主动式学习策略研究综述

张剑，潘晓衡，袁华强 / 东莞理工学院工程技术研究院

摘 要：近年来，主动式学习策略越来越受到研究者们的关注，并且有了许多重要的研究成果。其核心思想是通过选择有效的训练样本，使得机器学习算法能在使用更少训练数据的情况下，同样能达到良好的性能。首先对有关研究成果进行综述，然后展望未来主动式学习策略可能的应用前景。

关键词：主动式学习；确定性原则；成员性原则；样本选择；支持向量机

随着互联网技术的广泛应用，人们接触到的网络资讯日益膨胀，高效地处理海量信息并从中挖掘出有用信息的需求，显得日益迫切。研究者们提出了各种高效的机器学习算法，通过计算机进行信息处理与挖掘。但是这些机器学习算法的性能，却有赖于训练数据的质量和数量。所以标注训练数据的人工成本往往就成为机器学习算法应用的瓶颈。主动式学习策略提供了一种在保证机器学习算法性能的同时，通过选择有效未标注样本，进而最小化人工标注工作量的方法。主动式学习策略大体上可以分为：（1）基于确定性原则[1][2]；（2）基于成员性原则[3][4]；两类主流方法。它们已经被广泛地应用于网络上的自然语言处理任务。下面将针对这两类现有的主动式学习策略方法及其应用的研究成果进行综述；然后对主动式学习策略的应用研究进行展望。

1 基于确定性原则的主动式学习策略及应用

应用基于确定性原则的主动式学习策略进行机器学习算法训练时，首先需要选择小部分样本进行标注，这些样本被称为种子样本；然后使用这些种子样本训练出初始模型；接着使用初始模型，计算出其它未标注样本的确定性分值，再根据准则从中选择出一部分样本再进行标注；最后重新训练模型，如此反复迭代，直到结束条件满足。

Schohn&Cohn[1]提出了一种简单的主动式学习策略用于训练支持向量机，极大地提高了支持向量机的泛化能力，并在书面文档分类任务中进行了验证。研究结果发现，只用该主动式学习策略选择出来的小部分样本，训练出来的支持向量机模型，其性能强于使用整个数据库训练出来的模型。总所周知，支持向量机的训练时间随着训练数据量的增加而大幅度增加，因此，该主动式学习策略能更高效地训练高性能的支持向量机模型。Tong&Koller[2]也应用了该策略进行支持向量机的训练。

基于确定性原则的主动式学习策略已经应用于不同的自然语言处理任务中，比如：语音理解[5]，信息抽取[6]，多媒体检索[7]等。Turetal.[5]将基于确定性原则的主动式学习策略与半监督学习算法相结合，以进一步减少训练模型所需的标注样本，并在语音理解任务中进行了验证实验。基于确定性的主动式学习策略同样也被应用于自动语音识别任务当中[8][9]。

2 基于成员性原则的主动式学习策略及应用

应用基于确定性原则的主动式学习策略进行分类算法训练时[4][10]，首先选取若干组不同的分类算法，使用种子样本进行初始模型的训练；然后使用初始模型对未标注样本进行类别预测，再选取那些被不同算法训练出来的模型预测类别结果差异性的样本，进行人工标注；最后，将新标注的样本放入训练数据库，重新再训练模型，如此反复迭代，直到结束条件满足。

Seungetal.[11]提出了一种成员问询投票机制的主动式学习策略。Freundetal.[12]进一步分析了这一策略。他们通过从一组随机输入串中过滤信息量高的问询。研究结果显示：如果采用基于两位成员的委员会投票机制算法，它能够取得正向的信息增益，那么其预测误差将随着询问数目的增加而指数式的减小，特别是用于神经网络算法中感知元的学习。

Argamon-Engelson&Dagan[4]将这一策略进行了规范化，并应用于概率框架的分类算法训练当中。进一步他们引入了投票熵值用于量化委员会成员之间的分歧性。最后在词性标注任务中进行了实验验证。该策略的一个不足之处就是为了训练多种不同的分类器，需要将样本的特征集拆分为若干部分。这样，可能使得原本可以用于训练出一个高性能分类器的样本，最后只训练出若干个低性能的分类器。为了克服这一不足，Abe&Mamitsuka[13]提出了新的基于问询投票机制的策略，即：将问询投票与Boosting和封装机制相结合。

3 总结与展望

本文对近年来主动式学习策略应用领域所取得的研究成果进行了全面的综述。基于主动式学习策略的机器学习算法训练，能够帮助人们尽可能少的标注训练数据，更快速地训练出高性能的机器学习模型。在未来几年中，它将成为本领域研究热点和前沿。如何将该策略应用到更多不同的机器学习算法训练当中，尤其是在大数据背景下，如何克服模型训练效率低下的瓶颈等都将成为研究者关注的主要方向。

参考文献：

[1]G.SchohnandD.Cohn,Lessismore:Activelearningwithsupportvectormachines[C].inMachineLearning-InternationalWorkshopTHENConference-,2000,pp.839-846.

[2]S.TongandD.Koller,Supportvectormachineactivelearningwithapplicationstotextclassification[J].TheJournalofMachineLearningResearch,vol.2,pp.45-66,2002.

中图分类号：G42

下转第208页》》》》》》

案例或者说模型，以供不同学科的学生来学习。因此，设计这样使用不同专业的案例，能有效的培养学生的专业思维，这种思维形式其实就是一种计算思维，这种思维能音带学生逐步形成解决问题的思维，通过案例来透视问题的关键，并抓住哦解决问题的根本，把脉专业的动向。更为重要的是，这种与实践结合的计算思维能帮助学生解决实践中遇到的问题。为学生的实践打下了牢固的基础。甚至有些学生在学习中还会积极探索新的学习方法。

3.3 通过实验法培养学生的计算思维。实验法目前在理科，尤其是化学、物理、医学方面应用较为广泛，而在大学生计算机基础课程却运用的较为稀少。实验方法能够让学生形成一种计算思维，在做实验前会做出大胆的猜测，而在实验中会小心求证，并最终将实验得出的结论与最初的预测结果进行比对，这样一个过程，可以培养出学生客观、系统的计算思维。综上所述，如果能将这种实验法运用到大学生计算机基础课程中去，便能很好的培养学生的计算思维。在中实验法中，教师要引导学生积极的破除实验教学必须严密一幅理论的陈旧模式，让学生根据一定的依据大胆做出推测。并在实验中，综合运用自己此前所学到的知识，而能因此形成一种全面的计算思维，学习能力也能大大提高。

3.4 汇聚和整合教学资源。要创造适合培养学生计算思维的环境，必须要大力整合教学资源，为计算思维的培养和开展创造一定的物质条件。比如说，在现代化的信息化的时代背景下，提供网络技术和信息技术的支持自然是必不可少，要充分运用多媒体技术为学生营造出一种适合培养计算思维的氛围来，让学生能自由轻松的学习，通过各种案例开拓自己的眼界，提供多元的发展条件，学生可以利用技术上的优势结合自身的特点，寻找适合自己的多元发展目标。这样能很大的激发学生的潜能，让每个学生都形成属于自己的计算思维。

教师要积极的投身到教学中，为学生搭建一个颇具指导性的框架，同时也不要束缚学生发展，让学生感到束手束脚，相反要利用现代化的信息、网络优势让学生自己对周边的资源进行一定力所能及的整合，不断拓展自己的发展空间。

参考文献：

[1]朱鸣华,赵铭伟,赵晶,林鸿飞.计算机基础教学中计算思维能力培养的探讨[J].中国大学教学,2012(3):33-35.

[2]陈国良,董荣胜.计算思维与大学计算机基础教育[J].中国大学教学,2011(1):7-11.

[3]牟琴,谭良.计算思维的研究及其进展[J].计算机科学,2011(3):10-15.

[4]朱勇,张芳,李晓辉.农业院校大学生“计算思维”意识的培养[J].高等农业教育,2012(3):89-91.

[5]龚沛曾,杨志强.大学计算机基础教学中的计算思维培养[J].中国大学教学,2012(5):51-54.

作者简介：赵阳（1978-），男，山东微山人，硕士，讲师，主要研究方向:计算机应用技术、软件设计。

作者单位：山东农业工程学院信息科学与工程系，济南 250100

《《《《《《上接第206页

[3]A.McCallumandK.Nigam,EmployingEMinPool-basedActiveLearningforTextClassification[C].inProceedingsofICML,pp.350-358,1998.

[4]S.Argamon-EngelsonandI.Dagan,Committee-basedsampleselectionforprobabilisticclassifiers[J].JournalofArtificialIntelligenceResearch，vol.11,pp.335-360,1999.

[5]G.Tur,D.Hakkani-Tr,andR.E.Schapiro,CombiningActiveandSemisupervisedLearningforSpokenLanguageUnderstanding[J].SpeechCommunications，vol.45,pp.171-186,2005.

[6]D.Shen,J.Zhang,J.Su,G.Zhou,andC.Tan,Multi-criteria-basedActiveLearningforNamedEntityRecognition[C].inProceedingsof42thAnnualMeetingoftheAssociationforComputationalLinguistics.AssociationforComputationalLinguisticsMorristown,NJ,USA,2004.

[7]S.AyacheandG.Qu enot,Evaluationofactivelearningstrategiesforvideoindexing[J].SignalProcessing:ImageCommunication,vol.22,no.7-8,pp.692-704,2007.

[8]R.Rose,B.Juang,andC.Lee,Atrainingprocedureforverifyingstringhypothesesincontinuousspeechrecognition[C].inAcoustics,Speech,andSignalProcessing,1995.ICASSP-95.,1995InternationalConferenceon,vol.1,1995.

[9]R.ZhangandA.Rudnicky,Wordlevelconfidenceannotationusingcombinationsoffeatures[C].inSeventhEuropeanConferenceonSpeechCommunicationandTechnology.ISCA,2001.

[10]D.Cohn,L.Atlas,andR.Ladner，“Improvinggeneralizationwithactivelearning[J].MachineLearning,vol.15,no.2,pp.201-221,1994.

作者简介：张剑（1982-），男，江西南昌人，助理研究员，博士，研究方向：语音理解、语音文摘、自然语言理解、人工智能；潘晓衡（1983-），男，湖南湘潭人，工程师，硕士，研究方向：机器学习、智能计算、人工智能；袁华强（1966-），男，湖南湘潭人，教授，博士，研究方向：机器学习、人工智能。

作者单位：东莞理工学院工程技术研究院，广东东莞 523808

基金项目：广东省高等学校科技创新项目(2012KJCX0099)，2012年广东省自然科学基金博士启动基金(No.S2012040007560)，2012年东莞理工学院校博士启动基金(No.ZJ120408)。