什么人　工智能或许能帮助我们了解人类语言的起源 e

Kilpatrick, Alexander
名古屋商科大学 本研究尚未经过同行评审。请持怀疑态度。

摘要

本研究探讨了人工智能与声音象征相结合的研究中出现的一种有趣模式。在这些研究中，研究人员训练监督式机器学习算法，根据指称对象名称的发音对样本进行分类。机器学习算法能够高效地学习声音象征，但往往会偏向其中一个类别。这种模式是：当某个类别可以说代表着更大的威胁时，算法往往会过度预测样本属于该类别。本文基于错误管理理论提出一项假设：这种现象可能是适应谨慎行为偏好的证据。为检验这一假设，研究人员利用中文、日语和韩语宝可梦名称中的发音构建极端梯度提升（XGBoost）模型，并观察分类误差分布。

关键词：机器学习、语言演化、声音象征、XGBoost、宝可梦语音象征学

1. 引言

在过去一百年里，关于人类语言进化的理论研究一直被语言起源的手势优先假说 [1] 以及“声音与意义之间的关系是任意的”这一哲学观点[2]所掩盖。关于后者，越来越多的研究揭示了人类语言中存在系统性的声音象征关系。声音象征与任意性之间的争论至少可以追溯到柏拉图的对话篇《克拉底鲁篇》；该篇详细记录了哲学家之间的一场辩论，苏格拉底最终得出结论：形式与意义之间的关系有时是任意的，有时则不是[3]。现代研究表明，在这一点上苏格拉底是正确的：人类语言中的声音象征是随机的，而非规定性的；也就是说，它呈现出一种可进行统计分析的随机概率分布，但未必适用于所有情况[4]。有趣的是，研究发现，许多系统性的声音象征关系在不同语言中都成立。例如，一项研究[5] 使用无意义词刺激材料，考察中文、英语、日语和韩语使用者对与元音相关的大小判断。研究发现，在所有这些语言中，[i] 都与较小的物体相关，而[a] 则与较大的物体相关。这种模式通常用频率编码假说 [6] 来解释，该假说认为，大小会反映在声音的声学属性上。不过，有些跨语言声音象征关系可能并不反映大小，而是反映形状。例如，近期一项研究发现，在接受测试的25种语言中，有17种语言都稳定地呈现kiki/bouba效应——即无意义词kiki与尖刺形状相关，而无意义词bouba与圆形相关[7]。

另一种跨语言的声音象征模式，可能既不反映大小也不反映形状，在宝可梦语音象征学中也能找到。宝可梦语音象征学是对名为宝可梦的电子游戏角色名称中的声音象征模式进行分析。宝可梦已被翻译成多种语言，而且宝可梦的属性在不同语言中保持一致。Shih等人 [8] 有一项有趣的发现：宝可梦名称中的声音象征效应，在那些对游戏内生存能力至关重要的特征上更强。例如，战斗属性和性别都呈现出声音象征模式，但战斗属性上的模式比性别上的更强，因为性别对游戏内表现影响甚微。宝可梦的另一项属性是友好度，反映了每只宝可梦天生有多友善。在对中文、英语、法语、德语、日语和韩语宝可梦名称中的友好度与塞音之间的关系进行考察时， [9]，名称中 [p] 出现的次数，在所有语言中都是高友好度的重要预测指标；而名称中 [g] 出现的次数，则是在塞音存在浊音性音系对立的所有语言（即中文和韩语除外）中，低友好度的重要预测指标。这些模式的跨语言特性，以及因变量显然与大小和形状无关，促使作者谨慎地得出结论：这或许是存在一种共享声音象征系统的证据，而该系统可能是某种残留警报叫声系统的结果。这一观点也得到趋同进化效应的支持——不同物种在相似的环境压力下会进化出相似的特征——以及警报叫声系统在动物界的普遍存在。例如，撒哈拉以南非洲的非人灵长类动物，尤其是长尾猴 [10]

以及绿猴 [11]——拥有复杂的警报叫声系统，能够向同类传达捕食者的信息。

近期的宝可梦研究采用监督式机器学习算法，进一步考察宝可梦语料库中的模式，并测试这些模型的学习效率。研究人员使用随机森林（RF）算法，根据宝可梦名称在中文、日语和韩语中的构成音，将宝可梦分类为进化前和进化后两类 [12]。在宝可梦系列中，进化是角色“升级”的一种方式。宝可梦进化后，名称和外形都会改变，同时也会变得更强。研究发现，初始模型存在过拟合，因此研究人员使用分别随机抽取的子集构建了多个独立的随机森林，并对数据集进行交叉验证。所有随机森林的分类准确率都高于随机水平，但误差分布和特征重要性计算呈现出一个有趣的模式：进化前宝可梦的错误率往往较高，而对分类准确率有重要影响的特征往往在进化后类别中占比过高。表 1 展示了中文随机森林的混淆矩阵，表明该模型倾向于将样本过度分类到进化后类别。

		分类	
		进化前	进化后
样本	进化前	479	444
	进化后	300	667

表 1：中文宝可梦随机森林算法的混淆矩阵 [12]。

		分类	
		女性	Male
样本	女性	9539	12623
	Male	6461	28918

表 2：中文名字随机森林算法的混淆矩阵 [13]。

类似实验中也观察到了这种模式[13,14]：这些实验基于构成人类名字的声音构建随机森林，用以探究中文 [13] 和日语 [14]名字中的性别声音象征。表 2 展示了基于中文名字构建的随机森林的混淆矩阵。

尽管 [12] 和 [13] 检验的是双尾假设，但从单尾假设的角度解读本研究中的这些发现，或许能为声音象征带来新的见解。在单尾检验中，表1中的进化后类别和表2中的男性性别，可能分别被视为各分类类别中威胁性更强（因此也更有可能引发警报叫声）的一类。

那么，那些来自非威胁类别、却被归入威胁类别的样本，就属于假阳性（FP）错误；而那些具有威胁性的样本、却被归入非威胁类别的，则属于假阴性（FN）错误。按照这种解读，可以通过根据测试子集中的分布偏斜调整假阳性错误所占百分比，来计算假阳性错误率。以表1为例，进化后样本（967）比进化前样本（923）多出4.8%；假阳性错误率的计算方式是将少数类错误中的假阳性错误百分比乘以104.8%，以校正偏斜。表1所列算法的假阳性错误率为58%，表2所列算法的假阳性错误率为76%。需要指出的是，这些是汇总结果，只有少数随机森林的假阳性错误率不超过50%。有趣的是，重要特征的分布偏斜往往指向威胁类别。换言之，模型在做出分类决策时所依赖的重要特征，在进化后类别和男性类别中的分布更多。例如，表1所示中文宝可梦随机森林的最重要特征是降调、/ŋ/、/t͡ɕ/、/ɕ/和/e/。其中，只有/ɕ/在进化前类别中的分布更多；而只有降调和/ŋ/——这两个特征在进化后样本中占比过高——呈现出显著的特征重要性。表3报告了五个多个随机森林的模型准确率和假阳性错误率，所有模型都呈现出这一典型模式。这些只是部分示例；[12] 中还可以找到更多符合这一模式的例子。

数据集	因变量	准确率。	FP%	Ref.
中文宝可梦	进化	60%	58%	[12]
中文名字	性别	72%	76%	[13]
日本宝可梦	进化	68%	52%	[12]
日文名字	性别	76%	54%	[14]
韩国宝可梦	进化	60%	60%	[12]

表3：五个多个随机森林的模型准确率和假阳性错误率（假阳性率）。

在威胁传达中偏向假阳性错误的误差分布偏斜，可能具有进化优势。错误管理理论 [15], 以信号检测理论为基础，是一种用于评估生物体如何在不确定情况下做出决策的框架。EMT认为，进化选择会青睐那些选择能将错误代价降至最低的行为的生物体。例如，绿猴会针对豹、鹰和蛇发出声学上不同的警戒叫声[10]。幼年绿猴似乎天生就有这种倾向

从出生起就会大致将其他物种划分为捕食者和非捕食者，但它们往往会犯错，对鸽子和疣猪等非捕食者发出警戒叫声 [16]。幼年长尾猴将鹰警报发给鸽子，就是假阳性错误的一个例子；与幼猴犯下假阴性错误、未能发出捕食者出现的信号相比，这种错误的代价可能较低。从某些声音象征元素会在人类语言中传达威胁这一假设出发，错误管理理论可能解释了为何在机器学习算法中观察到假阳性错误占比过高。

2. 方法与材料

本研究使用从在线资料库中获取的日语、韩语和中文宝可梦名称，构建 XGBoost 模型 [17] 。本研究使用 XGBoost 模型而非随机森林模型，测试误差分布偏斜是否也会出现在随机森林算法之外。我们将名称转换为每个名称中各语音特征出现次数的计数。由此得到一个包含 898 个样本的数据集，其特征大多为空值；研究发现，这会导致基于决策树的算法过拟合 [12]。为解决这一问题，我们采用 k 折交叉验证（k = 3），即为每个模型构建三次迭代。每次迭代都使用不同的子集来训练（A+B、A+C 和 B+C）和测试（C、B 和 A）算法。报告的结果是这三次迭代的汇总结果，但统计假设检验则基于各次迭代的单独结果进行。对于每种语言，我们测试四个因变量：攻击（衡量每只宝可梦攻击威力的指标）、防御（衡量每只宝可梦防御能力的指标）、身高和体重。选择这些变量是因为理论上，每项变量值的增加都应代表威胁加大，但不同变量所代表的威胁增幅应有所不同。例如，战斗相关参数（攻击和防御）的增加，所代表的威胁增幅应大于体型相关参数（身高和体重）的增加。需要注意的是，在这些名称所对应的电子游戏中，攻击和防御会直接影响战斗，而身高和体重在游戏中不会产生任何影响。样本分类依据每个连续变量的中位数二分法确定，中位数样本从数据集中剔除；同时，对于每个因变量，通过随机剔除多数类别中的样本，使各类别保持平衡。本研究检验以下假设：假设 1：所有模型中的假阳性错误均大于 50%；假设 2：模型对战斗参数进行分类的模型，其假阳性错误将高于对体型相关参数进行分类的模型。数据可通过以下网址获取：https://tinyurl.com/mw9u43ya

3. 结果

所有机器学习算法的准确率均超过 50%。战斗模型的 FP 均超过 50%，但并非所有体型模型都是如此。为检验假设1，我们针对 50% 的随机概率计算了仅含截距的线性模型。为检验假设2，我们采用了简单线性模型。所有模型均使用各次独立迭代的数据（共 36 个算法），而非表 4 中呈现的汇总结果。

	攻击		防御	
	准确率	FP%	准确率	FP%
日语	59.15%	51.93%	59.16%	55.96%
中文	57.04%	57.08%	55.66%	53.35%
韩语	56.10%	55.47%	54.07%	55.64%
平均值	57.43%	54.83%	56.30%	54.98%
	身高		体重	
	准确率	FP%	准确率	FP%
日语	63.03%	48.82%	62.33%	47.97%
中文	59.27%	44.17%	57.51%	54.05%
韩语	57.77%	48.77%	55.38%	51.50%
平均值	60.03%	47.26%	58.41%	51.17%

表 4：XGBoost 模型的汇总准确率和假阳性错误率。

假设1）在四个因变量中，只有攻击模型显示出对FP错误分布的显著交互作用；$t(8) = 2.6$，$p = 0.031$。防御（$p = 0.145$）、身高（$p = 0.522$）和体重（$p = 0.699$）模型的结果均不显著。通过结合战斗变量和体型变量计算得出的仅含截距的模型显示，只有战斗变量显著；$t(17) = 2.8$，$p= 0.012$，而体型变量则不显著（$p = 0.757$）。假设2）简单线性模型显示，战斗变量中的 FP（M = 55%，SD = 7%）与体型变量中的 FP（M = 49%，SD =10%）存在显著差异；$t(34) =25.55$，$p < 0.001$。

战斗模型中假阳性错误增加的一种可能解释是越长越强原则 [19]，即研究发现，实力较强的宝可梦通常名字也更长。为探究这一原则可能产生的影响，我们进行了一系列简单线性回归分析，考察四个连续变量（包括此前省略的样本）与中文名称中除声调外的特征数量之间的关系。除中文数据集中的身高（$p = 0.67$）和体重（$p = 0.70$）外，所有线性回归分析均显示出显著正相关（$p < 0.05$）。综合分析结果见表5。

可变自由度		*F*	*p*	R^2
攻击	1, 2693	58.88	2.33E-14	0.021
防御	1, 2693	49.36	2.69E-12	0.018
身高	1, 2693	17.95	2.35E-05	0.007
体重	1, 2693	27.40	1.79E-07	0.010

表5：参数与名称长度关系的简单线性回归分析。

4. 讨论

与假设1相反，并非所有因变量的假阳性错误都超过50%。只有攻击模型与随机结果的差异达到显著水平；不过，攻击模型与防御模型合并后，战斗模型的结果达到显著水平。身高和体重对游戏内表现没有影响，这或许可以解释为何体型模型没有呈现战斗模型的模式。与假设2一致，战斗模型的假阳性错误率显著高于体型模型的假阳性错误率。尽管采用了不同的机器学习算法，本研究中的战斗模型所呈现的模式仍与表3中的模型相符。这些发现支持核心假设：依据威胁进行分类的算法中，假阳性错误所占比例会偏高。在警戒呼叫系统中，这意味着模糊不清的发声更可能被感知为警戒叫声，而非非警戒叫声。尽管这会导致更多假阳性错误，但这类错误的代价低于假阴性错误，因此这种策略在进化上具有优势。任何基于发声、且存在于现代语言中的进化适应，都意味着人类语言并非完全以动觉模态进化而来；手势优先假说主张人类语言完全以动觉模态进化，而这一观点近期受到批评 [20]。

尽管模型本身并未给出明确解释，但假阳性错误分布偏斜最可能的解释是：在威胁分类中起重要作用的特征，比在非威胁分类中起重要作用的特征更容易出现在非威胁样本中，而后者则不太容易出现在威胁样本中。传统统计分析很难观察到这种效应，因为除非专门对此进行检验，否则它很可能会表现为整个模型的效应量降低，具体取决于所采用的检验。[12], [13], [14], 中的交叉验证方法以及本研究中的交叉验证方法还有一项次要作用：它们会产生多个结果，可使用传统方法进行分析，从而检验是否存在显著的分布偏斜。

对此一种可能的替代解释是越长越强原则。名称越长，空值越少。基于决策树的算法构建许多决策树——这些决策树由随机自助采样样本构成 [21]。每棵树都由节点组成，每个节点都会测试随机选择的特征 [22]。如果节点只选择了具有空值的特征，那么该节点就没有作用，终端节点的结果必然要基于其他节点。有三个问题与这一解释有关：首先，目前尚不清楚这如何会导致算法过度分类到威胁类别；其次，控制每个节点所选特征数量的超参数会根据数据集进行调优，从而减少无作用节点；最后，线性回归分析显示，除中文外，所有变量与名称长度之间都存在显著的正相关。尽管长度在战斗参数中的影响更强，但合理的推断是，如果这一解释准确，那么名称长度增加应会使所有变量的假阳性错误增加，而不仅仅是战斗变量。

萨根标准认为，非凡的主张需要非凡的证据。本研究使用的是有限的电子游戏角色名称语料，而非自然语言，且仅考察东亚语言。与随机森林算法类似，XGBoost算法的内部运作机制难以观察；因此，算法为何偏向于产生假阳性错误并不完全清楚。本研究及其核心假设最显著的问题或许在于，它将机器学习算法拟人化了，也就是说，它假设那些导致人工智能增加假阳性错误的模式在人类身上也会产生相同影响。作者建议，在得出任何确切结论之前，应对这一领域开展更多研究。

4. 结论

本研究提出了 XGBoost 模型算法，旨在根据声音象征将宝可梦名称归类为攻击、防御、身高和体重较高或较低的类别。研究检验这些算法是否会偏向较高类别，因为有观点认为较高类别代表更强的威胁。结果发现，只有对战斗变量进行分类的模型，其假阳性错误分布才显著增加。这可能表明，人类语言具有一种内在特征，使人们更常将声音象征信息感知为威胁。从错误管理理论的角度来看，这在进化上具有益处，因为假阳性错误的代价低于假阴性错误。尽管仍建议开展更多研究，但这一发现有望显著影响我们对声音象征和人类语言起源的认识。

5. 参考文献

[1] 孔狄亚克，É.（2001）。《人类知识起源论》（H. Aasleff 译）。英国剑桥：剑桥大学出版社。[2] 萨丕尔，E. 1929。《语音象征研究》。《实验心理学杂志》12，225-239。[3] 塞德利，D.（2003）。柏拉图《克拉底鲁篇》。剑桥大学出版社。[4] 河原，S.、胜田，H.和熊谷，G.（2019）。使用最大熵模型解释声音象征的随机性。《开放语言学》，5(1)，109-120。[5] 篠原，K.、河原，S. 2010。跨语言声音象征研究：大小意象。《伯克利语言学会年会论文集》36(1)，396-410。[6] 奥哈拉，J. 1994。频率编码是声音象征性地运用音高的基础。载于《声音象征》，L. 欣顿、J. 尼古拉、J. J. 奥哈拉编，第325-347页。英国剑桥／纽约：剑桥大学出版社。[7] Ćwiek，A.、福克斯，S.、德拉克斯勒，C.、阿苏，E. L.、德迪乌，D.、Hiovain，K.……和温特，B.（2022）。布巴／基基效应在不同文化和书写系统中均十分稳健。《英国皇家学会哲学汇刊B辑》，377(1841)，20200390。[8] 施，S. S.、阿克曼，J.、赫马林，N.、因克拉斯，S.和卡维茨卡娅，D.（2018）。宝可梦名称：声音象征与宝可梦名称研究。《美国语言学会论文集》，3(1)，42-1。[9] 为评审而匿名处理。[10] 塞法斯，R. M.、切尼，D. L.和马勒，P.（1980）。长尾猴警报叫声：自由活动的灵长类动物中的语义交际。《动物行为》，28(4)，1070-1094。[11] 普赖斯，T.、韦德维茨，P.、切尼，D.、塞法斯，R.、哈默施密特，K.和菲舍尔，J.（2015）。重新研究绿猴：对警报叫声结构和情境特异性的定量分析。《科学报告》，5(1)，1-11。[12] 为评审而匿名处理。[13] 为评审而匿名处理。[14] 为评审而匿名处理。[15] 哈塞尔顿，M. G.和巴斯，D. M.（2009）。错误管理理论与错误信念的演化。《行为与脑科学》，32(6)，522-523。[16] 塞法斯，R.和切尼，D.（1990）。绿猴对自身及其他物种警报叫声的评估。《动物行为》，40(4)，754-764。[17] 陈，T.、何，T.、贝内斯蒂，M.、霍蒂洛维奇，V.、唐，Y.、赵，H.和陈，K.（2015）。Xgboost：极端梯度提升。R软件包版本1.7.1，1(4)，1-4。[18] 宝可梦百科。（2022）《宝可梦百科：由社群驱动的宝可梦百科全书》。可访问：https://bulbapedia.bulbagarden.net/（访问日期：07/Mar/2022）。[19] 河原，S.和熊谷，G.（2019）。在宝可梦名称中表达进化：实验性探索。《日本语言学杂志》，35(1)，3-38。[20] 肯登，A.（2017）。对语言起源“手势优先”假说的思考。《心理学通报与评论》，24(1)，163-170。

[21] Breiman, L. (1996). 装袋预测器. 机器学习, 24(2), 123-140.[22] Ho, T. K. (1998). 用于构建决策森林的随机子空间方法. IEEE 模式分析与机器智能汇刊, 20(8), 832- 844.