文章编号:1672-3961(2013)02-0018-05 DOI:10.6040/j.issn.1672-3961.3.2012.051

一种新的基于半监督的多标记学习算法

李雅林[1,2],张化祥[1,2*],冯新营[1,2]
(1. 山东师范大学信息科学与工程学院,山东 济南 250014;
2. 山东省分布式计算机软件新技术重点实验室,山东 济南 250014)

摘要:多标记学习中通常存在大量未标记示例,本研究结合协同训练(Co-training)方法充分利用数据集中的未标记示例,在数据集上选取局部 k-NN(k nearest neighbor)和全局 k-NN 进行训练得到两个分类器,分类器分别标记未标记示例并相互更新训练集。协同训练过程不断迭代进行,直至训练完成。试验结果表明,该方法性能均优于其他多标记学习算法。

关键词:半监督学习;多标记学习;局部 k-NN;全局 k-NN
中图分类号:TP301 **文献标志码**:A

A new multi-label learning algorithm based on semi-supervised learning

LI Ya-lin[1,2], ZHANG Hua-xiang[1,2*], FENG Xin-ying[1,2]
(1. School of Information Science & Engineering, Shandong Normal University, Jinan 250014, China;
2. Shandong Provincial Key Laboratory for Novel Distributed Computer Software Technology, Jinan 250014, China)

Abstract: Multi-label learning usually has many unlabeled samples. Combined with co-training method, this research made full use of the unlabeled sampled in dataset, selected the local k-NN(k nearest neighbor) and global k-NN for training to get two classifiers, which could label the unlabeled examples and could be added to the training set. The collaborative training process iterated continuously, until the training finished. The experimental results showed that this algorithm could outperform other multi-label learning algorithms.

Key words: semi-supervised learning; multi-label learning; local k-NN; global k-NN

0 引言

真实世界的对象通常并不只具有唯一的语义,而是多义性的。例如,在文本分类中,每个文档可能属于多个主题,如政府和卫生[1-2];在功能基因组中,每个基因可能与一个功能类集相关,如新陈代谢,转录和蛋白质合成[3];在场景分类中,每个场景图像可能同时属于几个语义类,如海滩和城市[4-7],这些都属于多标记问题。在多标记学习中,训练集中每个示例与多个类别标记相关,学习的目的是将所有合适的类别标记赋予未见示例。

多标记问题广泛存在于真实世界的诸多应用中。早期,多标记学习的研究主要是解决文本分类中遇到的歧义性问题[1-2,8-9]。经过近十年的发展,多标记学习技术已在多媒体内容自动标注[4,10-12]、生物信息学[3,13-14]、Web 挖掘[15-16]、信息检索[17-19]、个性化推荐[20-21]等领域得到了广泛应用。令 $X=\mathbf{R}^d$ 表示输入空间,$Y=\{1,2,\cdots,Q\}$ 表示类别标记集合。多标记训练集 $S=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}$,

收稿日期:2012-12-05 网络出版时间:2013-01-18 14:31
网络出版地址:http://www.cnki.net/kcms/detail/37.1391.T.20130118.1431.003.html
基金项目:国家自然科学基金资助项目(61170145);教育部高等学校博士点专项基金资助项目(20113704110001);山东省自然科学基金资助项目(ZR2010FM021)
作者简介:李雅林(1989 -),女,山东烟台人,硕士研究生,主要研究方向为机器学习与数据挖掘. E-mail:yalinli2012@163.com
*通讯作者:张化祥(1966 -),男,山东济宁人,教授,博士生导师,主要研究方向为机器学习,模式识别及 Web 挖掘等. E-mail:huaxzhang@163.com

$\boldsymbol{x}_i \in X$ 是单一示例,$\boldsymbol{y}_i \subseteq Y$ 是与 $\boldsymbol{x}_i$ 相关标记集,多标记学习目标是在训练集 S 上学习函数 $h: X \to 2^Y$,为未知样例预测标记集。一种简单直观的解决多标记学习问题的方法是将其分解为多个二分类问题,然而这种方法没有充分考虑标记之间的相关性。因此,多种新的解决方法相继提出。McCallum A 提出一种基于贝叶斯和 EM 算法的混合模型[1]用于多标记文本分类;Schapire R E 和 Singer Y 提出了一种基于集成学习(ensemble learning)的方法 BoosTexter[2],该方法是流行集成学习方法 AdaBoost[22]的扩展,训练过程中 BoosTexter 保持训练样例和相应标记的权重集;Elisseeff A 和 Weston J 提出了一种核方法[3],定义一个基于 Ranking Loss 的代价函数以及相应的边界(margin);Clare A 和 King R D 通过改变熵(entropy)的定义对 C4.5 决策树进行了改造[14],使其可以处理多标记数据;Cai L 和 Hofmann T 扩展支持向量机(SVM)[23]学习并结合文档分类问题中的分层分类的判别函数[24];Zhang Min-ling 和 Zhou Zhi-hua 提出一种新的算法 ML-kNN[25],也是一种解决多标记学习问题的有效方法,它将 kNN 算法和贝叶斯算法相结合来学习分类器,从而对多标记数据进行有效地分类。

在传统的有监督多标记学习中,分类器通过对大量有标记的训练样例进行学习,建立模型用于预测未见示例的标记。然而在现实问题中通常存在大量的未标记示例,有标记示例则比较少。显然,收集大量未标记示例相当容易,而获取大量有标记的示例则相对较为困难,如果只使用少量的有标记示例,那么所训练出的多标记学习系统往往很难具有强泛化能力;另一方面,如果仅使用少量"昂贵的"有标记示例而不利用大量"廉价的"未标记示例,则是对数据资源的极大浪费。半监督学习(semi-supervised learning)[26]试图利用大量的未标记示例辅助对少量有标记示例的学习,给定一个未知分布的有标记示例集 $L = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \cdots, (\boldsymbol{x}_{|L|}, \boldsymbol{y}_{|L|})\}$ 以及一个未标记示例集 $U = \{\boldsymbol{x'}_1, \boldsymbol{x'}_2, \cdots, \boldsymbol{x'}_{|U|}\}$,期望学得函数 $f: X \to Y$ 可以准确地对示例 $\boldsymbol{x}$ 预测其标记 $\boldsymbol{y}$。这里 $\boldsymbol{x}_i, \boldsymbol{x'}_j \in X$ 均为 d 维向量,$\boldsymbol{y}_i \in Y$ 为示例 $\boldsymbol{x}_i$ 的标记,$|L|$ 和 $|U|$ 分别为 L 和 U 的大小,即它们所包含的示例数。半监督学习对于减少标注代价、提高学习性能具有非常重大的实际意义。提有的多种半监督学习方法中 Co-training 算法[27]是一个很好的选择。Li G Z 等结合 Co-training 和 ML-k-NN 提出一种半监督多标记学习算法 COMN[28]用于基因功能分析,它在同样的数据集上利用两个具有不同参数集的 ML-k-NN 分类器,两个分类器标记未标记示例且相互更新训练数据,最终预测结果由两个分类器的组合决定。

本研究基于半监督提出一种新的多标记学习算法。首先选取局部 k-NN 训练得到一个分类器,标记置信度较高的未标记示例并添加到数据集中,再在新训练集上选取全局 k-NN 进行训练,分类器对置信度较高的示例进行标记,将标记的示例加入到训练集中,另一个分类器再利用这些新标记的示例进行更新。协同训练过程不断迭代进行,直至标记完成。实验结果表明,该算法在提高分类精度上具有较好的性能。

1　半监督学习 Co-training

协同训练(Co-training)是一个有效的半监督学习算法,提供一个利用未标记示例的框架来提高分类精度。Co-training 被分类为一个基于多视图的算法,Co-training 学习框架的形式化定义:示例空间两个视图在分布 D 上定义为 $X = (X^1 \cup X^2)$;$X_L = (\overrightarrow{\boldsymbol{x}_i^1}, \overrightarrow{\boldsymbol{x}_i^2}, \boldsymbol{y}_i)_{i=1}^{l}$ 是已标记的训练样例集,$X_U = (\overrightarrow{\boldsymbol{x}_j^1}, \overrightarrow{\boldsymbol{x}_j^2})_{j=l+1}^{l+u}$ 是未标记的训练集且 $l = |X_L|, u = |X_U|$。给出两种数据表示,Co-training 算法如下:已标记样例集 X_L 和未标记样例集 X_U,每个都由两个视图表示,训练 X_L 中各自的表示生成两个弱分类器。这些分类器标记 A 中的每个示例,A 是 X_U 的任意大小的子集。每个分类器选择置信度较高的 $n+p$ 个样例给予标记并扩展到 X_L 中。每个分类器选择 $n+p$ 个样例,U 中将减少 $2n+2p$ 个样例,因此必须从 X_U 随机寻找 $2n+2p$ 个样例重新填满。这一过程重复进行,最终训练获得两个半监督分类器。其中,n 和 p 的值用户定义,考虑数据的基础分布。

本研究将 Co-training 的思想应用于多标记学习中,结合两种不同的 k-NN 策略,从局部和全局的角度考虑训练集,从而提高分类精度。

2　新算法 ML-Co2k-NN

本研究从不同的角度寻找最近邻,将示例向量间的夹角关系作为选取 k 近邻的标准。首先将样本根据其属性转化为 n 维向量,向量之间位置关系除了使用空间距离表示之外还可以使用向量夹角 θ(向量 $\boldsymbol{x}$ 与向量 $\boldsymbol{y}$ 的夹角)进行描述:

$$\mathrm{sim}(\boldsymbol{x}, \boldsymbol{y}) = \cos\theta = \frac{\sum_{i=1}^{n} \boldsymbol{x}_i \times \boldsymbol{y}_i}{\sqrt{\sum_{i=1}^{n} \boldsymbol{x}_i^2} \sqrt{\sum_{i=1}^{n} \boldsymbol{y}_i^2}}, \quad (1)$$

式(1)中,$\cos\theta$ 的值越接近 1 则表明 $\boldsymbol{x}$,$\boldsymbol{y}$ 间的夹角越小,即两向量越接近,两个样例相似度越高。因此本研究以样例特征向量间的夹角 θ 的大小作为选取样本最近邻的标准,采用了基于向量夹角 θ 的 k 近邻多标记分类算法。

2.1 局部 k-NN

本研究将传统的 k-NN 分类器看作是局部 k-NN。分类器在预测新样例的标记时,考虑距目标样例最近的 k 个邻居点,这一方式选择的近邻反映的是目标样例的局部性。

2.2 全局 k-NN

全局 k-NN 分类器从另外一个不同的角度选择目标样例的最近邻,将训练集看作是一个整体来考虑。获得的邻居样例是训练样例的一个集合$\overrightarrow{\boldsymbol{x}_i} \in X_L$,且$\overrightarrow{\boldsymbol{x}_i}$的 k 局部近邻集要包含目标样例$\overrightarrow{\boldsymbol{x}_{\text{new}}}$。这种策略获得的近邻与局部方式的完全不同,一个直观简单的寻找全局 k-NN 的方法是首先按顺序对所有的样例确定其局部 k 近邻,然后判断近邻中是否包含目标样例。因此,与局部方法不同的还有全局 k-NN 方法检索到的近邻的数量不固定,$|N_k(\overrightarrow{\boldsymbol{x}_i})_{\text{global}}| \in [0 \cdots l]$。

总之,局部 k-NN 和全局 k-NN 是从不同的视角考虑问题,它们的结合会提高分类精度,特别是在判别边界上。

2.3 ML-Co2k-NN

ML-Co2kNN 算法结合两种 k-NN 策略与半监督学习方法 Co-training 应用于多标记学习中,将 ML-k-NN 作为基础分类器。具体算法步骤描述如下:首先在数据集已标记样例上训练局部 ML-k-NN 得到一个分类器 h_1,选择分类结果中置信度较高的前 m 个示例进行标记并重新添加到训练集中,再在新数据集上训练全局 ML-k-NN,得到的分类器 h_2 标记置信度较高的前 m 个示例,将标记的示例也加入到训练集中,另一个分类器再利用这些新标记的示例进行更新。协同训练过程不断迭代进行,直至训练完成。最终结果由两个分类器组合决定。m 由用户定义,迭代次数上限设置为 50。

ML-Co2k-NN 算法流程如下:

输入 标记样例 X_L,未标记样例 X_U,每次标记的样例数 m,近邻数 k,迭代次数 i

输出 组合分类器 h_3

过程

(1) for $i=1$ to 50 do;

(2) $h_1 \leftarrow$ML-Lk-NN$(X_L^{(1)},k)$;

(3) $h_2 \leftarrow$ML-Gk-NN$(X_L^{(2)},k)$;

(4) for each $\overrightarrow{\boldsymbol{x}_i} \in X_U$ do;

(5) 检索 ML-Lk-NN$(h_1,\overrightarrow{\boldsymbol{x}_i},k)$;

(6) 将置信度 $\boldsymbol{x}_i$ 添加到 Clist1 中;

(7) end;

(8) 标记 Clist1 中的前 m 个样例并加入 X_L;

(9) for each $\overrightarrow{\boldsymbol{x}_i} \in X_U$ do;

(10) 检索 ML-Gk-NN$(h_2,\overrightarrow{\boldsymbol{x}_i},k)$;

(11) 将置信度 $\boldsymbol{x}_i$ 添加到 Clist2 中;

(12) end;

(13) 标记 Clist2 中的前 m 个样例并加入 X_L;

(14) end;

(15) 定义 h_3:$P_{h_3}(c_j|X_L) \leftarrow P_{h_1}(c_j|X_L^{(1)}) * P_{h_2}(c_j|X_L^{(2)})$;

return h_3。

3 试验结果及分析

3.1 试验数据集介绍

本研究在两个数据集 Image[3] 和 Yeast[25] 上进行试验, ML-Co2k-NN 算法与目前已存在的半监督多标记学习算法 COMN 和自定义的局部 k-NN、全局 k-NN 自训练(Self-training)算法 ML-SLk-NN、ML-SGk-NN 进行对比。采用 Hamming loss、One-error、Coverage、Ranking loss、Average precision[29] 评测指标对算法分类结果进行评估。参数设置 $k=10$,m 设置为 20,最大循环次数 50。试验采用 10 倍交叉验证。

Image 数据集包含 2 000 个自然场景图像,分别属于类别沙漠, 高山, 海洋, 日出和树木。超过 22% 的图像同时属于多个类且每个图像平均与 1.24 个类标记有关。Yeast 数据集包含 2 417 个由 103 维特征向量表示的基因,存在 14 个可能的类别标记且每个基因平均与 4.24 个类相关。

3.2 试验结果分析

表 1 是在 Image 数据集上的对比试验结果。如表 1 所示,试验结果表明在所有的评价指标上,ML-Co2k-NN 算法结果均好于其他半监督多标记学习算法,协同训练算法结果均优于自训练算法。在 Image 数据集上局部自训练和全局自训练算法差异不大,COMN 选取参数不同的 k-NN 而 ML-Co2k-NN 算法选取的是局部 k-NN 和全局 k-NN,充分考虑数据特性并结合有效的半监督学习方法 Co-training,明显提高分类精度。

表 1　Image 数据集上的试验结果对比
Table 1　Comparison of experimental result based on image dataset

评测方法	算法			
	ML-Co2*k*-NN	ML-SL*k*-NN	ML-SG*k*-NN	COMN
Hamming loss ↓	0.122 0	0.163 0	0.161 0	0.138 0
One-error ↓	0.226 0	0.325 0	0.321 0	0.262 0
Coverage ↓	0.750 0	0.823 0	0.812 0	0.785 0
Ranking loss ↓	0.115 0	0.144 0	0.142 0	0.125 0
Average precision ↑	0.846 0	0.656 0	0.648 0	0.765 0

表 2 是在 Yeast 数据集上的试验结果。从表 2 可以看出,在所有的评价指标上,ML-Co2*k*-NN 算法结果相比其他算法均有提高,协同训练结果也均优于自训练算法。在 Yeast 数据集上,局部自训练优于全局自训练,其主要原因是由于基因数据具有流行结构, ML-Co2*k*-NN 算法选取局部*k*-NN,充分考虑数据局部特性,有利于提高分类精度。因此该算法相对于自训练方法和 COMN 具有明显的优势。

表 2　Yeast 数据集上的实验结果对比
Table 2　Comparison of experimental result based on Yeast dataset

评测方法	算法			
	ML-Co2*k*-NN	ML-SL*k*-NN	ML-SG*k*-NN	COMN
Hamming loss ↓	0.153 0	0.168 0	0.179 0	0.165 0
One-error ↓	0.216 0	0.260 0	0.311 0	0.242 0
Coverage ↓	6.373 0	6.695 0	6.939 0	6.563 0
Ranking loss ↓	0.166 0	0.175 8	0.188 0	0.172 0
Average precision ↑	0.865 0	0.826 0	0.798 0	0.830 0

4　总结

多标记学习训练集中存在大量的未标记示例,半监督学习用于考虑如何利用大量的未标记示例和少量的有标记示例进行训练,减少标注代价,提高学习性能。本研究基于半监督提出一种新的多标记学习算法,结合两种 ML-*k*-NN 选取策略在数据集上进行协同训练,协同训练过程不断迭代进行,直至标记完成。实验结果表明,ML-Co2*k*-NN 算法分类效果与其他多标记算法相比,分类性能均表现较好。下一步将寻找更好的半监督多标记学习方法。

参考文献:

[1] MCCALLUM A. Multi-label text classification with a mixture model trained by EM [C]//Proceedings of Working Notes of the AAAI'99 Workshop on Text Learning. Menlo Park, CA: AAAI, 1999:1-7.

[2] SCHAPIRE R E, SINGER Y. BoosTexter: a boosting-based system for text categorization [J]. Machine Learning, 2000, 39 (2/3):135-168.

[3] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification [C]//Proceedings of Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002:681-687.

[4] BOUTELL M R, LUO J, SHEN X, et al. Learning multi-label scene classification [J]. Pattern Recognition, 2004, 37(9):1757-1771.

[5] COMITE F D, GILLERON R, TOMMASI M. Learning multi-label alternating decision tree from texts and data [C]//Proceedings of the Lecture Notes in Computer Science. Berlin: Springer, 2003:35-49.

[6] CHEN M S, HAN J H, YU P S. Data mining: an overview from a database perspective [J]. IEEE Trans on Knowledge and Data Engineering, 1996, 8(6):866-883.

[7] 周志华,张敏灵. MIML:多示例多标记学习[J]. 机器学习及其应用,2009:218-234.
ZHOU Zhihua, ZHANG Minling. MIML: Multi-instance multi-label learning [J]. Machine Learning and Application, 2009:218-234.

[8] UEDA N, SAITO K. Parametric mixture models for multi-label text [C]//Proceedings of Neural Information Processing Systems 15. Cambridge, MA: MIT Press, 2003: 721-728.

[9] GAO S, WU W, LEE C H, et al. A MFoM learning approach to robust multi-class multi-label text categorization [C]//Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada: ACM Press, 2004:329-336.

[10] CARNEIRO G, CHAN A, MORENO P, et al. Supervised learning of semantic classes for image annotation and retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(3):394-410.

[11] QI G J, HUA X S, RUI Y, et al. Correlative multi-label video annotation [C]//Proceedings of the 15th ACM International Conference on Multimedia. New York, NY: ACM Press, 2007:17-26.

[12] SNOEK C G M, WORRING M, Van GEMERT J C, et al. The challenge problem for automated detection of 101 semantic concepts in multimedia [C]//Proceedings of the 14th ACM International Conference on Multimedia. New York, NY: ACM Press, 2006: 421-430.

[13] BARUTCUOGLU Z, SCHAPIRE R E, TROYANSKAYA O G. Hierarchical multi-label prediction of gene function [J]. Bioinformatics, 2006, 22(7): 830-836.

[14] CLARE A, KING R D. Knowledge discovery in multi-label phenotype data [C]//Proceedings of Lecture Notes in Computer Science. Heidelberg, Berlin: Springer,

2001:42-53.

[15] TANG L, RAJAN S, NARAYANAN V K. Large scale multi-label classification via metalabeler[C]//Proceedings of the 18th International Conference on World Wide Web. Madrid, Spain, NY: ACM Press, 2009:211-220.

[16] YANG B, SUN J T, WANG T, et al. Effective multi-label active learning for text classification[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, NY: ACM Press, 2009:917-926.

[17] YU K, YU S, TRESP V. Multi-label informed latent semantic indexing[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazil: ACM Press, 2005: 258-265.

[18] ZHU S, JI X, XU W, et al. Multi-labeled classification using maximum entropy method[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brazi: ACM Press, 2005:274-281.

[19] GOPAL S, YANG Y. Multi-label classification with meta-level features[C]//Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva, Switzerland: ACM Press, 2010:315-322.

[20] SONG Y, ZHANG L, GILES L C. A sparse Gaussian processes classification framework for fast tag suggestions[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management. Napa Valley, CA: ACM Press, 2008: 93-102.

[21] OZONAT K, YOUNG D. Towards a universal marketplace over the Web: statistical multi-label classification of service provider forms with simulated annealing[C]//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France: ACM Press, 2009:1295-1303.

[22] FREUND Y, SCHAPIRE R E. A decision theoretic generalization of on-line learning and an application to boosting[C]//Proceedings of Lecture Notes in Computer Science. Heidelberg, Berlin: Springer, 1995:23-37.

[23] 曹林林,张化祥,王至超. 一种基于信息熵数据修剪的支持向量机:EB-SVM[J]. 山东大学学报:理学版, 2012,47(5):59-62.
CAO Linlin, ZHANG Huaxiang, WANG Zhichao. A kind of support vector machine based on information entropy data pruning: EB-SVM [J]. Journal of Shandong University: Natural Science, 2012, 47(5):59-62.

[24] CAI L, HOFMANN T. Hierarchical document categorization with support vector machines[C]//Proceedings of the 13rd ACM International Conference on Information and Knowledge Management. Washington D C, USA: ACM Press, 2004:78-87.

[25] ZHANG Minling, ZHOU Zhihua. ML-KNN: a lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.

[26] SHAHSHAHANI B, LANDGREBE D. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon[J]. IEEE Transactions on Geo-science and Remote Sensing, 1994, 32(5): 1087-1095.

[27] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the 11th Annual Conference on Computational Learning Theory. Wisconsin, MI: ACM Press, 1998:92-100.

[28] LI G Z, YOU M, GE L, et al. Feature selection for semi-supervised multi-label learning with application to gene function analysis[C]//Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology. Niagara Falls, NY: ACM Press, 2010:354-357.

[29] SPYROMITROS E, TSOUMAKAS G, VLAHAVAS I. An empirical study of lazy multi-label classification algorithms[C]//Proceedings of the 5th Hellenic Conference on Artificial Intelligence. Berlin, Springer, 2008: 401-406.

(编辑:胡春霞)