Application of Machine Learning Algorithm in Human Resource Recommendation: From Tradition Machine Learning Algorithm to AutoML

Xiaoan Pan[1,*]

[1] *School of Software, Jiangxi Normal University, Nanchang, Jiangxi Province, China*
[*] *Email: panxiaoan@jxnu.edu.cn*

ABSTRACT
In the decades since artificial intelligence was proposed, machine learning, the main branch of artificial intelligence, has achieved remarkable results in various fields such as natural language processing, image recognition, and computer vision. This article mainly introduces machine learning from traditional machine learning algorithms to Automated Machine Learning (AutoML). The main structure of this article is as follows: 1) Principles of machine learning algorithms, 2) Application of various machine learning algorithms in human resource recommendation algorithms, 3) Simple application of AutoML algorithm in human resource recommendation algorithms. The advantages and limitations are briefly described at the end. It turns out that the simplicity of the AutoML algorithm makes it very easy to apply in vsarious fields. In the future, with the optimization of the algorithm, the generalization performance of the algorithm will be further improved.
Keywords: *human resource recommendation, machine learning, AutoML algorithms*

机器学习算法在人力资源管理中的应用：从传统机器学习到AutoML

潘晓安[1, *]

[1]*江西师范大学软件学院，南昌，江西省，中国*
[*]*邮箱: panxiaoan@jxnu.edu.cn*

摘要
自人工智能被提出的数十年来，作为人工智能主要分支的机器学习已在自然语言处理，图像识别，计算机视觉等各个领域取得了卓越的成效。本文主要介绍了机器学习由传统的机器学习算法到自动化机器学习（Automated Machine Learning,简称AutoML）。本文主要结构如下:1）机器学习算法原理，2）各种机器学习算法在人力资源推荐算法中的应用，3）AutoML算法在人力资源推荐算法中的简单应用。在最后简要介绍了优点和局限性。事实证明，AutoML算法的简便性使得它十分易于在各个领域中应用。将来，随着对算法的优化，将进一步提高算法的泛化性能。
关键词：*人力资源管理；机器学习；AutoML算法*

1. 引言

机器学习作为数据科学和人工智能的重要分支，正在被越来越多的研究人员和学生研究和探索。这是一个有着非常良好前景的领域，因为它可以应用于许多其他的领域，如统计，生物，医学，金融等。机器学习在这些领域迅速的推广开来，并得到了良好的应用。因此，它使我们得以发现人工智能技术的巨大潜力。什么是机器学习（ML）？在《机器学习》[1]中，作者周志华给出了以下定义：机器学习是一门致力于研究如何通过计算的手段，利用经验来改善系统自身的性能的学科。综上所述，机器学习的根本目的在于通过经验来自动提高计算机系

统的性能。根据提供的训练数据和反馈信息，机器学习可以分为监督学习、无监督学习、半监督学习和强化学习。

本文主要讨论监督学习中的分类问题和回归问题。从传统的机器学习到最近的自动化机器学习。每种算法都将以以下结构进行演示：1）算法原理，2）在人力资源推荐算法中的应用。因此，本文的主要目的是帮助其他研究者更好的理解不同机器学习算法在人力资源领域中的应用，以及AutoML算法在人力资源推荐算法的应用。

2. 传统机器学习

近20年来，统计学习代替符号学习成为机器学习算法中的主流学习算法。统计学习是以数据为驱动，方法为中心建立模型，对数据进行预测和分析。统计学习是以概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科。[2] 本章将通过一些特色鲜明的算法对它们进行说明。

2.1 监督学习

监督学习是统计学习中极其重要的分支，也是应用最广泛的部分。监督学习通过带标签的数据集（由特征向量撑起的特征空间）学习得到一个模型，从而通过已有的训练数据得到最佳模型，再通过模型预测对未知数据进行预测。接下来将简单介绍决策树及***的算法原理及应用。

2.1.1 决策树

决策树是一种既可以用于分类问题也可以用于回归问题的算法，本节主要讨论用于分类的决策树。决策树模型是一种树形结构的模型，它可以被理解为定义在特征空间和类空间上的条件概率分布。利用训练数据，最小化损失函数建立模型，且模型具有可读性，分类速度快[2]，而且树的构造不需要任何领域专家知识或参数设置，适合与探索性知识发现。当前有许多研究采用决策树算法，如电能消耗[3],乳腺癌预测[4]，事故发生频率[5]，人岗匹配[6]，工作态度[7]等。

2.1.2 决策树在HRM中的应用

Hamidah Jantan提出了人力资源管理中基于C4.5决策树的人才预测。该研究中，作者使用C4.5决策树算法，从员工绩效相关因素，如工作成果、知识技能、个人素质、以及活动和贡献等信息（包括背景信息）作为输入变量，将员工是否推荐晋升为标签（是/否）作为输出变量。

A. C4.5决策树

相较于采用信息增益的，偏好可取数值较多的特征的其他决策树，C4.5采用增益率来选择最优划分属性。

增益率定义为：$Gain\_ratio(\boldsymbol{D},a)=\frac{Gain(\boldsymbol{D},a)}{IV(a)}$ (1)

其中$\boldsymbol{D}$为数据集，$a\in A$为属性集，$Gain(\boldsymbol{D},a)$为信息增益，$IV(a)$随着可取数值的增大而增大。因此，增益率偏好可取数值较少的特征，所以C4.5决策树采用先从特征中选取信息增益高于平均水平的特征，再采用增益率从中选取特征。信息增益定义为：

$$Gain(\boldsymbol{D},a)=Ent(\boldsymbol{D})-\sum_{i=1}^{I}\frac{|\boldsymbol{D}^i|}{|\boldsymbol{D}|}Ent(\boldsymbol{D}^i) \quad (2)$$

其中$Ent(\boldsymbol{D})$为信息熵，信息熵是度量样本集合纯度最常用的一种指标，信息熵越小，则数据集的纯度越高。

B. 使用C4.5决策树对人才进行预测

1) 收集员工数据后进行数据预处理和数据清洗
2) 使用C4.5分类器为训练数据生成分类规则，建立模型。即使用设计好的特征，C4.5分类器对训练数据集和分类规则进行分析。在数据预处理阶段，将数据分为两组，一组为训练数据，一组为测试数据，比例为9：1.
3) 使用测试数据评估模型性能。

最后，将该模型嵌入智能决策支持系统，用于预测是否推荐员工晋升。

2.1.3 贝叶斯分类器

贝叶斯分类器是基于贝叶斯定理和特征条件独立的学习算法。贝叶斯算法通过训练数据和大数定律得到类先验概率，使用极大似然估计法来估算类条件概率，朴素贝叶斯(Naïve Bayes)分类器的表达式可以表示为：

$$h_{nb}(x)=arg\max_{c\in\mathcal{Y}}P(c)\prod_{i=1}^{d}P(x_i|c) \quad (3)$$

其中$\mathcal{Y}$为类别标记的集合，$P(c)$为类概率，d为样本的属性数，x_i为样本x在第i个属性上的取值。

贝叶斯分类器在许多研究中被采用，如疾病诊断[8]，文本分类[9]，人才职位匹配[10]等

2.1.4 贝叶斯分类器在人力资源管理中的应用

通过评估增益函数，将特征集分为两个子集。特征的一个子集用于构建决策表，另一子集用于构建初始贝叶斯模型。因此，该算法采用了一个前向选择搜索过程，其中在每个步骤中，选择一组属性由朴素贝叶斯建模，其余属性由决策表建模。在每个步骤中，算法都考虑从模型中完全删除特征。

A. DTNB（Decision Table Naïve Bayes）

决策表朴素贝叶斯（DTNB）的算法与独立决策表（DT）的算法大致相同。在搜索的每个点，它都会评估与将属性分为两个不相交的子集：一个用于DT，另一个用于NB。我们使用前向选择，其中在每个步骤中，所选属性由NB建模，其余属性由DT建模，所有属性最初由DT建模。留一法交叉验证的AUC用于基于组合模型生成

的概率估计来评估拆分的质量。整体的类概率$Q(y|X)$可以表示为：

$$Q(y|X) = \frac{\alpha \times Q_{DT}(y|X^{\top}) \times Q_{NB}(y|X^{\perp})}{Q(y)} \quad (4)$$

其中$Q_{DT}(y|X^{\top})$和$Q_{NB}(y|X^{\perp})$分别是从DT和NB获得的类概率估计值。

B. 使用DTNB匹配职业与求职者

使用三种不同的设定重复实验，每种设定都使用来自训练数据的不同样本。在所有设置中，预测类别都是完整数据中出现最频繁的25家公司之一。但是，在每组参数中，选择当前员工职位仅限于特定公司或大学的情况。在第一种设定中，当前职位是在最频繁的100所大学之一或在最频繁的100个公司之一中。在第二种设定中，当前职位是最频繁出现的100家公司之一。在第三种设定中，当前职位是25家公司中最频繁的公司之一，即那些也构成了类别标签集的公司。

该研究通过求职者过往在学校和公司的经历作为训练数据，将属性分为两个不相交的子集，使用DTNB训练模型，通过求职者信息和模型向求职者推荐职位。

2.2 无监督学习

不同于有监督学习，无监督学习的训练样本中是没有标签的。聚类是无监督学习中研究最多，应用最广的任务，其他常见的无监督学习任务还有密度估计和异常检测等。

2.2.1 聚类

聚类任务的目标是将数据集中的样本划分为数个（通常情况下）不相交的类，这个类也叫做簇（cluster）。通过划分，每个簇可能对应真实数据分布的类别，而这些类别对聚类算法来说是未知的，聚类任务只能根据数据潜在的特性将数据分为数个不同的簇，而这些簇分别属于哪一类需要自行把握。

2.2.2 k-means聚类算法在人力资源管理中的应用

A. k-means

K均值是一种基于原型的聚类（prototype-based clustering），通常情况下，此类算法先对模型进行初始化，然后对原型进行迭代更新求解。k均值首先从数据集中随机选择k个样本作为聚类中心$\{\mu_1, \mu_2 \dots \mu_k\}$，根据各样本距离聚类中心的距离$d$划入相应的簇$\lambda$，表达式如下

$$\lambda_j = \arg min_{i \in \{1,2,\dots k\}} d_{ji}$$
$$d_{ji} = \|x_j - \mu_i\| \quad (5)$$

根据划分好的簇重新计算聚类中心，表达式如下

$$\mu_i' = \frac{1}{|\lambda_i|} \sum_{x \in \lambda_i} x \quad (6)$$

重复以上两步至收敛。k均值聚类算法在管理问题上已有较多研究，如识别学习者群体[11]，评估员工满意度[12]，在学习曲线中生成同质工人组[13]。

B. 使用K-means聚类算法解决人员管理问题[14]

该研究使用从公司人力资源部数据库中收集的，包含详细说明层次结构和绩效评估的劳动力数据，在对数据进行归一化处理后，首先生成聚类，然后评估形成k个簇的效果，选择平局轮廓指数（SI）最高的簇数，即选择使得SI最高的k。最后确定每个簇内部的关键特征，并由专家组解释簇间差异。此研究的目的是改进员工福利，培训和薪酬，为该公司的人才保留做出了贡献。

3. AutoML

从传统机器学习方法来看，机器学习任务需要大量的人工干预，如提取特征，选择最佳模型，调节模型参数，专家分类等等。随着机器学习在众多领域中的成功应用，各行各业对于机器学习系统的需求与日俱增，但机器学习的高门槛令许多人望而却步，而AutoML的出现改变了这一现象，AutoML的目的是将上述步骤进行自动化学习且无需人工干预，对于各行各业不具备机器学习基础知识的使用者来说，AutoML能自动进行数据处理、算法选择、超参数优化等重要工作而无需深入了解及其学习的理论知识。因此，当前有许多企业致力于实现这一需求（如微软的Azure Machine Learning，DataRobot.com，谷歌的Prediction API和Amazon Machine Learning）。 机器学习服务的本质，就是在给定数据集的情况下，确定使用哪一种算法，以及是否对其特征进行预处理和如何进行预处理并进行超参数优化。更具体的地，AutoML问题可以描述为：

定义1：对于$i \in (1, n+m)$，令$x_i \in \mathbb{R}^d$表示特征向量，$y_i \in Y$表示样本i对应的目标值。给定训练数据$D_{train} = \{(\boldsymbol{x_1}, y_1), \dots, (\boldsymbol{x_n}, y_n)\}$和从相同的原始数据分布得到的测试数据集$D_{test} = \{(\boldsymbol{x_{n+1}}, y_{n+1}), \dots, (\boldsymbol{x_{n+m}}, y_{n+m})\}$的特征向量$\boldsymbol{x_{n+1}}, \dots, \boldsymbol{x_{n+m}}$以及资源预算$b$和损失指标$\mathcal{L}(\cdot,\cdot)$，AutoML问题的任务是（自动地）生成测试数据集的预测目标值$\hat{y}_{n+1}, \dots, \hat{y}_{n+m}$。AutoML得到的测试集目标值$\hat{y}_{n+1}, \dots, \hat{y}_{n+m}$的损失由$\frac{1}{m}\sum_{j=1}^{m} \mathcal{L}(\hat{y}_{n+j}, y_{n+j})$给出。[15]

但AutoML并不是万能的，它只在某些场景下可以实现全自动化工作，而且目前只在图像识别领域有较多的应用，在其他领域的应用较少。本文利用开源的机器学习框架ML.NET简单实现了基于AutoML的人才推荐算法。

4. 基于AutoML的人才推荐算法的实现

本次实验采用阿里天池竞赛的数据，其中包含求职者信息，职位信息及求职者是否投递、投递后HR是否满意。在实验中，我们将求职者信息和职位信息作为输入变量，求职者是否投递、HR是否满意作为输出变量，并按3：7的比例作为投递与否和满意与否的权值。最后使用以上数据建立自动化回归模型，得出结果如表1所示

表1 各回归模型评价指标

Trainer	Rsquare	Absolute loss	Squared loss	RMSE
Fast Forest Regression	0.1162	0.71	2.38	1.54
Fast Tree Tweedie Regression	0.0951	0.92	2.51	1.92
Lbfgs Poisson Regression	0.0607	0.81	2.42	1.93
Ols Regression	0.0217	0.97	2.67	2.01
Online Gradient Descent Regression	0.0007	1.15	3.39	1.99

在ML.NET中AutoML会在同一个模型使用不同的参数和阈值来训练，这里我们只给出每个模型中性能最好的一次数据，并按Rsquare值降序排列。我们可以发现Fast Forest Regression, Fast Tree Tweedie Regression的Rsquare相比于其他算法来说较高，但即使是最高的Fast Forest Regression也只有0.1162。由此可见回归模型对该问题的泛化性能较差，即数据间的线性相关性较差。由此可见，该问题不适用于回归模型，仍需要进一步测试其他模型的精度。

5. 结论

本文描述了从传统机器学习到自动化机器学习在人力资源管理中的应用。但是，从实验结果看出，仅使用ML.NET中的回归模型进行自动化机器学习显然是不够的，还应考虑使用其他的模型（如分类模型，深度学习等），以便使用该数据集更好的进行预测。

REFERENCES

[1] Zhou Z.H., Machine Learning, TSINGHUA UNIVERSITY PRESS, Beijing, 2016.

[2] Li Hang, Statistical Learning Method, TSINGHUA UNIVERSITY PRESS, Beijing, 2012.

[3] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, pp. 1761-1768, (2007).

[4] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," Artificial Intelligent in Medicine, vol. 34, pp. 113- 127, (2005).

[5] L. Y. Chang and W. C. Chen, "Data mining of tree-based models to analyze freeway accident frequency," Journal of Safety Research vol. 36, pp. 365-375, (2005).

[6] C. F. Chien and L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," Expert Systems and Applications, vol. 34, pp. 380-290, (2008).

[7] K. Y. Tung, I. C. Huang, S. L. Chen, and C. T. Shih, "Mining the Generation Xer's job attitudes by artificial neural network and decision tree-empirical evidence in Taiwan," Expert Systems and Applications, vol. 29, pp. 783-794, (2005).

[8] Hall, G. H., THE CLINICAL APPLICATION OF BAYES' THEOREM. The Lancet, 290(7515), 555–557. (1967)

[9] Jiang, L., Li, C., Wang, S., & Zhang, L., Deep feature weighting for naive Bayes and its application to text classification. Engineering Applications of Artificial Intelligence, 52, 26–39. (2016).

[10] I., Cambazoglu, B. B., & Gionis, A. Machine learned job recommendation. Proceedings of the Fifth ACM Conference on Recommender Systems - RecSys '11. (2011).

[11] Pimentel, E., França, V. & Omar, N. A identificação de grupos de aprendizes no ensino presencial utilizando técnicas de clusterização. *Proceedings of XIV Simpósio Brasileiro de Informática na Educação.* (2003).

[12] Horst, F., Martins Jr, C. & Souza, L. Técnicas de Data-Mining e reconhecimento de padrões aplicada na predição do nível de satisfação dos colaboradores de um hospital na cidade de Guarapuava/PR. *Proceedings of Congresso de Matemática Aplicada e Computacional.* (2012).

[13] Stroieke, R., Fogliatto, F. & Anzanello, M. Análise de conglomerados em curvas de aprendizado para formação de agrupamentos homogêneos de trabalhadores. *Production, 23,* 537-547. (2013).

[14] Todeschini B, Rodrigues C, Anzanello M, et al. Clustering tool usage to align a company strategy to its talent management needs[J]. European Journal of Applied Business and Management, 82-95,2016, 2(2).

[15] Feurer M, Klein A, Eggensperger K, et al. Efficient and robust automated machine learning[C]. Advances in neural information processing systems. 2015: 2962-2970.