多维高斯样本在无监督增量学习中的覆盖问题

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

重叠多变量高斯聚类在无监督在线学习中的度量方法

本文提出了一种新的检测多元高斯聚类重叠的度量方法。从数据流中进行在线学习的目标是创建能够根据流数据的概念漂移随时间适应的聚类、分类或回归模型。在聚类的情况下，这可能会导致大量可能重叠且应合并的簇。常用的分布不相似性测度由于无法考虑所有形状的簇以及其高计算需求，在基于数据流在线学习上下文中确定重叠簇方面并不充分。我们提出的不相似性测度专门设计用于检测重叠而非不相似性，并且与现有测度相比可以更快地计算。我们的方法比比较的方法快几倍，能够检测到重叠的簇同时避免合并正交的簇。

1 引言

自适应模型 [1] 是数学模型，其任务是实时调整结构和参数以预测因变量、生成数据或分类。观察系统处于变化的环境中，因此容易受到概念漂移和参数突发变化的影响。这种系统在现实世界中经常遇到，从工业交易中的不断变化的情况，金融交易和投资，到医疗数据流，社交网络和网络安全。关键要求是模型能够快速适应变化，并且计算效率高，因为数据量在数据流的情况下是无限的。这些系统通常使用无监督的数据生成方法来确定结构并提取数据中的信息。为此经常使用高斯混合 [2–6]，表示为多变量正态分布（多元正态分布）和数据协方差矩阵 $P \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$。随着簇数的增加，我们希望能够逼近更复杂的分布，即使观察过程不遵循高斯分布。

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.

在数据流及时学习的情况下，簇的数量会增加，并随着数据量的增加而开始重叠。有时需要合并这些簇以减少方法的计算复杂度并简化模型。在此过程中出现了一个关键问题：如何在即时学习中更有效地找到重叠的簇。文献中提出了几种检测簇重叠的方法，但这些方法尚未解决所有情况下的计算效率和重叠检测问题 [4–6]。这些问题在于它们专注于分布相似性而不是簇的重叠，因此无法检测到大小不同的簇之间的重叠。潜在地可以通过使用并行处理来提高合并速度，这允许在一步中合并更多的簇，但这一方面尚未在文献中得到研究。另一个关键问题是逐步合并两个或多个簇是否更有效。在第一种情况下需要多次检查簇的重叠，在第二种情况下则需要找到最佳的合并组合。希望一次能够合并尽可能多的簇，因为这是在图形卡上实现计算的最佳方法。问题在于寻找最大的群组（团）来合并是本质上递归的问题，并且目前还没有能够在多项式时间内找到最优解的算法（NP 难）。同时潜在使用图形卡 [7,8] 由于并行处理的潜力可以加速在大型图中的群组搜索。经典方法依赖于递归，这不适合在图形卡上计算。

研究的目标是开发一种检测重叠簇的方法，使高维和大量簇的计算既稳定又快速。我们感兴趣的是是否可以并行处理簇的重叠和合并过程，以实现高效利用图形卡。这些在最近十年内被证明比处理器计算更有效，特别是如果能够执行大量的矩阵操作而不涉及递归循环，并且不需要存储或同步中间步骤。本文提出了一种基于簇体积比率的新重叠测量方法。该方法在高维中计算效率和稳定，考虑了簇内的样本数量，可以检测小的簇并区分不同形状的簇。

2 方法论

集合粒子需要粒子覆盖的度量，选择适合合并的粒子的过程以及计算合并粒子的协方差矩阵。我们使用 $i \in \mathcal{I}$ 索引的粒子集，在这里 $\mathcal{I} = \{1, \ldots, c\}$ 是完全由分配给它的样本数量 n_i 定义的，协方差矩阵 $\mathbf{\Sigma}_i$ 和粒子集合的期望值或中心点 $\boldsymbol{\mu}_i$。我们感兴趣的是既能最好地描述粒子覆盖又在计算上最有效的度量。

2.1 分布相似性

设 $P \sim \mathcal{N}(\boldsymbol{\mu}_P, \mathbf{\Sigma}_P)$ 和 $Q \sim \mathcal{N}(\boldsymbol{\mu}_Q, \mathbf{\Sigma}_Q)$ 为多变量高斯分布，定义了相应的族群。我们讨论相似性和不一致性的度量（不相似性），这些度量用作覆盖族群的程度的指标。

2.1.1 巴塔恰里亚距离

巴特查里亚距离（B）常被用作种群覆盖度量，因为它是对称的，并且通过种群 [9–11] 的平均协方差矩阵间接归一化距离：

$$D_B(P\|Q) = \frac{1}{8}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P)^\top {\mathbf{\Sigma}_M}^{-1}(\boldsymbol{\mu}_Q - \boldsymbol{\mu}_P) + \frac{1}{2}\ln\left(\frac{\det \mathbf{\Sigma}_M}{\sqrt{\det \mathbf{\Sigma}_P \det \mathbf{\Sigma}_Q}}\right), \tag{1}$$

其中是 $\mathbf{\Sigma}_M = \frac{1}{2}(\mathbf{\Sigma}_P + \mathbf{\Sigma}_Q)$.

可以将表达式展开为对数行列式的和，其计算更为数值稳定。第二项的一个有趣特性是，分子等于平均值的行列式，分母则是协方差矩阵行列式的几何平均。

2.1.2 詹森-香农散度

杰森-香农散度（JS）是基尔沙克-莱布勒散度（KL）的扩展，它通过引入合并分布 $M \sim \mathcal{N}(\boldsymbol{\mu}_M, \mathbf{\Sigma}_M)$ 解决了 KL 散度不对称的问题，该合并分布表示两个分布的平均值，并允许对它们之间的距离进行对称测量 [12]：

$$D_{JS}(P\|Q) = \frac{1}{2}\big(D_{KL}(P\|M) + D_{KL}(Q\|M)\big), \tag{2}$$

$$D_{KL}(P\|M) = \frac{1}{2}\left(\operatorname{tr}\left(\mathbf{\Sigma}_M^{-1}\mathbf{\Sigma}_P\right) + \ln\left(\frac{\det \mathbf{\Sigma}_M}{\det \mathbf{\Sigma}_P}\right) + (\boldsymbol{\mu}_M - \boldsymbol{\mu}_P)^\top \mathbf{\Sigma}_M^{-1}(\boldsymbol{\mu}_M - \boldsymbol{\mu}_P) - d\right), \tag{3}$$

其中 $\boldsymbol{\mu}_M = \frac{1}{2}(\boldsymbol{\mu}_P + \boldsymbol{\mu}_Q)$ 和 $\mathbf{\Sigma}_M = \frac{1}{2}(\mathbf{\Sigma}_P + \mathbf{\Sigma}_Q)$。

2.1.3 瓦瑟斯坦距离（2-）

二维 Wasserstein 距离（W），基于分布间距离的平方，是将一个分布的质量最优传输到另一个分布的成本度量 [13]：

$$D_W(P \parallel Q) = \left\|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\right\|^2 + \operatorname{Tr}\left(\mathbf{\Sigma}_P + \mathbf{\Sigma}_Q - 2\left(\mathbf{\Sigma}_P^{1/2}\mathbf{\Sigma}_Q\mathbf{\Sigma}_P^{1/2}\right)^{1/2}\right), \tag{4}$$

2.1.4 梅拉覆盖 e 高斯

蜂群的叠加 eGauss+ [4] 基于合并前后蜂群体积的比率：

$$D_{eGauss+}(P\|Q) = \frac{\det(\mathbf{\Sigma}_M)}{\det(\mathbf{\Sigma}_P) + \det(\mathbf{\Sigma}_Q)}. \tag{5}$$

此时，合并后的协方差矩阵等于：

$$n_M = n_P + n_Q, \tag{6}$$

$$\boldsymbol{\mu}_M = \big(n_P\boldsymbol{\mu}_P + n_Q\boldsymbol{\mu}_Q\big)/n_M, \tag{7}$$

$$\mathbf{\Sigma}_M(n_M - 1) = (n_P - 1)\mathbf{\Sigma}_P + (n_Q - 1)\mathbf{\Sigma}_Q + \boldsymbol{M}_P^\top \boldsymbol{E}_P^\top \boldsymbol{E}_P \boldsymbol{M}_P + \boldsymbol{M}_Q^\top \boldsymbol{E}_Q^\top \boldsymbol{E}_Q \boldsymbol{M}_Q - \boldsymbol{M}_M^\top \boldsymbol{E}_M^\top \boldsymbol{E}_M \boldsymbol{M}_M, \tag{8}$$

其中 $\boldsymbol{M}_P = \boldsymbol{\mu}_P^\top \mathrm{I} \in \mathbb{R}^{d\times d}$ 表示对角矩阵，其对角线上包含蜂群中心的变量，d 是问题的维度或特征数，$\boldsymbol{E}_P \in \mathbb{R}^{n_P \times d}$ 是所有元素都等于 1 的矩阵，Q 和 M 也是如此。

其允许排除具有正交特征向量的聚类，但不包括对高维协方差矩阵的保护。在此上下文中，正交性意味着定义聚类的特征值在特征空间中彼此相对垂直。合并聚类的过程也很有趣，这不再仅仅是聚类的简单平均，而是考虑了协方差矩阵的定义和聚类中的样本数量。这使得可以赋予具有更多样本的聚类更大的影响，从而解决了在计算过程中，尽管另一度量有大量更多的样本，但分布或包含少量样本的聚类占据主导地位的问题。

2.1.5 建议的覆盖度量

我们想要一种衡量集群重叠度（而非相似度）的度量，并能够检测不同大小的集群（I），在较高维度下具有计算稳定性 d（II），并且计算效率高（III）。我们通过使用集群行列式的算术平均值来解决第一个要求，就像在 eGauss+度量中一样。我们可以通过计算 $\ln(\det(\Sigma))$ 而不是行列式 $\det(\Sigma)$

来解决第二个要求。我们得到了覆盖度量 (重叠):

$$D_O(P\|Q) = \ln\left(\frac{\det(\mathbf{\Sigma}_M)}{\frac{1}{2}\left(\det(\Sigma_P)+\det(\Sigma_Q)\right)}\right) = \ln(2)+\ln\det(\mathbf{\Sigma}_M)-\ln\left(e^{\ln\det(\mathbf{\Sigma}_P)}+e^{\ln\det(\mathbf{\Sigma}_Q)}\right), \tag{9}$$

其中合并协方差矩阵定义为:

$$(n_p+n_q-1)\mathbf{\Sigma}_M = (n_p-1)\mathbf{\Sigma}_p + (n_q-1)\mathbf{\Sigma}_q + + \frac{n_p n_q}{n_p+n_q}(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)(\boldsymbol{\mu}_p-\boldsymbol{\mu}_q)^\top. \tag{10}$$

合并协方差矩阵的公式基于简化版的方程 (8)。为了比较，我们测试了算术平均值作为 Bhattacharyya 度量 (aB) 的方法。Bhattacharyya 度量显示出很大的潜力，但当群集大小差异很大时，在第二项分母中的几何平均值会导致问题。仅这一点本身并不是缺点，如果我们用距离作为相似性度量，因为这两个集群确实不同，但它不允许检测到重叠的集群。通过在度量中用算术平均值替换几何平均值可以消除这个缺点。此外，我们可以通过使用包括群集中心距离信息的合并方法来避免计算第一项中的逆矩阵以获得新的合并矩阵。当合并群集的体积小于两个群集体积的平均值时，覆盖度量值将小于零。

第三个要求是最具挑战性的，因为体积、特征向量和矩阵逆的计算基于行列式的计算，这是一种计算上昂贵的操作。我们通过估计行列式的上限来解决这个需求而不是计算精确值。哈达玛不等式表明，对于任何正半定矩阵 $\Sigma \in \mathbb{R}^{d\times d}$，矩阵的行列式小于或等于其对角元素的乘积 [14]:

$$\det(\Sigma) \le \prod_{l=1}^{d}\sigma_{ll}. \tag{11}$$

使用哈达玛不等式 ($\tilde{\mathrm{O}}$) 的覆盖度量表示体积上界的估计。在该度量中，我们用估计代替精确的行列式计算，这大大加快了高维情况下的计算速度。这种方法的缺点是，它丧失了检测具有非常不同特征值簇的能力，因为失去了关于属性之间相关性的信息。

2.2 更大数量的群集合并

合群的效率可以通过一次合并更多的群体来提高，而不是迭代进行。在迭代合并中，需要多次连续计算相似度量，这可能非常低效，特别是在使用图形卡时。检查群组对之间的重叠是一个复杂的过程，尤其是当群组很大时，但我们可以通过使用额外的标准来选择合适的候选者。验证所有可能的群体组合以确定不同数量的重叠群组是低效的，甚至是不可能的。我们建议比较群组对并寻找最大互重叠群组集。关键在于找到最大的一组 (最大团) 适合合并的群体。本文的重点不是计算最大集合的效率，而是展示相似度量和一次步骤中合并多个群体的方法。我们使用邻近群体的矩阵表示 (邻接矩阵)，这是我们通过重叠测量获得的，并且使用任意经典算法来查找最大的集合。

为了合并更多的群体，我们需要一个新的方程。我们建议改进方程 (8) 以用于合并后的协方差矩阵，这允许合并更多群体。设 $\mathcal{J}\subseteq\mathcal{I}$ 是要合并的群体集。方程的推导基于 $\boldsymbol{X}_M^\top\boldsymbol{X}_M = \sum_{j\in\mathcal{J}}\boldsymbol{X}_j^\top\boldsymbol{X}_j$ 和 $\mathbf{\Sigma}_M = \frac{1}{(n_M-1)}(\boldsymbol{X}_M^\top\boldsymbol{X}_M - n_M\boldsymbol{\mu}_M\boldsymbol{\mu}_M^\top)$，其中 $\boldsymbol{X}_M^\top = \left[\boldsymbol{X}_j^\top : j\in\mathcal{J}\right] \in \mathbb{R}^{(\sum_j n_j)\times d}$ 表示包含所有群样本的数据矩阵。我们可以将其表述为一个不需要数据矩阵的方程:

$$(n_M-1)\mathbf{\Sigma}_M = \sum_{j\in\mathcal{J}}(n_j-1)\mathbf{\Sigma}_j + \sum_{j\in\mathcal{J}} n_j\boldsymbol{\mu}_j\boldsymbol{\mu}_j^\top - n_M\boldsymbol{\mu}_M\boldsymbol{\mu}_M^\top = \underbrace{\sum_{j\in\mathcal{J}}(n_j-1)\mathbf{\Sigma}_j}_{\text{varianca}} + \underbrace{\sum_{i\in\mathcal{J}}\sum_{j\in\mathcal{J},j>i}\frac{n_i n_j}{n_M}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^\top}_{\text{razdalja med središči}}, \tag{12}$$

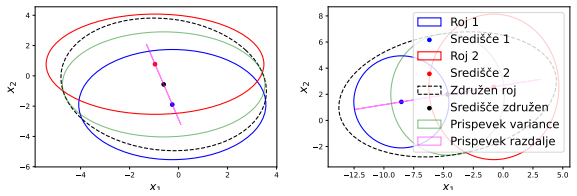
其中 $n_M = \sum_{j\in\mathcal{J}} n_j$ 和 $\boldsymbol{\mu}_M = \frac{1}{n_M}\sum_{j\in\mathcal{J}} n_j\boldsymbol{\mu}_j$。

推导过程因篇幅原因省略。最终方程的记录更具说明性，但对于较大的群数量而言计算效率不如初始形式。合并后的群由各群体协方差矩阵的贡献和中心之间的距离组成，后者是一阶项。此表达式允许可视化每个项对组合群的影响。

3 实验

我们感兴趣的是能够检测到重叠所有情况的发散度量，这种重叠在变化环境中实时学习的问题中经常出现。由于在数据流中的大量数据和计算需求方面，实时学习方法的速度至关重要，我们不仅关心发散值的准确性，还关心每个度量的计算复杂性，我们通过计算所需的时间来评估这一点。图 1展示了不同群集合并的情景：(a) 重叠群集，(b) 不重叠群集，(c) 大群集内的小群集和 (d) 具有非常不同的主导特征值但共享中心的群

集。在小群集中，希望度量显示重叠，在正交群集中，则希望度量显示没有重叠。这些群集是随机创建的。



(a) 覆盖范围。 (b) 不覆盖数组。

(c) 大和小鱼。 (d) 直角坐标系。

Slika 1: 显著的覆盖群示例在即时学习中被我们用于实验。展示了合并前的簇中心和 2σ 个椭圆簇以及合并后的簇。还显示了由于中心之间的距离和协方差矩阵贡献而产生的分离贡献。

在第二部分中，我们验证了使用我们的度量合并群集的情况，其中考虑了多个相互重叠的群集的覆盖情况。这些群集是通过随机中心和协方差矩阵生成的。合并候选者的选择是根据发散性指标来进行的，他们的索引被包含在连接矩阵中，该矩阵随后用于寻找最大的组。在此过程中，每个群集恰好出现在一个组中。所有群集组随后在一个步骤中合并。我们随机生成了群集，并通过可视化观察是否正确地选择了相互重叠的群集并将其组合在一起。

4 结果与讨论

在第一个实验中，我们在不同的场景下分析了各种覆盖度量。实验结果汇总在表 1中。所有指标都能成功区分群集之间的重叠和非重叠情况，这是意料之中的，因为这些指标通常用于此目的。它们之间的重要区别在于某些指标根据合并后群集的大小对距离进行归一化，并考虑了群集之间的比例关系，而不依赖于所比较的群集的大小。在分析小群集时，度量值应能明确区分重叠和非重叠情况，因为预计该度量将清楚地显示重叠现象。这个要求已在我们的所有指标以及 eGauss+ 指标中得到满足。值得注意的是，基于 Bhattacharyya 距离的我们的度量返回了与该距离非常相似的值，唯一的区别是它能够检测到小群集。

在最后一个场景中，我们处理了两个具有正交主特征向量的群集。合并这些群集将导致一个比原始两者都大的新群集。在这种情况下，不希望进行群集合并。这种场景被 eGauss+ 指标和我们的覆盖度量所成功识别，该度量并不包含对行列式上限估计的考量。

所建议的度量能够检测到低维和高维群集的所有场景，并且比比较的方法更快。Wasserstein 距离包括矩阵根，这是一项计算上昂贵的操作。此外，它不受特征定义域归一化的影响，这对于选择合并阈值是一个很大的问题。有趣的是，所有 Jensen-Shannon 散度和 Bhattacharyya 距离的值对于所有情况都是相同的，这是一个出乎意料的结果。这可能是由于选择了均值作为中心和协方差矩阵的原因，这是一种常见的做法。这也得到了分析推导的证实：

$$
\begin{aligned}
&D_{JS}(P\|Q) = \\
&\quad \frac{1}{4}\Bigg(\Big[\operatorname{tr}\left(\boldsymbol{\Sigma}_M^{-1}\boldsymbol{\Sigma}_P\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}_M^{-1}\boldsymbol{\Sigma}_Q\right) - 2d\Big] + \\
&\quad \left[(\boldsymbol{\mu}_M - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_M^{-1}(\boldsymbol{\mu}_M - \boldsymbol{\mu}_P) + \right. \\
&\quad \left.(\boldsymbol{\mu}_M - \boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_M^{-1}(\boldsymbol{\mu}_M - \boldsymbol{\mu}_Q)\right] + \\
&\quad \left[\ln\left(\frac{\det \boldsymbol{\Sigma}_M}{\det \boldsymbol{\Sigma}_P}\right) + \ln\left(\frac{\det \boldsymbol{\Sigma}_M}{\det \boldsymbol{\Sigma}_Q}\right)\right]\Bigg) = \\
&\quad = \frac{1}{4}\Bigg(\Big[\operatorname{tr}\left(\boldsymbol{\Sigma}_M^{-1}\frac{2(\boldsymbol{\Sigma}_P+\boldsymbol{\Sigma}_Q)}{2}\right) - 2d\Big] + \\
&\quad 2\Big(\frac{\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q}{2}\Big)^\top \boldsymbol{\Sigma}_M^{-1}\Big(\frac{\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q}{2}\Big) + \\
&\quad \Big[2\ln\det\boldsymbol{\Sigma}_M - \frac{1}{2}\big(\ln\det\boldsymbol{\Sigma}_P + \ln\det\boldsymbol{\Sigma}_Q\big)\Big]\Bigg) = \\
&\quad = 0 + \frac{1}{8}(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)^\top \boldsymbol{\Sigma}_M^{-1}(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q) + \\
&\quad \frac{1}{2}\ln\left(\frac{\det\boldsymbol{\Sigma}_M}{\sqrt{\det\boldsymbol{\Sigma}_P \det\boldsymbol{\Sigma}_Q}}\right) = D_B(P\|Q).
\end{aligned}
\tag{13}
$$

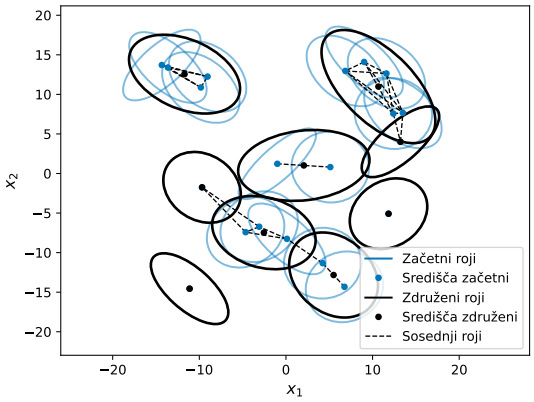
通过等式（12）获得的合并群集示例如图 1所示。显示了由于共同协方差矩阵中协方差矩阵和中心距离的不同贡献，提供了对群集合并机制工作的额外见解。在大多数情况下，中心之间的距离决定了是否认为两个群集是重叠的，除了共享同一中心的情况。在另一个实验中，我们讨论度量群集重叠是否合适。可以通过图 2来视觉评估覆盖度

d	实验	B		JS		W		eGauss+		**我们的** (aB)		O(**纳沙**)		Õ(**我们的**)	
		V	**时间** [μs]	V	**时间** [μs]	V	**时间** [μs]	V	**时间** [μs]	V	**时间** [μs]	V	**时间** [μs]	V	**时间** [μs]
2	prekrivanje ✓	0.06	149(2.8x)	0.06	79(1.5x)	1.25	162571(3010.6x)	0.48	217(4.0x)	0.03	68(1.3x)	0.04	54	0.05	55(1.0x)
	ni prekrivanja ✗	1.89	87(1.7x)	1.89	89(1.8x)	10.92	296(5.9x)	5.22	132(2.6x)	1.86	65(1.3x)	1.52	50	1.59	57(1.1x)
	majhen roj ✓	1.50	110(1.4x)	1.50	126(1.6x)	13.83	386(4.9x)	0.87	172(2.2x)	-0.29	78	0.51	96(1.2x)	0.51	92(1.2x)
	pravokotna ✗	0.85	114(1.5x)	0.85	120(1.6x)	1.22	406(5.5x)	2.72	152(2.1x)	0.85	74	1.69	96(1.3x)	0.84	86(1.2x)
100	prekrivanje ✓	1.45	1006(5.7x)	1.45	930(5.2x)	14.95	2254(12.7x)	23.63	1066(6.0x)	1.19	567(3.2x)	1.46	432(2.4x)	1.88	178
	ni prekrivanja ✗	117.08	720(3.0x)	117.08	922(3.8x)	96.23	2942(12.3x)	146.08	1687(7.0x)	116.24	639(2.7x)	4.27	588(2.5x)	91.78	240
	majhen roj ✓	75.38	582(2.8x)	75.38	838(4.0x)	666.22	3115(14.8x)	0.04	783(3.7x)	-25.99	553(2.6x)	-7.84	467(2.2x)	-7.27	211
	pravokotna ✗	1.14	451(2.6x)	1.14	994(5.6x)	1.39	13764(78.2x)	2.97	649(3.7x)	1.14	404(2.3x)	1.78	314(1.8x)	0.75	176

(**我们的**) – **我们的建议度量**，B – Bhattacharyjeva 距离, JS – Jensen-Shannon 散度, W – Wasserstein 距离, eGauss+ – eGauss+ [4] 的体积比, aB – Bhattacharyjeva 距离 + 算术平均值 + lndet(), O – 重叠度量,Õ– 重叠度量 + Hadamard 不等式

Tabela 1: 不同覆盖场景下维度高低的覆盖测量比较。当检测正确时，测量值用绿色标记。计算时间以微秒（μs）和最佳结果的倍数给出。

量的质量。可以看到，所建议的测量正确地检测到了所有相邻的集群，并选择了相互重叠的最大集群组，合并多个集群的等式 12是精确的。



Slika 2: 将多个随机生成的群集合并，其中我们使用了我们的覆盖度量来选择重叠的群集，并使用建议的方程来同时合并多个群集。

尽管正态分布是最常用的概率分布描述结构，但计算这些度量的计算成本很高。所有的度量都需要至少计算行列式、特征值或矩阵的逆运算，间接地包括了行列式的计算。有趣的是，聚类合并过程本身在计算上非常有效，而确定合并有意义的相似性度量的计算则复杂得多。如果我们只有聚类可用，就无法避免使用覆盖测量来计算合并的适当性。虽然建议的方法尚未完全解决覆盖聚类计算的成本问题，因为它仍然涉及矩阵行列式的计算，但是相对于可比方法而言，在大多数情况下，该建议的过程速度要快得多。

5 结论

在本文中，我们提出了一种新的高斯混合模型的覆盖度量，并建议了一个用于选择和合并更多聚类的新过程。这项研究的动机源自实时无监督学习，但提出的算法可以应用于使用高斯混合作为原型的各种聚类生成方法。提议的度量比其他覆盖方法快得多，并且能够在其他方法失败的情况下准确地识别出重叠和非重叠的聚类。未来的方法将被用于联邦学习（Federated Learning）中的模型合并，在这种情况下，每个数据所有者都会建立自己的局部聚类模型，然后在不涉及数据传输的情况下合并为一个全球模型。这样的系统的一个应用示例是在金融系统的欺诈交易数据实时生成或在网络安全性中检测漏洞。

Literatura

[1] Škrjanc, Igor and Iglesias, Jose and Sanchis, Araceli and Leite, Daniel and Lughofer, Edwin and Gomide, Fernando, “Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A Survey,” Information Sciences, vol. 490, p. 344 – 368, Jul. 2019.

[2] E. Lughofer and I. Škrjanc, “Online Active Learning for Evolving Error Feedback Fuzzy Models within a Multi-Innovation Context,” IEEE Transactions on Fuzzy Systems, p. 1 – 14, 2023.

[3] M. Ožbot, E. Lughofer, and I. Škrjanc, “Evolving Neuro-Fuzzy Systems-Based Design of Experiments in Process Identification,” IEEE Transactions on Fuzzy Systems, vol. 31, no. 6, p. 1995 – 2005, Jun. 2023.

[4] 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.

[5] D. Dovžan, V. Logar, and I. Škrjanc, "Implementation of an Evolving Fuzzy Model (eFuMo) in a Monitoring System for a Waste-Water Treatment Process," IEEE Transactions on Fuzzy Systems, vol. 23, no. 5, p. 1761 – 1776, Oct. 2015.

[6] 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.

[7] M. Almasri, Y.-H. Chang, I. E. Hajj, R. Nagi, J. Xiong, and W.-m. Hwu, " Parallelizing Maximal Clique Enumeration on GPUs," no. arXiv:2212.01473, Oct. 2023.

[8] Y.-W. Wei, W.-M. Chen, and H.-H. Tsai, "Accelerating the Bron-Kerbosch Algorithm for Maximal Clique Enumeration Using GPUs," IEEE Transactions on Parallel and Distributed Systems, vol. 32, no. 9, p. 2352 – 2366, Sep. 2021.

[9] E. Lughofer, C. Cernuda, S. Kindermann, and M. Pratama, "Generalized smart evolving fuzzy systems," Evolving Systems, vol. 6, no. 4, p. 269 – 292, Dec. 2015.

[10] E. Lughofer, M. Pratama, and I. Skrjanc, "Incremental Rule Splitting in Generalized Evolving Fuzzy Systems for Autonomous Drift Compensation," IEEE Transactions on Fuzzy Systems, vol. 26, no. 4, p. 1854 – 1865, Aug. 2018.

[11] I. Baidari and N. Honnikoll, "Bhattacharyya distance based concept drift detection method for evolving data stream," Expert Systems with Applications, vol. 183, p. 115303, Nov. 2021.

[12] L. Pardo, Statistical Inference Based on Divergence Measures. New York: Chapman and Hall/CRC, Oct. 2005.

[13] A. Salmona, J. Delon, and A. Desolneux, " Gromov-Wasserstein Distances between Gaussian Distributions," no. arXiv:2104.07970, Apr. 2021. [Online]. Available: http://arxiv.org/abs/2104.07970

[14] M. Różański, R. Wituła, and E. Hetmaniok, "More subtle versions of the Hadamard inequality," Linear Algebra and its Applications, vol. 532, p. 500 – 511, Nov. 2017.