

基于类属属性的多标记学习算法

吴磊, 张敏灵

显示缩略图

本文结构

1 LIFT算法

2 多标记类属属性机制

3 实验

3.1 数据集

3.2 实验设置

3.3 结果分析

4 总结

参考文献

LIFT的变体算法LIFT-INSIDIF.通过将示例 x 与每个标记的正样本集 $P_k=\{x_j|(x_j, Y_j)ID, l_k \neq Y_j\}$ 的中心求差<math>\frac{1}{|P_k|} \sum_{p \in P_k} p</math>, INSIDIF将示例 x 转换为由 q 个示例组成的示例包,

即, $\{x^{(k)}|1 \leq k \leq q\}$, 其中, $x^{(k)} = x - \frac{1}{|P_k|} \sum_{p \in P_k} p$. 由此可得LIFT-INSIDIF中针对每个标记的类属属性映射:

$$\varphi_k(x) = x^{(k)} = x - \frac{1}{|P_k|} \sum_{p \in P_k} p \tag{7}$$

· LIFT-MLF

另一种多标记属性转换机制是由原属性空间 X 构建元级属性(meta-level feature), 以此刻画示例和标记之间的关系. 在此, 我们使用算法MLF[1,6]中的元级属性构建方法, 实现LIFT的变体算法LIFT-MLF. MLF将每个示例 x 转换为 $q \times (3r+2)$ 维的元级属性向量 $[y(x, l_1), \dots, y(x, l_q)]$, 其中 $y(x, l_k) (1 \leq k \leq q)$ 是由示例 x 与 P_k 中的 r 个近邻样本构造的 $3r+2$ 维元级属性. 具体来说, $y(x, l_k)$ 是示例 x 与其 r 个近邻样本的L2距离、L1距离、余弦距离以及与 P_k 中心的L2距离、余弦距离的结合. 由此可得LIFT-MLF中针对每个标记的类属属性映射:

$$\varphi_k(x) = \psi(x, l_k) \tag{8}$$

由此可见, 以上介绍的BR, LIFT-MDDM, LIFT-INSIDIF, LIFT-MLF以及LIFT都是一阶算法[9], 未考虑标记之间的相关性. 同时, 上述算法的区别在于类属属性的映射方式不同, 其二类分类模型的训练过程完全一致.

3 实验

3.1 数据集

对于数据集 $S = \{(x_i, Y_i) | 1 \leq i \leq p\}$, 我们使用 $|S|, dim(S), L(S)$ 及 $F(S)$ 分别表示样本个数、属性个数、标记个数以及属性类型. 除此之外, 在此还给出了几种描述多标记数据集性质的统计量[9,17]:

· $LCard(S) = \frac{1}{p} \sum_{i=1}^p |Y_i|$:: 标记基数(label cardinality), 即每个样本具有的平均相关标记个数;

· $LDen(S) = \frac{1}{L(S)} \cdot LCard(S)$:: 标记密度(label density), 即由标记个数归一化的标记基数;

· $LDiv(S) = |\{Y | (x, Y) \in S\}|$:: 标记多样性(label diversity), 即数据集中不同标记集合的个数;

· $PLDiv(S) = \frac{1}{|S|} \cdot LDiv(S)$:: 归一化标记多样性, 即由样本个数归一化的相异标记集.

表 1按照样本数的多少概括了本文实验所用的数据集, 其中包括6个常规规模的数据集($|S| \leq 5000$)以及6个大规模的数据集($|S| > 5000$). 此外, 根据惯例对文本数据集rcv1(subset1), rcv1(subset2), eurlex(subject matter)以及tmc2007按照文档频率[1,8]进行了降维处理. 其中, 文档频率最高的2%的属性被保留作为最终属性.

	Table 1 Characteristics of experimental data sets
表 1 实验数据集性质	

从表 1可以看出: 这12个数据集涵盖了很多实际应用领域, 且其多标记性质各不相同. 由此可见, 本文实验所用的数据集是比较充分且全面的, 具有较强的概括性.

3.2 实验设置

由于每个样本可能具有多个标记, 故传统的单标记评价指标如精度(accuracy)、查准率(precision)、查全率(recall)等在多标记学习中不再适用[8]. 假设多标记测试集为 $T = \{(x_i, Y_i) | 1 \leq i \leq t\}$, 多标记学习系统返回的 q 个实值函数为 $\{f_1, f_2, \dots, f_q\}$. 在本文的实验中, 我们使用5种基于样本的多标记评价指标[3,19,20]来评价学习系统的性能:

· Hamming Loss

$$HLoss_T(h) = \frac{1}{t} \sum_{i=1}^t |h(x_i) \Delta Y_i|,$$

其中, $h(x_i) = \{l_k | f_k(x_i) > 0, 1 \leq k \leq q\}$ 对应于示例 x_i 的相关标记集合, Δ 返回两集合之间的对称差(symmetric difference). 该评价指标考察样本在单个标记上的误分类情况, 即: 相关标记被预测为无关标记, 或者无关标记被预测为相关标记. 该评价指标取值越小, 则系统性能越优, 其最优值为0.

· Ranking Loss

$$RLoss_T(f) = \frac{1}{t} \sum_{i=1}^t \frac{1}{|Y_i| \cdot \overline{Y_i}|} |\{(l_a, l_b) | f_a(x_i) \leq f_b(x_i), (l_a, l_b) \in Y_i \times \overline{Y_i}\}|.$$

该评价指标衡量错误排序的标记对的比例, 即, 无关标记排在相关标记之前的比例. 该评价指标取值越小, 则系统性能越优, 其最优值为0.

· One-Error

$$One-Error_T(f) = \frac{1}{t} \sum_{i=1}^t \left| \left[\arg \max_{l \in Y_i} f_l(x_i) \right] \notin Y_i \right|,$$

其中, 当谓词 p 成立时, $\$p$ 返回1; 否则返回0. 该评价指标衡量在样本的类别标记排序中, 排在第1位的标记不是

显示缩略图

本文结构

1 LIFT算法

2 多标记类属属性机制

3 实 验

3.1 数据集

3.2 实验设置

3.3 结果分析

4 总 结

参考文献

$$Coverage_T(f) = \frac{1}{t} \sum_{i=1}^t \max_{l \in Y_i} rank_f(x_i, l) - 1,$$

其中, $rank_f(x_i, l)$ 对应于标记/根据 $\{f_1(x_i), f_2(x_i), \dots, f_q(x_i)\}$ 的值降序排序后所处的位置. 该评价指标衡量在测试集 T 上遍历到样本所有相关标记的平均查找深度. 该评价指标取值越小, 则系统性能越优, 其最优值为 1.

· Average Precision

$$AvgPrec_T(f) = \frac{1}{t} \sum_{i=1}^t \frac{1}{|Y_i|} \sum_{l \in Y_i} \frac{|\{l' \mid rank_f(x_i, l') \leq rank_f(x_i, l), l' \in Y_i\}|}{rank_f(x_i, l)}.$$

该评价指标考察在样本的标记排序序列中, 排在该样本相关标记之前的标记仍为相关标记的比例. 该评价指标取值越大, 则系统性能优越, 其最优值为 1.

为便于对比, 本文通过除以标记个数 $L(S)$ 将评价指标 Coverage 规格化到区间 $[0, 1]$. 上述基于样本的评价指标首先衡量学习系统在每个测试样本上的分类性能, 然后返回整个测试集的均值作为最终结果.

此外, 本文还采用一种基于标记的评价指标 Macro-averaging AUC^[2,1]. 与基于样本的评价指标不同, 基于标记的评价指标首先衡量学习系统在每个标记上分类性能, 然后返回在所有标记上的均值作为最终结果.

· Macro-averaging AUC

$$AUC_{macro} = \frac{1}{q} \sum_{k=1}^q \frac{|\{(x', x'') \mid f_k(x') \geq f_k(x''), (x', x'') \in T_k \times \bar{T}_k\}|}{|T_k| |\bar{T}_k|},$$

其中, $T_k = \{x_i \mid l_k \in Y_i, 1 \leq i \leq t\}$, $\bar{T}_k = \{x_i \mid l_k \notin Y_i, 1 \leq i \leq t\}$ 分别对应于含有标记 l_k 的测试样本集合与不含有标记 l_k 的测试样本集合. 值得注意的是, 上述公式可由 AUC 与 Wilcoxon-Mann-Whitney 统计量的相互关系导出^[2,2]. 此外, 该评价指标取值越大, 系统的性能越优, 其最优值为 1.

本文将给出 6 种算法在上述 12 个数据集上的实验结果, 包括 BR, ML-kNN, LIFT, LIFT-MDDM, LIFT-INSDF 以及 LIFT-MLF. 根据文献^[9], 设定 LIFT 的参数 $r=0.1$; 根据文献^[2,3], 设定 ML-kNN 的近邻个数 $k=10$; 对于 LIFT-MDDM, 为与原文保持一致, 控制转换后特征空间维度 (即 $d\phi$) 的参数 thr 设定为 99.9%; 对于 LIFT-MLF, 我们采取原文的策略, 对近邻个数 r 在区间 $[10, 100]$ 进行调节, 并最终设定为 10. 此外, 为了对比的公平, 所有算法的基二类分类器均使用线性核 LIBSVM^[2,4].

3.3 结果分析

在每个数据集中, 我们采用无放回抽样选取 50% 的样本构成训练集, 余下的 50% 样本构成测试集, 抽样过程重复 10 次并记录 10 次实验的均值和方差.

本文共有两组实验: 第 1 组将没有使用类属属性机制的 BR 以及常用的 ML-kNN 算法与采用类属属性机制的算法 LIFT, LIFT-MDDM, LIFT-INSDF 以及 LIFT-MLF 进行对比, 目的在于验证类属属性机制对多标记学习的影响; 第 2 组将 LIFT 与算法 LIFT-MDDM, LIFT-INSDF, LIFT-MLF 进行对比, 目的在于验证 LIFT 所采用的类属属性构建方法的有效性.

表 2 与表 3 给出了 6 个算法在 12 个数据集上的实验结果, 实验结果采用均值±方差 (mean±std) 的形式表示. 对于每个评价指标, 表示其值越大, 性能越优; 相应地, “表示其值越小, 性能越优. 此外, 对比算法中性能的最好的结果使用黑体标出.

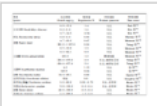


Table 2 Predictive performance of each comparing algorithm (mean±std) on the six regular-scale data sets

表 2 对比算法在 6 个常规规模数据集上的实验结果 (mean±std)

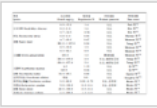


Table 3 Predictive performance of each comparing algorithm (mean±std) on the six large-scale data sets

表 3 对比算法在 6 个大规模数据集上的实验结果 (mean±std)

对于第 1 组实验, 我们分别统计了 BR, ML-kNN 与 LIFT 及其变体算法在 12 个数据集、6 个评价指标上的 72 个对比结果. 与 LIFT 和 LIFT-INSDF 相比, BR 没有胜出的情况; 与 LIFT-MDDM 相比, 胜出的情况仅占 5.56%; 与 LIFT-MLF 相比, 胜出的情况占 44.44%. 与 LIFT, LIFT-INSDF 以及 LIFT-MDDM 相比, ML-kNN 胜出的情况分别仅占 6.9%, 15.3% 以及 30.6%; 与 LIFT-MLF 相比, 胜出的情况占 66.7%. 从 BR, ML-kNN 与 LIFT, LIFT-INSDF, LIFT-MDDM 的对比结果可以看出, 有效的类属属性可以较大程度地提高多标记学习算法的学习性能. 值得注意的是, LIFT-MLF 的性能在一些数据集上比 BR, ML-kNN 的性能差. 一个可能的原因是, 该算法使用近邻样本构造元级属性的方式不能有效地构造标记的类属属性. 可见, 并非所有的属性转换机制都能十分有效地地构造类属属性.

如第 2 节所述, BR 可认为是 LIFT 及其变体算法的退化 (degenerated) 版本. 为了进一步验证类属属性对多标记学习算法的影响, 对于每个评价指标, 本文使用成对 T 检验 (检验水平 $\alpha=0.05$) 对 BR 与 LIFT 及其变体算法进行统计分析 (见表 4(a)). 表 4(a) 中的结果进一步证实了上述结论, 即: 类属属性是提高多标记学习系统泛化性能的可行途径, 但并非所有的多标记属性转换机制都能有效地构建类属属性.

显示缩略图

本文结构

1 LIFT算法

2 多标记类属属性机制

3 实验

3.1 数据集

3.2 实验设置

3.3 结果分析

4 总结

参考文献

基于类属属性的多标记学习算法

吴磊, 张敏灵



significance level $\alpha=0.05$

表 4(a) 对于每个评价指标,BR与使用类属属性机制算法在12个数据集上的成对T检验($\alpha=0.05$)结果(win/tie/lose)

与第1组实验相同,我们统计了LIFT及其3个变体算法在12个数据集、6个评价指标上的72个对比结果.其中,LIFT排在第1位的情况占87.5%,排在第2位和第3位的情况分别仅占11.11%和1.39%,且没有排在最后一位与倒数第2位的情况发生.由以上统计可以看出,LIFT的学习性能优秀,进而说明LIFT的类属属性构造方法是非常有效的.

参考第1组实验,本文使用成对T检验(检验水平 $\alpha=0.05$)对LIFT及其3种变体算法在每个数据集上的实验结果进行统计分析(见表 4(b)).从表 4(b)中可以看出:



Table 4(b) Paired T-test results (win/tie/lose) of comparing LIFT against its variants in terms of each criteria on the total twelve data sets at significance level $\alpha=0.05$

表 4(b) 对于每个评价指标,LIFT与其变体算法在12个数据集上的成对T检验($\alpha=0.05$)结果(win/tie/lose)

(a) 在绝大多数情况下,LIFT的分类性能优于其他3种变体算法——LIFT-INSDF,LIFT-MDDM以及LIFT-MLF.在很少数据集的个别评价指标上,LIFT的性能差于LIFT-INSDF以及LIFT-MDDM;

(b) 除了评价指标One-Error及Coverage,LIFT-INSDF与LIFT-MDDM在其余4个评价指标上的结果相当.LIFT与LIFT-INSDF的性能优于或至少相当于LIFT-MDDM的性能,这说明相比于通过属性选择赋予每个标记相同的类属属性(即公式(6)),为每个标记构造不同的类属属性(即公式(3)与公式(7))是一种更有效的属性转换机制;

(c) 与表 4(a)中得到的结论一致,在每个评价指标上,LIFT-MLF的性能都要差于其余3种算法.可能的原因是,该算法使用近邻样本构造元级属性(即公式(8))的方式不能有效地构造标记的类属属性.

概括来说,表 4(b)中的成对T检验统计结果证明了LIFT中的属性转换机制能非常有效地构造类别标记的类属属性.因此,LIFT的类属属性机制可以作为很多多标记学习算法的预处理(pre-processing)部分,以提高学习系统的学习性能.

4 总结

与许多关注标记空间的算法不同,本文考察在属性空间进行操作对多标记学习算法学习性能的影响.本文在12个数据集上进行了大量的实验,将LIFT及其变体算法与BR进行对比分析,结果显示:有效的类属属性构造,的确较大幅度地提高多标记学习算法的学习性能.同时,LIFT与其变体算法的统计检验结果表明,LIFT的属性转换机制可有效地构造类别标记的类属属性.

值得注意的是,本文中所有算法均忽略标记之间的相关性.未来,我们可以尝试将类属属性和标记之间的相关性相结合,一种直观的方式是,为与标记配对相关的分类器^[25]设计对应的二阶类属属性.

参考文献

[1] Liu DY, Yang K, Tang HY, Chen JZ, Yu QY, Chen GF. A convex evidence theory model. Journal of Computer Research & Development, 2000,37(2):175-181 (in Chinese with English abstract).

[2] Han JF. Research on the synthesis of comfort degree for fuzzy sensors based on temperature and humidity. Journal of Transducer Technology, 2002,21(6):19-24 (in Chinese with English abstract).

[3] Nakamura EF, Loureiro AAF, Frery AC. Information fusion for wireless sensor networks: Methods, models, and classifications. [ACM Computer Survey, 2007,39\(3\):9](#).

[4] Bahador K, Alaa K, Fakhreddine OK, Saiedeh NR. Multi-Sensor data fusion: A review of the state-of-the-art. [Information Fusion, 2013,14:28-44](#).

[5] Debasis B, Jaydip S. Internet of things: Applications and challenges in technology and standardization. [Wireless Personal Communications, 2011,58:49-69](#).

[6] Otman B, Yuan XH. Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory. [Information Fusion, 2007,8\(4\):379-386](#).

[7] Suvasini P, Amlan K, Shamik S, Majumdar AK. Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning. [Information Fusion, 2009,10\(4\):354-363](#).

[8] Yee L, Ji NN, Ma JH. An integrated information fusion approach based on the theory of evidence and group decision-making. [Information Fusion, 2013,14\(4\):410-422](#).

[9] Brice M, Michael AW, Joanne CW. An approach using Dempster-Shafer theory to fuse spatial data and satellite image derived crown metrics for estimation of forest stand leading species. [Information Fusion, 2013,14\(4\):384-395](#).

[10] Yang Y, Liu DY, Wu LZ, Wang WY. An important extension of an evidence-theory based inexact reasoning model. Chinese Journal of Computers, 1990,13(10):772-778 (in Chinese with English abstract).

[11] Liu DY, Ouyang JH, Tang HY, Chen JZ, Yu QY. Research on a simplified evidence theory model. Journal of Computer Research & Development, 1999,36(2):134-138 (in Chinese with English abstract).

[12] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large database. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 1996. 103-114.

[13] Han JW, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed., Beijing: China Machine Press, 2011. (in Chinese).

[14] Chiu T, Fang DP, Chen J. A robust and scalable clustering algorithm for mixed type attributes in large databases. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2001. 263-268.