一种基于均值更新的分类模型①

冯进玫[1,2]，卢志茂[1]，陈纯锴[1,2]

[1](哈尔滨工程大学 信息与通信工程学院, 哈尔滨 150001)
[2](黑龙江科技学院 电气与信息工程学院, 哈尔滨 150027)

摘 要：最小距离分类法和最近邻分类法是最简单、快速、有效的分类方法，但对噪声较敏感，对于训练样本很少或训练样本偏离类中心较远时，分类效果较差。针对这一问题,提出了基于均值更新（MU）的分类模型，通过不断扩大训练样本并更新均值中心来改善对测试数据的分类效果; 并在此基础上提出了基于均值更新的最小距离（MU-MD）分类模型，利用 MU 的分类结果重新计算各类的均值，然后采用最小距离法对所有测试样本重新进行划分，以确定最终的类别归属，这样可以部分纠正 MU 分类过程中的错分，进一步提高分类效果。

关键词：最小距离分类法; 均值更新; 训练样本; 测试样本

Classification Model Based on the Mean Update

FENG Jin-Mei[1,2], LU Zhi-Mao[1], CHEN Chun-Kai[1,2]

[1](College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China)
[2](College of Electric and Information Engineering, Heilongjiang Institute of Science and Technology, Harbin 150027, China)

Abstract: The minimum distance classification algorithm and the nearest neighbor classification algorithm are the simplest, most rapid and most effective classification methods, and they are more sensitive to the noise. But to the training samples in few or the training samples that are far from the cluster center, the classification results is poor. To solve this problem, this paper proposes a classification model based on the mean update (MU), by expanding the training sample and updating the mean center to improve the classification results of the test data; and on this basis, it proposes the MU-based minimum distance (MU-MD) classification model, and uses the MU's classification results to recalculate the mean of all test samples, then all test samples are re-divided by using the minimum distance method, so as to determine the final category attribution. This can partially correct misclassification in the MU category process and further improve the classification results.

Key words: the minimum distance classification algorithm; mean update; training samples; test samples

随着网络信息的迅猛发展，人们获取的信息量越来越大，信息分类对获得有用信息也显得越来越重要。分类是数据挖掘和机器学习中的一个重要研究课题，分类可以看作是一种决策，根据训练样本的属性对测试样本做出应属哪一类的决策。目前经典分类方法主要有最小距离、决策树、神经网络、k-最近邻、贝叶斯、支持向量机等方法，其中最小距离（MD）分类法和最近邻（NN）分类法是最简单、快速、易于实现的分类方法。在图像处理、语音识别及文本分类等领域普遍使用 MD、NN 及其改进算法[1-3]。MD 分类法的缺点是如果所选择的代表点并不能很好地代表各类，分类的错误率将很大。目前出现了很多改进算法，如通过加权最小距离选择更有效的距离度量[4]、采用遗传算法及采用自适应算法等方法[5]。NN 分类法是非参数法中最重要的方法之一，其分类效果取决于所获得的每类训练样本的空间分布，当训练样本的数量不足以表达全部样本的空间分布时，分类错误率较高[6]。通常分类方法都是针对训练样本较大的情况，所做实

① 收稿时间:2012-02-16;收到修改稿时间:2012-03-12

验的训练样本占到总样本的 20%以上，对于训练样本极少如仅占总样本的 4%或更少、训练样本偏离样本中心的情况的研究很少。本文针对这种情况及 MD 和 NN 分类器的不足提出了一种新的分类模型——均值更新(Means Update，MU)分类模型。在此模型的基础上进一步提出基于均值更新的最小距离(MU-MD)分类模型。

1　均值更新分类

1.1 均值更新分类模型

人类认识的客观世界包括已知世界和未知世界两部分。客观世界中的每一个事物都可以作为一个样本，每个样本都有一个用于表征本身类别归属的特征向量，第 i 个样本表示为 $X_i=(x_{i1}, x_{i2}, \cdots, x_{im})$，其中 x_{im} 表示第 i 个样本的第 m 个属性，将每个样本理解为样本空间的一个点。对于一个用于分类的数据集，本文称其所有的样本组成的空间为全样本空间（Complete sample space，CSS），记为 $CSS=\{X_i | i=1,2, \cdots, S_c\}$。将全样本空间中用于训练的样本组成的训练集作为训练样本空间（Training sample space，TRSS），用于测试的样本组成的训练样本集作为训练样本空间（Test sample space，TESS）。CSS 的大小为 S_c，TRSS 的大小为 S_{tr}，TESS 的大小为 S_{te}，显然有，$S_c = S_{tr} + S_{te}$。

人类对客观世界认识的初期是已知世界小于未知世界。随着认识程度的深入和范围的扩大，已知世界不断增大，而相对的未知世界不断减小。根据对客观世界认识的这一自然规律，本文在分类研究中选取的训练样本空间均小于测试样本空间，即 $S_{tr} < S_{te}$。通常增大与 TRSS 相对应的已知世界，也会提高其与客观世界的相似性，TRSS 的大小越接近于 CSS，TRSS 与 CSS 就越相似。已知世界到客观世界的相似性可以用距离衡量。同理，测试样本与训练样本之间的距离越小，其相似性就越大。很多文献通过实验得出的结论是训练数据规模增大后，分类精度也随之提高。但是通常人工增大训练数据的规模很困难，所以通常的处理策略是作为 TRSS 的已知世界保持大小不变，也就是说每个测试样本的类别决策只能从固定的 TRSS 里寻找答案。本文根据不断增大已知世界就可以不断增多信息并提高决策置信度的道理，设计了一种自动增大 TRSS 的分类模型——均值更新（MU）分类模型。MU 分类模型如图 1 所示，对每一个测试样本进行分类之后，将其放入到训练样本集中。随着分类的不断进行，训练样本不断增加，即对客观世界的认识范围不断扩大，认识程度不断加深，对样本进行分类的准确率呈上升趋势。本文提出的基于均值更新的最小距离(MU-MD)分类模型如图 2 所示，借助 MU 分类模型得到的分类结果，重新计算类均值，采用最小距离分类法对测试样本重新分类。此分类模型可以部分纠正 MU 分类过程中的错分，进一步提高分类正确率。

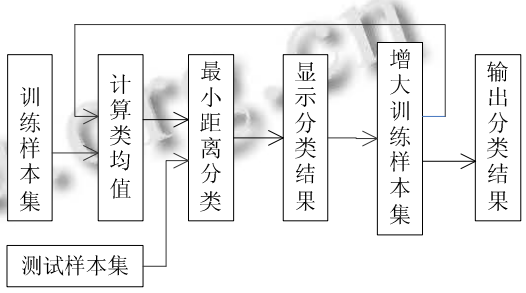


图 1　MU 分类模型

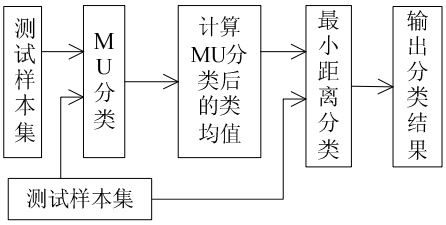


图 2　MU-MD 分类模型

1.2 均值更新分类的实现

TRSS 中每一个类别都对应一个已知世界，假设每类样本有标明类别的训练样本 N_i 个，i=1,2，…，c。c 表示类别数。每个类的训练样本 X_{is} 作为已知世界的初始样本，X_{is} 表示第 i 个模式类 ω_i 的第 s 个训练样本。已知世界以自然簇的形式存在，并且已知世界有一个代表点或基准模板用均值中心 M_i 表示。

$$M_i = \frac{1}{N_i}\sum_s X_{is} \tag{1}$$

依据最小距离规则计算一个未知样本 X 与各类均值中心之间的距离，这里采用欧氏距离度量。

$$d_i = \| X - M_i \| \quad , \qquad i=1,2, \cdots, c \tag{2}$$

式中 X 为测试样本的特征向量，M_i 为第 ω_i 类的均值的特征向量。

根据计算的距离值判定未知样本 X 的类别，确定 X 属于与之距离最近的代表点所属的类。

$$\omega_j = \arg\min_{\omega_i}(d_i) \quad (3)$$

其中 i=1,2，…，c； j∈i； ω_j 为类别的判别的结果。

未知世界的未知样本 X 成为已知世界的样本，即将测试样本 X 加入到训练样本集中，此时 X 所属的类别中的样本个数增加一个，更新类均值 M_i。

$$M_i = \frac{N_i * M_i + X}{N_i + 1} \quad (4)$$

依次对测试样本做上述的计算，将测试样本逐个划分到相应的类中。

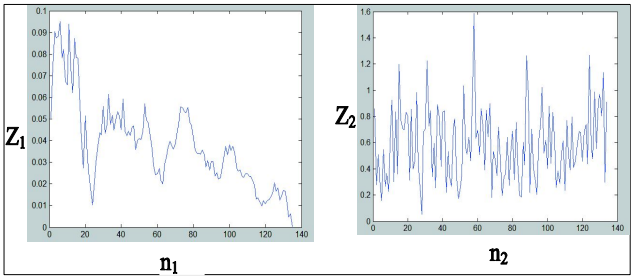
本文在 MU 分类模型基础上提出的 MU-MD 分类模型的实现是利用 MU 分类法的分类结果，再次计算类均值，对所有测试样本依据样本与类均值的最小距离重新进行划分，以确定最终的类别归属。

1.3 类代表点的变化特性分析

代表点选取的好坏直接影响分类的效果，运用本文提出的分类方法对球形分布、类的大小和密度均匀的 4k2_far.txt 数据集进行分类，分析分类过程中得到的类代表点的变化特性。

1.3.1 已知世界相对客观世界较小的情况

分析分类数据集中的一个类，以一个类为客观世界，计算其均值，以训练样本为已知世界，训练样本较少，仅取该类样本的 10%，计算其均值，然后计算两个均值的距离。再将测试样本一个一个地加入到已知世界中，每加入一个样本就更新已知世界的均值，并重新计算已知世界和客观世界的距离 Z_1，可以获得的距离个数 n_1 为 S_{te}+1 个。Z_1 的变化趋势如图 3（a）所示，随着已知世界的增大，与客观世界的相似性也增大，并且其分布接近于客观世界分布的机会大幅增加，Z_1 的振幅明显减弱。样本到已知世界均值的距离为 Z_2，可以获得的距离个数 n_2 为 S_{te} 个，其距离随已知世界增大的变化波形如图 3（b）所示。



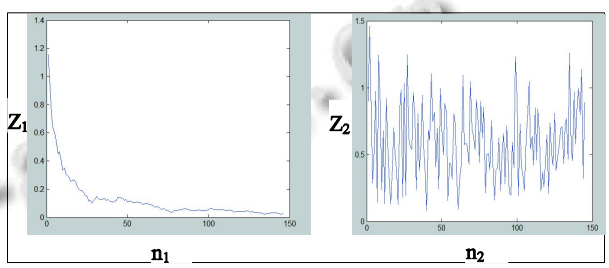
(a)均值更新的变化趋势图 (b)样本在均值附近的振荡波形图

图 3 训练样本占总样本 10%的类代表点的变化特性

如果 TRSS 到 CSS 的距离较大，以此为依据进行分类，那么错分在所难免。当一个簇加入错分的样本，就会给类均值造成影响，使得最终已知世界的均值偏离 CSS 的均值，误识率增加。因为本文设计的分类方法使用了先验知识，所以错分的几率不会很大，因此信噪比不会很低。当已知世界增大到一定程度时，波形的变化逐渐趋于平坦，几乎没有冲激成分，这说明处在类边缘的样本和噪声对均值的影响已经很微弱，不能引起大的波动，此时已知世界的均值已经很鲁棒。

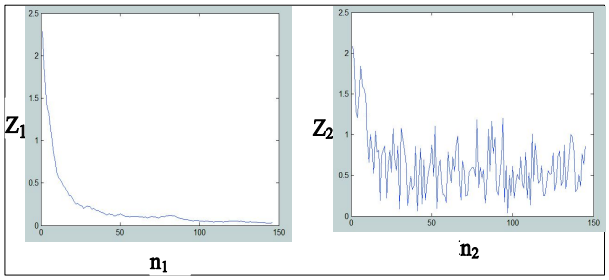
1.3.2 已知世界相对客观世界极小的情况

每类取距离全样本类均值最远的一小簇（4 个样本）作为训练样本。已知世界逐步增大后，它到客观世界的距离 Z_1 的变化趋势图如图 4（a）所示。波形在起始位置的振幅与图 3（a）相比较大，为 1.2，随着训练样本的增加，波形变化较快。波形图上振幅接近零的点，说明已知世界具有与 CSS 相同的分布，选择此时的样本组成训练样本集无疑对分类是有好处的。然而，我们除了控制训练集和测试集的比例以外不能干预抽样的结果，即便是某一个抽样与 CSS 的分布相同，我们也无从知道，除非通过最终分类的评价结果间接地获得这个信息。测试样本在其所属类的均值附近波动，Z_2 随已知世界增大的变化如图 4（b）所示。



(a)均值更新的变化趋势图 (b)样本在均值附近的振荡波形图

图 4 距离质心较远的 4 个训练样本的类代表点的变化特性



(a)均值更新的变化趋势图 (b)样本在均值附近的振荡波形图

图 5 噪声情况下类代表点的变化特性

1.3.3 在第 2 种情况下加入少量噪声

在类中加入噪声，噪声是其他类的一个样本点。由于噪声的存在，类均值发生很大变化，从图 5（a）可以看到波形在起始位置的振幅较大，为 2.5。随着训练样本的增加，波形变化较快，噪声影响逐渐减小。

测试样本在其所属类的均值附近波动，如图 5（b）所示。由于噪声的存在，测试样本与训练样本均值的距离在起始位置的振幅很大，约为 2.1，但波形的振幅逐渐变小，这表明本文提出的分类方法具有较好的克服噪声影响的能力。

2　算法比较与实验分析

为了验证本文提出的 MU 和 MU-MD 分类模型的性能，针对训练样本少的三种情况对 UCI 标准数据集中的几组数据集进行分类，并与 MD、NN 分类方法相比较。

2.1 全样本的 4%为训练样本情况

对六组数据集进行分类，每类取前 4%的样本作为训练样本集，其余的为测试样本集，对测试样本随机打乱其顺序，然后采用四种分类算法对其进行分类，最后计算出分类的正确率，其正确率如图 6 所示。图中纵坐标表示正确率，横坐标表示的六组数据集依次为 sonar、ionosphere、yeast、breast-cancer-wisconsin、leuk72_3k 和 iris 数据集。本文提出的分类模型在分类过程中，训练样本集随着分类的进行不断增大。对测试样本进行分类，其准确率通常是大训练样本集比小训练样本集高。从图 6 可以看出本文提出的算法的分类效果优于 MD 和 NN。

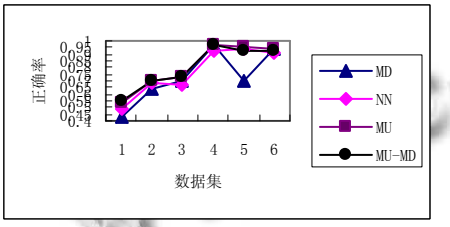


图 6　数据集的 4%为训练样本的分类正确率

2.2 训练样本更少的情况

每类取前两个样本作为训练样本集，其余为测试样本，对测试样本随机打乱其顺序后进行分类，计算正确率如图 7 所示。图中横坐标表示的六组数据集及数据集的次序与图 6 相同。从图中可以看到本文提出的两种算法在六组数据集上的分类效果除了 sonar 数据集外均好于最小距离分类法和最近邻分类法.

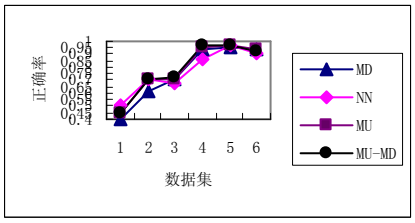


图 7　2 个样本为训练样本的分类正确率

2.3 训练样本距离质心较远且数量很少的情况

每类取距离较近且其中一个距离质心最远的四个样本作为训练样本集，其余的为测试样本，随机打乱测试样本的顺序，对其进行四种分类，分类的正确率如图 8 所示。图中横坐标表示的五组数据集依次为 balance-scale、pima-indians-diabetes、iris、breast-cancer-wisconsin 和 leuk72_3k 数据集.

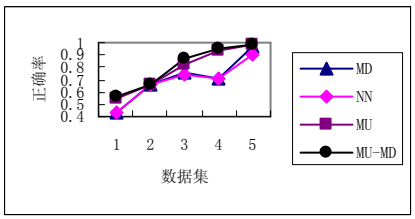


图 8　4 个训练样本且距离质心较远的分类正确率

在训练样本集固定，测试样本随机打乱顺序的情况下，重复进行多次分类，MD 和 NN 分类法分类的正确率始终保持不变，MU 和 MU-MD 分类模型的正确率会随测试样本的顺序变化有一定波动，本文提出的分类模型的正确率优于其他两种方法，优于的程度略有波动。由于训练样本偏离样本集的质心，MD、NN 误分类的概率较高，而本文提出的 MU 随着分类的进行，均值中心不断更新，均值中心对类别的代表性有逐渐变好的趋势，从而正确率要高于 MD 和 NN 分类法。MU-MD 分类模型使用的均值中心是 MU 分类后计算得到的结果，其值对类别的代表性远远优于由最初的训练样本计算得出的类均值，也比 MU 分类过程中计算得出的绝大多数均值中心的代表性好，所以其分类效果通常会比 MU 分类模型的好。从图 8 可以看到本文提出的两种分类模型的分类正确率远远高于 MD 和 NN 分类方法，尤其在 breast-cancer-wisconsin 数据集上的分类效果更为突出，并且 MU-MD 分类模型的分类效果优于

(下转第 135 页)

30%；IT 人员花在备份上的时间减少 50%。系统备份在不同数据量下的备份时间比较如图 6 所示。

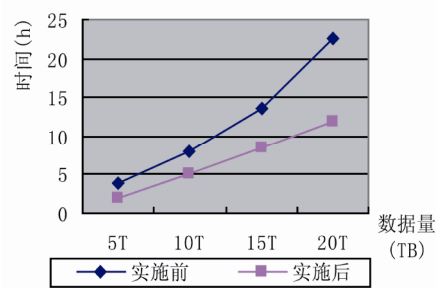


图 6 系统实施前后备份时间对比

2) 自动化、集中化的备份方式满足了系统管理员对系统文件、数据库数据的日常备份与恢复需求，且无须考虑备份介质的使用率和可用性问题，随时方便了解备份进度及状态，减少了系统管理员负担。

3) 数据备份和恢复时间大大降低，同时提高了备份的安全性和可靠性，从而大大增加了企业信息系统的安全性和抗灾害能力，也为后续的容灾系统提供技术基础。

4 结论

本文针对传统备份系统无法满足企业大规模、高效率和灵活性要求的现状，将云存储思想引入备份平台，提出了基于异步数据云的备份平台架构，以网络交换层为核心，将逻辑数据云和物理数据云接入网络，实现稳定、高效、安全的海量数据备份，同时针对不同备份对象制定灵活的备份策略，提高了备份和恢复速度，通过在浙江中烟的实施，使其信息化管理工作效率显著提高，取得了明显效益，也为企业建设容灾、减灾系统提供了基础。

参考文献

1 任欣,李涛,胡晓勤.远程文件备份与恢复系统的设计与实现.计算机工程,2009,35(10):112–114.
2 韩塞北.网络存储主流技术比较及分析.长春师范学院学报(自然科学版),2010,29(6):38–41.
3 田萍芳,鲁宏伟,秦磊华.网络存储技术.计算机与数字工程,2004,32(2):38–41.
4 刘金芝,余丹,朱率率.一种新的云存储服务模型研究.计算机应用研究,2011,28(5):1869–1872.
5 高建秀,吴振新,孙硕.云存储在数字资源长期保存中的应用探讨.现代图书情报技术,2010,(6):1–6.
6 边根庆,高松,邵必林.面向分散式存储的云存储安全架构。西安交通大学学报,2011,45(4):41–45.
7 李翠侠.基于混合云的高校图书馆存储方案研究.图书馆学研究(理论版),2011(3):68–71.
8 刘国萍,谭国权,杨明川.基于云存储的在线备份安全技术研究.电信科学,2010(9):79–83.

(上接第 126 页)

MU 分类模型。

3 结语

本文提出 MU 分类模型，通过不断更新均值中心可以有效的对测试数据进行准确的分类，该算法不受样本分布概率模型限制，不用控制参数，使用简单，在此基础上又提出了 MU-MD 分类模型。MU 分类模型和 MU-MD 分类模型在训练样本较少或训练样本的均值离质心较远的情况，与 MD 和 NN 分类法相比较，其准确率较高。最后，通过实验比较，验证了本文提出的分类模型的有效性,以及在训练样本较少、训练样本的均值偏离质心较远及噪声情况下分类的优越性。

参考文献

1 Toth D, Aach T. Improved minimum distance classification with Gaussian outlier detection for industrial inspection. Italy: Proc. of the 11th International Conference on Image Analysis and Processing. 2001.584-588.
2 Tan SB. An effective refinement strategy for KNN text classifier. Expert Systems with Applications, 2006,30(2): 290–298.
3 张孝飞,黄河燕.一种采用聚类技术改进的 KNN 文本分类方法.模式识别与人工智能,2009,(12).
4 任靖,李春平.最小距离分类器的改进算法——加权最小距离分类器.计算机应用,2005,25(5):992–994.
5 干正如,曾宪珪.基于自适应距离原理的自适应分类方法.江西理工大学学报,2007,(8).
6 Aeberhard S, Coomans D, Devel O. Comparative analysis of statistical pattern recognition methods in high dimensional setting. Pattern Recognition, 1994,27(8):1065–1077.