

多标签分类器准确性评估方法的研究

秦 锋, 黄 俊, 程泽凯, 杨 帆

(安徽工业大学 计算机学院, 安徽 马鞍山 243002)

摘 要:分类是数据挖掘领域研究的核心技术之一,分类器性能评估方法也是众多学者的研究热点之一。以往的分类器性能评估方法一般针对于单标签数据集,对于多标签问题并未涉及。文中主要针对多标签分类问题中的单实例情况,提出了一种多标签分类器准确性评估方法(EMOSIML)。该方法的思路是:如果分类器对一个多标签对象预测的类别标签是其属于的多个类别标签中的任何一个,则分类结果都是正确的。该方法用C#编程实现,并对朴素贝叶斯分类器进行分类器性能评估实验,实验结果表明,EMOSIML评估方法较传统的准确率评估方法更合理。

关键词:准确率评估;分类器评估;二类分类;多标签分类

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2010)01-0046-04

A Study on Accuracy Evaluation Method for Multi-Label Classifier

QIN Feng, HUANG Jun, CHENG Ze-kai, YANG Fan

(School of Computer Science, Anhui University of Technology, Ma'anshan 243002, China)

Abstract: Classification is one of the key techniques of data mining, and the classifier performance evaluation is also a hotspot. Previous classifier evaluation methods focused on single label data sets, and the multi label problem was not concerned. This article mainly aims at single instance multi label classification problem, proposes a more accurate and reasonable evaluation method(EMOSIML). The main idea of this method is that if the predict class forecasted by the classifier is one of the label set which the instance belongs to simultaneously. It is programmed by C#, in plain on the naive Bayesian classifier, and the experimental results show that it is more reasonable than traditional accuracy evaluation methods.

Key words: accurate assessment; classifier assessment; binary classify; multi label classify

0 引 言

分类^[1]是数据挖掘中应用领域极其广泛的重要技术之一,分类是根据数据集的特点构造一个分类器,利用该分类器对未知类别的对象赋予类别的一种技术。构造分类器的过程一般分为训练和测试两个步骤:在训练阶段,分析训练数据集的特点,为每个类别产生一个对相应数据集的准确描述或模型。在测试阶段,利用类别的描述或模型对测试集进行分类,测试其分类的准确度。

在传统的分类问题中,都是假设一个事物(对象)只用一组属性来表示,组成一个实例,并且这个实例只属于一个类标签,即单实例单标签 SISL(Single Instance Single Label)。在大多数模式识别问题中,例如

语音识别和指纹识别等,每一个实例只能唯一地对应一个类别标签。但是,由于客观事物本身的复杂性,一个事物(对象)可以用单个实例来表示,并且该实例属于多个类别标签,即单实例多标签 SIML(Single Instance Multi Label)^[2],比如文本分类中,一个文本同时属于新闻和经济。另外,还有多实例单标签 MISL(Multi Instance Single Label)^[3]和多实例多标签 MIML(Multi Instance Multi Label)^[4]。多实例单标签指的是一个事物(对象)可以用多组属性表示,由这些组属性可以组成多个实例,并且这些实例都对应同一个类标签,而多实例多标签是指由这些不同组属性构成的实例分别对应不同的类标签。图片分类问题中,一张图片的每个组成部分都可以代表一个实例,而且这些实例同时对应不同的类别标签,例如一张图片同时属于类别沙滩和城市。为解决以上不同分类问题,学者已经提出了一些有效的算法,如 AdaBoost、MM、AdaBoost、MR 和 BoosTexter^[2]算法,ML-kNN、MIML-BOOST、MIMLSVM 和 M³MIML^[4~6]等。这些算法能

收稿日期:2009-04-16;修回日期:2009-07-23

基金项目:安徽省自然科学基金资助项目(KJ2007A051)

作者简介:秦 锋(1962-),男,安徽和县人,教授,研究方向为人工智能、数据挖掘。

够解决多实例或多标签分类问题,为评估此类算法,已经提出了一些特殊的评估标准,如 Hamming Loss、One-error、ranking loss、Coverage 和 Average precision^[2,7]。在实际应用中,如果使用传统分类算法解决这些问题多标签分类问题,该如何正确地评估分类器的性能呢?

文中首先介绍一些传统的分类器评估标准,然后以二类分类问题为例,对 SIML 分类问题,通过朴素贝叶斯算法来构建分类模型,由传统的准确率评估方法推导出一种新的准确率度量方法 EMOSIML,最后对实验结果进行分析比较。

1 传统的分类器评估标准

传统的分类器评估标准有:准确率、灵敏性、特效性、精度和 F1 值等^[1]。灵敏性也称真正(识别)率,即正确识别正元组的百分比;特效性是指真负率,即正确识别负元组的百分比;准确率是正确识别正负元组的百分比,它是灵敏性和特效性的函数;精度是指正确识别的正元组在识别为正元组中所占的比率;F1 值是精度和灵敏性的调和平均数。可以用混淆矩阵来形象地说明这几个评价标准,表 1 是一个二分类问题的混淆矩阵。

表 1 混淆矩阵

		分类器预测类别	
		C1	C2
专家预测类别	C1	真正(t_pos)	假负(f_neg)
	C2	假正(f_pos)	真负(t_neg)

(1)灵敏性(Sensitivity): $Sensitivity = t_pos / pos$

(2)特效性(Specificity): $Specificity = t_neg / neg$

(3)精度(Precision): $Precision = t_pos / (t_pos + f_pos)$

(4)准确率(Accuracy): $Accuracy = Sensitivity * (pos / (pos + neg)) + Specificity * (neg / (pos + neg))$

(5)F1 值: $F1 = 2Sensitivity * Precision / (Sensitivity + precision)$

其中 $pos = t_pos + f_neg$, $neg = f_pos + t_neg$ 。

除以上评估标准之外,还有分类器的鲁棒性、速度、可伸缩性和可解释性等。鲁棒性是指当数据集中有噪声数据或缺失数据时,分类器能正确分类的能力;速度标准涉及产生和使用分类器的计算花费;当数据量很大时,能够有效构造分类器的能力称分类器的可伸缩性;可解释性指分类器提供的理解和洞察的水平,此标准受主观因素的影响较大,一般很难评估。在分类器评估过程中,准确率是最常用的评估标准。对于 SISL 分类问题,准确率评估只需考虑校验数据集中元

组的类标签和通过分类模型预测的类标签是否一致,如果一致则认为分类正确,反之则认为不正确。但是现实数据集中很多问题涉及多个类别标签分类问题, SIML 分类问题属于该种情况,而传统的分类算法对一个元组只能预测一个类标签。若该元组同时属于多个类别标签,分类器将该元组的类标签预测为它可能属于的多个类标签中的其它类标签,与该元组本身对应的类标签不同,则分类器在评估时将会给出错误的结果,从一定程度上影响分类器的评估效率。所以这种准确率评估方法对 SIML 分类问题是不适合的。以往分类器准确率评估的相关文献中对此类问题的评估并未涉及^[8-10],文中主要针对 SIML 数据集分类问题提出了一种新的准确性评估方法 EMOSIML(Evaluation Method on Single Instance Multi Label)。

2 准确率评估方法

2.1 多标签分类算法的评估标准

假定多标签数据集 D ,类标签集合 $C = \{C_1, C_2, C_3, C_4\}$,存在元组 X_i ,对应的类标签为 $Y_i = \{C_1, C_2\}$ 。利用多标签分类算法进行分类时,对一个元组 X_i 一次可以预测多个类标签,例如分类器的分类结果可能是:

$\{C_1, C_2\}$:完全正确

$\{C_1, C_3\}$:部分正确

$\{C_1\}$:部分正确

$\{C_1, C_3, C_4\}$:部分正确

$\{C_3, C_4\}$:完全错误

以上只列出了几种可能,分类结果的多样性决定了多标签分类算法评估的复杂性。相比传统的单标签分类器评估问题,多标签数据分类问题的评估更为复杂,现已提出了一些特殊的评估标准,如 One-error、Hamming Loss、Ranking loss、Coverage 和 Average precision^[2,7]等。Matthew 等人提出了一种 α 准确率评估方法^[11]:假设元组 X_i 对应的类标签集合为 Y_i , P_i 表示分类器得到的类标签集合,定义 $M_i = Y_i - P_i$,表示没有预测到的类标签集合, $F_i = P_i - Y_i$,表示预测的错误类标签集合。对每一个 X_i 分类的准确程度可用下式计算:

$$score(P_i) = (1 - \frac{|\beta M_i \cap \gamma F_i|}{|Y_i \cup P_i|})^\alpha$$

$$(\alpha \geq 0, \beta \geq 0, \gamma \leq 1, \beta = 1 \vee \gamma = 1)$$

该评估方法中的 α 称为宽恕率, α 的值越小准确率就越高。准确率的计算公式可表示为:

$$Accuracy = \frac{1}{|D|} \sum_{i=1}^{|D|} score(P_i)$$

2.2 传统的准确率评估方法

假定测试数据集为 D , $|D|$ 记为 D 的元组总数, $X_i (i = 1, 2, \dots, n)$ 为 D 中的元组, $n = |D|$, X_i 对应的类标签为 C_i , $C = \{1, 2\}$ 为类标签集合。传统分类器学习一个函数: $Y = F(X)$, 给定一个元组 X , 函数 F 返回 X 对应的类别 Y , $X \in D$, $Y \in C$ 。根据上面介绍的准确率计算方法, 准确率的计算可以表示为函数 F 的一个函数:

$$\text{Accuracy} = \left(\sum_{i=1}^{|D|} T(F(X_i), (C_i)) \right) / |D|$$

函数 $T(F(X_i), C_i)$: 如果 $F(X_i)$ 与 C_i 的值相同, 则返回值为 1, 否则返回 0。可以用下面的公式表示:

$$T(F(X_i), C_i) = \begin{cases} 1, & F(X_i) = C_i \\ 0, & F(X_i) \neq C_i \end{cases}$$

但是, 此计算公式只适合 SISL 分类问题。对于 SIML 问题这个准确率计算函数就不适合了, 因为存在元组同时属于两个类别的情况。例如, D 中存在 X_i 和 X_j 对应的类标签分别是类别 1 和类别 2, 且 $X_i = X_j$, $i \neq j$, 此时分类器只能将这两个元组同时预测为一个类别, 假设分类器将它们分为类别 1, 即 $F(X_i) = F(X_j) = 1$, 而记录 X_j 对应的类别 $C_j = 2 \neq F(X_j)$, 则认为该条记录分类错误。显然用该评估方法评估 SIML 分类问题是不适合的。为解决此问题, 文中提出了一种新的多标签分类器 EMOSIML 评估方法。

2.3 EMOSIML 评估方法

当一个元组属于多个类别标签时, 传统的分类器评估方法不能正确地评估分类器的准确率, 文中提出了一种准确率评估方法 EMOSIML, 该算法的设计思路是: 只要分类器将一个属于多个标签的元组分到它可能属于的多个类标签中的任何一个, 分类结果都是正确的。在 EMOSIML 评估方法中, 根据多标签数据集的特点, 将准确率评估方法中的函数 $T(F(X_i), C_i)$ 修改为: $T(F(X_i), C_i, \text{MLD})$, 其中, MLD 表示训练数据集中属于多个类标签元组的集合。如果 $F(X_i) = C_i$, 则返回 1; 如果 $F(X_i) \neq C_i$, 而 $X_i \in \text{MLD}$, 则返回 1; 如果 $F(X_i) \neq C_i$, $X_i \notin \text{MLD}$, 则返回 0。公式表示为:

$$T(F(X_i), C_i, \text{MLD}) = \begin{cases} 1, & F(X_i) = C_i, \text{ or } F(X_i) \neq C_i, X_i \in \text{MLD} \\ 0, & F(X_i) \neq C_i, X_i \notin \text{MLD} \end{cases}$$

由上可知, EMOSIML 评估方法的计算公式可以表示为:

$$\text{Accuracy} = \left(\sum_{i=1}^{|D|} T(F(X_i), C_i, \text{MLD}) \right) / |D|$$

文中主要解决的是二分类多标签分类评估问题,

X_i 只可能同时属于类别 1 和类别 2, 分类器对 X_i 预测的类标签只会是类别 1 和类别 2, 当 $F(X_i) \neq C_i$ 时, 此处只需判断 $X_i \in \text{MLD}$ 与否, 而不需要考虑分类器对元组 X_i 预测的具体类别标签。EMOSIML 评估算法的执行效率与准确性评估方法是同一数量级时间复杂度。

3 实验与结果分析

以上已经详细介绍了 EMOSIML 的思想和具体评估方法, 为了验证 EMOSIML 评估算法的有效性, 文中用 C# 语言编程实现了该评估算法。从 UCI (University of California in Irvine) 上下载标准数据集, 对数据集进行预处理, 将 SISL 分类评估问题拓展为 SIML 分类评估问题, 数据集的信息如表 2 所示。在训练之前, 如果训练集中有同一属性不同类别标签的元组, 则需要将这些元组保存一个数据结构中 (此处为二维表), 以便评估所用。具体的做法是: 数据集中存在元组 $X_i (i = 1, 2, \dots, n)$, 如果存在 $X_i = X_j (i \neq j)$ (对应字段的属性值相同), 但对应的类标签 $C_i \neq C_j$, 则将该组元组中的任何一个 X_i 存储在表 MLD (Multi Label Data) 中。最终将训练集中具有此类特性的元组全部存入表 MLD 中。

表 2 UCI 标准数据集的概况

数据集	属性数	类别数	实例总数
Pima	5	2	768
Diabetes	7	2	768
Heart	8	2	270
Glass2	6	2	163
Post operative	8	2	243

与此同时, 作者用 C# 语言开发了朴素贝叶斯分类器, 并以之为实验测试平台, 采用 5 叠交叉验证方法, 分别用传统的准确率评估方法和文中提出的 EMOSIML 评估方法来评估分类器的准确率, 实验结果如表 3 所示。

表 3 试验结果

数据集	Multi Label Correct	传统的评估方法	EMOSIML 评估方法
Pima	127	79.638%	96.186%
Diabetes	96	77.516%	90.065%
Heart	5	81.088%	82.971%
Glass2	21	83.411%	95.938%
Post operative	3	74.884%	76.549%

表3中,列 Multi Label Correct 表示在传统的准确率评估中,记为错误的实例数,而在文中提出的 EMOSIML 评估方法中认为是正确分类的,而且这些元组均同时属于两个类标签。对于数据集 Corral、Diabetes 和 Glass2, multi label correct 对应的元组在各训练集中所占的比率较大,所以用 EMOSIML 评估方法得到的准确率比 Accuracy 评估方法得到的准确率相对较高。而数据集 Heart 和 Post operative 中, multi label correct 对应的元组在各训练集中所占的比率较小,所以两种评估方法得到的准确率差异较小。

对于以上的五个数据集,在利用朴素贝叶斯分类模型进行分类时,由于数据集中存在元组同时属于多个类别标签,如果使用传统的准确率评估方法, multi label correct 这一列所标示数目的多标签元组,在对每个数据集的分类评估过程中都是被判为错误分类的,所以利用传统的准确率评估方法评估得到的分类器准确率相对较低,而 EMOSIML 评估方法能够很好地解决 SIML 分类评估问题,实验证明文中提出的准确率评估方法较合理。

4 结束语

文中首先对单实例多标签、多实例单标签和多实例多标签分类问题进行了简单的介绍,然后介绍传统的准确率评估方法。由此对 SIML 分类问题提出了一种新的评估方法 EMOSIML,该评估方法的思路是:只要分类器将一个对象分给它所属多个类标签中的任何一个,分类结果都是正确和合理的,再次介绍了该评估方法的计算方法,最后通过实验证明 EMOSIML 评估方法的有效性和合理性。试验所用到的数据集类别数均为两类,使用该评估方法评估两个以上类别的数据集构建的分类模型,也是值得研究的。同时,对 SIML、MISL 和 MIML 分类问题分类算法的设计和评估将是值得研究的一个方向。

下一步的工作:首先利用贝叶斯算法对多类别多

标签数据集构建分类器,并在 EMOSIML 的基础上进行改进评估方法,以适合评估多类别多标签分类问题。其次在现有的多标签分类算法的基础上,提出一种新的多标签分类算法。

参考文献:

- [1] Hanjiawei, Kamber M. Data Mining Concepts and Techniques [M]. [s. l.]: Morgan Kaufmann publishers, 2000.
- [2] Schapire R E, Singer Y. BoostTexter: A boosting-based system for text categorization[J]. Machine Learning, 2000, 39(2-3): 135-168.
- [3] Dietterich T G, Lathrop R H, Lozano-Perez T. Solving the multi-instance problem with axis-parallel rectangles[J]. Artificial Intelligence, 1997, 89(1-2): 31-71.
- [4] Zhou Z H, Zhang M L. Multi-instance multi-label learning with application to scene classification[M]//In Advances in Neural Information Processing Systems 19. Cambridge, MA: MIT Press, 2007: 1609-1616.
- [5] Zhang M L, Zhou Z H. M³MIML: A maximum margin method for multi-instance multi-label learning[C]//In: Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08). Pisa, Italy: [s. n.], 2008: 688-697.
- [6] Zhang M L, Zhou Z H. ML-kNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.
- [7] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions[J]. Machine Learning, 1999, 27(3): 297-336.
- [8] 程泽凯, 林士敏. 文本分类器的准确性评估方法[J]. 情报学报 2004, 23(5): 631-636.
- [9] 秦 锋, 杨 波, 程泽凯. 分类器性能评价标准研究[J]. 计算机技术与发展, 2006, 16(10): 85-88.
- [10] 秦 锋, 罗 慧, 程泽凯, 等. 一种新的基于 AUC 的多类分类评估方法[J]. 计算机工程与应用, 2008, 44(5): 194-196.
- [11] Boutell M R, Luo Jiebo, Shen Xipeng, et al. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37: 1757-1771.

(上接第45页)

- [3] 侯要红, 栗松涛. Java XML 应用程序设计[M]. 北京: 机械工业出版社, 2007.
- [4] Burke E M. Java 与 XSLT[M]. 高 伟, 英 宇, 译. 北京: 中国电力出版社, 2003.
- [5] Busatto G, Lohrey M, Maneth S. Efficient memory representation of XML document trees[J]. Information Systems, 2008, 33(4-5): 456-474.
- [6] Wang Fangju, Li Jing, Homayounfar H. A space efficient XML DOM parser[J]. Data & Knowledge Engineering, 2007, 60

(1): 185-207.

- [7] 蔚晓娟, 冉 静, 李爱华, 等. 基于 DOM 的 XML 解析与应用[J]. 计算机技术与发展, 2007, 17(4): 86-88.
- [8] 陈 娟, 李 晖, 鱼 雷. XML 文档快速解析技术研究[J]. 计算机技术与发展, 2007, 17(10): 40-42.
- [9] 陈 奇. XSLT、XPath 和 DOM 的应用研究[J]. 计算机工程, 2003, 29(3): 14-16.
- [10] 蔡 剑, 景 楠. Java 网络程序设计: J2EE(含 1.4 最新功能)[M]. 北京: 清华大学出版社, 2003.