联邦学习基于自发展的高斯生成模型

Miha Ožbot[1], Igor Škrjanc[2]

[1,2]Fakulteta za elektrotehniko, Univerza v Ljubljani, Slovenija

[1]miha.ozbot@fe.uni-lj.si, [2]igor.skrjanc@fe.uni-lj.si

基于自演化高斯聚类的联邦学习

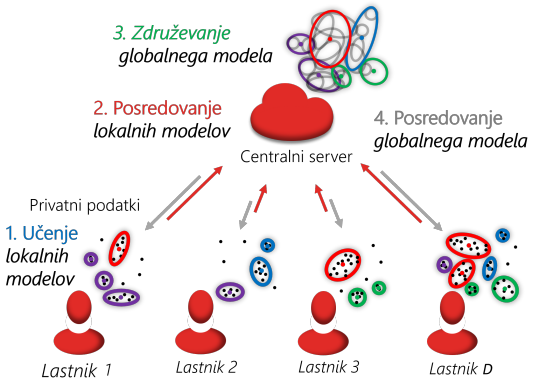
在本研究中，我们提出了一个在联邦学习背景下的进化模糊系统，该系统能够随着新簇的添加动态适应，因此不需要预先选择簇的数量。与传统方法不同，联邦学习允许模型在客户端设备上进行本地训练，仅将模型参数共享给中央服务器而不是数据本身。我们的方法使用 PyTorch 实现，并已在聚类和分类任务中进行了测试。结果表明，我们的方法在几个知名的 UCI 数据集上的表现优于已建立的分类方法。尽管由于重叠条件计算而导致计算强度大，但所提出的方法在去中心化数据处理方面显示出了显著的优势。

1 引入

由于对隐私和数据安全的担忧日益增加，对于旨在实现模型训练过程去中心化的机器学习方法的需求也在增长。联邦学习（联邦学习 FL）[1]在无法因隐私顾虑、监管限制或大量生成的数据而在中央服务器集中收集数据的情况下，是一个有吸引力的解决方案。所有数据不再集中在中央服务器上，而是由数据所有者进行训练 [2]，仅将局部模型更新发送到中心服务器合并，而不是原始数据。

主要贡献是自适应软系统（演化模糊系统 – EFS）[3–6] 的调整以适用于联邦学习。它们为我们在 FL 中遇到的多个关键挑战提供了解决方案。无监督孵化的主要挑战在于需要预先确定孵化数量，这在 FL 中特别困难，在那里每个所有者都需要做出单独决定 [7]。相反，自适应软孵化动态地添加新的孵化，完全避免了这个问题 [5]。自适应系统没有慢收敛的问题，在非独立同分布（非独立同分布 – Non-IID）数据的情况下也是如此 [8]，因为这些孵化在局部是有效的——如果多个所有者的孵化重叠，它们会合并，否则保持分离。自适应系统可以在一次遍历数据时取得良好的结果，因为它们设计用于开放循环中的高效学习。可以通过群集合并机制将局部模型整合到全局模型中，这在自适应软系统中使用。此外，软系统的本质特征是其清晰的结构、基于集群的成员关系、如果-那么规则和本地线性规则，这些都允许对模型推断结果进行解释。

Prvi avtor se zahvaljuje Javni agenciji za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (Slovenian Research and Innovation Agency – ARIS) za finančno podporo, projekt P2-0219.



Slika 1: 建议的联邦学习算法：(1) 每个数据所有者在其私有数据上训练自己的本地模型。(2) 这些本地模型的参数被发送到中央服务器。(3) 在服务器上，将这些本地模型合并成一个全局模型。(4) 然后将这个合并后的全局模型返回给每个本地节点。

推荐的算法如图 1所示。我们关注的是自我进化软分类器（演化模糊分类器 – EFC），它们在 [3] 中被提出，使用了 eClass 分类器系列。作者提出了两种类型的分类器：Class-0，其中类别标签直接用作输出，以及 Class-1，它使用回归来确定后续部分的局部线性模型。后者的示例是 pClass 模型 [4]，该模型使用高斯椭球定义群集。类似于我

们的方法使用群集体积的概念，但计算样本隶属度和规则移除机制有所不同。类似地，在 [6] 中提出了 EFC，它利用无监督的生成来组合数据，并采用实时主动学习选择最具有信息量的样本供专家标注。该方法允许更多的用户标记他们的数据并创建分类器，然后根据共识将这些分类器合并到一个集成中。

柔性模型已在联邦学习中用于多种不同目的。例如，在 [9] 中，训练了不同的联邦学习模型，其中使用柔性模型来确定模型选择。在 [10] 中引入了一种分布式的柔性神经网络，它通过自适应机制添加新规则和停用现有规则来处理非均质和不确定性数据。此外，还为每个所有者提供了规则子集的选择，这对于联邦学习非常重要。除此之外，柔性神经网络最近已在分布式无监督学习 [11] 中使用了 K-中心点聚类（K-均值），并在分布式半监督学习中被采用。据称，在 [7] 中提出了带 c-中心的联邦软聚类方法（模糊 c 均值 FCM）用于无监督聚类，以应对非均匀数据。有趣的是，最后的研究作者强调选择聚类数量作为未解决的问题。这些方法的弱点在于使用了轴平行高斯簇来定义其先验，这比椭球表示 [4] 提供的信息少。更为关键的问题是假设簇的数量固定不变，不允许在学习阶段之后添加新的簇。

2 方法论

自适应软模型的基本构建块是高斯团，定义为中心 $\underline{\mu}_i \in \mathbb{R}^D$，协方差矩阵 $\underline{\Sigma}_i \in \mathbb{R}^{D\times D}$ 和样本数量 $n_i \in \mathbb{N}$，其中 $i \in \mathcal{I}, \mathcal{I} = 1, ..., c$ 是团的索引。为了计算样本与每个聚类之间的距离，我们使用了马哈拉诺比斯距离 $d_i^2 \in \mathbb{R}^+$，然后用它来计算样本对每个规则的隶属函数 $\gamma_i \in [0, 1]$:

$$\gamma_i = \exp\Big(-\frac{1}{D}(\underline{x}-\underline{\mu}_i)^\top \underline{\Sigma}_i^{-1}(\underline{x}-\underline{\mu}_i)\Big). \qquad (1)$$

在分类中，我们的数据集包含 M 个类别，其中为每个类别学习了自己的分类器（一对一）。每个软规则由一个聚类和后续部分组成，后者包含了类别编码 $\underline{\theta}_i \in \{0, 1\}^M$。这种编码是一个 M 维的二进制向量，每一维唯一地代表了一个类别。在快速模式下，工作系统评估每个类别的数据方差，并根据 Fisher 的估计值（费舍尔得分）的阈值 κ_F 选择最具有信息量的特征，该估计值是分布重叠程度的度量。分类器的输出对应于样本 [3, 4] 具有最高隶属度的类别:

$$\hat{\underline{y}}(\underline{x}) = \underline{\theta}_{i^*}, \quad \text{kjer je} \quad i^* = \underset{i\in\mathcal{I}}{\operatorname{argmax}}\ \gamma_i(\underline{x}). \qquad (2)$$

如果条件 $\gamma_{i^*} > \exp(-\mathrm{N}_\sigma^2/D)$ 得到满足，则可以通过现有的软规则来描述样本。参数 i^* 是通过方程 [5] 递增学习的:

$$\underline{e}_{i^*} = \underline{x} - \underline{\mu}_{i^*}, \qquad (3)$$

$$\underline{\mu}_{i^*} = \underline{\mu}_{i^*} + \underline{e}_{i^*}/(n_{i^*}+1), \qquad (4)$$

$$\underline{S}_{i^*} = \underline{S}_{i^*} + \underline{e}_{i^*}\big(\underline{x}-\underline{\mu}_{i^*}\big)^\top, \qquad (5)$$

$$n_{i^*} = n_{i^*} + 1, \qquad (6)$$

如果无法用现有规则描述模式，则添加新规则 $i^* = c+1$，其中心位于 $\underline{\mu}_{i^*} = \underline{x}, n_{i^*} = 1$ 和 $\underline{\Sigma}_{i^*} = \mathrm{diag}\,(\underline{\sigma}^2/\mathrm{N}_r)$，其中 $\underline{\sigma}^2 \in R^D$ 是所有数据方差的估计值，$\mathrm{N}_r \in \mathbb{R}^+$ 是影响簇大小的量化数。激活值我们设置为 $\mathrm{N}_\sigma = \sqrt{D}$。

在实时学习过程中可能会出现较多重叠群集，因此自我发展的软系统采用群集合并机制。该机制合并满足覆盖条件 [5] 的群对 $p, q \in \{i \mid \gamma_i > \exp(-\mathrm{N}_\sigma^2/D)\}$: $V_{pq}/(V_p+V_q) < \kappa_m^D$，其中 pq 表示合并后的群集，而 $V_i = \frac{2\pi^{D/2}}{d\Gamma(D/2)}|\underline{\Sigma}_i|$ 是表示群集的超椭球体的空间。我们并行计算所有群组合的这个条件。中心和合并群集的样本数量计算为 [5]:

$$\underline{\mu}_{pq} = \frac{n_p\underline{\mu}_p + n_q\underline{\mu}_q}{n_{pq}}, \ \text{kjer je}\ n_{pq} = n_p + n_q, \qquad (7)$$

新的协方差矩阵 Σ_{pq} 则计算为:

$$\begin{aligned}(n_{pq}-1)\underline{\Sigma}_{pq} = {} & (n_p-1)\underline{\Sigma}_p + (n_q-1)\underline{\Sigma}_q + \\ & + \frac{n_p n_q}{n_{pq}}(\underline{\mu}_p-\underline{\mu}_q)(\underline{\mu}_p-\underline{\mu}_q)^\top.\end{aligned} \qquad (8)$$

该表示允许在不保存每个群集样本的情况下合并群集，从而实现服务器上的群集合并，在那里样本不可用。此计算同时针对所有候选群集组合进行。确定协方差矩阵后，计算体积条件，并将最合适的候选人合并。一个出现在群集合并中的问题是，一个群集可能变得过大并压倒其他 [4]。为了避免这种情况，在合并时我们限制了群集的大小为原型群集体积的倍数。

在使用私有数据训练后，数据所有者将模型参数提交到服务器，在那里所有本地模型被添加到全局模型并执行合并机制。系统通过计算所有群集中心到其他群集的距离来选择合并候选人。

除了使用正确的合并规则外，我们还采用了基于群集老化的方法，定义为学习过程中规则最近激活时间以来的时间。在服务器上的合并后，移除最老的规则，从而允许系统删除孤立规则并保留最相关的规则。全局模型随后被传输回本地节点。

3 实验

我们实施了联邦生成和分类的实验，采用了提出的方法。该方法通过 PyTorch 库实现，支持在 CPU 或 GPU 上运行。生成是所提方法的关键部分，无论我们的目标是生成还是分类。我们在 2D 数据集 [1] 上进行了生成实验，这些数据集具有高斯过程分布和任意分布的数据集。数据随机分配给 3 个所有者。每个参与者构建自己的局部模型。然后在一轮通信中将模型聚合到服务器上。设计参数需要根据任务和数据集进行一些调整。在此过程中，我们将所有数据集的参数设置为 κ_n=N$_r$ 和 κ_m=1.5。参数 N_r（量化空间）是针对每个数据集通过实验确定的。

研究的第二部分集中在扩展繁殖方法的应用，用于分类。我们建议的方法与现有的分类方法进行了比较：XGBoost、逻辑回归、朴素贝叶斯、K 近邻（k-最近邻 – KNN）、支持向量机（支持向量机 – SVM）、决策树以及自适应模糊分类器 ALMMo-1 [12]。使用的数据集可在 UCI[2] 存储库中找到。我们设定的样本最小数量为 κ_n=1，其他方法参数根据每个数据集进行了经验调整。对于我们的方法，我们将训练数据分给 3 个所有者，而其他方法不是联邦化的。实验采用 3 倍交叉验证（K 折）进行，并重复了 10 次。

4 结果与讨论

出生结果如图 2所示。所提方法很好地描述了基于高斯过程的数据。但对于重叠数据的鲁棒性较差，由于聚类合并的方法不允许具有相似特征向量的同心簇。对于任意分布的数据，单个高斯簇无法完全描述实际的簇。该方法的主要优点是不依赖于样本数量，请参见 S4。两个簇的空间比较在更高维度中变得不可信，因为协方差矩阵的空间由特征值的乘积决定。特征值持续增加或减少会导致空间差异显著增大，即使簇在所有维度上的形状看起来几乎相同也是如此。在这个实验中我们只改变了空间量化，因此出现了一个问题，是否可以在操作过程中根据最小簇的大小自动选择该参数。

分类结果如表 1所示。可以看出，所提方法在大多数选定的数据集上达到了最高的准确性，或者与其它方法相当。虽然 ALMMo 方法可以对多个类别进行分类，但其在二元分类中表现最佳。我们的方法在数值特征上的精度优于分类特征，尽管采用了编码。然而，我们的方法比实现于良好优化库中的比较方法要慢得多。我们方法的大部分计算时间是由于循环中覆盖条件的计算结果，包括多次计算行列式。此计算需要针对多种群组合进行，以找到最适合合并的一个。虽然结合机制会同时为所有群对进行计算，但该方法基于在线学习，依次处理每个样本。图形卡在可以并行执行大量操作且无需存储中间值的情况下运行速度较快，这与在线学习相悖。对于较小的数据集如 Iris 和 Wine，此方法在处理器上更快，但对于其他较大的数据集则在图形卡上更快。

5 结论

在研究中，我们调整了用于识别发展中数据流系统的算法，并将其应用于联邦学习。该方法的关键优势在于不依赖于群数的先验知识且局部模型直接包含在全球模型中。此外，系统可以随时间变化，如果预期的群数发生变化也是如此。此方法的主要限制是其有限的并行计算能力；虽然大多数算法部分已向量化，但增量聚类并未实现。合并群是最耗时的部分。尽管群合并过程简单，但覆盖度量包含要合并的所有群组合的行列式计算。结果表明，所提方法在标准数据集上可达到与公认分类器相当的结果。未来的工作包括将此方法与其他联邦学习和分类技术进行比较。半监督学习具有巨大潜力，它同时包括标记和未标记的数据。

[1] https://cs.uef.fi/sipu/datasets

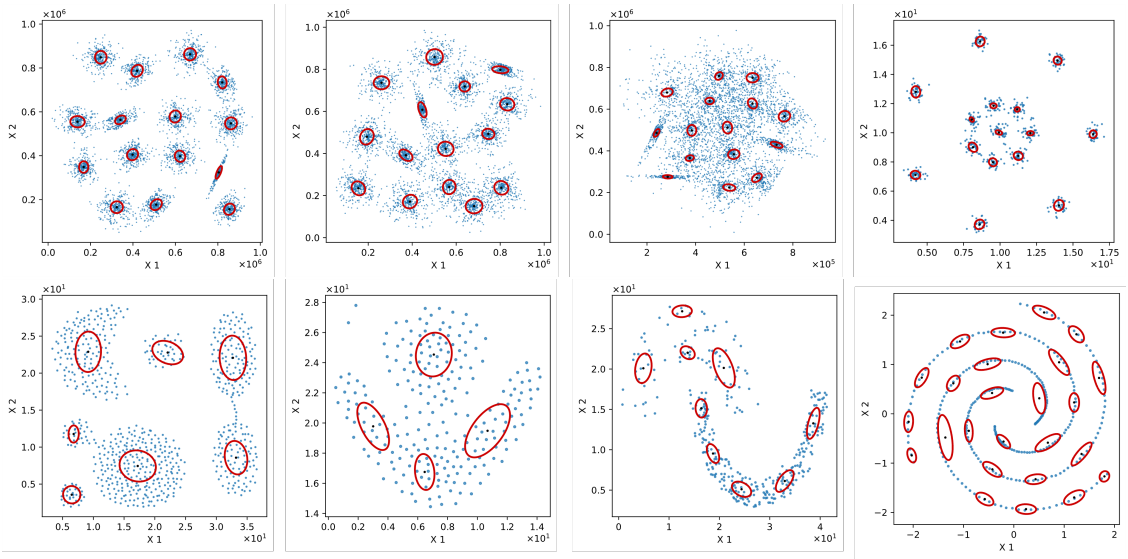
[2] https://archive.ics.uci.edu/数据集

Literatura

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” no. arXiv:1602.05629, Jan. 2023.

[2] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated Learning: Strategies for Improving Communication Efficiency,” no. arXiv:1610.05492, Oct. 2017.

[3] P. Angelov and X. Zhou, “ Evolving Fuzzy-Rule-Based Classifiers From Data Streams,” IEEE



Slika 2: 联邦生成合成数据集的聚类（从上到下，从左至右）：S1, S2, S4, R15；以及非高斯分布的数据集（从下至上，从左至右）：Aggregation, Flame, Jain, Spiral。显示了数据（蓝色）和 2σ 椭圆，代表聚类（红色）。

数据集	仅仅	方法							
		XGBoost	Logistic Regression	Naive Bayes	KNN	SVM(RBF)	Decision Tree	ALMMo-1	Naša
Perunika (Iris)	Natančnost [%]	94.1±2.5	95.1±2.2	95.2±2.0	94.9±2.0	95.7±2.2	93.9±3.0	66.7±6.3	96.5±1.8
	F1 score [%]	94.0±2.5	95.1±2.2	95.2±2.0	94.9±2.0	95.7±2.2	93.9±3.0	55.8±7.5	96.5±1.8
	ROC AUC [%]	98.2±1.1	99.7±0.3	99.5±0.4	99.0±1.0	99.8±0.2	95.5±2.3	50.0±0.0	99.9±0.2
	Čas / vzorec [ms]	0.24±0.05	0.01±0.02	0.00±0.00	0.01±0.02	0.01±0.00	0.00±0.00	0.76±0.11	1.44±0.08
Vino (Wine)	Natančnost [%]	96.7±1.8	98.0±1.6	97.5±1.5	95.6±2.2	98.2±1.4	90.3±3.9	68.5±3.7	98.3±1.6
	F1 score [%]	96.7±1.8	98.0±1.7	97.5±1.5	95.6±2.3	98.2±1.4	90.3±3.9	59.1±4.2	98.3±1.6
	ROC AUC [%]	99.7±0.3	100.0±0.1	99.9±0.2	99.4±0.8	100.0±0.1	92.6±3.0	50.0±0.0	99.9±0.1
	Čas / vzorec [ms]	0.18±0.03	0.01±0.01	0.00±0.00	0.01±0.01	0.01±0.00	0.00±0.00	0.72±0.10	1.28±0.05
Bolezni srca (Heart disease)	Natančnost [%]	79.0±3.7	83.2±4.2	83.5±3.7	82.0±3.3	82.3±3.7	72.4±4.4	83.5±3.8	83.1±3.8
	F1 score [%]	79.0±3.7	83.1±4.3	83.4±3.7	82.0±3.3	82.2±3.7	72.4±4.3	83.4±3.9	81.6±3.7
	ROC AUC [%]	87.7±3.0	90.2±2.7	89.6±2.6	88.5±2.9	89.6±2.7	72.5±4.3	50.0±0.0	89.4±2.5
	Čas / vzorec [ms]	0.09±0.03	0.00±0.01	0.00±0.00	0.01±0.02	0.01±0.01	0.00±0.01	0.69±0.14	1.13±0.08
Rak dojke (Brest cancer)	Natančnost [%]	96.2±1.2	97.7±0.8	93.3±1.4	96.4±1.0	97.4±0.8	92.6±2.1	95.6±1.2	96.4±1.5
	F1 score [%]	96.2±1.1	97.7±0.8	93.3±1.4	96.3±1.0	97.4±0.8	92.6±2.1	95.5±1.2	95.2±1.9
	ROC AUC [%]	99.1±0.5	99.5±0.3	98.5±0.6	98.4±0.6	99.5±0.3	92.2±2.2	50.0±0.0	99.4±0.5
	Čas / vzorec [ms]	0.06±0.02	0.00±0.00	0.00±0.00	0.02±0.01	0.01±0.01	0.01±0.01	0.75±0.13	1.02±0.06
Številke (Digits)	Natančnost [%]	96.1±1.0	96.7±0.7	78.5±3.6	97.1±0.7	98.0±0.5	84.5±1.5	11.5±1.2	98.1±0.4
	F1 score [%]	96.1±0.9	96.7±0.7	78.4±3.7	97.1±0.7	98.0±0.5	84.5±1.5	4.3±0.7	98.1±0.4
	ROC AUC [%]	99.9±0.0	99.9±0.0	97.2±0.4	99.6±0.1	99.9±0.0	91.4±0.8	50.0±0.0	99.8±0.1
	Čas / vzorec [ms]	0.13±0.02	0.01±0.00	0.00±0.00	0.02±0.00	0.08±0.00	0.00±0.00	1.25±0.15	1.45±0.10
Avtizem (Autism)	Natančnost [%]	100.0±0.0	100.0±0.0	96.6±1.1	96.1±1.2	99.2±1.1	100.0±0.0	95.5±1.4	100.0±0.0
	F1 score [%]	100.0±0.0	100.0±0.0	96.6±1.1	96.1±1.2	99.2±1.1	100.0±0.0	95.5±1.4	100.0±0.0
	ROC AUC [%]	100.0±0.0	100.0±0.0	99.1±0.9	99.3±0.4	100.0±0.1	100.0±0.0	50.0±0.0	100.0±0.0
	Čas / vzorec [ms]	0.03±0.01	0.01±0.01	0.00±0.00	0.02±0.01	0.01±0.01	0.00±0.00	0.97±0.18	0.56±0.10

Tabela 1: 与不同数据集的分类问题推荐方法的比较。展示了精度（准确性）的标准差平均值 ±，F1 评分（F1 分数），接收者操作特性曲线下的面积（接收者操作特性曲线下的面积 –ROC 曲线下面积）以及每个样本的学习时间，以毫秒为单位。

Transactions on Fuzzy Systems, vol. 16, no. 6, p. 1462 – 1475, Dec. 2008.

[4] M. Pratama, S. G. Anavatti, M. Joo, and E. D. Lughofer, “ pClass: An Effective Classifier for Streaming Examples,” IEEE Transactions on Fuzzy Systems, vol. 23, no. 2, p. 369 – 386, Apr. 2015.

[5] I. Škrjanc, “Cluster-Volume-Based Merging Approach for Incrementally Evolving Fuzzy Gaussian Clustering-eGAUSS+,” IEEE Transactions on Fuzzy

Systems, vol. 28, no. 9, p. 2222 – 2231, Sep. 2020.

[6] E. Lughofer, "Evolving multi-user fuzzy classifier systems integrating human uncertainty and expert knowledge," Information Sciences, vol. 596, p. 30 – 52, Jun. 2022.

[7] M. Stallmann and A. Wilbik, " On a Framework for Federated Cluster Analysis," Applied Sciences, vol. 12, no. 2020, p. 10455, Jan. 2022.

[8] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the Convergence of FedAvg on Non-IID Data," no. arXiv:1907.02189, Jun. 2020.

[9] D. Poap, " Fuzzy Consensus With Federated Learning Method in Medical Systems," IEEE Access, vol. 9, p. 150383 – 150392, 2021.

[10] L. Zhang, Y. Shi, Y.-C. Chang, and C.-T. Lin, "Federated Fuzzy Neural Network with Evolutionary Rule Learning," IEEE Transactions on Fuzzy Systems, vol. 31, no. 5, p. 1653 – 1664, May 2023.

[11] Y. Shi, C.-T. Lin, Y.-C. Chang, W. Ding, Y. Shi, and X. Yao, "Consensus Learning for Distributed Fuzzy Neural Network in Big Data Environment," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 5, no. 1, p. 29 – 41, Feb. 2021.

[12] P. P. Angelov, X. Gu, and J. C. Principe, "Autonomous learning multimodel systems from data streams," IEEE Transactions on Fuzzy Systems, vol. 26, no. 4, p. 2213 – 2224, Aug. 2018.