文章编号：1008-0562(2010)05-0970-04

基于约束的模糊分类算法改进研究

令狐大智[1]，武新丽[2]，王东红[3]
（1.广西大学 商学院，广西 南宁 530004；2.广西大学 行健文理学院，广西 南宁 530005；
3.广西财经学院 工商管理系，广西 南宁 530003）

摘　要：针对基于约束的模糊数据归类算法的不足，构建模式自学习调节模型，并提出基于约束规则的模式微调算法 CIP、基于自学习的分类规则优化更新算法 OCRS。新算法基于模式间相关性、距离和支持度等因素，建立模式自主更新标准和算法协调机制。实验研究表明：新算法在准确度相同的情况下，增加了算法的识别率、自学习能力和鲁棒性。

关键词：数据挖掘；自学习；模糊集；聚类分析；模式优化
中图分类号：TP311；TP393　　**文献标识码**：A

An improved restraint-based fuzzy classification algorithm

LINGHU Dazhi[1]，WU Xinli[2]，WANG Donghong[3]
(1.School of Business, Guangxi University, Nanning 530004, China; 2. School of Xingjian Art & Science, Guangxi University, Nanning 530005, China; 3. Department of Business & Administration, Guangxi College of Finance and Economics, Nanning 530003, China)

Abstract：In order to overcome the shortcoings in restraint-based fuzzy data classification algorithm, a model of pattern self-learning is developed in this study. Also, a constraint-based inching pattern algorithm (CIP) and optimized classification rule algorithm based on self-learning (OCRS) are proposaed. The new algorithms incorporate the standard of pattern self-update and the coordination mechanism which are based on various factors, such as models correlation, distance and support etc. An experimental study shows that the new algorithms enhance the ability of self-learning and the robustness.

Key words: data mining; self-learning; fuzzy; clustering; optimized pattern

0　引　言

模糊分类分析作为当前非常重要的研究和应用课题[1-9]。模糊分类分析的关键是检测准确度、对新数据的识别能力和一定的泛化能力。许多学者对模糊分类器的泛化能力进行了研究[1-9]，从拓扑角度分，目前常见的有三种模糊分类器结构：超矩形、多面体型和椭圆型。文献[2,4]分别从矩形、椭圆形拓扑角度对模糊分类进行了研究。文献[3,7]提出动态生成簇方法自主调整分类器；文献[8]提出在动态聚类过程中增加考虑误分目的地实现分类器自调整，使分类器具有了一定的自学习和自适应能力，但由于建立的基础使其拥有较大的模式库规模并限制了它们的检测效率。文献[1,5,6]从图形之外考虑，分别从遗传算法、神经网络和实体化约束的角度给出了模糊分类算法，增强了算法的普适性，降低了模式库的规模。文献[1,5]中的方法存在对训练集数据要求高、训练时间长、处理速度慢、结果不易解读的问题；文献[6]中方法虽较好地缓解了这些问题，但对训练集数据敏感、缺乏自学习和自适应能力。

本文通过改进文献[6]中算法，建立模式自学习调节模型。在研究中，本文借鉴了包分类和聚类分析方面的一些方法和系统思想[9-11]。

1　基于约束规则的模式微调

EFCBA 方法希望利用多点间的约束实现新数据的有效分类[6]。在研究中我们发现，当训练数据集样本数量有限、模式代表性不足或者出现特殊数据时，分类器的识别率就会降低。针对这些问题，本文提出基于约束规则的分类器微调方法。

收稿日期：2010-06-15
基金项目：广西自然科学基金资助项目（0640026）；广西青年科技基金资助项目（桂科青 0991097）；广西教育厅基金资助项目（200705LX300）
作者简介：令狐大智（1979-），男，山西 运城人，副教授，主要从事模式识别、信息安全方向的研究。本文编校：焦　丽

1.1　相关概念

定义 1　相似程度二元组{step,simulation}：指某个记录与给定模式集相似的程度，利用相似等级step[6]和等级相似程度 simulation[6]综合表示。

定义 2　特殊数据：即特殊类型数据，指其经分类后得到的相似程度二元组{step,simulation}中最大相似等级及其等级相似度都较低的数据。

定义 3　中心支持度 $T(R)$：对于 A 类数据集，经 R_a 检测到的数据量与 A 类集中总数据量的比值。

$$T(\mathrm{R}) = \mathrm{COUNT\_DETEC(A)/COUNT(A)} \quad (2)$$

R_a 为代表数据五元组[6]；$T(A \in B)$，表示用 B 的聚类中心测算 A 类数据，其结果为真的比例。

1.2　基于约束规则的模式微调

基于约束规则的模式微调 CIP（Constraint-based Inching Pattern Algorithm）是以虚拟数字实体化观点[6]、五元组约束规则[6]为基础，利用相似程度二元组和特殊数据实现对五元组微调。算法分结果判断和模式微调两部分。

分类结果判断是针对相似程度二元组中相似等级和等级相似程度，使用基于弹性理论的数据归类判定算法 EDCA[6]计算出特殊数据及其所属类别。模式微调是按照新加入点对五元组的影响来修改五元组中的 A、B、C 三点或者$|AD|$、$|AB|$、$|BD|$、$|CA|$、$|CB|$、$|CD|$6 个差异度参数。

算法　基于约束规则的模式微调方法 CIP

//利用模式和特殊数据微调分类器

//输入：特殊类型数据 F、相应模式库模式 Ru

//输出：调整后的模式 Ru'（五元组）

(1)获取特殊类型数据 F

(2)计算 F 与五元组中各中心记录的距离，记为$|DF|$、$|AF|$、$|BF|$、$|CF||$，同时获取类别信息

(3)模式微调

①若$|DF|>|DA|$，则更新 A 点；否则，进入 B；②若$|AF|+|DF|>|AB|+|BD|$，则更新 B 点；否则，进入 C；③若$|AF|+|BF|+|DF|>|AC|+|CD|+|BC|$，则更新 C 点，否则，进入 D；④当运行完当前调整后，再次运行 A～C 步，直至判断结果都为否，则进入 4）

(4)更新模式库

2　基于自学习的分类规则优化更新

在样本数据分布较复杂的情况下，异类样本所形成的集合图形往往会出现重叠[3]、图形上存在较大拐点或者发现传统的两类数据在概念上可以泛化为同类等情况，因此类仅用单一 EFCBA[6]模式描述的代表性较差。

针对重叠问题，文献[3]通过设定初始模糊模式库、误分数量阈值，将误分数据当作一个新的子类，并生成新的模式，然后在不改变原有模式的情况下调整模式，直至识别率达到最高。我们研究发现当某类的描述模式过于一般化时，不仅产生前述问题，还会发生数据误分多样化的问题。如针对图 1，会产生如下误分情况：①B 类数据被误分入 A 类数据；②A 类数据中包含 B 类数据；③A 类数据被误分入 B 类；④B 类数据中包含 A 类数据；⑤A 类数据被误分入 C 类；⑥C 类数据中包含 A 类数据；⑦C 类数据被误分入 A 类；⑧A 类数据中包含 C 类数据。

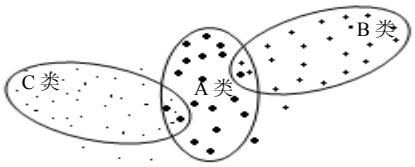

图 1　二维平面中的三类样本

Fig.1 sample of three classes in 2-D space

根据上述情况，本文提出基于自学习的分类规则优化更新算法 OCRS(Optimized Classification Rule Algorithm Base on Self-learning)，从误分源类、误分目标类及误分程度三方面寻找误分原因，并采用如下综合解决方案：一是采用文献[3]的方法，从误分源着手将被误分数据当作一类；二是将样本误分目标类分成两类；三是将公共部分合成一类，从原有类中剥离；四是两类数据合为一类。

2.1　相关概念

定义 4　模式相似度 RD_{ab}：指 A、B 两个模式间的相似程度，用模式所识别的数据比例来表示。

$$R_{ab} = MIN\left(\frac{A_b}{A}, \frac{B_a}{B}\right), \forall a, b \in (1 \cdots n), a \neq b \quad (3)$$

其中，A、B 为两类数据各自的总数；A_b 表示本为 B 类数据但被识别为 A 类的数据数目；B_a 表示本为 A 类但被识别为 B 类的数目；　n 为模式数。

定义 5　数据集模式相似程度 Rd：用建立的模式库检测数据集中的每类，判定其类别，若两者可以合并，则将其相似性记为 Rd。

2.2 基于自学习的分类规则优化更新算法描述

模式的更新过程就是模式比较、合并、简化过程。本文从数据集整体出发，综合考虑支持度、相似度、模式相似度、类别因素等指标后进行模式改进，是一种专家指导、逐步求精、自主学习的方法。

算法 2 首先进行阈值λi 求解，λi 是模式合并的比较基础，它通过定义 5 获取；然后利用λi 判定两模式是否能够进行合并，进入模式改进部分。

算法 2：基于自学习的分类规则优化更新算法

```
//输入：模式集 R、数据集 Mu、阈值λi
//输出：模式是否进行了调整
bool change_pattern(R,Mu, λi)
{ //获取λi，若不存在，先进行λi 计算
   Bool flag=false;
   R1=copy(R);//复制模式集
   If(λi==0)
     λi=get_namuna(R,Mu);
   if(λi==-1)
     return flag;//退出本次模式改进
   while(R.next!=null)
   {//针对模式库进行合并可能性判定，利用定义 5 计算两模式相似度 Rdab，当 Rdab>λi 时将它们合并
      while(R1.next!=null)
         { if(R.next!=R1.next)
         {//计算两模式间的相似度
        min=Rdab(R.next,R1.next);
//对两个要合并的模式查找原属聚集，将两个聚集合并到一起，调用算法 EFCBA 重新生成该模式，并将新的模式命名为原来两个模式名的合名
            if(min>λi)
              {//合并类别，构建新模式，更新模式库
              Change_class(R.next,R1.next);
              EFCBA(R.next,R1.next);
              del_pattern(R.next,R1.next);
                  flag=true;} }
       R1=R1.next; }
      R=R.next;// 直至所有模式都参与比较
       }Return flag;}
```

3 模式自学习调节模型

在实际运行过程中，首先使用 EDCA[6]算法将数据进行归类检测得到相似程度集，然后根据相似程度，将数据分为已知、未知、特殊三类后再进行处理。对已知类型数据按相关响应方法处理；未知类型数据进行累积，按一定策略经由模糊聚类算法进行类别确定，考虑新模式的生成和调优；对特殊数据，依据检测结果进行相关模式的调整更新。

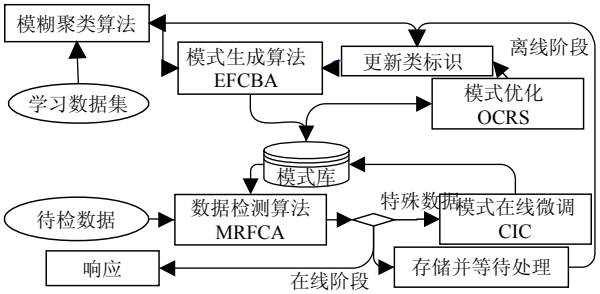

图 2 模式自学习调节模型

Fig.2 Pattern optimization model based self-learning

4 算法实验分析

本文采用 KDD98[13]数据集，结合文献 6 所用方法，在 Intel P8700 2.53G CPU、2G 内存的实验平台下进行对比测试。实验中构造三组数据集：

Test1 用于分类器微调测试，由 1000 行带分类标识 back 类记录组成；Test2 包含 100 行不带分类标识的 back 类新记录；Test3 由 back、normal、guest 三类数据组成，各 1000 行，共 3000 行记录。

4.1 基于约束规则的分类器微调算法测试

使用 Test1 数据建立分类模式，同时使用该模式检测 Test1，进行模式的微调，建立新模式，用两个模式分别检测 Test1 和 Test2。

表 1 模式微调检测结果对比表

Tab.1 result tables of model tuning

模式	检测对象	识别率/%	误检数	特殊	用时/μs
文献 6	Test1	100	0	0	10
	Test2	96	4	2	2
CIP	Test1	100	0	0	9
	Test2	99	1	0	2

从表 1 知进行微调后，模式检测效果明显提升。

4.2 基于自学习的分类规则优化更新测试

（1）数据集及模式变化

使用 test3 数据集中各类数据的 70 %构造模式，针对所有数据进行数据归类分析。

从表 2 观察，back、normal、guest 三类数据集

间彼此有部分数据被检入其它类，在使用其它类模式做检测时发现弱相关性。根据优化算法，guest 类型数据集中有 100 条记录应被拆分入 back 类数据集，同时还有 100 条记录在检测后应被独立分为一类；在 back、guest 与 normal 的对比中，有大约 5‰ 的数据被归入 normal 类型，根据检测分析对这部分数据不做处理。因此，back 类、guest 类的聚类中心被调整；增加一个新类；模式数目达到 4 组。

（2）模式调整前后检测结果对比

从表 3 可知，优化后的模式代表性更强，识别率高于优化前，对特殊样本的识别能力也大大增强，检测时间有所减少。同时，由于选取数据的缘由，模式规模略有增加。

表 2 数据集及模式变化检测情况表

Tab.2 result tables of data sets and model

类别标识	类名	数目	识别率/%	归 *A* 数目*	归 *B* 数目*	归 *C* 数目*	*A* 检测率/%	*B* 检测率/%	*C* 检测率/%
A	back	1000	99.1	991	1	8	99.1	0.1	0.2
B	normal	1000	99	1	990	9	0.2	99	0.8
C	guest	1000	98.4	26	5	984	1.2	0.2	98.4

*数据由于存在被分入多类的情况，因此条目数统计有重复。

表 3 模式优化后检测结果对比表

Tab.3 result tables of data sets and model

模式	检测对象	数据量	识别率/%	误检数	特殊样本数	检测时间/μs
文献 6 方法	back	1000	99.1	9	2	10
	normal	1000	99	10	3	11
	guest	1000	98.4	200*	1	8
本文方法	back(n)	1100	99.3	8	0	12
	normal	1000	99	10	1	10
	guest(n)	800	100	0	0	4
	new	100	100	0	0	2

*在 guest 检测中部分数据被归为两个类，同时算入识别率和误检数。

5 结　论

本文以基于相互约束的模糊数据归类算法为基础，提出基于约束规则的模式微调算法 CIP 和基于自学习的分类规则优化更新算法 OCRS，搭建模式自学习调节模型。通过在数据检测阶段，引入 CIP 方法，缓解训练数据集样本数量有限、模式代表性不足及存在特殊数据时造成的分类器识别率低的问题；在分类模型生成阶段，引入 OCRS 方法，缓解集合图型重叠、拐点、类别概念上融合等问题。本文基于 KDD98 数据集进行了对比实验分析，结果表明方法在保证算法效率、模式规模的情况下增加了算法的识别率、自学习能力和鲁棒性。

参考文献：

[1] 李继东,张学杰.基于遗传算法的多维模糊分类器构造的研究[J],软件学报,2005,16(05),779-785.

[2] Cordon O,del Jesus M J.Genetic learning of fuzzy classification systems cooperating with fuzzy reasoning methods[J].International Journal of Intelligent Systems,1998,13(10 / 11):l025-l053.

[3] Simpson P K.Fuzzy min-max neural networks-part classification[J].IEEE Trans. Neual Networks.1992.3(5):776-782.

[4] Abe Shigeo.Training of a fuzzy classifier with ellipsoidal regions by dynamic cluster generation.[C].Second International Conference on Knowledge Based Intelligent Electronic System,1998:126-131.

[5] Uebele F,Abe S,Lan M S.A Neural Network-Based Fuzzy Classifier.IEEE Trans on Systems, M an, and Cybernetics,1999,25(2):353-361.

[6] 令狐大智,李陶深.新的聚类中心构造算法及类别判定方法[J],计算机工程与设计,2008,29(9):2320-2323.

[7] 阳爱民,胡运发.一种基于椭圆区域的进化式模糊分类系统[J].模式识别与人工智能,2005,18(6):698-707.

[8] 腾明贵,吴正龙,熊范纶.一种通过动态聚类训练椭圆形模糊分类器的方法[J].小型微型计算机系统,2004,25(11):1990-1994.

[9] 陈　兵,潘宇科.一种采用启发式分割点计算的包分类算法[J].电子与信息学报,2009,31(7):1594-1599.

[10] 程舒通,徐从富,但红卫.频繁模式聚类算法改进研究[J].计算机工程与应用,2008,44(1):162-164.

[11] 李洪兴,汪群,段钦治,等.工程模糊数学方法及应用[M].天津:天津科学技术出版社,1993.

[12] 令狐大智.基于 Fuzzy Cluster 的入侵检测引擎研究[D].广西大学,2006.

[13] KDD Cup 1998.http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html[EB/OL].