改进的知识特征驱动的任务分解模型

凡少强[1] 王国胤[1] 李美争[2]
(重庆邮电大学计算智能重庆市重点实验室 重庆 400065)[1]
(西南交通大学信息科学与技术学院 成都 610031)[2]

摘 要 任务分解被广泛应用于解决大而复杂的问题,学者们已经提出了很多分解模型。知识特征驱动的任务分解模型在无需过多先验知识的情况下,就可以将原始问题分解成一系列子问题,然而这种分解方式却没有考虑对子问题噪点进行处理。在知识特征驱动下,利用马氏距离可以去除子问题的噪点,并对子问题空间进行扩充,这就得到了一种去除噪点的知识特征驱动的任务分解模型。该模型在处理双螺旋问题、UCI abalone数据集、UCI yeast数据集时,都得到了较高的精度,说明了其可行性和有效性。

关键词 任务分解,知识特征驱动,马氏距离,自动分解
中图法分类号 TP181 **文献标识码** A

Improved Knowledge Characteristic-driven Task Decomposition Model

FAN Shao-qiang[1] WANG Guo-yin[1] LI Mei-zheng[2]
(Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)[1]
(School of Information Science & Technology, Southwest Jiaotong University, Chengdu 610031, China)[2]

Abstract Task decomposition is widely used to solve those large-scale and complex problems. Many researchers have presented their task decompositions. Knowledge characteristic-driven task decomposition model can divide the original problem into several smaller tasks without much prior experience. But this model forgets to treat the noisy point of subtask. Inserting a process of getting rid of noisy point and expanding the subtask, an improved knowledge characteristic-driven task decomposition model was obtained. We carried some experiments on two-spiral problem, UCI abalone data set and UCI yeast data set. The results show that our method can get a better accuracy.

Keywords Task decomposition, Knowledge characteristic-driven, Mahalanobis distance, Automatic decomposition

分类是数据挖掘中最重要的任务之一,被广泛应用于医疗诊断、金融分析、模式识别、基因分析、文本分类、语音识别等领域。常见的分类算法有决策树、贝叶斯、人工神经网络、k-临近、支持向量机SVM、基于关联规则的分类、集成学习、adaboost、bagging、基于案例的推理、遗传算法、粗糙集方法、模糊集方法等[1],然而这些分类算法在遇到现实世界的复杂问题或多类问题时总是显得心有余而力不足。为了克服这些缺点,学者们提出了许多任务分解的方法——将一个大而复杂的问题分解成几个规模较小的问题。解决这些规模较小的问题比解决原始问题要简单许多,小问题的求解恰恰又是许多经典分类算法的强项,通过组合这些小问题的解就可以给出初始问题的求解。任务分解最核心的问题之一就是如何将一个大而复杂的问题分解成多个小而简单的子问题。许多学者对此做出了研究,这些研究方法大致可以分成以下几类。

A. 专家式分解

在学习之前,领域专家就将问题分解成一系列的子问题[2-4]。基于这种方法的神经网络结构被用来解决倒车问题[5]和遥感信息处理[6]。这种方法的缺点是必须对相关问题有充足的先验知识。

B. 按类别分解

在学习之前,按照训练数据中各类别之间的固有关系将问题分解成一系列的子问题。常见的分解策略有:

a)一对多(One-Versus-All, OVA)。对于一个 k 类问题,OVA将第 i 类作为一个类别,其余 $k-1$ 类作为一类,训练一个二类分类器。这样共需要训练 k 个二类分类器。OVA更多地用于处理静态数据集。Rifkin和Klautau研究发现,当分类器合理组合时OVA可以得到与其他分类模式相同的精度[7]。Perdisci等将OVA应用到计算机网络异常发现当中[8]。Dain和Zach将adaboost作为二类分类器,其在笔迹识别中得到很好的实验结果[9]。然而OVA在解决多类问题时往往会造成子数据的严重不平衡,并且有时这种方法效率并不高,因为它并没有减少子问题的样本规模。

b) 一对一(One-Versus-One, OVO)。对任意 $i, j(1\leqslant i, j\leqslant k, i\neq j)$ 类样本训练一个分类器,因此 k 个类别的样本就需要设计 $k(k-1)/2$ 个分类器。

相比于专家式分解,按类别将一个问题分解则不需要相

到稿日期:2013-05-15 返修日期:2013-07-26 本文受国家自然科学基金(61073146,61272060)资助。
凡少强(1987-),男,硕士生,主要研究方向为数据挖掘,E-mail:osfan@qq.com;王国胤(1970-),男,博士,教授,主要研究方向为粗糙集、粒计算、机器学习、数据挖掘、知识技术和认知计算;李美争(1984-),女,博士生,主要研究方向为数据挖掘、粗糙集、粒计算。

关问题领域的专家知识，一般人都可以做到。但是当数据的类别较多时，求解原始问题时，OVO 不可避免会产生庞大而又复杂的组合计算。

C. 自动分解

学习原始问题时，利用一个分类器，将全部样本分解成两个部分：正确学习的部分和未正确学习部分；然后重新训练分类器递归学习未正确学习的样本，这样原始问题相应被分解成一系列的子问题[10-12,22]。这种知识特征驱动方式分解原始问题的方法，在处理不确定问题[13]或很难获得先验知识的问题时，可以根据知识本身的特性，自动将问题分解。但是这种方法在分解问题的过程中会产生碎片，即子问题的噪点。

为了克服知识特征驱动的任务分解过程中的噪点问题，本文提出了一种改进的知识特征驱动任务分解模型，其在知识特征驱动下得到子问题时，根据马氏距离将学习到的子问题中远离子问题的样本从中剔除，同时将未正确学习的数据中与子问题距离较近的数据合并到子问题当中，这样就有效地解决了自动分解过程中产生的碎片。在双螺旋、abalone 问题、yeast 数据集上的实验证明了本文方法的可行性。

本文第 1 节主要介绍了知识特征驱动的数据挖掘和马氏距离等相关概念；第 2 节给出了本文提出的改进的知识特征驱动任务分解模型与算法；对本文算法验证的仿真实验在第 3 节中给出；最后总结本文所做的工作。

1 相关概念

1.1 知识特征驱动的数据挖掘

传统的数据挖掘认为数据挖掘(data mining)是一个从大规模数据库中抽取有效的、隐含的、以前未知的、有潜在使用价值的信息的过程[14]。根据对数据挖掘本质和数据挖掘基本问题的研究，王国胤等指出所谓数据挖掘就是知识转换的过程[15]，并提出了面向领域的数据驱动的数据挖掘(3DM)，即知识特征驱动的数据挖掘(Knowledge Characteristic-Driven Data Mining，KCDDM)。王燕等根据对此的理解，用一系列的算法证明了该模型的有效性[16]。实现知识特征驱动的知识发现的核心是找到数据的特征并对其进行有效的度量，然后在知识本身所蕴含的这种特征的引导下，利用任何一种经典的数据挖掘及计算智能的方法，结合先验知识、用户的兴趣、领域约束等输入条件，完成数据挖掘的任务[16]。而在整个数据挖掘的过程中知识本身的基本特征是不变的。知识特征驱动的知识发现主要包括两个方面的内容：数据驱动的数据挖掘和面向领域的数据挖掘。

1.1.1 数据驱动的数据挖掘

知识存在于多种载体中，如数据、符号、自然语言等等。知识蕴含在数据中称为知识以数据的形式存在。从图 1[15]可以看出，数据挖掘其实就是一种知识转换的过程——将知识从人类难以理解的数据形式转换为人类可以理解的符号形式，期间并不产生任何新的数据。这就像将一本英文书籍翻译成中文的过程，而书籍里的知识本身保持不变，改变的只是知识的编码形式(语言)。为了保持知识不变，需要取得数据形式知识的某些属性，然后利用这些属性来控制数据挖掘的过程。这就是数据驱动数据挖掘模型的核心思想。

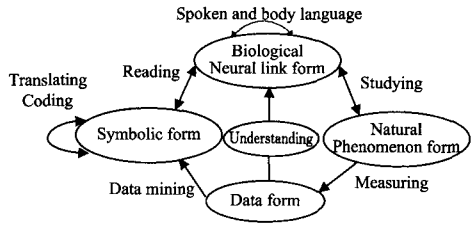

图 1 知识不同形式的转换

1.1.2 面向领域的数据挖掘

数据驱动的数据挖掘解决了如何从原始数据中挖掘大量的知识，但是数据挖掘应该沿着什么“方向”去挖掘？这就需要在数据挖掘中加入领域知识和用户兴趣等。Kuntz 等开发了以人为中心的方法来揭示社交规则，在数据挖掘过程中他们将用户作为驱动数据挖掘的启发算法[17]。Han 和 Lakshmanan 将基于约束和多维挖掘集合成一个框架，为有效和高效的数据分析和挖掘提供了一个互动的环境[18]。为了创建词汇知识库，Patrick 等利用领域专家来鉴定文本中字典流的构造元素，提出了一个半自主学习的方法[19]。面向领域驱动的数据挖掘有如下特点：

1)领域驱动的数据挖掘过程是基于约束的。

2)在领域驱动的数据挖掘过程中考虑到了用户的兴趣。

3)在领域驱动的数据挖掘过程中加入了领域专家的先验知识。

4)在领域驱动的数据挖掘过程中提供了用户和机器之间的交互。

数据驱动的数据挖掘和领域驱动的数据挖掘并不是矛盾的，而是可以将它们有机地结合起来，这就是知识特征驱动的数据挖掘。

1.2 马氏距离

马氏距离(Mahalanobis Distance)是印度数学家马哈拉诺比斯(Mahalanobis)在 1936 年首先提出的[20]，它采用样本协方差来计算两个样本之间的距离，是一种有效的计算两个未知样本集相似度的方法。

定义 1 设 T 是 p 维总体，数学期望为 μ，协方差矩阵为 Σ，定义 p 维样本 X 到总体 T 的马氏距离为：

$$D_M(X,G)=\sqrt{(X-\mu)'\Sigma^{-1}(X-\mu)} \tag{1}$$

设 T_1，T_2 为两个不同的 p 维总体，数学期望分别为 μ_1 和 μ_2，协方差矩阵分别为 Σ_1 和 Σ_2。设 X 为一个待判样本：

$D_M(X)=(X-\mu_2)'\Sigma_1^{-1}(X-\mu_2)-(X-\mu_1)'\Sigma_2^{-1}(X-\mu_1)$ 判断准则为：

$$\begin{cases} X\in T_1, & \text{若 } D_M(x)\geqslant 0 \\ X\in T_2, & \text{若 } D_M(x)<0 \end{cases}$$

马氏距离不受量纲的影响，即两点之间的马氏距离与原始数据的测量单位无关。这也是本文选择马氏距离而不是欧氏距离的最主要原因。

2 改进的知识特征驱动任务分解模型

知识特征驱动的任务分解(Knowledge Characteristic-Driven Task Decomposition，KCDTD)模型能够自动地将复杂学习任务分解为简单学习任务，经分别用不同的分类器进行处理，各分类器模块以并联方式构成整个知识特征驱动系统[22]。KCDTD 由控制分类器(Control classifier，CC)、识别控制器(Recognition classifier，RC) RC_i ($i=1,2,\cdots,p$) 和 LS_i

(或加法器、乘法器)构成,如图 2 所示。控制分类器的功能实际上就是对问题空间进行粗分解,判断输入向量 X 的归属(如 $X\in S_i$),相应地闭合逻辑开关 LS_i,然后将识别分类器 RC_i 的识别结果作为系统的有效输入,其余识别分类器不工作,因为识别分类器 RN_i 的功能实际上只能正确处理子问题 S_i。

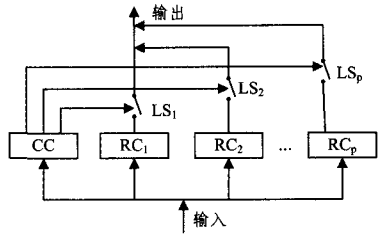

图 2 知识特征驱动结构图

知识特征驱动的任务分解方法得到的问题子空间是一组相似的数据,在实验过程中可以发现问题子空间会不可避免地出现些许噪点,即控制分类器会出现误分类的情况。由图 2 可以看出,控制分类器和识别分类器是串行的,控制分类器的错误会在识别分类器中被放大。为此本文提出了一种改进的处理子问题空间噪点的知识特征驱动任务分解(Improved Knowledge Characteristic-Driven Task Decomposition, IKCDTD)模型。初始问题 T 在知识特征驱动分解方式下得到分类器 RS_i 以及 RS_i 识别的问题子空间 S_i,令 $S_j=T-S_i$,$i\neq j$,KCDTD 只是简单地将原始问题划分为两个子问题空间 S_i 和 S_j,然后递归学习 S_j,如图 3(a)所示。

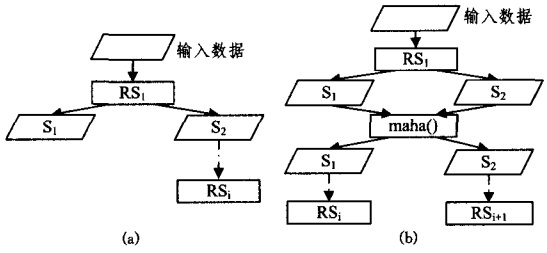

图 3 一次学习过程

2.1 IKCDTD 学习算法

设 T 为训练样本集,它包含 n_1 个样本,

$T=\{(X_j,Y_j)\mid j=1,2,\cdots,n_1\}$

其中,X_j 为第 j 个样本的输入向量,Y_j 为第 j 个样本的目标输出向量。整个学习过程如图 4 所示,详细算法如下:

1)令 $m=0,i=0,k=0,S_k=T$;

2)$i=i+1$,用所有训练样本训练一个分类器 RS_i,RS_i 能够正确学习到的数据为:

$S_{ai}=\{(X,Y)\mid \forall X_j->RS_i=Y_j\}$

其中 $X_j->RS_i=Y_j$ 表示 RS_i 对输入向量令 X_j 的预测与实际输出向量 Y_j 相符。令

$S_{bi}=\{(X,Y)\mid \forall X_j->RS_i\neq Y_j\}$

其中,$X_j->RS_i\neq Y_j$ 表示 RS_i 对输入向量 X_j 的预测与实际输出向量 Y_j 不相符。考虑到可能存在分类器对有些数据无法预测,设这类数据为 S_{ci},则显然有:

$S_k=S_{ai}\cup S_{bi}\cup S_{ci}$

3)计算样本集 S_{ai} 中的所有样本到 S_{ai} 的马氏距离,求得马氏距离的均值 μ_{ai}、最小值 $\min_{ai}$ 和最大值 $\max_{ai}$,令:$RAG_{ai}=\min_{ai}+\delta*(\mu_{ai}-\min_{ai})$,其中 δ 可根据人的先验知识确定。

3.1)去噪点:当$(X_j,Y_j)\in S_{ai}$ 且 $D_M(X_j)\geqslant 0$ 时,

$S_{ai}'=\{(X_j,Y_j)\mid \min_{ai}\leqslant D_M(X_j,S_{ai}')\leqslant RAG_{ai}\}$

$S_{bi}'=\{(X_j,Y_j)\mid \forall (X_j,Y_j)\in S_{bi}$ 或 $D_M(X_j,S_{ai}')\leqslant RAG_{ai}\}$

3.2)扩充:当$(X_j,Y_j)\in S_{bi}$ 且 $D_M(X_j)\geqslant 0$

$S_{bi}'=\{(X_j,Y_j)\mid \min_{ai}\leqslant D_M(X_j,S_{ai}')\leqslant \delta*RAG_{ai}\}$

3.3)得到两个新的样本集:

$S_{ai}''=S_{ai}'\cup S_{bi}'$

$S_{bi}''=S_{bi}-S_{bi}'$

4) $i=i+1$,用分类器 RS_i 学习 S_{ai}''。若学习满意,$m=m+1$,保存 RS_i 为 RC_m;否则 $k=k+1$,$S_k=S_{ai}''$,转向 2)。$i=i+1$,用分类器 RS_i 学习 S_{bi}''。若学习满意,$m=m+1$,保存 RS_i 为 RC_m;否则 $k=k+1$,$S_k=S_{bi}''$,转向 2)。

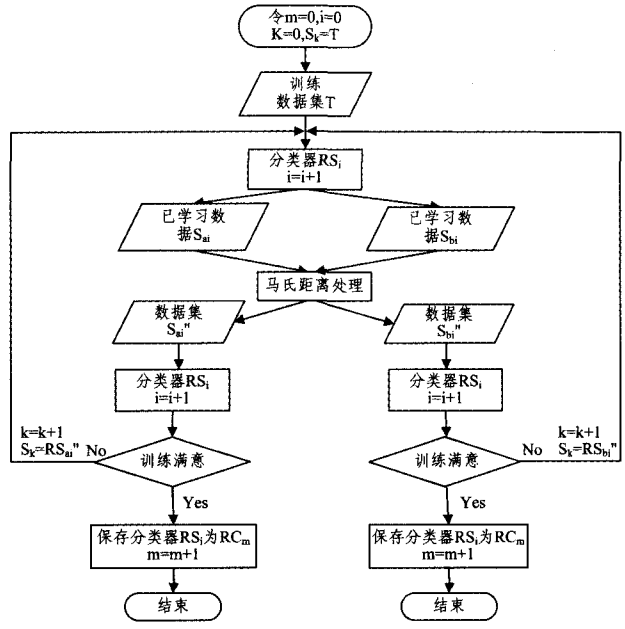

图 4 IKCDTD 学习原始训练集

2.2 IKCDTD 预测算法

预测数据集经过控制网络被分成两份,一份交由第一个学习子问题训练得到的分类器来学习,一份交由第一个未学习子问题训练得到的分类器学习。若将 IKCDTD 的预测过程画成一个二叉树的话,左支树代表学习到的问题的分类器,右支树代表未学习问题训练的分类器,节点代表学习分类器,则二叉树每个节点的左子树的深度不会超过 2,因为左侧的数据是由学习到的数据扩展而来的,一般这些数据不会超过两个同类分类器的学习能力,且二叉树的叶节点总是识别分类器,非叶节点总是控制分类器。一个常见的 IKCDTD 模型的结构如图 5 所示,预测数据经过控制分类器 CC_1 将预测数据分成两份,分别交由 CC_2 和 CC_3 来处理,CC_2 和 CC_3 再将任务下发,最终由识别分类器 RC_1、RC_2、RC_3、RC_4 等识别学习。一个大而复杂的任务就这样被分配成一系列小而简单的子任务,然后将 RC_1、RC_2、RC_3、RC_4 等子任务的解组合起来,就成为原始复杂问题的解。

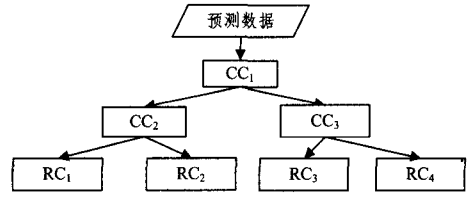

图 5 数据预测过程

3 实验仿真结果

3.1 双螺旋问题

双螺旋问题是监督学习算法的一个典型参照，对 BP 神经网络来说，双螺旋问题已经可以用来体现新的神经网络体系结构的性能。本实验所用到的双螺旋数据见参考文献[21]。图 6 显示了这两类螺旋：红色的圈表示类别 0 共 97 个实例，绿色的圈表示类别 1 共 97 个实例。通常认为单个隐含层的标准 back-program 算法对双螺旋问题很难达到良好而稳定的表现。国内外很多学者专家对该问题提出了各种解法；王国胤教授等采用 PNN 架构的筛选模型[22]；Timothy Bender 等利用自动分割模块的方法逐步求解整个双螺旋问题[23]；Chihiro Ikuta 等采用 Glial Network 与 Chaos 和 neuron 构建的 MLP[24]来巧妙地解决双螺旋问题。

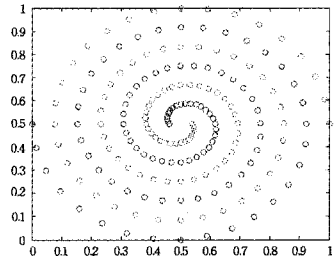

图 6 双螺旋数据的平面显示

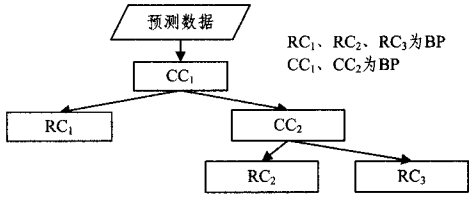

图 7 双螺旋数据训练的 IKCDTD 结构

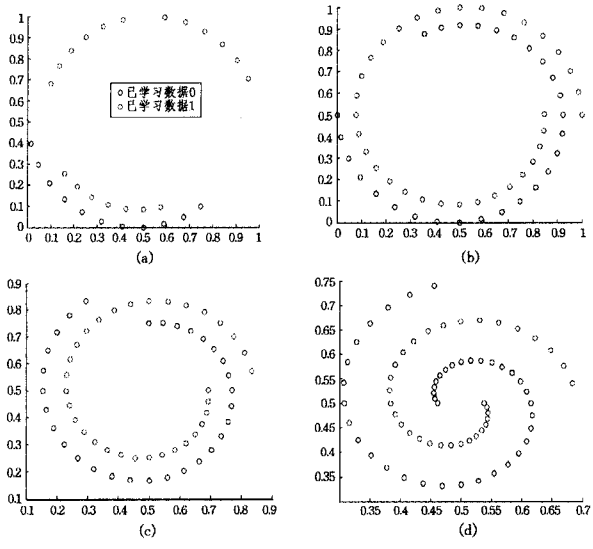

图 8 双螺旋学习过程

本文采用 IKCDTD 模型对该问题进行求解时，训练双螺旋数据后得到的 IKCDTD 模型的结构如图 7 所示。每个识别分类器 BP(RC_1,RC_2,RC_3)含有两个输入，隐含层采用 5 个神经元结点和一个输出，控制分类器 BP(CC_1,CC_2)的隐含层采用 10 个神经元结点。第一个识别分类器学习到的数据如图 8(a)所示，经过马氏距离处理之后的数据如图 8(b)所示。可以看出经马氏距离处理之后的数据更加集中且不会遗漏集中数据中的部分数据。图 8(c)和图 8(d)分别为第二次、第三次学习利用马氏距离处理后的数据。

图 9(a)显示了 IKCDTD 模型对双螺旋问题的处理结果，可以看出 IKCDTD 模型比较符合双螺旋的弧度走向、螺旋之间宽度相对平均、边缘平滑且没有出现误分类的情况。IKCDTD 模型训练较少的神经网络，共需要训练 5 个神经网络。图 9(b)是文献[22]的实验结果，其中红色方框内的区域分类效果明显不理想。图 9(c)是文献[23]采用 41 个神经网络所得到的实验结果，其边缘出现很多尖角形成锯齿状，且螺旋红色区块内的宽度太小，跟其他区域有明显差距，分类不甚理想。图 9(d)是文献[23]采用 17 个神经网络得到的实验结果，且左上角红色方框内已经出现了错误预测。

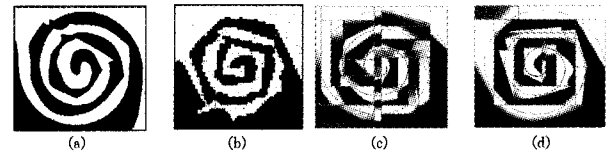

图 9 双螺旋实验结果比较

3.2 UCI Abalone 数据集

abalone 数据集根据测量鲍鱼环来反映鲍鱼的年龄，鲍鱼属无脊椎动物，在部分地区已经绝种，因此鲍鱼年龄的较好预测对生态学具有很大的影响。abalone 数据集包含 4177 个实例和 9 个属性。D. Clark, Z. Schreterd 等[25]将分类合并为 3 类(类 1—8、类 9 和类 10、类 11 以上)，采用隐含层含有 5 个神经元结点的 Back Propagation 神经网络的正确率可以达到 65%，也是最好的学习结果，其次 C4.5 的准确率可以达到 59%。Klaus Truemper[26]通过线性组合较小数量的属性的离散化，提高了算法的稳定性和可理解性，对 abalone 数据集的预测正确率提高至 72%。Sevilla 等[27]为数据密集型的基于案件推理的系统提出了一种灵活的 CBR 结构，其获取子任务的方法是在不固定的区域让专家构建一个 CBRS，该方法预测 abalone 数据集的正确率在 63.5%到 82.6%之间。

本文采用 IKCDTD 模型处理 abalone 问题时，子任务的识别分类器采用的是隐含层含有 5 个结点的 Back Propagation 神经网络，控制分类器采用的是隐含层含有 10 个结点的 BP 神经网络，正确率可以达到 80%左右，处理该问题的模型如图 10 所示。

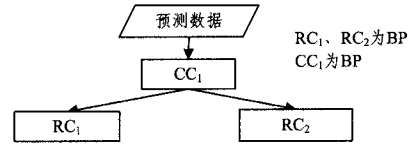

图 10 abalone 训练的 IKCDTD 结构

学习过程大致如下：第一次学习全部数据的 2051 个，剩余的 1081 个样本未学习，经马氏距离处理噪点和扩展之后，第一部分划分到 1622 个样本，第二部分划分到 1511 个样本，然后训练两个识别分类器 RC_1 和 RC_2 即可各自学习第一部分和第二部分，然后根据第一部分和第二部分的数据可以训练得到控制分类器 CC_1，预测样本经过控制分类器 CC_1 划分给 RC_1 分类器 592 个样本，RC_1 正确识别 592 个，CC_1 划分给 RC_2 分类器 452 个样本，RC_2 正确学习 363 个。若识别分类

器和控制分类器都采用 C4.5，IKCDTD 训练过程和结构跟采用 BP 的相仿，正确率为 80.84%，比经典算法的准确率提高了约 20%，比 D. Clark 和 Sevilla 的算法提高了约 10%。与 Klaus、D. Clark 和 Sevilla 的比较如表 1 所列。

表 1　IKCDTD 等 abalone 数据集的比较

abalone	Klaus/BP	Klaus/C4.5	Clark
正确率	65.61%	59.2%	72%
	Sevilla	IKCDTD/BP	IKCDTD/C4.5
正确率	63.5%～82.6%	80%	80.84%

3.3　UCI Yeast 数据集

yeast 数据集包括了 1484 个记录，包含 10 类，本文选择其中最多的 3 类：CYT、NUC、MID，共 1136 个记录。原始数据集的训练集和预测集比例为 40:15，为了与文献[26]作比较，根据文献[21]将数据的 50%作为训练器，其余 50%作为预测集。IKCDTD 模型的控制分类器和识别控制器都采用 C4.5 算法，对 yeast 数据集训练的 IKCDTD 模型结构图如图 11 所示。

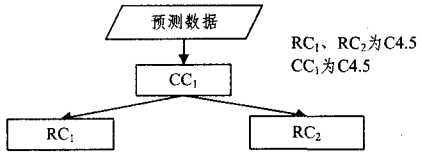

图 11　yeast 训练的 IKCDTD 结构

实验结果对比如表 2 所列。

表 2　TDDKC 等酵母菌数据集的比较

酵母菌类	CYT	NUC	MID	总体
C4.5	38.79	49.77	45.08	44.29
Sevilla	60	64	78	65.20
TDDKC	52.16	60.47	52.46	55.36

通过表 2 可以看出，控制分类器和识别控制器均采用 C4.5 的 TDDKC 模型，虽然与 Sevilla 论文中的实验结果相比逊色不少，但是每个类别和总体的正确率都高于传统的C4.5，本文算法整体预测准确率也比传统方法得到了提高。

结束语　知识特征驱动的任务分解可以根据知识本身的特性自动将复杂问题分解，在得到问题子空间后再经过马氏距离去噪和扩充，使得问题子空间更加集中和相似，最终使得控制网络的误分率降低。这种改进的知识特征驱动模型在双螺旋、abalone 和 yeast 数据集的实验上都达到了较高的精度。本文的所有试验中的控制分类器和识别控制器都是同构的，在未来的工作中，我们将研究该模型对分类器异构情况会产生怎样的效果。

参 考 文 献

[1] Han Jia-wei, Kamber V. 数据挖掘：概念与技术（第 2 版）[M]. 2007:277-280

[2] Zhu L, Yuan G, Du Q. An efficient explicit/implicit domain decomposition method for convection-diffusion equations [J]. Numerical Methods for Partial Differential Equations, 2010, 26(4): 852-873

[3] Bazán F S V, Gratton S. An Explicit Jordan Decomposition of Companion Matrices [J]. TEMA Tend. Mat. Apl. Comput, 2006, 7:209-218

[4] Rajkumar Murthy B E. Parallel alternating explicit implicit domain decomposition algorithm [D]. Texas Tech University, 2006

[5] Jenkins R E, Yuhas B P. A simplified neural network solution through problem decomposition: The case of the truck backer-upper [J]. IEEE Transactions on Neural Networks, 1993, 4(4): 718-720

[6] Thiria S, Mejia C, Badran F, et al. Multimodular architecture for remote sensing operations [C]// Moody J E, Hanson S J, Lippmann R P, eds. Advances in Neural Information Processing Systems4. San Mateo, CA: Morgan Kaufmann, 1992: 675-68

[7] Rifkin R, Klautau A. In defense of one-vs-all classification[J]. The Journal of Machine Learning Research, 2004, 5: 101-141

[8] Perdisci R, Gu G, Lee W. Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems [C]// Data Mining, 2006. ICDM'06. Sixth International Conference on. IEEE, 2006: 488-498

[9] Martlnez J, Iglesias C, Matlas J M, et al. DAGSVM Multiclass algorithm based on SVM binary classifiers with 1vsAll approach to the slate tile classification problem[C]// Modelling for Engineering and Human Behaviour. 2012: 115

[10] McGill K C, Cummins K L, Dorfman L J. Automatic decomposition of the clinical electromyogram[J]. IEEE Transactions on Biomedical Engineering, 1985(7): 470-477

[11] Cobo L C, Isbell C L Jr, Thomaz A L. Automatic task decomposition and state abstraction from demonstration[C]// Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems. 2012: 483-490

[12] Hasan M K, Apu M S, Molla M K I. A robust method for parameter estimation of AR systems using empirical mode decomposition[J]. Signal, Image and Video Processing, 2010, 4(4): 451-461

[13] 王国胤，何晓. 一种不确定性条件下的自主式知识学习模型[J]. 软件学报，2003，14(6)：1096-1102

[14] 史忠植. 知识发现[M]. 北京：清华大学出版社，2002

[15] Wang G, Wang Y. 3DM: domain-oriented data-driven data mining[J]. Fundamenta Informaticae, 2009, 90(4): 395-426

[16] 王燕，申元霞，陶春梅. 面向领域的数据驱动自主式知识获取模型及实现[J]. 重庆邮电大学学报：自然科学版，2009，21(4)：502-506

[17] Kuntz P, Guillet F, Lehn R, et al. A user-driven process for mining association rules [M]// Principles of data mining and knowledge discovery. Berlin Heidelberg: Springer, 2000: 483-489

[18] Han J, Lakshmanan L V S, Ng R T. Constraint-based, multidimensional data mining [J]. Computer, 1999, 32(8): 46-50

[19] Patrick J, Palko D, Munro R, et al. User driven example-based training for creating lexical knowledgebases [C]// Australasian Natural Language Processing Workshop. Canberra, Australia, 2002: 17-24

[20] Mahalanobis P C. On the generalized distance in statistics[J]. Proceedings of the national institute of sciences of India, 1936, 2(1): 49-55

[21] http://emergent.brynmawr.edu/eprg/? page=TwoSpiralsProblem

[22] 王国胤，施鸿宝，邓伟. 基于 NARA 模型和筛选方法的并行神经网络体系结构[J]. 计算机学报，1996，19(9)：679-686

[23] Bender, Timothy, et al. Partitioning Strategies for Modular Neural Networks[C]// International Joint Conference on Neural Networks (IJCNN 2009). Atlanta GA, 2009: 296-301

（下转第 99 页）

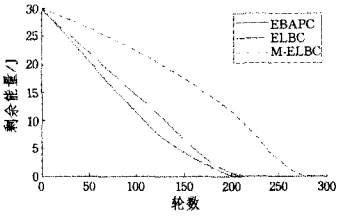

图 2　网络剩余能量情况

实验 3 主要评价算法对节点平均能耗的影响。从图 3 可以看出，ELBC 算法节点平均能耗优于 EBAPC 算法，且分布更均匀；而 M-ELBC 算法节点平均能耗明显低于 ELBC 算法和 EBAPC 算法，说明 M-ELBC 算法节点能量利用更加合理、高效，ELBC 算法次之。

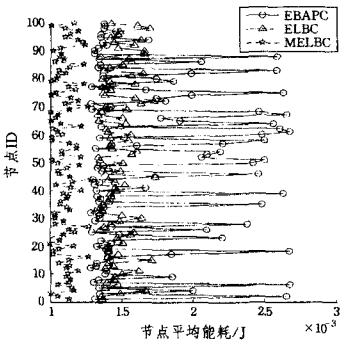

图 3　节点平均能耗

实验 4 主要评价基站位置对算法的影响。从图 4 可以看出，随着基站从(120,50)移向(180,50)，基站距离节点区域越来越远，网络死亡节点出现的时间越来越早，网络死亡时间也越来越早。可以看出，ELBC 算法和 EBAPC 算法在维持网络生存时间方面比较相近，而 ELBC 算法相较于 EBAPC 算法，能够有效地延缓死亡节点的出现时间；同时，可以看出 M-ELBC 算法在维持网络生存时间和延缓死亡节点的出现方面都明显优于 ELBC 算法和 EBAPC 算法。

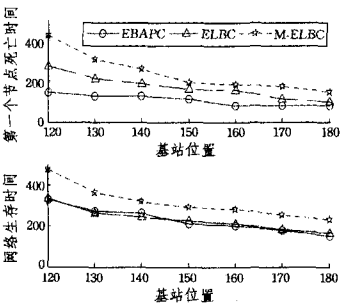

图 4　基站位置对算法的影响

结束语　通过分析网络整体能量状况和基站位置对选择策略的影响，并结合偏向参数的优化，提出了基于能量等级的 ELBC 分簇算法和多跳 M-ELBC 分簇算法。仿真实验表明，ELBC 算法在均衡能量消耗和避免死亡节点过早出现方面有很大的改进；多跳 M-ELBC 算法中节点能量消耗更均衡，有效地推迟了死亡节点的出现，能量利用更高效，延长了网络的生存时间。但是 ELBC 算法和 M-ELBC 算法在选择簇头时需要进行多次迭代，占用较多的存储空间和算法执行时间，这是下一步需要解决的问题。

参 考 文 献

[1] 张少军. 无线传感器网络技术及应用[M]. 北京：中国电力出版社，2010

[2] Heinzelman W, Chandrakasan A, Balakrishnan H. Energy-efficient communication protocol for wireless microsensor networks [C]// Proceedings of the Hawaii International Conference on System Sciences. 2000:3005-3014

[3] Younis O, Fahmy S. HEED: a hybrid, energy-efficient, distributed clustering apprpach for ad hoc sensor networks[J]. IEEE Transactions on Mobile Computing, 2004, 3(4): 366-379

[4] Lindsey S, Raghavendra C S. PEGASIS: power-efficient gathering in sensor information systems[C]// Proc. of IEEE Aerospace Conference. 2002(3): 1125-1130

[5] Heinzelman W, Chandraksasan A, Balakrishnan H. An application-specific protocol architecture for wireless microsensor networks[J]. IEEE Transactions on Wireless Communications, 2002, 1(4): 660-670

[6] Zytoune O, Fakhri Y, Aboutajdine D. A balanced cost cluster-heads selection algorithm for wireless sensor networks[J]. International Journal of Computer Science, 2009, 4(1): 21-24

[7] Wang Ning-bo, Zhu Hao. An energy efficient algorithm based on LEACH protocol[C]// 2012 International Conference on Industrial Control and Electronics Engineering. 2012: 339-342

[8] 冯江，吴春春. 基于能耗均衡的 WSN 多跳分簇路由算法[J]. 计算机工程，2012，38(16)：104-107

[9] 李岩，张曦煌，李彦中. LEACH-EE—基于 LEACH 协议的高效聚类路由算法[J]. 计算机应用，2007，27(5)：1103-1105

[10] 陈培培，张华忠. MHST-LEACH—基于 LEACH-EE 高效聚类路由算法[J]. 计算机工程与应用，2011，47(1)：120-122

[11] 胡艳华，张建军. LEACH 协议的簇头多跳(LEACH-M)改进算法[J]. 计算机工程与应用，2009，45(34)：107-109

[12] 崔可想，李志华. 基于能量的 EBAPC 分簇网络拓扑控制算法[J]. 计算机工程，2012，38(23)：104-108

[13] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976

[14] Li Zhi-hua, Li Peng-fei, Yin Xi, et al. Clustering network topology control method based on responsiblility transmission[J]. International Journal of Intelligence Science, 2012, 2(4): 128-134

[15] 邹瑜，彭舰，黎红友. 一种基于分层无线传感器网络的路由算法[J]. 计算机科学，2012，39(10)：65-68

(上接第 95 页)

[24] Ikuta C, Uwate Y, Nishio Y. Chaos Glial Network Connected to Multi-Layer Perceptron for Solving Two-Spiral Problem[C]// Proc. ISCAS'10. May 2010

[25] Clark D, Schreter Z, Adams A. A quantitative comparison of dystal and back propagation[C]// Proc. Austr. Conf. Neural Netw(ACNN). 1996

[26] Truemper K. Improved comprehensibility and reliability of explanations via restricted half space discretization[C]// Proceedings of International Conference on Machine Learning and Data Mining(MLDM 2009). 2009

[27] Sevilla Villanueva B, Sánchez Marrè M. Providing intelligent decision support systems with flexible data-intensive case-based reasoning[C]// International Congress on Environmental Modelling and Software. 2012