打开 TAR 黑箱：利用模糊 ARTMAP 神经网络开发用于电子取证的可解释系统

Charles Courchaine
国立大学 美国
charles
@courchaine.dev

Ricky J. Sethi菲奇
堡州立大学美国国立大学美国
rickys@sethi.org

arXiv:2305.04237v1 [cs.IR] 7 May 2023

CCS 概念、】【
• **信息系统** → 检索结果评估；信息检索。

关键词
TAR、法律、电子取证、模糊 ART-MAP

ACM 引用格式：
Charles Courchaine 和 Ricky J. Sethi。2023 年。揭开 TAR 黑箱：使用模糊 ARTMAP 神经网络开发用于电子取证的可解释系统。载于《欧洲信息检索会议论文集（*ECIR '23*）》。*ACM*，纽约，美国纽约州，*3* 页。*https://doi.org/ XXXXXXX.XXXXXXX*

1 引言

技术辅助审查（TAR）利用信息检索系统，在语料库中找出所有或几乎所有相关文档，并帮助减少查找这些文档所需的人力 [7, 9, 20]。TAR 系统应用于对召回率要求较高的任务，例如电子取证、系统性文献综述、循证医学和信息测试集标注 [9, 20]。这些系统通常采用文档分类器和主动学习组件，以选择应由人工审查的文档[8, 21]。能够解释文档相关性预测是如何以及为何得出的 TAR 系统，是帮助律师履行对客户的道德义务，并使客户能够充分参与这一过程的重要工具 [12]。尽管可解释的 TAR 系统具有诸多益处，现有系统却无法说明文档为何被归类为相关，因此从业者通常仍将这些系统视为“黑箱” [7, 17]。

虽然已有少数研究尝试提高 TAR 系统的可解释性，但它们关注的是从文档中提取片段作为解释机制，而非直接解释相关性模型 [7, 17]。因此，我们考察了可解释的模糊 ART-MAP 算法。模糊 ART-MAP 算法学习到的模型可以直接从几何角度进行解读。

允许为个人或课堂使用而免费制作本作品的全部或部分数字或纸质副本，前提是这些副本不得用于营利或商业用途，且副本须附有本声明和首页的完整引文。对于本作品中归其他方而非 ACM 所有的内容，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、转载、发布到服务器或重新分发到列表，须事先获得具体许可和/或支付费用。请通过 permissions@acm.org 申请许可。
*ECIR '23，2023*年*4*月*2*日至*6*日，都柏林，爱尔兰© 2023 计算机协会。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/XXXXXXX.XXXXXXX

[4, 19] 或根据所使用的特征表示为一组模糊如果-那么规则 [5, 6], 。

我们对可解释的模糊 ART-MAP 算法在 TAR 领域中的性能进行了初步评估，发现其召回率和精确率表现稳健 [10]。基于这些初步结果所展现的优势，我们现已继续开展这项基础研究，具体包括：

- 执行超参数扫描以优化参数
- 针对 20Newsgroups、Reuters-21578、RCV1-v2 和杰布·布什电子邮件语料库评估该系统的召回率、精确率和 F_1，并
- 生成文档相关性的如果-那么规则

虽然这些语料库并非专门针对法律领域，但 RCV1-v2 和杰布·布什电子邮件语料库经常用于电子证据开示评估 [20, 22]，因为法律事务通常具有保密性 [7, 9]，其语料库无法获取。20Newsgroups 语料库通常作为基于 ART 的算法的测试语料库使用 [18, 19]；它和 Reuters-21578 语料库也常用于评估文本分类算法 [1]。

2 模糊 ART-MAP

自适应共振理论（ART）描述了大脑如何在非平稳世界中学习和预测 [14]。该理论对大脑如何快速学习新信息，同时不遗忘先前学到的信息进行了建模。ART 已应用于众多神经网络架构，支持监督学习、无监督学习和强化学习应用 [3]。模糊 ART 是 ART 的一种神经网络算法实现，它采用模糊集理论中的运算符，具体而言是模糊与运算符，以处理实值特征 [4]。模糊 ART 算法的监督版本是模糊 ART-MAP 算法，可在输入与类别之间建立映射。通过在模糊 ART-MAP 神经网络算法中整合模糊集理论与 ART 动态机制，可以对所学习的模型进行多种解读。模型所学到的内容随后可以表示为模糊的“如果—那么”基于文本的规则，或以几何形式呈现 [4, 13]。

然而，要利用这种几何解释，输入必须采用补码编码。补码编码是使用模糊 ART-MAP[4] 时的一种归一化方法，即将输入向量 $\boldsymbol{x}$ 与其补码 $\boldsymbol{x}$连接起来，得到一个 $\boldsymbol{I} = [\boldsymbol{x}, \boldsymbol{x}]$ 的输入。因此，模糊 ART-MAP算法学习到的类别可以解释为n维超矩形 [4, 19]。在解释该模型时

从几何角度看，向量前半部分（即非补码编码部分）学习到的权重构成超矩形的一个顶点，而向量后半部分（即补码编码部分）则构成另一个顶点，如图 1 所示。

3 方法

对于 20Newsgroups、Reuters-21578、RCV1-v2 和杰布·布什电子邮件语料库，我们在较小的语料库中使用了词频-逆文档频率特征，并在所有语料库中使用了 300 维版本的 GloVe 和 Word2Vec 向量化表示。20Newsgroups 中的所有主题、Reuters-21578 中的 120 个主题，以及 RCV1-v2 和杰布·布什电子邮件语料库中的各 30 个主题均用于评估；由于内存限制，RCV1-v2 和杰布·布什语料库分别按每个 [22] 下采样至 20% 和 50%，同时保留每个主题的总体普遍性。对于每个主题，无论语料库大小如何，模糊 ART-MAP 算法均使用 10 个相关文档和 90 个不相关文档进行训练；对于较小的语料库，审阅以每批 100 个文档进行，对于较大的语料库，则以每批 1,000 个文档进行。当算法预测语料库未评估部分中不再有相关文档时，便结束对每个主题的文档审阅。模糊 ART-MAP 算法经过修改后，会报告与被预测为相关的文档相关的模糊子集隶属度 [4, 16]，随后使用该模糊子集隶属度对文档进行排序，以开展主动学习。基于对模糊 ARTMAP 神经网络算法超参数的扫描结果（评估了警戒参数（ρ）和学习率（β）的不同组合），警戒参数设为 .95，并选用了 1.0 的快速学习率。

针对 20Newsgroups 数据集中的个人电脑硬件类别文档，我们生成了一条用于词频-逆文档频率向量化的如果-那么规则概念验证，见表 1。词频-逆文档频率特征以斜体显示，流行程度以粗体显示。在此示例中，流行程度被量化为三个等级：罕见、较为普遍和高度普遍。此外，

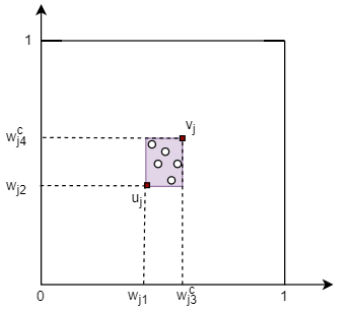

图1：采用补码编码且输入为二维时，由权重向量 j^{th} 表示的类别可在几何上解释为一个矩形，其顶点为 $\mathbf{u}_j$ 和 $\mathbf{v}_j$，其中 $\mathbf{u}_j$ 对应向量的第一和第二个位置，而 $\mathbf{v}_j$ 对应补码编码后的第三和第四个位置。矩形内部的圆圈表示落在该类别边界范围内的输入。

图 1 展示并讨论了几何解释的一个示例。

4 结果与讨论

考虑所有语料库和向量化方式，基于模糊 ART-MAP 的系统有 31% 的时间达到了 100% 的召回率，并有 67% 的时间达到了建议的 75% [15] 或更高召回率下限，这一点从表 2 中的召回率、精确率和 F_1 的中位数可以看出。对于 Reuters-21578 和 20Newsgroups 语料库，不同向量化器之间的召回率差异具有统计显著性，这基于 Friedman 检验 [11]，$p < .001$（$\chi^3(2)=25.09$ 和 $\chi^3(2)=34.9$）。事后 Nemenyi 检验 [11] 表明，词频-逆文档频率与 GloVe 和 Word2Vec 之间存在差异，其平均差异和统计显著性如表 3 所示。根据平均差异，词频-逆文档频率相较于 GloVe 和 Word2Vec 的向量化具有实际意义。在 RCV1-v2 或杰布·布什电子邮件语料库中，GloVe 与 Word2Vec 之间不存在统计或实际差异。

这些结果表明，召回率整体表现稳健，尤其是采用词频-逆文档频率向量化时。除杰布·布什电子邮件以及 20Newsgroups 的 GloVe 向量化外，召回率中位数均达到 75% 或更高。在 20Newsgroups 和杰布·布什电子邮件这两类较为非正式的语料库中，GloVe 和 Word2Vec 特征的表现不尽理想。不过，这可能是因为词频-逆文档频率具有语料库特异性，而 GloVe 和 Word2Vec 使用的是现成的词汇表。这表明，生成针对特定语料库的 GloVe 和 Word2Vec 表示，可能比使用默认词汇表效果更好。未来的研究可以优化词频-逆文档频率向量化的 If-Then 规则生成，并为 Word2Vec 和 GloVe 向量化提供文本及图形解释。此外，探索针对特定语料库的 Word2Vec 和 GloVe 版本，或许能使召回率达到词频-逆文档频率的水平，从而提供一种效率更高且同样稳健的方案。

虽然如果-那么规则和图形表示是公认的可解释性方法，但目前对于可解释人工智能领域尚无公认的定量指标[2]；此外，也没有针对电子证据发现 TAR 中现有可解释性尝试的定性或定量用户研究 [7, 17]。因此，这很可能是未来另一项富有成效的研究方向。

结论：这项基础性研究进一步证实了在 TAR 领域中使用模糊 ARTMAP 神经网络作为分类算法的可行性。尽管在提升召回率表现和可解释性方面仍存在进一步研究的空间，但本研究取得的稳健召回率结果，以及针对词频-逆文档频率向量化所开展的如果-那么规则概念验证，均有力地表明，基于模糊 ARTMAP 的 TAR 系统有望成为传统“黑箱”TAR 系统的一种切实可行且具有可解释性的替代方案。

参考文献

[1] Berna Altınel 和 Murat Can Ganiz。2018。语义文本分类：过去与近期进展综述。《信息处理与管理》54, 6（11月 2018 1129 1153 10 1016 2018 08 001），–。https://doi.org/。/j.ipm.。[2] 。Alejandro Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Ben- netot、Siham Tabik、Alberto Barbado、Salvador Garcia、Sergio Gil-Lopez、Daniel Molina、Richard Benjamins、Raja Chatila 和 Francisco Herrera。2020。可解释人工智能（可解释人工智能）：概念、分类体系、机遇与挑战——迈向负责任人工智能。《信息融合》58（2020年6月），82–115。https://doi.org/10.1016/j.inffus.2019.12.012

表 1：个人电脑硬件的规则输出摘录

文档相关	
IF	*advance* 在文档中**很少**普遍存在
and	*apr* 在文档中**很少** 出现
and	*bogus* 在文档中**很少**普遍存在。
and	*browning* 在**很大程度上**的文档中普遍存在
and	加州州立大学 在文档中**很少**普遍存在
and	驱动器 在文档中**较为**普遍
...	

表 2：按语料库-向量化器划分的中位数指标

语料库	向量化器	召回率	精确率	F_1
20新闻组	GloVe	0.57	0.522	0.434
	Word2Vec	0.772	0.41	0.523
	词频-逆文档频率	0.94	0.367	0.53
杰布·布什电子邮件	GloVe	0.622	0.07	0.125
	Word2Vec	0.593	0.055	0.098
RCV1-v2	GloVe	0.764	0.211	0.324
	Word2Vec	0.752	0.187	0.292
Reuters-21578	GloVe	0.909	0.384	0.526
	Word2Vec	0.931	0.514	0.624
	词频-逆文档频率	0.92	0.733	0.759

表 3：平均召回率差异

	Reuters-21578	20Newsgroups
tf-idf-GloVe	0.085**	0.451**
tf-idf-Word2Vec	0.069*	0.171**

*p < .05，**p < .01

[3] Leonardo Enzo Brito da Silva、Islam Elnabarawy 和 Donald C. Wunsch。2019。面向工程应用的自适应共振理论神经网络模型综述。神经网络120（2019年12月），167–203。https://doi.org/10.1016/ j.neunet. 2019 09 012 。。Gail A. Carpenter、Stephen Grossberg、Natalya Markuzon、John H. Reynolds 和 David B. Rosen。1992年9月。模糊 ART-MAP：用于模拟多维映射增量式监督学习的神经网络架构。*IEEE*神经网络汇刊3, 5（1992年9月），698–713。https://doi.org/10. 1109/72.159059[5] Gail A Carpenter 和 Ah-Hwee Tan。1993。规则提取、模糊 ART-MAP 与医学数据库。载于世界神经网络大会论文集。Erlbaum Associates，美国俄勒冈州波特兰，501–506。[6] Gail A. Carpenter 和 Ah-Hwee Tan。1995。规则提取：从神经架构到符号表示。*Connection Science* 7, 1（1995年1月），3–27。https://doi.org/ 10 1080 09540099508915655 。/ Rishi Chhatwal、Peter Gronvall、Nathaniel Huber-Fliflet、Robert Keeling、Jian- ping Zhang 和 Haozhen Zhao。2018。法律文档审查中的可解释文本分类：可解释预测编码案例研究。载于*2018 IEEE*国际大数据会议（*Big Data*）。IEEE，美国华盛顿州西雅图，1905–1911。https://doi.org/ 10 1109 2018 8622073 。/BigData。。Rishi Chhatwal、Nathaniel Huber-Fliflet、Robert Keeling、Jianping Zhang 和 Haozhen Zhao。2017。法律文档审查中主动学习策略的实证评估。载于*2017 IEEE*国际大数据会议（*Big Data*）。IEEE，美国马萨诸塞州波士顿， 1428 1437 10 1109 2017 8258076 –。https://doi.org/。/BigData。。Gordon V. Cormack 和 Maura R. Grossman。。技术辅助审查中持续主动学习的自主性与可靠性。*arXiv:* *1504 06868 .cs*（4月）。[ ] 2015 1504 06868 [ ] arXiv: . cs[10] Charles Courchaine 和 Ricky Sethi, J. 2022。模糊法律：旨在创建新型可解释技术辅助审查系统的探索。载于 *2022*年*IEEE*国际大数据会议（*Big Data*）. IEEE，日本大阪， 1218 1223 10 1109 55660 2022 10020503 – . https://doi.org/ . / 2006[11] BigData . . Janez Demsar. . 多个数据集上的分类器统计比较。《机器学习研究杂志》7（2006年12月），1–30。[12] Seth Katsuya Endo. 2018. 技术不透明性与程序不公。《波士顿学院法律评论》 59, 3 2018 822 875 59 3 2（3月），– . 2020[13] https://doi.org/bclr/vol /iss / Stephen Grossberg. . 通往可解释人工智能与自主自适应智能的道路：深度学习、自适应共振，以及关于感知、情绪和行动的模型。《神经机器人学前沿》14（2020年6月），36。https://doi.org/ 10 3389 2020 00036 . /fnbot. . 2021[14] Stephen Grossberg. . 迈向自主自适应智能：基于大脑如何形成心智的神经模型进行构建。《*IEEE*系统、人类与控制论汇刊：系统》51, 1（2021年1月），51–75。https://doi.org/10.1109/ TSMC.2020.3041476[15] Robert Keeling、Rishi Chhatwal、Peter Gronvall 和 Nathaniel Huber-Fliflet. 2020. 人类对抗机器：重申人类律师在法律文档审查中的优势，并考察算法证据开示方法的局限性。《里士满法律与技术杂志》26, 3（2020），65。[16] Bart Kosko. 1986. 模糊熵与条件化。《信息科学》40, 2（1986年12月），165–174。https://doi.org/10.1016/0020-0255(86)90006-X [17] Christian J. Mahoney、Jianping Zhang、Nathaniel Huber-Fliflet、Peter Gronvall 和 Haozhen Zhao. 2019. 法律文档审查中可解释文本分类的框架。载于*2019*年*IEEE*国际大数据会议（*Big Data*）。IEEE，美国加利福尼亚州洛杉矶，1858–1867。https://doi.org/10.1109/ BigData 47090 2019 9005659 . . Dušan Marček 和 Michal Rojček. 2017. ART神经网络中的类别激增问题。《匈牙利理工学报》14, 5（2017），15。[19] Lei Meng、Ah-Hwee Tan 和 Donald C. Wunsch II. 2019. 用于社交媒体分析的自适应共振理论（*ART*）。Springer International Publishing，瑞士尚贝里， 45 89 10 1007 978 3 030 02985 2_3 – . https://doi.org/ . / - - - - Eugene Yang、David D. Lewis 和 Ophir Frieder. . 一种在技术辅助审查中结合关键词与训练数据的正则化方法。载于第十七届人工智能与法律国际会议论文集。ACM，加拿大魁北克省蒙特利尔，153–162。https://doi.org/10.1145/3322640. 3326713[21] Eugene Yang、David D. Lewis 和 Ophir Frieder. 2021. 在法律文档审查工作流程中实现成本最小化。*arXiv:2106. 09866 [cs]*（2021年6月）。https: //doi.org/ 10 1145 3469096 3469872 2106 09866 [ ] . / . arXiv: . cs Eugene Yang、Sean MacAvaney、David D. Lewis 和 Ophir Frieder. 2021. Goldilocks：针对技术辅助审查对BERT进行恰到好处的调优。*arXiv:2105.01044 [cs]*（2021年5月）。arXiv:2105.01044 [cs]