中图法分类号: TP391.41 文献标志码: A 文章编号: 1006-8961(2011)11-2036-05
论文索引信息:朱明旱, 邵湘怡, 罗大庸. 人脑半监督学习机理分类法[J]. 中国图象图形学报,2011,16(11):2036-2040

人脑半监督学习机理分类法

朱明旱[1)],邵湘怡[1)],罗大庸[2)]
[1)](湖南文理学院电气与信息工程学院,常德 415000) [2)](中南大学信息科学与工程学院,长沙 410083)

摘 要: 针对 NN(nearest neighbor)和 kNN(k-nearest neighbor)方法在标记样本较少时,分类正确率不高的缺陷,根据人脑分类样本时,自觉地利用未标记样本的半监督学习机理,提出一种人脑半监督学习机理分类方法。该方法利用未标记样本间的近邻关系,减少了标记样本数量对分类正确率的影响程度。在 MNIST 手写体数字库和 ORL 人脸库上的样本分类实验表明,在标记样本数较少的情况下,该方法的分类正确率比 NN 和 kNN 方法高得多。

关键词: NN 分类方法;半监督学习机理;半监督分类

Classification method using human brain semi-supervised learning mechanism

Zhu Minghan[1)], Shao Xiangyi[1)], Luo Dayong[2)]
[1)](*College of Communication and Electric Engineering, Hunan University of Arts and Science, Changde* 415000 *China*)
[2)](*College of Information Science and Engineering, Central South University, Changsha* 410083 *China*)

Abstract: Aimed at the problem that nearest neighbor method and k-nearest neighbor method can't obtain better classification effectiveness when there aren't enough labeled examples, a semi-supervised classification method is proposed in this paper. The method is based on the mechanism that unlabeled samples were used if human classify pattern involuntary. The method utilizes the nearest neighbor relationship between unlabeled samples to reduce the influence of the number of labeled samples on classification accuracy. The experimental results using the MNIST database of handwritten digits and the ORL face database show the method has higher classification accuracy than the nearest neighbor method and the k-nearest neighbor method if there aren't enough labeled samples.

Keywords: nearest neighbor classification method; semi-supervised learning mechanism; semi-supervised classification

0 引 言

在模式识别领域中,NN(nearest neighbor)[1]和 kNN(k-nearest neighbor)[2]这两种非参数化方法已被广泛应用于样本分类。实际上,kNN 是 NN 的一个推广。NN 规则在对样本进行分类时,将测试样本归为与它最接近的那个训练样本的类。kNN 方法则将测试样本归为与它最接近的 k 个近邻训练样本中,出现最多的那个类别。为了减小 k 值对分类正确率的影响,Hechenbichler 等人[3]提出基于距离的加权 kNN 方法。该方法根据各近邻样本到测试样本的距离大小,赋予 k 个近邻样本不同的权值。距离越小权值越大,相反,距离越大,所赋的权值越小。这样,近邻样本与测试样本的相似程度,就能通过权值的大小来体现。即使 k 值很大,对测试样本

收稿日期:2010-09-27;**修回日期**:2010-11-15
基金项目:湖南省教育厅优秀青年基金项目(10B074);校级优秀青年基金项目(YXQ0905)。
第一作者简介:朱明旱(1974—),男,副教授。2009 年于中南大学获控制科学与工程专业博士学位,主要的研究方向为图像处理、模式识别。E-mail:zhumh_123@163.com。

分类起决定作用的,仍是与它相距较近的那些样本。这种加权 kNN 方法的分类准确率,对 k 值的选取不再敏感,表现出了较好的鲁棒性。为了突出重要特征在分类中所起的作用,一些研究者们又提出基于特征的加权 kNN 方法[4-7]。前期我们也对 kNN 方法进行了研究,并提出针对序列样本分类的加权 kNN 方法[8]。

但是,无论是 NN、kNN 还是加权 kNN 方法,为了获得较好的分类效果,都需要有足够多的标记样本(理论上要求标记的样本无限多),否则分类的正确性很差[9]。图 1 展示了标记样本个数对样本分类的影响。图中"●"和"★"分别表示不同类别的标记样本,"○"和"☆"分别表示相应类别的测试样本。(a)中,由于标记样本比较充足,NN、kNN 和加权 kNN 都可以将"①"正确地归为"●"所属的类;而(b)中,由于标记样本不够多,"①"被错误地归为了"★"所属的类。我们知道,在许多模式分类中,获得足够多的标记样本是相当困难的。这时用 NN、kNN 和加权 kNN 方法,就很难取得较好的分类效果。

(a) 标记样本充足

(b) 标记样本不够

图 1　标记样本个数对样本分类的影响

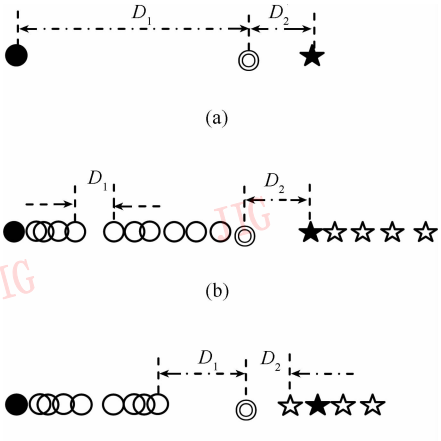
Fig. 1　The influence of the number of labeled samples on classification accuracy

由于非标记样本的获得相对比较容易,利用非标记样本来提高分类正确率的半监督分类方法引起了许多学者的关注,并成了模式分类领域的研究热点。根据决策边界不应在高密度区,而应在低密度区的假设,Joachims 提出 TSVM(transductive suppor vector machines)[10]。但由于精确地对 TSVM 求解是一个 NP-hard 问题,它在实际应用中不得不用近似算法来处理[11-12]。根据处于相同簇中的相邻样本具有相同类别的假设,研究者们提出一些基于图的分类方法[13-15]。但是这些方法有的不能直接分类在训练中没有出现的样本[13-14],有的则涉及最优参数的估计[15]。针对序列样本的分类,Li Wei 等人则提出一种半监督 NN 方法[16],通过不断把距标记样本最近的那个未标记样本加入训练集中,从而实现未标记样本的分类。但是,这个不断加入的过程何时应该停止,Li Wei 等人也没有找到一个可执行的标准。

根据人脑的半监督学习机理,并利用 NN 分类方法的思想,提出一种人脑半监督学习机理分类法。该方法既不需要待测样本参与训练,也不涉及最优参数估计,是一种非参数的半监督分类方法。在标记样本较少的情况下,能够利用未标记样本间的邻近关系,提高分类的正确率。

1　人脑半监督学习机理

Zhu 等人发现人脑分类空间中的样本时,自觉地考虑了未标记样本的分布状况,执行的是半监督分类[17],这里我们用图 2 简要地说明人脑的这种半监督学习机理。在图 2 中,"●"和"★"分别表示两类已标记样本,"○"和"☆"分别表示未标记样本,"◎"表示测试样本。(a)中只有两个标记样本和一个测试样本,(b)(c)中还有一些未标记样本,但它们的分布状况不一样。这里考虑对测试样本"◎"的分类,观察后我们会将(a)(c)中的"◎"归为"★"所对应的类,而将图(b)中的"◎"归为"●"所对应的类。



(a)

(b)

(c)

图 2　未标记样本对分类的影响

Fig. 2　The influence of unlabeled samples on classification accuracy

在图 2(a)—(c)中,测试样本、已标记样本,以及测试样本与标记样本的距离均一样,但分类结果却不完全一样。可见,人脑在对测试样本"◎"进行分类决策时,并没有直接将该样本与各标记样本的距离作为分类的依据。实际上,人脑在对测试样本"◎"进行分类决策时,是以图中的距离 D_1 和 D_2 作为分类依据的,也就是把测试样本"◎"到标记样本

“●”和“★”的路径上，最稀疏处的样本距离作为了分类依据，这就是人脑的半监督学习机理。如果存在着 p 条从“◎”到“●”的路径，且每条路径上最稀疏处样本间的距离为 $d_i(i=1,\cdots,p)$，那么人脑会将 d_i 中的最小值作为 D_1，即

$$D_1 = \min\{d_i, i=1,\cdots,p\} \tag{1}$$

不难发现，若用人脑的这种半监督学习机理来对图 1 中的测试样本“①”分类，即使所标记的样本不够多，如(b)所示，也仍不会将它错判为“★”所属的类。下面介绍本文根据人脑的半监督学习机理所提出的分类方法。

2 半监督分类方法

为了较好地阐述该方法的思想实质，这里先定义样本与样本间距离，以及样本与集合间距离的度量函数。用欧氏距离来度量两样本间距离，用样本到集合中各样本欧氏距离的最小值作为样本与集合之间的距离。它们各自的定义如下：

定义 1 d 维空间中的两个向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$，它们间的欧氏距离为

$$d(\boldsymbol{a},\boldsymbol{b}) = [\sum_{l=1}^{d}(a_l - b_l)^2]^{1/2} \tag{2}$$

这里 a_l 和 b_l 分别表示向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的第 l 个分量。

定义 2 设 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ 为 n 个 d 维向量组成的向量集，向量 $\boldsymbol{a}$ 与向量集 $\boldsymbol{X}$ 的距离为

$$D(\boldsymbol{a},\boldsymbol{X}) = \min\{d(\boldsymbol{a},\boldsymbol{x}_i), i=1,\cdots,n\} \tag{3}$$

提出的半监督分类方法，首先计算出待测样本与各类训练集的距离；然后根据未标记样本间的近邻关系，不断地更新测试样本与各类训练集的距离；最后根据得到的最终距离确定出测试样本的类别。

设 $\{(\boldsymbol{L}_j,\omega_j), j=1,\cdots,m\}$ 为 m 个不同类别的训练集，ω_j 是训练集 $\boldsymbol{L}_j$ 的类别标记。$\boldsymbol{U}=\{(\boldsymbol{a}_i,y_i), i=1,\cdots,n\}$ 为测试集，y_i 为 $\boldsymbol{a}_i$ 的类别标记，均是待测的。用该分类方法判断样本 $(\boldsymbol{a}_i, 1\leqslant i\leqslant n)$ 类别的步骤如下：

1)根据定义 2 的向量与向量集的距离，求出 $\boldsymbol{a}_i$ 到所有训练集 $\boldsymbol{L}_j$ 的距离 $\{D(\boldsymbol{a}_i,\boldsymbol{L}_j), (j=1,\cdots,m)\}$；

2)将测试样本 $\boldsymbol{a}_i$ 从测试集 $\boldsymbol{U}$ 中去掉，剩余测试样本组成的集合记为 $\boldsymbol{U}'$；

3)利用测试集 $\boldsymbol{U}'$ 中的样本，更新 $\{D(\boldsymbol{a}_i,\boldsymbol{L}_j), (j=1,\cdots,m)\}$ 值，更新算法如下：

(1) begin initialize $\boldsymbol{a}\leftarrow\boldsymbol{a}_i$, $\{D(\boldsymbol{a},\boldsymbol{L}_j)\leftarrow D(\boldsymbol{a}_i,\boldsymbol{L}_j),(j=1,\cdots,m)\}$, $\boldsymbol{U}'$

(2) do 从 $\boldsymbol{U}'$ 中找出 $\boldsymbol{a}$ 的最近邻样本 $\boldsymbol{a}'$，求出 $\boldsymbol{a}'$ 到 $\boldsymbol{a}$ 的欧氏距离 $d(\boldsymbol{a},\boldsymbol{a}')$，同时也计算 $\boldsymbol{a}'$ 到所有训练集 $\boldsymbol{L}_j$ 的距离 $\{D(\boldsymbol{a}',\boldsymbol{L}_j), j=1,\cdots,m\}$

(3) If $d(\boldsymbol{a},\boldsymbol{a}')\leqslant\min\{D(\boldsymbol{a},\boldsymbol{L}_j), (j=1,\cdots,m)\}$

(4) then $D(\boldsymbol{a},\boldsymbol{L}_j)\leftarrow\min\{D(\boldsymbol{a},\boldsymbol{L}_j),D(\boldsymbol{a}',\boldsymbol{L}_j)\},(j=1,\cdots,m)$

(5) else 转移到第 8 步

(6) $\boldsymbol{a}\leftarrow\boldsymbol{a}'$，并将 $\boldsymbol{a}'$ 从 $\boldsymbol{U}'$ 中去掉

(7) until $\boldsymbol{U}'$ 为空集

(8) return $\{D(\boldsymbol{a},\boldsymbol{L}_j),(j=1,\cdots,m)\}$

(9) end

4)根据最终的 $\{D(\boldsymbol{a},\boldsymbol{L}_j),(j=1,\cdots,m)\}$ 值，确定 $\boldsymbol{a}_i$ 所属的类别，即

$$y_i = \arg\min_{\omega_c}\{D(\boldsymbol{a},\boldsymbol{L}_j),(j=1,\cdots,m)\} \tag{4}$$

3 实验与分析

我们分别用 NN、距离加权 kNN 和本文提出的半监督分类方法，在 MNIST 手写体数字库[18]和 ORL 人脸库[19]上进行了字符和人脸的识别实验。在 MNIST 数字库上进行了两类样本的分类实验，而在 ORL 人脸库上则进行了多类样本的分类实验。

3.1 实验 1

MNIST 数字库共包含手写体数字“0～9”的 60 000 个训练样本和 10 000 个测试样本，每个样本均为 28×28 的灰度图像，图 3 为该数据库中的 20 个样本。

图 3 MNIST 数据库中的样本图像

Fig. 3 Sample images from MNIST database

从 MNIST 数据库中，取出“2”和“3”的图像各 1 200幅，共 2 400 幅。先用 2DPCA(two-dimensional principal component analysis)[20]将所有的图像都降到 140 维，然后进行 8 次字符识别实验。第 1 次取每个字符的 10 幅图像，共 20 幅图像作为标记样本

集,剩余的 2 380 幅为测试集。第 2 次取每个字符的 20 幅图像,共 40 幅图像作为标记样本集,剩余的 2 360 幅为测试集。其余各次的标记样本数、测试样本数,以及每次实验的识别率数据如表 1 所示。

表 1　在 MNIST 数据库上 NN、kNN 与本文方法的识别率(k=5)

Tab. 1　Recognition accuracy of NN, kNN and method proposed in the paper on the MNIST database

/%

标记样本数	20	40	60	80	100	400	1 200	2 000
测试样本数	2 380	2 360	2 340	2 320	2 300	2 000	1 200	400
NN	91.43	93.14	93.59	93.62	94.83	96.80	97.67	98.50
kNN	90.55	89.58	91.88	92.97	94.43	95.93	96.41	97.34
本文方法	97.05	97.12	96.92	97.50	98.50	97.05	97.67	98.50

3.2　实验 2

ORL 人脸库由 40 人,每人 10 幅图像组成,图像大小为 112×92 像素。其中有些图像拍摄于不同时期,脸部表情和脸部细节有着不同程度的变化,例如,笑与不笑,眼睛睁与闭,眼镜戴与不戴;人脸深度旋转和平面旋转可达 20°,人脸的尺度变化多达 10% 。图 4 为 ORL 人脸库中的 4 幅样本图像。

图 4　ORL 人脸库中的样本图像

Fig. 4　Sample images from ORL face database

先用 2DPCA 将所有的图像都降到 1 120 维,然后进行了 9 次人脸识别实验。第 1 次将每人的第 1 幅图像,共 40 幅作为标记样本集,剩余的 360 幅作为测试集。第 2 次将每人的前 2 幅图像,总共 80 幅作为标记样本集,剩余的 320 幅作为测试集。直到第 9 次将每人的前 9 幅图像,总共 360 幅作为标记样本集,剩余的 40 幅作为测试集。表 2 为每次实验的标记样本、测试样本,以及对应的识别率数据(其中 kNN 的 k 取最优值)。

3.3　实验分析

在实验 1 的 8 次字符识别实验中,本文提出方法的识别率最高,NN 的次之,kNN(k = 5)的最低。提出方法的识别率比 NN 的分别高出了 5.62%、3.98%、3.33%、3.88%、3.67%、0.25%、0、0。在标记样本不超过 100 的情况下,该方法的识别率比 NN 的高出了 3% 以上。特别是在标记样本数只有 20 个的情况下,比 NN 的高出了 5.62% 。而当标记样本数为 1 200 和 2 000 时,两方法的识别率一样。

表 2　在 ORL 人脸库上 NN、kNN 与本文方法的识别率

Tab. 2　Recognition accuracy of NN, kNN and method proposed in the paper on the ORL face database

/%

标记样本数	40	80	120	160	200	240	280	320	360
测试样本数	360	320	280	240	200	160	120	80	40
NN	71.67	81.25	84.64	88.33	90.50	96.25	96.67	96.25	95.00
kNN	71.67	81.25	84.64	88.33	90.50	96.25	96.67	96.25	95.00
本文方法	74.17	85.00	87.14	90.00	96.00	96.88	96.67	96.25	95.00

在实验 2 的 9 次人脸识别实验中,提出方法的识别率比 NN 和 kNN(k 取最优值)的分别高出了 2.5%、3.5%、2.5%、1.67%、5.5%、0.63%、0、−0.42%、0。在标记样本没超过 280 时,该方法的识别率比 NN 和 kNN 的都高。当标记样本数为 320 时,比 NN 和 kNN 的稍低了 0.42%,标记样本为 360 时,三者的识别率一样。

综合实验 1 和实验 2 的结果,不难得到这样的结

论,在标记样本较少的情况下,该方法体表现出了一定的优越性,与NN和kNN相比,它的识别效果要好。在标记样本比较充足的情况下,该方法的识别效果与NN和kNN(k取最优值)差别十分小。

4 结 论

针对NN和kNN方法在标记样本较少时,分类正确率不高的问题,提出一种人脑半监督学习机理分类法。该方法在对测试样本进行分类时,利用未标记样本间的近邻关系,缩小了测试样本到训练集的距离,使分类的决策依据更加合理。从而使得在标记样本数较少的情况,仍然能够获得比较好的分类效果。实质上,本文的半监督分类方法,通过利用测试集中的未标记样本,减小了样本分类时,分类正确性对训练集大小的依赖程度。

参考文献(References)

[1] Cover T M, Hart P E. Nearest neighbor pattern classification [J]. IEEE Transactions on Information Theory, 1967, 13(1): 21-27.

[2] Edward A P, Frederick P F. A generalized k-nearest neighbor rule [J]. Information and Control, 1970, 16(2): 128-152.

[3] Hechenbichler K, Schliep K. Weighted k-nearest neighbor techniques and ordinal classification: discussion paper 399 [R]. Munich, Germany: Ludwig-Maximilians University, 2004.

[4] Chen Zhenzhou, Li Lei, Yao Zhengán. Feature-weighted k-nearest neighbor algorithm with SVM [J]. Acta Scientiarum Naturalium Universitatis Sunyatseni, 2005, 44(1): 17-20. [陈振洲,李磊,姚正安. 基于SVM的特征加权kNN算法[J]. 中山大学学报:自然科学版,2005,44(1): 17-20.]

[5] Liu Ming, Yuan Baozong, Tang Xiaofang. A new approach to determine the similarity parameters in evidence-theoretic k-NN rule [J]. Acta Electronica Sinica, 2005, 33(4): 766-768. [刘明,袁保宗,唐晓芳. 证据理论k-NN规则中确定相似度参数的新方法[J]. 电子学报,2005,33(4): 766-768.]

[6] Vivencio D P, Hruschka E R, Nicoletti M C, et al. Feature-weighted k-nearest neighbor classifier [C]//Proceedings of the IEEE Symposium on Foundations of Computational Intelligence. Washington DC, USA: IEEE Communications Society, 2007, 481-486.

[7] Sun Yan, Lv Shipin, Tang Yiyuan. No previous ordering for kNN algorithm [J]. Journal of Chinese Computer Systems, 2008, 29(4): 682-686. [孙岩,吕世聘,唐一源. 无先序条件约束的kNN算法[J]. 小型微型计算机系统,2008,29(4): 682-686.]

[8] Zhu Minghan, Luo Dayong. A sequential weighted k-nearest neighbor classification method [J]. Acta Electronica Sinica, 2009, 37(11): 2584-2588. [朱明早,罗大庸. 一种序列的加权kNN分类方法[J]. 电子学报,2009,37(11): 2584-2588.]

[9] Duda R O, Hart P E, Stork D G. Pattern Classification [M]. 2 ed. Beijing: China Machine Prees, 2003, 146-151. [Duda R O, Hart P E, Stork D G. 模式分类[M]. 2版. 李宏东,姚天翔译. 北京: 机械工业出版社,2003,146-151.]

[10] 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.

[11] Bennett K P, Demiriz A. Semi-supervised support vector machines [C]//Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 1999, 368-374.

[12] Chapelle O, Zien A. Semi-supervised classification by low density separation [C]// Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. New Jersey: The Society for Artificial Intelligence and Statistics, 2005, 57-64.

[13] Chapelle O, Weston J, Schölkopf B. Cluster kernels for semi-supervised learning [C]//Proceedings of the 2003 Conference on Advances in Neural Information Processing Systems 15. Cambridge, MA: MIT Press, 2003, 585-592.

[14] Belkin M, Matveeva I, Niyogi P. Regularization and semi-supervised learning on large graphs [C]// Proceedings 17th Annual Conference on Learning Theory. Berlin, Heidelberg, Germany: Springer Press, 2004, 624-638.

[15] Balaji K, David W, Ya X, et al. On semi-supervised classification [C]//Proceedings of Neural Information Processing Systems Conferences, Vancouver, Canada: British Columbia, 2005, 721-728.

[16] Li Wei, Eamonn Keogh. Semi-supervised time series classification [C]//Proceedings of the 12th ACM SIGKDD International, ACM. New York, NY, USA, 2006, 748-753.

[17] Zhu Xiaojin, Timothy Rogers, Ruichen Qian, et al. Humans perform semi-supervised classification too [C]// Proceedings of the 22nd National Conference on Artificial Intelligence. Menlo Park, Calif: AAAI Press, 2007, 864-869.

[18] Yann Lecun, Corinna Cortes. The MNIST Database of Handwritten Digits [EB/OL]. [2010-04-15]. http://yann.lecun.com/exdb/mnist/.

[19] AT & T Laboratories Cambridge. The ORL Database of Faces [EB/OL]. [2010-04-15]. http://www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html/.

[20] Yang J, Zhang D. Two-dimensional PCA: a new approach to appearance-based face representation and recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(1): 131-137.