一种基于向量夹角的 k 近邻多标记文本分类算法

广 凯 潘金贵

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 在多标记学习中,一个示例可以有多个概念标记。学习系统的目标是通过对由多标记样本组成的训练集进行学习,以尽可能正确地预测未知样本所对应的概念标记集。k 近邻算法已被应用到多标记学习中,该算法将测试示例转化为多维向量,根据其 k 个近邻样本的标记向量来确定该测试示例的标记向量。传统的 k 近邻算法是基于向量的空间距离来选取近邻,而在自然语言处理中,文本间的相似度常用文本向量的夹角来表示,所以本文将文本向量间的夹角关系作为选取 k 近邻的标准并结合 k 近邻算法提出了一种多标记文本学习算法。实验表明,该算法在文档分类的准确率上体现出较好的性能。

关键词 机器学习,多标记学习,文本分类

An kNN Algorithm Based on Vector Angle for Multi-label Text Categorization

GUANG Kai PAN Jin-Gui

(State Key Lab For Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Nanjing 210093)

Abstract In multi-label learning, each instance in the training set is associated with a set of labels, and the task is to output a label set whose size is unknown a priori for each unseen instance. k nearest neighbors (kNN) algorithm is recently applied to multi-label categorization. In detail, each instance is transformed into a vector and the label vector of the test instance is determined by its k nearest neighbors, which are chosen by the Euclidean distance of a couple of vectors. In this paper, a multi-label lazy learning approach named θ-MLkNN is presented, which is derived from the traditional k nearest neighbor (kNN) algorithm. Instead, we select the k nearest neighbors by the angle of two vectors. Experiments on a real-world text data set show that θ-MLkNN achieves better precision to traditional MLkNN algorithms.

Keywords Machine learning, Multi-label learning, Text categorization

多标记学习问题在真实世界中随处可见。例如,在文档分类问题中,每个文档可能同时隶属于多个主题;在生物信息学中,每个蛋白质分子在与更大的细胞耦合时可能具有多种不同的功能;在场景分类问题中,每幅图像可能同时隶属于多种场景。在以上的这些问题中,一个示例可同时具有多个概念标记,学习系统的目标是通过对由多标记样本组成的训练集进行学习,以尽可能正确地预测未知样本所对应的概念标记集。

如果限定每个样本只对应一个概念标记,则传统的两类或多类学习问题均可以看作多标记学习问题的特例。然而,多标记学习的一般性也使得其比传统的监督学习问题更难解决。一种直观的解决多标记学习问题的方法是将其分解为若干个独立的二值分类问题进行求解,然而这种方法没有充分考虑各个样本所对应的多个概念标记之间的关系。幸运的是,目前为止已经出现了多种多标记学习算法,如基于决策树的多标记学习算法[1,2]和基于支持向量机的多标记学习算法[3]。最近,传统的 k 近邻(kNN)学习算法也被应用到多标记学习中。而在自然语言处理中,文本间的相似度常用文本向量的夹角来表示,所以本文将文本向量间的夹角关系作为选取 k 近邻的标准并结合 k 近邻算法提出了一种多标记文本学习算法,将其应用到文档分类这一特殊的多标记学习问题,并与传统的 k 近邻的多标记学习算法进行了实验比较。

1 多标记学习

多标记学习起源于文档分类研究中遇到的歧义性问题。R. E. Schapire 和 Y. Singer 为了将同一个文档归入多个类别,提出了一种基于集成学习(ensemble learning)[4,5]的方法 BoosTexter[6],该方法实际上是对 AdaBoost[7] 的扩展,它在训练过程中不仅要改变训练示例的权重,还要改变概念标记的权重。在此之后,多标记学习引起了很多学者的关注。

另外,A. K. McCallum 也提出了一种 Bayes 和 EM[8] 相结合的方法[9],他用一个混合模型来表示文档类别,然后利用 EM 对混合权和每一类的混合成分中字的分布进行学习。2001 年,A. Elisseeff 和 J. Weston 通过定义一个称为 Ranking Loss 的代价函数以及相应的边际(margin)[3],提出了一种基于支持向量机的方法,有助于缓解 BoosTexter 在训练集较小时容易过配(overfitting)的问题。同年,A. Clare 和 R. D. King 通过改变熵(entropy)的定义[1],对 C4.5 决策树进行了改造,使其可以处理多标记数据。2002 年,N. Ueda 和 K. Saito 假设多标记文本有一个特征字的混合出现在每个类的单标记文档中[10],从而在 Bag-of-Words 的表示下,提出了两种概率发生模型(probabilistic generative model)。2003 年,

广 凯 硕士研究生,主要研究领域为机器学习、信息检索;潘金贵 教授,博士生导师,主要研究领域为多媒体信息处理技术、多媒体远程教育系统。

F. D. Comité 等人对交替决策树(Alternating Decision Trees)[11]进行了改造,使其可以处理多标记数据。同年,B. Lauser 和 A. Hotho 利用 SVM 来建立多语言环境下多标记文档的自动索引[12]。最近,Zhang 等人将传统的 k 近邻算法引入到多标记学习[13],体现出良好的性能。

2 θ-MLkNN 算法

2.1 k 最近近邻算法

k 最近近邻(k-Nearest Neighbor)算法先将样本根据其属性转化为 n 维向量,对于每一个样本,找出与它"最近"的 k 个邻居,然后根据 k 个邻居的标记来确定该样本的标记。在传统的 kNN 算法中,这里的"最近"采用两点间的欧几里德距离(Euclidean distance)进行计算:即对于两个样本 $X=\{x_1, x_2, \cdots, x_n\}$,$Y=\{y_1, y_2, \cdots, y_n\}$有以下公式:

$$E\_dis(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{1}$$

公式(1)计算了向量的空间距离,其值越小,则两样本的空间距离越"近"。

当样本转化为 n 维向量后,向量之间位置关系除了使用空间距离表示以外,还可以使用另一个向量间的重要关系——向量夹角 θ(向量 X 与向量 Y 的夹角)进行描述:

$$sim(X,Y)=\cos\theta=\frac{\sum_{i=1}^{n}x_i\times y_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}} \tag{2}$$

在式(2)中,$\cos\theta$ 的值越接近 1,则表明 X,Y 间的夹角越小,即两向量越接近。在自然语言处理中,文本向量间夹角的余弦值常被作为两文本的相似度,所以本文针对文本分类这一特定的领域,以样本向量间的夹角 θ 的大小来作为选取样本最近近邻的标准,提出了基于向量夹角 θ 的 k 最近近邻多标记文本分类算法(θ-MLkNN)。

2.2 θ-MLkNN 算法

首先我们给出一些相关的符号。

定义 1 假设给定一个示例 x 和 $\vec{y}_x$ 表示 x 的标记向量,令 $\vec{N}_x$ 表示 x 的第 j 个近邻示例向量,$\vec{N}_x^j(l)$表示 x 的第 j 个近邻上第 l 个标记的值,若该近邻有标记 l,则其值为 1,否则为 0。$\vec{C}_x(l)$用于统计 x 的近邻中属于第 l 类的数量,则可定义其为

$$\vec{C}_x(l)=\sum_{i=1}^{k}N_x^i(l) \tag{3}$$

定义 2 $E_j^l(j\in\{0,\cdots,k\})$表示示例 t 的 k 个近邻中恰有 j 个示例具有标记 l 这一事件。

又令 H_1^l 表示 t 有标记 l 这一事件,而 H_0^l 表示 t 没有标记 l 这一事件。则在该算法中,测试示例的类向量 $\vec{y}_t$ 将根据如下的最大化后验概率原则所确定:

$$\vec{y}_t(l)=\arg\max_{b\in\{0,1\}}P(H_b^l\mid E^l_{\vec{C}_t(l)}) \tag{4}$$

其中 $l\in Y$,该式确定了测试示例 t 的类向量是使得 $P(H_b^l\mid E^l_{\vec{c}_t(l)})(b\in\{0,1\})$的值最大时的取值。

基于贝叶斯规则(Bayesian rule),式(4)可以改写如下:

$$\vec{y}_t(l)=\arg\max_{b\in\{0,1\}}\frac{p(H_b^l)P(E^l_{\vec{c}_t(l)})\mid H_b^l)}{P(E^l_{\vec{c}_t(l)})}=\arg\max_{b\in\{0,1\}}P(H_b^l)P(E^l_{\vec{c}_t(l)}\mid H_b^l) \tag{5}$$

由此,我们给出 θ-MLkNN 算法基本步骤(具体算法如算法 1 所示):

第一步:从训练集合统计得出 $P(H_b^l)$ $(l\in y, b\in\{0,1\})$;

第二步:对于训练集合中的每一个示例 x 根据式(2)计算出 x 的 k 个近邻向量 $\vec{N}_x^j(j\in\{1,\cdots,k\})$。

第三步:根据第二步得到的 $\vec{N}_x^j$ 根据式(3)计算出 $\vec{C}_x(l)$,再利用 $\vec{C}_x(l)$和训练集合计算出 $P(E_j^l\mid H_b^l)$ $(j\in\{0,\cdots,k\}, b\in\{0,1\})$。

第四步:对于测试样本 t,根据式(2)计算集合 t 的近邻集合 $\vec{N}_t^j(j\in\{1,\cdots,k\})$,再根据式(5)计算 t 的标记向量。

算法 1 给出了 k 近邻多标记算法:(1)~(3)计算 $P(H_0^l)$ 和 $P(H_1^l)$,(4)~(11)计算得出 $\vec{N}_x^j$,(12)~(23)计算后验概率 $P(E_j^l\mid H_b^l)$ $(j\in\{0,\cdots,k\}, b\in\{0,1\})$,(24)~(25)根据最大化后验概率确定测试样本的标记向量。

算法 1 算法的伪码表示

```
Step 1:compute P(H_b^l)P(H_1^l)
(1)for l∈Y
(2)   P(H_1^l)=(Σ_{i=1}^{m} y_{x_i}(l))/m;
(3)   P(H_0^l)=1-P(H_1^l);
Step 2:compute N_x^j
(4)for i=1:m
(5)   for j=i:m
(6)      temp1=min_{p=1}^{k}(Sim(X_i, N_{x_i}^p));
(7)      index1=arg min_{p=1}^{k}(Sim(X_i, N_{x_i}^p));
(8)      temp2=min_{p=1}^{k}(Sim(X_j, N_{X_i}^p));
(9)      index2=arg min_{p=1}^{k}(Sim(X_j, N_{X_j}^p));
(10)     if(Sim(X_i, X_j)>temp1) N_{X_i}^{index1}=X_j;
(11)     if(Sim(i, X_j)>temp2) N_{x_j}^{index2}=X_i;
Step 3  compute P(E^l_{c_x(l)}|H_b^l)P(E^l_{c_x(l)}|H_1^l)
(12)for l∈Y
(13)   for j=0:k
(14)      labeled[j]=0;
(15)      unlabeled[j]=0;
(16)   for i=1:m
(17)      C_x(l)=Σ_{i=1}^{k} N_x^i(l);
(18)      if(y_{x_i}(l)==1)
(19)         labeled[C_x(l)]++;
(20)      else unlabeled[C_x(l)]++;
(21)   for j=1:k
(22)      P(E_j^l|H_1^l)=labeled[j]/Σ_{p=0}^{k} labeled[p];
(23)      P(E_j^l|H_0^l)=unlabeled[j]/Σ_{p=0}^{k} unlabeled[p];
Step 4:test
(24)C_t(l)=Σ_{i=1}^{k} N_t^i(l);
(25)y_t(l)=arg max_{b∈{0,1}} P(H_b^l)P(E^l_{c_t(l)}|H_b^l);
```

3 实验

3.1 实验环境

本文的实验数据采用了在文档分类领域中常用的 Reuters21578 新闻数据集,该数据集包含了 21578 篇路透社在 1987 年播出的新闻专线稿件[1)]。

在利用 θ-MLkNN 在该数据集上进行学习之前,对该文本数据集进行预处理。本文采用 Bag-of-Words 的文本表示,即将每篇文档表示成为一个数值向量,向量每一维上的取值对应于词汇表中的词在该文档中出现的次数。

首先,对每篇文档的正文进行相关的规范化,将所有字符全部转化为小写,同时将所有文章中出现的数字串用统一的符号表示。此外,去掉正文中出现的标点符号以及 SMART 功能词表(stop list)中出现的所有功能词。

(下转封 3)

1)http://www.daviddlewis.com/resources/testcollections/reuters21578/

image matrix: two-dimensional Fisherfaces. In: ICSP'04 Proceedings, 2004. 1419～1422
15 Etemad K, Chellapa R. Face recognition using discriminant eigenvectors [C]. In: Proc. of ICASSP, 1996
16 Turk M A, Pentland A P. Face recognition using eigenfaces [C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Akanitos, 1991. 586～591
17 Pentland A P, Moghaddam B, Starner T. View-based and modular eigenspaces for face recognition [C]. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Seatle, WA, 1994. 84～91
18 Blanz V, Vetter T. Face Recognition Based on Fitting a 3D Morphable Model [J]. IEEE Transactions on PAMI, 2003, 25(9): 1063 ～ 1074
19 Pablo N, Javier R. Analysis and comparison of eigenspace-based face recognition approaches [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2002, 16(7): 817～830
20 Chen Song-can, Liu Jun, Zhou Zhi-hua. Making FLDA applicable to face recognition with one sample per person [J]. Pattern Recognition, 2004, 37(7): 1553～1555
21 Monwar M, Pau P P, Islam W, et al. A Real-Time Face Recognition Approach from Video Sequence using Skin Color Model and Eigenface Method [C]. In: Canadian Conference on Electrical and Computer Engineering, 2006. 2181 ～ 2185
22 Wang Jie, Plataniotis K N, Venetsanopoulos A N. Selecting discriminant eigenfaces for face recognition [J]. Pattern Recognition Letters. Available online 26 January 2005
23 周激流，张晔，郭晶，等. 基于活动轮廓模型的人脸特征提取方法的研究[J]. 中国图象图形学报，2000，5(4)：341～344
24 王刚，陈世福，陈兆乾，等. 基于分布式工作流技术的校园应用软件集成模型研究[J]. 计算机科学，2005，32(8)：94～96，108
25 张生亮，陈伏兵，杨静宇. 对单训练样本的人脸识别问题的研究[J]. 计算机科学，2006，33 (2)：225～229
26 周晓飞，姜文瀚，杨静宇. 11 范数最近邻凸包分类器在人脸识别中的应用[J]. 计算机科学，2007，34 (4)：234～235，238

(上接第 206 页)

该数据集具有 9 大类概念，数据集中的每篇文档均隶属于若干的概念类别。统计每个类别所包含的文档数，然后将包含文档数最多的 l 个类($l=3,4,\cdots,9$)组合在一起，由此可以得到 7 个不同的数据集，分别记为 num3，num4，num5，num6，num7，num8，num9。对于每一数据集，统计其中每个词汇的文档频率(document frequency)，即包含该词汇的文档数目。根据每个词汇的统计结果，取文档频率排在前 2% 的词汇构成最终的词汇表。数据集中的每篇文档即可采用 Bag-of-Word 的方法表示成为一个数值向量以供多标记学习算法处理。表 1 给出了数据集的描述信息。

表 1　实验数据集的描述信息

数据集	文本数	词汇表大小	每篇文档平均标记数
num3	7258	530	1.0074
num 4	8078	599	1.0140
num 5	8655	651	1.0207
num 6	8817	663	1.0352
num 7	9021	678	1.0375
num 8	9158	683	1.0396
num 9	9190	686	1.0480

3.2　实验结果

本文将 θ-MLkNN 算法与基于空间距离的 ML-kNN 算法进行了实验比较。对于每个数据集而言，本文采用与文[13]相同的实验策略，在每个数据集上进行 3 倍交叉运算(3-fold cross-validation)并给出平均结果。限于时间因素，本文对两种算法都取近邻数为 10。

我们对实验结果重点考察标记的准确率，采用以下公式：

$$avr\_pre=\frac{1}{n}\sum_{i=1}^{n}\frac{|\{l|l\in L_i,l\in L'_i\}|}{|L_i|} \qquad (6)$$

公式(6)中 L_i 为测试样本 i 的标记集合，L'_i 为 i 经分类器标记后的标记集合。表 2 为我们对上述 Reuters21578 进行实验后的结果。

表 2　θ-MLkNN 与 ML-kNN 实验结果对比

Avr_pre	θ-MLkNN	ML-kNN
num3	0.904622	0.849195
num4	0.88479	0.837702
num5	0.881038	0.836508
num6	0.862281	0.825364
num7	0.852382	0.814089
num8	0.846534	0.799531
num9	0.837955	0.784393

由表(2)可以明显看出，基于向量角度 θ 的 k 近邻算法在对于文档分类这一多标记学习中的准确度要高于传统的基于 Euclidean 距离的 k 近邻算法。

结语　本文针对多标记文本学习提出了一种基于向量角度距离的 k 近邻算法，该算法基于传统 k 近邻算法思想，并在此基础上对近邻选取上以向量夹角来代替空间距离，通过实验可以看出，得到了较好的准确度。进一步的工作将考虑在标记数量更大的环境中对该算法进行测试，并将其应用在其他多标记学习中，与传统的多标记学习进行对比。

参 考 文 献

1 Clare A, King R D. Knowledge discovery in multi-label phenotype data. In: De Raedt L, Siebes A, eds. Lecture Notes in Computer Science 2168, Berlin: Springer, 2001. 42～53
2 Comité F D, Gilleron R, Tommasi M. Learning multi-label alternating decision trees fromtexts and data. In: Perner P, Rosenfeld A, eds. Lecture Notes in Computer Science 2734, Berlin: Springer, 2003. 35～49
3 Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Dietterich T G, Becker S, Ghahramani Z, eds. Advances in Neural Information Processing Systems 14, Cambridge, MA: MIT Press, 2002. 681～687
4 Dietterich T G. Ensemble learning. In: Arbib M A, ed. The Handbook of Brain Theory and Neural Networks. 2nd edition. Cambridge, MA: MIT Press, 2002
5 Zhou Z-H, Chen S-F. Neural network ensemble. Journal of Computers, 2003, 25(1): 1～8
6 Schapire R E, Singer Y. BoosTexter: a boosting-based system for text categorization. Machine Learning, 2000, 39(2-3): 135～168
7 Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting. In: Vitányi P M B, ed. Lecture Notes in Computer Science 904, Berlin: Springer, 1995. 23～37
8 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistics Society - B, 39(1): 1～38
9 McCallum A. Multi-label text classification with a mixture model trained by EM. In: Working Notes of the AAAI'99 Workshop on Text Learning, Orlando, FL, 1999
10 Ueda N, Saito K. Parametric mixture models for multi-label text. In: Becker S, Thrun S, Obermayer K, eds. Advances in Neural Information Processing Systems 15, Cambridge, MA: MIT Press, 2003. 721～728
11 Freund Y, Mason L. The alternating decision tree learning algorithm. In: Proceedings of the 16th International Conference on Machine Learning, Bled, Slovenia, 1999. 124～133
12 Lauser B, Hotho A. Automatic multi-label subject indexing in a multilingual environment. In: Koch T, Sølvberg I, eds. Lecture Notes in Computer Science 2769, Berlin: Springer, 2003. 140～151
13 Zhang M L, Zhou Z H. ML-kNN: A lazy learning approach to multi-label learning. Pattern Recognition (PRJ), 2007, 40(7): 2038～2048