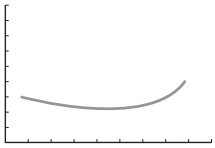
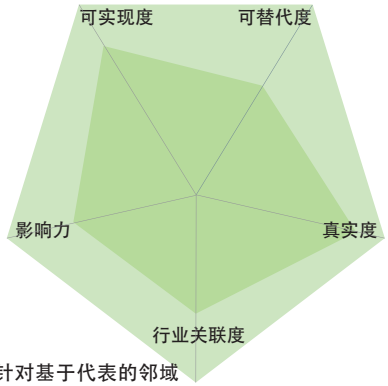
DOI：10.3969/j.issn.1001- 8972.2020.18.037





link appraisement industry

王 轩[1] 顾 峰[1] 孙远秋[2] 王水仙[3] 龚 涛[2] 冉志成[2]
1. 西南石油大学网络与信息化中心；2. 西南石油大学计算机科学学院；3. 西南石油大学经济管理学院

point

本文针对基于代表的邻域覆盖粗糙集分类算法的研究工作，进行了梳理、归纳和总结。在机器学习中的基于代表的分类领域起到推动作用。

如付诸现实将有助于降低分类的误分类代价、以及分类成本、提高分类精度，有助于算法的研究进展。

通讯作者：王轩（1991-）男，西南石油大学助理实验师，主要研究方向：机器学习、粗糙集、计算机网络；顾峰（1981-）男，西南石油大学实验师，主要研究方向：计算机网络与安全；孙远秋（1999-）男，西南石油大学本科在读，主要研究方向：计算机网络；王水仙（1999-），西南石油大学本科在读，主要学习方向：计算机网络，市场营销，企业管理；龚涛（1998-）男，西南石油大学本科在读，主要学习方向：后端开发；冉志成（1998-）男，西南石油大学本科在读，主要学习方向：安卓开发。
基金项目：国家自然科学基金资助项目（61902328）；四川省科技厅应用基础研究项目（2019YJ0314）；西南石油大学课外开放实验项目立项（KSP19G05）

代表选举分类综述

随着机器学习和大数据研究的进步与发展，分类问题作为其重要研究课题之一，也取得了极大的成功。2015 年张等在邻域覆盖粗糙集的基础上，提出基于代表的邻域覆盖粗糙集分类算法（Representative — based classification through covering — based neighborhood rough sets RC — CNRS）。RC — CNRS 算法实现了覆盖粗糙集在分类中的应用。自论文发表以来，基于代表的分类算法成为热门研究课题之一。

RC — CNRS 算法认为，距离最近的有效代表决定待分类样本类标签。RC — CNRS 算法利用邻域覆盖粗糙集，从训练集中选出代表，对待分类样本进行预测。算法分为代表选举和标签预测两个阶段。总体来看，目前对 RC — CNRS 算法的研究主要是对标签预测的研究。从学者们的具体研究进展来看，结合属性加权、代价敏感、交叉验证、主动学习等方面的研究工作不断涌现，形成了一批有价值的研究成果。接下来本文将结合学者们的具体研究工作，进行归纳总结和分析。

不同相似度比较

RC — CNRS 算法中，相似度计算是分类器训练阶段的核心内容。相似度决定了邻域和代表的选举过程。在 A comparison study of similarity measures for covering — based neighborhood classifiers 一文中，刘福伦采用多种相似度计算方式，对 RC — CNRS 算法的分类性能进行了研究。RC — CNRS 算法中采用的 Overlap 相似度，加上文献中选用的 Eskin，Goodall3，Good all4，IOF，OF 五种相似度，共 6 种相似度计算方法。刘福伦等的研究指出，不同相似度下，代表样本的最小相似度阈值不尽相同，这也导致算法分类精度不同。其工作从平均代表阈值和分类精度两个方面进行了对比实验。

分类精度对比实验表明，没有一种相似度计算方法能使 RC — CNRS 算法在所有数据上精度都有提升。选择相似度标准，要根据数据集自身特性进行选择。总体来说，Overlap，IOF，OF 三种相似度计算方法效果不错，可使 RC — CNRS 在大部分数据集上都取得较高的分类精度。特别是 Overlap，计算方法简单，且不需要先验知识。

平均代表阈值对比实验表明，不同相似度计算方法，造成分类精度不同的原因主要有两个方面。其一是改变了代表的最小相似度阈值，进而影响待分类样本与代表的距离计算。另一个原因是，不同相似度计算方法改变了代表数量，影响了待分类样本对有效代表的选择。

标签预测策略对比

RC — CNRS 算法的标签预测阶段，针对冲突对象，采用简单投票来处理。在邻域覆盖分类的两种加权策略这篇文章中，提出了属性加权策略和相似度策略。文章在 UCI 标准数据集上对三种策略进行对比实验。结果显示，两种策略对 RC — CNRS 性能有不同程度的提升，其中属性加权策略表现更好。

师彦文等在代表选举的分类策略对比这篇文章中，提出了基于规模策略和基于密度策略。在文中所用数据集上，基于密度策略最优，基于规模策略次之。总体来说，文中提出

策略对原算法分类精度提升幅度不大。

对比三篇文章的 5 种标签预测策略，笔者认为，属性加权策略分类精度最高，简单投票策略更简单、复杂度更低。从策略对比来看，属性加权策略对于 UCI 数据集中的游戏数据集有更好的分类性能，在游戏数据集上，属性加权策略的优势更加明显；从三篇文章与其他分类算法的对比实验，可以看出 RC － CNRS 算法对生命数据集的分类效果不错。

代价敏感

针对不同的分类问题，误分类对结果造成的影响是不同的。例如医生将一名感冒患者诊断为脑炎，和将脑炎患者诊断为感冒，都是误诊，可造成的影响却相差甚远。针对此类问题，闵帆教授在文献中提出了代价敏感，把分类器的性能衡量方式由分类精度转换为平均误分类代价。误分类代价向量往往由专家设定。

刘福伦等在代价敏感代表选举的邻域覆盖粗糙集分类方法一文中，将代价敏感与代表选举分类方法相结合。为代表选举分类方法的应用与发展做了重要工作。文中的类标签预测遵循两个原则：一是距离待分类样本近的代表影响类标签的预测；二是类标签的预测倾向于获取更低的误分类代价。分类模型的训练阶段，误分类代价不产生影响。文献的工作主要回答了两个问题：（1）对于代价敏感的代表选举分类算法，更大规模的训练集，会得到更低的平均误分类代价。（2）误分类代价向量的设置，会影响最终的平均误分类代价，并且其影响和训练集规模以及数据集实例分布有关。

与误分类代价相结合，使基于代表选举的邻域覆盖分类算法具有更好的分类性能。必要时舍弃一点分类精度，以保证平均误分类代价降至最低。分类精度和平均误分类代价追求的目的是一致的，都是为了提高分类器的性能。代价敏感使 RC － CNRS 更贴近了应用，对算法的应用和推广具有重要意义。

交叉验证与集成学习

集成学习是通过结合多个弱分类器来完成学习任务，往往能取得更好的性能。在基于代表的留一法集成学习分类中，研究者结合留一法进行交叉验证，运用集成学习的思想，来限制 RC － CNRS 算法中抽样不均对算法产生的影响。

通过文献的研究工作，分析可以得出以下结论：（1）抽样不均对 RC － CNRS 算法的分类精度有显著影响，交叉验证能有效提升分类性能。（2）对于较大数据集来说，RC － CNRS 算法受抽样不均的影响较小，此时采用留一法交叉验证对算法性能提升不明显，用集成学习的思想反而增加了算法的复杂度。此时就要综合考虑分类性能和算法复杂度来选择了。

同样也可以采用 k － fold 交叉验证，结合集成学习对 RC － CNRS 进行研究。在研究的时候要综合考虑集成后的分类性能和算法的复杂度。利用 k － fold 交叉验证进行研究时，k 值也是影响算法分类精度的重要因素之一。

主动学习策略

主动学习是通过现有知识，主动的选择要学习的补充知识的算法。实际的研究和应用中，获得类标签往往要付出昂贵的代价。主动学习能在一定程度上降低获取标签时付出的资源和代价。在基于覆盖约简的符号型数据分类及主动学习中，刘福伦提出了五种种主动学习策略。其中前三种是基于代表性的主动学习策略，另外两种为基于不确定性的主动学习策略。

刘福伦的研究工作，验证了主动学习策略在 RC － CNRS 上的有效性。实验结果显示，基于不确定性的异质 QBC 策略更适合应用于 RC － CNRS 的分类工作。它在实验所用数据集上，优于另外四种主动学习策略，并且在和其他经典的主动学习算法做对比实验时，也能取得更加稳定的分类效果。

小结及展望

本文总结了现有的对 RC － CNRS 算法的研究工作，并对各位学者的研究工作做了对比分析。对相似度计算的比较和研究，看出不同的数据集适用于不同的相似度计算方法。在运用时可以根据数据集特性选择对应的相似度计算方法。对标签预测策略所做的工作，分析出了不同标签预测策略下，RC － CNRS 算法适用的部分数据集类型。代价敏感与 RC － CNRS 结合的研究，对算法应用于实际场景有重要作用。可以根据误分类对不同行业产生的影响，由专家设置不同代价，进而将 RC － CNRS 算法应用于实际的生活生产中。利用交叉验证，结合集成学习对 RC － CNRS 算法的研究，看出抽样不均对算法的分类精度有一定的影响。结合主动学习的研究证明了主动学习策略在 RC － CNRS 算法上的有效性，并找到了一种效果很好的主动学习策略。

根据以上的研究，可以看出基于代表的分类算法，更适合代表性强的数据集。离散度较高的数据集要用 RC － CNRS 算法进行分类，最好采用交叉验证和集成学习的方法进一步处理。标签预测策略的研究，找到了一种能使算法分类精度提高的方法——采用属性加权然后再进行预测。

主动学习和代价敏感使 RC － CNRS 应用于实际场景成为现实。针对现实场景中，获取标签往往要花费较多的资源；分类错误也往往会产生不同的影响效果，结合主动学习和代价敏感的研究对算法的实际应用具有重要意义。进一步工作中可以重点研究主动学习和代价敏感在其中的应用。