基于粗糙度的近似概念格动态分类集成学习模型研究与应用

丁卫平[1,2,3] 王建东[2] 朱 浩[1,2] 管致锦[1] 施 佺[1]

(南通大学计算机科学与技术学院 南通 226019)[1] (南京航空航天大学信息科学与技术学院 南京 210016)[2]

(苏州大学江苏省计算机信息处理技术重点实验室 苏州 215006)[3]

摘 要 概念格(Galois格)是一种进行数据分类学习的有效工具,然而建格规模庞大使分类效率和准确率受到较大影响。将粗糙度理论应用到概念格分类问题研究中,提出一种新型的近似概念格动态建格和分类挖掘集成学习模型(CACLR)。该模型在粗糙度区间根据样本空间分布构建多个相对独立分布且比较精确的近似概念格分类器,能及时消除建格过程中大量与分类知识无关的节点,有效缩减原格规模,融合得到的分类挖掘集成学习模型,具有较好的粗糙分类精度和知识预测学习能力。最后进行CACLR分类集成学习模型在标准UCI数据集中的对比实验,有效验证了该模型的实用价值。

关键词 粗糙度,近似概念格,集成学习,分类挖掘

中图法分类号 TP301.6 **文献标识码** A

Research and Application of Dynamical Classification Model for Ensemble Learning Based on Approximation Concept Lattice of Roughness

DING Wei-ping[1,2,3] WANG Jian-dong[2] ZHU Hao[1,2] GUAN Zhi-jin[1] SHI Quan[1]

(School of Computer Science and Technology, Nantong University, Nantong 226019, China)[1]

(College of Information Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)[2]

(Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou 215006, China)[3]

Abstract Concept lattice is an effective tool for data classification, but classification efficiency and precision are effected by its large scale. In this paper, rough sets theory was applied into the classification research of concept lattice, and a dynamical classification model(named CACLR) for ensemble learning based on approximation concept lattice of roughness was put forward. This model can constructe some identical approximation concept lattice classifiers of independent distribution and much precision according to the instance spatial configuration at the scope of roughness. And it can eliminate independent nodes in time during approximation concept lattice constructed, reduce the scale of concept lattice effectively. The multi-combination model for ensemble learning has robustness at the accuracy of rough classification and the efficiency of knowledge prediction. In the last part of this paper, the experiments tested on the UCI benchmark data sets were carried on and performance results of were given, which prove the practical value of CACLR model.

Keywords Roughness, Approximation concept lattice, Ensemble learning, Classification mining

20世纪90年代以来,分类学习是机器学习领域中一个非常重要的研究课题,其主要过程是先根据数据集的特点构造一个分类器,然后利用分类器对未知类别的样本赋予类别。构造分类器的过程一般分为模型训练和使用模型测试分类两个步骤:在训练阶段,通过分析训练数据集的特点为每个类别产生一个对相应数据集的准确描述或模型。在测试阶段,利用类别的描述或模型对测试数据集进行分类,测试其分类准确度。但是目前大多数分类算法(如决策树归纳法[1]、人工神经网络[2]和支持向量机[3]等)构造的分类器在训练数据集上和测试数据集上仅强调样本个体,构建单一分类器,用于训练学习集上的预测结果不准确且对"过拟合"数据集分类效果不佳。集成学习(Ensemble Learning)分类是在给定的训练集上构造多个性能较好而又相对独立的分类器,再通过决策优化(Decision Optimization)或覆盖优化(Coverage Optimization)

到稿日期:2009-08-31 返修日期:2009-10-03 本文受江苏省高校自然科学研究项目(09KJD520008),南通市应用研究计划项目(K2008031,K2008018),苏州大学江苏省计算机信息处理技术重点实验室开放课题,南通大学自然科学基金项目(05Z061),南通大学通信与信息系统学科科技创新基金资助。

丁卫平(1979—),男,博士生,讲师,CCF会员,主要研究方向为粗糙集、概念格和电子病历挖掘等,E-mail:dwp9988@163.com;王建东(1945—),男,教授,博士生导师,主要研究方向为人工智能和知识工程等;朱 浩(1977—),男,博士生,主要研究方向为网络安全等;管致锦(1962—),男,博士,教授,主要研究方向为量子可逆计算等;施 佺(1975—),男,博士生,副教授,硕士生导师,主要研究方向为数据挖掘和数据库技术等。

两种手段将若干弱学习器的预测能力进行综合，以达到优化学习系统得到最终的判定结果，该方法已成为当前机器学习领域热点[4]。

概念格（Galois 格）也称形式概念分析，由德国教授 Wille. R 于 1982 年首先提出，它是一种非常适合用来进行知识分类和学习的有效工具。概念格根据数据集中对象与属性之间的二元关系建立概念层次结构，生动简洁地体现了各概念间的泛化和特化关系[5]。概念格本质上描述了对象与属性之间的联系，每个概念都是对象（外延）与属性（内涵）的统一体。然而由于概念格的完备性和精确性，构造概念格中概念数目是形式背景大小的指数级函数，随着数据量巨增，庞大的格节点构造及其存储成为制约概念格的最主要因素，将大大影响概念格的分类性能。

粗糙集是处理不确定性和模糊性知识的有效工具，粗糙集的不确定性主要来自两个方面：一是由不可分辨关系将对象分成一些等价类，同一个等价类中元素就不可分辨而产生不确定性；二是粗糙集近似域中的下近似与上近似不相等时，存在边界，即不确定性也就存在。1991 年 Pawlak. Z 提出关于粗糙集中不确定性测量来用粗糙度方法，通过近似空间中下近似与上近似的近似域来确定粗糙度，较好地解决了由近似域所引起的不确定性问题[6]。该方法在粗糙集处理不确定领域引起较多的关注，有很多学者以此为基础提出一些改进方法等。

基于上述研究问题，本文将粗糙度等理论应用到概念格分类学习问题研究以弥补概念分类学习过程中的不足，并建立动态分类集成学习模型，主要研究内容和创新成果如下：(1)提出一种新型的基于粗糙度范围的动态近似概念格建格方法，动态消除大量与分类知识无关的数据节点，有效缩减原格规模；(2)根据样本空间分布构建多个相对独立且比较精确的近似概念格分类器，分别进行样本的分类学习；(3)通过特征重组和格空间优化，利用典型的 AdaBoost. MH 算法进行多个同类分类器集成，在符合一定的评价标准下得到基于粗糙度的近似概念格分类挖掘集成学习 CACLR 模型（Classification Model based on Approximation Concept Lattice of Roughness）；(4)通过对比实验验证该模型集成学习的准确性、高效性和适用性。

1 概念格构造和集成学习算法研究现状

概念格的每个节点是一个形式概念，由两部分组成：外延和内涵，外延即概念所覆盖的实例，内涵即该概念覆盖实例的共同特征。另外，概念格通过 Hasse 图生动、简洁地体现了这些概念之间的泛化和特化关系。属性值集合（即项目集）之间的关系在概念格之间得到了充分体现，从数据集（概念格中称为形式背景）中生成概念格的过程实质上是一种概念聚类过程[5]。在概念格分析中，概念格建格具有很重要的地位。对于同一批数据，所生成的格不受数据或属性排列次序的影响，具有唯一性，这是概念格最大的优点之一。自从概念格被提出以来，其建造算法一直成为学术界研究的热点问题，不同的学者在自己的研究领域提出了不同的建格方法。

目前概念格建格方法主要分为两个方面：批处理算法和增量算法。批处理算法目前分为 3 类，即从顶向下算法、自底而上算法和枚举算法。它们的思想和优缺点分析如下：从顶向下算法首先构造格的最上层节点，再逐渐往下构造，其代表有 Bordat 算法[7]，该算法很简洁、直观，并且易于并行化；缺点是重复生成了许多重复节点，每个节点都以其父节点个数的次数被重复生成；自底而上算法则相反，首先构造底部的节点，再一层一层向上建造，其代表有 Chein 算法[8]，该算法看起来比较容易理解，但是会在同层和下层中产生大量的冗余对，效率较差；枚举算法则是按照一定顺序枚举格的所有节点，然后生成 Hasse 图，即节点间关系，其代表有 Ganter 算法[9]，该算法通过拓扑排序方式建立概念格，速度相对较高，但其生成 Hasse 图概念间的关系不明显，难理解。增量算法把当前要插入的对象和概念格中所有的概念进行相交，由交的结果把格中节点分为不变节点、更新节点和新增节点，然后对其采取不同的处理方法。增量算法（含改进）目前也有很多算法，主要区别在于连接边的方法不同，其最具代表的有 Godin 算法[10]，该算法通过对与格相交后生成的 3 类节点进行不同处理来较好地建立格结构，但在搜索产生子格节点和新生格节点直接子节点时需要遍历所有节点，算法的时间性能和空间性能相对比较弱。目前也有很多其他建格算法，但大多数都是上述算法的改进或者演化。

集成学习算法是利用多个学习器的互补学习功能，可获得比仅使用单一学习器更强的泛化能力，能较好地减少分类误差。经过十几年的不懈努力，集成学习由最初的萌芽逐步发展起来，很多研究者提出了各种不同算法，其中以 Boosting 和 Bgaging 最为典型，Boosting 的基本思想是把一个弱学习器通过逐阶段最小化一个特定的误差函数的梯度下降从而转化为一个任意高精度的学习器，Bgaging 的基本设计思想是改变样本分布，提高错误样本概率，使下一次的弱学习机能够集中精力针对那些困难样本进行学习。本文采用以 Boosting 算法为基础的改进的典型 AdaBoost. MH 算法，它将一个多类问题转化为一系列两分类问题，组合相同类型“狭义”的学习分类器对同一个问题进行学习。其具体实施过程为：首先给每一个训练样例赋予相同的权重，然后训练第一个基分类器并用它来测试训练集，对于那些分类错误的测试样例提高其权重，最后用调整后的带权训练集训练第二个基分类器，重复这个过程直到符合一定标准的组合建立集成学习分类器[11,12]。

在概念格建格过程，最坏的情况是概念格中节点数按指数级快速增长，用该格进行知识学习的能力较差。因此对大规模数据集，必须控制格中节点的增长。对于批处理建格算法，可在建格过程中引入一个支持度门限。对于支持度小于门限的节点，不予继续展开而达到剪枝的目的。对于增量式算法，需要在建格过程中进行反复修剪，情形就较为复杂。近年来有些学者提出利用内涵缩减来简化概念格[13]，基于搜索空间划分进行概念格生成[14]以及概念格的横向合并等[15]。由于概念格的完备性和精确性大大制约了概念格构造效率，因此利用修剪等方法简化格有时效果也不是太好。我们知道粗糙度利用下近似与上近似的近似域范围来确定其值，带有一定的不确定性和模糊性，因此可以开展两者技术的融合，在粗糙度一定区间范围内研究动态近似概念格建格，建立独立分布且比较精确的基本近似概念格分类器，并利用 AdaBoost. MH 算法将多个分类器进行集成学习，这给概念格和分类学习的研究必然带来新的思路。

2 基于粗糙度的近似概念格构造理论基础

2.1 粗糙度

定义 1[6] 令 $X\subseteq U$，R 是 U 上的一个等价关系。当 X 为 R 的某些等价类的并时，称 X 是 R 可定义的(R-definable)，否则称 X 是 R 不可定义的(R-undefinable)。R 可定义集称为 R 精确集，R 不可定义集称为 R 粗糙集。粗糙集可以用两个精确集，即粗糙集的下近似和上近似来描述。

任意给定一个集合 $X\subseteq U$，如果使用 R 等价类无法精确描述 X，则 X 就是 R 的粗糙集；反之 X 是 R 的精确集。包含在 X 中的最大可定义集称为 X 的 R 下近似(Lower Approximation)：

$$\underline{R}(X)=\{x\in U|[x]_R\subseteq X\}$$

包含 X 的最小可定义集称为 X 的 R 上近似(Upper Approximation)：

$$\overline{R}(X)=\{x\in U|[x]_R\cap X\neq\phi\}$$

X 的边界域 $B_n(X)=\overline{R}(X)-\underline{R}(X)$，也可以将粗糙集表示为：$POS_R(X)=\underline{R}(X)$ 为 X 的 R 正域，$NEG_R(X)=U-\overline{R}(X)$ 为 X 的 R 负域。

定义 2[16,17] 设近似空间 (U,R)，U 为有限非空集合，R 为 U 上的一个等价关系，$X\subseteq U$，为了测量粗糙集中的不确定性，Z. Pawlak 用粗糙度来反映知识的不完全程度，其定义如下：

$$\rho_R(X)=1-\alpha_R(X)=1-\frac{|\underline{R}(X)|}{|\overline{R}(X)|}$$

式中，$\underline{R}(X)$，$\overline{R}(X)$ 分别为 X 的下、上近似。

如果 $\rho_R(X)=0$，则集合 X 关于 R 是普遍集合；如果 $\rho_R(X)>0$，则集合 X 关于 R 是粗糙的。

显然 $0\leqslant\rho_R(X)\leqslant 1$，粗糙集中不确定越大，$\rho_R(X)$ 也就越大，因而可用粗糙度来测量粗糙集中的不确定性。

定义 3 在决策信息系统 $S=(U,C\cup D,V,f)$ 中，$\emptyset\neq P\subseteq C$，任意属性 $a\in C-P$ 关于属性集 P 相对于 D 的粗糙近似重要度 $SGF(a,P,D)$ 定义为：

$$SGF(a,P,D)=(1-\mu(a,P\cup\{a\},D))(E(P)-E(P\cup\{a\}))$$

由上述定义可知，$SGF(a,P,D)$ 的值越大，说明在已知 P 的条件下，属性 $a\in C-P$ 关于知识 P 就越重要。

定义 4 对于分类规则 $A\Rightarrow B$，设粗糙对象 RO 为可能属于 B 类的对象集，精确对象 PO 为精确属于 B 类的对象集，则分类规则 A 中条件的对象满足 B 的可能粗糙度可定义为：

$$\eta=\rho_R(PO/RO)=\rho_R(PORO)/\rho_R(RO)=\rho_R(PO)/\rho_R(RO)$$

对分类规则 $A\Rightarrow B$，其粗糙度 η 的取值区间为 $[0,1]$。

当 $\eta=0$ 时，说明无精确属于 B 类的对象，该规则无意义；

当 $\eta=1$ 时，说明可能属于 B 类的对象全都精确地属于 B，该规则退化为传统的分类规则。

粗糙度越接近于 0，规则不确定程度越大，反之，规则越精确。

2.2 近似概念格分类

定义 5[5] 给定形式背景 K 上的一个序偶 (X,S)，其中 $X\subseteq U$，$S\subseteq D$，如果满足 $f(X)=S$ 并且 $X=g(S)$，则称 (X,S) 为一个概念。X 称为概念 (X,S) 的外延，S 称为概念 (X,S) 的内涵。形式背景 K 中的所有概念及概念之间的偏序关系构成的结构称为概念格，记作 $L(U,D,R)$，其中每个概念被看作概念格中的一个节点。

定义 6[18] 任何形式概念的外延可看成是一个可定义集，借助粗糙集上下近似算子，对任意一个对象集 $X\subseteq U$，其对象的可定义集就是形式概念的外延，概念格的外延可用上下逼近来描述。在近似空间 $apr=(U,L)$，基于粗糙集理论的关于对象子集 X 的上下近似概念格可定义为：

$$\overline{laprX}=extern(\wedge\{(A,B)|(A,B)\in L,X\subseteq A\})=\bigcap\{A|(A,B)\in L,X\subseteq A\}$$

$$\underline{laprX}=extern(\vee\{(A,B)|(A,B)\in L,A\subseteq X\})=(\bigcup\{A|(A,B)\in L,A\subseteq X\})^{**}$$

定理 1[13] 对于概念格中的概念结点 $C_1=(U_1,A_1)$，$A_2\in P(A)$，且 $A_2<A_1$，称 A_2 是结点 (U_1,A_1) 概念的内涵缩减当且仅当

(1) $g(A_2)=g(A_1)=U_1$；

(2) $\forall A_3<A_2$，$g(A_3)\supset g(A_2)=U_1$。

条件(1)称为外延不变性，即内涵 A_1 和它的内涵缩减 A_2 具有相同的外延。

条件(2)称为内涵缩减的最小性，即从中除去任意一个原子特征都会导致外延的增加。

根据上述定理，设形式背景 $K=(U,A,R)$，其上概念结点 $C_1=(U_1,A_1)$，设特征子集 A_2 是其内涵缩减，$A_2\leqslant A_1$，设结点 $C_1=(U_1,A_1)$ 的所有直接父结点为 $\{C_i=(U_i,A_i)|i=3,4,\cdots,p\}$，则对任意 $i=3,4,\cdots,p$，$A_i<A_2$，否则由对象映射函数 g 的单调减性，$g(A_2)\supseteq g(A_i)=U_i\supset U_1$。

定理 2[19] 如果格中结点 $H=(X,X')$ 有 d 个双亲结点 $M_1=(Y_1,Y_1')$，$M_2=(Y_2,Y_2')$，$\cdots$，$M_d=(Y_d,Y_d')$，则 H 产生规则前件的多个描述符，则对任意 $\forall p\in\{X'-(Y_1'\cup Y_2'\cup Y_3'\cup\cdots\cup Y_d')\}$，都存在一条无冗余规则 $p\Rightarrow X'-p$。

定义 7 当 $\theta\rightarrow\varphi$ 为一个分类决策规则，且 θ 和 φ 分别为 C 基本公式和 D 基本公式，$C,D\subseteq A$，则分类决策规则 $\theta\rightarrow\varphi$ 称为 CD 基本决策规则。

当 $\theta_1\rightarrow\varphi,\theta_2\rightarrow\varphi,\cdots,\theta_n\rightarrow\varphi$ 均为基本分类决策规则时，决策规则 $\theta_1\vee\theta_2\vee\cdots\vee\theta_n\rightarrow\varphi$ 称为基本决策规则 $\theta_1\rightarrow\varphi,\theta_2\rightarrow\varphi,\cdots,\theta_n\rightarrow\varphi$ 的组合，简称为组合分类规则。

3 CACLR 分类集成学习模型研究

3.1 CACLR 集成学习分类模型设计思想

集成学习分类是在给定的训练集上构造多个性能较好而又相对独立的基本分类器，再采用一定策略将各个分类器融合对未知样本进行训练学习，从而得到新的判定结果。本文设计的 CACLR 集成学习分类模型克服了利用单个概念格分类模型的缺点，具有较好的分类精度。其基本思想如下：

首先对原测试样本集进行数据预处理，消除噪音和无关数据，其次利用粗糙集对原概念格进行上下近似，从而得到近似概念格并进行格的约简，动态消除大量与分类知识无关的数据节点缩减原格规模；然后在近似域中利用粗糙度划分多个样本空间，构建多个相对独立且比较精确的近似概念格基本分类器，分别进行所在空间域的样本分类学习；再次通过特征重组和格空间优化，利用集成学习中典型的 AdaBoost. MH 算法进行多个同类分类器集成，在符合一定的评价标准下融合，得到分类挖掘集成学习模型。此时的模型就是本文提出

的 CACLR 模型，最后利用该模型进行未知大规模样本的分类学习和预测，具体如图 1 所示。

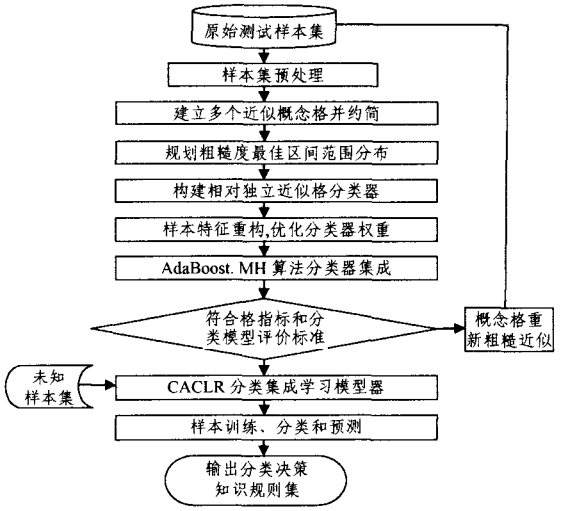

图 1　CACLR 集成学习模型流程图

3.2　CACLR 分类模型嵌入的核心算法设计

在 CACLR 集成学习分类模型中有两个核心的步骤：一是利用粗糙度区间选择合适样本空间分布，建立近似格建格；二是利用概念集成学习分类模型对样本进行分类和预测，得出分类决策知识规则集，下面的算法 1 和算法 2 进行了相关描述。

算法 1　粗糙度近似概念格建格

1. Input Concept Lattice L, Rough Thresholding $[\eta,\gamma]$
2. L constructs the corresponding formal context $L=(U,M,I)$
3. For each node $h\in L$
4. S=GenSubNodes(h)
5. For each node $h_c\in S$
6. If $h_c\notin L$
7. $L=L\cup\{h_c\}$
8. Add side $h_c\leftarrow h$ in concept lattice L
9. $S=S\cup\{h_c\}$
10. $\forall$ Rough thresholding subset $[\eta_1,\gamma_1]\subset[\eta,\gamma]$
11. For each$[\eta,\gamma]\in L(U,M,I)$
12. If $\gamma\cap V_d\neq\emptyset$ then
13. For each$((\eta,\gamma),(\eta_1,\gamma_1))\in L$
14. $S_i:=S\cup\{(\gamma-\gamma_1)-V_d\}$
15. NEXT
16. If $d_t\neq V_d\cap\gamma$ then
17. $S_i:=S-\{\gamma\rightarrow(d=d_t)\}$
18. $S\leftarrow\gamma(L(U,M,L)$
19. Output approximation concept lattice of rough entropy S

算法 2　CACLR 模型分类学习与知识提取

1. Approximation Concept Lattice of Rough Entropy S
2. Compute relative attribute reduction of approximation concept lattice S
3. Combine the same attribute reduction sets, $R=\emptyset$
4. For each attribute $a\in L$
5. Compute rough entropy $E(\{a\};D)$
6. $M\leftarrow\{a\in L|\min\{E(\{a\};D)\}\}$
7. Choose front attribute $a\in M$
8. Descend tiered operating with attribute a
9. Compute $U/\{a\}=\{U_1,U_2,\cdots,U_s\}$
10. Construct decision information system $S^*=(U_1,U_2,\cdots,U_s,L,D,V,f)$
11. For i=1 to s Do
12. {If $U_i\subseteq D_j\in U/D$
13. Then classification decision rulesets $(ai,V_{ai})\wedge\cdots\wedge(ak,V_{ak})\rightarrow(D_j,V_d)$
((a_i,V_{ai}) is superior attribute and attribute-value pais;
$(a_k,V_{ak})$is self attribute and attribute-value pairs:
(D_j,V_d) is decision attribute an attribute-value pairs)
14. Else $R=R\cup\{U_i\}$ }
15. If $R\neq\emptyset$, $\forall U_i\in R$
16. Perform$L=L-\{a\}$
17. Construct tiered decision information system $S^*=(U_i,L,D,V,f)$
18. Compute approximate importance and dependence degree
And gain classification decision rulesets R^*
19. Output classification decision rulesets R^*

4　模型实验分析与验证

4.1　UCI 数据集上分类精度实验

实验环境为 Inter Pentium 3. 06GHz CPU，内存 1024MB，Microsof Window XP 操作系统，DBMS 为 ORACLE9i，用 Java 实现了 CACLR 集成学习模型和其中相关算法。测试数据为标准 UCI 机器学习数据库[20]中的数据集。图 2 和图 3 为选取的 Chess，Letter 数据集，在设定的粗糙度区间范围内进行 CACLR 模型的分类精度测试实验，数据量设定为 4000 条、8000 条、16000 条、32000 条和 64000 条等中规模数据。

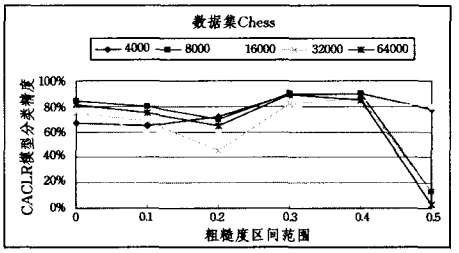

图 2　数据集 Chess 分类精度图

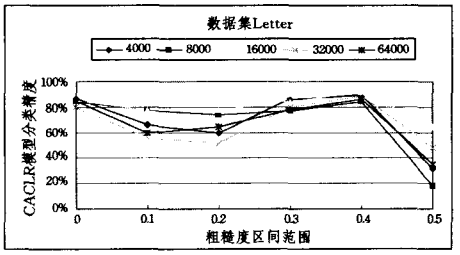

图 3　数据集 Letter 分类精度图

从以上粗糙分类精度图可以看出，数据集 Chess，Letter 在(0，0. 2)区间上曲线呈明显的下降趋势，数据的分类精度呈现不平稳性；在(0. 2，0. 3)区间，曲线有明显上升趋势，特别在区间(0. 3，0. 4)达到最高值，而在(0. 4，0. 5)区间呈明显下降趋势。这说明当粗糙度区间取值过小时，格结构由于完全反映知识的特征，造成 CACLR 模型分类精度过度拟合，影响分类挖掘的准确率；反之当粗糙度过大(0. 4，0. 5)，则数据中有用信息被忽略，造成知识过度丢失，分类准确率急剧下降。上

述实验表明当粗糙度在区间(0.3,0.4)取值,CACLR 模型能较好地消除噪音对所建概念格结构的影响,提高数据空间域的样本分类精度。

4.2 近似概念格与原概念格相关性能对比实验

概念格虽然是一种进行数据分类学习的有效工具,然而在处理大规模数据量时,建格规模庞大会影响分类效率和精度。在 CACLR 集成学习模型,我们利用粗糙集对原概念格进行上下近似得到近似概念格,在近似域中利用粗糙度划分多个样本空间,构建多个相对独立且比较精确的近似概念格基本分类器,动态消除大量与分类知识无关的数据结点,缩减原格规模,提高了分类挖掘的效率。我们设粗糙度在区间(0.3,0.4),数据规模为 64000,进行单个独立基近似格与原概念格在格节点、建格时间、分类学习时间以及测试和训练精度的比较,结果如表 1 所列。

表 1 单个独立基近似格与原概念格相关性能比较

Dataset	格结点数		建格时间(s)		分类学习时间(s)		测试集精度(%)(表示为均值±标准差)		训练集精度(%)(表示为均值±标准差)	
	原概念格	近似格	原概念格	近似格	原概念格	近似格	原概念格	近似格	原概念格	近似格
Breast-Cancer	12560	8603	343	230	215	178	83.42	85.52	75.07	79.78
Chess	7890	3670	162	187	167	169	75.56	77.78	80.08	83.79
Letter	5603	4930	165	124	143	112	81.58	86.49	72.79	85.37
Bypa	8756	6201	209	189	178	127	79.56	81.34	73.89	77.54
Soybean	4978	4209	179	168	137	121	79.90	81.23	81.67	82.08
Tic-Tac-Toe	8960	7120	434	269	398	278	67.12	87.32	60.74	84.46

4.3 CACLR 与典型分类模型对比实验

分类学习根据给定的样本及其类标号设计出分类判别函数,从而能对新样本的类别做出正确的预测。目前典型分类算法有 SMO,MIND 和 RBFN 算法等,将这 3 种算法皆采用 Java 语言实现,设计多个独立基分类器,同样采用 AdaBoost.MH 算法进行多个同类分类器集成。当数据规模为 16000 且在每个数据集上都进行 20 次实验取平均值,每种模型的集成学习分类精度如表 2 所列。

表 2 CACLR 模型与典型算法分类精度(%)比较

Dataset	SMO	MIND	RBFN	CACLR
Vote	75.43	79.69	80.23	83.72
Zoo	82.67	79.50	76.45	81.63
Australian	77.23	66.98	69.59	79.18
Ionosphere	82.40	84.08	78.09	86.42
Hypothyroid	80.89	79.12	77.56	82.03

从表 2 可以看出 CACLR 分类模型具有明显的优势,虽然在数据集 Zoo 上分类精度不是最佳,但基本接近 SMO 的最佳值。

在理论分析和实验结果的基础上,权衡建格效率和分类学习的准确率,我们建议在粗糙度区间[0.35,0.4]之间选取粗糙度区间范围,构建相对独立近似概念格基本分类器,然后进行特征重组和格空间优化,建立 CACLR 分类集成学习模型,进行样本空间域的动态分类学习和预测。

结束语 本文将粗糙度等理论应用到概念格分类学习问题研究以弥补概念分类学习过程中的不足,提出一种新型的基于粗糙度的近似概念格分类挖掘集成学习 CACLR 模型,该模型在粗糙度区间范围构建多个相对独立的近似概念格分类器,动态消除大量与分类知识无关的数据结点,有效缩减原格规模;通过特征重组和格空间优化,采用 AdaBoost.MH 算法进行多个同类分类器集成融合得到集成学习模型。通过 UCI 数据集上不同比较实验均表明:CACLR 分类精度较高、误差较小。与其它算法构建的分类器相比,CACLR 模型具有相当或更好的分类效果。

本文采用的独立基分类器为同类分类器。若要进一步提高分类效率,就要考虑分类器差异互补的特点,尽量使用多种不同类型的分类器和较好的分类集成算法。由于过度拟合数据产生的原因非常复杂,如噪音数据影响、分类模型的过度复杂性等,如何采用分类集成学习器更好地对过度拟合数据的概念格进行构造、分类和学习,是以后需要研究的工作。

当然,如何将 CACLR 模型应用到与文献[21,22]所描述的类似的电子病历系统中,进行大规模的临床医学病历的分类、预测和智能诊断,也是作者目前正在积极努力探索的领域。

参 考 文 献

[1] Joshi M, Karypis G, Kumar M. ScalParC: a new scalable and efficient parallel classification algorithm for mining large dataset s [A]// Int'l Parallel Processing Symposium[C]. Orlando, Florida, USA, April 1998: 573-579

[2] Haykin S. Neural Networks: A Comprehensive Foundation[M]. New Jersey: Prentice Hall, 1999

[3] Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition[J]. Data Mining and Knowledge Discovery, 1998: 121-167

[4] Dietterich T G. Ensembles in Machine Learning. Multiple Classifier Systems[J]. Lecture Notes in Computer Science, 2000: 1-15

[5] Wille R. Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts[M]// I. Rival, eds. Ordered Sets Reidel, Dordrecht, 1982: 445-470

[6] Pawlak Z. Rough sets: theoretical aspects of reasoning about data[M]. Dordrecht: Kluwer Acasemic Publisher, 1991

[7] Bordat J P. Calcul pratique du treillis de galoisd' unecorrespondence[J]. Mathematiqueset Sciences, 1986, 96: 31-47

[8] Chein M. Algorithm de recherch des sous-matrices premieres d' unevatrice[J]. Bull. Math. Soc. Sci. R. S. Roumanie, 1969, 3 (61): 1-25

[9] Ganter B, Wille R. Formal concept analysis: mathematical foundations[M]. Berlin: Springer, 1999

[10] Godin R, Mili H, Mineau G W, et al. Design of class hierarchies based on concept (Galois) lattices[J]. Theory and application of object systems, 1997, 4(2): 117-134

[11] Schapire R, Singer Y. Improved boosting algorithms using confidence rated predict ions[J]. MachineLearning, 1999, 37(3): 297-336

(下转第 232 页)

2。这是因为:使用 Tree-to-String 方法获取翻译等价对时,由于不受目标语句法分析精度及双语语法体系一致性的制约,因此等价对中包含的噪声较小,进而从中获取的翻译模板质量较高。

从表 3 中可以发现,实验 3 的译文评测分数接近实验 1。这是由于模板自动获取要受到词对齐工具和汉语句法分析器精度的制约,因而自动获取的模板其质量还没有达到手工书写的。但从模板获取的时间和人工劳动强度而言是可取的,同时随着词对齐工具和汉语句法分析器性能的逐步提升,以及翻译等价对、翻译模板优化方法研究的深入,自动获取的模板质量会进一步提高。

从图 5 中可以发现,实验 3 使用的模板数目要远超过实验 1,这是由于利用机器学习方法还不能获取像语言工程师所编写的具有高度抽象概括的翻译模板。但从模板获取的代价和译文质量提升方面来讲,本文所提出的自动模板获取方法是可取的。

为了比较手工书写的模板与自动获取模板之间的一致性,在实验 4 中,将实验 3 所获取的 12361 条模板加入系统原始模板库中,去掉其中冲突和冗余的部分,并对开放测试语料进行翻译。其译文评测分数如表 3 所列,实验 4 使用的模板数目如图 5 所示。

实验 4 与实验 1 相比,开放测试语料的 3 元 Nist 评测分数提高了 0.1365,上升了 3.41%。这表明:利用本文的方法所获取的模板与手工书写模板之间具有较好的一致性。

结束语 本文使用 Tree-to-String 方法从汉、英双语句对中抽取翻译等价对,利用错误驱动的机器学习方法获取翻译模板。将自动获取的模板用于 MTS2005,并使用 1200 句旅游领域的测试语料进行开放测试。实验结果表明:本文所提的模板获取方法其性能要好于传统方法,当新获取的模板加入系统原始模板库后,表现出了较好的一致性。

参 考 文 献

[1] Imamura K, Okuma H. Example-based Machine Translation Based on Syntactic Transfer with Statistical Models[C]// The 20th International Conference on Computational Linguistics. 2004:99-105

[2] Zettlemoyer L S, Moore R C. Selective Phrase Pair Extraction for Improved Statistical Machine Translation[C]// Proceedings of NAACL HLT 2007. 2007:209-212

[3] Meyers A, Yangarber R. Deriving Transfer Rules from Dominance-Preserving Alignments[C]// The 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics. 1998:843-847

[4] 屈刚,陈笑蓉,陆汝占. 基于有效句型的英汉双语短语对齐[J]. 计算机研究与发展,2003,40(2):143-149

[5] Wong Fai, Hu Dong-cheng. A Flexible Example Annotation Schema: Translation Corresponding Tree Representation[C]// The 20th International Conference on Computational Linguistics. 2004:1079-1085

[6] 张春祥. 基于短语评价的翻译知识自动获取研究[D]. 哈尔滨:哈尔滨工业大学,2007

[7] 吕雅娟,赵铁军,李生,等. 统计和词典方法相结合的双语语料库词对齐[C]// 第六届计算语言学联合学术会议. 2001:108-115

[8] Xue Yongzeng, Zhao Tiejun. Research On Sports News Oriented Skeleton Machine Translation[C]// The 7th International Conference for Young Computer Scientists. 2003:310-312

[9] Doddington G. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics[C]// Proceedings of ARPA Workshop on Human Language Technology. 2002

[10] Papineni K, Roukos S. BLEU: a method for automatic evaluation of machine translation[C]// Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311-318

(上接第 178 页)

[12] Breiman L. Bagging predictor [J]. MachineLearning, 1996, 26 (1):5-24

[13] 谢志鹏,刘宗田. 概念格节点的内涵缩减及其计算[J]. 计算机工程,2001,27(3):9-10

[14] 齐红,刘大有,胡成全,等. 基于搜索空间划分的概念生成算法[J]. 软件学报,2005,16(12):2029-2035

[15] 李云,刘宗田,陈崚,等. 多概念格的横向合并算法[J]. 电子学报,2004,32(11):1849-1854

[16] 刘贵龙. 粗糙集的粗糙度[J]. 计算机科学,2004,31(3):140-141,153

[17] 王玲芝. 粗糙模糊集的贴近度[J]. 四川师范大学学报:自然科学版,2002,25(5):476-478

[18] 丁卫平,管致锦,石振国. 基于扩展粗糙集模型近似概念格规则挖掘研究[J]. 南京邮电大学学报:自然科学版,2009,29(2):10-15

[19] 王志海,胡可云,胡学钢,等. 概念格上规则提取的一般算法与渐进式算法[J]. 计算机学报,1999,22(11):66-70

[20] UCI 机器学习数据库. ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[21] 丁卫平. 关联规则挖掘 Apriori 算法的改进及其应用研究[J]. 南通大学学报:自然科学版,2008,7(1):50-53

[22] 丁卫平,顾春华,石振国,等. 基于形式概念分析的不完备电子病历系统粗糙挖掘研究[J]. 计算机科学,2009,36(10):230-233