l_+ + l_-}C$，$C_- = \frac{l_+}{l_+ + l_-}C$[7]。其中 l_+ 和 l_- 分别是正类训练点和负类训练点的个数。C_+ 是对正类点集的惩罚参数，C_- 是对负类点集的惩罚参数。此时支持向量机的原始问题变形为

$$\min_{w\in H, b\in R, \xi\in R^l} \frac{1}{2}\|w\|^2 + \sum_{y_i=1} C_+\xi_i + \sum_{y_i=-1} C_-\xi_i$$

$$s.t.\ y_i((w \cdot x_i)+b) \geqslant 1-\xi_i, i=1,\cdots,l \tag{8}$$

$$\xi_i \geqslant 0, i=1,\cdots,l$$

其对偶问题为加权 SVM(CSVM)：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i,x_j) - \sum_{j=1}^{l}\alpha_j$$

$$s.t. \sum_{i=1}^{l} y_i \alpha_i = 0 \tag{9}$$

$$0 \leqslant \alpha_i \leqslant C_+, y_i = 1$$

$$0 \leqslant \alpha_i \leqslant C_-, y_i = -1$$

由式(8)得到启发，对不同类别的样本加不同的惩罚因子，可以改善分类器的分类精度。对于聚类后得到的不同类样本，同样可以对其加以不同的惩罚，得到下面的二次规划问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i,x_j) - \sum_{j=1}^{l}\alpha_j$$

$$s.t. \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \leqslant \alpha_i \leqslant C_{class1},\ Class\ Index = class1;$$

$$0 \leqslant \alpha_i \leqslant C_{class2}, Class\ Index = class2; \tag{10}$$

$$\cdots\cdots$$

$$0 \leqslant \alpha_i \leqslant C_{classN}, Class\ Index = classN;$$

$$C_i = \frac{l_1+,\cdots,l_{i-1},l_{i+1},\cdots,l_n}{l_1+,\cdots,+l_n}$$

其中，α 是拉格朗日乘子，$class1,\cdots,classN$ 指聚类后的类别，$Class\ Index$ 指类别标记，$l_1,\cdots,l_n$ 指每类中的样本点数目，C_i 表示每类的惩罚因子。

2.2 K 均值聚类多核 SVM(KM-MKSVM)

将式(3)，(10)结合起来考虑，可以得到 K 均值多核 SVM：

$$\begin{cases} \max \frac{1}{2}\sum_{i=1}^{l}\alpha_i - \sum_{i,j=1}^{l}\alpha_i\alpha_j(\sum\mu_d K_d) y_i y_j \\ s.t. \sum_{i=1}^{l}\alpha_i y_i = 0, \\ 0 \leqslant \alpha_i \leqslant C_{class1}, Class\ Index = class1; \\ 0 \leqslant \alpha_i \leqslant C_{class2}, Class\ Index = class2; \\ \cdots\cdots \\ 0 \leqslant \alpha_i \leqslant C_{classN}, Class\ Index = classN; \\ \sum\mu_d = 1 \\ C_i = \frac{l_1+,\cdots,l_{i-1},l_{i+1},\cdots,l_n}{l_1+,\cdots,+l_n} \end{cases} \tag{11}$$

其中，α 是拉格朗日乘子，$class1,\cdots,classN$ 指聚类后的类别，$Class\ Index$ 指类别标记，$l_1,\cdots,l_n$ 指每类中的样本点数目，C_i 表示每类的惩罚因子，K_d 分别表示不同的核函数，μ_d 表示核函数的权衡因子。

3 数据实验

3.1 DDSM 数据库数据

实验采用美国南佛罗里达州立大学 DDSM 数据库[9]，共测试 239 幅图像。从中提取 1064 个 ROI，其中正样本 496 个、负样本 568 个，共提取 63 个特征，如表 1 所列。

表 1 乳腺癌微钙化簇检中采用的特征

特征类别	特征(维数)
灰度特征	对比度；灰度均值；方差；三阶距；四阶距；平均梯度；区域边界的平均梯度；不变矩(7 维)
几何特征	圆形度；紧缩度；球状性；傅立叶描述子
纹理特征	基于纹理能量图；能量图均值；能量图方差；基于灰度共生矩阵；能量；熵；对比度；基于小波变换；能量(9 维)；熵(9 维)

3.2 实验结果

从图 1 我们看到，负类样本在聚类时类内差异与类间差异之比几乎没有变化，即负类样本分布比较均匀。而正类样本在聚类中，聚类数目对聚类结果影响很大，即正类样本分布不均匀。图 2 中的 10 条曲线分别是随机选取 10 组训练样本进行聚类得到的评价曲线。当聚类数目达到 8 个以上时，评价值变得稳定，因此选 8 为聚类点数。

实验当中选取了 4 种核函数，分别是：

线性核函数 $k(x,x') = x \cdot x'$

多项式核函数 $k(x,x') = (x \cdot x' + 1)^a$

高斯核函数 $k(x,x') = \exp(-(x-x') \cdot (x-x')'/2 \cdot b^2)$

Sigmoid 核函数 $k(x,x')=\tan h(p_c \cdot x \cdot x'+p_d)$

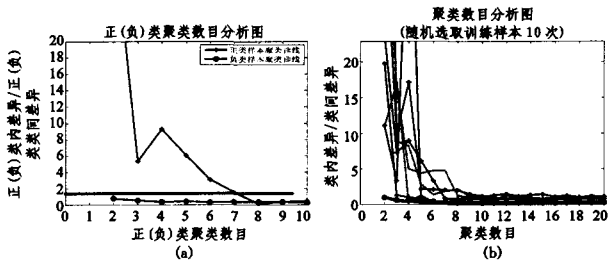

图 1 (a)聚类数目不同时正负类样本评价值对比图;(b)聚类数目不同时评价值的稳定性图

在实验过程中采用基于主动反馈学习的样本选择方法[13]。设 P 是训练样本集,S 是当前迭代的训练样本集,新被选择的样本满足以下条件:

$$\min_{x_i \in Q \setminus S}((1-\lambda_1-\lambda_2)|d(x_i)|+\lambda_1 \max_{x_j \in S}|k(x_j,x_i)|+\lambda_2 \max_{x_j \in SV}|k(x_j,x_i)|) \quad (12)$$

其中 $d(x_i)=\alpha^* * diag(y) * K(x_j,x_i)$,$\theta$ 是权衡因子且 $0 \leqslant \theta \leqslant 1$,$K=\sum_{d=1}^{D}\mu_d K_d$。当分类器的分类性能稳定的时候,反馈学习终止。实验结果表明,KM-MKSVM 可以提高乳腺微钙化簇检对率,如表 2 所列。

表 2 不同 SVM 的性能对比表

分类器	核函数	核参数	检对率
SVM	高斯核	b=0.5	76.51%
CSVM	高斯核	b=0.5	78.80%
KSVM	高斯核	b=0.5	79.40%
MKSVM	线性核+多项式	a=1.7	77.16%
MKSVM	线性核+高斯	b=0.5	79.09%
MKSVM	线性核+sigmoid	c=4.1;d=10	79.53%
MKSVM	多项式+高斯核	b=0.5;a=1.7	77.59%
MKSVM	多项式+sigmoid	a=1.7;c=4.1;d=10	76.51%
MKSVM	高斯核+sigmoid	b=0.5;c=4.1;d=10	79.31%
KM-MKSVM	线性核+多项式	b=0.5	81.47%
KM-MKSVM	线性核+高斯	b=0.5	79.09%
KM-MKSVM	线性核 sigmoid	c=4.1;d=10	81.25%
KM-MKSVM	多项式+高斯	a=1.7;b=0.5	80.18%
KM-MKSVM	多项式+sigmoid	a=1.7;c=4.1;d=10	81.68%
KM-MKSVM	高斯核+sigmoid	b=0.5;c=4.1;d=10	82.11%

结束语 为提高微钙化簇的检对率,考虑到乳腺微钙化簇正负类样本分布不平衡以及特征多样性的特点,提出基于 K 均值聚类的多核 SVM。本方法将正类和负类样本分别聚类,对不同类样本加不同的惩罚因子,用主动反馈学习的方法选择训练样本,最后用多核 SVM 训练空间模型。实验结果表明,该方法可以有效地提高乳腺微钙化簇的检对率。

参 考 文 献

[1] Thangavel K, Karnan M, Sivakumar R, et al. Automatic Detection of Microcalcification in Mammograms-A Review[J]. International Journal on Graphics Vision and Image. Processing, 2005,5(5):31-61

[2] Bazzani A, et al. A SVM classifier to separate false signals from microcalcifications in digital mammograms[J]. Phys. Med. Biol, 2001,46(6):1651-1663

[3] Papadopoulos A, Fotiadis D I, Likas A. Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines[J]. Artificial Intelligence in Medicine, 2004,34(2):141-150

[4] Wei Liyang, Yang Yongyi, et al. A Study on Several Machine-learning Methods for Classification of Malignant and Benign Clustered Microcalcifications[J]. March. IEEE Transaction on Medical Imaging, 2005,4(3):371-380

[5] El-Naqa I, et al. A support vector machine aproach for detection of microcalcifications[J]. IEEE Transactions on Medical Imaging, 2002,21(12):1552-1563

[6] Li Ying, Jiang Jianmin. Combination of SVM Knowledge for Microcalcification Detection in Digital Mammograms[C]// IDEAL, 2004, LNCS 3177. 2004:359-365

[7] 邓乃扬,田英杰. 数据挖掘中的新方法-支持向量机[M]. 北京:科学出版社,2004:369-370

[8] 边肇祺,张学工. 模式识别(第二版)[M]. 北京:清华大学出版社,2000:280-281

[9] Rose C, Turi D, Williams A, et al. Digital Database for Screening Mammography[DB/OL]. http: // marathon. csee. usf. edu/ Mammography/Database. html, 1998

[10] Wang Jiaqi, Wu Xindong, Zhang Chengqi. Support Vector Machines Based on K-means Clustering for Real-time Business Intelligence Systems[J]. International Journal of Business Intelligence and Data Mining, 2005,1(1):54-64

[11] Li Maokuan, Cheng Yusheng, Zhao Honghai. Unlabeled Data Classification via Support Vector Machines and k-means Clustering[C]// Proceedings of the International Conference on Computer Graphics, Imaging and Visualization. 2004:183-186

[12] Lanckriet G, et al. Learning the Kernel Matrix with Semi-definite Programming[J]. Journal of Machine Learning Research, 2004:27-72

[13] Jiang J, Ip H H S. Active Learning with SVM[M]. Rabuñal J R, Dorado J, Pazos A, eds. Encyclopedia of Artificial Intelligence: Information Science Reference, 2008

(上接第 207 页)

[11] Chen L, Shen J, Qin L. A method for solving optimization problem in continuous space by using ant colony algorithm[J]. Journal of Software, 2002,13(12):2317-2322 (in Chinese)

[12] Wu B, Shi ZZ. An ant colony algorithm based partition algorithm for TSP[J]. Chinese Journal of Computers, 2001,24(12):1328-1333 (in Chinese)

[13] Zhang J H, Gao Q S, Xu X H. A self-adaptive ant colony algorithm[J]. Control Theory And Applications, 2000, 17(1): 1-3 (in Chinese)

[14] Dorigo M, Maniezzo V, Colorni A. Ant System: An autocatalytic optimizing process[R]. 91-106. 1991

[15] Ding Jian-li, Chen Zeng-qiang, Yuan Zhu-zhi. Dynamin optimization routing method base on ant adaptive algorithm[J]. Control and Decision, 2003,18(6):751-753 (in Chinese)

[16] Zhang Su-bing, Lu Guo-ying, LIU Ze-min, et al. QoS Routing Based on Ant-Algorithm[J]. Journal of Circuits and Systems, 2000,5(1)

[17] Xiang F, Junzhou L, Jieyi W, et al. QoS Routing Based on Genetic Algorithm[J]. Computer Communications, 1999, 22: 1392-1399

[18] Wang Z, Crowcroft J. Quality of Service Routing for Supporting Multimedia Applications[J]. JSAC, 1996,14(7):1228-1234

[19] Schoonderwoerd R, Holland O, Bruten J. Ant-based Load Balancing in Telecommunications Networks[J]. Adaptive Behavior, 1996,5(2):169-207