doi:10.11959/j.issn.1000-436x.2015263

基于隐含信息的半监督学习方法研究

刘国栋[1]，许静[1]，张国兵[2]
(1. 南开大学 计算机与控制工程学院，天津 300071；2. 北京航空航天大学 电子信息工程学院，北京 100191)

摘 要：研究了基于隐含信息的半监督学习方法，并将该方法应用于支持向量机和随机森林模型。利用 UCI 数据库中的数据验证了基于此方法的支持向量机和随机森林的精度。在此基础上，将此种方法应用于肺音识别领域，利用实际的肺音数据对此方法处理实际问题的效果进行了验证，同时实验分析了无标记样本的数量以及质量对此方法的影响。

关键词：半监督学习；肺音；隐含信息
中图分类号：TP181 **文献标识码**：A

Study of implicit information semi-supervised learning algorithm

LIU Guo-dong[1], XU Jing[1], ZHANG Guo-bing[2]
(1.College of Computer and Control Engineering, Nankai University, Tianjin 300071, China;
2. School of Electronic and Information Engineering, Beihang University, Beijing 100191, China)

Abstract: Implicit information semi supervised learning algorithm was studied. The implicit information semi supervised learning algorithm was used in support vector machine and random forest, which were called semi-SVM and semi-RF. The semi-SVM and semi-RF were evaluated by using UCI, the experimental results show that the semi-SVM and semi-RF are more effective and more precise. The semi-SVM and semi-RF were applied to classifying lung sounds, and verified the effect by using the actual lung sounds data. the quantity and quality of samples affect semi-SVM and semi-RF were analyzed.

Key words: semi-supervised learning; lung sounds; implicit information

1 引言

半监督学习技术是在有监督学习和无监督学习技术的基础上提出的，与传统的机器学习技术相比，半监督学习技术更适合解决实际中少量有标记样本与大量无标记样本并存的问题，能够更好地利用大量未标记样本数据帮助提高学习的性能[1~3]。因此，半监督学习技术成为机器学习领域中重点研究方向之一。

半监督学习技术基本做法是根据数据的分布模型的假设，借用未标记样本与已标记样本之间的某种关系，建立对未标记样本进行标记的模型。然后利用标记后的样本对分类器进行训练，以提高分类器的分类性能。从文献[4～6]中提到的自训练(self-training)方法开始，以及后来的模型生成方法[7]、基于图和流形的半监督学习方法[8,9]、协调训练法[10,11]等，这些半监督学习方法都需要以某种假设为支撑对未标记样本进行标记。主要有聚类假设[12]、流形假设[13]和局部与全局一致性假设[15]。基于标记法的半监督方法概念清晰，易于实现，但是利用少量的标记样本对大量无标记样本进行标记时不可避免地会出现标记错误的情况，从而造成样本数据被污染，造成最终训练的分类器性能下降。

文献[15]从另一个角度提出基于隐含信息特征的半监督学习（FSSL, formative semi-supervised learning）方法。该方法充分利用已标记样本和无标记样本之间隐含的关联关系，将这种关联关系作为已标记样本的扩展特征，构成新的有标记的样本集。将拓展特征后获得的新的训练样本集重新进行训练，得到新的分类器。基于隐含信息属性的半监督方法不仅避免了通过假设对未标记的样本进行标记的弊端，同时又充分、客观地利用了未标记样本的信息。

收稿日期：2015-05-14；修回日期：2015-09-27

本文主要研究 FSSL 学习方法。将 FSSL 方法应用于支持向量机（SVM, support vector machine）和随机森林（RF, rondom forest）模型。利用通用的 UCI 数据库中的数据验证了 FSSL 方法提高分类器精度的效果。另一方面，针对实际肺音数据中存在大量无标记样本的实际情况，首次将基于 FSSL 方法的 SVM 和 RF 模型应用于肺音识别领域。实验结果表明，该方法可以对无标记肺音数据进行正确的分类。

2 基于隐含信息属性的半监督学习方法

2.1 训练样本隐含属性的计算

设 S 是 $|S|$ 个数据样本的集合，$S=(X,Y)=\{(x_i,y_i)|x_i\in R^n,y_i\in R,i=1,2,\cdots,|S|\}$，$|S|=L+U$ 为两部分，分别为已标记数据 $L=\{x_1,x_2,\cdots,x_i,\cdots,x_L\}$ 和无标记数据 $U=\{x_1,x_2,\cdots,x_j,\cdots,x_U\}$，且 $L\ll U$。对于半监督学习而言，关键是在如何客观地利用未标记样本集 $U=\{x_1,x_2,\cdots,x_j,\cdots,x_U\}$ 的信息。下面首先研究如何获得 L 和 U 的关联属性。

由于 L 和 U 都是来自同一分布的样本集，因此可以看作都是在随机变量 Z 的基础上产生的。对于每一个给定的 Z_k，都有 $P(l_i|z_k)$ 和 $P(u_j|z_k)$ 与之互相关联，因此 $P(z_k|l_i)$ 中必然包含了集合 U 中的样本信息。因此只需要求出 $P(z_k|l_i)$ 的值，然后将这些隐含信息属性值附加到 L 集合中的样本属性中，新的有标记的样本集中必然包含了未标记样本的信息，如表 1 所示。

表 1　含有隐含信息属性的样本

L	原有的属性值				类标记	隐含信息属性值		
l_i	f_1	f_2	…	f_s	Y	$P(z_1\|l_i)$	…	$P(z_k\|l_i)$
…	…	…	…	…	…	…	…	…

现在的问题就转化为如何求出后验概率 $P(z_k|l_i)$ 的值。根据最大似然原则，$P(z_k|l_i)$ 的值可以使 $P(L,U,Z)$ 的值达到最大，即 $\log P(L,U,Z)$ 取最大值。

$$\rho=\log P(L,U,Z)=\sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}\log P(l_i,u_j,z_k)$$
$$=\sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}\log[P(z_k|l_i,u_j)P(l_i,u_j)] \tag{1}$$

其中，$\theta_{l_i}^{u_j}$ ($i=1,\cdots,L, j=1,\cdots,U$)表示为已标记样本集与无标记样本集中的样本之间的度量关系，矩阵 $\boldsymbol{\theta}=\theta_{l_i}^{u_j}$ 的值为 2 个相关样本之间的欧式距离，其中

$$P(l_i,u_j)=P(l_i)P(u_j|l_i)=P(l_i)\sum_{k=1}^{K}P(u_j|z_k)P(z_k|l_i) \tag{2}$$

利用贝叶斯公式有

$$P(z_k|l_i)=\frac{P(l_i|z_k)P(z_k)}{P(l_i)} \tag{3}$$

$$P(z_k|l_i,u_j)=\frac{P(l_i,u_j|z_k)P(z_k)}{\sum_{t=1}^{K}P(l_i,u_j|z_t)P(z_t)} \tag{4}$$

由于 L 和 U 在以隐含信息属性 Z 为条件的概率是相互独立的，所以式（4）可以变换为

$$P(z_k|l_i,u_j)=\frac{P(l_i|z_k)P(u_j|z_k)P(z_k)}{\sum_{t=1}^{K}P(l_i|z_t)P(u_j|z_t)P(z_t)} \tag{5}$$

根据式（2）～式（5）可以使式（1）进一步变换为

$$\rho=\log P(L,U,Z)$$
$$=\sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}\{\log[P(l_i|z_k)P(u_j|z_k)P(z_k)]\}\cdot$$
$$P(z_k|l_i,u_j)+\sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}\log[P(z_k|l_i,u_j)] \tag{6}$$

因为

$$\sum_{k=1}^{K}P(z_k)=1,\ \sum_{i=1}^{L}P(l_i|z_k)=1,\ \sum_{j=1}^{U}P(u_j|z_k)=1$$

所以可以构建一个拉格朗日函数

$$\Lambda(\rho,\lambda,\mu,\nu)=\rho+\lambda\sum_{k=1}^{K}(1-P(z_k))+$$
$$\mu\sum_{i=1}^{L}(1-P(l_i|z_k))+\nu\sum_{j=1}^{U}(1-P(u_j|z_k)) \tag{7}$$

对式（7）求解为

$$P(z_k)=\frac{\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}P(z_k|l_i,u_j)}{\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}}$$

$$P(l_i|z_k)=\frac{\sum_{j=1}^{U}\theta_{l_i}^{u_j}P(z_k|l_i,u_j)}{\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}P(z_k|l_i,u_j)}$$

$$P(u_j \mid z_k)=\frac{\sum_{i=1}^{L}\theta_{l_i}^{u_j}P(z_k \mid l_i,u_j)}{\sum_{i=1}^{L}\sum_{j=1}^{U}\theta_{l_i}^{u_j}P(z_k \mid l_i,u_j)}$$

2.2　待分类样本的隐含信息属性的计算与求解

对于附加隐含信息属性值后的样本训练得到的支持向量分类机为 $f(L')$，利用 $f(L')$ 是无法对待分类样本进行分类的，因为它们不在同样的空间内，所以需要将待分类样本的隐含信息属性也求出来，使它们所处的映射空间一致。

对待分类样本集也可以采用上面的算法进行求解，但是为了减少计算量，采用近似的方法求解待分类样本集的隐含信息属性值，如图 1 所示。

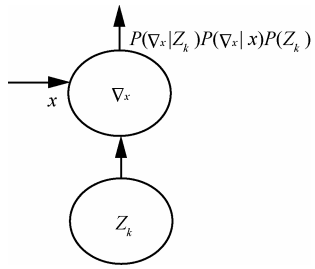

图 1　待分类样本隐含信息属性求解示意

对于待分类样本 x，利用它周围最近的 ∇_x 个已标记样本点作为估计 $P(z_k \mid x)$ 的依据，根据贝叶斯公式可以求得

$$P(z_k \mid x)=\frac{P(x \mid z_k)P(z_k)}{\sum_{t=1}^{K}P(x \mid z_t)P(z_t)}=\frac{\sum_{i,l_i\in\nabla_x}^{|\nabla_x|}P(l_i \mid z_k)P(x \mid l_i)P(z_k)}{|\nabla_x|\sum_{t}^{K}\sum_{i,l_i\in\nabla_x}^{|\nabla_x|}P(l_i \mid z_t)P(x \mid l_i)P(z_t)}$$

其中，$P(x \mid z_k)$ 表示待分类样本 x 最近的 ∇_x 个已标记样本点的平均值代替，即

$$\sum_{i,l_i\in\nabla_x}^{|\nabla_x|}\frac{P(l_i \mid z_k)P(x \mid l_i)P(z_k)}{\nabla_x}$$

$P(x \mid l_i)$ 表示待分类样本 x 与已标记样本 $l_i(l_i\in\nabla_x)$ 之间的关联度，它的值可以利用 2 个样本之间的欧式距离代替。

该算法中 K、∇_x 2 个参数值的确定十分关键，如果它们取值不当，就会降低半监督学习的效果。根据文献[15]的验证结果，一般 K 为所分类样本集数据种类的 2 倍比较合适，$|\nabla_x|$ 取较小的值比较合适。

3　实验分析

本节的实验包括 2 个部分，首先利用 UCI 数据库中的数据验证基于隐含属性的半监督学习方法的精度，然后将半监督学习方法应用于肺音数据的识别，验证其实际工程应用效果。

3.1　UCI 数据验证

本节从 UCI 数据库中选出 9 种数据，并采用交叉验证的方法分别对上述方法进行验证。半监督学习 SVM、SVM 和半监督 RF 实验数据及结果对比如表 2 所示。

表 2　半监督学习 SVM 与 SVM 和半监督 RF 实验数据

数据名称	已标记	未标记	测试样本	SVM	半监督 SVM	半监督 RF
Haberman*	156	100	50	75.22	75.78	75.24
Liver*	175	100	70	70.75	72.14	70.56
Breast	479	120	100	95.83	96.12	95.69
Ttt	558	200	200	93.50	93.54	89.15
German redit*	600	200	200	83.95	86.78	77.89
Spect heart	147	70	50	83.95	86.18	81.28
Chess	1 696	1 000	500	92.45	94.57	91.57
Sonar*	108	60	40	75.78	83.25	83.63

从实验分析得出，半监督学习的测试结果基本上都优于传统支持向量机测试的结果。差值的大小反映了已标记样本与未标记样本之间的关联度的大小，也就是无标记样本带有多少新的信息。经过实验分析，可以看出基于隐含信息属性的半监督学习方法是有效的。

3.2　肺音数据识别

3.2.1　实验数据

人体随时都在进行着呼吸运动，身体在同外界进行着换气行为，这一过程与肺部密切联系在一起，它们相互作用产生各种声音，这些声音中含有丰富的信息，有病理的也有生理的，这些信息对于研究肺部的病理音是一个重要的指标。肺音信号是一种非线性信号，具有复杂性、多样性、随机性等特点，因此如何采集与利用肺音数据，对人体肺部功能进行检测和监测是一项具有挑

战性的工作。

对于肺部声音分类识别的研究，多年来一直受到人们的高度关注，使用的识别方法不断涌现，对于肺音识别分类算法的分类依据也不统一，特别是随着人工智能技术的发展，有研究者已将分类技术应用于肺音识别领域。通过对肺音的分类来识别不同的肺音，是一种重要的肺音识别方法。对于分类方法而言，在训练分类器时，要求训练样本包含决策属性。可是肺音数据中会存在大量无标记的数据，即没有标记是否是肺部存在问题的数据，这些数据中包含了有价值的信息，需要挖掘提取。针对这个问题，本文采取基于 FSSL 方法的 SVM 和 RF 模型，实时吸收大量无标记样本所含有的新信息，提高 SVM 和 RF 应用于肺音识别领域的应用性能。

在此实验中利用肺音识别问题验证基于隐含信息的半监督学习方法的实际应用效果。肺音数据为来自于自主设计开发的肺音数据采集系统所采集的数据，在呼吸科专家的支持下，对所采集人群的地域、职业、年龄都进行了考虑，保证数据的真实性、有效性和代表性。采集人数：1 826 位，采集部位：9 个（喉　部（FT）、右上部（RU）、左上部（LU）、右中部（RM）、左中部（LM）、右下部（RD）、左下部（LD）、右后部（RB）、左后部（LB））采集总数据量达到 49 302 个样本数据。

3.2.2 肺音特征提取

在模式识别中，特征提取对于分类效果的好坏起着决定性作用。

从图 2 和图 3 中显示了几例正常和异常的上肺右侧肺音的频谱图，对比频谱图可以看出，正常和异常的肺部的肺音频谱是有很大差异的，频谱特征从原始信号直接变换而来，原始信息保留较好，因此在本文的研究中采用频谱图中的特征作为肺音识别特征。

在实验中对正常与异常的上肺部所对应肺音的最大功率处的频率（PF）、平均频率（MNF）、肺音总强度（LSI）、总功率的 25%、50%、75% 处的频率 Q25、Q50、Q75 差异进行了分析。从实验可以看出，在这些参数上 2 类肺音的数据存在着一定的差异，如图 4 所示，因此在本文中选取 PF、MNF、Q25、Q50、Q75 作为肺音识别的特征参数。

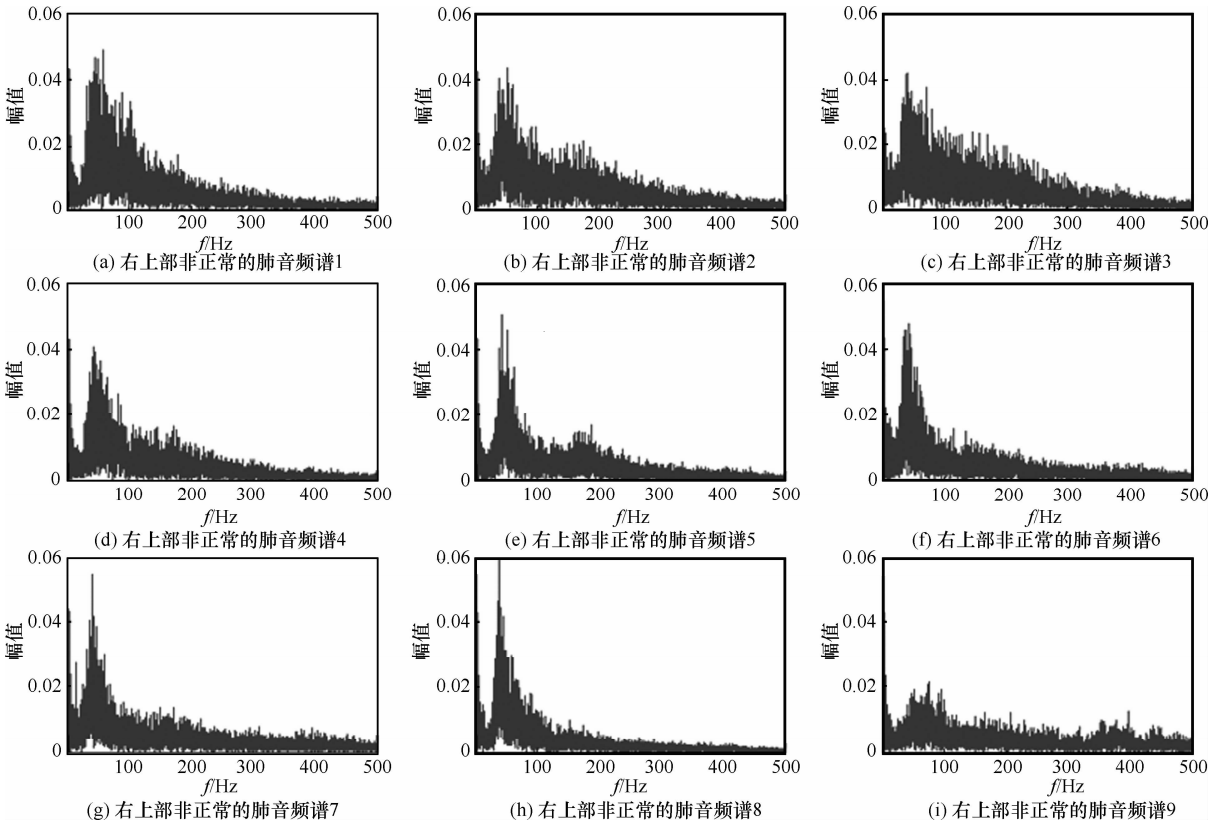

图 2　右上部非正常(RU_nonNormal)的肺音频谱

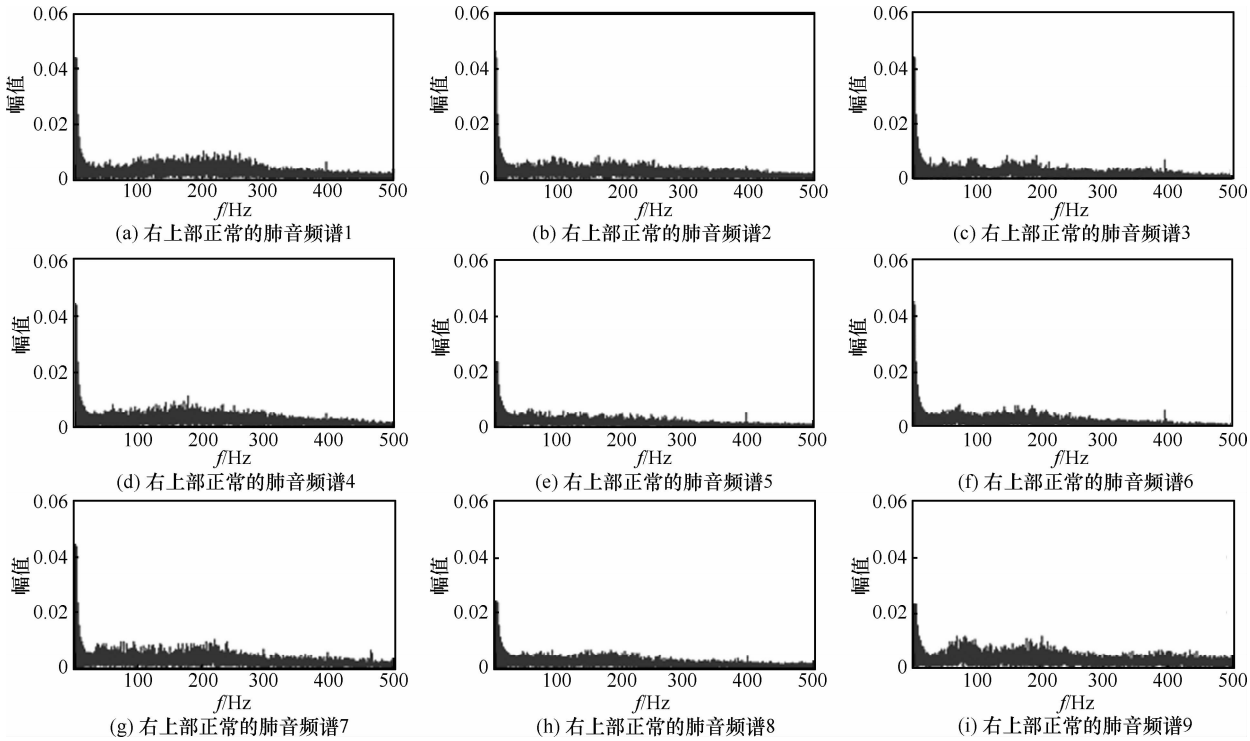

图 3　右上部正常(RU Normal)的肺音频谱

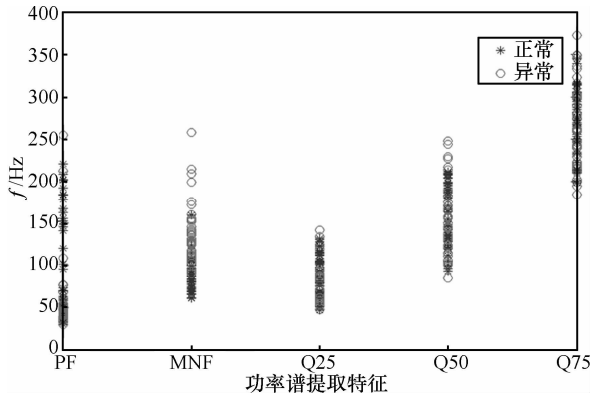

图 4　RU 数据功率谱的特征分布

表 3　　SVM 与半监督 SVM 实验数据

数据	SVM 精度/%	半监督 SVM 精度/%	半监督 RF 精度/%
FT	72.50	76.67	73.67
LB	72.18	74.44	74.49
LD	65.56	66.73	65.23
LM	69.58	71.91	70.67
LU	60.42	61.46	61.05
RB	63.33	69.44	67.31
RD	76.67	78.92	79.56
RM	72.22	75.00	75.04
RU	87.96	88.77	89.65

3.2.3　实验分析

1) 识别效果分析

在此实验中将半监督 SVM 模型及半监督 RF 模型应用于肺音识别领域。在此实验中利用 20%的样本作为有标记样本；将 30%的样本的决策属性去掉，作为无标记样本；30%的样本作为测试样本。实验结果如表 3 所示，从实验结果可以看出由于新信息的增加，半监督 SVM 的精度比标准 SVM 有一定的提高。

2) ROC 曲线

根据实际应用的需求，对于肺音识别问题，人们总是希望不要漏掉肺部有问题的样本。因此在实验中设不正常的肺音为正类，正常的肺音为负类。TP 表示正确识别出不正常肺音样本的能力，即命中率；FP 表示错分正常肺音样本的情况，即误报率。

下面通过 ROC 曲线分析肺音识别时命中率和误报率之间的关系，其中

$$FP=\frac{\text{错分负例个数}}{\text{负例总数}}$$

$$TP=\frac{\text{正确分类正例个数}}{\text{正例总数}}$$

从表 4 可以看出，基于 FSSL 的半监督学习技术，可以提高命中率，即识别非正常肺音样本的能力，但对误报率没有改变。

表 4　　命中率和误报率的关系

RU	*TP*/%	*FP*/%
SVM	91.75	12.78
SEMI_SVM	93.44	12.78
SEMI_RF	92.19	12.78

3) 影响因素分析

下面利用上肺的数据分析无标记样本的数量与半监督 SVM 性能的关系，实验结果如表 5 所示。从实验结果可以看出，只有无标记样本量增加到一定程度时，才对半监督 SVM 的性能有影响。

表 5　半监督 SVM 分类精度与无标记样本数量的关系

无标记样本量/%	半监督 SVM 分类精度/%
5	87.96
10	87.96
15	87.99
20	88.03
25	88.25
30	88.77

半监督 SVM 的性能并不一定高于 SVM，其性能还取决于无标记样本是否能提供新的信息量。为了验证这个性质，在原始 SVM 所得的支持向量的周围生成一些随机样本，作为无标记上肺肺音样本，利用这些样本进行半监督学习，实验汇总结果如表 6 所示。实验结果表明离支持向量越近的点所能提供的新信息越少，对半监督 SVM 的性能几乎没有影响。但是随着距离的增加，使样本点落入另外一类，从而增加了错分点，也使半监督 SVM 性能下降。由此可见，半监督 SVM 分类精度与无标记样本质量关系密切，对实际应用具有一定的影响。

表 6　半监督 SVM 分类精度与无标记样本质量的关系

随机样本距离支持向量距离	半监督 SVM 分类精度
0.01	87.96
0.05	88
0.10	88.23
0.50	87.46

4　结束语

本文研究了基于隐含信息属性的半监督学习方法，并将其应用于 SVM 和 RF 模型。为了验证基于隐含信息的半监督 SVM 和半监督 RF 算法，首先利用 UCI 数据库对算法的精度进行了验证，然后利用实际采集的肺音数据的识别问题验证了其解决实际工程问题的效果。同时分析了未标记样本的数量以及未标记样本的质量与基于隐含信息的半监督学习方法的关系。可以看出，无标记样本的数量和质量对于此方法的实际应用效果影响较大，无标记样本数量对于分类性能的影响比其质量影响要大。因此，在保证具有一定数量的无标记样本的情况下，此方法对于实际应用是比较适合的，能够在实际应用中发挥其独特的优势。

参考文献：

[1] YANG L X, YANG S Y, Semi-supervised hyperspectral image classification using spatio-spectral laplacian support vector machine[J]. IEEE Geoscience and Remote Sensing Letters, 2014, 11(3): 651-656.

[2] LIU L C, HSAIO W H, LEE C H, *et al*. Semi-supervised text classification with universum learning[J]. IEEE Transactions on Cybernetics, DOI: 10.1109/TCYB.2015.2403573, 2015.

[3] ONOFREY J A, Low-dimensional non-rigid image registration using statistical deformation models from semi-supervised training data[J]. IEEE Transactions on Medical Imaging, DOI: 10.1109/ TMI. 2015.2404572, 2015.

[4] ZHOU Z H. Co-training paradigm in semi-supervised learning[A] Proceeding of the Chinese Workshop on Machine Learning and Applications[C]. Nanjing, China，2007. 261-267.

[5] ZHOU D Y，BOUSQUET O，LAL T N，*et al*. Learning with local and global consistency[J]. Advances in Neural Information Processing System, 2004，3(21):23-28.

[6] WU C M，WANG X D, BAI D Y. Fast incremental learning algorithm of SVM on KKT conditions[A]. 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery[C]. 2009. 551-554.

[7] JAKKOLA T, HAUSSLER D. Exploiting generative models in discriminative classifiers [A]. Advances in Neural Information Processing Systems[C]. Cambridge, MA: The MIT Press, 1999. 487-493.

[8] FRALICK S C. Learning to recognize patterns without a teacher[J]. IEEE Transactions on Information Theory, 1967,13(1):57-64.

[9] AGRAWALA A K. Learning with a probabilistic teacher[J]. IEEE Transactions on Information Theory, 1970, 16(4):373-379.

[10] HOLUB A，WELLING M, PERONA P. Exploiting unlabeled data for hybrid object classification[A]. Proc of the 17th Annual Conference on Neural Information Processing Systems[C]. 2005.165-171.

[11] WANG W，ZHOU Z H. Analyzing co-training style algorithms[A]. Proc of the 18th Conference on Machine Learning[C]. 2007.454-465.

[12] KUBAT M, HOLTE B C, MATWIN S. Machine learning for the detection of oil spills in satellite radar images[J]. Machine Learning，1998,30(2):195-215.

[13] JOACHIMS T. Transductive inference for text classification using support vect ormachines[A]. Machine Learning-international Workshop Then Conference[C]. Morgan Kaufmann Publishers, INC, 1999. 200-209.

[14] BLUM A, LAFFERTY J, RWEBANGIRA M R, *et al*. Semi- supervised learning using randomized mincuts[A]. Proceedings of the Twenty-first International Conference on Machine Learning[C]. ACM, 2004. 13.

[15] ZHU X Q. Cross-domain semi-supervised learning using feature formulation[J].IEEE Transactions on Systems MAN, and Cybernertics-Part B: Cybernetics, 2011, 41(6):1627-1638.

作者简介：

刘国栋（1966-），男，山东乐陵人，南开大学博士生，主要研究方向为软件工程、模式识别。

许静（1967-），女，天津人，博士，南开大学教授、博士生导师，主要研究方向为大数据分析、软件安全、软件测试等。

张国兵（1979-），男，河北邢台人，硕士，主要研究方向为自动化测试、智能识别。