一种针对多关系数据的半监督协同训练算法

王 娇[1] 罗四维[2] 王 立[1]

(中央广播电视大学计算机科学与技术系 北京 100031)[1]

(北京交通大学计算机应用研究所 北京 100044)[2]

摘 要 半监督学习是机器学习领域的研究热点。协同训练研究数据有多个特征集时的半监督学习问题。将图表示法引入协同训练,使用多个图结构表示多关系数据。在每个图上进行半监督学习,在多个图之间进行协同学习,使多个图上的学习器对数据的预测一致。创新性地提出一种针对多关系数据的半监督协同训练算法,并从概率角度分析学习过程。在真实数据集上的实验表明,提出的算法处理多关系数据时具有较好的性能。

关键词 机器学习,半监督学习,协同训练,多关系数据

中图法分类号 TP181 **文献标识码** A

New Co-training Algorithm for Multi-relation Data

WANG Jiao[1] LUO Si-wei[2] WANG Li[1]

(Department of Computer Science and Technology, The Open University of China, Beijing 100031, China)[1]

(Research Institute of Computer Application, Beijing Jiaotong University, Beijing 100044, China)[2]

Abstract Semi-supervised learning is a hot research topic of machine learning. Co-training is a multi-view semi-supervised learning method. Graph representation was introduced to co-training where multiple graphs were used to represent multi-relation data. Semi-supervised learning was processed on each graph while co-training was conducted between graphs to ensure the predictions of graphs are the same. A new co-training algorithm for multi-relation data was proposed, and it was analyzed from the viewpoint of probability. Encouraging experimental results are gotten from real world multi-relation dataset.

Keywords Machine learning, Semi-supervised learning, Co-training, Multi-relation data

1 引言

半监督学习综合利用标记数据和未标记数据进行学习,以提高学习器的性能,是近年来机器学习领域的研究热点[1,2]。协同训练是一类半监督学习算法,最早由 Blum and Mitchell[3] 提出,它针对数据有多个视图的情况,使用未标记数据来辅助学习。Blum and Mitchell 在研究网页分类问题时,将网页本身包含的信息作为一个视图,将超链接所包含的信息作为另一个视图,从而用两个视图进行半监督学习,获得了较好的效果,协同训练算法由此受到研究者的关注[4,5]。

机器学习中常用图结构对数据进行建模。图可以用二元组 $G=(V,E)$ 表示,其中 V 代表图上的顶点,E 代表图上的边。图上的顶点和实际问题中的数据点一一对应,图上的边表示数据之间的关系,而关系的强弱则用边的权值 W 来表示。图的学习能够通过与其对应的矩阵运算来进行,因而基于图的机器学习方法具有较好的数学基础,近年来基于图结构的半监督学习方法也日益增多[6]。用图结构表示数据的优点在于图能够刻画数据之间的关系。

越来越多的实际问题涉及分析与处理多关系的数据。例如对于学术论文的分类问题,一篇学术论文包括标题、作者、摘要、参考文献等多个描述,论文之间既有相似关系,也有参考文献之间的引用关系,还有作者之间的合作关系。如何对这类多关系数据进行学习是机器学习面临的挑战之一。对此问题的研究工作主要集中在聚类问题上,例如可以使用相容二部图[7]对多关系数据进行联合聚类。

本文把图表示法引入半监督协同训练中,研究多关系数据的分类问题。将多关系数据用不同的图结构表示,提出一种基于图的半监督协同训练算法。在每个图上使用半监督学习方法进行学习,并在多个图之间协同学习,以使多个图上的学习器对未标记数据的预测相似。从概率角度分析多个图上的学习过程。在真实数据上的实验表明,提出的算法比单个图上的学习算法有更好的分类性能。

2 一种针对多关系数据的半监督协同训练算法

2.1 数据的图表示

基于图结构的机器学习方法首先要构建一个图,图上的顶点表示数据点,图上的边表示数据点之间的关系,而边的权值大小则表示数据点之间关系的强弱。一般假设连接权值大

本文受国家自然基金项目(60975078)资助。

王 娇(1982-),女,博士,讲师,主要研究方向为机器学习、模式识别,E-mail:wang0828@gmail.com;罗四维(1943-),男,博士,教授,主要研究方向为神经计算;王 立(1978-),男,博士,讲师,主要研究方向为并行计算。

的数据具有较高的相似性，应当具有相同的类别。由于图上的边能够反映数据之间的关系，因此通过图的构建能够把数据集的结构信息包含到图的结构之中。

本文用多个图结构来表示多关系数据。例如对于学术论文数据集，可以根据论文之间的相似关系构建一个图，再根据参考文献之间的引用关系构建一个图，从而用两个图结构来表示同一事物内不同的关系。在不同图上的数据点可以具有不同的特征集，因此使用多个图结构不仅能够表示多关系数据，而且能够表示更复杂的数据。对多关系数据构建的图结构如图1所示。

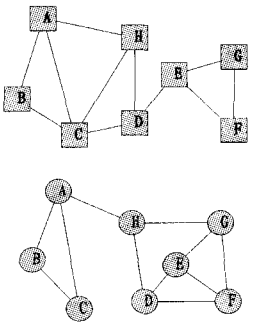

图1 数据的图表示

图1中有8个数据，分别用A～H这8个字母表示，左边是根据论文之间的相似关系构建的图，右边是根据参考文献之间的引用关系构建的图。两个图结构可能会有所不同，例如左图中C和H之间有边相连，右图中却没有，这表示论文C和H之间内容有相似关系，但参考文献没有引用关系。通过构建多个图结构，把复杂数据以及数据之间的复杂关系都表示在图中。

2.2 图上的半监督学习

给定数据集 X，其由 l 个标记数据 $\{(x_1,y_1),\cdots,(x_l,y_l)\}$ 和 u 个未标记数据 $\{x_{l+1},\cdots,x_{l+u}\}$ 组成。记 $n=l+u$，分类函数为 f，记 $F=[f(x_1),\cdots,f(x_n)]^{\mathrm{T}}$，定义类别矩阵 Y 为

$$Y_{ij}=\begin{cases}1, & y_i=j\\ 0, & y_i\neq j\end{cases} \tag{1}$$

基于图的半监督学习以标记数据为“源”，使标记数据的类别信息在图上传播，从而赋予每个未标记数据某个类别值。具体地，图上的随机游走算法[8]步骤如下：

(1)构造权值矩阵

$$W_{ij}=\begin{cases}\exp(-\|x_i-x_j\|^2/2\sigma^2), & i\neq j\\ 0, & i=j\end{cases}$$

(2)构造对角矩阵 $D_{ij}=\begin{cases}0, & i\neq j\\ \sum_j W_{ij}, & i=j\end{cases}$，定义矩阵 $S=D^{-1/2}WD^{-1/2}$；

(3)迭代计算 $F(t+1)=\alpha SF(t)+(1-\alpha)Y$，直至收敛，其中参数 $\alpha\in(0,1)$；

(4)设序列 $\{F(t)\}$ 收敛到 F^*，则未标记数据的类别为 $y_i=\arg\max_j F_{ij}^*$。

上述类别标签的传播过程是收敛的，即给定图结构和标记数据，未标记数据的类别 y_U 是确定的。不失一般性，假设 $F(0)=Y$，通过步骤(3)可得

$$F(t)=(\alpha S)^{t-1}Y+(1-\alpha)\sum_{i=0}^{t-1}(\alpha S)^iY \tag{2}$$

因为 $\alpha\in(0,1)$，而矩阵 S 的特征值在 $[-1,1]$ 之间，所以

$$\lim_{t\to\infty}(\alpha S)^{t-1}=0 \tag{3}$$

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha S)^i=(I-\alpha S)^{-1} \tag{4}$$

综上可以得到

$$F^*=\lim_{t\to\infty}F(t)=(1-\alpha)(I-\alpha S)^{-1}Y \tag{5}$$

谱图理论[9]的发展使得可以从谱方法的角度研究图的性质。定义图的Laplacian矩阵为

$$\Delta=D^{-1/2}(D-W)D^{-1/2} \tag{6}$$

则图上半监督学习的目标函数可以写为

$$\arg\min_f((y-f)^2+\mu f^{\mathrm{T}}\Delta f) \tag{7}$$

式中，$\mu>0$。文献[8]证明了式(7)的解与式(5)相同。式(7)中第一项使函数 f 在标记数据上的预测值尽可能与标记数据的真正类别值 y 近似；第二项的意义是使图上权值大的节点之间的预测值尽可能相近，其等价于

$$f^{\mathrm{T}}\Delta f=\sum_{i,j}(f_i-f_j)W_{ij} \tag{8}$$

单个图上的半监督学习过程如式(7)所示。以图上标记数据的标记信息 y 为“源”，通过反映图结构的Laplacian矩阵 Δ，将标记信息在图上进行传播，从而预测未标记数据的标记值。

2.3 多个图之间的协同学习

本文研究在数据由多个图结构表示的情况下，如何进行半监督学习。用 $X^{(v)}$ 表示在第 v 个图上的数据 X，其中 $v\in(1,2,\cdots,V)$。假设一部分数据点 x_L 的类别已知，其它数据点 x_U 的类别未知，为了得到 x_U 的类别标签，在每个图上分别构造权值矩阵为

$$W_{ij}^{(v)}=\begin{cases}\exp(-\|x_i^{(v)}-x_j^{(v)}\|^2/2\sigma^2), & i\neq j\\ 0, & i=j\end{cases} \tag{9}$$

根据权值矩阵，分别构造对角矩阵和Laplacian矩阵为

$$D_{ij}^{(v)}=\begin{cases}0, & i\neq j\\ \sum_j W_{ij}^{(v)}, & i=j\end{cases} \tag{10}$$

$$\Delta^{(v)}=\frac{1}{\sqrt{D^{(v)}}}(D^{(v)}-W^{(v)})\frac{1}{\sqrt{D^{(v)}}} \tag{11}$$

则在各个图结构上，以数据点 $x_L^{(v)}$ 为“源”的标签传播过程可以写为下面的最优化问题，即

$$\arg\min_{f^{(v)}}(y-f^{(v)}(x^{(v)}))^2+\mu^{(v)}f^{(v)\mathrm{T}}\Delta^{(v)}f^{(v)} \tag{12}$$

在构建的多个图结构中相应的顶点是同一数据，例如图1中左、右两个图中的顶点A都代表同一篇论文，所以在这多个图中相应顶点的学习结果应该尽可能相同。这可以形式化地表述为：对某数据点 x，在不同图上的标签传递结果 $f^{(q)}(x^{(q)})$、$f^{(v)}(x^{(v)})$ 满足

$$\min(f^{(q)}(x^{(q)})-f^{(v)}(x^{(v)}))^2 \tag{13}$$

式中，$q,v\in(1,2,\cdots,V)$ 且 $q\neq v$。式(13)体现了多个图之间的协同学习，协同学习使同一个数据点在多个图上的学习结果尽可能相似。

综合单个图上的半监督学习和多个图之间的协同学习，得到基于图的半监督协同训练算法(Multi-graph Co-training，简称MC算法)，其目标函数为

$$\arg\min_f \sum_{q,v=1}^{V}[(y-f^{(v)}(x^{(v)}))^2+\mu^{(v)}f^{(v)\mathrm{T}}\Delta^{(v)}f^{(v)}+\frac{1}{2}\lambda(f^{(q)}(x^{(q)})-f^{(v)}(x^{(v)}))^2] \tag{14}$$

式中，$q\neq v$。式(14)中前两项表示每个图上的半监督学习，后一项表示多个图之间的协同学习。通过协同学习能够使多个图之间互相提供信息，每个图上的学习器都用其它图上的信息修正自己的学习过程，从而使整个学习过程综合考虑了多个图结构。多个图的半监督学习由每个图上的半监督学习和多个图之间的协同学习组成，这使得标记信息在每个图上传播的同时，考虑标记信息在其它图上的传播情况，从而使学习结果是多个图上的综合优化结果。

从概率的角度解释上述学习过程。在每个图上，定义一个高斯随机场，即

$$p(f^{(v)})=\frac{1}{Z}\exp(-\beta^{(v)}E(f^{(v)})) \tag{15}$$

式中，E 是能量函数，β 是常数，而 Z 为配分函数，且

$$Z=\int_{f^{(v)}{}_L=Y_L}\exp(-\beta^{(v)}E(f^{(v)}))\mathrm{d}f^{(v)} \tag{16}$$

式中，Y_L 表示式(1)定义的类别矩阵中，标记数据所对应的那部分矩阵。学习的目标是希望求解 $p(f^{(v)}|Y_L)$。引入式(11)所示图的 Laplacian 后，能量函数可以表示为

$$E(f^{(v)})=f^{(v)\mathrm{T}}\Delta^{(v)}f^{(v)} \tag{17}$$

则函数 $f^{(v)}$ 的高斯随机场为

$$p(f^{(v)})=\frac{1}{Z}\exp(-\beta^{(v)}f^{(v)\mathrm{T}}\Delta^{(v)}f^{(v)}) \tag{18}$$

为了把协同学习项(即式(13))表示为二次型的形式，记 $f^{(v)}=\tilde{\omega}^{(v)\mathrm{T}}x^{(v)}$，则在数据有两个视图时，式(13)可写为

$$\sum_i(\tilde{\omega}^{(1)\mathrm{T}}x_i^{(1)}-\tilde{\omega}^{(2)\mathrm{T}}x_i^{(2)})^2=\tilde{\omega}_C^{\mathrm{T}}C\tilde{\omega}_C \tag{19}$$

式中，$\tilde{\omega}_C=[\tilde{\omega}^{(1)\mathrm{T}}\tilde{\omega}^{(2)\mathrm{T}}]^{\mathrm{T}}$，$C=\sum_i[x_i^{(1)\mathrm{T}}(-x_i^{(2)\mathrm{T}})]^{\mathrm{T}}[x_i^{(1)\mathrm{T}}(-x_i^{(2)\mathrm{T}})]$。

把式(19)看作协同学习项的能量函数，则其对应的高斯先验为

$$p(\tilde{\omega}_C)\propto\exp(-\beta_C\tilde{\omega}_C{}^{\mathrm{T}}C\tilde{\omega}_C) \tag{20}$$

记 $A^{(v)}=x^{(v)\mathrm{T}}\Delta^{(v)}x^{(v)}$，单个图上学习的高斯先验可写为

$$p(\tilde{\omega}^{(v)})\propto\exp(-\beta\tilde{\omega}^{(v)\mathrm{T}}x^{(v)\mathrm{T}}\Delta^{(v)}x^{(v)}\tilde{\omega}^{(v)})\propto\exp(-\beta\tilde{\omega}^{(v)\mathrm{T}}A^{(v)}\tilde{\omega}^{(v)}) \tag{21}$$

把每个图上的学习和协同学习结合起来，可以得到

$$p(\tilde{\omega})\propto\exp(-\beta^{(v)}\tilde{\omega}^{(v)\mathrm{T}}A^{(v)}\tilde{\omega}^{(v)}-\beta_C\tilde{\omega}_C^{\mathrm{T}}C\tilde{\omega}_C) \tag{22}$$

这样，通过式(22)，可以求解 $p(f^{(v)}|Y_L)=p(\tilde{\omega}^{(v)\mathrm{T}}x^{(v)}|Y)$，得到把标记信息作为先验信息的条件下，图上的学习函数 $f^{(v)}$ 对未标记数据的分类结果。

3 相关工作

Zhu 等人[10]和 Zhang 等人[11]在研究网页分类问题时，把网页之间的链接关系用有向图表示，然后用此有向图辅助对网页的分类。具体地，这类方法将网页中的文本内容用向量模型表示，将网页之间的链接用有向图表示，在对网页文本进行分类时，考虑网页链接提供的信息，将文本的核学习与图的核学习结合在一起。这类方法研究两种核函数的结合，与本文中的研究内容有所不同。

Zhou 等人[12]在研究多视图的谱聚类问题时，使用了多个图结构，在每个图上计算节点的出度、入度和状态转移概率，并将多个图上的对应值加权，以综合多个图上的信息。其研究对象虽然是多个图上的半监督学习，但该文献主要解决谱聚类问题，而本文主要解决分类问题。

Argyriou 等人[13]在研究如何构建更优的图结构时，考虑了多个图的学习问题。单个图上的学习目标函数见式(7)，其中 Δ 表示图的 Laplacian，一般取 $\Delta=D-W$，W 是图上的权值矩阵，D 是 W 的对角阵。Argyriou 等人[13]将多个图上的 Laplacian 相加，即 $\Delta=\Delta^{(1)}+\Delta^{(2)}$，然后再用式(7)进行学习。这相当于先把多个图合并成一个图，再利用基于图的学习方法进行学习。Tsuda 等人[14]在研究蛋白质的分类问题时也使用了多个图，并将图的 Laplacian 相加。这种方法存在的问题是，如果这多个图上的 Laplacian 不可比，则不能将它们简单相加；而且这样把多个图合并为一个图，也失去了协同训练带来的优点。

本文针对多关系数据构建多个图结构，然后分别在每个图上进行半监督学习，并约束这多个图上的学习器对同一数据的预测相似，以达到协同学习的目的。这样的学习方法一方面使不具有可比性的数据对象在不同图上可以单独进行学习；另一方面使多个图之间共同学习，综合优化多个视图上的学习函数。本文中的学习算法结合了图学习法的优点和协同训练的优点，从而能够提升每个图上的学习器性能。

4 实验结果与分析

对 MC 算法进行实验分析，实验所用数据集为 Cora 论文数据集[15]。该数据集由 30714 篇学术论文组成，其中关于机器学习的论文分为 7 个类别，分别为案例学习、遗传算法、神经网络、概率方法、增强学习、规则学习、理论研究，具体情况见表 1。在表 1 所列的机器学习类数据集上进行实验。

表 1 Cora 中的机器学习数据集

类别	数据个数	所占比例
Case-based	402	11.2%
Genetic Algorithms	551	15.4%
Neural Networks	1064	29.7%
Probabilistic Methods	529	14.8%
Reinforcement Learning	335	9.3%
Rule Learning	230	6.4%
Theory	472	13.2%

Cora 数据集中的每篇论文由标题、作者、摘要、参考文献等内容构成。如前文中所述，论文之间存在多种关系，例如文章的相似关系、作者的合作关系、参考文献的引用关系等。综合考虑这些关系，构建多个图结构以描述论文之间的不同关系。

使用 Rainbow 软件包[16]对数据进行处理，把每篇论文用一个向量表示，得到的向量由 20000 个特征组成，按照 Rainbow 软件包中提供的标准办法把每篇论文表示成 $tf\cdot idf$ 向量形式。使用论文之间文本的相似性构建第一个图结构，相似性的计算见式(9)；使用论文作者的合作关系构建第二个图结构，如果某两篇论文有共同作者，则这两篇论文之间有边连接，并设置此条边的权值为 1，否则权值为 0；使用论文参考文献的引用关系构建第 3 个图结构，如果某两篇论文有共同的参考文献，则这两篇论文之间有边连接，并设置此条边的权值

为1，否则权值为0。通过上述过程得到3个不同的图，其权值矩阵分别表示为 $W^{(1)}$、$W^{(2)}$、$W^{(3)}$。

对于表1所列的机器学习数据集，将其按标记率 η 随机划分为标记数据集 L 和未标记数据集 U，标记率 η 代表标记数据在训练集中所占的比例。Cora数据集对每篇论文都提供了类别信息，若划分到未标记数据集的数据点 $x_i \in U$，在训练时隐藏其类别信息，在训练结束之后对比学习到的 $f(x_i)$ 和其类别信息 y_i 是否相同，并以此计算错误率。在实验中对 η 的不同取值进行了考察。注意在数据集的划分过程中保持类别比例不变。每个实验重复10次，并取10次运行结果的平均值。

比较在3个图上分别训练得到的分类器性能，综合3个图训练得到的分类器性能。用 *content*、*author*、*refe* 分别表示使用权值矩阵 $W^{(1)}$、$W^{(2)}$、$W^{(3)}$ 训练得到的分类器，用 *MC* 代表本文中多个图的半监督协同训练算法得到的分类器，*MC* 算法对某数据点 x_i 的分类结果取各视图上分类结果的平均，即 $f(x_i)=\frac{1}{V}\sum_{v=1}^{V} f^{(v)}(x_i)$。考察标记率 η 从0.1增加到0.6的过程中，上述分类器的错误率变化情况，得到的结果如图2所示。

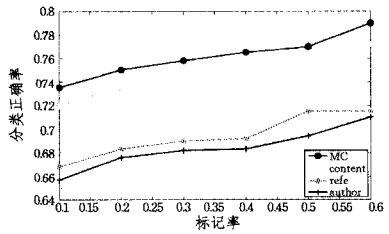

图2 Cora数据集上学习结果曲线图

图2中横坐标表示标记率 η，纵坐标表示分类正确率。从图中可以看出，当标记数据增加时，各个图上的学习器性能均有所提高，但综合多个图学习得到的分类器性能始终好于使用单个图学习得到的分类器性能。具体情况是，*MC* 的分类正确率比单独使用论文内容训练的分类器 *content* 的正确率平均高2.75%，比单独使用参考文献引用关系训练的分类器 *refe* 高9.73%，比单独使用作者合作关系关系训练的分类器 *author* 高11.15%。分析其原因，用数据之间的不同关系构建多个图，并综合多个图进行学习，这种方法利用了数据中蕴含的多种信息，所以比只使用数据之间的一种关系学习有更好的学习效果。这说明当数据之间存在多种关系时，使用多个图结构来描述数据之间的不同关系，并综合多个图结构进行半监督学习，能够提高学习性能。

结束语 图表示法能够较好地表示数据之间的关系。在半监督学习中，通过图表示法可以将标记和未标记数据的结构信息都包含在图中，而图上的半监督学习能够收敛到最优解。本文通过多个图结构表示多关系的数据，将基于图的半监督学习引入协同训练，创新性地提出一种多个图的协同训练算法，以解决多关系数据的学习问题。在真实数据集上的实验表明，本文提出的算法对多关系数据有较好的预测性能。

参 考 文 献

[1] Olivier C, Bernhard S, Alexander Z. Semi-Supervised Learning [M]. Cambridge, MA: MIT Press, 2006

[2] 周志华，尹学松，肖宇，等. 半监督学习专刊[J]. 软件学报，2008，19(11)：2789-2868

[3] Avrim B, Tom M. Combining labeled and unlabeled data with co-training[C] // Proceedings of the 11th Annual Conference on Learning Theory. Madison, WI, 1998: 92-100

[4] 周志华. 半监督学习中的协同训练风范[M]// 周志华，王珏. 机器学习及其应用. 北京：清华大学出版社，2007：259-275

[5] Zhou Zhi-hua, Li Ming. Semi-supervised learning by disagreement[J]. Knowledge and Information Systems, 2010, 24(3): 415-439

[6] Avrim B, Shuchi C. Learning from labeled and unlabeled data using graph mincuts[C] // Proceedings of the 18th International Conference on Machine Learning. Williamston, MA, 2001: 19-26

[7] 刘铁岩，高斌. 高阶异构数据挖掘[M]// 周志华，王珏. 机器学习及其应用. 北京：清华大学出版社，2007：28-48

[8] Zhou Deng-yong, Bernhard S. Learning from labeled and unlabeled data using random walks[C] // Proceedings of the 26th German Association for Pattern Recognition DAGM Symposium. Berlin, Germany, Springer, 2004: 237-244

[9] Chung Fan R K. Spectral Graph Theory[M]. American Mathematical Society, 1997

[10] Zhu Sheng-huo, Yu Kai, Chi Yun, et al. Combining content and link for classification using matrix factorization[C] // Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. Amsterdam, The Netherlands, 2007: 487-494

[11] Zhang Tong, Alexandrin P, Byron D. Linear prediction models with graph regularization for web-page categorization[C]// Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining. Philadelphia, PA, USA ACM Press, 2006: 821-826

[12] Zhou Deng-yong, Burges Christopher J C. Spectral clustering and transductive learning with multiple views[C] // Proceedings of the 24th International Conference on Machine Learning. 2007

[13] Andreas A, Mark H, Massimiliano P. Combining graph laplacians for semi-supervised learning[C] // Advances in Neural Information Processing Systems. 2005

[14] Koji T, Hyunjung S, Bernhard S. Fast protein classification with multiple networks[J]. Bioinformatics, 2005, 21(2): 59-65

[15] McCallum A, Nigam K, Rennie J, et al. Automating the contruction of internet portals with machine learning[J]. Information Retrieval Journal, 2000, 3: 127-163

[16] Andrew M C. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering [EB/OL]. http://www. cs. cmu. edu/wmccallum/bow, 1996