计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.191100094

# 训练样本数据选择方法研究综述

周 玉 任钦差 牛会宾
华北水利水电大学电力学院 郑州 450011

**摘 要** 机器学习作为数据挖掘中一种重要的工具,不只是对人的认知学习过程的探索,还包括对数据的分析处理。面对大量数据的挑战,目前一部分学者专注于机器学习算法的改进和开拓,另一部分研究人员则致力于样本数据的选择和数据集的缩减,这两方面的研究工作是并行的。训练样本数据选择是机器学习的一个研究热点,通过对样本数据的有效选择,提取更具有信息量的样本,剔除冗余样本和噪声数据,从而提高训练样本质量,进而获得更好的学习性能。文中就目前存在的样本数据选择方法进行综述研究,从基于抽样的方法、基于聚类的方法、基于近邻分类规则的方法这三大类以及其他相关数据选择方法4个方面对目前存在的方法进行总结和分析对比,并对训练样本数据选择方法存在的问题和未来研究方向提出一些总结和展望。

**关键词**:训练样本;数据选择;机器学习;神经网络;支持向量机
**中图法分类号** TP181

## Research on Training Sample Data Selection Methods

ZHOU Yu,REN Qin-chai and NIU Hui-bin
School of Electric Power,North China University of Water Resources and Electric Power,Zhengzhou 450011,China

**Abstract** Machine learning,as an important tool in data mining,not only explores the cognitive learning process of human beings,but also includes the analysis and processing of data. Faced with the challenge of massive data,at present,some researches focus on the improvement and development of machine learning algorithm,while others focus on the selection of sample data and the reduction of data set. The two aspects of researches work in parallel. The selection of training sample data is a research hotspot of machine learning. By effectively selecting sample data,extracting more informative samples,eliminating redundant samples and noise data,thus improving the quality of training samples and obtaining better learning performance. In this paper,the existing methods of sample data selection are reviewed,and the existing methods are carried out in four aspects:sampling-based method,cluster-based method,nearest neighbor classification rule-based method and other related data selection methods. Summarize and analyze the comparison,and put forward some conclusions and prospects for the problems existing in the training sample data selection method and future research directions.

**Keywords** Training sample,Data selection,Machine learning,Neural networks,Support vector machines

## 1 引言

随着数字技术与网络的广泛普及,大量的数据被收集,数据大爆炸时代已经到来。Szalay 和 Gray 将这种现象描述为“淹没在数据中”[1]。数据爆炸性增长引出的一个非常自然的问题是“收集了大量数据后,我们如何处理它?”原始数据很少直接使用,人工分析也根本无法应对数据的快速增长。数据挖掘和知识发现(Knolwedge Discovery and Data Mining,KDD)作为一个新兴领域出现,用以解决这一问题,其包括数据库、统计学和机器学习等学科。KDD 试图挖掘出原始数据隐藏的信息为科学发现和商业智能化提供帮助。

KDD 过程包括数据选择、预处理、数据挖掘、解释和评估[2]。机器学习是数据挖掘中一种重要的工具,不只是对人类认知学习过程的探索,还包括对数据的分析处理。做好数据选择,是机器学习成功的关键之一。面对大量数据的挑战,目前一部分学者专注于机器学习算法的改进和开拓,另一部分研究人员则致力于样本数据的选择,这两方面研究工作是并行的。通过对样本数据的有效选择,提取信息量大的样本,剔除冗余样本和噪声数据,从而提高训练样本质量,进而获得更好的学习性能。

许多因素导致必须进行数据选择。首先,数据收集的目的不纯粹是用于数据挖掘或某个特定算法;其次,在收集和记录期间存在缺失数据、冗余数据和错误数据;最后,有时数据可能过于庞大而无法处理。针对数据选择,一种情况是通过使用不同的特征选择方法[3-7]来解决问题,其主要任务是删除对机器学习意义不大的数据集的列,进行特征向量的缩减;而

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:河南省高等学校青年骨干教师培养计划(2018GGJS079);国家自然科学基金(U1504622,31671580)
This work was supported by the Project of Training Young Backbone Teachers in Colleges and Universities of Henan Province (2018GGJS079) and National Natural Science Foundation of China (U1504622,31671580).
通信作者:周玉(zhouyu_beijing@126.com)

另一种情况则是通过减少数据集中无关紧要的行来进行数据选择，即对样本数据的个数进行缩减。在查阅文献过程中发现，第二种情况的研究已经引起越来越多学者的关注并逐渐成为一种新趋势。这种趋势的存在有很多原因：首先，选择行涉及数据缩减的某些方面特征选择无法覆盖，比如特征选择只是选择属性，并没有去除错误的数据，错误数据参加训练将直接影响训练所得模型的准确性；其次，现在可以通过先进的统计知识和积累的经验来进行数据行的选择，使得训练样本的选择更加高效；最后，这样做在机器学习训练中有许多优势，比如选择后的数据更具代表性，进行训练时所需时间更短。本文主要对第二种训练样本数据选择方法进行分析总结。

训练样本数据选择就是从原始训练样本集 $T$ 中尽量多地删除冗余样本和噪声样本，然后得到一个压缩集 $T'$，使选择后得到的样本集 $T'$ 能够保持或者改善分类器的分类性能。具体定义如下：设初始训练集为 $\{X,P\}=\{(x_1,p_1),(x_2,p_2),\cdots,(x_n,p_n)\}$，表示有 $n$ 个训练样本，其中 $\{X_n\}_{i=1}^n$ 为样本的特征属性，$\{P_n\}_{i=1}^n$ 为样本的类别标签。存在一个最小的压缩集 $\{X',P'\}\subset\{X,P\}$，使得任意 $\{x,p\}\subset\{X,P\}$，$P=Classifier(x,\{X',P\})$ 表示测试样本 $x$ 使用分类器在压缩集 $\{X',P'\}$ 上的分类结果。

样本数据选择有以下突出功能。1）数据选择使不可能成为可能。众所周知，分类算法在某种程度上都受到其处理数据大小、类型、格式等能力的限制。当数据集太大且在没有减少数据的情况下，可能无法运行分类算法或运行缓慢，或者无法有效地进行训练。2）数据几乎包括域中的所有内容，但一个应用程序通常只涉及域的一个方面，我们只需关注需要的方面即可，这样搜索就会更加集中，效率更高。3）“垃圾入，垃圾出”（garbage in，garbage out）原理几乎适用于所有分类算法，因此在训练之前清理数据是至关重要的。训练前进行数据选择的意义在于通过选择重要数据，删除不相关的数据（包括噪声和冗余数据）来获得高质量的数据，以此达到提高分类结果准确率，改善分类性能并降低分类器训练成本的目的。

在数据选择时，经常要对数据规模和应用效果进行权衡。从这个意义上讲，数据选择是一种优化问题，它试图在尽量最小化样本量的同时保持准确率等指标。数据选择的目的是通过选择获得最具有代表信息的较少样本。本文旨在对机器学习（支持向量机、神经网络、K 近邻算法、决策树等）相关领域的训练样本数据选择方法进行综述研究和分析对比，并对数据选择方法的现状和未来的研究趋势进行探讨。

## 2　样本数据选择方法

目前存在的训练样本数据选择方法多种多样，根据选择中用到的数据选择策略和过程的不同，其主要分为以下 3 种方法：基于抽样的方法、基于聚类的方法以及基于近邻分类规则的方法。除此之外，还有其他相关数据选择方法。

### 2.1　基于抽样的方法

抽样方法是一种简单实用的方法[8]，其在所有系统中都可以使用。抽样方法是对分类器进行训练时缩减训练集的最基本策略，在训练支持向量机中很多学者应用了抽样的方法。为了减小训练集，Balcazar 等提出了一种随机抽样算法[9]，即从整个训练集中随机生成一个子集作为训练样本。该方法简单易行、速度快，得到了广泛使用。其虽然能有效减少样本数量，但不能保证分类器的泛化能力。在此基础上，一些学者还提出了其他一些基于抽样的方法，包括 Ferragut 的 SSVM[10] 和 Lee 的 RSVM[11] 等。文献[10]利用现有的 SVM 训练算法，建立了一种新的 SVM 训练算法，即使用随机抽样数据来训练多个子 SVM，然后利用子 SVM 对训练样本进行过滤、整合，再进行训练。该方法更好地将 SVM 扩展到了大型数据集的应用中。实验表明，与原有的支持向量机算法以及级联支持向量机相比，该方法具有更快的速度和更高的精度。文献[11]均匀、随机选择一个压缩集来进行训练，这种简化的核技术已经成功地应用于其他基于内核的学习算法中。这些方法已被证实了比选择前支持向量机的训练速度更快，而且从文献[12-13]的工作中发现，均匀随机抽样是最优鲁棒性方案。然而学者们发现，尽管这些方法操作简单且成本较低，但在大多数情况下通过抽样方法选择数据进行训练得到的算法分类精度的标准偏差很大。因为随机选择样本得到高质量样本的可能性较低，其得到的样本是随机不确定的。结果的不确定性导致学者逐渐摒弃这一方法，开始尝试别的方法或把抽样方法与其他方法融合使用。随着研究的发展，抽样方法多与聚类等方法结合使用，降低了数据选择的偶然性。

### 2.2　与聚类相关的方法

随机选择训练样本进行训练不能保证分类器在实际识别任务中可靠地工作，因为不同的训练集可能会产生不同的分类错误。因此，越来越多的学者提出通过考虑数据的分布特征来有效地选择训练集，其中通过聚类方法来进行数据选择获得了越来越多学者的青睐。

通过聚类方法可以快速从大量数据中分块地进行数据选择。Zhang 等[14]通过对训练集中的各类样本分别进行多次聚类，并对聚类结果进行融合获得最终聚类，然后通过判断聚类内样本的离散度和覆盖区域选取一定数量的样本。由于同一聚类中的样本具有较高的相似性，因此进行样本选择时遵循的基本原则就是放弃大部分的内部样本，保留边界样本。该方法保留每个聚类的边界样本，然后删除内部样本，确保了学习的性能，提高了训练效率。但如果训练样本中包含噪声数据，该方法的结果将受到严重影响，需要进一步与去噪结合来改善算法性能。Almeida 等[15]提出了一种基于 K 均值聚类的支持向量机训练方法 SVM-KM，其先利用 K 均值进行聚类[16]，然后将只由属于同一类标签的样本形成的集群视为冗余样本而忽略，只使用其集群中心。相反，具有多个类标签的集群被保留并添加到重构子集中，其主要思想是从类边缘聚类矢量，并在其附近保留矢量。该方法使支持向量机训练集中的向量数目更少，训练时间也更短；在不降低支持向量机的泛化能力的同时，缩短了训练时间。但在使用 K-means 算法的同时，输出集群的个数选择是一个困难的问题。此外，K-means 的不同初始点将导致完全不同的分区。因此，基于此聚类技术的结果是不稳定的。Guan 等[17]利用模糊聚类来提高监督学习的性能，提出了 3 种基于模糊聚类方法的数据选择机制：基于中心的选择，基于边界的选择和混合选择。基于中心的选择是将每个群集中具有高隶属度的样本作为训练数据；基于边界的选择是提取群集之间的样本，即边界数据作

为训练样本；混合选择是基于中心的选择和基于边界的选择的组合。该方法使用 FCM 来实现数据选择机制，并通过 BP 神经网络进行验证。实验证明，与随机选择相比，混合选择可以有效地提高数据集的学习性能。我们可以看出，基于聚类的方法多是通过聚类算法寻找边界样本，去除对结果影响不大的冗余样本来进行数据选择。Zhou 等[18]通过阴影集进行边界数据的选择，首先通过 FCM 聚类得到样本数据的最优模糊划分矩阵，基于此得到相应的阴影集[19]，最后通过阴影集构造核数据和边界数据并进行数据选择。利用该方法分别对 BP 网络、LVQ 网络和 ENN(Extension Neural Network)等分类器进行实验，实验结果验证了该方法能够在保证泛化能力的同时减少样本数量和训练所需时间。与上面数据选择方法需要在整个数据集上进行聚类不同，文献[20]给出了一种基于聚类的多类实例选择(MCIS)算法。MCIS 利用聚类技术，采用了远离正类聚类中心的正实例和靠近这些中心的负实例在边界附近的假设。这里聚类的目的是提高实例选择的效率，而不是像以前的方法那样直接从集群中选择实例。选择一个给定的 C 类作为正类，MCIS 首先对 C 类进行聚类，得到一定数量的聚类中心；然后以这些中心为参考点，删除靠近这些中心的正实例，并在中心附近选择距离最近的负实例来获取边界数据进行支持向量机的训练。MCIS 算法的具体过程如算法 1 所示。

**算法 1** MCIS 算法

输入：训练集 T，包括 L 类、N 个样本，选择样本比例 r；

输出：$S_c$，当 C 类被视为正类、其余类被视为负类时所选的样本集($1\leqslant c\leqslant L$)；

1. for each class C, $1\leqslant c\leqslant L$；
2. $S_c\leftarrow\{x|x$ is from class c$\}$；
3. 在 C 类上执行 K 均值聚类，获得聚类中心 $M_1, M_2, \cdots, M_k$；
4. for each class $M_i$，计算 $M_i$ 与每个样本 x 之间的距离 $d(M_i, x)$；
5. if $N_c\geqslant r\cdot N/3+k$；
6. 在 $S_c$ 中获得最接近 $M_i$ 的 $\overline{n_c}$ 个样本，并把它们从 $S_c$ 中删除；这里 $\overline{n_c}=(N_c-r\cdot N/3)/K$；
7. end if
8. for each class $l\neq c, 1\leqslant l\leqslant L$；
9. 使用最小距离度量找到 l 类中的 $n_l$ 个实例，并把它们加入 $S_c$ 中，这里有：

$$n_l=(r\cdot N-n_c)/(K\cdot(L-1))$$

$$n_c=\begin{cases}r\cdot N/3, & N_c\geqslant r\cdot N/3+K\\ N_C, & \text{其他}\end{cases}$$

10. end for
11. end for
12. end for

文献[21]进一步探讨了在一个集群中存在聚类和分散的数据点，认为聚类中心点周围的数据点是位于聚类内层的密集数据点，这些数据点不包涵 sv(支持向量)被删除；而分散的数据点是集群外层的稀疏数据点，这些数据点具有 sv，因此被保留。同时，提出利用 Fisher 判别比确定密集数据点与分散数据点之间的边界；通过去除集群中的冗余密集数据点，加快支持向量机的训练过程。

聚类技术在主动学习中同样有很多应用，特别是在文本分类中结合支持向量机的应用[22-23]。通过研究我们发现，对于支持向量机[24]，进行数据选择主要是寻找边界数据即支持向量进行训练，以达到减少训练数据和缩短训练时间的目的；对于神经网络分类器，多是通过获得边界数据，结合中心数据或随机数据进行训练，达到减少训练数据和缩短训练时间的目的。但大多聚类方法的前提条件都是类数据不重合的情况下进行的，对有重叠的数据集的数据选择方法并不多，还有待研究，而且通过聚类选择的子集中可能包含噪声样本，因此去噪算法与聚类技术的结合也是研究的一个方向：先通过去噪算法预处理样本，再通过聚类进行边界样本的选择，以达到提高算法性能的目的。

### 2.3 针对近邻分类规则的数据选择方法

近邻分类(Nearest Neighbor Classification)一直是数据挖掘领域学者们研究的重要课题之一。该方法是一种惰性学习(Lazy Learning)方法，也是一种非参数分类方法，主要通过计算测试样本和训练样本间的距离进行分类，在文本分类[25]、字符识别[26]、时间序列[27]等诸多领域应用广泛。

第一个针对近邻规则的样本选择算法是由 Hart 在 1968 年提出的压缩最近邻规则 (Condensed Nearest Neighbor, CNN)[28]。CNN 算法从训练集中随机抽取一个实例组成新的数据集；然后将训练集中不能被新数据集正确分类的数据添加到该集合中，直到全部能正确分类为止。该算法选择靠近类边界的实例，缩小了训练集的规模。CNN 算法的具体过程如算法 2 所示。

**算法 2** CNN 算法

输入：训练样本 T；

输出：压缩集 D；

结束条件：$T=\varnothing$ 或 T 中所有样本都能被 D 用 K-NN 正确分类；

初始化：从 T 中随机选择一个样本加入 D 中；

当不满足结束条件时：

{ 对 T 中每一个样本 $x\in T$，用压缩集 D 对 K-NN 进行分类；

若 x 能够正确分类，则 $T=T-x$；

若 x 不能正确分类，则 $D=D\cup x$；

}

输出 D

然而，CNN 所找到的子集并不是最小的，而且对样例出现的先后顺序及噪音点很敏感，它所选择的样本离分类边界较远，噪音也会被近邻样本误分类。虽然 CNN 算法性能不佳，但该模型启发了新方法的构建，一系列基于近邻分类规则的数据选择方法不断涌现。文献[29]提出了约简近邻法(Reduced Nearest Neighbor, RNN)，与 CNN 不同的是其是从整个训练集开始通过去除那些不降低分类识别率的样本来获得压缩集，而不是从新的样本集开始。文献[30]提出了选择最近邻(Selective Nearest Neighbor, SNN)算法，所选压缩集中的所有样本必须比其他类中的任何样本更接近同一类的选择性邻居。Dasarathy 提出了 MCS (Minimal Consistent Set)算法[31]，该算法是在最近异类样本(Nearest Unlike Neighbor, NUN)和最近异类样本子集(Nearest Unlike Neighbor Set, NUNS)的理论基础上，通过在同类邻域内使用投票的方法，每个样本都给与自己异类距离最近的同类样本投票，得票数越多的样本与周围样本的差异越大，即能更好地区分周围样本；再将得票数最多的样本添加到最小一致集中进行下一轮投票操作，终止条件是最小一致集中的样本不再变化。MCS 算法对样本较少的数据集作用比较好，但所挑选的压缩集并不是最小压缩集，不能完全保证逼近分类边界。文献[32]提出了 FCNN(Fast Condensed Nearest Neighbor)算法，其思想

是把每类的类心作为初始压缩集，然后把每类中与当前压缩集同类距离最近且不能正确分类的样本加入压缩集，直到全部正确分类为止。此方法在大规模数据集上也取得了良好的效果。FCNN 算法的运用与选取样本的前后顺序无关，其时间复杂度相对较低，与 MCS 相比更适用于样本数较多的样本；但通过 FCNN 选取的训练集依旧很大，且无法避免噪声样本的选取。文献[33]提出了一种基于邻域属性的模式选择(Neighborhood Property based Pattern Selection，NPPS)算法，该算法通过 3 个邻域属性来进行数据的选择。第一个属性用于识别位于决策边界附近的样本；第二个属性用于删除位于决策边界错误一侧的样本；第三个属性用于跳过不必要的距离计算，从而加快了样本选择过程。该方法降低了计算复杂度，考虑到了噪声点的影响。文献[34]基于 CNN 算法提出了一种基于局部均值和类全局信息改进的最近邻原型选择(LCNN)算法，其利用近邻局部均值和类全局信息设定不同的阈值来进行数据选择，对 CNN 对样本顺序依赖的缺点进行了改进。以上算法均是学者们在 CNN 的基础上进行研究改进的，也存在噪声数据集的问题，且 RNN，SNN 和 MCS 等的计算成本比 CNN 算法还高。

CNN 相关算法主要是通过考虑关键样本即决策边界的样本进行数据选择，抗噪能力较弱；除了这些考虑关键样本的方法外，基于近邻规则通过去除噪声样本的数据选择方法同样值得注意。Wilson 在 1972 年提出剪辑近邻法(Edited Nearest Neighbor，ENN)[35]，该算法是通过 K 近邻分类规则对训练样本进行分类，把错误分类的样本删除，以达到去除噪声样本、清理分类边界的作用。ENN 算法的具体过程如算法 3 所示。

**算法 3**　ENN 算法

输入：训练样本 T；

输出：压缩集 D；

初始化：D＝T；

对于每一个 x∈D，进行以下操作：

　　计算 x 的 K 个近邻样本，找出 K 个近邻样本的类别中每种类别对应样本的个数；

　　判断 x 是否与找出的最多数量类别样本为同类；

　　若不属于同类，将 x 从 D 中删除，反之保留；

结束，输出压缩集 D

基于 ENN，文献[36]提出了 RENN(Reduced ENN)算法，重复使用 ENN 规则，直到压缩集不再变化；进一步地，提出了剪辑所有近邻法(ALLKNN Editing)，通过对所有 $K$ 的取值都使用 ENN 进行剪辑来获得更小压缩比的数据集。文献[37]提出了一种新的剪辑近邻规则，在提出的规则中，对于任意的 $k$，在通过剪辑后的参考集中，每个样本 $y$ 必须被它的 $k$ 个最近的邻居包围，且这些邻居都属于 $y$ 所属的类。也有一些学者通过互近邻的原理来去除噪声数据，进行数据选择[38]。以上与 ENN 相关的算法都是删除噪声但保留了训练集内部的样本，计算复杂度并没有过多减少，冗余样本也都大量保留，因此其针对小型数据效果较好但不适用于中大型数据集的数据选择。

无论是与 CNN 相关，亦或是与 ENN 相关的样本选择算法，都没有全面考虑样本的特性。文献[39]提出基于最近异类的 NES 样本选择算法，针对噪声样本会导致训练过程复杂且训练结果正确率不高的问题，NES 首先对噪声数据进行了剔除。去噪过程基于近邻规则，首先对全部样本点 $x$ 寻找 $k$ 最近邻 kNN($x$)，计算 $x$ 的 $k$ 近邻中类标签与其一致的样本所占比例，若比例低于设定阈值，则删除该样本点，否则保留。通过设置参数可以调整去噪度，参数大都依据样本集设置。然后，NES 根据各个样本点的最近异类信息，将训练集划分为相互独立的多个子集，并在每个子集中进行样本选择。随后，作者又在其基础上进行了改进，提出基于 $\theta$-Net 的 NENet 样本选择算法。该方法通过有效的数据选择提高了数据的压缩率和算法的准确率，但是在参数 $k$ 和 $\theta$ 的设置方面还存在一些问题，而且在使样本选择算法能够启发式地调整不同子集中的样本选择参数方面还有待研究。

### 2.4　其他方法

除了上述 3 类基于统一基础思想的方法外，在减少数据方面，近年来涌现出了越来越多的数据选择方法，包括与决策树(Decision Tree，DT)、几何、进化算法、距离等相关的方法。

文献[40]结合了决策树的方法，使用决策树来形成被视为聚类的区域。与特定的聚类方法(如 K-mean 或 FCM)不同，这些聚类方法通过使用距离或密度度量对样本进行分组，从而显式地形成聚类；DT 则使用类似于基尼系数或熵的纯度作为分类依据，其优点是不需要提前指定聚类集群的数量，而是把叶子当作聚类的簇，然后通过有指导的随机算法抽取各簇中的向量组成新的训练集，通过使用决策树可以一定程度地解决聚类算法收敛缓慢的问题。文献[41]提出了一种基于决策树的大数据集 SVM 分类的数据选择方法，通过从支持向量机第一阶段得到的 SV 和非 SV 建立决策树，进而应用决策树进行数据过滤，使用决策树来识别与计算的 SV 具有相似特征的对象，然后从原始数据集中删除不太重要的对象。实验证明了该算法在实际数据集的训练时间上具有优越性。

与几何学相关的数据选择方法，一种是基于计算最优分离超平面即等价于寻找凸壳上最接近的一对点[42-44]，另一种是基于类质心的快速训练集缩减算法，根据样本的几何分布去除大部分的非支持向量[45-46]。此外，文献[47-48]通过计算样本向量之间夹角的余弦值来进行边界样本的挑选，文献[48]采用余弦相似度当作样本数据之间远近程度的衡量标准，余弦夹角更多体现的是方向上的不同，其对向量空间坐标位置不敏感，因此可通过向量单位化降低计算复杂性，提高样本选择效率。文献[49]提出了一种基于邻域样本密度的 SVDD 样本剪辑方法，选取一个固定大小的邻域，并计算邻域里面的样本数，根据确定的阈值选择丢弃还是保留邻域样本，所保留的样本接近边缘。文献[50]采取了类似的方法，通过确定一个固定邻域球来进行边界数据的选择。与文献[49-50]不同，文献[51]提出一种壳状数据选择方法，其以样本数据中心向量为中心，以 $R$ 为半径、寻找该圆形区域内的冗余数据点并进行筛选。将类中的样本向量点 $x$ 到中心向量的距离与筛选半径 $R$ 进行比较，若其比 $R$ 小，则剔除该向量点，重新计算筛选后的中心向量对样本点进行筛选，直到筛选不到点或该类别中剩余的点达到截断阈值 $T$ 为止。随着每轮迭代后中心向量位置的变化，中心向量周围的点不断被删除，最终留下的点大部分是分布在该类边缘的向量点，其形状在向

量空间中类似于一个“壳”。相比于同类的其他几何型数据选择方法，该方法几乎完整地保留了样本的边界数据，即壳边缘，具有更高的准确率。实验表明，相较于基于近邻的样本选择算法和基于聚类的样本选择算法，壳状数据选择算法具有更高的运行效率，可以更好地适用于大规模训练集；但该类方法不适用于环形数据集或分布过于分散的特殊数据集。

进化算法进行实例选择应用最多的是遗传算法。文献[52]提出用遗传算法(Genetic Algorithm，GA)来为分类器找到较好的数据原型。文献[53]又提出了利用遗传算法和进化策略(Evolution Strategy，ES)相结合的算法 IFS-IBGAES 来解决实例选择中的选择问题和特征权重问题。在 IFS-IBGAES 算法中，GA 的目的是从原始数据集中选择原型实例，而 ES 的目的是利用该技术对特征进行加权。文献[55]在 AGA(Adaptive Genetic Algorithm)[54]的基础上提出了一种新的动态自适应遗传算法 DAGA(Dynamically Adaptive Genetic Algorithm)来选择有价值的训练集，并证明了 DAGA 不仅可以快速选择训练数据，而且可以在不需要任何先验信息的情况下动态确定所需训练集的大小。文献[56]又提出了交替遗传算法 ALGA(Alternating Genetic Algorithm)来优化支持向量机模型与支持向量机训练集。遗传算法主要是通过适应度函数对选择的样本进行评价，通过染色体的交叉变异来进行选择。其思路是维持满足目标的种群数，并不断地迭代进化。一般目标函数与分类准确率和样本压缩率有关，但是基于遗传算法的样本选择计算复杂度大都比较高。除了遗传算法外，蚁群算法也在实例选择中得到应用，文献[57]提出了一种基于蚁群算法原理的实例约简方法 ACO-IR，该方法的主要思想是，每只蚂蚁从原训练集中构造一个候选的约简集，蚂蚁完成旅行后，通过对原训练集中的所有实例进行分类，并对预测精度进行检验，计算出约简集的适应度。

目前大多数基于距离的方法都是基于两个观察结果：1)标签不同且最接近的样本最有助于决策边界的形成[58-60]；2)远离样本边界的实例无助于决策边界的定义。

## 3 分析讨论

高质量地对训练样本进行数据选择在机器学习中扮演的角色越来越重要，通过对样本数据进行选择可以删除冗余样本、噪声数据和错误样本，有效减少训练样本的数量，提高训练样本的质量，进而提高机器学习的性能。本文系统研究分析了样本数据选择的各种方法和策略，从基于抽样的数据选择方法、与聚类相关的数据选择方法、针对近邻分类规则的数据选择方法和其他方法这 4 个方面对数据选择方法进行了总结概括。通过研究得出数据选择的主要目的在于：1)去除噪声；2)压缩数据即去除冗余数据；3)还原样本原型。基于抽样的方法虽然可以起到压缩数据的目的，但是由于其对数据进行选择时具有随机性，目前较少单独使用，多与其他方法结合使用。基于聚类的方法通过分析样本的整体分布结构，对边界样本进行挑选，选择的样本能够很好地还原样本的原型，因为分类超曲面主要由边界数据决定，但是运用聚类算法的前提是类之间不存在大量的数据重合。针对近邻规则的方法主要针对近邻分类规则来进行数据选择，研究主要从去除冗余数据和去除边界噪声数据两大方向进行。其他方法的研究目的也是在于寻找对分类起决定作用的样本。进行数据选择后的数据有助于提升分类器的性能。

## 4 存在的问题及研究展望

本文针对机器学习涉及到的样本数据选择方法进行了总结概括，阐述了各类样本数据选择方法的过程与达到的效果。通过对数据选择方法的研究进行回顾阐述，发现其存在的问题有以下几点。

1)基于抽样的样本选择方法虽然操作简单，但因抽样随机性比较大，结果偏差较大。

2)用聚类方法进行数据选择大多是在不同类数据集不存在重合的情况下进行的，且要分成簇的个数不易确定；对有重叠的数据集的数据选择方法并不多，还有待研究，而且噪声样本的存在对基于聚类算法的数据选择方法的影响较大。

3)部分数据选择方法是针对特定分类器的，其对别的分类器的适用性还有待研究。

4)不同训练样本数据选择方法根据应用场景的不同，其好坏的判定标准并不统一。

5)数据选择方法的参数设置问题一直是一个研究热点，如何更好地确定算法的参数还有待持续研究。

从目前数据选择方法研究中存在的问题来看，未来可从下面几方面进行进一步的研究。

1)抽样方法针对样本量特别大的训练集能节省大量时间，因此样本量特别大时可以把抽样方法与聚类等方法结合使用，降低抽样的随机性。

2)大部分的样本选择方法都证明了边界数据在分类器训练中起到了关键作用，因此如何通过各种方法更加清晰地确定边界数据依然是研究的重点，且边界样本与噪声样本的区分至关重要。

3)面对部分数据选择方法只针对特定分类器，或针对特定分布的样本集，如何进行方法融合来选择数据值得学者们研究。

4)部分数据选择算法只剔除冗余样本，部分数据选择方法只剔除噪声数据，如何在控制时间复杂度以及确保精度的前提下把两种方法结合使用，针对不同训练样本集确定每种方法剔除样本的比例是学者研究的重点。

5)大部分数据选择方法都涉及参数设置这一问题，如何结合经验以及数据本身的结构特征等设置对实验结果最佳的参数依旧是很多学者研究的重点。

目前，数据选择已成为进行分类器训练前一项重要的预处理步骤，做好数据选择既可以提高分类性能，又可以节省训练时间，其正在越来越多的工程实际应用中得到使用。

**结束语** 在日常生活和科研领域的多个方面，分类问题的应用越来越多，因此分类器的研究成为了机器学习中的一个研究热点。分类器模型的构建离不开训练样本的支持，因此对训练样本进行选择成为除了对分类器模型进行改进外的另一个研究趋势。本文系统地介绍了目前数据选择方法的研究现状，通过分析总结对存在的问题和未来的研究方向进行了展望。可以预见，在当下这种数据爆炸的时代，对数据选择方法进行研究充满着机遇与挑战，有很大的应用空间。

## 参 考 文 献


[1] SZALAY A，GRAY J. Drowning in data [OL]. https://www.sciam.com/explorations /1999/.

[2] FAYYAD U M，PIATETSKY-SHAPIRO G，SMYTH P. From data mining to knowledge discovery：an overview [M] // Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence，1996.

[3] BLUM A L，LANGLEY P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence，1997，97(1/2)：245-271.

[4] BARBU A，SHE Y，DING L，et al. Feature Selection with Annealing for Computer Vision and Big Data Learning[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，39(2)：272-286.

[5] LIU Y，BI J W，FAN Z P. Multi-class sentiment classification：The experimental comparisons of feature selection and machine learning algorithms [J]. Expert Systems with Applications，2017，80：323-339.

[6] DASGUPTA A，DRINEAS P，HARB B，et al. Feature selection methods for text classification [C] // Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM，2007.

[7] LIU H. Feature Selection for Knowledge Discovery and Data Mining[M]. Kluwer Academic Publishers，1998.

[8] KIVINEN J，MANNILA H. The power of sampling in knowledge discovery [C] // Proceedings of the thirteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. ACM，1994：77-85.

[9] BALCÁZAR J，DAI Y，WATANABE O. A random sampling technique for training support vector machines [C] // International Conference on Algorithmic Learning Theory. Springer-Verlag，2001.

[10] FERRAGUT E M，LASKA J. Randomized Sampling for Large Data Applications of SVM [C] // International Conference on Machine Learning & Applications. IEEE Computer Society，2012.

[11] LEE Y J，MANGASARIAN O L. RSVM：reduced support vector machines[C] // SIAM International Conference on Data Mining. 2001.

[12] LEE Y J，HUANG S Y. Reduced Support Vector Machines：A Statistical Theory[J]. IEEE Transactions on Neural Networks，2007，18(1)：1-13.

[13] LI X，CERVANTES J，YU W. Fast classification for large data sets via random selection clustering and Support Vector Machines[M]. IOS Press，2012.

[14] ZHANG L，GUO J. A Method for the Selection of Training Samples Based on Boundary Samples[J]. Journal of Beijing University of Posts and Telecommunications，2006，29(4)：77-80.

[15] ALMEIDA M B D，BRAGA A D P，BRAGA J P. SVM-KM：Speeding SVMs learning with a priori cluster selection and k-means[C] // Brazilian Symposium on Neural Networks. IEEE，2000.

[16] LLOYD S P. Least squares quantization in PCM [J]. IEEE Trans，1982，28(2)：129-137.

[17] GUAN D，YUAN W，LEE Y K，et al. Improving supervised learning performance by using fuzzy clustering method to select training data[J]. Journal of Intelligent & Fuzzy Systems，2008，19(4)：321-334.

[18] ZHOU Y，ZHU A F，ZHOU L，et al. Sample data selection method for neural network classifiers[J]. Journal of Huazhong University of Science and Technology(Natural Science Edition)，2012，40(6)：39-43.

[19] PEDRYCZ W. From fuzzy sets to shadowed sets：Interpretation and computing[J]. International Journal of Intelligent Systems，2010，24(1)：48-61.

[20] CHEN J，ZHANG C，XUE X，et al. Fast instance selection for speeding up support vector machines[J]. Knowledge-Based Systems，2013，45(3)：1-7.

[21] SHEN X J，MU L，LI Z，et al. Large-scale support vector machine classification with redundant data reduction [J]. Neurocomputing，2016，172：189-197.

[22] KANG J，RYU K R，KWON H C. Using Cluster - Based Sampling to Select Initial Training Set for Active Learning in Text Classification[C] // Pacific-asia Conference on Knowledge Discovery & Data Mining. Springer Berlin Heidelberg，2004.

[23] XU Z，YU K，TRESP V，et al. Representative sampling for text classification using support vector machines [C] // European Conference on Ir Research. Springer-Verlag，2003.

[24] VAPNIK V N，VAPNIK V. Statistical Learning Theory [J]. John Wiley and Sons，Inc.，1998.

[25] WAN C H，LEE L H，RAJKUMAR R，et al. A hybrid text classication approach with low dependency on parameter by integrating k-nearest neighbor and support vector machine[J]. Expert Systems with Applications，2012，39(15)：11880-11888.

[26] MATEI R，POP P C，VÂLEAN H . Optical character recognition in real environments using neural networks and k-nearest neighbor[J]. Applied Intelligence，2013，39(4)：739-748.

[27] GONZÁLEZ M，BERGMEIR C，TRIGUERO I，et al. On the stopping criteria for k-nearest neighbor in positive unlabeled time series classification problems [J]. Information Sciences，2016，328：42-59.

[28] HART B P E. The condensed nearest neighbor rule [J]. IEEE Transactions on Information Theory，1968，14(3)：515-516.

[29] GATES G W. The reduced nearest neighbor rule (Corresp.) [J]. IEEE Transactions on Information Theory，1972，18(3)：431-433.

[30] RITTER G L，WOODRUFF H B，LOWRY S R，et al. An algorithm for a selective nearest neighbor decision rule (Corresp.) [J]. IEEE Transactions on Information Theory，1975，21(6)：665-669.

[31] DASARATHY B V. Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design[J]. IEEE Transactions on Systems Man & Cybernetics，1994，24(3)：511-517.

[32] ANGIULLI F. Fast condensed nearest neighbor rule[C] // International Conference on Machine Learning. ACM，2005.

[33] SHIN H，CHO S. Neighborhood Property-Based Pattern Selection for Support Vector Machines [J]. Neural Computation，2007，19(3)：816-855.

[34] LI J, WANG Y P. A Fast Neighbor Prototype Selection Algorithm Based on Local Mean and Class Global Information [J]. Acta Automatica Sinica, 2014, 40(6): 1116-1125.

[35] WILSON D L. Asymptotic properties of nearest neighbor rules using edited data[J]. IEEE Transactions on Systems Man & Cybernetics, 1972, SMC-2(3): 408-421.

[36] TOMEK I. An Experiment with the Edited Nearest-Neighbor Rule[J]. IEEE Transactions on Systems Man & Cybernetics, 2007, SMC-6(6): 448-452.

[37] HATTORI K, TAKAHASHI M. A new edited k-nearest neighbor rule in the pattern classification problem[J]. Pattern Recognition, 1999, 33(3): 521-528.

[38] SHI X X, HU X G, LIN Y J. K-nearest neighbor classification algorithm combined with mutual neighbors and credibility[J]. Journal of Hefei University of Technology, 2014, 37(9): 1055-1058.

[39] YU G H. Instance Selection for Complex Classification [D]. Tianjin: Tianjin University, 2014.

[40] LOPEZCHAU A, GARCIA L L, CERVANTES J, et al. Data Selection Using Decision Tree for SVM Classification [C]// IEEE International Conference on Tools with Artificial Intelligence. IEEE Computer Society, 2012.

[41] CERVANTES J, LAMONT F G, LÓPEZ-CHAU A, et al. Data selection based on decision tree for SVM classification on large data sets[J]. Applied Soft Computing, 2015, 37(C): 787-798.

[42] YANG M H, AHUJA N. A Geometric Approach to Train Support Vector Machines[J]. Proc. IEEE Conf. Computer Vision & Pattern Rec, 2000, 1(6): 430-437.

[43] CRISP D J, BURGES C J C. A geometric interpretation of ν-SVM classifiers[C]// International Conference on Neural Information Processing Systems. MIT Press, 1999.

[44] PENG X. Efficient geometric algorithms for support vector machine classifier[C]// Sixth International Conference on Natural Computation. IEEE, 2010.

[45] LUO Y, YI W, HE D, et al. Fast reduction for large-scale training data set[J]. Journal of Southwest Jiaotong University, 2007, 42(4): 468-460.

[46] LIU C, WANG W, WANG M, et al. An efficient instance selection algorithm to reconstruct training set for support vector machine[J]. Knowledge-Based Systems, 2017, 116(1): 58-73.

[47] ZHU F, YE N, YU W, et al. Boundary detection and sample reduction for one-class Support Vector Machines[J]. Neurocomputing, 2014, 123: 166-173.

[48] LI C L, LIU Z D, HUI K H. Boundary Sample Selection Method Based on Cosine Similarity [J]. Computer and Modernization, 2017(8): 66-70.

[49] ZHANG A A, ZHENG P, FANG L, et al. A Sample Reduction Method for SVDD and Its Application[J]. Jiangxi Science, 2014, 32(6): 884-889.

[50] PAN D, YIN Y, SUN Y, et al. Sample Selection in Support Vector Machines: A Fixed Neighborhood Sphere Approach[C]// 2016 3rd International Conference on Information Science and Control Engineering (ICISCE). IEEE, 2016.

[51] LIU C, WANG W, WANG M, et al. An efficient instance selection algorithm to reconstruct training set for support vector machine[J]. Knowledge-Based Systems, 2017, 116(1): 58-73.

[52] KANGAS J. Prototype Search for a Nearest Neighbor Classifier by a Genetic Algorithm[C]// International Conference on Computational Intelligence & Multimedia Applications. IEEE, 1999.

[53] AMIREZ-CRUZ J F, FUENTES O, ALARCON-AQUINO V, et al. Instance Selection and Feature Weighting Using Evolutionary Algo-rithms[C]// 2006 15th International Conference on Computing. IEEE, 2006.

[54] NALEPA J, KAWULOK M. Adaptive Genetic Algorithm to Select Training Data for Support Vector Machines[M]// Applications of Evolutionary Computation. Springer Berlin Heidelberg, 2014.

[55] KAWULOK M, NALEPA J. Dynamically Adaptive Genetic Algorithm to Select Training Data for SVMs[M]// Advances in Artificial Intelligence—IBERAMIA 2014. 2014.

[56] KAWULOK M, NALEPA J, DUDZIK W. An Alternating Genetic Algorithm for Selecting SVM Model and Training Set [C]// Mexican Conference on Pattern Recognition. Cham: Springer, 2017.

[57] OTHMAN O M. Instance-Reduction Method based on Ant Colony Optimization[C]// Proceedings of the 2018 10th International Con-ference on Machine Learning and Computing. ACM, 2018: 47-53.

[58] WANG J, NESKOVIC P, COOPERL N. Selecting Data for Fast Support Vector Machines Training [M]// Trends in Neural Computation. 2007.

[59] HARA K, NAKAYAMA K, KARAF A A M. A Training Data Selection In On-Line Training For Multilayer Neural Networks [C]// IEEE World Congress on IEEE International Joint Conference on Neural Networks. IEEE, 2017.

[60] WANG Z Y, WANG M W, ZUO J L, et al. The New Boundary Sample Selection Method and Its Application in the Text Classification [J]. Journal of Jiangxi Normal University (Natural Science Edition), 2019, 43(1): 76-83.

**ZHOU Yu**, born in 1979, Ph.D, associate professor. His main research interests include intelligence computing and intelligent control.