基于属性代表的多粒度集成分类算法

张清华[1),2)] 支学超[1),2)] 王国胤[1),2)] 杨 帆[3)] 薛付忠[3)]

[1)](旅游多源数据感知与决策技术文化和旅游部重点实验室 重庆 400065)

[2)](重庆邮电大学计算智能重庆市重点实验室 重庆 400065)

[3)](山东大学公共卫生学院 济南 250000)

摘 要 面对复杂多变的信息系统,传统的机器学习多分类模型无法实现一个动态分类的过程.序贯三支决策作为一种多粒度分类算法,常用于解决多粒度空间下动态分类问题.然而,序贯三支决策在粗粒度空间下容易产生决策冲突,在细粒度空间下要考虑很多属性导致其分类效率不高以及无法对最终未分类对象进行处理.因此,本文结合集成学习和粒计算的思想提出了一种基于属性代表的多粒度集成分类算法.首先,通过选择每一粒层中分类能力较强的属性作为属性代表来构建分类器,形成基于属性代表的集成分类器.其次,通过评分表保留粗粒度空间下分类器的分类意见以减少细粒度下需要考虑的属性个数.最后,采用"相对最优"的策略,将反对率最少的决策类作为最终未分类对象的分类结果.通过实验验证,本文方法相比于序贯三支决策以及其他机器学习的多分类算法具有较好的鲁棒性、分类效率以及分类性能.

关键词 动态分类;序贯三支决策;集成学习;属性代表;多粒度

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.01712

Multi-Granularity Ensemble Classification Algorithm Based on Attribute Representation

ZHANG Qing-Hua[1),2)] ZHI Xue-Chao[1),2)] WANG Guo-Yin[1),2)] YANG Fan[3)] XUE Fu-Zhong[3)]

[1)](*Key Laboratory of Tourism Multisource Data Perception and Decision, Ministry of Culture and Tourism, Chongqing* 400065)

[2)](*Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing* 400065)

[3)](*School of Public Health, Shandong University, Shandong* 25000)

Abstract In the face of the complex and changeable information systems, in the field of machine learning traditional multi-classification models cannot achieve a dynamic classification process, and it cannot solve some problems such as disease diagnosis. Because some diagnosis procedures are too expensive, it is necessary to judge whether the patient is likely to be ill through some preliminary diagnosis, thereby reducing the cost of the process. Sequential three-way decisions as a multi-granularity classification algorithm, which is used to solve dynamic classification problems in multi-granularity space. The sequential three-way decision model sorts attributes by balancing the cost of decision results and decision process, then a multi-level granularity space is constructed. With the injection of information in turn, objects that meet the conditions are classified at different granularity levels. It can be said that the sequential three-way decision model solves the problem of excessive costs for decision process. Therefore, many scholars at home and abroad have optimized the sequential three-way decision model from perspective of cost-sensitive. However,

收稿日期:2021-06-07;在线发布日期:2022-01-21.本课题得到国家重点研究发展计划(2020YFC2003502)、国家自然科学基金(61876201)、重庆市自然科学基金(cstc2019jcyj-cxttX0002)资助.**张清华**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为粗糙集、模糊集、粒计算和不确定性信息处理.E-mail:zhangqh@cqupt.edu.cn.**支学超**,硕士研究生,主要研究方向为粗糙集、集成学习和不确定性信息处理.**王国胤**,博士,教授,博士生导师,中国计算机学会(CCF)会士,长江学者,主要研究领域为粗糙集、粒计算、知识技术、数据挖掘、神经网络以及认知计算.**杨 帆**,博士,教授助理,主要研究方向为深度因果学习、贝叶斯因果推理、生物医学.**薛付忠**,博士,教授,博士生导师,中华预防医学会健康保险专业委员会主任委员,主要研究领域为深度因果学习、贝叶斯因果推理及生物医学.

in some cases, the sequential three-way decision model in the coarse granularity space is prone to decision conflicts, that is, the same object gets multiple different classification results. Therefore, many attributes must be considered in the fine granularity space, which leads to low classification efficiency. Because of the lack of more information and corresponding strategies, the sequential three-way decision model is unable to process the final unclassified objects. Therefore, this paper combines the ideas of ensemble learning and granular computing to propose a multi-granularity ensemble classification algorithm based on attribute representation. Firstly, constructing a classifier by selecting the attribute representatives of each granularity layer to form an ensemble classifier based on attribute representatives. By synthesizing the different opinions of the classifier which is constructed by attribute representation, the generation of decision conflicts in each granularity layer can be effectively reduced. Secondly, the classification opinions of classifiers in the coarse granularity space are retained through the scoring table to reduce the number of attributes that need to be considered in the fine granularity space. The retained score can make the ensemble classifier which is constructed by attribute representation in the fine granularity space to avoid more likely errors, thereby obtaining a more confident classification result. Finally, there may still be cases where some objects are not classified after all the information has been injected, so the "relatively optimal" strategy is adopted, and the decision class with the least objection rate is used as the final classification result of unclassified objects. In order to verify the validity of the model in this paper, the 14 UCI data sets and 6 real data sets which are related to medical diagnosis are used to conduct horizontal and vertical comparison experiments respectively. Among them, the horizontal comparison experiment includes ten popular multi-classification algorithms. Through experiments, the proposed method in this paper has better robustness, classification efficiency and classification performance than the sequential three-way decisions and other machine learning multi-classification algorithms. Moreover, the multi-granularity ensemble classification algorithm based on attribute representation has improved significantly in the real data sets of medical diagnosis.

Keywords　dynamic classification; sequential three-way decisions; ensemble learning; attribute representation; multi-granularity

1　引　言

近年来，随着信息量爆炸式增长，数据中所蕴含的知识信息越来越复杂多元化. 尤其在面对多分类问题时，往往需要利用机器学习算法[1]构建复杂的分类模型去应对不同的变化，其每一次的改进都需要耗费很大的代价. 因此面对复杂分类问题的求解时，人们可能会采用"分而治之"的思想来将复杂的分类问题转化为多个简单的二分类问题，分别使用单分类模型对多个简单的二分类问题进行求解，最终通过合理的策略将多个简单二分类问题的解融合得到多分类问题的最终解. 如今，基于集成学习[2]思想的多分类集成算法[3-4]正是解决这类问题的典型算法，其经常被应用于金融、医疗等任务[5-9].

Yao[10]基于概率粗糙集理论提出三支决策(Three-way Decisions，3WD)，其结合"分而治之"、"三分而治"的思想，用于处理复杂动态的不确定性问题. 近些年，三支决策模型作为一个符合人类决策认知的人工智能算法已成功应用在情感分析、医疗系统、属性约简等领域[11-19]. 然而，现实的决策往往是一个动态演变的过程，因此 Yao 结合粒计算的思想提出了动态的三支决策方法——序贯三支决策(Sequential Three-way Decisions，S3WD)[20-21]. 其主要思想为在粗粒度空间下由于信息系统中具有较大的不确定性，因此只对信息充分的对象采取相应的决策，而信息不充分的剩余对象将保留到更细粒度空间，等获取到更多信息时再进行决策. 利用序贯三支决策处理多分类问题[22]的方法主要是对于每个决策类分别定义它的二分类，即将 m 个分类看作是 m 个二分类问题，依次对 m 个二分类进行三支决策，从而达到每一粒层的多分类效果. S3WD 作为一

个动态的多粒度分类算法很好地改善了传统的静态二支决策的不足,并更加符合当下动态多元化问题的需求. Yang 等人[23]考虑到规则冲突等问题结合矩阵决策的方法提出一种新的 S3WD 模型. Li 等人[24]利用 S3WD 解决分类领域中代价敏感的人脸识别问题. Ju 等人[25]通过选择合适的粒度空间以及属性约简方法,提高 S3WD 的分类效率和性能. Zhang 等人[26]结合自主纠错的思想,减少 S3WD 在粗粒度空间下造成的分类错误. Lang 等人[27]结合模糊集的思想,找出优化决策冲突的方案. Xu 等人[28]考虑到传统的 S3WD 用于多分类问题比较耗时,提出了一种新的基于增量式的多分类算法. Ye 等人[29]引入区间直觉模糊集,重新构造三支多分类方法,解决了传统模型在处理多分类问题时人工参数过多、计算复杂、决策冗余等问题.

然而,在多层粒度结构下的 S3WD 仍然有三个不足:(1)在粗粒度空间下,考虑属性较少且属性分类能力较弱时,容易产生决策冲突;(2)随着由粗粒度空间到细粒度空间,在粗粒度空间下已经参与过分类的属性需要和新加入的属性一起划分等价类(论域中满足相同条件划分的子类),因此当较细粒度空间下待分类对象较多时,模型的分类效率不高;(3)当所有属性考虑完之后仍还存在未分类对象时,缺乏合理的策略对最终未分类对象进行处理.

基于上述 S3WD 在多分类中存在的问题,本文提出了一个基于属性代表的多粒度集成分类算法(Multi-Granularity Ensemble Classification algorithm based on Attribute Representation, AR-MGEC). 首先,为了解决在粗粒度空间下属性较少且属性分类能力较弱容易产生决策冲突的问题,本文采用计算更加灵活的基尼增益作为属性重要度来选择属性代表,结合集成学习的思想,综合每个属性代表的分类器的分类意见进行决策. 其次,为了提高模型的分类效率,将粗粒度空间下属性代表的分类意见保存在评分表里,而不用在细粒度空间下继续考虑已经参与过分类的属性,并且保证每个属性代表进行单独分类从而减少等价类的划分,进一步的提高模型的分类效率. 最后,对于仍然未完成分类的极少部分对象,本文提出了一个符合人类认知的策略即选择最终评分表中反对率最少(相对最优)的类别作为该对象最终的分类,以此得到一个合理且有效的分类结果. 经过实验验证,AR-MGEC 在面对不同的情况下,鲁棒性和分类性能要优于 S3WD,并且随着数据量和属性个数的增加,该模型的运行时间要明显优于 S3WD 模型.

综上,本文的主要贡献在于:

(1)提出了一种基于属性代表分类的新集成方式,解决了多粒度分类算法在粗粒度下容易做出超过一定容忍程度的错误分类结果的问题,并利用综合评判的方法减少决策冲突,加快分类效率;

(2)保留了粗粒度空间下属性代表的分类意见作为细粒度空间下综合评判的一种参考,不需要对已经参与过分类的属性重复计算,从而加快了较细粒度下分类的效率;

(3)提出评分保留制度来寻找近似最优解,从而弥补了多粒度分类算法在最细粒度空间上缺乏对未分类对象进行最终分配的不足;

(4)针对当前绝大多数的多分类算法主要在单粒度层次上进行分类决策,本文提出了基于属性代表的多粒度集成分类算法,实现了在不同粒度层次上做出不同阶段的决策判断,最终获得一个更加稳定且有效的分类结果. 通过实验表明了该算法相对于其他机器学习的多分类算法的优势.

2 背景知识与相关工作

为了便于本文的描述,粒计算、序贯三支决策以及集成学习的背景知识将会在本节进行简要地介绍.

2.1 粒计算

粒计算[30]是一种符合人类认知思维的人工智能方法论,其通常被用于复杂问题的处理和分析. 多粒度结构是由不同级别的粒层组成的一个多层次结构,每一个粒层由一组具有相似信息粒度的粒子组成. 信息粒是知识表示和处理的关键组成部分,而多层信息粒度结构对于问题的描述和总体的解决方案至关重要. 随着论域中信息的增加,这些信息粒会进一步分解成更小或者更细粒度,实现了粗粒度空间向细粒度空间的渐近. 多层粒度结构会形成多层次的决策过程,其与三支决策的结合,构成了渐近计算的序贯三支决策. 序贯三支决策就是运用了粒计算的思想,实现了由粗粒度空间到细粒度空间渐近的动态决策过程,能够有效地处理和分析复杂问题.

2.2 序贯三支决策

三支决策是由粗糙集[31-33]延伸出来解决不确定性问题的方法,其采用最小期望总体风险决策策略,通过阈值 α 和 β 将论域划分为三个互不相交的三个区域(正域,边界域,负域). 序贯三支决策是结合粒计算的思想,实现了动态三支决策的效果. 在每一粒

层中，根据现有信息作出相对可靠的正向和负向策略，而将无法进行决策的对象将划分到边界域，留到下一粒层进行再次决策，并在更细粒度空间下获取更充分的信息再进行相应的划分，其主要定义如下：

定义 1[20]. 给定决策信息表 $S=(U,At\cup D,V,f)$，其中 U 表示有限非空对象的集合，D 为决策属性集，At 为条件属性集 $C_1\subset C_2\subset\cdots\subset C_n=At(|At|=m)$ 为属性集序列，其中 $C_1\neq\varnothing$. V 是属性值集合，$f:U\times At\to V$ 是一个完整的信息函数，代表论域结合条件属性划分出的三个区域的值. 多层次粒度结构表示为 $GS=\{GS_1,GS_2,\cdots,GS_n\}$，则基于等价关系 U/C_i 的第 i 个粒层为 $GS_i=(U_i,C_i\cup D,V_i,f_i)$. 给定阈值对 (α_i,β_i) 满足 $1>\alpha_i>\beta_i>0$，在第 i 个粒层中，正域、边界域以及负域的划分如下：

$$POS_{(\alpha_i,\beta_i)}(X_i)=\{x\in U_i\mid \Pr(X_i\mid[x]_{C_i})\geq\alpha_i\}\quad(1)$$

$$BND_{(\alpha_i,\beta_i)}(X_i)=\{x\in U_i\mid \beta_i<\Pr(X_i\mid[x]_{C_i})<\alpha_i\}\quad(2)$$

$$NEG_{(\alpha_i,\beta_i)}(X_i)=\{x\in U_i\mid \Pr(X_i\mid[x]_{C_i})\leq\beta_i\}\quad(3)$$

其中，X_i 表示第 i 个粒层的目标概念，$X_i\subseteq U_i$，U_i 表示第 i 粒层中待决策对象的集合，即第 i 个粒层的论域.

经过粒层 GS_i 决策之后，得到边界域 $BND_{(\alpha_i,\beta_i)}(X_i)$，对于 $BND_{(\alpha_i,\beta_i)}(X_i)$ 中的对象将作为下一粒层 GS_{i+1} 的论域 U_{i+1} 进行再次决策，满足 $U_n\subset\cdots\subset U_2\subset U_1$. 此外，粒层 GS_{i+1} 的目标概念为 $X_{i+1}=X_i\cap BND_{(\alpha_i,\beta_i)}(X_i)$ 满足 $X_1\subset X_2\subset\cdots\subset X_n$. 当边界域中对象个数为 0 或者划分到最细粒层 GS_m 时，则序贯三支决策完成决策过程，其剩余未决策的对象将等到新的信息加入时再作处理.

2.3　集成学习

集成学习[34]的主要思想是通过设置相应的规则生成多个学习器，再通过集成投票等策略将每个学习器的结果进行综合考虑，最后输出相应的结果. 通常，良好的集成学习器需要同时具备较高的准确率和多样性. 集成学习按照体分类器之间的种类关系可以划分为同态集成学习和异态集成学习两种[35]. 同态集成学习是指使用同种类的分类器，只是修改了分类器的参数，通过设置不同的参数来达到不同分类效果，从而通过每个节点上的分类结果，集成为最终的分类结果. 异态集成学习则表示使用不同种类的分类器作为分类节点进行集成，其中异态集成学习最具代表性的是元学习(Meta Learning)和叠加(Stack Generalization). 此外，Bagging、Boosting、投票表决法等也是集成学习极具代表性的算法.

根据 Bagging[36]中采用的方法，学习器的分类结果按照“少数服从多数”的投票规则，取得票最多的类别作为最终决定. 假设学习器误差独立，当每个学习器的错误率低于随机分类的错误率，则 Bagging 得到结果的错误率将低于单一学习器的错误率均值，且在 n 趋近于无穷时，Bagging 结果的错误率趋近于理论最小错误率. 因此，集成方法通过集成多个学习器获得一个强学习器的分类结果相比于单个学习器效果要更好. 在实际问题中，很多研究学者通过结合集成学习的方法来提高模型的性能.

集成学习算法之间的区别主要在三个方面：提供给个体学习器的训练数据不同，产生个体学习器的过程不同，学习结果的组合方式不同[37]. 其本质为增强集成学习的多样性以此提高集成学习的泛化能力.

2.4　多分类算法

多分类问题是数据挖掘领域中常见的问题之一，目前已经有学者提出很多相关的多分类算法，比如朴素贝叶斯分类算法(Naive Bayes，NB)、K 近邻(K-Nearest Neighbor，KNN)、逻辑回归(Logistic Regression，LR)、决策树(Decision Tree，DT)、支持向量机(Support Vector Machines，SVM)、梯度提升树算法(Gradient Boosting Decision Tree，GBDT).

NB 算法是一类基于概率统计知识的多分类算法，通过计算分类对象属于各个类的概率来选择可能性最大的类作为分类结果. 而 LR 算法在 NB 算法基础之上加入了关联规则的考虑，在算法的训练阶段，通过关联规则来优化在模型训练过程中寻找更加有益的项集. 虽然这些多分类算法的分类效率较高，但是其在复杂的多分类场景下分类效果不尽人意. 而 SVM 作为统计学衍生出的新的通用的分类算法，其通过结构化风险最小化原则来实现更好的模型训练效果，解决了小样本的机器学习问题、提高了模型的泛化能力、解决了非线性、高维等问题.

3　基于属性代表的多粒度集成分类器

在不同粒度空间下的多分类问题中，每一粒层中的对象都有可能被分到多个决策类，从而产生大量的决策冲突，很多对象往往因为无法得到唯一的分类结果，从而需要延续到下一粒层进行再次分类. 因此，多分类下的冲突分析问题[38]一直都是研究学者想要解决的一大热点问题. 然而，很多研究学者都

是基于调整阈值的角度去减少冲突的发生,这就导致面对不同问题需要不停地调整改变阈值的策略来配合数据的变化.虽然通过不停地更换策略对决策冲突进行分析处理,在多分类问题上取得了一定的效果.但是在如今波动频繁、海量数据云集的大数据时代下,这样频繁改变策略是不足以满足现实需求的,更需要一个稳定且有效的策略来减少决策冲突的发生.

当面对医疗诊断的问题时,在粗粒度空间下依靠极少的信息对患者进行强行分类是很危险的,并且不同的疾病都会伴有相同的症状,比如使用发热的症状直接将患者分类到某种病上是明显不现实的,因为症状中带有发热的疾病有很多,例如感冒、肺炎、鼻炎、肠胃炎、病毒性脑炎、新型冠状病毒肺炎等等,所以往往需要进一步的检测是否有其他症状或者异常的身体指标才能更精确地对患者进行诊断.下面本文将通过例 1 来分析该类问题.

例 1. 给定决策信息表 $S=(U,At\cup D,V,f)$,其中 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$ 为 10 个到医院进行初步诊断的患者,$At=\{$全身,头部,腹部,胸部,皮肤$\}$为患者的条件属性集即症状,$D=\{d\}$为决策属性集,即患者具体患了何种病.表 1 给出了具体信息表的细节部分.

表 1 决策信息表

U	全身	头部	腹部	胸部	皮肤	d
x_1	发热	鼻塞	正常	正常	正常	感冒
x_2	正常	头晕 头痛	正常	正常	皮肤 干燥	正常
x_3	贫血	正常	腹痛 出血	正常	正常	胃癌
x_4	四肢 麻木	头晕 头痛	正常	胸闷	正常	高血压
x_5	正常	正常	正常	胸闷	正常	正常
x_6	发热	头晕 头痛	腹痛 出血	正常	正常	结直肠癌
x_7	四肢 麻木	正常	腹痛	正常	皮肤 感染	糖尿病
x_8	正常	头晕 头痛	正常	胸闷	正常	冠心病
x_9	发热	鼻塞	正常	胸闷	正常	肺炎
x_{10}	发热	鼻塞	正常	正常	正常	鼻炎

根据信息表里的信息,10 个患者基于决策属性集 $D=\{d\}$,可以划分为 9 个类 $\Omega=\{$感冒,正常,胃癌,高血压,结直肠癌,糖尿病,冠心病,肺炎,鼻炎$\}$.根据定义 1,可以构建该信息表下的多层次粒度结构 $GS=\{GS_1,GS_2,GS_3,GS_4,GS_5\}$,其中 GS_i 表示第 i 个粒层.基于多层次粒度结构,得到条件属性序列 $C_1\subset C_2\subset\cdots\subset C_n=At(|At|=m)$ 单粒层中的属性集为 $C_1=\{$全身$\}$,$C_2=\{$全身,头部$\}$,$C_3=\{$全身,头部,腹部$\}$,$C_4=\{$全身,头部,腹部,胸部$\}$,$C_5=\{$全身,头部,腹部,胸部,皮肤$\}$.利用多粒度序贯三支决策模型,可以得到如下基于决策信息表 2 的分类过程.

表 2 多粒度序贯三支决策模型的分类过程

U	GS_1	GS_2	GS_3	GS_4	GS_5
x_1	—	—	—	—	—
x_2	正常	正常	正常	正常	正常
x_3	胃癌	胃癌	胃癌	胃癌	胃癌
x_4	—	高血压	高血压	高血压	高血压
x_5	正常	正常	正常	正常	正常
x_6	—	结直肠癌	结直肠癌	结直肠癌	结直肠癌
x_7	—	糖尿病	糖尿病	糖尿病	糖尿病
x_8	正常	正常	正常	正常	正常
x_9	—	—	—	肺炎	肺炎
x_{10}	—	—	—	—	—

在第一粒层 GS_1 中,患者 x_8 本该诊断为冠心病,因为没有考虑到其他症状就被误诊为正常,这在医学诊断领域是非常严重的错误.并且由于多粒度的特性,这个误分类结果将一直留存,无法更改.随着粒层的递进,依然有全身和头部属性都相同的对象无法区分,如第二粒层 GS_2 中头部症状同时为鼻塞的疾病就有肺炎、鼻炎和感冒,因此产生了决策冲突,导致三个患者 x_1,x_9 和 x_{10} 都无法进行划分,只能留待下一粒层.最后,所有粒层分类完成后仍残留两个对象 x_1 和 x_{10} 没有得到分类结果,而序贯三支决策模型对这类对象也没有策略进行后续处理.然而在医疗领域,医生不可能让患者先回去,等再出现新的症状再来进行相应的诊断,因此需要全新的方法解决这一问题.

综合上述实例和分析可得,基于多粒度的序贯三支决策模型进行多分类会有以下问题:

(1)在粗粒度空间下,属性信息较少,很容易造成决策冲突;

(2)随着对象个数的减少,弱分类条件属性很难对剩余对象进行分类;

(3)没有相应的策略对剩余未分类对象进行最终分类,并且保证该分类结果的准确率和合理性.

因此基于上述问题,本文提出了一个新的多分类模型—基于属性代表的多粒度集成分类器,首先本文利用决策树中计算更加迅速的基尼增益作为分类属性的属性重要度,选择出该粒层中分类能力较强的最佳属性代表.其次,本文通过引入集成学习的思想,构建每一粒层中的属性代表集成分类器,利用属性代表给出的不同分类意见进行综合评判,并且每一粒度空间下的分类结果将被保存在评分表,通

过利用上一粒层留下的分类意见作为该粒层进行分类的参考，以此来减少该粒层中的分类难度，即利用“排除法”的思想将前$(i-1)$粒层反对的不可能的分类排除在外，在剩余可能的分类中选择合适的分类. 最后，本文利用最终的评分对仍未分类的对象选择一个“相对最优”的分类结果.

3.1　基于基尼增益的属性代表选择策略

在多分类问题中，基于序贯三支决策形成的多层次粒度结构，本文由粗粒度空间到细粒度空间组建了多粒度集成分类器，其多分类粒度结构表示为$MCGS=\{Mcgs_1,Mcgs_2,\cdots,Mcgs_n\}$，其中$Mcgs_i$，$i=1,2,\cdots,n$表示第$i$个粒层结构. 在不同的粒度空间下，条件属性的分类能力是有明显差异的，而要想选择出分类能力突出的属性代表，需要利用客观合理的指标对条件属性的分类能力进行相应地刻画. 因此，本文引入了决策树中常用的基尼增益来刻画条件属性的属性重要度，并将每个条件属性的基尼增益作为其分类结果的权重. 下面将给出每个粒结构下，属性代表的选择过程.

定义 2.　给定粒层$Mcgs_i$下的决策信息表$S_i=(U_i,At_i\cup D,V_i,f_i)$，$i=1,2,\cdots,n$，其中$U_i$表示该粒层中剩余待分类对象的集合，$D=\{d\}$为决策属性集，$At_i$为该粒层中剩余的条件属性的集合. $\Omega=\{D^1,D^2,\cdots,D^v\}$为$Mcgs_i$下的决策类集合，根据决策属性集$D=\{d\}$形成$v$个不相交的决策类，对于决策类$D^l(l=1,2,\cdots,v)$形成二分类$\{D^l,D^{l^C}\}$. 利用基尼增益获得该粒层中每个条件属性的分类能力ca_i^j，其中i表示第i个粒层，j表示第j个条件属性，$j=1,2,\cdots,m$. 其具体计算公式如下：

$$Gini(U_i)=1-\sum_{l=1}^{v}P(u_i^l)^2 \tag{4}$$

$$Gini(U_i|a_i^j)=1-\sum_{l=1}^{v}P(u_i^l)\sum_{q=1}^{Q}P(u_i^l|v_i^{j^q})^2 \tag{5}$$

$$ca_i^j=Gains_{Gini}(U_i,a_i^j)=Gini(U_i)-Gini(U_i|a_i^j) \tag{6}$$

其中，u_i^l第i个粒层中论域U_i满足决策类$d=D^l$的对象，$v_i^{j^q}$表示第i个粒层中论域U_i满足条件属性a_i^j为第q个属性值$a_i^{j^q}$的对象，Q为条件属性的属性值个数.

通过比较分类能力，可以将该粒层中剩余的条件属性进行排序，并挑选出合适的属性代表构建该粒层中的集成分类器. 首先，该模型会优先选择分类能力最强的条件属性构建集成分类器，为了防止该粒层中权重最大的属性做出错误的决定并且无法纠正，出于分类鲁棒性的考虑设置了属性代表的选择策略，即按照分类能力的大小依次选择属性代表构建集成分类器，直到分类能力最强的属性代表权重小于其他属性代表的分类权重之和，利用“一超多强”的思想，通过其他属性代表的分类结果来纠正权重最大的属性代表错误的决定.

第i粒层中挑选出的属性代表的集合记为At_r_i，对每一个属性代表分类能力进行归一化处理作为其该属性代表的分类器的权重，则所有属性代表的分类权重集合为$CW_i=\{cw_i^1,cw_i^2,\cdots,cw_i^z\}$，其中$z$表示该粒层中属性代表的个数，$cw_i^j$表示第$i$粒层中第$j$个属性代表的分类权重. 其分类器的分类权重计算公式可表示为

$$cw_i^j=ca_i^j\Big/\sum_{t=1}^{z}ca_i^t \tag{7}$$

3.2　基于属性代表的集成分类器

根据 3.1 节的基于基尼增益的属性代表选择策略，本文首先对$MCGS=\{Mcgs_1,Mcgs_2,\cdots,Mcgs_n\}$的每一粒层构建了相应的属性代表集成分类器. 其次，将每一个决策类当作一个二分类，利用三支决策的思想，对每个二分类进行正域、负域、边界域的划分. 最后，分类器对每个对象的决策结果按照正域、负域、边界域进行三种不同的操作，形成如下多粒度集成分类器的分类规则：

定义 3.　给定第i个粒层$Mcgs_i$下的决策信息表$S_i=(U_i,At_i\cup D,V_i,f_i)$，$i=1,2,\cdots,n$，其中$U_i$表示该粒层中剩余待分类对象的集合，$D=\{d\}$为决策属性集，$At_i$为该粒层中条件属性的集合，$At_r_i$为该粒层中属性代表的集合. $\Omega=\{D^1,D^2,\cdots,D^v\}$为$Mcgs_i$上的决策类集合，根据决策属性集$D=\{d\}$形成$v$个不相交的决策类，对于决策类$D^l(l=1,2,\cdots,v)$形成二分类$\{D^l,D^{l^C}\}$. 对于$\forall x\in U_i$，$a_i^j\in At\_r_i$，得到该对象$x$相对于决策类$D^l$的隶属度

$$\Pr_{a_i^j}^{D^l}(X_i^{D^l}|[x])=\frac{|X_i^{D^l}\cap[x]|}{|[x]|},$$

其中$X_i^{D^l}$为论域U_i在决策类D^l下的等价类. 基于决策阈值对(α,β)和决策理论粗糙集的思想，可得粒层$Mcgs_i$下的最小风险决策规则：

(P) 如果$\Pr_{a_i^j}^{D^l}(X_i^{D^l}|[x])\geq\alpha$，则表示属性$a_i^j$赞成对象$x$划分到$D^l$类；

(N) 如果$\beta<\Pr_{a_i^j}^{D^l}(X_i^{D^l}|[x])<\alpha$，则属性$a_i^j$对于对象$x$是否划分到$D^l$类表示中立；

(B) 如果$\Pr_{a_i^j}^{D^l}(X_i^{D^l}|[x])\leq\beta$，则表示属性$a_i^j$反对对象$x$划分到$D^l$类.

基于上述的分类规则，在粒层$Mcgs_i$下可以得

到基于属性代表和 v 个决策类的正域、边界域、负域，则每个论域 U_i 中的对象都能获得对应 v 个决策类的评价(赞成、中立、反对). 对象 x_k，$k=1,2,\cdots,n$ 对于 D^l 决策类的评分可表示为 $Score_{D^l}(x_k)$，即评分规则可表示为：

(P) 假设对象 x_k 获得了属性代表 a_i^j 对于决策类 D^l 的赞成意见，则对象 x_k 在决策类 D^l 上的评分 $Score_{D^l}(x_k)$ 将加上属性代表 a_i^j 的分类权重 cw_i^j；

(N) 假设对象 x_k 获得了属性代表 a_i^j 对于决策类 D^l 的中立意见，则对象 x_k 在决策类 D^l 上的评分 $Score_{D^l}(x_k)$ 将保持不变；

(B) 假设对象 x_k 获得了属性代表 a_i^j 对于决策类 D^l 的反对意见，则对象 x_k 在决策类 D^l 上的评分 $Score_{D^l}(x_k)$ 将减去属性代表 a_i^j 的分类权重 cw_i^j.

基于上一粒层留下的评分，再综合这一粒层中属性代表给出的评价，优先选出对象 x 的评分最高、唯一且大于 0 的类，由于可能被多个属性代表投上反对票，因此对象 x 面对一个或者多个类的评分将会成为负值，而这些负值的评分将大大降低下一粒层的决策难度. 因为利用一些比较大的负值可以实现“排除法的思想”，即利用前人的宝贵经验规避相应的风险. 此外，为了刻画每个对象对于 v 个不同的决策类的评分，本文构建了每个对象对于 v 个决策类的评分表，如表 3 所示.

表 3 评分表

U	D^1	D^2	…	D^v
x_1	$Score_{D^1}(x_1)$	$Score_{D^2}(x_1)$	…	$Score_{D^v}(x_1)$
x_2	$Score_{D^1}(x_2)$	$Score_{D^2}(x_2)$	…	$Score_{D^v}(x_2)$
…	…	…	…	…
x_n	$Score_{D^1}(x_n)$	$Score_{D^2}(x_n)$	…	$Score_{D^v}(x_n)$

当每个粒层中的集成分类器分类完成后，评分表也随之更新. 在 $Mcgs_i$ 粒层中，基于更新完成的评分表，对论域 U_i 中剩余待分类对象 x 进行分类，取对象 x 相对于 v 个不同的决策类中最大且唯一的评分所代表的类作为该对象具体的分类，其最大且唯一的评分为

$$Score_{D^t}(x)=\max\{Score_{D^1}(x),Score_{D^2}(x),\cdots,Score_{D^v}(x)\},\ \forall i,t\in[1,v]\text{且}\ i\neq t,Score_{D^t}(x)\neq Score_{D^i}(x) \tag{8}$$

基于该最大且唯一的评分，根据“少数服从多数”的原则，可以得到在 $Mcgs_i$ 粒层中的分类规则为

(1) 如果对象 x 的评分最大值存在、不冲突且 $Score_{D^t}(x)>0$ 时，将对象 x 划分为 D^t 类；

(2) 如果对象 x 的评分最大值存在但冲突时，将对象 x 划分到下一粒层的论域 U_{i+1} 中；

(3) 如果对象 x 的评分最大值存在、不冲突但 $Score_{D^t}(x)\leq 0$ 时，将对象 x 划分到下一粒层的论域 U_{i+1} 中.

根据以上分析，AR-MGEC 算法的分类规则具有以下优点：

(1) 利用多个属性代表的集成分类器的分类意见，大大降低了决策冲突的产生，使得论域 U_i 中剩余待分类对象 x 能得到快速分类；

(2) 每个对象的分类结果是经过属性代表的分类意见综合评判所得，提高了分类结果的合理性和准确率；

(3) 剩余待分类对象 x 的评分将保留下一粒层进行结算，利用“排除法”的思想，把不可能的分类排除在外不予考虑，提高了分类的效率，降低了剩余待分类对象的分类难度.

综上所述，本文通过多分类粒度结构 *MCGS* 中每一粒层的属性代表集成分类器取得相应的分类结果和评分表. 然而在现实的多分类问题中，庞大信息量的数据集中存在很多冗余的信息，因此即便本文通过多粒度集成分类，依然存在极小部分对象无法被分类的情况. 针对这一问题，本文利用每一粒层都进行更新的评分表设置了快速且有效的最终分类策略. 对于最终未分类对象 x，通过比较它的 v 个不同的决策类的评分，利用“相对最优”的思想，从对象 x 的所有评分中选择反对率最少的决策类作为该对象的最终分类，并且这个结果是从粗粒度空间渐进到细粒度空间的属性代表分类意见的统一，可以看作为所有条件属性的意见统一，具有一定的可信度. 下面本文将通过例 2 来分析和对比 AR-MGEC 和 S3WD 在同样的信息表下的决策过程.

例 2. 基于表 2 的决策信息表 $S=(U,At\cup D,V,f)$，其中 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}\}$ 为 10 个到医院进行初步诊断的患者，$At=$ {全身，头部，腹部，胸部，皮肤} 为患者的条件属性集即症状，$D=\{d\}$ 为决策属性集，即患者的患病种类. 根据 3.1 节中属性代表的选择策略和定义 2 的基尼增益的计算公式，可以得出患者的条件属性的基尼增益如表 4 所示.

表 4 基尼增益表

At	全身	头部	腹部	胸部	皮肤
$Gains_{Gini}(U,a_i)$	0.3165	0.2557	0.1904	0.0730	0.1643

根据表 4 基尼增益表，在 $Mcgs_1$ 粒层中选择了 $At\_r_1=$｛全身，头部，腹部｝为该粒层的属性代表集，其分类权重如表 5 所示.

表 5　$Mcgs_1$ 粒层中属性代表的分类权重表

At_r_1	全身	头部	腹部
cw_1^j	0.4150	0.3353	0.2597

在 $Mcgs_1$ 粒层中，由于 x_4 对象‘全身’的属性代表的属性值为四肢麻木，而属性值为四肢麻木的决策类有高血压、糖尿病，这种情况在 S3WD 的多分类方法中肯定是会发生决策冲突的，然而在本模型中很好地规避了这一问题. 在本模型中，这只表示对象 x_4 相对于这两个决策类都获得了‘全身’属性代表的支持，即 $Score_{高血压}(x_4)$、$Score_{糖尿病}(x_4)$ 都加上‘全身’属性代表的分类权重 $cw_1^{全身}$，而对象 x_4 相对于其他决策类的评分将减去 $cw_1^{全身}$；其‘头部’属性代表的属性值为头晕、头痛，而属性值为头晕、头痛的决策类有正常、高血压、结直肠癌、冠心病，则 $Score_{正常}(x_4)$、$Score_{高血压}(x_4)$ 等都加上‘头部’属性代表的分类权重 $cw_1^{头部}$；其‘腹部’属性代表的属性值为正常，而属性值为正常的决策类有感冒、正常、高血压、冠心病、肺炎、鼻炎，则 $Score_{感冒}(x_4)$、$Score_{高血压}(x_4)$ 等都得加上‘头部’属性代表的分类权重 $cw_1^{头部}$，以此类推，其 $Mcgs_1$ 粒层中最终的评分表 $Score\_T_i$ 如表 6 所示.

表 6　$Mcgs_1$ 粒层的评分表 $Score\_T_i$

U	正常	感冒	胃癌	高血压	结直肠癌	糖尿病	冠心病	肺炎	鼻炎
x_1	−0.5005	1.0	−1.0	−0.5005	−0.1698	−1.0	−0.5005	1.0	0.3295
x_2	1.0	−0.5005	−1.0	0.1698	−0.3295	−1.0	1.0	−0.5005	−0.5005
x_3	−0.3295	−1.0	−1.0	−1.0	−0.5005	−0.3295	−1.0	−1.0	−1.0
x_4	0.1698	−0.5005	−1.0	1.0	−0.3295	−0.1698	0.1698	−0.5005	−0.5005
x_5	1.0	−0.5005	−0.3295	−0.5005	−1.0	−0.3295	0.3295	−0.5005	−0.5005
x_6	−0.3295	−0.1698	−0.5005	−0.3295	1.0	−1.0	−0.3295	−0.1698	−0.1698
x_7	−0.3295	−1.0	−0.3295	0.1698	−1.0	1.0	−1.0	−1.0	−1.0
x_8	1.0	−0.5005	−1.0	0.1698	−0.3295	−1.0	1.0	−0.5005	−0.5005
x_9	−0.5005	1.0	−1.0	−0.5005	−0.1698	−1.0	−0.5005	1.0	0.3295
x_{10}	−0.5005	0.3295	−1.0	−0.5005	−0.1698	−1.0	−0.5005	0.3295	1.0

通过 $Mcgs_1$ 的评分表，可以得到 6 个患者的正确分类，以及 10 个患者的分类情况｛x_1：—，x_2：—，x_3：胃癌，x_4：高血压，x_5：正常，x_6：结直肠癌，x_7：糖尿病，x_8：—，x_9：—，x_{10}：鼻炎｝，成功完成 6 个对象的正确分类. 随着 $Mcgs_2$ 粒度空间下属性代表 $At\_r_2=$｛皮肤，胸部｝的分类器进行更新评分表，本模型成功完成了所有患者的正确分类，不仅解决了传统多粒度序贯三支决策模型多分类方法中决策冲突频发的问题，并且解决了随着粒度的细分，只需要利用评分表就能获得之前利用过的属性进行决策，而不需要将之前的属性重新考虑，大大提高了决策的效率，并且最终利用了评分表的相对最优解较好解决了最终无法分类的对象的具体分配问题.

从模型的执行效率来看，本文提出的 AR-MGEC 算法虽然需要在每一粒层都需要重新计算属性重要度对粒层进行排序. 当属性较多的情况下，重新计算属性重要度会造成一定的开销，因此可以采取相应的“优中选优”策略去避免这一问题，在每一粒层只考虑重新计算上一粒层表现较好且没被选中的属性，这样就可以大大避免属性过多导致的计算量增大的问题. 相比于传统的 S3WD 多分类算法，AR-MGEC 算法在每一粒层只需要属性进行单独分类，而不需要组合在一起划分等价类来计算隶属度. 假定数据集中有 4 个属性，且每个属性都有 10 个不同的属性值，基于 S3WD 的思想，需要将 4 个属性的不同情况进行排列组合，总共需要划分出 $10^4=10000$ 个等价类，相当于每种属性不同情况的乘法. 而 AR-MGEC 算法只需要每个属性单独考虑，即划分出 $(10+10+10+10)=40$ 个等价类，相当于每种属性不同情况的加法. 因此，当属性或属性值个数较多时，本文算法相较于 S3WD，能大大降低等价类划分的时间，以及基于等价类计算隶属度的迭代次数，从而减少了分类模型的运行时间.

综上所述，本文提出了一种新的集成策略，将单一属性视为独立的弱分类器，摆脱了原来集成方法需要搭建不同的分类模型或是不同参数的分类模型的思想. 在此基础上，本文进一步提出了分类器搭配策略，通过选择每一粒层中理论最优的分类器组合，完成多层粒度空间下最优分类结果的计算，让分类效率和分类性能都能获得一定程度上的收益.

3.3　算法流程

基于 3.1 节、3.2 节的内容，构建了下列基于属性代表的多粒度集成分类算法. 其中，图 1 为 AR-MGEC

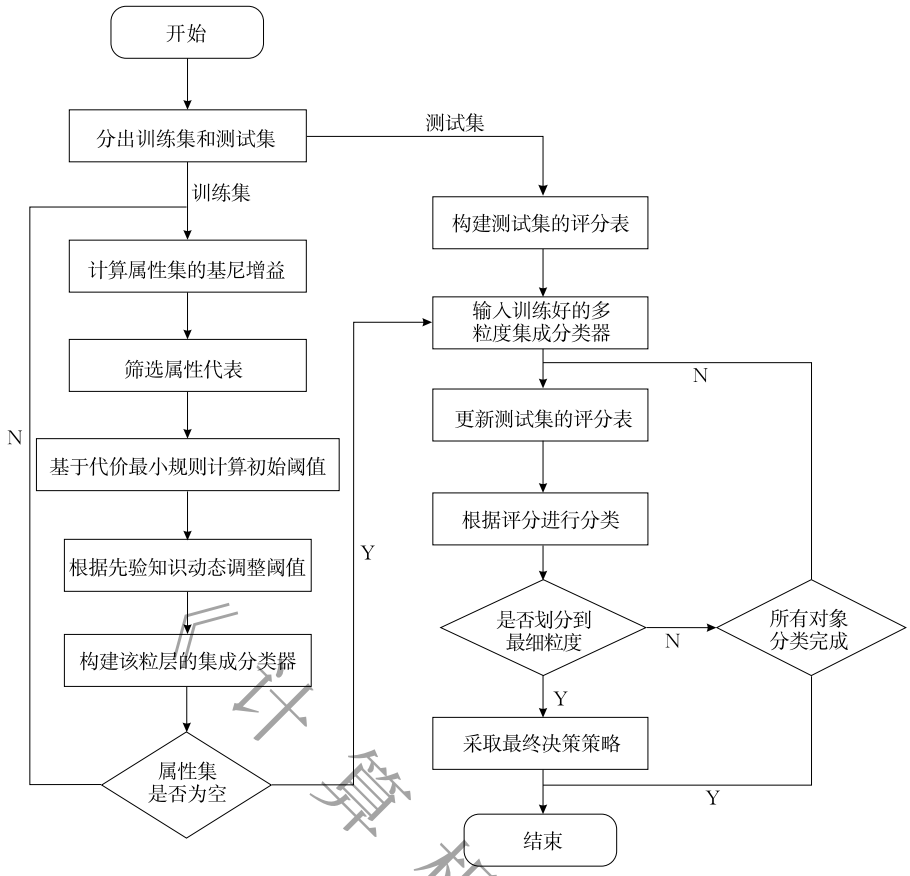

图 1 AR-MGEC 算法流程图

算法流程图,算法 1 和算法 2 分别为粒层 $Mcgs_i$ 中的属性代表选择和基于属性代表的集成分类.

算法 1 通过计算粒层 $Mcgs_i$ 中属性集 At_i 的基尼增益,通过鲁棒性策略选择相应的属性代表 At_r_i,及属性代表的归一化后的分类权重集 CW_i,其中算法 1 的时间复杂度为 $O(|U_i| * |At_i|)$,可简化为 $O(n)$.

算法 1. 粒层 $Mcgs_i$ 中的属性代表选择.

输入:$S_i=(U_i, At_i \cup D, V_i, f_i)$

输出:At_r_i, CW_i

1. 计算论域 U_i 的先验基尼系数 $Gini(U_i)$;
2. FOR each $a_i^j \in At_i$ do
3. 计算 a_i^j 的条件基尼系数 $Gini(U_i|a_i^j)$;
4. 计算 a_i^j 的基尼增益 $Gini(U_i, a_i^j)$;
5. END FOR
6. 根据鲁棒性策略选择属性代表加入 At_r_i;
7. FOR each $a_i^j \in At_i$ do
8. 计算 a_i^j 的分类权重 $cw_i^j=\dfrac{Gini(U_i, a_i^j)}{\sum_{t=1}^{z} Gini(U_i, a_i^t)}$;
9. 将权重加入 CW_i 中
10. END FOR
11. RENTURN At_r_i, CW_i

算法 2. 粒层 $Mcgs_i$ 中基于属性代表的集成分类.

输入:$S_i=(U_i, At_i \cup D, V_i, f_i), At\_r_i, CW_i, (\alpha, \beta), Score\_T_{i-1}$

输出:$Score\_T_i, res$

1. FOR each $a_i^j \in At\_r_i$ do
2. 构建属性代表 a_i^j 的弱分类器;
3. IF $\Pr_{a_i^j}^{D^l}(X_{D^l}|[x]) \geq \alpha$ THEN
4. $Score_{D^l}(x) += cw_i^j$;
5. END IF
6. IF $\Pr_{a_i^j}^{D^l}(X_{D^l}|[x]) \leq \beta$ THEN
7. $Score_{D^l}(x) -= cw_i^j$;
8. 根据分类意见更新评分表 $Score\_T_i$;
9. END FOR
10. FOR each $D^l \in D$ do
11. IF $Score_{D^l}(x)$最大且唯一 THEN
12. 将 x 划分为 D^l 类;
13. END FOR
14. RETURN $Score\_T_i, res$

算法 2 通过粒层 $Mcgs_i$ 的属性代表 At_r_i 构建集成分类器,利用三支决策的思想设置正域、负域、边界域的三种不同执行策略,通过综合属性代表不

同的分类意见更新评分表 $Score\_T_i$，并根据这一粒层更新好的评分表对论域 U_i 中的对象进行相应的分类，其时间复杂度为 $O(|U_i| * (|At\_r_i| + |D|))$，可简化为 $O(n)$.

综合上述，AR-MGEC 算法总的时间复杂度为 $O(|U_i| * (|At_i| + |At\_r_i| + |D|))$，当属性个数和决策类个数远小于对象个数的数据集中，该算法的时间复杂度为 $O(n)$. 而当属性个数和决策类个数大于等于对象个数的数据集中，该算法的时间复杂度为 $O(n^2)$.

4　实验与分析

4.1　实验设置

本节对本文实验所用的数据集和实验方法做出了介绍，并对实验结果进行了相应的分析. Lenses、Zoo、Hayes-roth、Lymphography、Tae、Soybean(Large)、Dermatology、Balance-scale、Audit Data、Cmc、Car-Evalution、Nursery、Adult 为本文实验所用到的 14 个 UCI 标准数据集，以及 6 个来自英国的 UK Biobank 由不同病的患者组合的医疗数据集(随机抽取). 为了更好地展现实验效果，对这 20 个数据集进行了相应的数据处理. Hayes-roth 中的 name 属性为每个对象的独立属性，因此在做实验之前对其做删除处理. Cmc、Audit Data、PhishingData、Hayes-roth 和 Adult 的 UCI 数据集以及 6 个来自 UKB 的真实数据集中部分条件属性为连续性数据，因此做了一定的离散化处理，即利用最大最小值通过分段化成离散的属性值. 其实验环境为 macOS10.14.4 系统和 8 GB RAM，2.3 GHz CPU，双核处理器，以及编程语言为 Python. UCI 标准数据集和真实数据集的基本信息如表 7 和表 8 所示.

表 7　UCI 标准数据集的描述

ID	Dataset	$\|At\|$	$\|U\|$	$\|D\|$
1	Lenses	4	24	3
2	Zoo	17	101	7
3	Hayes-roth	5	132	3
4	Lymphography	18	148	4
5	Tae	5	151	3
6	Soybean(Large)	35	307	19
7	Dermatology	33	366	6
8	Balance-scale	4	625	3
9	Audit Data	17	776	2
10	Phishingdata	9	1353	3
11	Cmc	9	1473	3
12	Car-evalution	6	1728	4
13	Nursery	8	12960	5
14	Adult	14	48842	2

表 8　UKB 真实数据集的描述

ID	Dataset	$\|At\|$	$\|U\|$	$\|D\|$
1	UKB-1	19	1155	3
2	UKB-2	18	1509	4
3	UKB-3	18	702	3
4	UKB-4	18	839	3
5	UKB-5	16	2073	5
6	UKB-6	19	1205	3

数据可用性声明：本文实验使用的 UK Biobank(简称 UKB)来源于使用 ID：51470. 研究者可以通过官方网站(https://www.ukbiobank.ac.uk/)申请 UKB 数据的使用.

4.2　分类性能对比与分析

为了验证 AR-MGEC 的有效性，在本文实验中，通过与同为多粒度分类算法的序贯三支多分类在每一粒层的执行时间以及最终的分类准确率、Kappa 分类一致性系数、宏观分类精度、宏观分类召回率、宏观 $F1$ 值以及总耗时进行纵向对比和分析. 其 UCI 数据集和真实数据集下分类准确率、Kappa 系数、分类精度、分类召回率、$F1$ 值的比较如表 9、表 10 所示.

表 9　UCI 数据集下分类性能纵向对比

Dataset	分类准确率		Kappa 分类一致性系数		分类精度		分类召回率		$F1$ 值	
	S3WD	AR-MGEC	S3WD	AR-MGEC	S3WD	AR-MGEC	S3WD	AR-MGEC	S3WD	AR-MGEC
Lenses	0.9166	**0.9583**	0.8339	**0.9250**	0.8500	**0.9778**	**0.9608**	0.9444	0.9020	**0.9608**
Zoo	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
Hayes-roth	0.6439	**0.7575**	0.4637	**0.6140**	0.6242	**0.7085**	0.7737	**0.8224**	**0.6909**	**0.7612**
Lymphography	0.7905	**0.8446**	0.6018	**0.6926**	**0.7683**	0.5543	**0.7709**	0.6726	**0.7696**	0.6077
Tae	0.7748	**0.8543**	0.6615	**0.7812**	0.8213	**0.8538**	0.8264	**0.8590**	0.8239	**0.8564**
Soybean(Large)	**0.9674**	0.9609	**0.9643**	0.9572	0.9829	**0.9842**	0.9794	**0.9879**	0.9811	**0.9860**
Dermatology	0.9699	**0.9781**	0.9624	**0.9726**	0.8819	**0.9728**	0.9387	**0.9807**	0.9094	**0.9767**
Balance-scale	0.7328	**0.8416**	0.5061	**0.7062**	0.5470	**0.6087**	**0.82638**	0.5611	**0.6582**	0.5839
Audit Data	0.9510	**1.0000**	0.8966	**1.0000**	0.9539	**1.0000**	0.9438	**1.0000**	0.9488	**1.0000**
PhishingData	0.8499	**0.8536**	0.7196	**0.7252**	**0.6351**	0.6156	**0.8978**	0.5673	**0.7439**	0.5905
Cmc	0.5227	**0.5605**	0.2675	**0.2713**	0.5054	**0.5161**	0.5188	**0.5464**	0.5120	**0.5309**
Car-Evalution	0.8310	**0.8472**	0.5353	**0.6672**	0.4386	**0.4571**	**0.7013**	0.3932	**0.5397**	0.4227
Nursery	0.8076	**0.8588**	0.7191	**0.7907**	**0.8845**	0.5317	**0.9078**	0.5192	**0.8960**	0.5254
Adult	**0.8417**	0.8246	0.5059	**0.5476**	0.5597	**0.7606**	0.4756	**0.7916**	0.5143	**0.7758**

表 10 真实数据集下分类性能纵向对比

Dataset	分类准确率		Kappa分类一致性系数		分类精度		分类召回率		F1 值	
	S3WD	AR-MGEC	S3WD	AR-MGEC	S3WD	AR-MGEC	S3WD	AR-MGEC	S3WD	AR-MGEC
UKB-1	0.5755	**0.5903**	0.3496	**0.3754**	0.5729	**0.5924**	0.5646	**0.5860**	0.5687	**0.5892**
UKB-2	0.4871	**0.5149**	0.2467	**0.2869**	0.3980	**0.4179**	0.3040	**0.4011**	0.3447	**0.4093**
UKB-3	0.6199	**0.6667**	0.1389	**0.1592**	0.6891	**0.7811**	0.3992	**0.3917**	0.5055	**0.5218**
UKB-4	0.5387	**0.6758**	0.1578	**0.2781**	0.4862	**0.7880**	0.3294	**0.4359**	0.3928	**0.5613**
UKB-5	0.7086	**0.7660**	0.6275	**0.7012**	0.7629	**0.7722**	0.7092	**0.7516**	0.7350	**0.7618**
UKB-6	0.5029	**0.5826**	0.1536	**0.2887**	0.6140	**0.6863**	0.4126	**0.4757**	0.4936	**0.5619**

由结果可知,基于属性代表的多粒度集成分类算法的准确率和 Kappa 分类一致性系数在多个数据集中相比于传统多粒度序贯三支分类算法有了一定的提升.尤其在 PhishingData 数据集中,每一粒层的属性代表很好地通过协作实现了高准确率的分类效果.然而,传统多粒度序贯三支分类算法在 14 个 UCI 数据集上的实验结果不太稳定,这是由于该算法在粗粒度空间下造成了一定的错误,并且在划分到最细粒度空间之后仍存在一些未分类对象无法处理.

4.3 分类效率对比与分析

分类效率是分类模型好坏的主要指标之一,表 11 展示了两个多粒度分类算法在 14 个 UCI 数据集的 4 个分类粒层 level-1、level-2、level-3、level-4 下的执行时间以及总耗时.

表 11 14 个分类粒层下 UCI 数据集的执行时间以及总耗时

Dataset		S3WD	AR-MGEC	Dataset		S3WD	AR-MGEC
lenses	level-1	0.0076	0.0077	Zoo	level-1	0.0327	0.0213
	level-2	0.0073	0.0086		level-2	0.0305	0.0100
	level-3	0.0137	0		level-3	0.0256	0.0087
	level-4	0.0221	0		level-4	0.0086	0
	总耗时	0.0509	**0.0164**		总耗时	0.0975	**0.0401**
Hayes-roth	level-1	0.0090	0.0213	Lymphography	level-1	0.0314	0.0221
	level-2	0.0266	0.0158		level-2	0.0108	0.0124
	level-3	0.0247	0		level-3	0.0035	0
	level-4	0.0914	0		level-4	0.0060	0
	总耗时	0.1519	**0.0372**		总耗时	0.0518	**0.0346**
Tae	level-1	0.0576	0.1393	Soybean(Large)	level-1	1.2980	0.1197
	level-2	0.0416	0.0041		level-2	0.0502	0.0650
	level-3	0.0208	0		level-3	0.0072	0.0451
	level-4	0.0367	0		level-4	0.0086	0.0252
	总耗时	0.1568	**0.1434**		总耗时	1.3641	**0.2251**
Dermatology	level-1	0.5536	0.0383	Balance-scale	level-1	0.0122	0.0342
	level-2	0.1584	0.0347		level-2	0.0209	0.0090
	level-3	0.0053	0.0366		level-3	0.0338	0
	level-4	0.0071	0.0225		level-4	0.0404	0
	总耗时	0.7246	**0.1323**		总耗时	0.1074	**0.0433**
Audit Data	level-1	1.0322	0.9319	PhishingData	level-1	0.0493	0.0303
	level-2	0.0445	0		level-2	0.0661	0
	level-3	0.0141	0		level-3	0.0114	0
	level-4	0.0047	0		level-4	0.0025	0
	总耗时	1.0957	**0.9319**		总耗时	0.0694	**0.0303**
Cmc	level-1	0.0292	0.0633	Car-evalution	level-1	0.0180	0.0352
	level-2	0.0678	0.0161		level-2	0.0364	0.0225
	level-3	0.0460	0.0231		level-3	0.1328	0.0114
	level-4	0.0289	0.0137		level-4	0.6657	0
	总耗时	0.1720	**0.1163**		总耗时	0.8531	**0.0691**
Nursery	level-1	0.0899	0.1982	adult	level-1	3.9504	0.3393
	level-2	1.1740	0.1064		level-2	8.6991	0.2775
	level-3	6.4487	0.0757		level-3	2.1222	0.3127
	level-4	16.0040	0		level-4	2.2129	0.1872
	总耗时	23.7160	**0.3803**		总耗时	16.984	**1.1168**

表 12　6 个分类粒层下 UKB 数据集的执行时间以及总耗时

Dataset		S3WD	AR-MGEC	Dataset		S3WD	AR-MGEC
UKB-1	level-1	0.1305	0.0197	UKB-2	level-1	0.3241	0.0202
	level-2	0.1654	0.3115		level-2	0.0758	0.1313
	level-3	0.2002	0.0621		level-3	0.0039	0.0932
	level-4	0.0487	0.0569		level-4	0.0046	0.0248
	总耗时	0.5448	**0.4502**		总耗时	0.4084	**0.2695**
UKB-3	level-1	0.1019	0.1019	UKB-4	level-1	0.0866	0.1254
	level-2	0.0111	0.0379		level-2	0.0263	0.0668
	level-3	0.0036	0.0209		level-3	0.0038	0.0233
	level-4	0.0041	0.0568		level-4	0.0048	0.0511
	总耗时	**0.1207**	0.2175		总耗时	**0.1215**	0.2666
UKB-5	level-1	0.2684	2.1861	UKB-6	level-1	0.1142	0.1164
	level-2	2.043	0.1733		level-2	0.1642	0.0760
	level-3	0.2017	0.1093		level-3	0.1146	0.0603
	level-4	0.0047	0.0637		level-4	0.0051	0.0275
	总耗时	**2.5178**	2.5324		总耗时	0.3981	**0.2802**

由表 12 可得，随着属性值个数和数据量的增加，AR-MGEC 相比于 S3WD 能够更快的收敛以及执行效率更高. AR-MGEC 采用保留上一粒层更新的评分表来代替属性，并且让每个粒层中的属性代表集成多分类器单独分类，不仅减少了决策冲突的发生，还减少了等价类的划分次数即减少了算法的迭代次数. 因此，在数据量和属性较多的数据集时，采用 AR-MGEC 算法会更加快速且有效.

为了更好地体现 AR-MGEC 算法相比于 S3WD 在不同情况下的分类性能和分类效率的提升，本文通过表 13、表 14 和图 2、图 3 来展示 AR-MGEC 算法在 14 个 UCI 标准数据集以及 6 个真实数据集 UKB 上分类正确率、Kappa 分类一致性系数、宏观分类精度、宏观分类召回率、宏观 $F1$ 值以及模型总耗时的提升比率.

表 13　14 个 UCI 数据集下 AR-MGEC 算法的提升比率

Dataset	分类正确率/%	Kappa 一致性/%	分类精度/%	分类召回率/%	$F1$-Score/%	总耗时/%
Lenses	45.5	10.9	15.0	−1.7	6.5	67.7
Zoo	0.0	0.0	0.0	0.0	0.0	58.9
Hayes-roth	17.6	32.4	13.5	6.2	10.1	75.5
Lymphography	6.8	15.1	−27.8	−12.7	−21.0	33.3
Tae	4.0	6.8	3.9	3.9	3.9	8.5
Soybean(Large)	−0.6	−0.7	0.1	0.8	0.5	81.3
Dermatology	0.8	1.06	10.3	4.4	7.4	81.7
Balance-scale	14.8	39.5	−32.1	−32.1	−11.2	59.7
Audit Data	5.1	11.5	4.8	5.9	5.3	14.9
PhishingData	0.4	0.8	−3.0	−36.8	−20.6	56.3
Cmc	7.2	1.4	2.1	5.3	3.6	32.3
Car-evalution	1.9	24.6	4.2	−43.9	−21.6	91.9
Nursery	6.3	10.0	−39.8	−42.8	−41.3	98.4
Adult	−2.0	8.2	35.8	66.4	50.8	93.4
平均提升	7.7	11.5	−0.9	−5.4	−1.9	61.0

表 14　6 个真实数据集 UKB 下 AR-MGEC 算法的提升比率

Dataset	分类正确率/%	Kappa 一致性/%	分类精度/%	分类召回率/%	$F1$-Score/%	总耗时/%
UKB-1	2.6	7.4	3.4	3.8	3.6	16.1
UKB-2	5.7	16.3	5.0	31.9	18.7	28.5
UKB-3	7.5	14.6	13.4	−1.9	3.2	−64.8
UKB-4	25.4	76.2	62.1	32.3	42.9	6.1
UKB-5	8.1	11.7	1.2	6.0	3.6	−16.9
UKB-6	15.8	88.0	11.8	15.3	13.8	25.2
平均提升	10.6	24.8	14.6	11.9	12.0	8.6

从表 13、表 14 和图 2、图 3 中可以看出，本文提出的 AR-MGEC 算法无论是在 UCI 标准数据集上还是在医疗真实数据 UKB 上相较于传统的多粒度分类算法 S3WD，在分类正确率和 Kappa 分类一致

性系数都有小幅度的提升，在分类效率上有大幅度的提升，并且真实数据集下的宏观分类精度、宏观分类召回率以及宏观 $F1$ 值也有一定的提升. 其中，UCI 标准数据集下分类正确率的提升比率范围为 −0.6%～45.5%，平均为 7.7%；Kappa 分类一致性系数的提升比率范围为 −0.7%～39.5%，平均为 11.5%；分类效率的提升比率范围为 8.5%～98.4%，平均为 61%. 真实 UKB 数据集下，本文算法的分类正确率平均提升比例为 10.6%，Kappa 一致性分类系数提升比例为 24.8%，宏观分类精度提升比例为 14.6%，宏观分类召回率提升比例为 11.9%，宏观 $F1$-Score1 的提升比例为 12.0%，总耗时的提升比例为 8.6%. UCI 标准数据集下，在处理 3 个数据量最大且属性个数较多的 Car-evalution、Nursery 和 Adult 数据集时，AR-MGEC 算法在保证分类性能提升的基础上大大提高了分类效率，尤其是在 Nursery 数据集中，分类效率提高了近百倍，只需要传统多粒度分类算法 S3WD 总耗时的 1.6%. 综合上述可得，AR-MGEC 算法保证了分类性能提升的同时，大大提高了多粒度分类算法的分类效率，因此本文提出的 AR-MGEC 算法更加适合处理时下数据量庞大且复杂多元的分类问题.

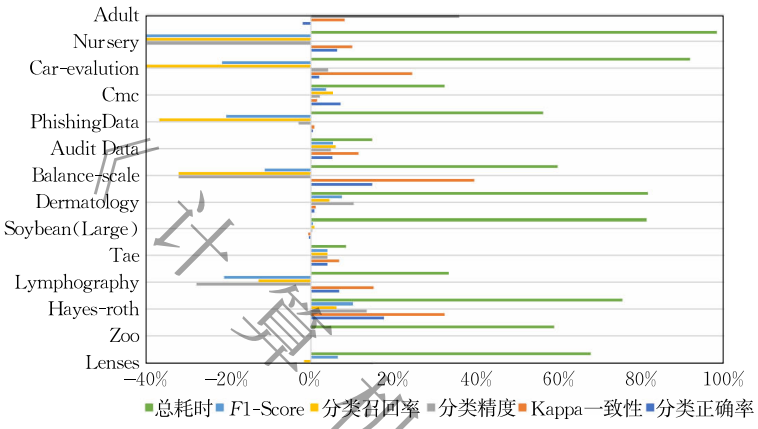

图 2 14 个 UCI 标准数据集下的提升比率

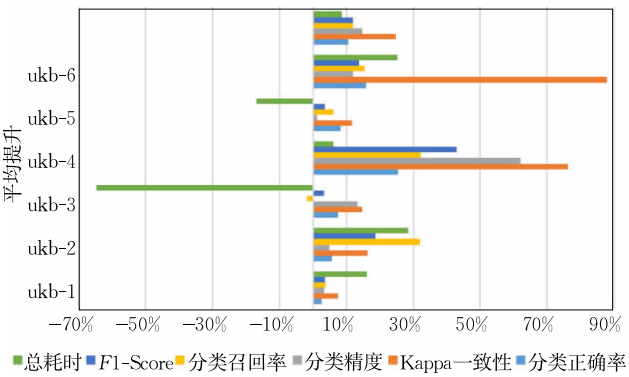

图 3 6 个真实 UKB 数据集下的提升比率

4.4 评分保留法的分类性能检验

AR-MGEC 中的评分保留法是指将粗粒度下属性代表的分类意见作为评分保留到更细粒层，细粒度下的属性代表只需要考虑分类意见而不用再利用之前的属性代表进行分类，这样不仅可以提高细粒度下分类的效率，又能保障分类的性能.

为了验证该方法的有效性，本文利用 AR-MGEC 模型在 14 个 UCI 数据集上进行了保留评分和不保留评分两种方法的对比试验，其实验结果如表 15 所示.

表 15 评分保留前后的分类性能对比

Dataset	分类正确率		Kappa 分类一致性系数	
	不保留	保留	不保留	保留
Lenses	0.7500	**0.9583**	0.5789	**0.9250**
Zoo	0.8713	**1.0000**	0.8206	**1.0000**
Hayes-roth	**0.7575**	**0.7575**	**0.6140**	**0.6140**
Lymphography	**0.8446**	**0.8446**	**0.6926**	**0.6926**
Tae	**0.8543**	**0.8543**	**0.7812**	**0.7812**
Soybean(Large)	0.7524	**0.9609**	0.7268	**0.9572**
Dermatology	0.9344	**0.9781**	0.9168	**0.9726**
Balance-scale	**0.8416**	**0.8416**	**0.7062**	**0.7062**
Audit Data	**1.0000**	**1.0000**	**1.0000**	**1.0000**
PhishingData	**0.8536**	**0.8536**	**0.7252**	**0.7252**
Cmc	0.5247	**0.5605**	0.2532	**0.2713**
Car-Evalution	**0.8472**	**0.8472**	**0.6672**	**0.6672**
Nursery	0.6625	**0.8588**	0.4958	**0.7907**
Adult	**0.8246**	**0.8246**	**0.5476**	**0.5476**

从表 15 可以看出，在部分数据集上，采用评分保留相比于不采用评分保留的分类性能有一定的提升，而其他数据集的分类结果也保持不变. 其中，Lenses、Zoo、Soybean(Large)、Nursery 等数据集下的分类性能提升最大，也表示粗粒度空间下的分类意见虽然没有对部分对象成功完成分类，但保留下来的评分大大降低了更细粒度空间下的分类难度. 因此，利用“评分保留法”将粗粒度空间下的属性代

表的分类意见保留到细粒度空间进行参考，对整个模型处理分类问题是有效的.

4.5 横向对比实验分析

为了进一步提高本文算法实验的有效性，以及体现与其他机器学习的多分类算法的优势. 在本节，本文分别与机器学习的 7 种主流多分类算法（NB、KNN、LR、RF、DT、SVM、GBDT）在 14 个 UCI 标准数据集以及 6 个 UKB 真实数据集，进行了横向对比实验. 下面表 16～35 分别展示了 UCI 数据集和真实数据集的分类结果.

表 16　UCI 数据集-Lenses 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.9583**	**0.9250**	**0.9778**	**0.9444**	**0.9608**	0.0164
NB	0.7500	0.0000	0.2500	0.3333	0.2857	0.0143
KNN	0.7500	0.3846	0.4444	0.6111	0.5146	0.1026
LR	0.7500	0.2381	0.2857	0.3333	0.3077	0.0172
RF	0.8750	0.7419	0.8333	0.9444	0.8854	0.0452
DT	0.8750	0.6190	0.6190	0.6667	0.6420	0.0065
SVM	0.7500	0.0000	0.2500	0.3333	0.2857	**0.0063**
GBDT	0.8750	0.6190	0.6190	0.6667	0.6420	0.3673

表 17　UCI 数据集-Zoo 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.0401
NB	0.9032	0.8731	0.6735	0.6429	0.6578	0.0149
KNN	0.9355	0.9164	0.8571	0.7857	0.8199	0.1214
LR	0.9355	0.9153	0.9524	0.9167	0.9342	0.0313
RF	0.9355	0.9168	0.8333	0.8095	0.8213	0.0461
DT	0.9032	0.8731	0.6735	0.6429	0.6578	0.0087
SVM	0.9032	0.8748	0.8095	0.7571	0.7825	**0.0148**
GBDT	0.9677	0.9578	0.7143	0.7143	0.7143	0.8885

表 18　UCI 数据集-Hayes roth 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	0.6439	0.4637	0.6242	0.7737	0.6910	0.1519
NB	0.7576	0.6140	0.7085	0.8224	0.7612	0.0372
KNN	0.5250	0.2423	0.6794	0.5280	0.5942	0.0118
LR	0.6250	0.3782	0.7432	0.5619	0.6400	0.1045
RF	0.5250	0.2445	0.5556	0.5423	0.5489	0.0154
DT	**0.8750**	**0.8022**	**0.9037**	**0.9037**	**0.9037**	0.0460
SVM	0.8000	0.6844	0.8471	0.8444	0.8458	**0.0067**
GBDT	0.8250	0.7197	0.8584	0.8556	0.8570	0.0085

表 19　UCI 数据集-Lymphography 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.8446**	**0.6926**	0.5543	0.6726	0.6077	0.0346
NB	0.7556	0.5152	0.5020	0.5111	0.5065	0.0179
KNN	0.7556	0.5190	0.5020	0.5139	0.5079	0.1218
LR	0.7778	0.5706	**0.8472**	**0.8472**	**0.8472**	0.0383
RF	0.7333	0.5055	0.3930	0.3750	0.3838	0.0505
DT	0.7333	0.5187	0.4900	0.6104	0.5436	**0.0094**
SVM	0.8000	0.6033	0.5324	0.5417	0.5370	0.0135
GBDT	0.8222	0.6515	0.5474	0.5611	0.5542	0.5864

表 20　UCI 数据集-Tae 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.8543**	**0.7812**	**0.8538**	**0.8590**	**0.8564**	0.1435
NB	0.3913	0.1099	0.3275	0.4191	0.3677	0.0109
KNN	0.5000	0.2622	0.5116	0.5224	0.5169	0.1098
LR	0.5435	0.3273	0.5535	0.5797	0.5663	0.0396
RF	0.5870	0.3648	0.5824	0.5827	0.5826	0.0456
DT	0.5435	0.3055	0.5353	0.5415	0.5384	**0.0098**
SVM	0.3913	0.1481	0.2778	0.4602	0.3464	0.0129
GBDT	0.7174	0.5534	0.7372	0.6727	0.7035	0.4271

表 21　UCI 数据集-Soybean 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.9609**	**0.9572**	**0.9842**	**0.9879**	**0.9860**	0.2551
NB	0.8226	0.8048	0.9081	0.9237	0.9158	0.0124
KNN	0.8548	0.8381	0.8568	0.8629	0.8598	0.1185
LR	0.9355	0.9282	0.9804	0.9786	0.9795	0.0789
RF	0.7903	0.7666	0.8099	0.8080	0.8090	**0.0556**
DT	0.8065	0.7840	0.8162	0.8737	0.8440	0.0097
SVM	0.8548	0.8371	0.8309	0.8551	0.8428	0.0652
GBDT	0.8710	0.8565	0.9112	0.9178	0.9145	3.4412

表 22　UCI 数据集-Dermatology 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.9781**	**0.9726**	**0.9728**	**0.9807**	**0.9767**	0.1323
NB	0.8727	0.8350	0.8239	0.8380	0.8309	**0.0068**
KNN	0.9818	0.9763	0.9778	0.9815	0.9796	0.0173
LR	0.9727	0.9644	0.9688	0.9722	0.9705	0.0411
RF	0.9118	0.8851	0.8897	0.8808	0.8851	0.0113
DT	0.9386	0.9204	0.9386	0.9425	0.9405	0.0079
SVM	0.9818	0.9763	0.9778	0.9815	0.9796	0.0258
GBDT	0.9545	0.9409	0.9565	0.9639	0.9602	0.9799

表 23　UCI 数据集-Balance scale 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	0.8416	0.7062	0.6088	0.5611	0.5840	0.0433
NB	0.8840	0.7841	0.5937	0.6359	0.6141	**0.0102**
KNN	0.8320	0.6984	0.6090	0.6142	0.6116	0.0221
LR	0.8920	0.8030	**0.7952**	**0.7040**	**0.7469**	0.0261
RF	0.7476	0.5675	0.5787	0.5607	0.5695	0.0135
DT	0.7580	0.5930	0.6046	0.5779	0.5910	0.0128
SVM	**0.9000**	**0.8149**	0.7720	0.6630	0.7133	0.0294
GBDT	0.8320	0.7019	0.5969	0.5988	0.5978	0.4334

表 24　UCI 数据集-Audit 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.0438
NB	0.9957	0.9908	0.9966	0.9942	0.9954	**0.0066**
KNN	0.8798	0.7561	0.8737	0.8999	0.8866	0.0223
LR	0.9871	0.9726	0.9831	0.9898	0.9865	0.0289
RF	0.9906	0.9797	0.9901	0.9899	0.9900	0.0120
DT	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.0074
SVM	0.7253	0.4074	0.7047	0.7027	0.7037	0.0578
GBDT	**1.0000**	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.1448

表 25 UCI 数据集-PhishingData 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	0.8537	0.7253	0.6157	0.5674	0.5905	0.0304
NB	0.8251	0.6812	0.6926	0.6478	0.6695	**0.0077**
KNN	0.8596	0.7466	0.8251	0.7485	0.7849	0.0304
LR	0.8522	0.7261	0.7161	0.6429	0.6775	0.0169
RF	0.8456	0.7252	0.7869	0.7761	0.7812	0.0126
DT	0.8693	0.7662	0.8547	0.8289	0.8416	0.0084
SVM	0.8842	0.7862	0.8405	0.6914	0.7587	0.1427
GBDT	**0.9113**	**0.8419**	**0.8949**	**0.8612**	**0.8777**	0.6988

表 26 UCI 数据集-Cmc 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.5587**	**0.2997**	0.5162	**0.5465**	**0.5309**	0.1164
NB	0.4729	0.2255	0.4898	0.4995	0.4946	**0.0067**
KNN	0.5136	0.2453	0.4987	0.4928	0.4957	0.0271
LR	0.4977	0.2128	0.5030	0.4755	0.4889	0.0480
RF	0.4702	0.1861	0.4552	0.4546	0.4549	0.0117
DT	0.4992	0.2304	0.4854	0.4861	0.4858	0.0094
SVM	0.4819	0.1688	0.3213	0.4266	0.3665	0.3365
GBDT	0.5407	0.2935	**0.5290**	0.5248	0.5269	0.6617

表 27 UCI 数据集-Car evalution 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	0.8472	0.6672	0.4571	0.3932	0.4227	0.0692
NB	0.6977	0.0095	0.3246	0.2518	0.2836	0.0158
KNN	0.7801	0.4569	0.7131	0.4790	0.5731	0.1587
LR	0.6833	0.1261	0.3428	0.3284	0.3354	0.0652
RF	0.7650	0.5042	0.4892	0.5349	0.5110	0.0497
DT	0.9356	0.8625	0.8126	0.8367	0.8245	**0.0106**
SVM	0.7419	0.2384	0.5754	0.3845	0.4609	0.0612
GBDT	**0.9487**	**0.8885**	**0.8760**	**0.8539**	**0.8648**	0.6333

表 28 UCI 数据集-Nursery 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	0.8650	0.7999	0.5317	0.5192	0.5254	0.3804
NB	0.6358	0.5279	0.5352	0.5781	0.5558	**0.0191**
KNN	0.9396	0.9109	0.7485	0.6686	0.7063	0.2927
LR	0.7577	0.6409	0.4878	0.4683	0.4779	0.1917
RF	0.8897	0.8381	0.6611	0.6609	0.6610	0.0226
DT	0.9906	0.9862	0.7874	0.7832	0.7853	0.0198
SVM	0.9398	0.9108	0.7645	0.5900	0.6660	4.8455
GBDT	**0.9988**	**0.9983**	**0.9991**	**0.9864**	**0.9927**	5.7159

表 29 UCI 数据集-Audit 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.8246**	**0.5476**	0.7606	**0.7917**	**0.7759**	1.1169
NB	0.7624	0.0000	0.3812	0.5000	0.4326	**0.0354**
KNN	0.8054	0.4412	0.7309	0.7120	0.7213	1.0972
LR	0.7587	0.0064	0.5391	0.5021	0.5200	0.2226
RF	0.8070	0.4376	0.7336	0.7073	0.7202	0.0488
DT	0.8150	0.4583	0.7465	0.7160	0.7309	0.0645
SVM	0.7624	0.0000	0.3812	0.5000	0.4326	208.5963
GBDT	0.8310	0.4862	**0.7783**	0.7207	0.7484	3.1166

表 30 真实数据集-UKB1 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.5903**	**0.3754**	**0.5924**	**0.5860**	**0.5892**	0.4325
NB	0.4186	0.0204	0.5831	0.3450	0.4335	**0.0091**
KNN	0.3876	0.0675	0.3773	0.3799	0.3786	0.0219
LR	0.5426	0.3104	0.5444	0.5526	0.5485	0.0555
RF	0.4349	0.1435	0.4324	0.4390	0.4357	0.0149
DT	0.4291	0.1456	0.4247	0.4528	0.4383	0.0156
SVM	0.4031	0.1223	0.1354	0.3270	0.1915	0.4900
GBDT	0.5581	0.3289	0.5598	0.5642	0.5620	0.7912

表 31 真实数据集-UKB2 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.5149**	**0.2869**	**0.4179**	**0.4011**	**0.4093**	0.2459
NB	0.0728	0.0095	0.3230	0.2739	0.2964	**0.0119**
KNN	0.3907	0.1122	0.3446	0.3233	0.3336	0.0355
LR	0.4901	0.2545	0.3810	0.3886	0.3847	0.0637
RF	0.4053	0.1352	0.3536	0.3342	0.3420	0.0171
DT	0.4038	0.1375	0.3335	0.3363	0.3349	0.0133
SVM	0.4702	0.2250	0.4082	0.3720	0.3893	0.5503
GBDT	0.4868	0.2472	0.3664	0.3825	0.3743	1.0980

表 32 真实数据集-UKB3 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.6667**	**0.1592**	**0.7811**	**0.3917**	**0.5218**	0.2160
NB	0.5975	0.0090	0.2715	0.3298	0.2978	**0.0115**
KNN	0.5409	0.0280	0.2951	0.3293	0.3113	0.0251
LR	0.6164	0.0425	0.3125	0.3449	0.3279	0.0499
RF	0.4686	0.0216	0.3159	0.3196	0.3177	0.0151
DT	0.4355	0.0333	0.3048	0.3010	0.3029	0.0118
SVM	0.6387	0.0086	0.3799	0.3364	0.3569	0.1352
GBDT	0.5588	0.0417	0.3071	0.3424	0.3238	0.6526

表 33 真实数据集-UKB4 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.6758**	**0.2781**	**0.7880**	**0.4359**	**0.5613**	0.2954
NB	0.4167	0.0150	0.1431	0.3196	0.1977	**0.0101**
NN	0.5060	0.0051	0.3276	0.3325	0.3300	0.0227
LR	0.6012	0.1632	0.4032	0.3867	0.3948	0.0490
RF	0.4842	0.0068	0.3371	0.3434	0.3398	0.0148
DT	0.4801	0.0179	0.3348	0.3234	0.3290	0.0122
SVM	0.5952	0.0186	0.3663	0.3390	0.3521	0.1421
GBDT	0.5908	0.1702	0.3951	0.3855	0.3902	0.5566

表 34 真实数据集-UKB5 的对比分析

指标	正确率	Kappa	精度	召回率	*F*1	时间
AR-MGEC	**0.7660**	**0.7012**	**0.7722**	**0.7516**	**0.7618**	0.2988
NB	0.1923	0.0530	0.3787	0.2494	0.3008	**0.0072**
KNN	0.3365	0.1587	0.3135	0.3102	0.3119	0.0187
LR	0.3365	0.1519	0.2508	0.3042	0.2749	0.1081
RF	0.3082	0.1294	0.3082	0.3025	0.3052	0.0162
DT	0.3337	0.1592	0.3396	0.3242	0.3317	0.0205
SVM	0.2212	0.0082	0.0907	0.2061	0.1260	1.7021
GBDT	0.4519	0.3054	0.4396	0.4273	0.4334	3.1913

表 35　真实数据集-UKB6 的对比分析

指标	正确率	Kappa	精度	召回率	$F1$	时间
AR-MGEC	**0.5826**	**0.2887**	**0.6863**	**0.4757**	**0.5619**	0.2836
NB	0.3651	0.0261	0.2639	0.3546	0.3026	**0.0129**
KNN	0.3693	0.0193	0.3397	0.3459	0.3428	0.0415
LR	0.4896	0.1506	0.6604	0.4115	0.5071	0.0572
RF	0.3983	0.0636	0.3755	0.3756	0.3755	0.0155
DT	0.3786	0.0494	0.3672	0.3706	0.3689	0.0132
SVM	0.4523	0.1047	0.3155	0.3869	0.3476	0.3381
GBDT	0.4425	0.1003	0.3852	0.3875	0.3863	0.6887

从以上横向对比实验数据可以看出，在 UCI 标准数据集上，大部分情况下本文的 AR-MGEC 算法还是有不错的性能收益，并且是在保证总执行时间保持在同类算法中较好的水准. 而在真实医疗数据 UKB 上，本文提出的算法保证了良好的分类性能，以及较好的执行效率，可见本文算法即便与同类算法相比较，在真实医疗数据集上，也有一定的价值. 因此，未来可以将 AR-MGEC 拓展到更多的多分类问题里，以此提供一个更加有效的模型去处理真实情况的多分类问题.

综上所述，AR-MGEC 算法通过符合人类认知的策略选择每一粒层的属性代表减少决策冲突的产生，并利用保留评分表的形式代替原有属性划分等价类，以此达到稳定、快速的集成分类效果. AR-MGEC 算法通过综合属性代表的分类意见来减少每一粒层中的冲突个数以达到快速地收敛，并且本算法利用粒层更新完成后的评分表，采取相对最优的策略，将最终未分类的对象划分到其最有可能的分类. 最终，本文模型在一定程度上提高了多粒度分类算法的效率、鲁棒性、分类正确率、Kappa 一致性系数，并在真实数据集下提升了宏观精度、召回率、$F1$-Score 值，以及面对当下不同分类问题的适用性.

5　结　论

本文为了减少多粒度分类算法的每个粒层中决策冲突的产生，提高分类效率以及实现对最终未分类对象的处理，结合集成学习和粒计算的思想，提出一个全新的基于属性代表的多粒度集成分类算法. 该算法可以综合考虑属性代表的不同分类意见，提高分类结果的正确率以及减少划分等价类的时间. 经过实验验证，AR-MGEC 算法相比于 S3WD 算法更加适用于结构化数据、数据量大、属性个数多且属性之间关系复杂的分类场景. 在未来的研究工作中，我们将进一步对该模型进行完善以及拓广. 我们希望本文提出的基于属性代表的多粒度集成分类算法能够推动多粒度分类算法在如医疗诊断之类的实际问题中具有更好的效果.

参　考　文　献

[1] Zhou Zhi-Hua. Machine Learning. Beijing: Tsinghua University Press, 2016(in Chinese)
(周志华. 机器学习. 北京：清华大学出版社，2016)

[2] Dietterich T G. Ensemble methods in machine learning// Proceedings of the 1st International Workshop on Multiple Classifier Systems (MCS 2000). Berlin, Germany: Springer-Verlag, 2000: 1-15

[3] Galar M, Fernández A, Barrenechea E, et al. An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. Pattern Recognition, 2011, 44(8): 1761-1776

[4] Yu Si-Hao, Guo Jia-Feng, Fan Yi-Xing, et al. Multi classifier ensemble algorithm based on knowledge-line memory. Chinese Journal of Computers, 2021, 44(3): 462-475(in Chinese)
(于思皓，郭嘉丰，范意兴等. 基于知识线记忆的多分类器集成算法. 计算机学报，2021，44(3)：462-475)

[5] West D, Dellana S, Qian J. Neural network ensemble strategies for financial decision applications. Computers & Operations Research, 2005, 32(10): 2543-2559

[6] Florez-Lopez R, Ramon-Jeronimo J M. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal. Expert Systems with Applications, 2015, 42(13): 5737-5753

[7] Zhang Y, Zhang B, Coenen F, et al. One-class kernel subspace ensemble for medical image classification. EURASIP Journal on Advances in Signal Processing, 2014, 2014(1): 1-13

[8] Muzammal M, Talat R, Sodhro A H, et al. A multi-sensor data fusion enabled ensemble approach for medical data from body sensor networks. Information Fusion, 2020, 53: 155-164

[9] Fraz M M, Remagnino P, Hoppe A. An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Transactions on Biomedical Engineering, 2012, 59(9): 2538-2548

[10] Yao Y Y. Three-way decisions with probabilistic rough sets. Information Sciences, 2010, 180(3): 341-353

[11] Fan Qin, Liu Dun, Ye Xiao-Qing. Cost-sensitive text sentiment analysis based on sequential three-way decision. Pattern Recognition and Artificial Intelligence, 2020, 33(8): 732-742(in Chinese)
(范琴，刘盾，叶晓庆. 基于序贯三支决策的代价敏感文本情感分析方法. 模式识别与人工智能，2020，33(8)：732-742)

[12] Yang Xin, Liu Dun, Li Qiu-Ke, et al. Sequential three-way sentiment analysis based on temporal-spatial multi-granularity. Pattern Recognition and Artificial Intelligence, 2020, 33(8): 743-752(in Chinese)

(杨新，刘盾，李楸柯等. 基于时空多粒度的序贯三支情感分析. 模式识别与人工智能，2020，33(8)：743-752)

[13] Yao J T，Azam N. Web-based medical decision support systems for three-way medical decision making with game-theoretic rough sets. IEEE Transactions on Fuzzy Systems，2014，23(1)：3-15

[14] Hu J，Yang Y，Chen X. A novel TODIM method-based three-way decision model for medical treatment selection. International Journal of Fuzzy Systems，2018，20(4)：1240-1255

[15] Ma X A，Yao Y Y. Three-way decision perspectives on class-specific attribute reducts. Information Sciences，2018，450：227-245

[16] Ren R S，Wei L. The attribute reductions of three-way concept lattices. Knowledge-Based Systems，2016，99：92-102

[17] Li W W，Jia X Y，Wang L，et al. Multi-objective attribute reduction in three-way decision-theoretic rough set model. International Journal of Approximate Reasoning，2019，105：327-341

[18] Cheng Y L，Zhang Q H，Wang G Y，et al. Optimal scale selection and attribute reduction in multi-scale decision tables based on three-way decision. Information Sciences，2020，541：36-59

[19] Xie Qin，Zhang Qing-Hua，Wang Guo-Yin. An adaptive three-way spam filter with similarity measure. Journal of Computer Research and Development，2019，56(11)：2410-2423(in Chinese)
(谢秦，张清华，王国胤. 基于相似度量的自适应三支垃圾邮件过滤器. 计算机研究与发展，2019，56(11)：2410-2423)

[20] Yao Y Y，Deng X. Sequential three-way decisions with probabilistic rough sets//Proceedings of the 10th IEEE International Conference on Cognitive Informatics and Cognitive Computing. Banff，Canada，2011：120-125

[21] Yao Y Y. Granular computing and sequential three-way decisions//Proceedings of the 8th International Conference on Rough Sets and Knowledge Technology. Berlin，Germany：Springer，2013：16-27

[22] Xu Y，Tang J X，Wang X S. Three sequential multi-class three-way decision models. Information Sciences，2020，537：62-90

[23] Yang X，Li T，Fujita H，et al. A sequential three-way approach to multi-class decision. International Journal of Approximate Reasoning，2019，104：108-125

[24] Li H X，Zhang L B，Huang B，et al. Sequential three-way decision and granulation for cost-sensitive face recognition. Knowledge-Based Systems，2016，91：241-251

[25] Ju H G，Pedrycz W，Li H X，et al. Sequential three-way classifier with justifiable granularity. Knowledge Based Systems，2019，163：103-119

[26] Zhang Q H，Huang Z K，Wang G Y. A novel sequential three-way decision model with autonomous error correction. Knowledge-Based Systems，2020，212(106526)

[27] Lang G M，Miao D Q，Fujita H. Three-way group conflict analysis based on Pythagorean fuzzy set theory. IEEE Transactions on Fuzzy Systems，2020，28(3)：447-461

[28] Xu Yi，Wang Xu-Sheng. Multi-category classification model and multilevel incremental algorithms. Journal of Frontiers of Computer Science and Technology，2019，13(8)：1431-1440 (in Chinese)
(徐怡，王旭生. 多类分类模型和多层次增量算法. 计算机科学与探索，2019，13(8)：1431-1440)

[29] Ye D，Liang D，Li T，et al. Multi-classification decision-making method for interval-valued intuitionistic fuzzy three-way decisions and its application in the group decision-making. International Journal of Machine Learning and Cybernetics，2020，12(1)：1-27

[30] Yao J T，Vasilakos A V，Pedrycz W. Granular computing：Perspectives and challenges. IEEE Transactions on Cybernetics，2013，43(6)：1977-1989

[31] Pawlak Z，Skowron A. Rudiments of rough sets. Information Sciences，2007，177(1)：3-27

[32] Wang Guo-Yin，Yao Yi-Yu，Yu Hong. A survey on rough set theory and applications. Chinese Journal of Computers，2009，32(7)：1229-1246(in Chinese)
(王国胤，姚一豫，于洪. 粗糙集理论与应用研究综述. 计算机学报，2009，32(7)：1229-1246)

[33] Wang Guo-Yin，Zhang Qing-Hua. Uncertainty of rough sets in different knowledge granularities. Chinese Journal of Computers，2008，31(9)：1588-1598(in Chinese)
(王国胤，张清华. 不同知识粒度下粗糙集的不确定性研究. 计算机学报，2008，31(9)：1588-1598)

[34] Zhang Chun-Xia，Zhang Jiang-She. A Survey of Selective Ensemble Learning Algorithms. Chinese Journal of Computers，2011，34(8)：1399-1410(in Chinese)
(张春霞，张讲社. 选择性集成学习算法综述. 计算机学报，2011，34(8)：1399-1410)

[35] Yu S X，Cao S M. Feature Selection and Classifier Ensembles：A Study on Hyper Spectral Remote Sensing Data [D]. University of Antwerp，Antwerp，Belgium，2003

[36] Breiman L. Bagging predictors. Machine Learning，1996，24(2)：123-140

[37] Sun Bo，Wang Jian-Dong，Chen Hai-Yan，et al. Diversity measures in ensemble learning. Control and Decision，2014，29(3)：385-395(in Chinese)
(孙博，王建东，陈海燕等. 集成学习中的多样性度量. 控制与决策，2014，29(3)：385-395)

[38] Sun B Z，Chen X T，Zhang L Y，et al. Three-way decision making approach to conflict analysis and resolution using probabilistic rough set over two universes. Information Sciences，2020，507：809-822

ZHANG Qing-Hua, Ph. D., professor, Ph. D. supervisor. His main research interests include rough sets, fuzzy sets, granular computing and uncertain information processing.

ZHI Xue-Chao, M. S. candidate. His main research interests include rough sets, ensemble learning and uncertain information processing.

WANG Guo-Yin, Ph. D., professor, Ph. D. supervisor. His main research interests include rough sets, granular computing, knowledge technology, data mining, neural network, and cognitive computing.

YANG Fan, Ph. D., assistant professor. His main research interests include deep causal learning, Bayesian causal inference, and biomedicine.

XUE Fu-Zhong, Ph. D., professor, Ph. D. supervisor. His main research interests include deep causal learning, bayesian causal inference, and biomedicine.

Background

Multi-classification has always been a common problem in real life. When dealing with multi-classification problems, researchers often simplify them into simple problems and then solve them. Now, mainstream multi-classification models are all handled according to this idea. Ensemble learning is a mainstream machine learning method, in the field of multi-classification, it is also a good method to improve classification performance. The multi-classifier system is the product of ensemble learning applied to multi-classification problems. As an important branch of the hybrid system, it aims to solve complex real-life multi-classification problems, such as medical diagnosis, email filtering, sentiment analysis, face recognition, etc. The multi-classifier system uses the training set to train each classification learner, and then integrates the classification results of the trained classifiers through voting and other integration methods to obtain a better result than a single classifier.

Granular computing, as an artificial intelligence algorithm that conforms to human cognitive thinking, is usually used for processing and analysis of complex problems. Different levels of granularity layers form a multi-layered granularity structure, and each granularity layer is composed of a group of particles with similar information granules. The multi-level granularity structure will form a multi-level decision process, and its combination with the three-way decisions constitutes a sequential three-way decision of asymptotic calculation. Sequential three-way decision is a multi-granularity classification algorithm, which realizes the dynamic decision process from coarse granules space to fine granules space asymptotically, and can effectively deal with complex multi-classification problems. However, when dealing with multi-classification problems, the sequential three-way decision model is prone to decision conflicts in the coarse granules space; the classification efficiency of the model is not high in the fine granules space; and it lacks a good strategy to deal with the final unclassified objects. This paper combines the ideas of ensemble learning and granular computing, and proposes a new multi-granularity ensemble classification algorithm based on attribute representation.

Compared with the classic multi-granularity classification algorithm, this algorithm has better classification performance and efficiency, especially in the face of data sets with a large amount of data and more attribute values. Under the coarse granules space, the model selects the attribute representation in the current granularity layer to construct an ensemble classifier through corresponding strategies, and reduces the generation of decision conflicts in the granularity layer by synthesizing the classification opinions of the attribute representation. Under the fine granules space, the model retains the classification opinions of the attribute by scoring, so as to improve the classification efficiency and performance in this granularity level. For the final unclassified object, the relatively optimal classification result is selected through the scoring table. Therefore, the model proposed in this article can solve the problem that other algorithms cannot be completely classified and cannot be correctly classified without sufficient information. It is especially suitable for use in medical prediction and other similar fields, because all the indicators in medical diagnosis cannot be obtained in the first time. Prioritizing selection of several attributes means that patients will be tested, and through further diagnosis, corresponding tests will be carried out. This paper is supported by the National Key Research and Development Program (No. 2020YFC2003502), the National Natural Science Foundation of China (No. 61876201), and the Chongqing Natural Science Foundation (No. cstc2019jcyj-cxttX0002).