摘 要

成对约束作为一种监督信息，比标签信息更加泛化，被应用于各类机器学习问题的解决。但是，可获得的初始成对约束关系是非常有限的，为了获得更多的成对约束关系，许多成对约束传播（Pairwise Constraint Propagation, PCP）方法应运而生。在现有的成对约束传播方法中，通常是在初始获得的成对约束关系是精确可靠的基础上，研究如何通过传播学习来增加成对约束的数量。然而，随着大数据和人工智能的快速发展，在现实场景中，数据信息的不确定性更加明显。对于成对约束而言，除了精确的信息以外，往往还存在着一些不确定的信息，这些信息对于成对约束的传播也具有重要意义。因此，如何利用这些不确定的成对约束信息来提高成对约束的传播效果是一个有待解决的问题。针对这一问题，本文的研究如下：

首先，通常情况下，可以获得的精确的成对约束关系极少，例如，当初始成对约束信息来源于用户反馈时，由于其没有掌握专业的知识和完善的判断方法，很难对提供的成对约束信息具有十足的把握，但是往往可以提供一些关于两个数据属于同一类别的可能性。针对这种指示两个样本类别相同的可能性的信息，提出了一种含不确定必须链接的成对约束的双对抗流形传播方法（Pairwise Constraint Propagation with Uncertain Must-link, UMPCP）。主要思想是在原有的精确的初始成对约束的基础上，使用一个矩阵添加一些关于不确定必须链接（must-link）的可能性，这个可能性在传播的过程中受不能链接（cannot-link）矩阵的对抗影响，另外，必须链接和不能链接本身在传播中也存在对抗关系，这两种对抗相结合，形成一个新的对抗强度，从而作用于成对约束的流形正则化传播中。通过实验证明，必须链接的可能性信息有效作用于传播过程，在提高了成对约束传播效果的同时，更符合现实场景的应用。

值得注意的是，不确定必须链接的加入只对必须链接的传播效果具有显著影响，而想要提高成对约束整体的传播效果，需要对不确定信息挖掘得更全面，所以，我们又关注到不能链接的可能性信息，并提出了一种不确定成对约束的双对抗流形传播方法（Uncertain Pairwise Constraint Propagation, UPCP）。在这一方法中，模仿 UMPCP 模型中对必须链接不确定性的定义，加入不能链接的可能性信息，将其中的双对抗结构，改造成两种可能性之间的对抗与两种约束之间的对抗相结合的形式，使必须链接和不能链接的对抗强度在竞争中趋于最小化。实验结果表明，UPCP 方法在提高传播精度的同时，克服了 UMPCP 模型对不确定信息应用的片面性。

关键词：成对约束；不确定信息；对抗关系；流形正则化；成对约束传播

Abstract

Pairwise constraints, as a type of supervised information, are more generalized than label information and are applied to solve various machine learning problems. However, the initial pairwise constraint relationships available are very limited. In order to obtain more pairwise constraint relationships, many pairwise constraint propagation (PCP) methods have emerged. In existing pairwise constraint propagation methods, it is usually studied how to increase the number of pairwise constraints through propagation learning based on the accurate and reliable initial obtained pairwise constraint relationships. However, with the rapid development of big data and artificial intelligence, the uncertainty of data information becomes more apparent in real scenarios. For pairwise constraints, in addition to precise information, there is often some uncertain information that is also of great significance for the propagation of pairwise constraints. Therefore, how to utilize these uncertain pairwise constraint information to improve the propagation effect of pairwise constraints is a problem that needs to be solved. In response to this issue, the research in this article is as follows:

Firstly, in general, there are very few precise pairwise constraint relationships that can be obtained. For example, when the initial pairwise constraint information comes from user feedback, it is difficult to have a full grasp of the provided pairwise constraint information due to the lack of professional knowledge and comprehensive judgment methods. However, it often provides some possibility that two data belong to the same category. A pairwise constraint propagation with uncertain must-link (UMPCP) method is proposed to address the possibility of two sample categories being the same. The main idea is to use a matrix to add some possibilities about uncertain must-links based on the original precise initial pairwise constraints. This possibility is influenced by the adversarial effects of the cannot-link matrix during the propagation process. In addition, there is also an adversarial relationship between the must-link and the cannot-link in the propagation process. These two adversaries are combined to form a new adversarial strength. Thus, it acts on the propagation of manifold regularization with pairwise constraints. Through experiments, it has been proven that the possibility information of must-link effectively acts on the propagation process, improving the propagation effect of pairwise constraints while being more in line with real-world applications.

It is worth noting that the addition of uncertain must-links only has a significant impact on the propagation effect of must-links. To improve the overall propagation effect of pairwise constraints, it is necessary to mine uncertain information more comprehensively. Therefore, we also focus on the possibility information of cannot-links and propose an uncertain pairwise constraint propagation (UPCP) method for uncertain pairwise constraints. In this method, by imitating the definition of must-link uncertainty in the UMPCP model, the possibility

information of cannot-link is added, and the dual adversarial structure is transformed into a combination of the adversarial between two possibilities and the adversarial between two constraints, so as to minimize the intensity of must-link and cannot-link adversarial in competition. The experimental results show that the UPCP method improves propagation accuracy while overcoming the one-sidedness of UMPCP model in applying uncertain information.

Keywords: pairwise constraint; uncertain information; adversarial relationship; manifold regularization; pairwise constraint propagation

目　录

第一章 绪论.....1
1.1 研究背景及意义.....1
1.2 国内外研究现状.....3
1.2.1 单模态数据中成对约束的应用.....3
1.2.2 多模态数据中成对约束的应用.....5
1.3 本文主要研究内容和章节安排.....6
第二章 相关理论介绍.....8
2.1 成对约束.....8
2.2 流形正则化.....10
2.2.1 流形假设.....10
2.2.2 k 近邻图的构造.....11
2.2.3 图的正则化.....12
2.3 成对约束传播.....13
2.3.1 成对约束传播的定义及作用.....13
2.3.2 成对约束传播与半监督学习的关系.....13
2.3.3 成对约束传播方法介绍.....14
2.3.4 成对约束传播方法的评价指标.....16
2.4 本章小结.....17
第三章 含不确定必须链接的成对约束的双对抗流形传播方法.....18
3.1 引言.....18
3.2 含不确定必须链接的成对约束的双对抗流形传播模型.....19
3.3 UMPCP 模型的优化求解.....21
3.3.1 优化求解过程.....21
3.3.2 UMPCP 算法总结.....23
3.4 实验过程与结果分析.....25
3.4.1 数据集与实验环境介绍.....25
3.4.2 评价指标.....25
3.4.3 参数设置.....26
3.4.4 实验设计与结果.....26
3.5 本章小结.....30
第四章 不确定成对约束的双对抗流形传播方法.....31
4.1 引言.....31
4.2 不确定成对约束的双对抗流形传播模型.....32
4.2.1 UPCP 模型的建立.....32

4.2.2 UPCP 模型的优化求解.....34
4.2.3 UPCP 算法总结.....35
4.3 UPCP 算法的收敛性分析.....36
4.4 实验过程与结果分析.....38
4.4.1 数据集介绍.....38
4.4.2 评价指标与参数设置.....39
4.4.3 实验设计.....39
4.4.4 实验结果.....40
4.5 本章小结.....45
第五章 工作总结与展望.....47
5.1 论文工作总结.....47
5.2 未来工作展望.....47
参考文献.....49

第一章 绪论

1.1 研究背景及意义

机器学习作为人工智能的核心技术，旨在让计算机通过学习经验数据，自动改进和发展算法，并从中提取出规律和模式，以实现预测和决策的能力。在当今科技飞速发展的时代背景下，机器学习涉及的范围非常广泛，更是包括医疗、交通等与人类生活密切相关的领域。机器学习的研究背景可以追溯到二十世纪五十年代，但是当时的计算机能力和存储空间有限，难以处理复杂的数据和算法，直到近些年才取得了可喜的进展，众多机器学习算法喷涌而出。机器学习算法目前可以分为三大类，分别为监督学习（Supervised Learning）、无监督学习（Unsupervised Learning）和半监督学习（Semi-Supervised Learning）。顾名思义，三类算法的区分标准在于监督信息的有无，其中监督学习是利用大量有标记的样本训练模型，并用训练好的模型对新的输入进行预测；无监督学习是指对没有标记的样本数据进行学习，从中寻找规律，发现数据的内在结构；半监督学习则是介于监督学习和无监督学习之间，常用于有少量的有标记样本和大量的无标记样本的情况。三类机器学习算法各有各的特点，其具体应用也有所区别，图 1-1 中直观地比较了三类算法。在现实场景中，信息纷繁复杂，受各种现实条件的制约，很多数据信息的获取并不容易，并且获取到的数据信息的准确度也有所差异，多数情况下我们可获得的监督信息比较符合半监督学习的前提条件，相比之下，半监督学习在现实应用中具有广泛的应用场景，涉猎领域较多，发展前景广阔。

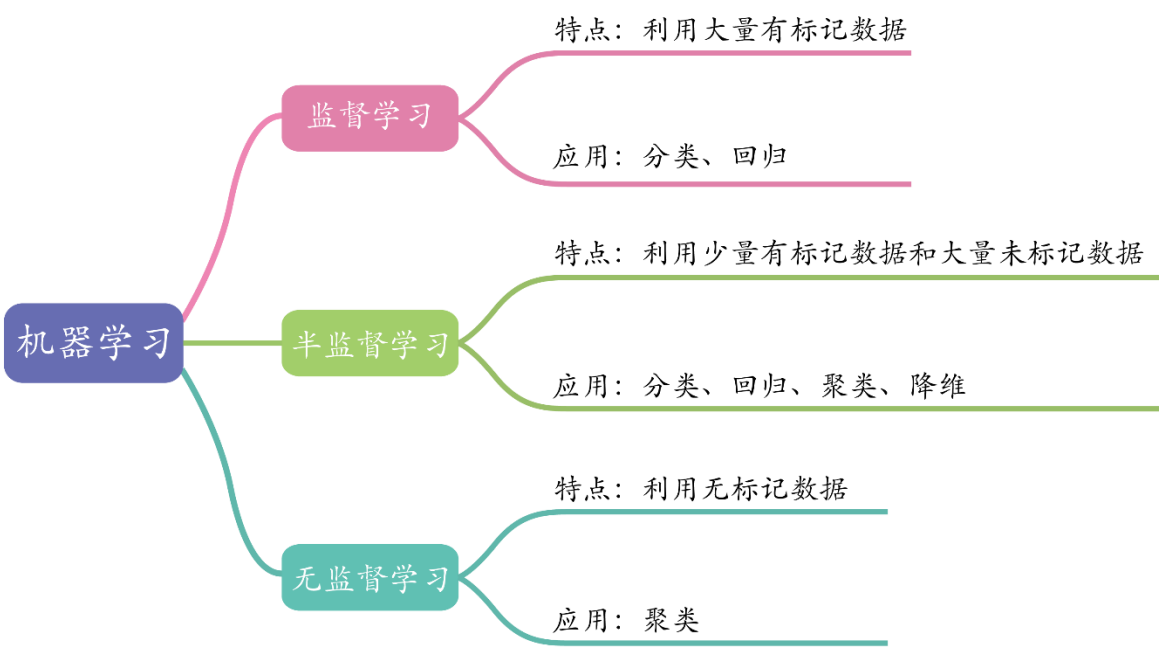
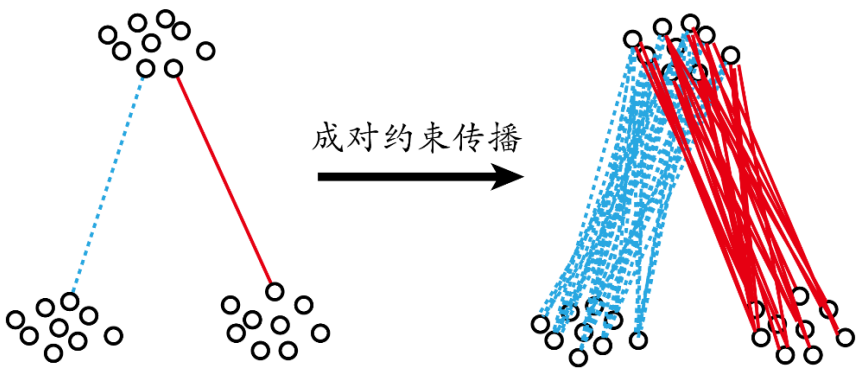


图 1-1 机器学习算法分类

常用的监督信息有标签信息、特征信息和成对约束信息三种类型。类标签很显然是标记样本属于哪一类别的信息，每个样本对应一个标签；特征信息通常是一个向量，用于描述样本的属性特征；成对约束信息与标签信息类似，都是描述有关样本类别的信息，但是没有标签信息那么具体，它是表示两个样本之间的关系，即仅提供两个样本属于同一类或者不属于同一类的信息，而不提供样本属于哪一类的信息。相比之下，成对约束对信息的挖掘更浅显、更概括，提供的信息也更弱，是一种边信息，所以成对约束也被称作弱监督信息[1]。但是正因为成对约束的上述特点，其相较于标签信息来说更容易获取，可以节省大量的人力、物力和时间成本，所以在近年来被广泛应用于机器学习任务中，例如约束聚类[2-4]、度量学习[5, 6]、特征选择[7, 8]、降维[9]和图像分割[10, 11]等。

成对约束信息通常可以通过领域知识[12, 13]、机器学习或者用户反馈来获得，例如，对于用户来说，当场景中包含了大量的目标（如图像）时，虽然让其直接给出每个目标的类别是很困难的，但是他们往往可以判断出两个目标是否属于同一类别。尽管如此，由于很多数据存在缺失、信息不完整，或者在一些场景中数据涉及到个人隐私而受保护等原因，在实际应用中，可获得的初始成对约束关系仍然是极少的。我们不能期望少量的成对约束信息能够告诉我们一切，所以如何在初始成对约束的数量有限的情况下，使机器学习任务获得更好的性能，是一个非常有意义和研究价值的问题。

在这样的背景下，“成对约束传播”的概念被提出，用来扩充成对约束的数量，以充分利用初始少量的成对约束信息。成对约束数量的扩充过程可以简单表示为图 1-2。成对约束传播相比于标签传播更具挑战性，因为对于一个具有 n 个数据点的数据集来说，成对约束传播需要推测出 $O(n^2)$ 的成对约束关系，而标签传播只需要推测出 $O(n)$ 的数据标签[14]。近年来，针对单模态数据和多模态数据提出了许多成对约束传播方法，这些方法在模式识别、计算机视觉和多媒体内容分析等不同的挑战性任务中取得了较好的结果。



(a)初始成对约束　　(b)传播后的成对约束

图 1-2 成对约束关系传播。黑色圆圈代表数据点，蓝色虚线和红色实线分别代表两种约束关系。

通过分析总结，我们观察到现有的成对约束传播方法大多都是使用精确的初始成对约束信息，即初始成对约束是一定存在的、确定的，其约束关系必须满足，不可违反，通常也称之为硬成对约束[15]。但是，我们正处于信息爆炸的时代，首先，信息的来源呈

现多元化，信息可以来自互联网、传媒、社交网络等各种渠道，不同来源的信息通常具有不同的可信度和权威性，其次，信息的格式呈现多元化，信息可以采用不同的格式表示，例如文字、图片、音频、视频等，不同的人对于不同形式的信息的获取程度、关注内容各有差异等，以上所呈现的信息系统复杂多样的现状，不可避免地造成信息模糊不确定的现象。受不确定信息的启发，我们关注到成对约束传播方法的初始成对约束信息的不确定问题，原有的成对约束传播方法对信息精确度的高要求限制了初始成对约束的表达形式，即仅用 1 和 0 两个元素来表示成对约束是否存在，其单一的形式也进而限制了成对约束传播在现实场景中的应用。

针对上述问题，我们对一些潜在的初始成对约束信息进行了研究，并挖掘出一部分不确定的初始成对约束信息，包含必须链接的可能性信息和不能链接的可能性信息。为了能够同时将确定的初始成对约束信息和不确定的初始成对约束信息应用于成对约束传播过程中，本文提出了不确定成对约束的对抗流形传播算法。该算法打破了原有算法在现实应用中的局限性，能够充分利用现实场景中获取的初始成对约束的可能性信息，一定程度上解决了初始成对约束的数量有限的问题，并且成对约束信息的不确定特征更加贴合实际，拓宽了成对约束传播在现实场景中的应用渠道。实验证明，本文提出的算法能够有效利用不确定成对约束信息提高成对约束传播的精度，因此，本文的研究具有一定的实际价值。

1.2 国内外研究现状

下面我们将针对单模态数据和多模态数据两个角度，分别介绍一下当前国内外对成对约束以及成对约束传播的研究情况及取得的成就。

1.2.1 单模态数据中成对约束的应用

所谓单模态数据是指数据集中每个数据点的表示形式只有一种。关于成对约束在各类机器学习任务中的应用研究，以及对于成对约束传播方法的探索，大多都是在单模态数据中进行的。下面将介绍一下成对约束在聚类、度量学习等机器学习任务中的探索与研究。

成对约束被广泛用于增强聚类效果，通常把这种方法称为约束聚类或半监督聚类。基于成对约束传播的约束聚类，其实是利用约束传播后的较多的成对约束关系，来指导整个数据集进行聚类分析。Kamvar 等人[16]将成对约束应用于聚类中时，是在传统谱聚类的基础上修改相似度矩阵的定义，当两个样本之间是必须链接关系时，相似度矩阵中对应的两个元素设置为 1，而当两个样本之间是不能链接关系时，则设置为 0，然后基于修改后的数据相似度利用谱聚类方法完成聚类分析。这是在聚类中应用成对约束最直接的方法，但是，由于该方法中刚性地把数据间的相似度调整为 1 或 0，其性能并不能达到令人满意的效果。Li 等人[17]提出了基于谱正则的约束聚类算法（Constrained Clustering with Spectral Regularization，CCSR），其主要思想是从一个光滑的数据表示空间中选择一个与给定的成对约束关系最一致的数据表示空间，通过这种方式，成对约束

可以传播到无约束对象，算法框架被表述为一个小的半定规划问题，其求解复杂度与数据集中的样本数目和给定的成对约束数目无关，可以扩展到大规模问题上。在 Li 等人的另一项工作中[18]，提出了一个谱核学习框架用来做约束聚类（Constrained Clustering by Spectral Kernel Learning，CCSKL），其与 CCSR 方法类似，关键思想是训练谱核以遵守给定的成对约束关系，谱核在一定程度上能够保持图的光滑结构，这个过程被表示为一个凸二次规划问题，可以有效地进行最优求解。Yang 等人[19]通过鼓励与一个必须链接相关联的两个样本的低维表示彼此相似，提出了一种基于非负矩阵分解（Nonnegative Matrix Factorization，NMF）的方法，但是这个方法中忽略了不能链接关系。为了将不能链接关系也融入到 NMF 中，Zhang 等人[20]提出了一种新的基于 NMF 的约束聚类框架，该框架强制必须链接上的两个样本之间的相似性近似为 1，不能链接上的两个样本之间的相似性近似为 0，并将约束直接集成到目标函数中。但是，当可用的成对约束数量不足时，该算法会降低 NMF 的聚类性能。Cucuringu 等人[21]通过将约束聚类公式化为广义特征值的问题，来同时利用必须链接和不能链接。另外，还可以通过利用成对约束对聚类的损失函数进行正则化来提高聚类的性能，例如，Jia 等人[22]基于给定的成对约束，构造了一个具有反块对角特征的矩阵，并进一步利用该矩阵进行正则化，整个模型被表述为一个约束优化问题，是一种半监督谱聚类模型。Jia 等人[23]还提出了基于图拉普拉斯主成分分析（Graph-Laplacian Principal Component Analysis，GLPCA）的凸约束聚类方法，该方法使用图拉普拉斯主成分分析来提取特征，并将数据投影到低维空间中，以成对约束的形式引入少量的监督信息来正则化低维表示之间的相似性，从而使得到的低维表示具有很强的可区分性，该方法可以很好的捕捉输入数据样本的局部和全局特征，但是它只使用了不能链接信息而忽略了必须链接信息的价值。

当成对约束应用于度量学习时，通过保持属于必须链接关系的两个数据点彼此接近，而使属于不能链接关系的两个数据点彼此分离，来保持语义相似性，从而学习到最优距离。基于成对约束的度量学习最早是由 Xing 等人[24]提出的，其主要是利用了必须链接关系，如果两个样本之间存在必须链接关系，那么这两个样本之间的距离保持相近，在提出的模型中，把度量表示成一个马氏（Mahalanobis）距离，通过增加同类样本之间的成对约束求解一个凸优化问题。Schultz 等人[25]认为 Xing 等人直接让属于必须链接的样本之间的距离小于一个值的方法过于刚性，进而提出了一种改进方案，让样本对之间进行比较，具体来说，如果样本 a 和 b 存在必须链接关系，而 b 和 c 存在不能链接关系，那么可以认为在学到的 Mahalanobis 距离度量中，a 和 b 之间的距离要比 b 和 c 之间的距离近。之后，在成对约束的方向上相继出现了一些度量学习方法。Bar-Hillel 等人[26]使用必须链接关系来学习数据点之间的最优 Mahalanobis 距离，提出了相关成分分析（Relevant Component Analysis，RCA）算法，RCA 是一个建立在信息论基础上的优化问题。随后，Davis 等人[27]也提出了一种信息论度量学习方法（Information-Theoretic Metric Learning，ITML）来学习 Mahalanobis 距离函数，该方法中将问题公式化为在距离函数的约束下最小化两个多元高斯之间的微分相对熵，并将这个问题的求解表示为一个特殊的 Bregman 优化问题，该方法可以处理各种各样的约束，包括必须链接、不能链

接以及距离对之间的关系。Weinberger 等人[28]提出了一种称为大间隔最近邻（Large Margin Nearest Neighbor，LMNN）的距离度量学习方法，它认为 Mahalanobis 距离可以等价地被视为在使用欧几里德距离的 k 近邻（k-Nearest Neighbor, k-NN）分类之前的输入空间的全局线性变换，线性变换是通过最小化两项组成的损失函数而得到的，第一项是惩罚同一类中被期望作为 k 近邻的样本之间的大距离，第二项是惩罚标签不匹配的样本之间的小距离。最小化这些项将产生输入空间的线性变换，这样可以增加其 k 近邻具有匹配标签的训练样本的数量。Chechik 等人[29]提出了一种在线可伸缩图像相似性估计算法（Online Algorithm for Scalable Image Similarity，OASIS），可应用于大规模的图像数据集。在该算法中，从共享公共标签或响应于公共文本查询而检索的图像对中提取相似性信息。因此，OASIS 可以学习一个独立于类的相似性度量，而不需要类标签。在 Yu 等人[30]的工作中，提出了一种语义保持距离度量学习（Semantic Preserving Distance Metric Learning，SP-DML）算法，利用用户提供的成对约束来描述两幅图像之间的语义关系，在统一的特征空间中挖掘视觉特征与成对约束的互补性。在该空间中，学习的距离度量可以用于测量两个图像之间的相异度，具体地说，SP-DML 采用的流形结构是由图像的视觉特征来揭示的，为了在距离度量学习中集成语义内容，SP-DML 利用成对约束构建语义块，并对齐这些块以获得新特征空间的最优距离度量。

另外，在特征选择[31-33]中，成对约束还可以用来指导模型选择哪些特征进行建模，基于约束关系来限制某些特征之间的相关性或相似性，常见的基于成对约束的特征选择方法包括成对特征选择、成对约束规则学习和成对稀疏建模等。在降维[34-36]中，成对约束主要通过控制特征之间的相关性来减少数据的维度，常见的基于成对约束的降维方法有主成分分析、因子分析和独立成分分析等。

与此同时，国内对于成对约束在机器学习任务中的应用也做了一些研究。在聚类中，王纵虎等人[37]为了在文本聚类算法中应用成对约束，将成对约束信息扩展后嵌入文档相似度矩阵，提出了一种半监督文本聚类算法。李晁铭等人[38]为了利用少量成对约束指导聚类，引入样本交叉熵表达成对约束信息，提出基于成对约束的交叉熵半监督聚类算法。陶性留等人[39]采用非负矩阵分解提取样本的本质特征，并且加入成对约束条件指导聚类过程进行模糊聚类。在徐圣兵等人的工作中[40]，也提出了一种模糊聚类方法，利用对象交叉熵测度工具，提出最小−最大交叉熵隶属度学习准则，并作为成对约束信息测度项引入到成对约束核聚类的目标函数中。肖成龙等人[41]在谱图理论的图正则项中引入成对约束与流形正则化构造相似度矩阵，通过与自编码器的联合学习，提出基于流形正则化与成对约束的深度半监督谱聚类算法。另外，还有黄华等人[42] 在半监督分类方法中引入成对约束信息，刘笑等人[43]利用成对约束进行度量学习等。

1.2.2 多模态数据中成对约束的应用

多模态数据[14]是指由两种或两种以上模态组成的数据。具体来说，对于一个描述对象，可以通过不同领域或视角获取到数据，并把描述这些数据的每一个领域或视角叫做一个模态。前面介绍的都是单模态数据中成对约束的应用方法，而在现实场景中，还存

在着许多具有多种模态的数据集，尤其是随着多媒体技术的应用，日常生活中出现了大量的多媒体数据，这些多媒体数据都属于多模态数据[14]。例如，来自社交媒体（例如微博）的图片，通常可以分别基于视觉特征和用户提供的文本标签来使用两个单独的模态进行表示。因此，对于成对约束在多模态数据中的应用问题，国内外也做了广泛的研究。

Fu 等人[44]最初提出了一种多模态成对约束传播方法（Multi-Modal Pairwise Constraint Propagation，MMCP），其为多模态数据的每个模态都构建了一个相似图，然后在这些图上定义了随机游走过程，可以使用随机游走过程图来计算多个图上的任意两个节点之间的转移概率，根据这种转移概率，认为任意一对具有高的转移概率的节点应该具有关于数据集中其他数据点相似的成对约束配置，即具有相似的成对约束传播结果。这一思想可以表示为一系列独立的基于多重图的约束传播子问题，再应用图正则化框架通过一系列二次优化来解决这些子问题。该方法证明了约束传播子问题集可以统一为一个单一的二次优化问题来求解，并且存在一个封闭解。但是，该方法具有一定的局限性，它无法处理数据点在某些模态上缺失的情况，而且 MMCP 本质上是将多个模态图融合到一起，而不是在每一个模态上建立独立的学习。针对 MMCP 存在的问题，Fu 等人又进行了另一项工作[45]，提出了一种模态一致性学习框架，首先将每一个数据模态上分离的成对约束关系合并到一起，形成一个单独的针对多模态数据集的优化问题，然后引入了一个新的一致性正则化项，迫使不同数据模态上的传播结果尽可能保持一致，在此基础上，形成两个定义良好的多模态算法，这两个算法可以通过交替优化的方式进行求解，这样就可以将初始的成对约束关系传播到整个多模态数据集中。Lu 等人[46]考虑到了跨模态的异构成对约束传播的问题，关注到不同数据模态的约束设置不一定是一致的。该文章将异构约束传播分解为一组独立的半监督学习子问题，然后通过基于图的标签传播来有效地解决这些子问题。更重要的是，该工作中进一步开发了一种约束稀疏表示方法，用于在使用同构成对约束的每个数据模态上构造关联图。也就是说，该方法可以同时利用异构和同构的成对约束。在 Lu 等人的另一项工作中[47]，提出了一种用于多视图数据上的视图内和视图间约束传播的统一框架。视图间的成对约束是指在来自不同视图的一对数据点上定义的必须链接和不能链接关系，但是由于来自不同视图的数据点之间的相似性通常是未知的，所以视图间的成对约束传播具有很大的挑战性，为了解决这一问题，文中将视图内或视图间的成对约束传播问题分解为多个半监督学习子问题，将这些子问题统一地表示为正则化能量泛函的极小化问题，为了进一步将它们集成到一个统一的框架中，该方法中利用视图内成对约束传播的结果来调整每个视图的相似性矩阵，然后使用调整后的相似性矩阵进行视图间的成对约束传播。

1.3 本文主要研究内容和章节安排

本文研究的内容主要分为五章，每章具体内容概括如下:

第一章为绪论部分，首先介绍了本文的研究背景以及研究意义，随后分析了本文课题的国内外研究现状和发展趋势。研究背景及意义部分阐述了成对约束在机器学习任务

中的重要性，并且针对现有成对约束传播方法分析了不足之处，本文主要从现实应用性出发，对成对约束传播的初始成对约束信息进行研究。国内外研究现状部分从单模态数据和多模态数据两个方面分别介绍了成对约束的应用与发展。最后，对本文的研究内容和框架安排做了简单介绍。

第二章主要介绍了一些与本文研究内容相关基础理论和常用方法。首先介绍了成对约束的定义、表示方法以及自身特点。然后介绍了流形正则化的构造过程，包括本质思想、前提假设和基础图构造等。随后介绍了成对约束传播的定义以及作用，并且列举出现有的具有代表性的几种成对约束传播方法，分析了其优点与不足，并给出了成对约束传播方法常用的评价指标，为下文算法的提出奠定了理论基础。

第三章中考虑到现实场景中可获得的成对约束的不确定性问题，将成对约束的对抗流形传播方法中引入了必须链接的可能性信息，提出了一种含不确定必须链接的成对约束对抗流形传播方法。该方法考虑了必须链接和不能链接的差异性，使两种约束关系分开进行传播，同时加入不确定的必须链接，使其与确定的两种约束关系一起，为传播过程提供更多初始信息，共同作用于传播结果。并且在多个数据集上的结果中表明，加入的必须链接的可能性信息对成对约束传播效果的提高具有重要作用。

第四章是进一步对有关成对约束的不确定信息进行探索，在第三章的基础上，又加入了有关不能链接的一些潜在的可能性信息，提出了不确定成对约束的双对抗流形传播方法。该方法把初始的必须链接可能性和不能链接可能性都融入到传播过程中，另外使两种可能性之间形成对抗，加之原有的约束对抗，在传播过程中一起调整对抗强度，调节传播性能。通过实验证明，不能链接的可能性信息在原模型的基础上又为成对约束传播提供了一部分初始信息，不仅可以提高传播精度，也实现了对不确定信息更详尽的应用。

第五章对本文的研究内容及取得的效果进行了归纳总结，并进一步讨论了当前研究中仍然存在的问题和不足，进而对未来的研究方向及目标做出大致规划。

第二章 相关理论介绍

2.1 成对约束

成对约束的作用类似于类标签，是一种监督信息，但是通常情况下，在现实生活中类标签的获得往往需要花费大量的人力、物力，而获得一对数据点间的二元关系会容易许多，所以在这种背景下，Wagstaff 提出“成对约束”的概念[48]，来表示一对数据点间的关系。成对约束包括两种约束关系，一种关系表示的是两个数据点一定属于同一类别，被称为必须链接，另一种关系则表示两个数据点一定不属于同一类别，被称为不能链接。

根据以上描述，我们得知这两种约束关系是确定的、强制的关系，那么可以采用二元组的形式来表示两种约束，并将同种约束存储在集合中，所以设必须链接的集合为 M，不能链接的集合为 C，两个集合的数学表示如下：

$$M=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\,\boldsymbol{x}_i,\boldsymbol{x}_j\text{属于同一类}\} \tag{2.1}$$

$$C=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\,\boldsymbol{x}_i,\boldsymbol{x}_j\text{属于不同类}\} \tag{2.2}$$

其中 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 是来自同一数据集的两个数据点。在应用成对约束关系时，若已知一对数据点 $(\boldsymbol{x}_i,\boldsymbol{x}_j)$，如果有 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M$，则称 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间存在必须链接，二者一定属于同一类别，对应的，如果有 $(\boldsymbol{x}_i,\boldsymbol{x}_j)\in C$，则称 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间存在不能链接，二者一定属于不同类别。除此之外，两种约束关系也可以在同一个矩阵中表示，如下所示：

$$Z_{i,j}=\begin{cases}1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\text{属于同一类}\\ -1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\text{属于不同类}\\ 0, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\text{未标记}\end{cases} \tag{2.3}$$

矩阵 $\boldsymbol{Z}$ 中的 1、-1、0 元素分别表示当前位置所关联两个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间是必须链接、不能链接和未标记关系。图 2-1 中将两种约束关系进行了具象的表示。

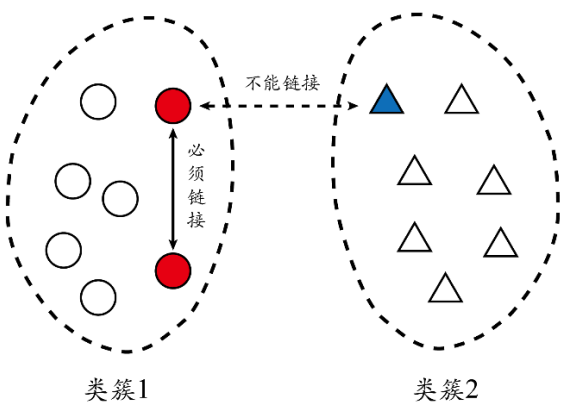


图 2-1 成对约束关系

同样作为监督信息，类标签与成对约束之间存在一定的推导关系，但是这种关系是单向性的。一般情况下，可以从数据的类标签中推出一对数据点的成对约束关系，即当两个数据点的类标签相同时，可以推出二者之间是必须链接关系，而当两个数据点有不同的类标签时，则二者之间是不能链接关系。但是反过来，这种推导关系就不成立，比如已知两个数据点之间存在必须链接关系，但是我们并不可以得知两个数据点的类标签是什么、二者具体是属于哪一类的信息，尤其对于不能链接来说这种现象更明显。从类标签与成对约束的关系来看，与类标签相比较，成对约束的信息强度更弱，是一种弱监督信息。

通过对成对约束特点的观察，我们可以总结出其具有以下两个特性：

1、不可传递性。如图 2-2 所示，图中 3 张图片的图像标注已经给出，通过标注我们可以获得 3 张图片之间的成对约束关系。(a)和(c)两张图片的标注中都有 plane 和 sky 的描述，所以认为二者之间存在必须链接关系，同时(a)和(b)、(b)和(c)并没有相同的标注信息，可以认为二者之间存在不能链接关系。很显然，(a)和(b)、(b)和(c)之间的不能链接关系并不能传递推出(a)和(c)之间存在不能链接关系，所以我们称成对约束具有不可传递性。



plane, sky, city (a)　　car, tree, road (b)　　plane, sky, cloud (c)

图 2-2 成对约束的不可传递性图。必须链接：(a,c)；不能链接：(a,b)，(b,c)。

2、对称性[49]。如果已知数据点 $\boldsymbol{x}_i$ 和数据点 $\boldsymbol{x}_j$ 之间存在某种约束关系，那么等价于数据点 $\boldsymbol{x}_j$ 和数据点 $\boldsymbol{x}_i$ 之间也存在相同的约束关系，这就是成对约束的对称性。其数学表示为：

$$(\boldsymbol{x}_i, \boldsymbol{x}_j) \in M \ \ \& \ \ (\boldsymbol{x}_j, \boldsymbol{x}_i) \in M \tag{2.4}$$

$$(\boldsymbol{x}_i, \boldsymbol{x}_j) \in C \ \ \& \ \ (\boldsymbol{x}_j, \boldsymbol{x}_i) \in C \tag{2.5}$$

成对约束的对称性在前面介绍的成对约束关系的矩阵表示形式中会呈现得更加清晰，图 2-3 中给出一个关于 7 个数据点的成对约束矩阵，很明显，矩阵以红线圈出的主对角线为对称轴，上三角和下三角元素形成对称。另外，对称性在成对约束传播中也是不可忽略的重要特性。

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0	-1	1	0	0	0	1
x_2	-1	0	0	-1	0	0	0
x_3	1	0	0	0	0	-1	0
x_4	0	-1	0	0	0	0	0
x_5	0	0	0	0	0	0	0
x_6	0	0	-1	0	0	0	-1
x_7	1	0	0	0	0	-1	0

图 2-3 成对约束矩阵的对称性

2.2 流形正则化

2.2.1 流形假设

流形正则化其实就是在机器学习问题中的正则化项中加入与流形相关的项，利用数据分布的几何结构，起到半监督的作用。在半监督学习中有三种常用的基本假设来建立预测样本和学习目标之间的关系，分别为平滑假设（Smoothness Assumption）[50]、聚类假设（Cluster Assumption）[51]和流形假设（Manifold Assumption）[52]。平滑假设是指位于稠密数据区域的两个距离很近的样本的类标签相似，也就是说，当两个样本被稠密数据区域中的边连接时，它们有很大可能有相同的类标签；相反地，当两个样本被稀疏数据区域分开时，它们的类标签趋于不同。聚类假设是指当两个样本位于同一聚类簇时，它们在很大的概率下有相同的类标签。流形假设是将高维数据嵌入到低维流形中，当两个样本位于低维流形中的一个小局部邻域内时，它们具有相似的类标签。从本质上说，这三类假设是一致的，只是各自关注的重点不同，其中流形假设更具普遍性，流形正则化也正是基于流形假设提出的。

具体来讲，流形假设是假定数据分布的几何结构由基础流形支持，例如黎曼流形（Riemannian manifold）[53]等，流形假设主要考虑的是模型的局部特性，它假设数据在高维空间中的分布不是随机的，而是在局部具有欧氏空间的性质，能够保持局部的平滑性，也就是说，处在一个很小的局部邻域内的样本有着相似的性质，从而使其标签也相似。合理使用流形假设，可以尽可能多地利用无监督数据，让数据空间变得更加稠密，从而有助于更加准确地刻画局部区域的特性，使决策函数可以更好地进行数据拟合，从而提高机器学习任务的性能。

通常情况下，我们可以使用图来描述流形。假设给定一个由 n 个 m 维数据样本组成的数据集 $\boldsymbol{X}$，将其第 i 列记为 $\boldsymbol{x}_i \in \mathbb{R}^{m\times 1}$，那么输入数据表示为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}\in\mathbb{R}^{m\times n}$。我们假定数据集 $\boldsymbol{X}$ 的数据分布的几何结构由基础流形支持，可以将其映射为一个无向加权图 $G=<v,\varepsilon>$，每一个顶点代表一个数据样本，所以 v 是由 $\boldsymbol{X}$ 中所有数据样本组成的

顶点集，两个样本之间的相似度用加权边来度量，所以 ε 是描述 $\boldsymbol{X}$ 中两两样本之间相似程度的边集。由于流形假设要求图中的标签平滑地变化，所以，如果在图中有两个顶点由一条强边连接起来，那么这两个样本通常会具有相似的标签。

2.2.2 *k* 近邻图的构造

对于上一小节中提到的图 G，有几种构造方式，比如 ε 近邻图、全连接图、k 近邻图等，由于稀疏图通常不会导致较大的计算量，可以带来更好的性能，所以在流形正则化中采用 k 近邻法构造数据点的局部近邻关系，并把图中样本之间的边权关系，记录于数据集的关联矩阵中，这样的图称之为 k-NN 图。

k-NN 图的构造过程如下：首先选择一个合适的 k 值，在数据集 $\boldsymbol{X}$ 中寻找与每个样本最靠近的 k 个样本，以 $\boldsymbol{x}_i$ 为例，分别计算 $\boldsymbol{x}_i$ 与与其他样本之间的距离，并按照距离递增的次序排序，选取与 $\boldsymbol{x}_i$ 距离最小的 k 个样本，并将其记录为 $N_k(\boldsymbol{x}_i)$。其中的距离计算一般采用欧氏距离，计算公式为

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\sum_{l=1}^{m}(x_{l,i}-x_{l,j})^2} \tag{2.6}$$

其中 m 表示样本数据的维度。当每个样本的 k 个近邻确定后，每个样本在图中对应的顶点与其他 k 个顶点相连即形成边，由此生成 k-NN 图。图 2-4 展示了 k-NN 图构造的一个例子，当 k=3 时，$\boldsymbol{x}_i$ 的 k 近邻为实线圆圈内的 3 个数据点，当 k=5 时，$\boldsymbol{x}_i$ 的 k 近邻为虚线圆圈内的 5 个数据点。

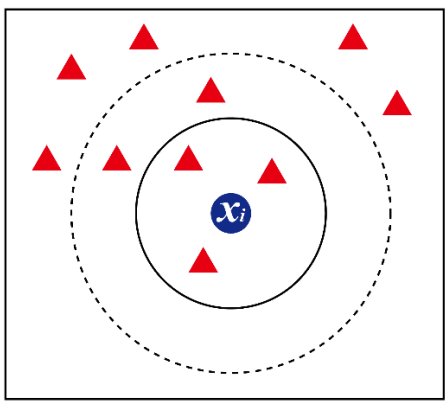


图 2-4 k-NN 图构造过程示例

k-NN 图中两个顶点之间的边权值是非负的，可以采用一个边权矩阵 $\boldsymbol{W}\in\mathbb{R}^{n\times n}$ 来记录，计算方法通常使用径向基核函数(Radial Basis Function, RBF)计算，生成一个位于区间 $(0,1]$ 的实值，其定义如下：

$$\boldsymbol{W}_{i,j}=\begin{cases}\exp(-\dfrac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}), & \text{if}\quad \boldsymbol{x}_j\in N_k(\boldsymbol{x}_i)\\ 0, & \text{otherwise}\end{cases} \tag{2.7}$$

其中 $\sigma>0$ 是高斯核的宽度。如公式所示，当 $\boldsymbol{x}_j$ 属于 $\boldsymbol{x}_i$ 的 k 个近邻之一时，如果两个样

本之间的距离越小，其构成的边权值越大，当 $\boldsymbol{x}_j$ 不属于 $\boldsymbol{x}_i$ 的 k 个近邻之一时，两个样本之间不存在边，则 $\boldsymbol{W}$ 中对应的元素均简单地设置为 0，这样可以减少图构造时噪声的产生。另外，为了避免求解过程中出现自增强效应，权值矩阵主对角线上的元素都设定为 $W_{i,i}=0$。因为 k-NN 图是对称的，需要用公式 $\boldsymbol{W}=(\boldsymbol{W}+\boldsymbol{W}^T)/2$ 对初步获得的权值矩阵进行对称处理。

2.2.3 图的正则化

流形正则化的本质思想就是通过流形学习，迫使嵌入的数据样本与临近点具有相似的性质，在新的投影空间中能够保持原来的局部线性结构，从而得到非线性的流形全局结构。已知上述 $\boldsymbol{W}_{i,j}$ 表示 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 两个数据样本之间在流形结构上的相似度，给定 k-NN 图，流形正则化的描述及推导过程如下：

$$\begin{aligned}
&\sum_{i,j}^{n}\boldsymbol{W}_{i,j}\,\|\,\boldsymbol{x}_i-\boldsymbol{x}_j\,\|^2\\
&=\sum_{i,j}^{n}\boldsymbol{W}_{i,j}\boldsymbol{x}_i^T\boldsymbol{x}_i+\sum_{i,j}^{n}\boldsymbol{W}_{i,j}\boldsymbol{x}_j^T\boldsymbol{x}_j-2\sum_{i,j}^{n}\boldsymbol{W}_{i,j}\boldsymbol{x}_i^T\boldsymbol{x}_j\\
&=\sum_{i}^{n}\boldsymbol{D}_{i,i}\boldsymbol{x}_i^T\boldsymbol{x}_i+\sum_{j}^{n}\boldsymbol{D}_{j,j}\boldsymbol{x}_j^T\boldsymbol{x}_j-2\sum_{i,j}^{n}\boldsymbol{W}_{i,j}\boldsymbol{x}_i^T\boldsymbol{x}_j\\
&=2\sum_{i}^{n}\boldsymbol{D}_{i,i}\boldsymbol{x}_i^T\boldsymbol{x}_i-2\sum_{i,j}^{n}\boldsymbol{W}_{i,j}\boldsymbol{x}_i^T\boldsymbol{x}_j\\
&=2tr(\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T)-2tr(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^T)\\
&=2tr(\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^T)
\end{aligned}\tag{2.8}$$

其中 $\boldsymbol{D}=\mathrm{diag}(\boldsymbol{d}_1,\boldsymbol{d}_2,...,\boldsymbol{d}_n)$ 是一个对角矩阵，它的第 i 个对角元素计算为 $\boldsymbol{d}_i=\sum_{j=1}^{n}W_{i,j}$ [54]，也就是边权值矩阵 $\boldsymbol{W}$ 对应行的元素的和；$\boldsymbol{L}\in\mathbb{R}^{n\times n}$ 是拉普拉斯矩阵，由于 $\boldsymbol{W}_{i,i}=0$，所以其结构可以表示为：

$$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}=\begin{pmatrix}
\sum_{j=1}^{n}\boldsymbol{W}_{1,j} & -\boldsymbol{W}_{1,2} & -\boldsymbol{W}_{1,3} & \cdots & -\boldsymbol{W}_{1,n}\\
-\boldsymbol{W}_{2,1} & \sum_{j=1}^{n}\boldsymbol{W}_{2,j} & -\boldsymbol{W}_{2,3} & \cdots & -\boldsymbol{W}_{2,n}\\
-\boldsymbol{W}_{3,1} & -\boldsymbol{W}_{3,2} & \sum_{j=1}^{n}\boldsymbol{W}_{3,j} & & \vdots\\
\vdots & \vdots & & \ddots & -\boldsymbol{W}_{n-1,n}\\
-\boldsymbol{W}_{n,1} & -\boldsymbol{W}_{n,2} & \cdots & -\boldsymbol{W}_{n,n-1} & \sum_{j=1}^{n}\boldsymbol{W}_{n,j}
\end{pmatrix}\tag{2.9}$$

$tr(\cdot)$ 表示矩阵的迹(trace)，其定义为矩阵主对角线上各个元素的总和，即 $tr(\boldsymbol{Z})=\sum_{i=1}^{n}\boldsymbol{Z}_{i,i}$。流形正则化项可以作为一种平滑项，提升模型的泛化能力。

2.3 成对约束传播

2.3.1 成对约束传播的定义及作用

在解决实际问题时，初始可获得的成对约束关系是非常有限的，而成对约束作为先验知识对于提高聚类、特征选择等机器学习任务的性能具有重要作用。所以，如何获得大量的成对约束关系是一个备受关注的问题，许多成对约束传播方法[55, 56]也随之提出。成对约束传播方法的原理是通过将已知的少量的初始成对约束关系在数据集上进行传递，来获取大量的可靠的数据点之间的成对约束信息，再利用这样大量的成对约束关系完成机器学习任务。很显然，相比初始的少量的成对约束信息，传播后大量的可靠的成对约束信息将给机器学习任务带来更高的性能。成对约束传播已然成为充分利用有限数量的成对约束的一种有效方法。

成对约束传播的具体定义可以这样描述：当给定一个包含 n 个数据样本的数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}\in\mathbb{R}^{m\times n}$ 时，已知其中少量的成对数据样本之间的关系，即初始成对约束信息，分别保存在必须链接集合 M 和不能链接集合 C 中，其他未知的成对约束关系就是成对约束传播中需要求解的未知量。当 M 和 C 中的初始成对约束关系通过一个模型泛化到整个数据集中时，使得未知的成对约束关系都推测出一个估计值，可以通过这个估计值来判断每对约束属于必须链接关系还是不能链接关系，从而增加可靠的成对约束的数量，这就是成对约束的传播问题。

2.3.2 成对约束传播与半监督学习的关系

从成对约束传播问题的初始条件来看，成对约束传播具有半监督学习的结构[57]。因为在半监督学习问题中，是利用少量的有标签数据去预测大量未标记数据的标签，类似的，成对约束传播问题也是利用少量的初始成对约束关系，去预测数据集中大量的未知成对约束关系。

可以通过以下简单的示例来更加清晰地呈现成对约束传播的半监督结构。依据 2.1 节中给出的成对约束矩阵定义，图 2-5 中展现的是一个数据集的成对约束矩阵，其中的 1、-1、0 元素分别表示当前位置所关联的两个数据点之间是必须链接关系、不能链接关系和未标记关系。数据集中一共包含 7 个数据点，当我们关注于其中某一个数据点时，例如图中做出标记的 $\boldsymbol{x}_3$，关于 $\boldsymbol{x}_3$ 的成对约束传播是整个数据集的一个成对约束传播子问题，在 $\boldsymbol{x}_3$ 与其他数据点的成对约束关系中，有少量关系已知，大部分是未知的，这个子问题实际上就具有半监督结构，是一个关于两类的半监督学习问题[14]，其中元素 1 代表的必须链接关系是正类，而元素-1 代表的不能链接关系就是负类，元素 0 的位置就是即将被分类的成对约束关系，已知的成对约束关系在垂直方向和水平方向上进行传播，得到关于 $\boldsymbol{x}_3$ 的未知成对约束关系的预测结果，也是半监督学习的结果。当数据集中关于所有数据点的成对约束传播子问题都得到求解后，所有的解合并在一起便是整个数据集的成对约束传播结果。

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0	-1	1	0	0	0	1
x_2	-1	0	0	-1	0	0	0
x_3	1	0	0	0	0	-1	0
x_4	0	-1	0	0	0	0	0
x_5	0	0	0	0	0	0	0
x_6	0	0	-1	0	0	0	-1
x_7	1	0	0	0	0	-1	0

图 2-5 成对约束传播的半监督结构

数据之间的成对约束传播问题虽然类似于半监督学习问题，但也存在着明显的区别，因为我们前面提到过，成对约束具有对称性。成对约束的对称性不仅表现在初始的成对约束信息中，同样在传播后的成对约束关系也应该有明显的体现，具体来说，如果在求解的结果中显示 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间存在必须链接关系，那么对应的，结果中的 $\boldsymbol{x}_j$ 和 $\boldsymbol{x}_i$ 之间也应表现为必须链接关系，不能链接关系也相同。所以，在成对约束传播模型求解时，必须把对称性考虑在内。

2.3.3 成对约束传播方法介绍

随着成对约束传播问题得到的关注越来越多，许多成对约束传播方法被提出，在这一节中，将简单介绍几个具有代表性的成对约束传播方法。

（1） 基于半定规划(Semi-definite Programming，SDP)的成对约束传播方法

Li 等人[58]最初提出了一种成对约束传播方法，如果两个数据点之间如果有必须链接关系，那么它们之间的相似度就取最大值(比如 1)，相反的，如果是不能链接关系，那么相似度就取最小值(比如 0)。具体在模型表示中，将该方法描述为一个半定规划问题，公式如下：

$$\begin{aligned} &\min_{\boldsymbol{K}} tr(\boldsymbol{LK}) \\ &\text{s.t. } k_{i,i}=1,\ i=1,2,\ldots,n \\ &\quad k_{i,j}=1, \forall(x_i,x_j)\in M \\ &\quad k_{i,j}=0, \forall(x_i,x_j)\in C \\ &\quad \boldsymbol{K}\geq 0 \end{aligned} \tag{2.10}$$

其中，$\boldsymbol{L}\in\mathbb{R}^{n\times n}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ 是标准化后的拉普拉斯矩阵，它是对称的并且半正定的，$\boldsymbol{K}\in\mathbb{R}^{n\times n}$ 是传播后的相似度矩阵。对于 $\boldsymbol{K}$ 中的元素，当其值为 1 时，表明对应于一个必须链接，其值为 0 时，则对应于一个不能链接。

该模型可以用于多类分类，并且可以在统一的框架中处理数据标签、成对约束或两种信息的混合，其有效性在分类和约束聚类等应用中得到了验证，但是，半正定约束会导致极大的计算成本，计算复杂度高达 $O(n^4)$，这使得模型不具有实际应用性。

（2） 流形嵌入的成对约束传播方法

为了降低基于 SDP 的成对约束传播方法的高计算复杂度，Lu 等人[59]提出了一种穷举且有效的成对约束传播方法(Exhaustive and Efficient Constraint Propagation，E^{2}CP)。该方法假设两个相似的样本具有相似的成对关系，将成对约束传播问题分解为一组独立的半监督分类子问题，模型表达为

$$\min_{\boldsymbol{F}} \left\| \boldsymbol{F}-\boldsymbol{Z} \right\|_F^2+\frac{\mu}{2}tr(\boldsymbol{F}^T\boldsymbol{L}\boldsymbol{F}+\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^T) \tag{2.11}$$

其中 $\boldsymbol{Z}\in\mathbb{R}^{n\times n}$ 是初始的成对约束关系矩阵，其定义为

$$Z_{i,j}=\begin{cases} 1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in M \\ -1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in C \\ 0, & \text{otherwise} \end{cases} \tag{2.12}$$

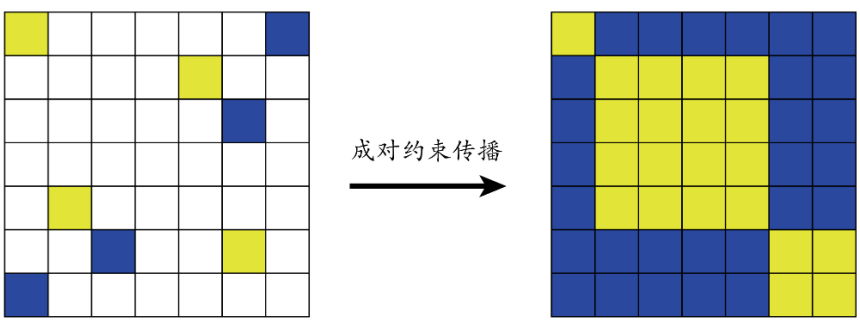
另外，$\boldsymbol{F}\in\mathbb{R}^{n\times n}$ 是传播后的成对约束关系矩阵，μ 是一个正则化参数，用来调整正则化项在模型中影响的大小。该方法的时间复杂度是 $O(n^2)$，与基于 SDP 的方法相比时间成本要低很多。

（3） 基于低秩建模的成对约束传播方法

在 E^{2}CP 中采用了流形正则化，是利用了两个相似样本应该具有相似的成对约束关系的假设，这种原理也可以用低秩假设来表示，也就是假设成对约束关系矩阵是一个低秩矩阵。基于低秩假设，Fu[60]提出了一种低秩矩阵补全的成对约束传播方法，其中初始的成对约束关系作为观测变量，未知的成对约束关系作为缺失变量。该方法的数学表达为：

$$\begin{aligned} &\min_{\boldsymbol{F},\boldsymbol{A},\boldsymbol{E}} rank([\boldsymbol{A},\boldsymbol{F}])+\lambda\left\|\boldsymbol{E}\right\|_1 \\ &s.t.\ \ \boldsymbol{D}=\boldsymbol{A}+\boldsymbol{E} \\ &\quad \boldsymbol{F}_{i,j}=\boldsymbol{Y}_{i,j},\forall(x_i,x_j)\in M\cup C \end{aligned} \tag{2.13}$$

其中 $\boldsymbol{Y}\in\mathbb{R}^{n\times n}$ 是初始的成对约束关系矩阵，其定义与 E^{2}CP 方法中的矩阵 $\boldsymbol{Z}$ 相同。$\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 是输入数据的低秩分量，$\boldsymbol{E}\in\mathbb{R}^{n\times n}$ 是噪声分量，$\boldsymbol{F}\in\mathbb{R}^{n\times n}$ 则是传播后的成对约束关系矩阵。联合低秩项 $rank([\boldsymbol{A},\boldsymbol{F}])$ 表示使两个矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{F}$ 共同低秩。低秩补全的过程展示在图 2-6 中。该模型虽然取得了较好的效果，但是需要有足够多的初始成对约束才能正常工作。



(a)初始成对约束矩阵　　(b)低秩补全矩阵

图 2-6 低秩补全示意图。黄色和蓝色分别代表成对约束矩阵中的+1（必须链接）和-1（不能链接），空白处代表待预测的成对约束关系。

2.3.4 成对约束传播方法的评价指标

为了判断成对约束传播算法传播效果的好坏，通常采用的几种评价指标是准确率、精确率、召回率和 F1 分数。在介绍这几种评价指标之前，需要先了解一下计算各个评价指标时所涉及的关于预测结果的统计术语。

表 2-1 二分类混淆矩阵

真实情况	预测情况	
	正例	负例
正例	TP（真正例）	FN（假负例）
负例	FP（假正例）	TN（真负例）

如表 2-1 所示，几个术语的定义如下：TP（True Positive）表示实际为正例，在预测结果中也被正确预测为正例的样本数量；FP（False Positive）表示实际为负例，在预测结果中被错误预测为正例的样本数量；TN（True Negative）表示实际为负例，在预测结果中也被正确预测为负例的样本数量；FN（False Negative）表示实际为正例，在预测结果中被错误预测为负例的样本数量。

（1）准确率（accuracy）

accuracy 是最常见也是最基本的评价指标，判定被正确预测样本占所有预测样本的比例，其计算公式为 $\text{accuracy}=\frac{\text{TP+TN}}{\text{TP+FP+TN+FN}}$。但是，在正负样本不平衡的情况下，其参考价值并不高。

（2）精确率（precision）

precision 是相对于预测结果而言的，用来表示在所有预测为正例的样本中，实际为正例的样本的比例，计算公式为 $\text{precision}=\frac{\text{TP}}{\text{TP+FP}}$，其可以理解为对模型预测结果的可信度度量。

（3）召回率（recall）

recall 是相对于真实标记而言的，用来表示实际为正例的样本中，被预测正确的样本所占的比重，其计算公式为 $\text{recall}=\frac{\text{TP}}{\text{TP+FN}}$，相当于对模型预测结果覆盖范围的考察。

（4）F1 分数（F1-score）

precision 和 recall 两个指标通常是此消彼长的，很难兼得，在大规模数据集中相互制约，这样就需要综合考虑，最常见的方法就是 F1 值，其计算公式为 $\text{F1}=\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}$，F1 值综合了 precision 和 recall 两个指标，对模型预测结果的评价更全面。

需要注意的是，以上评价指标各自的侧重点不同，所以对于不同问题，要根据实际情况进行选择和调整。举例来说，在垃圾邮件屏蔽系统中，垃圾邮件为正例，正常邮件为负例，其目的追求正常邮件可以被接收到，即使有一部分垃圾邮件被判定为正常邮件（FN），也不希望正常邮件进入垃圾箱（FP），即 FP 的值要尽量小，那么该系统需要尽可能提高 precision 值，尽管要牺牲一部分 recall 值。

2.4 本章小结

本章是本文研究内容的基础理论，首先介绍了成对约束的定义表示以及特点，然后介绍了流形正则化的前提假设、构造过程以及数学表达，接着介绍了成对约束传播的原理、作用，以及现有的具有代表性的成对约束传播方法，并给出了成对约束传播算法的评价指标的详细内容，以上理论介绍都为下文算法的提出奠定了基础。

第三章 含不确定必须链接的成对约束的双对抗流形传播方法

3.1 引言

成对约束被认为是一种弱监督信息[61]，因为相比于标签信息，成对约束不能提供显式的类信息，而是一种综合的监督信息。具体来说，成对约束不是标记每一个数据样本属于哪一类，而是标记一对数据样本之间的关系，即由必须链接和不能链接两类约束信息来说明两个数据样本是否属于同一类。成对约束可以作为一种先验知识改进模型来提高机器学习任务的性能。在很多机器学习任务中，成对约束比类标签使用起来更简单。但是通常可用的成对约束通常是非常有限的，因此成对约束的传播就至关重要。成对约束传播就是将有限可用的成对约束关系从已知约束的数据样本对传播到未知约束的数据样本对，从而获得大量可靠的成对约束信息，应用于各种机器学习任务后，能够显著提高其性能。

为了将初始成对约束信息传播到整个数据集中，前人做了大量的工作，提出很多方法。例如，Lu 等人[62]通过高斯核函数将已知的约束传播到其他相似的未知约束的数据点，但是该方法做了特定假设，仅针对两个类的聚类问题。Li 等人[58]将成对约束传播表示为半定规划过程，不局限于两个类的问题，但是半定规划的求解过程会产生极大的开销，计算复杂度达到$O(n^4)$。Yu 等人[63]把成对约束传播表示为一个约束优化问题，但只局限于必须链接的优化。Lu 等人[64]为了克服上述问题，提出一种不限于两类问题和必须链接，并且能够降低计算复杂度的方法，将成对约束传播分解为一组相互独立的半监督分类子问题。为了增强传播中两种约束的可辨别性，Jia 等人[65]提出一种基于对抗流形正则化的成对约束传播模型，将必须链接和不能链接分开进行传播，并且加入其对抗关系。

前面提到的很多成对约束传播模型都是利用精确的初始成对约束来传播，即要求获得的初始约束关系是百分之百存在，并且在传播过程中要严格遵守的，也就是硬成对约束。而在现实世界中，想要获取精确的成对约束存在很大的困难，很多约束关系来自现实环境，其具有弱关系[33]，所以成对约束关系存在不确定性。很显然只追求精确的初始成对约束时并没有考虑到现实中获取成对约束的精准性，导致在现实情境下的传播会忽略一部分不精确信息。例如当初始成对约束信息是从用户反馈中获得时，用户对于信息的判断并不权威，信息的精准性不能保证，其给出的初始约束关系也达不到硬成对约束的标准。假设现有两个病例的症状信息，不同个体出现的症状差异和病患个人表达能力的偏差等都会成为判断两个病例是否患同一种病的不确定因素，我们不能得出两个病例一定患有同一种病的结论，但可以判断二者患有同一种病的可能性。

针对上述问题，在获取初始成对约束时，允许不确定性将是一种好的解决方案。在本章中，我们关注于必须链接的不确定性，并以对必须链接进行放松的方式引入不确定

信息，为此，定义了一种表达必须链接不确定度的新形式，但其并没有摒弃传统必须链接的经典定义，而是在保留精确的成对约束的情况下，容许加入一部分未被利用的必须链接的可能性，综合全部的初始约束信息，并将其应用于必须链接与不能链接分开进行传播的对抗流形正则化模型中。该方法实现了对初始成对约束信息的扩充，也考虑到了必须链接与不能链接的差异性，同时也贴合现实场景，具有更加广泛的应用范围。

本章的主要贡献如下：

1、关注到从现实场景中想要获取两个对象之间明确的约束关系的困难，提出不确定必须链接的新形式来描述两个对象属于同一类的可能性大小。

2、将必须链接与不能链接的传播分开进行，在保持二者差异性的同时，加入二者的对抗关系，并使新加入的不确定必须链接也参与对抗过程，形成一个双对抗的新结构，提出了含不确定必须链接的成对约束的双对抗流形传播方法。

3、在多个数据集上的结果表明，不确定必须链接信息的加入有效改善了成对约束对抗流形传播的效果。

本章的其余部分组织如下：第二节介绍了 UMPCP 模型的建立过程，第三节中给出了算法优化求解的方法，并给出了算法求解流程，第四节展示了实验验证过程，并给出结果分析，最后，在第五节中对本章提出的方法做了总结。

3.2 含不确定必须链接的成对约束的双对抗流形传播模型

为了将不确定的必须链接信息融入到成对约束的传播中，首先，我们对式(2.1)中必须链接的集合定义做出修改，将 M 改为用来记录初始的确定的必须链接和不确定的必须链接的集合，也可以理解为在理想状态下为必须链接的样本对组成的集合，其表示如下：

$$M=\{(\boldsymbol{x}_i,\boldsymbol{x}_j)\,|\,\boldsymbol{x}_i,\boldsymbol{x}_j\text{理想状态下属于同一类}\} \tag{3.1}$$

另外，不能链接集合 C 的定义保持不变。

不同的链接信息对于提高成对约束传播的准确性都至关重要，如果将必须链接和不能链接定义在同一个矩阵中，如公式(2.3)所示，一定程度上忽略了必须链接和不能链接的差异性，因此，为了实现两种约束关系能够各自进行传播，给定一个由 n 个 m 维数据样本组成的输入数据矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_n\}\in\mathbb{R}^{m\times n}$，根据已获得的初始成对约束信息，定义两个矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ 和 $\boldsymbol{B}\in\mathbb{R}^{n\times n}$，分别记录初始的必须链接信息和不能链接信息：

$$\boldsymbol{A}_{i,j}=\begin{cases}1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in M\\ 0, & \text{otherwise}\end{cases} \tag{3.2}$$

$$\boldsymbol{B}_{i,j}=\begin{cases}1, & (\boldsymbol{x}_i,\boldsymbol{x}_j)\in C\\ 0, & \text{otherwise}\end{cases} \tag{3.3}$$

以上两个矩阵的定义虽然在形式上看起来相同，但是由于必须链接集合 M 的改变，两个矩阵内包含的信息存在一定区别，其中矩阵 $\boldsymbol{A}$ 中值为 1 的元素包括确定的必须链接和不确定的必须链接两种，而矩阵 $\boldsymbol{B}$ 中值为 1 的元素仅包括确定的不能链接。

为了将矩阵 $\boldsymbol{A}$ 中的确定信息与不确定信息做区分，并且能够表达不确定度的大小，下

面，我们先定义一个新的概念：

定义 1 当两个数据样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 在理想状态下存在必须链接时，根据客观经验，可以判断两个样本属于同一类的可能性为 $p_{i,j}$，那么称这样表示的必须链接为不确定必须链接。

根据定义 1，可以定义一个相似可能性矩阵 $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ 来表示初始必须链接的不确定度，其表示为：

$$\boldsymbol{U}_{i,j} = \begin{cases} u_{i,j}, & \text{if } \boldsymbol{A}_{i,j} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{3.4}$$

其中 $u_{i,j}$ 是根据与现实场景相关的知识判断得到，例如用户反馈、专家知识等途径，其数值是区间 $(0.5,1]$ 内的一个实值。具体来说，$u_{i,j}$ 所在位置关联的两个数据样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间可能存在初始必须链接关系，$u_{i,j}$ 的数值是衡量该必须链接存在的可能性大小。因为只有在可能性大于 0.5 时，两个数据样本之间存在必须链接的不确定性才相对可靠，故其取值范围设定在 $(0.5,1]$。值得注意的是，当 $u_{i,j}$ 的值取 1 时，那么其关联的两个数据样本之间的必须链接是确定的。

为了记录传播中的成对约束，我们使用两个非负矩阵 $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ 和 $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ 分别记录传播过程中的必须链接和不能链接，在基于数据样本构建的 k-NN 图的指导下，根据流形正则化，将必须链接和不能链接的传播分别表示为：

$$\begin{aligned} &\min tr(\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}^T) \\ &\text{s.t. } \boldsymbol{K} \geq 0 \\ &\quad \boldsymbol{K}_{i,j} = \boldsymbol{A}_{i,j}, \text{ if } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in M \end{aligned} \tag{3.5}$$

$$\begin{aligned} &\min tr(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^T) \\ &\text{s.t. } \boldsymbol{F} \geq 0 \\ &\quad \boldsymbol{F}_{i,j} = \boldsymbol{B}_{i,j}, \text{ if } (\boldsymbol{x}_i, \boldsymbol{x}_j) \in C \end{aligned} \tag{3.6}$$

其中 $\boldsymbol{L}$ 是拉普拉斯矩阵。

对于同一对样本，其间存在必须链接关系或不能链接关系是相对的，从而形成一种对抗关系，我们将这种对抗关系扩展到所有样本对中时，利用 Hadamard 乘积，可以表示整体的对抗强度，其数学表达为 $\|\boldsymbol{K} \circ \boldsymbol{F}\| = \sum_{i,j=1}^{n} \boldsymbol{K}_{i,j}\boldsymbol{F}_{i,j}$，其中 $\circ$ 符号即表示 Hadamard 乘积。根据 Hadamard 乘积的定义，对抗过程的原理是当同一对样本之间的 $\boldsymbol{K}_{i,j}$ 和 $\boldsymbol{F}_{i,j}$ 值差异较大时，表示样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的约束关系偏向于必须链接或不能链接其中一种，其乘积 $\boldsymbol{K}_{i,j}\boldsymbol{F}_{i,j}$ 的值就会较小，表明二者之间的对抗关系越弱，反之则对抗关系越强。

相应地，还需要考虑必须链接不确定度的传播，定义一个矩阵 $\tilde{\boldsymbol{U}} \in \mathbb{R}^{n\times n}$ 来记录传播中的相似可能性。利用初始必须链接的相似可能性矩阵 $\boldsymbol{U}$ 中的信息，可以将可能性传播表示为 $\|\tilde{\boldsymbol{U}} - \boldsymbol{U}\|^2$ 的形式，其可以将初始已知的有关必须链接的可能性信息泛化到其他未知可能性的必须链接中。另外，由于必须链接和不能链接之间存在对抗关系，那么必须链接

的可能性与不能链接之间也不可避免地存在对抗，同样采用 Hadamard 乘积，可以将这个对抗关系表示为$\|\boldsymbol{U}\circ\boldsymbol{F}\|$。那么，综合提出的两种对抗关系，可以使其相结合形成一种新的双对抗结构，并产生一种综合的对抗强度，作用于成对约束的传播过程。在数学表达中这种双对抗关系可以有多种形式，我们采用$\|\boldsymbol{U}\circ\boldsymbol{F}\|\|\boldsymbol{K}\circ\boldsymbol{F}\|$的形式来描述它。这种形式不仅能够综合不确定信息产生更真实的对抗强度，而且便于模型的交替优化求解，可以有效调节成对约束的传播效果。

综合两类约束关系的传播和必须链接可能性的传播，再加入双对抗过程，并且使其最小化，可以使相近的成对约束具有相似的约束关系，这样可以将模型表述为：

$$\begin{aligned}&\min\|\boldsymbol{U}\circ\boldsymbol{F}\|\|\boldsymbol{K}\circ\boldsymbol{F}\|+\alpha tr(\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}^T)+\beta tr(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^T)+\lambda\left\|\tilde{\boldsymbol{U}}-\boldsymbol{U}\right\|^2\\&\text{s.t. }\boldsymbol{K}\geq 0,\boldsymbol{F}\geq 0\\&\quad\boldsymbol{K}_{i,j}=\boldsymbol{A}_{i,j},\boldsymbol{F}_{i,j}=\boldsymbol{B}_{i,j},\ \text{if }(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M\cup C\end{aligned}\tag{3.7}$$

其中α和β是两个用来控制必须链接和不能链接传播强度的参数。因为每个数据集中各类样本所占的比例不同，样本之间潜在的必须链接和不能链接的比例也不能确定，为了避免某一种链接关系占比过高而主导传播过程，我们采用α和β以加权的形式来平衡必须链接与不能链接的传播强度。而λ是调节必须链接的可能性信息在模型中的贡献权重的参数。

3.3 UMPCP 模型的优化求解

3.3.1 优化求解过程

为了有效地处理模型约束条件中的等式约束，我们定义了一个位置矩阵$\boldsymbol{P}\in\mathbb{R}^{n\times n}$来协助模型的优化，其定义形式如下：

$$\boldsymbol{P}_{i,j}=\begin{cases}1,\ \text{if }(\boldsymbol{x}_i,\boldsymbol{x}_j)\in M\cup C\\0,\ \text{otherwise}\end{cases}\tag{3.8}$$

如矩阵表达式所示，$\boldsymbol{P}$记录了初始成对约束的位置，当$\boldsymbol{P}_{i,j}$的值为 1 时，表明数据样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的初始成对约束信息已知，即包含获取的确定的必须链接、不确定的必须链接和确定的不能链接的位置，当$\boldsymbol{P}_{i,j}$的值为 0 时，数据样本$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的初始成对约束信息未知。

因为已知成对约束很少而未知成对约束很多，为缓解过拟合，利用$\boldsymbol{P}$将两个等式约束表示成$\|\boldsymbol{P}\circ(\boldsymbol{K}-\boldsymbol{A})\|^2$和$\|\boldsymbol{P}\circ(\boldsymbol{F}-\boldsymbol{B})\|^2$的形式，引入已知的初始成对约束信息，迫使预测情况与真实情况尽可能接近，将模型优化为

$$\begin{aligned}\min_{\boldsymbol{K},\boldsymbol{F},\boldsymbol{U}}&\|\boldsymbol{U}\circ\boldsymbol{F}\|\|\boldsymbol{K}\circ\boldsymbol{F}\|+\alpha tr(\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}^T)+\beta tr(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^T)+\\&\gamma(\|\boldsymbol{P}\circ(\boldsymbol{K}-\boldsymbol{A})\|^2+\|\boldsymbol{P}\circ(\boldsymbol{F}-\boldsymbol{B})\|^2)+\lambda\left\|\tilde{\boldsymbol{U}}-\boldsymbol{U}\right\|^2\\&\text{s.t. }\boldsymbol{K}\geq 0,\boldsymbol{F}\geq 0\end{aligned}\tag{3.9}$$

为求解传播结果矩阵，对于模型中的不等式约束条件，使用拉格朗日乘数法进行处理，将上式的拉格朗日函数表示为：

$$\begin{aligned} O(\boldsymbol{K},\boldsymbol{F},\boldsymbol{U}) = &\left\|\boldsymbol{U}\circ\boldsymbol{F}\right\|\left\|\boldsymbol{K}\circ\boldsymbol{F}\right\| + \alpha tr(\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}^T) + \beta tr(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^T) + \\ &\gamma(\left\|\boldsymbol{P}\circ(\boldsymbol{K}-\boldsymbol{A})\right\|^2 + \left\|\boldsymbol{P}\circ(\boldsymbol{F}-\boldsymbol{B})\right\|^2) + \\ &\lambda\left\|\tilde{\boldsymbol{U}}-\boldsymbol{U}\right\|^2 + \sum_{i,j=1}^{n}\Phi_{i,j}\boldsymbol{K}_{i,j} + \sum_{i,j=1}^{n}\varphi_{i,j}\boldsymbol{F}_{i,j} \end{aligned} \tag{3.10}$$

其中 $\Phi\in\mathbb{R}^{n\times n}$ 是对应 $\boldsymbol{K}\geq 0$ 约束关系的拉格朗日乘数矩阵，$\varphi\in\mathbb{R}^{n\times n}$ 则是对应 $\boldsymbol{F}\geq 0$ 约束关系的拉格朗日乘数矩阵。

由于 $\sum_{i,j=1}^{n}\Phi_{i,j}\boldsymbol{K}_{i,j}$ 的值实际上就是矩阵 Φ 和矩阵 $\boldsymbol{K}$ 相乘后的结果矩阵的主对角线元素之和，所以可以应用矩阵的迹来替换 $\sum_{i,j=1}^{n}\Phi_{i,j}\boldsymbol{K}_{i,j}$ 这一项，表示为 $tr(\Phi^T\boldsymbol{K})$，同样地，$\sum_{i,j=1}^{n}\varphi_{i,j}\boldsymbol{F}_{i,j}$ 可以替换为 $tr(\varphi^T\boldsymbol{F})$。这样一来，拉格朗日函数的表示变为：

$$\begin{aligned} O(\boldsymbol{K},\boldsymbol{F},\boldsymbol{U}) = &\left\|\boldsymbol{U}\circ\boldsymbol{F}\right\|\left\|\boldsymbol{K}\circ\boldsymbol{F}\right\| + \alpha tr(\boldsymbol{K}\boldsymbol{L}\boldsymbol{K}^T) + \beta tr(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^T) + \\ &\gamma(\left\|\boldsymbol{P}\circ(\boldsymbol{K}-\boldsymbol{A})\right\|^2 + \left\|\boldsymbol{P}\circ(\boldsymbol{F}-\boldsymbol{B})\right\|^2) + \\ &\lambda\left\|\tilde{\boldsymbol{U}}-\boldsymbol{U}\right\|^2 + tr(\Phi^T\boldsymbol{K}) + tr(\varphi^T\boldsymbol{F}) \end{aligned} \tag{3.11}$$

首先，对于上述拉格朗日函数分别求出关于链接矩阵 $\boldsymbol{K}$ 和 $\boldsymbol{F}$ 的一阶偏导，可得：

$$\begin{aligned} \frac{\partial O}{\partial \boldsymbol{K}_{i,j}} &= (\left\|\tilde{\boldsymbol{U}}\circ\boldsymbol{F}\right\|\boldsymbol{F} + \alpha\boldsymbol{K}(\boldsymbol{L}+\boldsymbol{L}^T) + 2\gamma\boldsymbol{P}\circ(\boldsymbol{K}-\boldsymbol{A}) + \Phi)_{i,j} \\ &= (\left\|\tilde{\boldsymbol{U}}\circ\boldsymbol{F}\right\|\boldsymbol{F} + 2\alpha\boldsymbol{K}\boldsymbol{L} + 2\gamma\boldsymbol{P}\circ\boldsymbol{K} - 2\gamma\boldsymbol{P}\circ\boldsymbol{A} + \Phi)_{i,j}, \forall\{i,j\} \end{aligned} \tag{3.12}$$

$$\begin{aligned} \frac{\partial O}{\partial \boldsymbol{F}_{i,j}} &= (\tilde{\boldsymbol{U}}\boldsymbol{K} + \beta\boldsymbol{F}(\boldsymbol{L}+\boldsymbol{L}^T) + 2\gamma\boldsymbol{P}\circ(\boldsymbol{F}-\boldsymbol{B}) + \varphi)_{i,j} \\ &= (\tilde{\boldsymbol{U}}\boldsymbol{K} + 2\beta\boldsymbol{F}\boldsymbol{L} + 2\gamma\boldsymbol{P}\circ\boldsymbol{F} - 2\gamma\boldsymbol{P}\circ\boldsymbol{B} + \varphi)_{i,j}, \forall\{i,j\} \end{aligned} \tag{3.13}$$

由 KKT（Karush-Kuhn-Tucker）条件可知，关于必须链接矩阵 $\boldsymbol{K}$ 和不能链接矩阵 $\boldsymbol{F}$ 的一阶偏导均等于 0，即 $\frac{\partial O}{\partial \boldsymbol{K}_{i,j}}=0$，$\frac{\partial O}{\partial \boldsymbol{F}_{i,j}}=0$，在求解等式时，在两个等式两边分别乘以 $\boldsymbol{K}_{i,j}$ 和 $\boldsymbol{F}_{i,j}$，可得等式如下：

$$(\left\|\tilde{\boldsymbol{U}}\circ\boldsymbol{F}\right\|\boldsymbol{F})_{i,j}\boldsymbol{K}_{i,j} + 2\alpha(\boldsymbol{K}\boldsymbol{L})_{i,j}\boldsymbol{K}_{i,j} + 2\gamma(\boldsymbol{P}\circ\boldsymbol{K})_{i,j}\boldsymbol{K}_{i,j} - 2\gamma(\boldsymbol{P}\circ\boldsymbol{A})_{i,j}\boldsymbol{K}_{i,j} + \Phi_{i,j}\boldsymbol{K}_{i,j} = 0 \tag{3.14}$$

$$(\tilde{\boldsymbol{U}}\boldsymbol{K})_{i,j}\boldsymbol{F}_{i,j} + 2\beta(\boldsymbol{F}\boldsymbol{L})_{i,j}\boldsymbol{F}_{i,j} + 2\gamma(\boldsymbol{P}\circ\boldsymbol{F})_{i,j}\boldsymbol{F}_{i,j} - 2\gamma(\boldsymbol{P}\circ\boldsymbol{B})_{i,j}\boldsymbol{F}_{i,j} + \varphi_{i,j}\boldsymbol{F}_{i,j} = 0 \tag{3.15}$$

由于在另一 KKT 条件中 $\Phi_{i,j}\boldsymbol{K}_{i,j}=0$，$\varphi_{i,j}\boldsymbol{F}_{i,j}=0$，并且将拉普拉斯矩阵 $\boldsymbol{L}$ 使用对角矩阵 $\boldsymbol{D}$ 和权值矩阵 $\boldsymbol{W}$ 进行替换，即 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，所以上述等式可以变化为：

$$\begin{aligned} &(\left\|\tilde{\boldsymbol{U}}\circ\boldsymbol{F}\right\|\boldsymbol{F})_{i,j}\boldsymbol{K}_{i,j} + 2\alpha(\boldsymbol{K}\boldsymbol{D})_{i,j}\boldsymbol{K}_{i,j} + 2\gamma(\boldsymbol{P}\circ\boldsymbol{K})_{i,j}\boldsymbol{K}_{i,j} - \\ &2\alpha(\boldsymbol{K}\boldsymbol{W})_{i,j}\boldsymbol{K}_{i,j} - 2\gamma(\boldsymbol{P}\circ\boldsymbol{A})_{i,j}\boldsymbol{K}_{i,j} = 0 \end{aligned} \tag{3.16}$$

$$(\tilde{\boldsymbol{U}}\boldsymbol{K})_{i,j}\boldsymbol{F}_{i,j}+2\beta(\boldsymbol{F}\boldsymbol{D})_{i,j}\boldsymbol{F}_{i,j}+2\gamma(\boldsymbol{P}\circ\boldsymbol{F})_{i,j}\boldsymbol{F}_{i,j}-$$
$$2\beta(\boldsymbol{F}\boldsymbol{W})_{i,j}\boldsymbol{F}_{i,j}-2\gamma(\boldsymbol{P}\circ\boldsymbol{B})_{i,j}\boldsymbol{F}_{i,j}=0 \tag{3.17}$$

依据上述等式关系，可以求出必须链接矩阵 $\boldsymbol{K}$ 和不能链接矩阵 $\boldsymbol{F}$ 的更新规则，其分别表示为：

$$\boldsymbol{K}_{i,j}=\boldsymbol{K}_{i,j}\frac{(2\gamma\boldsymbol{P}\circ\boldsymbol{A}+2\alpha\boldsymbol{K}\boldsymbol{W})_{i,j}}{(\left\|\tilde{\boldsymbol{U}}\circ\boldsymbol{F}\right\|\boldsymbol{F}+2\alpha\boldsymbol{K}\boldsymbol{D}+2\gamma\boldsymbol{P}\circ\boldsymbol{K})_{i,j}} \tag{3.18}$$

$$\boldsymbol{F}_{i,j}=\boldsymbol{F}_{i,j}\frac{(2\gamma\boldsymbol{P}\circ\boldsymbol{B}+2\beta\boldsymbol{F}\boldsymbol{W})_{i,j}}{(\tilde{\boldsymbol{U}}\boldsymbol{K}+2\beta\boldsymbol{F}\boldsymbol{D}+2\gamma\boldsymbol{P}\circ\boldsymbol{F})_{i,j}} \tag{3.19}$$

由于成对约束具有对称性，传播后的链接矩阵也应该保持对称，所以需要在传播后用公式 $(\boldsymbol{K}+\boldsymbol{K}^T)/2$ 和 $(\boldsymbol{F}+\boldsymbol{F}^T)/2$ 处理两个矩阵以保持对称性。

接下来，将求解必须链接的可能性矩阵 $\tilde{\boldsymbol{U}}$ 的更新规则，同样，先求拉格朗日函数 O 关于 $\tilde{\boldsymbol{U}}$ 的一阶偏导，可得：

$$\frac{\partial O}{\partial\tilde{\boldsymbol{U}}_{i.j}}=(\boldsymbol{F}\left\|\boldsymbol{K}\circ\boldsymbol{F}\right\|+2\lambda\tilde{\boldsymbol{U}}-2\lambda\boldsymbol{U})_{i,j},\forall\{i,j\} \tag{3.20}$$

根据 KKT 条件 $\frac{\partial O}{\partial\boldsymbol{U}_{i,j}}=0$，可以求出 $\tilde{\boldsymbol{U}}$ 的更新规则为：

$$\tilde{\boldsymbol{U}}_{i,j}=\boldsymbol{U}_{i,j}-\frac{(\boldsymbol{F}\left\|\boldsymbol{K}\circ\boldsymbol{F}\right\|)_{i,j}}{2\lambda} \tag{3.21}$$

由于相似可能性矩阵是非负的对称矩阵，更新后的 $\tilde{\boldsymbol{U}}$ 需要归一化为 0 到 1 之间，并处理成对称矩阵。

3.3.2 UMPCP 算法总结

综上，UMPCP 算法的详细求解步骤可以总结如下：

算法 3-1 UMPCP 算法

输入：参数 α，β，γ，λ，数据集的 k-NN 图结构的边权矩阵 $\boldsymbol{W}$，初始成对约束的位置矩阵 $\boldsymbol{P}$，初始必须链接矩阵 $\boldsymbol{A}$ 和初始不能链接矩阵 $\boldsymbol{B}$，必须链接的初始相似可能性矩阵 $\boldsymbol{U}$

初始化：$\boldsymbol{K}=\mathrm{rand}^{+}(n,n)$，$\boldsymbol{F}=\mathrm{rand}^{+}(n,n)$

$\tilde{\boldsymbol{U}}=\mathrm{rand}^{+}(n,n)$

（$\mathrm{rand}^{+}(n,n)$ 返回一个正的 $n\times n$ 的随机矩阵 ）

Step1: 计算对角矩阵 $\boldsymbol{D}$

计算拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$

Step2: 迭代更新 $\boldsymbol{K}$，$\boldsymbol{F}$，$\tilde{\boldsymbol{U}}$ 矩阵

While 式(3.9)不收敛 do

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/245023003003011114