基于主动半监督学习的智能电网信调日志分类

年素磊[1] 黎 铭[1] 杜 科[2] 姜 远[1] 林为民[2] 郭经红[2]
(南京大学软件新技术国家重点实验室 南京 210093)[1] (中国电力科学研究院 南京 211106)[2]

摘 要 智能电网的通信调度系统是智能电网正常运行的保证。为保证系统正确运行,值班员需要对电网信调系统的运行状态、突发事件、事故故障以及相应的处理方案进行记录。为帮助管理者及时了解智能电网信息调度系统的工作情况,发现潜在安全隐患,通常需要为这些日志数据标注其日志类型,以方便管理者查询和检索,因此,要求智能电网信息调度系统能够自动对每天记录的各种日志根据管理需要进行分类。对大量根据值班员自己理解和习惯撰写的日志进行自动分类,需要对由信息调度专家提供类型标注的大量日志数据进行学习。然而因人工阅读标注耗时、耗力,故在实际应用中往往仅能提供少量的标注,从而影响自动分类的性能。针对这一问题,提出了基于主动半监督学习的日志自动分类方法,该方法一方面利用主动学习找出对学习最有帮助的日志,获得其类型标注;另一方面,通过利用大量缺乏类型标注的日志进一步提升学习性能。在国家电网的智能电网信息调度日志数据上的应用结果表明,基于主动半监督学习,可获得比现有方法更优的日志自动分类性能。

关键词 数据挖掘,机器学习,主动半监督学习,信调日志分类,智能电网
中图法分类号 TP181 **文献标识码** A

Classifying Communication Dispatch System Logs of Smart Grid Based on Active Semi-supervised Learning

NIAN Su-lei[1] LI Ming[1] DU Ke[2] JIANG Yuan[1] LIN Wei-min[2] GUO Jing-hong[2]
(National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)[1]
(China Electric Power Research Institute, Nanjing 211106, China)[2]

Abstract Communication dispatch system is the guarantee for the normal operation of the smart grid. In order to ensure the correct operation of the system, man on duty need to record the operational status, emergencies, the accident fault as well as the corresponding treatment program of the communication dispatch system of smart grid. To help managers to keep up with the working status of system, for finding the potential security risks, the logs need to be labeled for certain types, to facilitate managers to query and retrieve, so the communication dispatch system needs to be able to automatically classify recorded logs according to various demands of management. However, the automatic classification for logs recorded by attendants in terms of their own understanding and habits, needs to learn from a large number of labeled logs data provided by information scheduling experts. Since manually reading to label is a time-consuming and labor-intensive process, only a small amount of labels are often provided in practical applications, thus affecting the performance of the automatic classification. In terms of this limitation, this paper proposed an automated classification method based on active semi-supervised learning. This method, on one hand, acquires the labels of logs that can improve the classifier most through active learning, on the other hand, further enhance learning performance by the use of larges number of unlabeled logs. The results of application on logs of communication dispatch system of national smart grid show that the method based on active semi-supervised learning can achieve better performance than existing methods.

Keywords Data mining, Machine learning, Active semi-supervised learning, Log classification, Smart grid

1 引言

智能电网是未来电网的发展方向,其实现基础是双向、实时、集成的通信系统。通信系统调度则是为确保信息通信系统安全、高效、经济、优质运行,对信息通信系统运行进行的组织、指挥和协调的总称,是公司信息及通信系统运行、检修以及客服的核心与枢纽。

为了保证通信调度系统正确运行,值班人员需要对信调系统的运行状态和发生的各类事件进行记录,形成值班日志。日志内容包括调度系统目前是否正常运行,如果有故障发生,故障的现象、原因及解决过程是怎样的,系统的例行检修计划执行得如何,遇到了什么紧急但未对系统造成重大影响的事

到稿日期:2012-02-27 返修日期:2012-07-14 本文受国家自然科学基金(60903103),国家电网公司科技项目(EPRIXXKJ[2012]2918)资助。
年素磊(1990—),男,硕士生,主要研究方向为机器学习、数据挖掘等;黎 铭(1980—),男,博士,副教授,主要研究方向为机器学习、数据挖掘、软件挖掘等,E-mail:lim@lamda.nju.edu.cn(通信作者);杜 科(1977—),男,博士,工程师,主要研究方向为电力系统自动化及信号处理等;姜 远(1976—),女,博士,教授,博士生导师,主要研究方向为机器学习、信息检索、数据挖掘等;林为民(1965—),男,硕士,教授级高工,主要研究方向为电力系统自动化及电网信息安全等;郭经红(1967—),男,博士,教授级高工,主要研究方向为电力系统自动化及其通信等。

件，以及总部对各分公司进行的调度联络信息等。由于很多事件（比如故障）从发生到解决跨越了很长时间，因此日志是对这类事件的一个很好的概括和追踪。信调系统的管理者可以通过查看日志来了解系统的工作情况，检查故障解决的进度，发现潜在的安全隐患。由于日志都是以流水账的方式进行记录，因此关于同类事件、同一事件的日志内容可能分散在多条记录，非常不便于管理者查询。如果能为这些日志数据标注类型，则可根据其类别进行查询和检索，例如，管理者可以查看事故的日志。因此，智能电网信息调度系统需要能够对每天记录的各种日志根据管理需要进行自动分类。

然而，这些大量的值班日志都是值班人员根据自己的理解和习惯撰写的，不同的人对同一事件的记录方式和描述侧重点有所不同，对同一个故障的危害程度的认识也不相同。要对其自动分类，需要首先在带有类型标注的日志数据上进行学习。为了达到好的分类效果，往往需要大量的带类型标注的日志。而对这些日志数据进行准确的类型标注，只能由信息调度专家来完成。由于专家人工阅读日志再去标注耗时、耗力，因此在实际应用中往往仅能提供少量的标注日志，从而影响自动分类的性能。针对这一问题，本文提出了基于主动半监督学习的日志自动分类方法。该方法一方面利用主动学习找出对学习最有帮助的日志，并将其提交给专家来标注其类型；另一方面利用大量无类型标注的日志进行学习，以进一步提升学习性能。在国家电网的智能电网信息调度日志数据上的应用结果表明，本文的主动半监督日志分类方法可获得比现有的监督文本分类方法更优的日志自动分类性能。

本文第 2 节介绍相关工作；第 3 节介绍主动半监督日志分类方法；第 4 节汇报实验；最后是结束语。

2 相关工作

2.1 文本分类

文本分类经过多年的发展，已成功应用在了很多领域，包括新闻的分类[1]、电子文档的分类[2]、网页的分类[3]等。Pazzani 等[4]运用文本分类技术，通过对用户上网时的网页评级进行学习，自动地向用户推荐感兴趣的内容。Zhang 等[5]通过采用最大化熵模型，结合文本分类技术，实现了对垃圾邮件的过滤。

大多数领域的文本分类处理的文本对象都比较长，而随着短信、twitter、微博和即时通信工具等的普及与流行，短文本内容（如微博长度上限为 140 字）飞速增加，传统的文本分类技术无法完全适用于短文本内容的处理。因此，最近一些短文本分类技术被提了出来，包括 Sriram 等人[6]以 twitter 特定领域中部分用户的个人信息及所发文本作为特征样本，将短文本划分到预定义的分类之中；Liu 等人[7]通过在不同对话中选取大信息量的词语，再用知网拓展这些词语的语义特性来提取新的微博短文本特征等。上述方法尽管已能根据微博短文本的特点提供相应的处理办法，但大多仅能针对特定领域的短文本进行处理，而全面、高效地进行短文本分类仍需要更加系统地研究。

2.2 半监督学习

半监督学习是利用未标记数据的一种主流学习技术，它能够在不加外界干预的情况下，自动地利用少量已标记数据和大量未标记数据进行学习。目前半监督学习方法大致可以分为以下 4 类：基于生成式模型的方法[8,9]、基于低密度划分的方法[10,11]、基于图和流形正则化的方法[12,13]以及基于不一致性的方法[14-17]。

在这些方法中，与本文最相关的是基于不一致性的方法。最早的基于不一致性的方法是协同训练（co-training），它是由 Blum 和 Mitchell[14]提出的，该算法要求问题域存在两个充分冗余（sufficient and redundant）视图，即两个不同的属性子集，每个子集都包含了足够训练出完美学习器的信息。该算法在每个视图上利用有标记样本训练一个分类器，然后两个学习器通过对未标记数据进行预测，互相提供新的有标记训练样本。为了放松上述约束条件，Goldman 等人[15]提出了一种改进算法，其不受属性集划分的约束，可用整个属性集训练两个分类器，但对分类器有了限制，导致算法时间复杂度较高。Zhou 和 Li[16]于 2005 年提出了一种既不要求充分冗余视图，也不要求使用不同类型分类器的 tri-training 算法[16]。该算法的一个显著特点是在单一视图上使用了 3 个分类器，该方法的优点在于不需要两个视图，因此削弱了协同训练的视图完备独立的前提假设。在对未标记样例进行分类时，tri-training 算法不再像以往算法那样挑选一个分类器来使用，而是通过使用集成学习中的投票法，将 3 个分类器组成一个集成学习器来实现。之后为了提高集成学习的效率，Li 和 Zhou[17]又于 2007 年提出了多分类器半监督学习方法 CoForest。

2.3 主动半监督学习

主动学习[18]是另一种有效地利用未标记样本的技术，它假设学习器对环境有一定的控制能力，可以“主动地”向学习器之外的某个“神谕”（oracle）进行查询来获得训练例的标记。主动学习有两类主要的方法：一种是挑选“最不能确定的”实例（uncertainty sampling），例如 Lewis 等[1]训练单个的分类器，然后挑选此分类器最不确定的未标记样本来查询。另一种是“基于委员会的”方法（committee-based sampling），例如文献[19,20]构建一个包含多个分类器的“委员会”，然后挑选那些“委员会”成员预测分歧最大的未标记样本来查询。

将主动学习与半监督学习结合可以有效地利用两种方法的优势来提高对未标记样本的利用率。Muslea 等人[21]将主动学习与半监督学习结合，用于多视图学习问题中，提出了更鲁棒的多视图学习方法。Zhu 等人[22]使用高斯随机场以及谐波函数来进行半监督学习，他们首先基于训练例建立一个图，图中每个结点就是一个（有标记或未标记）示例，然后求解根据流形假设定义的能量函数的最优值，从而获得对未标记示例的最优标记。Zhou 等人[23]将主动学习与半监督学习集成进基于内容的图像检索的相关反馈过程中，减少了用户需要提交的相关反馈的标记样本，增强了图像检索的质量。在垃圾邮件过滤问题中，Xu 等人[24]通过结合主动学习与半监督学习，成功解决了过滤器模型更新的问题，更新过程中仅需要用户提供少量标注即可完成。

3 主动半监督日志分类方法

3.1 数据表示

信调日志记录有固定的格式，每一条记录都由 5 部分组成，表 1 中为两条典型的日志记录。对于每条记录，本文仅取“问题情况”和“处理情况”，并将其拼接起来，用来生成特征向量。选定一组日志集合，首先进行中文分词，本文使用的分词

系统为中科院计算所开发的汉语词法分析系统 ICTCLAS;分词后对记录集中出现的所有词进行统计,去掉无意义的虚词、连词、符号等停止词(stopword),然后生成记录集的字典(dictionary);将字典中的词作为向量空间模型(VSM)的特征项,并用式(1)TF-IDF 进行加权。

$$TFIDF_{i,j}=TF_{i,j}\times IDF_i \tag{1}$$

式中,$TF_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$

$$IDF_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|}$$

$n_{i,j}$表示词语 t_i 在日志 d_j 中出现的次数,$|D|$表示日志总数,$|\{j:t_i\in d_j\}|$表示包含词语 t_i 的日志数。

表 1 日志记录示例

问题情况	处理情况	前班遗留	状态	记录人	标签
右屏多项显示无数据	29 日,9 点右屏显示中病毒木马感染数无数据,被攻击数无显示,联系可视化李浩处理,经查,数据库中无 ISS 最新数据,最后一次传送数据的时间是 28 日中午 14:00,联系 ISS 项目组检查,请接班人员继续跟踪。	是	执行中	陈剑	事故
今日检修情况	1 日应执行计划检修 2 项,实际执行 2 项,检修执行前后各单位均电话汇报。	否	完结	徐晓飞	非事故

3.2 主动半监督学习方法

在进行分类时,本文采用了 Li 等[26]最近提出的 ACoForest 算法。ACoForest 以一个非常有名的集成学习算法随机森林(Random Forest)[25]为基础,将主动学习与半监督学习成功地结合起来,它的伪代码如表 2 所列。首先,在有标记的训练集上学习一个随机森林。然后,在学习算法的每一轮迭代中,用学好的随机森林先标记所有的未标记样本,标记时随机森林中所有的随机树会独立地对样本进行预测,预测越一致表示标记的置信度越高,随后逆排序取出预测最不一致的 M 个样本,请用户标记。最后,用每棵随机树的补集成对未标记样本进行标记,把置信度高的(大于给定阈值)样本加入这棵树的标记样本集中,重新训练这棵树,依此迭代直到式(2)不再满足[26]。进行日志分类时,采用每棵随机树分别对日志类别进行预测,然后通过多数投票决定最后日志类别的方式。

表 2 ACoForest 算法伪代码[26]

算法:ACoForest
输入:
有标记数据集 L,无标记数据集 U,置信度阈值 θ,随机树的个数 N,每一轮迭代中要查询的样本的个数 M
过程:
1. 构建一个有 N 棵随机树的随机森林 $H=\{h_1,h_2,\cdots,h_N\}$
2. 重复 3—11 直到随机树都不再发生变化
3. 更新迭代轮数 t(t=1,2,…)
4. 在 U 中选出 M 个无标记样本,他们是用随机森林分类时分歧最大的 M 个
5. 请求用户标记这 M 个样本,并将其放入 L 中
6. 对于 H 中的每一棵随机树 h_i,做 7—11 步
7. 构建补集成 H_{-i}
8. 用 H_{-i}标记所有的无标记数据,然后估计标记的置信度
9. 把标记置信度大于阈值 θ 的无标记样本加入到一个新的有标记样本集合 $L'_{t,i}$
10. 对 $L'_{t,i}$采样使式(2)成立。如果它不成立,跳过第 11 步
11. 使用 $L\cup L'_{t,i}$学一棵随机树来更新 h_i
输出:
H,它的预测是由所有的成员树进行多数投票产生的

$$\frac{\hat{e}_{i,t}}{\hat{e}_{i,t-1}}<\frac{W_{i,t-1}}{W_{i,t}}<1 \tag{2}$$

4 实验

实验使用的数据集为从国家电网公司收集到的 6 个月的智能电网信息调度日志,表 3 展示的是对应月份的日志数、平均日志的长度(用字数描述)以及事故与非事故的比例(其中信调系统中无 2012 年 4 月的数据)。在将日志转换成向量空间模型表示后,每条日志表示成一个 1824 维的向量。

表 3 6 个月份对应的日志数、平均日志的长度(用字数描述)以及事故与非事故的比例

时间	日志数	平均日志长度	事故/非事故比例
2011 年 11 月	219	51	94/125
2011 年 12 月	156	57	49/107
2012 年 01 月	198	90	55/143
2012 年 02 月	192	129	49/143
2012 年 03 月	175	132	68/107
2012 年 05 月	135	138	56/79

本文将 ACoForest 和 4 种学习算法在日志分类问题上的性能进行了对比。CoForest[17]是 Li 和 Zhou 提出的半监督学习方法,它与 ACoForest 的区别在于没有主动学习的部分,另外 3 种算法分别是 J4.8 决策树[27]、支持向量机[28]和随机森林[25]。所有算法均基于 WEKA[29]机器学习软件包实现,参数采用 WEKA 的默认参数设置。

为评价分类性能,本文使用了 F-Measure 度量。若记召回率为 R(Recall)、准确率为 P(Precision),则 F-Measure 可以表示为

$$F\text{-}Measure=\frac{2RP}{R+P} \tag{3}$$

信调系统收集到 11 月份的数据后,先在其上随机采样一些样本(记为 a 个)作为标记样本,采样的比率为 0.3,其余的仍为未标记样本,经过这样划分后得到的样本集作为初始的训练集,在其上训练 ACoForest。训练过程中,ACoForest 会使用主动学习方法,再次请求得到一些样本的标记,每次请求的个数设为 20 个(记总共请求了 b 个)。训练完成后,将样本集中的所有未标记样本作为测试样本,用训练好的模型进行预测。而对比的 4 种算法则在原始的 11 月份的数据集上随机采样一定样本作为标记样本,其数量与 ACoForest 总共使用的标记样本数(即 $a+b$ 个)相同,用这些标记样本作为训练集,剩下的未标记样本则作为测试集。

当 12 月份的数据到来后,将其作为未标记样本合并到总体样本集中,然后在扩大后的总体样本集上重新训练 ACoForest。训练过程中 ACoForest 会再次请求得到一些样本的标记,每次请求个数仍设为 20(记总共请求了 c 个)。训练结束后,将总体样本集中的所有未标记样本作为测试样本,用训练好的模型进行预测。之后到来的 4 个月份的数据的处理情况与 12 月份的做法相同。对比的 4 种算法则在新数据合并到总体样本集后,在未标记的样本集中随机采样一定样本作为标记样本,其数量与 ACoForest 使用主动学习请求的样本数相同(即 c 个),然后在新的标记样本集上重新训练,剩下的未标记样本则作为测试集。

实验重复了 100 轮，记录对比方法的 F-Measure 均值(如图 1 所示)。从图 1 可以看出，在标记相同数量的样本的情况下，ACoForest 能提供一个更好的解决方案，成对 t 检验表明，ACoForest 的性能显著优于其他算法。注意到随着月份的增加，所有算法的性能都表现出一定程度的下降，原因可能是加入了难以分类的日志，为此本文分别在每个月的数据上进行交叉验证，结果如表 4 所列。可以看到，2 月、3 月的数据本身就比其他月份的数据更难分类，因此图 1 中的所有算法在 2 月、3 月数据加入后，性能都有所下降，而在 5 月份的数据加入后，性能又开始有所回升。

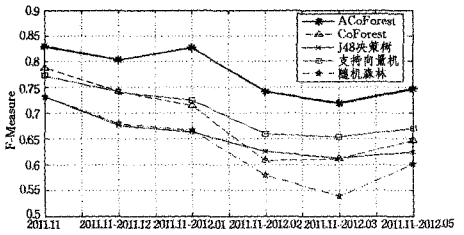

图 1　ACoForest 与 4 种学习算法在日志分类问题上的性能对比

表 4　支持向量机、决策树和随机森林分别在 6 个月的数据上做十折交叉验证的 F-Measure 值

时间	支持向量机	决策树	随机森林
2011 年 11 月	.777	.760	.782
2011 年 12 月	.742	.687	.739
2012 年 01 月	.725	.691	.525
2012 年 02 月	.511	.500	.413
2012 年 03 月	.662	.529	.542
2012 年 05 月	.873	.807	.737

由于人工标记的代价高，在实际应用中往往仅能得到少量的标记样本，因此希望主动挑选出的样本能有效地提高分类效果。为了验证 ACoForest 主动挑选样本的效用，本文在上述数据集上进行了另一个实验。实验设置大致相同，由于 ACoForest 在最后预测时会利用半监督学习的效果，为了公平地比较，排除半监督学习对最后预测的影响，本文选用 2 个相同的随机森林算法做比较，其中一个算法使用 ACoForest 主动挑选出的标记样本做训练，在总体样本集中剩下的样本上做测试；而另一个算法则在相同大小的随机采样的训练集上做训练，在总体样本集中剩下的样本上做测试。

实验结果如图 2 所示，其中红线表示的是随机森林在 ACoForest 主动挑选出的标记样本上做训练后得到的分类结果，而绿线表示的是随机森林在随机采样得到的相同大小的训练集上训练后得到的分类结果。对结果的成对 t 检验表明，采用 ACoForest 主动挑选样本训练的随机森林性能显著优于采用随机采样策略挑选样本的随机森林。这说明，ACoForest 主动挑选出的样本质量很高，能有效地提高分类效果，当仅能得到少量的标记样本时，能更充分地利用人工标记的宝贵机会。

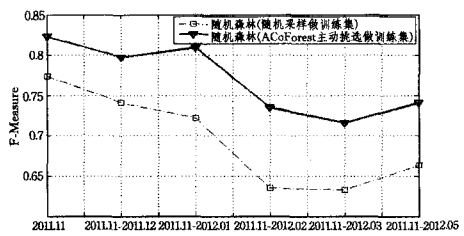

图 2　主动挑选样本和随机挑选样本的效用对比

结束语　本文提出了一种基于主动半监督学习的智能电网信调系统日志分类方法。该方法首先将日志记录进行分词、提取字典，然后用 TF-IDF 进行加权，将日志记录转换成向量空间模型中的特征向量表示，最后使用基于主动半监督学习的方法来对向量表示的日志进行自动分类。该方法一方面利用主动学习找出对学习最有帮助的日志，获得其类型标注；另一方面，通过利用大量缺乏类型标注的日志进一步提升学习性能。在国家电网公司的智能电网信息调度日志数据上的应用结果表明，本文提出的方法能获得比现有方法更优的日志自动分类性能。

将此方法应用于智能电网的信息调度系统中，是接下来要进行的工作。在将日志内容表示成特征向量的过程中，本文并没有做特殊的特征提取，如何选择对分类更有意义、表达力更强的特征是一个十分值得研究的方向。

参 考 文 献

[1] Lewis D, Gale W. A sequential algorithm for training text classifiers[C]// Proceedings of the 17th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 1994: 3-12

[2] Kim S, Han K, Rim H, et al. Some effective techniques for naive bayes text classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(11): 1457-1466

[3] Craven M, DiPasquo D, Freitag D, et al. Learning to extract symbolic knowledge from the World Wide Web[C]// Proceedings of the 15th National Conference on Artificial Intelligence. Madison, Wisconsin, USA, 1998: 509-516

[4] Pazzani M J, Muramatsu J, Billsus D. Syskill & webert: Identifying interesting web sites[C]// Proceedings of the 13th National Conference on Artificial Intelligence. Portland, Oregon, 1996: 54-61

[5] Zhang L, Yao T. Filtering junk mail with a maximum entropy model[C]// Proceedings of 20th International Conference on Computer Processing of Oriental Languages. Shenyang, China, 2003: 446-453

[6] Sriram B, Fuhry D, Demir E, et al. Short text classification in twitter to improve information filtering[C]// Proceedings of the 33th International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland, 2010: 841-842

[7] Liu Z, Yu W, Chen W, et al. Short text feature selection for micro-blog mining[C]// Proceedings of International Conference on Computational Intelligence and Software Engineering. Wuhan, China, 2010: 1-4

[8] Nigam K, McCallum A K, Thrun S, et al. Text classification from labeled and unlabeled documents using EM[J]. Machine Learning, 2000, 39(2/3): 103-134

[9] Miller D J, Uyar H S. A mixture of experts classifier with learning based on both labeled and unlabeled data[C]// Mozer M C, Jordan M I, Petsche T, eds. Advances in Neural Information Processing Systems 9. Cambridge MA: MIT Press, 1997: 571-577

[10] Joachims T. Transductive inference for text classification using support vector machines[C]// Proceedings of the 16th International Conference on Machine Learning. Bled, Slovenia, 1999: 200-209

(下转第 207 页)

大大提高获得最优解的概率。利用动态粒子群算法建立网格工作流调度问题的目标模型，并从跨时间粒度、跨时区、跨工作时间3个方面对工作流服务主体优选方法进行了讨论分析。实验结果表明，该方法比其他应用网格工作流调度的算法具有更短的执行时间和费用，实现了时间-费用的双重优化，具有更高的效率、更好的优越性。

参 考 文 献

[1] De P, Dunne E J, Ghosh J B, et al. The discrete time-cost tradeoff problem revisited[J]. European Journal of Operational Research, 1995, 81(2): 225-238

[2] Buyya R, Abramson D, Giddy J, et al. Economic models for resource management and scheduling in grid computing[J]. Concurrency and Computation: Practice and Experience Journal (Special Issue on Grid Computing Environments), 2002, 14(13-15): 1507-1542

[3] Lin M, Lin Z X. A cost-effective critical path approach for service priority selections in grid computing economy[J]. Decision Support Systems, 2006, 42(3): 1628-1640

[4] Yu J, Buyya R, Tham C K. Cost-based scheduling of workflow applications on utility grids[C]// Proceedings of the 1st IEEE International Conference on e-Science and Grid Computing. Melbourne, Australia, 2005

[5] 郭文彩，杨扬. 基于遗传算法的网格服务工作流调度的研究[J]. 计算机应用，2006，26(1)：54-56

[6] Robinson J, Ragnat-Samii Y. Particle swarm optimization in electromagnetics[J]. IEEE Transaction Antennas ProPag, 2004, 52(2): 397-407

[7] Huang T, Mohan A S. A hybrid boundary condition for robust particle swarm optimization [J]. Antennas and Wireless Propagation Letters, 2005, 4(1): 112-117

[8] Mikki S, Kishk. An improved particle swam optimization technique using hard boundary conditions[J]. Microwave Opt Technol Lett Sep, 2005, 46(5): 422-426

[9] Liu J X, Zhou C J. Researeh on the Workday Model in Business Service Grid Environment[C]// Proc. of the 2005 International Workshop on Workflow Management Systems in Grid Environment. Changsha: IEEE Publisher, 2006

(上接第170页)

[11] Chapelle O, Zien A. Semi-supervised classification by low density separation[C]// Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. Barbados, 2005: 57-64

[12] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions[C]// Proceedings of the 20th International Conference on Machine Learning. Washington DC, USA, 2003: 912-919

[13] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[C]// Thrun S, Saul L, Scholkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge, MA: MIT Press, 2004: 321-328

[14] Blum A, Mitchell T. Combining labeled and unlabeled data with co-training[C]// Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, Wisconsin, USA, 1998: 92-100

[15] Goldman S, Zhou Y. Enhancing supervised learning with unlabeled data[C]// Proceedings of the 17th International Conference on Machine Learning. Stanford, CA, USA, 2000: 327-334

[16] Zhou Z H, Li M. Tri-training: Exploiting unlabeled data using three classifiers[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529-1541

[17] Li M, Zhou Z H. Improve computer-aided diagnosis with machine learning techniques using undiagnosed samples[J]. IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans, 2007, 37(6): 1088-1098

[18] Settles B. Active learning literature survey [R]. Computer Sciences Technical Report 1648, University of Wisconsin-Madison, 2009

[19] Abe N, Mamitsuka H. Query learning strategies using boosting and bagging[C]// Proceedings of the 15th International Conference on Machine Learning. Madison, Wisconsin, USA, 1998: 1-9

[20] Seung H, Opper M, Sompolinsky H. Query by committee[C]// Proceedings of the 5th Annual Conference on Computational Learning Theory. Pittsburgh, PA, 1992: 287-294

[21] Muslea I, Minton S, Knoblock C A. Active + semi-supervised learning = robust multi-view learning[C]// Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 435-442

[22] Zhu X, Lafferty J, Ghahramani Z. Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions[C]// Proceedings of ICML 2003 workshop on The Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. Washington DC, 2003

[23] Zhou Z H, Chen K J, Dai H B. Enhancing relevance feedback in image retrieval using unlabeled data[J]. ACM Transactions on Information Systems, 2006, 24(2): 219-244

[24] Xu J M, Fumera G, Roli F, et al. Training SpamAssassin with active semi-supervised learning [C]// Proceedings of the 6th Conference on Email and Anti-Spam, Mountain View. CA, 2009

[25] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32

[26] Li M, Zhang H, Wu R, et al. Sample-based software defect prediction with active and semi-supervised learning[J]. Automated Software Engineering, 2012, 19(2): 201-230

[27] Quinlan J R. C4. 5: programs for machine learning[Z]. Morgan Kaufmann, San Francisco, CA, USA, 1993

[28] Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273-297

[29] Witten I H, Frank E. Data Mining: Practical machine learning tools and techniques with Java implementations [Z]. Morgan Kaufmann, San Francisco, CA, USA, 2000