

一种基于模糊近邻标签传递的半监督分类算法

郝建柏, 陈贤富, 黄双福, 杨 俊

(中国科学技术大学 电子科学与技术系, 安徽 合肥 230027)

摘 要: 提出了样本分布无关, 模型简单, 单控制参数的模糊近邻标签传递算法. 该算法依据样本与其 k 个近邻的模糊相似性连接, 使类别标签从标签数据向未标签数据传递, 实现未标签数据的分类. 最后, 通过人工合成数据和 UCI 数据集中数据的分类实验验证了该算法的简单有效性.

关键词: 半监督学习; 模糊计算; 相似性连接; 标签传递

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2010)02-0030-04

Semi-supervised Classification Algorithm Using Fuzzy Nearest Neighborhood Label Propagation

HAO Jian-bai, CHEN Xian-fu, HUANG Shuang-fu, YANG Jun

(Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: A Fuzzy Nearest Neighborhood Label Propagation algorithm (FNNLP) is proposed, which is sample distribution independent and is simple with only one parameter. Based on the similarity connections of the k nearest data, FNNLP classifies the unlabeled data by making the labels propagate from labeled data to unlabeled ones. Finally, two promising experiment results of synthesized data and UCI dataset classifications verify the simplicity and effectiveness of FNNLP.

Key words: semi-supervised learning; fuzzy computing; similarity connection; label propagation

1 引言

与有监督学习和无监督学习相比, 半监督学习 (Semi-Supervised Learning, SSL) 是在训练过程中同时采用已知类别的训练数据和未知类别的测试数据, 因此, 它可以利用更多的信息, 如数据的分布特性等. 在总数据量较大, 而标签数据数量相对较小的情况下, SSL 能获得更好的学习效果, 因此受到了广泛关注^[1-2].

现有的半监督学习算法主要有自训练算法 (Self-training) 和低密度划分算法 (Low Density Separation, LDS). Self-training, 如模型生成算法 (Generative Models, GM) 需要类条件概率密度函数 $p(x|w_i)$ 的形式作为先验知识, 然后采用期望最大化算法 (Expectation Maximization, EM) 估计 $p(x|w_i)$ 的参数. 但是, 这种先验知识往往很难获得. LDS 算

法, 如传导支持向量机 (Transductive SVM, TSVM; 又称 Semi-Supervised SVM, S3VM) 在训练时, 经未标签数据的引导使分类超平面移到数据分布稀疏的区域. 但是, TSVM 控制参数较多, 控制参数的选择会耗费大量时间和精力, 且 TSVM 训练过程慢, 难于处理大批量数据^[3].

针对以上两类半监督学习算法的问题, 提出了一种基于模糊近邻标签传递的半监督分类算法 (Fuzzy Nearest Neighborhood Label Propagation, FNNLP). 通过对比实验验证了 FNNLP 算法的简单有效性及在标签数据稀少时分类性能的优越性.

2 基于 Self-training 的标签传递

首先作如下符号定义: 数据集 $X = \{x_1, x_2, \dots, x_l, x_{l+1}, \dots, x_{l+u}\}$, 其中 $x_i \in R^d, 1 \leq i \leq l+u$. $L = \{x_1, x_2, \dots, x_l\}$, 表示标签数据集. $U = \{x_{l+1},$

$X_{l+2}, \dots, X_{l+u}$, 表示未标签数据集. L 和 U 是 X 的子集, 且 $L \cup U = X$. 设数据类别数为 $c (c \geq 2)$, 类别集合为 $W = \{w_1, w_2, \dots, w_c\}$.

Self-training 算法的学习过程为用初始化的 L 训练最初的分类器, 并用分类器分类 U 中数据, 给分类信心度最高的数据添加类别标签, 且将它们从 U 移到 L 中; 然后利用新的 L 重新训练分类器, 并再次分类 U 中数据. 重复如上过程, 直到满足训练结束条件^[4]. 把 Self-training 算法用于标签传递, 首先须计算数据间的相似关系, 文中采用 RBF 核函数^[5], 对任意两数据 x_i, x_j , 其相似性 s_{ij} 定义为

$$s_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2 / 2\sigma^2) & i \neq j \\ 0 & i = j \end{cases} \quad (1)$$

则可计算出任意未标签数据与标签数据的相似性, 得 $u \times l$ 的矩阵 S_{UL} . 然后, 求出 S_{UL} 中最大的 k 个值, 找出相应未标签数据, 即求出与标签数据最近的 k 个未标签数据, 把它们从 U 移到 L 中. 循环以上过程, 直到 U 为空集.

基于 Self-training 的标签传递算法在学习过程中不断扩大标签数据集, 但是它缺少控制错误分类积累的机制. 在标签传递时, 一旦出现错误分类情况, 算法无法避免和控制错分标签的继续传递. 同时, 由于 U 与 L 的不断更新, S_{UL} 也需重新计算, 因此该算法计算量非常大, 在一些对计算实时性要求较高的场合不适用.

3 FNNLP 算法

FNNLP 算法在处理未标签数据类别时采用模糊化的方法, 即未标签数据的类别不再是确定的, 而是采用 $[0, 1]$ 区间上的模糊隶属度表示. FNNLP 算法中, 未标签数据不是从标签数据处学习直接获得类别标签, 而是从其 k 个近邻处学习获得模糊的类别标签, 算法的最后一步去除模糊化, 实现未标签数据的分类. 类别标签模糊化后, FNNLP 算法避免了标签数据集的扩大, 减少了错误分类扩散的概率.

为便于说明 $c > 2$ 时的分类问题, 作如下的符号定义: 数据 x_i 的类别用 $1 \times c$ 的向量 f_i 表示, 若 $x_i \in L$, 则 f_i 的第 j 列元素的值定义为

$$f_{ij} = \begin{cases} 1 & x_i \in w_j \\ 0 & x_i \notin w_j \end{cases} \quad (2)$$

若 $x_i \in U$, 则 x_i 属于 w_j 类的隶属度 $f_{ij} \in [0, 1]$. X 中数据的类别标签用 $(l + u) \times c$ 的矩阵 f_X 表示; f_L, f_U 分别为 f_X 的前 l 行、后 u 行子矩阵.

FNNLP 算法中, 未标签数据的类别标签从其 k

个近邻处传递而来, 算法的最终目的是实现模糊的类别标签在所有未标签数据上平滑分布, 即对于任意 $x_i \in U$, 其类别 f_i 均满足:

$$\sum_{x_j \in N(x_i)} s_{ij} (f_i - f_j) = 0 \quad (3)$$

式中, $N(x_i)$ 表示 x_i 的 k 个近邻组成的集合. 由式 (3) 可得

$$f_i = \sum_{x_j \in N(x_i)} \frac{s_{ij}}{\sum_{x_j \in N(x_i)} s_{ij}} f_j = \sum_{x_j \in N(x_i)} p_{ij} f_j \quad (4)$$

其中,

$$p_{ij} = \begin{cases} s_{ij} / \sum_{x_j \in N(x_i)} s_{ij}, & x_j \in N(x_i) \\ 0, & x_j \notin N(x_i) \end{cases} \quad (5)$$

p_{ij} 表示归一化的相似性. 对任意 $x_i \in U, x_j \in X$, 由式 (5) 可得 $u \times (l + u)$ 的矩阵 P , P 表示 U 中数据与 X 中数据的 k 近邻关系. 此时, 式 (4) 可用矩阵形式表示为

$$f_U = P \times f_X \quad (6)$$

由式 (6) 可知, FNNLP 算法中未标签数据的类别标签在不停被更新, 类别标签以标签数据为起点, 最先传递到其近邻未标签数据, 而后, 再传递给次级近邻, 标签传递的最终状态为所有未标签数据类别不再变化, 即类别标签在所有未标签数据上实现平滑分布.

由上可得出 FNNLP 算法的迭代实现步骤如下:

- (1) 循环变量 $i = 0$, 初始化 $f_U^{(i)} = 0$;
- (2) 根据式 (5) 计算 P ;
- (3) $i = i + 1$, 利用 $f_U^{(i)} = P \times f_X^{(i-1)}$ 求解 $f_U^{(i)}$;
- (4) 重复步骤 (2)、(3), 直到 $f_U^{(i)} = f_U^{(i-1)}$;
- (5) 分类未标签数据: 查找矩阵 f_U 各行中的最大值, 求其对应的列号 j , 并把相应未标签数据分为 w_j 类.

从上面算法步骤可以看出, FNNLP 算法在每次迭代时并不更新 U 和 L , 无需反复计算归一化相似性矩阵 P , 相比基于 Self-training 的标签传递算法, FNNLP 算法计算量较小, 在一些对实时性要求较高的场合比较适用.

4 实验结果与分析

为验证 FNNLP 算法的有效性, 文中采用了两个实验. 第一个是人工合成的“双环”数据分类实验, 用于说明该算法分类时的简单有效性. 第二个是在 UCI 数据集集中的 Ionosphere、Wine 和 Iris 上进行

的分类实验,用于证明该算法处理实际数据的能力.同时,为了证明在标签数据稀少的情况下 FNNLP 算法能获得良好的分类效果,选择了 k 近邻法(k Nearest Neighborhood, kNN)^[6] 和 TSVM^[7] 两种分类器作对比实验.

4.1 人工合成数据实验

本实验使用线性不可分的“双环”分布数据,内环数据 50 个,外环数据 100 个,标签数据每类仅一个,如图 1(a) 中空心圆和星号所示,其余点为未标签数据.图 1(b)~图 1(d) 分别为采用 FNNLP 算法($k=7$)、kNN($k=1$)、TSVM(RBF 核, $\sigma=10$) 进行分类的结果,三种方法的分类精度分别为 100%、50.67%、83.33%. 对比可知,该实验中,FNNLP 算法获得了最优的分类效果.

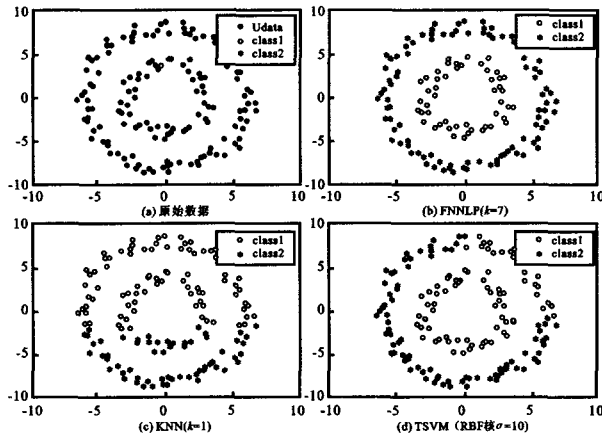


图1 人工合成数据分类实验

4.2 UCI 数据集实验

实验数据为来自 UCI 数据集的 Ionosphere、Wine 和 Iris,三种数据的信息如表 1 所示.

表1 Ionosphere、Wine 和 Iris 数据信息

数据集	样本总数	特征数	类别数
Ionosphere	351	34	2
Wine	178	13	3
Iris	150	4	3

采用 FNNLP($k=9$)、kNN($k=1$)、TSVM(RBF 核, $\sigma=12.5$) 三种算法对 Ionosphere 数据分类,结果如图 2 所示;FNNLP($k=5$)、kNN($k=1$)、TSVM(RBF 核, $\sigma=6.7$) 三种算法对 Wine 数据分类,结果如图 3 所示;FNNLP($k=4$)、kNN($k=1$)、TSVM(RBF 核, $\sigma=9.5$) 三种算法对 Iris 数据分类,结果如图 4 所示. 实验中,标签数据从数据集中随机抽取且各类别数据数量相等,其余的数据作为未标签数据用于测试,给出了 50 次实验结果的均值曲线.

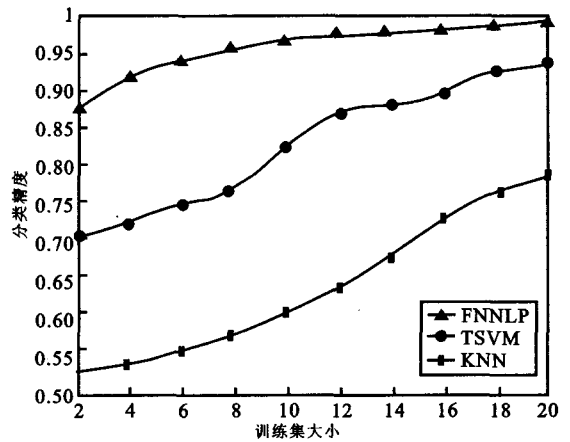


图2 Ionosphere 数据分类结果

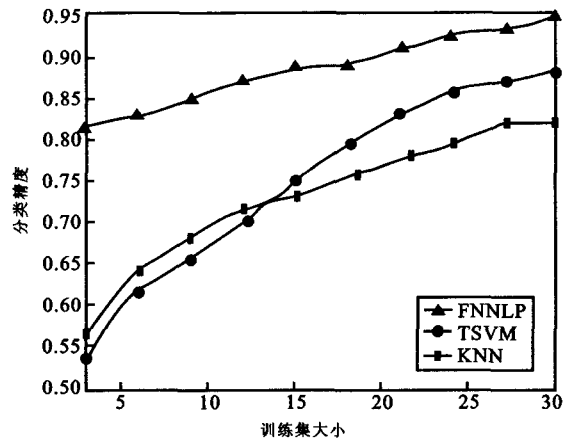


图3 Wine 数据分类结果

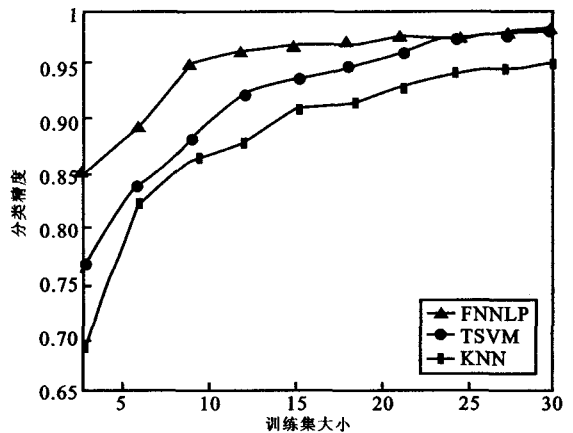


图4 Iris 数据分类结果

对比三个数据分类实验的结果可以看出, FNNLP 算法取得了最优的分类结果,因为,在标签数据稀少时,标签数据不足以表示样本的实际空间分布, kNN 就无法获得较好的分类效果. TSVM 利用了未标签数据,但标签数据过少时, SVM 往往学习不充分,难以获得良好的分类效果. 相比之下,即使标签数据稀少, FNNLP 算法仍能依靠数据间丰富的模糊相似性连接关系,使类别标签在数据间传递,取得较好的分类效果.

5 结束语

文中提出了一种基于模糊近邻标签传递的半监督分类算法.相比 GM 和 TSVM,该算法不受样本分布概率模型限制,控制参数少,使用简单.FNNLP 算法可以有效利用近邻数据间的模糊相似性关系,使类别标签从标签数据传递到未标签数据,实现分类目的,且能有效控制错误分类的扩散.最后,通过人工合成数据分类和 UCI 数据集上的实验验证了该算法的有效性,以及在标签数据较少情况下分类性能的优越性.

参考文献:

- [1] Zhu X, Ghahramani Z, Lafferty J. Semi-supervised learning using gaussian fields and harmonic functions[C] // Proceedings of the Twentieth International Conference on Machine Learning. Washington DC, USA, 2003:328-335.
- [2] 肖宇,于剑.基于近邻传播算法的半监督聚类[J].软件学报,2008,19(11):2803-2813.
- [3] Camps-Valls G, Marsheva T, Zhou D. Semisupervised graph-based hyperspectral image classification[J]. IEEE Trans. on Geoscience and Remote Sensing, 2007, 45(10): 3044-3054.
- [4] Rosenberg C, Hebert M, Schneiderman H. Semi-supervised selftraining of object detection models[J]. Application of Computer Vision, 2005, 1(5): 29-36.
- [5] 朱根标,张凤鸣,董群立.基于核函数和相似度的动态聚类算法[J].微电子学与计算机,2006,23(3):178-184.
- [6] 贾可亮,樊孝忠,许进忠.基于 KNN 的汉语问句分类[J].微电子学与计算机,2008,25(1):156-158.
- [7] Joachims T. Transductive inference for text classification using support vector machines[C] // Proceedings of the Sixteenth International Conference on Machine Learning. San Francisco, USA, 1999:200-209.

作者简介:

郝建柏 男,(1984-),硕士研究生.研究方向为模式识别、人工智能.
陈贤富 男,(1963-),博士,副教授.研究方向为机器学习、智能信息处理.
黄双福 男,(1983-),硕士研究生.研究方向为模式识别、人工智能.
杨俊 男,(1983-),硕士研究生.研究方向为模式识别、机器学习.

(上接第29页)

- [5] Wenze Qu, Jie Chen, Zhongrui Qi, et al. Low complexity cell search algorithm for mobile OFDMA systems[C] // Proc. of Wireless Communications, Networking and Mobile Computing. China, Dalian, 2008: 1-4.
- [6] 元中瑞. WiMAX 基带信号处理算法与原型机设计的研究[D].北京:中国科学院微电子研究所,2008.

作者简介:

曲文泽 男,(1983-),硕士研究生.研究方向为通信基带算

法及 SOC 设计.

刘海洋 男,(1980-),博士研究生.研究方向为通信基带算法及 SOC 设计.
邱昕 男,(1981-),助理研究员.研究方向为通信基带算法及 SOC 设计.
元中瑞 男,(1979-),助理研究员.研究方向为通信基带算法及 SOC 设计.
陈杰 男,(1964-),博士生导师.研究方向为通信基带 SOC 设计、多媒体处理器 SOC 设计与高性能数字信号处理器的研发.