基于粗糙集的加权KNN数据分类算法

刘继宇[1] 王 强[1] 罗朝晖[1] 宋 浩[1] 张绿云[2]
(广西师范大学计算机科学与信息工程学院 桂林 541004)[1]
(河池学院计算机与信息工程学院 宜州 546300)[2]

摘 要 粗糙集是处理不精确、不确定性问题的基本方法之一。采用粗糙集理论与方法进行数据分析具有不必具备数据集的先验知识、不需人为设定参数等优点,因而它被广泛应用于模式识别与数据挖掘领域。针对粗糙集训练过程中从未遇到过的样本的分类问题进行了探讨,根据条件属性的重要性确定加权系数,采用加权KNN的方法来解决无法与决策规则精确匹配的样本分类问题,并与加权最小距离方法进行了对比实验;同时对其他一些现有的粗糙集值约简算法进行了分析与研究,提出了不同的观点。对UCI多个数据集的大量数据进行了实验,并与近期文献中的多种算法进行了性能对比,实验结果表明,提出的算法的总体效果优于其他算法。

关键词 粗糙集,加权KNN,加权最小距离,属性值约简
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.10.057

Weighted KNN Data Classification Algorithm Based on Rough Set

LIU Ji-yu[1] WANG Qiang[1] LUO Zhao-hui[1] SONG Hao[1] ZHANG Lv-yun[2]
(College of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China)[1]
(College of Computer and Information Engineering, Hechi University, Yizhou 546300, China)[2]

Abstract Rough set is one of the basic methods in dealing with the imprecise or indefinite problems. For its advantages that the priori knowledge about analyzing dataset isn't necessary and the parameters analysis needn't to be set artificially, rough set is widely used in pattern recognition and data mining fields. For rough set theory, a core problem is how to classify the sample which has never been met in the process of training. This problem was discussed in detail in this paper. According to the importance of the condition attributes, a weighted KNN algorithm was proposed to classify the samples which can't precisely match to decision rules, and the contrast test with the weighted minimum distance (WMD) method was made to show the efficiency of our algorithm. At the same time, the existing algorithms about the attribute value reduction in rough set were analyzed and another point of view was put forward. The experiments on several UCI data sets and comparison with various existing algorithms proposed recently show that our algorithm is superior to these algorithms in overall effect.

Keywords Rough set, Weighted KNN, Weighted minimal distance, Attribute value reduction

1 引言

粗糙集理论是由波兰数学家Pawlak Z[1]于1982年提出的。由于粗糙集理论具备处理不精确、不确定性问题的能力,采用粗糙集方法进行数据分析具有不需要有关数据集的先验知识和人为设定参数等优点,因此它被广泛应用于模式识别[2,3]、知识发现[4]、聚类[5]与数据挖掘领域,并取得了丰硕的成果。

粗糙集理论引入不可区分关系作为基础,并在此基础上定义了上、下近似集等概念,通过上下近似集可以方便地求出边界区域,从而有效地刻画出具有不精确、不确定性的数据对象。粗糙集基本方法是通过属性约简和属性值约简等步骤提取信息系统的决策规则,并通过这些决策规则对测试样本进行分类识别。在此过程中,测试样本与决策规则的匹配关系往往会出现以下3种基本情况:①测试样本与规则库当中的某一条规则匹配;②测试样本与规则库当中的多个规则匹配;③测试样本不能与规则库当中任何规则进行匹配。

安利平等[6]分别对上述3种情况进行了讨论;马峻等[7]分析了这3种情况,并主要对第②种情况进行了讨论。但他们对第③种情况提出的解决策略的最终效果都不理想。由于第①种情况不存在不确定性,不需要过多讨论。因此,本文首先采用最优匹配原则来解决第②种情况,与此同时,对现有的一些粗糙集值约简算法进行了分析与研究,并提出观点。其次,针对第③种情况进行了较为深入的探讨,根据训练样本条件属性的重要性确定加权系数,采用加权KNN的方法来解决测试样本无法与规则库中任何规则精确匹配的分类问题。

到稿日期:2014-10-20 返修日期:2015-01-28 本文受国家自然科学基金地区项目(61165009),国家自然科学基金(61365009)资助。
刘继宇(1989—),男,硕士生,主要研究方向为模式识别与计算机视觉;王 强(1952—),男,博士,教授,硕士生导师,主要研究方向为模式识别、人工智能、计算机视觉及其应用,E-mail:qwang@mailbox.gxnu.edu.cn(通信作者);罗朝晖(1987—),男,硕士生,主要研究方向为模式识别与计算机视觉;宋 浩(1988—),男,硕士,主要研究方向为模式识别和计算机视觉;张绿云(1987—),女,硕士,主要研究方向为图像识别算法。

本文第 2 节介绍粗糙集相关定义和算法，并提出利用最优匹配的原则解决上述第②种情况；第 3 节针对上述第③种情况，提出加权 KNN 算法；第 4 节介绍实验分析与结果；最后总结全文。

2 粗糙集相关定义及算法

本节首先列举粗糙集理论中与本文的改进算法相关的一些定义，其次着重介绍粗糙集属性约简和属性值约简算法。这些定义和算法在王国胤[8]、张文修[9]等编著的书籍中有详细的介绍，在此分述如下。

2.1 粗糙集相关定义

定义 1(知识表达系统) 一个知识表达系统形式上是一个四元组 $S=(U,A,V,f)$，其中 $U=\{x_1,x_2,\cdots,x_n\}$ 为样本集，即论域；A 是属性的非空有限集合，$A=C\cup D$，且 $C\cap D=\varnothing$，子集 C 和 D 分别称为条件属性集和决策属性集；$V=\bigcup_{a\in A}V_a$ V_a 是属性值的集合，其中 V_a 是属性 a 的值域；$f=U\times A\rightarrow V$ 是一个信息函数，它指定 U 中每一个样本 x 的属性值。

定义 2(不可分辨关系) 根据定义 1 中的知识表达系统，对于每一个属性子集 $B\subseteq A$，不可分辨二元关系可定义为：

$$IND(B)=\{(x,y)\mid(x,y)\in U^2,\forall b\in B\text{ 有}(b(x)=b(y))\} \tag{1}$$

显然，$IND(B)$ 是一个等价关系，且 $IND(B)=\bigcap_{b\in B}IND(\{b\})$。

定义 3(属性约简(绝对约简)) 设 U 为一个论域，P 和 Q 为定义在 U 上的两个等价关系簇，且属性集合 $Q\subseteq P$，如果 $IND(P)=IND(Q)$，并且 Q 是独立的，则称 Q 是 P 的一个绝对约简。

定义 4(区分矩阵) 在定义 1 的知识表达系统中，$a_i(x_i)$ 是样本 x_i 在属性 a_i 上的取值，$d(x_i)$ 是样本 x_i 的决策属性的取值。$M_D(i,j)$ 表示区分矩阵中第 i 行第 j 列的元素，则区分矩阵 $M_D(i,j)$ 的定义为：

$$M_D(i,j)=\begin{cases}\{a_k\mid a_k\in C\wedge a_k(x_i)\neq a_k(x_j)\}, & d(x_i)\neq d(x_j)\\ 0, & d(x_i)=d(x_j)\end{cases} \tag{2}$$

其中 $i,j=1,2,\cdots,n$，n 表示对象的个数。

定义 5(二进制区分矩阵) 给定信息系统 $S=\langle U,C\cup D,V,f\rangle$，其对应的二进制区分矩阵中的元素记为 $M[(x_i,x_j),a_k]$，则有：

$$M[(x_i,x_j),a_k]=\begin{cases}1, & a_k\in C,f(x_i,a_k)\neq f(x_j,a_k)\\ & \text{且 } f(x_i,d)\neq f(x_j,d)\\ 0, & \text{其他}\end{cases} \tag{3}$$

其中，$i,j=1,2,\cdots,n,k=1,2\cdots m$，$n$ 表示对象的个数，m 表示条件属性的个数。

2.2 属性约简算法

属性约简是数据规约的一种形式，在数据挖掘中一般作为一种数据的预处理。其作用是在保持数据集的不可分辨关系不变的前提下，删除不必要或者不相关的属性，从而减少数据冗余量，提高效率。本文提出的加权 KNN 算法是对属性约简之后的结果进行加权，且属性重要性为加权 KNN 算法的权值计算提供依据。

Hu X 等[10]在 Swiniarski R W 等[11]提出的区分矩阵的基础上，将信息系统中所有有关属性的区分信息都浓缩进一个矩阵中，利用区分矩阵来进行属性约简。Felix R 等[12]提出了二进制区分矩阵的观点，根据该矩阵可以判别出决策表是否一致，且相对于传统的区分矩阵可以减少至少一半的存储空间。本文在文献[13,14]的基础之上，增加了对不一致情况的考虑，即两个样本所有条件属性的值都相同，决策属性不同的情况。这种情况在二进制区分矩阵中对应某行全为 0，因此无法删除该行。本文属性约简算法的主要思想：首先根据信息系统决策表求取正域，判断是否存在边界区域，若出现边界区域则采用边界规则进行处理，再求取正域的二进制区分矩阵；若不存在边界区域，可直接得到二进制区分矩阵。其次，在该矩阵中，将条件属性值为 1 的个数作为属性的重要性，并以此作为启发式的条件进行属性约简。

为了介绍属性约简算法的过程，本文通过一个实例说明如何由原信息表求出二进制区分矩阵，设表 1 为某信息系统的决策表。其中 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$ 表示论域，A_1,A_2,A_3,A_4,A_5 表示条件属性，D 表示决策属性。

表 1 信息系统决策表

U	A_1	A_2	A_3	A_4	A_5	D
x_1	0	0	0	0	0	0
x_2	1	1	0	1	0	0
x_3	1	1	0	2	0	2
x_4	2	1	0	2	0	2
x_5	2	2	0	1	0	1
x_6	2	2	0	0	0	1
x_7	2	0	0	1	0	1
x_8	2	0	0	0	0	1
x_9	0	1	1	0	0	1
x_{10}	0	1	0	0	1	1

对信息系统决策表求取正域，并通过式(3)计算正域得到二进制区分矩阵，如表 2 所列。

表 2 二进制区分矩阵

U′	A_1	A_2	A_3	A_4	A_5
(x_1,x_3)	1	1	0	1	0
(x_1,x_4)	1	1	0	1	0
(x_1,x_5)	1	1	0	1	0
(x_1,x_6)	1	1	0	0	0
(x_1,x_7)	1	0	0	1	0
(x_1,x_8)	1	0	0	0	0
(x_1,x_9)	0	1	1	0	0
(x_1,x_{10})	0	1	0	0	1
(x_2,x_3)	0	0	0	1	0
(x_2,x_4)	1	0	0	1	0
(x_2,x_5)	1	1	0	0	0
(x_2,x_6)	1	1	0	1	0
(x_2,x_7)	1	1	0	0	0
(x_2,x_8)	1	1	0	1	0
(x_2,x_9)	1	0	1	1	0
(x_2,x_{10})	1	0	0	1	1
(x_3,x_5)	1	1	0	1	0
(x_3,x_6)	1	1	0	1	0
(x_3,x_7)	1	1	0	1	0
(x_3,x_8)	1	1	0	1	0
(x_3,x_9)	1	0	1	1	0
(x_3,x_{10})	1	0	0	1	1
(x_4,x_5)	0	1	0	1	0
(x_4,x_6)	0	1	0	1	0
(x_4,x_7)	0	1	0	1	0
(x_4,x_8)	0	1	0	1	0
(x_4,x_9)	1	0	1	1	0
(x_4,x_{10})	1	0	0	1	1

表 2 中对应的行有且仅有一个 1 的行为核属性行，该行

中值为1的那个属性即为核属性。而属性重要性 I_i 可通过式(4)计算得到：

$$I_i=|V_i=1| \tag{4}$$

其中，V_i 表示属性 i 的值，| |表示个数。则 A_1 的属性重要性为 $I_1=21$，A_2 的属性重要性为 $I_2=18$，A_3 的属性重要性为 $I_3=4$，A_4 的属性重要性为 $I_4=22$，A_5 的属性重要性为 $I_5=4$。

提出的属性约简步骤为：

①创建一个空的属性集合 $red(A)$，用于保存约简后应保留的属性；并按本文方法求取信息系统的二进制区分矩阵。

②由于属性核是必不可少的属性，先将其放入应保留的属性集合 $red(A)$，并在二进制区分矩阵(见表2)中将核属性值为1的行去除。

③在剩下的信息表中，选择属性重要性最大的属性 A_i，即 $\max(I_i)$。并将选择的属性 A_i 放入 $red(A)$，在新的信息表中去除该属性值为1的行，重复执行本步骤，直到信息表的所有行都被删除为止。

④最后属性约简的结果为应保留的属性集合 $red(A)$。

表1通过属性约简之后的结果如表3所列。

表3　属性约简之后的决策表

U	A_1	A_2	A_4	D
x_1	0	0	0	0
x_2	1	1	1	0
x_3	1	1	2	2
x_4	2	1	2	2
x_5	2	2	1	1
x_6	2	2	0	1
x_7	2	0	1	1
x_8	2	0	0	1
x_9	0	1	0	1
x_{10}	0	1	0	1

2.3　值约简算法

粗糙集理论不仅能去除数据集中冗余的属性，还能从信息系统的决策表中提取知识规则。实际上，粗糙集的规则提取过程包括属性约简与属性值约简。本节针对属性值约简过程进行了分析与研究。基于粗糙集的属性值约简是删除决策规则的冗余属性值，使其最简化。

常犁云等[15]提出的值约简方法的主要思想是对属性约简之后的决策表进行逐列考察，去除某列后，判断是否存在冲突或者重复等情况，并分别对每一种情况进行处理，得到最终的化简结果；鄂旭等[16]首先通过对信息表进行属性约简，再对属性约简之后的信息表计算每条规则的核值和简化属性值集，最后导出规则；张利[17]首先通过约简得到决策表的属性核值，再利用条件属性与决策属性之间的互信息增量来度量属性值的重要度，随后再进行启发式值约简。

本文对常犁云等的值约简算法进行分析与研究，针对其中的步骤①提出不同的理解方式。原文值约简的步骤如下：

①对信息表中条件属性进行逐列考察，除去该列后，若产生冲突记录，则保留冲突记录的原属性值；若未产生冲突但含有重复记录，则将重复记录的该属性值标记为“*”；对于其他记录，将属性值标记为“?”。

②删除可能产生的重复记录，并考察每条含有标记“?”的记录。若仅由未标记的属性值即可判断出决策，则将“?”标记为“*”，否则，修改为原属性值；若某条记录的所有条件属性均被标记，则将标记有“?”的属性项修改为原属性值。

③删除所有条件属性均被标记为“*”的记录及可能产生的重复记录。

④如果两条记录仅有一个条件属性值不同，且其中一条记录该属性被标记为“*”，那么，该记录如果可由未标记的属性值判断出决策，则删除另外一条记录；否则，删除本记录。

原文步骤①逐列考察条件属性，即去除某属性列后，根据被标记前的原属性值考察剩下的条件属性集合对整个论域的区分能力。在未产生冲突但含有重复记录时，原文直接将重复记录的该属性值标记为“*”。然而，在这个过程中未考虑到已考察过且被标记为“*”的属性的影响，使得已丧失区分能力的属性成为考察当前属性的依据，从而在某些情况下导致错误与冲突的产生。

因此，本文对上述步骤①做出如下修改：按照原文方式采用未被标记的原信息表对条件属性进行逐列考察，除去该列后，若产生冲突记录，则保留冲突记录的原该属性值；若未产生冲突但含有重复记录，则逐一考察每个重复记录已考察过的属性中是否存在已经被标记为“*”的，若不存在，则将该重复记录的该属性标记为“*”；若存在，则利用未被标记为“*”的属性判断是否与其他的记录发生冲突，若产生冲突，则保留重复记录的该属性原值，若未产生冲突，则将重复记录的该属性值标记为“*”；对于其他记录，则将属性值标记为“?”。

本文测试过程是先将测试样本通过属性约简，再将属性约简之后的测试样本与规则库当中的规则进行逐一匹配。匹配的方法是判断测试样本与某规则所有未被标记的条件属性的值是否都对应相等，若均相等，则表明匹配成功；否则，不匹配。当匹配的规则存在多个，即引言中提及的第②种情况时，则采用最优匹配的原则进行匹配。具体地，在成功匹配的规则中，将测试样本与未被标记的条件属性个数最多的规则进行匹配，并将其判为该规则所属的类别。

为了充分说明问题，下面通过实例进行说明。将属性约简之后的信息表(见表3)根据原文中的理解方式进行值约简，得到表4。

表4　按原文中的理解方式进行值约简的结果

	A_1	A_2	A_4	D
1	0	0	*	0
2	1	*	1	0
3	*	*	2	2
4	2	*	*	1
5	0	1	*	1

将训练样本作为测试样本，采用最优匹配原则对规则库(见表4)进行验证。然而在这里却存在一个冲突现象，即当训练样本 x_4(见表5)充当测试样本时，按照上面产生的规则进行识别，将会产生冲突现象。

表5　训练样本充当测试样本

U	A_1	A_2	A_3	A_4	A_5	D
x_4	2	1	0	2	0	2

此时，与测试样本匹配的规则如表6所列。

表6　x_4 匹配的规则

	A_1	A_2	A_4	D
3	*	*	2	2
4	2	*	*	1

从表6可以看出冲突已经产生。

将属性约简之后的信息表(见表3)根据本文提出的修改方法进行值约简,得到表7。

表7 按本文提出的方法进行值约简的结果

	A_1	A_2	A_4	D
1	0	0	*	0
2	1	*	1	0
3	*	*	2	2
4	2	*	1	1
5	2	*	0	1
6	0	1	*	1

按照表7的规则库,训练样本充当测试样本时,采用最优匹配原则对规则库(见表7)进行测试验证,原训练数据集将不存在冲突现象。实验证明此方法能有效地解决可能存在的冲突问题。

3 加权KNN算法

传统的分类算法(如BP神经网络、支持向量机(SVM)、决策树(C4.5)等)通常分为3个阶段:第一,建立算法模型;第二,利用一部分的数据对算法模型进行训练;第三,用训练好的模型对一些未知的样本进行分类并评估模型的性能。然而传统的方法中,预测阶段往往对那些在训练过程中从未遇见过的样本不能进行很好的分类。因此提出的加权KNN[18]算法是针对粗糙集算法遇到规则库中没有任何规则与测试样本匹配,即引言中提及的第③种情况时进行识别分类的方法。本文提出的加权KNN算法具体实现过程如下。

①求属性的权值

通过属性的重要性,可以求出各属性的权值比重W_i,如式(5)所示:

$$W_i=\frac{I_i}{\sum_{i=1}^{n} I_i} \tag{5}$$

其中,n表示的属性的个数,I_i表示属性i的重要性。

②求测试样本与原训练集所有样本之间的加权距离

在测试的过程中,若测试样本与规则库中的规则都无法精确匹配,则表明该测试样本是从未遇见过的新的样本,在这种情况下,通过式(6)计算测试样本与训练集所有样本之间的欧氏距离并加权得到最终的距离。

$$d_i(x)=\sqrt{\sum_{j=1}^{n} W_j\times(x_j-V_{ij})^2} \tag{6}$$

其中,$i=1,2,\cdots,N$,$j=1,2,\cdots,n$,V_{ij}表示训练样本i的第j个条件属性的值,x_j表示测试样本x的第j个条件属性的值,W_j表示第j个属性的权值,N表示训练样本的个数,n表示经过属性约简之后的属性个数。

③根据式(6)得出与测试样本距离最小的K个训练样本并设置k值

通过式(6)求出测试样本与训练集中所有样本之间的加权距离d后,按照距离最小的原则选取该测试样本在训练样本中的k个近邻。

④统计

通过步骤③得到k个近邻后,统计各类别样本出现的次数,求出次数最多的类别,将测试样本的类别判为该类别。若不同类别样本在k个近邻中出现次数相同且均为最多,则回到步骤③继续求最小距离的训练样本,直到能统计得到出现训练样本次数最多的类别时才退出循环,并将测试样本判为该类别。

我们将此加权KNN算法分别应用于6个UCI数据集(详见实验数据的介绍)上,同时与加权最小距离算法进行对比实验,取得了较好的效果。

4 实验分析与结果

本节通过实验来验证提出的方法的性能,并采用Astudillo C S A等[19]的10-折交叉验证的方法来验证。实验最后给出了其他几种监督算法:贝叶斯网络(BN)、朴素贝叶斯(NB)、决策树(C4.5)、k-近邻(KNN)、学习矢量量化(LVQ)以及半监督算法:基于树的拓扑结构的自组织映射(TTOSOM15)、自组织映射(SOM)的结果。10-折交叉验证是将原始的数据集随机分成十等份,依次用每一份作为测试集,其他的剩下的9份作为训练集,从而得到10次交叉验证的结果,最终求10次交叉验证结果的平均值。本文采用同样的方法进行验证。

4.1 实验数据的介绍

实验数据是UCI数据库中的几个数据集,这几个数据集来自不同的领域。第一个数据集是非常有影响力的Iris集。第二个数据集是威斯康辛州诊断乳腺癌(WDBC)集,第三个数据集是wine数据集,其数据是来自种植在同一地区3个不同品种的意大利葡萄酒的化学分析的结果。第四个数据集是Yeast数据集,对于分类器来讲,它代表一个更具挑战性的环境,包含分为10个不同类别的1484个实例,目的是预测蛋白质的细胞定位的问题。第五个数据集为Wine_Quality数据集,目的是基于化学信息预测红酒的质量。第六个数据集是glass的数据集,它包含6个类型的玻璃(即钠、铁、钾等)。各数据集如表8所列。

表8 选择的数据集

Datasets	Instances	Attributes	Classes	Problem Type
Iris	150	4	3	Classification
WDBC	569	30	2	Classification
glass	214	9	6	Classification
wine	178	13	3	Classification
Wine_Quality	1599	11	6	Classification/ regression
Yeast	1484	8	10	Classification

4.2 对比实验与结果

为了能充分说明加权KNN算法对粗糙集训练过程中从未遇见过的样本的分类问题的有效性,本文将加权KNN算法与加权最小距离算法[20]以及近期文献中几个较为优秀的同类算法分别应用在UCI的6个数据集上,并进行实验对比。实验结果表明加权KNN算法具有更好的分类能力。

4.2.1 加权KNN与加权最小距离算法的实验比较

在选取的6个UCI数据集上分别采用加权KNN和加权最小距离的方法,通过10-折交叉验证得到实验结果,如图1—图3所示。其中图1是将加权KNN的方法与粗糙集相结合,并应用到6个UCI数据集中进行实验得到的结果;图2是用加权最小距离的方法替换加权KNN,并与粗糙集相结合应用到6个UCI数据集当中进行实验得到的结果。图3是加权KNN与加权最小距离算法进行比较得到的实验结果。注:图1—图3的横坐标的"次数"表示10-折交叉重复验证的第k次。

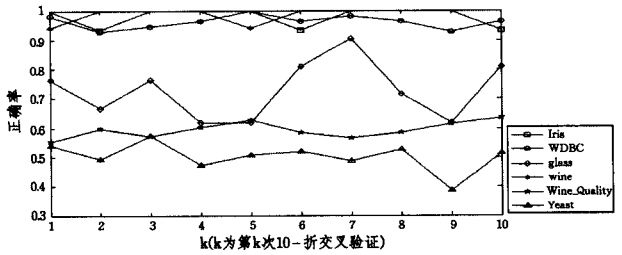

图 1　各数据集采用加权 KNN 算法的 10-折交叉验证的结果

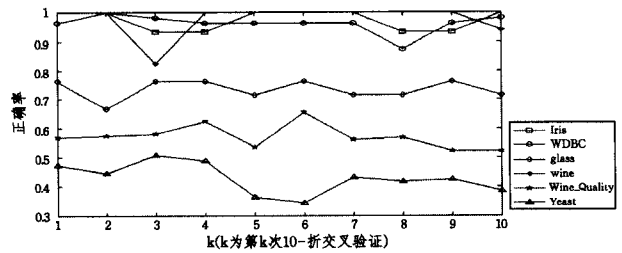

图 2　各数据集采用加权最小距离算法的 10-折交叉验证的结果

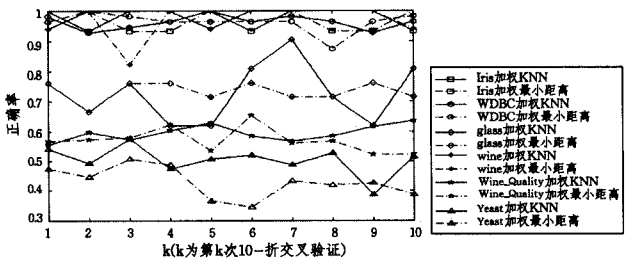

图 3　各数据采用加权 KNN 和加权最小距离算法的实验对比

通过将加权 KNN 和加权最小距离方法分别与粗糙集相结合，并应用于 6 个 UCI 数据集上，得到各数据集的平均识别率和正确率，如表 9、表 10 所列。

表 9　采用加权 KNN 算法的平均识别率和平均正确率

指标 \ 数据集	Iris	WDBC	glass	wine	Wine_Quality	Yeast
平均识别率	100%	100%	100%	100%	100%	100%
平均正确率	98.00%	96.25%	73.33%	98.82%	59.43%	50.20%

表 10　采用加权最小距离算法的平均识别率和平均正确率

指标 \ 数据集	Iris	WDBC	glass	wine	Wine_Quality	Yeast
平均识别率	100%	100%	100%	100%	100%	100%
平均正确率	97.33%	96.25%	73.33%	97.65%	56.98%	42.84%

实验结果表明，对于分类效果较好的数据集（如 Iris、WDBC、glass、wine），由于数据集的分布是椭圆形的，各类中心分布比较明确，因此加权 KNN 和加权最小距离都能很好地将数据进行分类，其实验效果相差不大。但对于比较有挑战性的数据集（如 Yeast、Wine_Quality），由于数据集的各类分布不是典型的椭圆形分布，此时加权最小距离并没有考虑到数据集的分布和局部特征，而加权 KNN 却能很好地表现出它的优势，其实验效果也表明加权 KNN 优于加权最小距离。

4.2.2　本文算法与几种较新的同类算法的实验比较

最后将本文方法与 Astudillo C S A 等[19]的基于树的拓扑结构的自组织映射（TTOSOM15）方法及其文中列举的半监督分类方法：自组织映射（SOM），监督分类算法：贝叶斯网络（BN）、朴素贝叶斯（NB）、决策树（C4.5）、k-近邻（KNN）、学习矢量量化（LVQ）等进行比较和分析，结果如表 11 所列。

表 11　本文的方法与文献中的方法的正确率对比（单位：%）

数据集 \ 方法	本文方法	TTOSOM15	BN	NB	C4.5	KNN	LVQ	SOM
Iris	98.00	96.67	92.67	96.00	96.00	95.33	96.00	84.67
WDBC	96.25	93.32	95.08	93.15	93.15	96.66	92.09	90.51
glass	73.33	67.29	71.96	49.07	67.76	67.76	61.22	63.08
wine	98.82	89.33	98.88	97.19	93.82	94.94	74.16	67.98
Wine_Quality	59.43	51.91	57.72	55.03	62.91	57.79	44.15	46.16
Yeast	50.20	54.18	56.74	57.61	55.86	54.78	24.33	49.59
平均百分比	79.34	75.45	78.84	74.68	78.25	77.87	65.32	67.00

表 11 表明，本文算法整体上优于其他各类算法，主要原因在于：①本文算法能针对测试样本在训练样本集中从未遇见过的情况，通过加权 KNN 有效地进行样本的分类。而 TTOSOM15、BN、NB、C4.5、KNN、LVQ、SOM 却不能很好地处理这种情况。②本文算法中，由于规则获取是建立在信息表的基础之上，通过启发式约简方式来获取规则，若测试样本是包含在训练样本集内的，则本文算法能够确保分类的正确性，相对于 TTOSOM15、BN、NB、KNN、SOM 分类方法具有较高的稳定性。

结束语　本文针对粗糙集训练过程中从未遇见过的样本的分类问题进行了探讨，并提出加权 KNN 的方法进行解决；为了说明加权 KNN 算法的有效性，通过与加权最小距离算法以及近期文献中较为先进的同类算法进行对比实验，结果表明加权 KNN 具有更好的实验效果。同时，对现有的粗糙集值约简算法进行了分析与研究，指出现有的值约简算法可能存在的错误和冲突，并给出改进措施。

参 考 文 献

[1] Pawlak Z. Rough sets: Theoretical aspects of reasoning about data[M]. Dordrecht & Boston: Kluwer Academic Publishers, 1991

[2] Theodoridis S, Koutroumbas K. 模式识别（第 2 版）[M]. 李晶皎，朱志良，王爱侠，等译. 北京：电子工业出版社，2004

[3] Mitra S. An Evolutionary Rough Partitive Clustering Pattern [J]. Recognition Letters, 2004, 25(12): 1439-1449

[4] Gibert K, Rodríguez-Silva G, Rodríguez-Roda I. Knowledge discovery with clustering based on rules by states: A water treatment application [J]. Environmental Modelling & Software, 2010, 25(6): 712-723

[5] Lai J Z C, Juan E Y T, Lai F J C. Rough clustering using generalized fuzzy clustering algorithm [J]. Pattern Recognition, 2013, 46(9): 2538-2547

[6] 安利平，陈增强，袁著祉. 基于粗集理论的多属性决策分析[J]. 控制与决策，2005，20(3)：294-298

An Li-ping, Chen Zeng-qiang, Yuan Zhu-zhi. Multi attribute decision analysis based on rough set theory [J]. Control and Decision, 2005, 20(3): 294-298

[7] 马峻，吉晓民. 利用粗糙集理论实现工艺决策的冲突消解[J]. 计算机辅助设计与图形学报，2005，17(3)：600-604
Ma Jun, Ji Xiao-min. Implementation of Conflict Resolution for Process Decision Based on Rough Theory [J]. Journal of Computer Aided Design & Computer Graphics, 2005, 17(3): 600-604

[8] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001
Wang Guo-yin. Rough set theory and knowledge acquisition [M]. Xi'an: Xi'an Jiaotong University Press, 2001

[9] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001
Zhang Wen-xiu, Wu Wei-zhi, Liang Ji-ye, et al. The rough set theory and method [M]. Beijing: Science Press, 2001

[10] Hu X, Cercone N. Learning in relational databases: a rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338

[11] Swiniarski R W, Skowron A. Rough set methods in feature selection and recognition[J]. Pattern Recognition Letters, 2003, 24(6): 833-849

[12] Felix R, Ushio T. Rough sets-based machine learning using a binary discernibility matrix[C] // Proceedings of the Second International Conference on Intelligent Processing and Manufacturing of Materials, 1999 (IPMM'99). IEEE, 1999: 299-305

[13] 杨萍，李济生，黄永宣. 一种基于二进制区分矩阵的属性约简算法[J]. 信息与控制，2009，38(1)：70-74
Yang Ping, Li Ji-sheng, Huang Yong-xuan. A attribute reduction algorithm based on binary Discernibility Matrix [J]. Information and Control, 2009, 38(1): 70-74

[14] 张颖淳，苏伯洪，曹娟. 基于粗糙集的属性约简在数据挖掘中的应用研究[J]. 计算机科学，2013，40(8)：223-226
Zhang Ying-chun, Su Bo-hong, Cao Juan. Study on application of Attributive Reduction Based on Rough set in Data mining [J]. Computer Science, 2013, 40(8): 223-226

[15] 常犁云，王国胤，吴渝. 一种基于 Rough Set 理论的属性约简及规则提取方法[J]. 软件学报，1999，10(11)：1206-1211
Chang Li-yun, Wang Guo-yin, Wu Yu. A Method of Attribute Reduction and Rule Extraction Based on Rough Set Theory[J]. Journal of Software, 1999, 10(11): 1206-1211

[16] 鄂旭，邵良杉，张毅智，等. 一种基于粗糙集理论的规则提取方法[J]. 计算机科学，2011，38(1)：232-235
E Xu, Shao Liang-shan, Zhang Yi-zhi, et al. Method of Rule Extraction Based on Rough Set Theory [J]. Computer Science, 2011, 38(1): 232-235

[17] 张利，卢秀颖，吴华玉，等. 基于粗糙集的启发式值约简的改进算法[J]. 仪器仪表学报，2009(1)：82-85
Zhang Li, Lu Xiu-ying, Wu Hua-yu, et al. Improved heuristic algorithm used in attribute value reduction of rough set [J]. Chinese Journal of Scientific Instrument, 2009(1): 82-85

[18] Suresh B V, Viswanath P. Rough-fuzzy weighted k-nearest leader classifier for large data sets[J]. Pattern Recognition, 2009, 42(9): 1719-1731

[19] Astudillo C S A, Oommen B J. On achieving semi-supervised pattern recognition by utilizing tree-based SOMs[J]. Pattern Recognition, 2013, 46(1): 293-304

[20] 任靖，李春平. 最小距离分类器的改进算法-加权最小距离分类器[J]. 计算机应用，2005，25(5)：992-994
Ren Jing, Li Chun-ping. Improved minimum distance classifier-weighted minimum distance classifier [J]. Computer Applications, 2005, 25(5): 992-994

(上接第 280 页)

[4] Ogilvie P, Callan J. Combining document representations for known-item search[C] // Proceedings of the 26th ACM SIGIR. Toronto, Canada, 2003: 143-150

[5] Kim J, Xue X, Croft W B. A Probabilistic Retrieval Model for Semistructured Data[C] // Proceedings of the 31th ECIR. Toulouse, France, 2009: 228-239

[6] Kim J, Croft W B. A Field Relevance Model for Structured Document Retrieval[C] // Proceedings of the 34th ECIR. Barcelona, Spain, 2012: 97-108

[7] Itakura K Y, Clarke C L. A framework for BM25F-based XML retrieval [C] // Proceedings of the 33rd ACM SIGIR. Geneva, Switzerland, 2010: 843-844

[8] 刘德喜，万常选，刘喜平，等. 基于结点权重模型的 XML 片段检索策略[J]. 计算机学报，2013，36(8)：1729-1744
Liu, De-xi, Wan Chang-xuan, Liu Xi-ping, et al. A Snipet Retrieval Strategy Based on Element Weighting Model[J]. Chinese Journal of Computers, 2013, 36(8): 1729-1744

[9] Yi X, Allan J, Croft W B. Matching resumes and jobs based on relevance models[C] // Proceedings of the 30th ACM SIGIR. Amsterdam, 2007: 809-810

[10] Zhao L, Callan J. Effective and Efficient Structured Retrieval [C] // Proceedings of the 18th ACM CIKM. Hong Kong, China, 2009: 1573-1576

[11] Blei D M, Ng A Y, Jordan M I. Latent Dirichletallocation[J]. Journal of Machine Learning Research, 2003, 3(4/5): 993-1022

[12] Yi X, Allan J. A Comparative Study of Utilizing Topic Models for Information Retrieval[C] // Proceedings of the 31th ECIR. Toulouse, France, 2009: 29-41

[13] Lavrenko V, Croft W B. Relevance-based language models[C] // Proceedings of the 24th ACM SIGIR. New Orleans, Louisiana, USA, 2001: 120-127

[14] Ganguly D, Leveling J, Jones G J F. An LDA-smoothed relevance model for document expansion: a case study for spoken document retrieval[C] // Proceedings of the 36th SIGIR. Dublin, Ireland, 2013: 1057-1060

[15] Bai J, Song D, Bruza P, et al. Query Expansion Using Term Relationships in Language Models for Information Retrieval[C] // Proceedings of the 14th CIKM. Bremen, Germany, 2005: 688-695

[16] Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation[C] // Proceedings of SIGMOD. Dallas, Texas, USA, 2000: 1-12

[17] Liang Y, Liu T, Ni W. Augmented Vector Space Model for Passage Intention Classification in Chinese Agricultural Prescription Documents[J]. Journal of Computational Information Systems, 2014, 10(1): 101-108

[18] Songa M, Song I-Y, Hu X, et al. Integration of association rules and ontologies for semantic query expansion[J]. Data & Knowledge Engineering, 2007, 63(1): 63-75