基于多值分解和多类标学习的分类框架设计①

沈良忠 陈胜凯 胡捷臻 (温州大学 城市学院 浙江 温州 325000)

摘 要: 多值多类标的数据分类是研究一个样本不但同时属于多个类别，而且在某些属性下也可能存在多个取值的问题。提出了一种结合多值分解和多类标学习的多值多类标分类框架（MDML），采用 4 种不同的多值分解策略，将问题转化为多类标问题，然后利用 3 种经典的多类标算法进行学习。实验结果表明，MDML 与已有的多值多类标决策树算法相比，有效地提高了分类的性能，而且不同的组合方法适用于不同特点的数据集。

关键词: 分类；多值属性分解；多类标数据；数据转化

Framework of Classification Based on Multi-Value Decomposition and Multi-Label Learning

SHEN Liang-Zhong, CHEN Sheng-Kai, HU Jie-Zhen

(City College, Wenzhou University, Wenzhou 325000, China)

Abstract: Classification of multi-valued and multi-labeled data is about a sample which is not only associated with a set of labels, but also with several values that include some attributes. This paper proposes a multi-valued and multi-labeled learning framework that combines multi-value decomposition with multi-label learning (MDML), using four strategies to deal with multi-valued attributes and three classical, multi-label algorithms to learn. Experimental results demonstrate that MDML significantly outperforms the decision tree based method. Meanwhile, combined methods can be applied to various types of datasets.

Keywords: classification; multi-label data; multi-valued attribute decomposition; data transformation

1 引言

多值多类标的数据分类是研究一个样本在条件属性和分类属性下取值可能为多个的分类问题[1]，如表 1 给出一个有关报刊征订的多值多类标数据的例子。一个客户可能有多个爱好而且曾经订过不同的刊物，因此每条客户记录在“爱好”和“刊物类型”属性下可能有多个取值。如果报刊征订部门希望根据这些历史记录来预测客户可能订购哪些类型的刊物，这就是一个典型的多值多类标分类问题。

表 1 有关报刊征订的具有多值属性的多类标数

id	性别	年龄	爱好	刊物类型
1	女	24	旅游、看电视	C2、C3
2	女	50	看电视	C1
3	男	51	看电视	C1、C2

为了解决多值多类标的数据分类问题，一种最直接的方法就是将它转化为单值单类标的数据，但是目前还没有专门针对转化的方法进行充分的研究，当多值属性数目比较多的时候，不当的方法将导致数据的激增，而且转化后的数据集，很多样本的区别仅仅在于多值属性下的取值不同，这将对最终构建的分类器的性能产生不利的影响。目前解决该问题的方法主要是基于多值多类标决策树的方法[1,2]，但是分类精度普遍偏低；而多类标分类算法[3]没有考虑多值属性的问题。本文提出了一种结合多值分解和多类标算法的学习框架，有效提高了多值多类标的数据分类精度。

2 多值多类标分类算法

2.1 多值属性分解策略

假设 D 表示包含 N 条多值多类标数据的训练集

① 收稿时间:2010-02-24;收到修改稿时间:2010-04-04

合，d_i 表示第 i 个样本，D’表示经过多值分解的单值多类标测试数据集；A 表示包含全部属性的集合，$|A|=n$，A_j 表示第 j 个属性，$WH(A_j)$表示 A_j 取值的全集，$MAX(A_j)$表示样本在 A_j 下最大的取值个数，$MAX(A_j) \leqq |WH(A_j)|$；V_i 表示 d_i 在 A_j 下的一个取值集合，$n_{i,j}$ 表示 d_i 在 A_j 的取值个数，$n_{i,j} \leqq MAX(A_j)$；L 表示包含所有类标的集合，$|L|=k$。

由于多类标的数据可以看作多值属性多类标的数据在多值属性下取值个数为 1 的一个特例，因此可以通过多值属性分解的方法，将其转化为多类标的数据进行学习，按照考虑多值之间相关信息的不同，下面给出 4 种多值分解策略：

1) 多值拆分的方法(MD)：对于样本 d_i 在所有的多值属性下，按照不同的取值，MD 将其分解为多条数据，那么 d_i 将分解为，其中当 A_j 为单值属性时，$n_{i,j}=1$，当 A_j 为多值属性时 $n_{i,j} \geqq 1$。MD 将多值看作是相互独立的，将它们之间进行拆分形成多条数据，这些数据属于相同的多类标集合，区别仅在于多值属性下取值不同。经过 MD 分解之后，数据集合 D 的大小变为原来的倍，这大大增加了数据的存储空间，也将影响算法的运行效率。

2) 随机取值的方法(RD)：RD 采用随机取 1 的策略作为样本在当前多值属性下的取值，从而避免了分解后数据存在的取值重叠问题。RD 的优点在于简单，不增加数据集合的大小，缺点在于可能会丢弃一些对于判定类别的重要的取值。

3) 定义新值的方法(ND)。ND 将每个可能出现的多值的取值集合看作一个新的值，它考虑了多值集合内部各个取值之间的相关信息，但忽略了多值集合之间的相关关系。如果多值集合之间存在相同的取值，那么它们之间就必然存在一定的相关关系，但是经过 ND 转换之后的新值就消除了这种相关信息，因此造成了部分重要信息的丢失。

4) 数据编码的方法(CD)。CD 将多值属性看作多个子属性的集合，它将当前多值属性下的每个可能的取值都看作是平等的。对于每一个 V_iA_j，它将被转化为一个长度为$|A_j|$的一个 0，1 向量，如果 V_i 包含 A_j 中的第 m 个值，则 K=1，否则 K =0。CD 的优点在于更多保留了多值之间的相关信息，但是它将多值属性的个数增大到倍，因此将对算法的运行效率产生较大的影响。

2.2 多类标学习算法

目前的多类标算法主要可以分为问题转化和算法改进两种方法[3]。问题转化的方法是通过一定的类标处理策略，将多类标问题转化为单类标问题，其中使用最广泛的两种方法是 BR 和 LP。BR 采用 1-VS-ALL 的策略将多类标问题转化为多个独立的单类标问题进行学习，由于它忽略了类标之间的相关关系，影响了最终建立的分类器的性能。LP 考虑了类标之间的相关关系，将不同的类标组合看作新类别，那么一个包含 K 个类标的数据集合，理论上它的类别数可能达到 2^K-1。当 K 比较大的时候会造成每种类标组合的样本数目很少，这直接影响了分类器的性能。算法改进是通过对已有的算法拓展使之能够处理多类标分类问题。Clare 等人对 C4.5 算法进行拓展，提出了多类标熵公式来处理多类标的数据分类[4]；Elisseeff 等人提出一种基于类标分级的算法来处理多类标分类[5]；周志华等人考虑类标的先验知识，对每个类标进行 KNN 学习，提出了 ML-KNN 算法[6]，其他的算法还包括多类标支持向量机、多类标神经网络等。

2.3 MDML 学习框架

由于 BR 和 LP 是问题转化方法中最具有代表性的两种算法，ML-KNN 是算法改进方法中高性能算法的代表，因此选择 BR、LP 和 ML-KNN 算法进行学习。整个学习系统可以分为 3 个部分：①多值属性分解；②多类标学习；③测试集合类标预测，如图 1 所示。

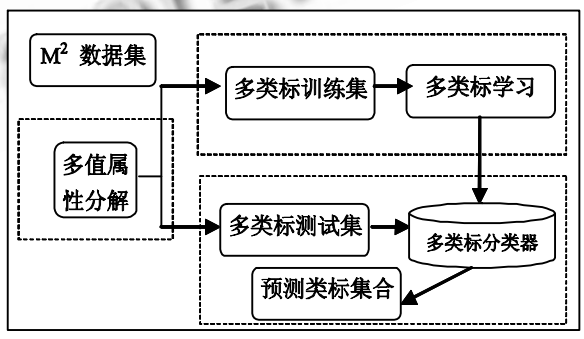

图 1　MDML 学习系统

3　实验和结果分析

3.1 数据集合

实验采用和文献 2 一样的合成数据集合进行训练和测试，数据集合的说明如表 2 所示。

表 2　实验数据属性的描述

属性	属性类型	取值个	取值范围
年龄	连续型、单	1	20～80 的整数
汽车	离散型、多	1～3	1～20 的整数
学历	离散型、单	1	1～5 的整数
性别	离散型、单	1	1，2
爱好	离散型、多	1～5	1～20 的整数
职业	离散型、多	1～2	1～10 的整数
婚姻	离散型、单	1	1～3 的整数
收入	连续型、单	1	20,000~150,000

其中，属性“爱好”、”“汽车”和“职业”是多值属性，其他的为单值属性，属性“年龄”和“收入”为连续型属性，其他的为离散型数据。表 3 给出根据属性值将客户分成 5 个不同类型的分类规则。

表 3　根据属性值判断客户类型的规则

类标	规则
C1	[性别=0∧收入∈[20000,100000)∧汽车∈[1,4]] ∨ [性别=1∧收入∈[100000,150000)∧汽车∈[5,8]] ∨职业∈{1,6}
C2	[[年龄∈[20,40) 汽车∈[1,4]] ∨[[年龄∈[40,60) 汽车∈[5,6]] ∨职业=2
C3	[婚姻=1∧收入∈[20000,100000)∧(爱好∈[1,3]∨汽车=4)] ∨ [婚姻=∈{2,3}∧收入∈[100000,150000)∧(爱好∈[5,7]∨汽车=8)] ∨职业∈[3,5]
C4	[婚姻=1∧收入∈[20000,40000)] ∨[婚姻=2∧收入∈[40000,80000)] ∨职业∈{7,8}
C5	[教育=1∧爱好∈[1,3] ∧婚姻=1] ∨[教育=3∧爱好∈[4,8] ∧婚姻∈{2,3}] ∨职业∈{9,10}

3.2 评价指标

假设 $y_i \subseteq L$ (L 为类标总集)是对应样本 d_i 的预测类标，真实类标集为 s_i，那么对于测试数据集合 D’，本文选择以下 3 个指标[3]对多值多类标的分类结果进行评价。

汉明距离：

$$Hloss(D') = 1 - \frac{1}{|D'|}\sum_{i=1}^{|D'|}\frac{|s_i \oplus y_i|}{|L|} \qquad (1)$$

正确率：

$$Acc(D') = \frac{1}{|D'|}\sum_{i=1}^{|D'|}\frac{|s_i \cap y_i|}{|s_i \cup y_i|} \qquad (2)$$

此外，为了对正确率 p_i 和召回率 r_i 有一个综合的度量，文中还选择了 F_1 度量：

$$F_1(D') = \frac{1}{|D'|}\sum_{i=1}^{|D'|}\frac{2 \times p_i \times r_i}{p_i + r_i} \qquad (3)$$

由于相似度的定义和汉明距离以及正确率的定义比较接近，因此文中不单独计算它的实验结果值。

3.3 实验设计

在算法对比中，对于多值多类标决策树算法，本文选择具有代表性且分类精度较高的算法 SSC。SSC 算法、数据产生和转化的程序都是在 Matlab 7.0 下实现的，对于 BR，LP 以及 ML-KNN 部分的实验都是在 Weka[7]平台上进行。

对于 SSC 参数设置如下：ub=4，Sup_{min}=50%，$Diff_{min}$=15%，Num_{min}=2，=0.5。在 BR 和 LP 中，选择 SMO[3]作为基分类器，ML-KNN 中的近邻数目设置为 10。实验中数据集合的数目依次为 2000，4000，6000，8000，采用十折交叉的方法进行结果的验证。实验计算机配置为 Inter(R) Dual E2180，CPU 频率为 2.00GHZ，内存为 1GB。

表 4 给出了 SSC 和 ML-KNN 的实验结果，MK1 代表 MD+ML-KNN 的组合，MK2 代表 RD/ND/CD+ML-KNN 的组合，粗体表示在当前算法组内性能最好。

表 4　SSC 和 ML-KNN 的实验结果

评价指标	算法	N=2000	N=4000	N=6000	N=8000
Hoss	SSC	0.3072	0.2934	0.2845	0.2827
	MK1	0.0452	0.0417	0.0382	0.0352
	MK2	0.1494	0.1502	0.1493	0.1513
Acc	SSC	0.6990	0.7144	0.6983	0.6976
	MK1	0.8922	0.9141	0.9332	0.9367
	MK2	0.6984	0.7160	0.7152	0.7183
F_1	SSC	0.7308	0.7315	0.7293	0.7285
	MK1	0.9350	0.9487	0.9612	0.9676
	MK2	0.8213	0.8182	0.8202	0.8294

由于在实验中 RD/ND/CD+ML-KNN 的组合几乎得到相同的实验结果，因此表 4 中只给出一组平均

结果，根据表 4 的实验结果，分析如下：

1) MK2 中不同组合取得相同结果的原因：由于实验数据集合属性数目和类标数目较少，虽然同一个数据集合在 RD、ND 和 CD 下，属性分解之后的数据集合属性个数分别变为 9，9，56，但是样本之间的欧式距离却几乎保持不变，因此造成了不同算法在同一个数据集合上得到相近或者相同的结果；同时由于 ML-KNN 是懒惰学习算法，它预测样本的类标集合不是基于构造好的分类模型，而是每次去寻找样本对应的 K 个近邻，而随着样本数目的变化，K 近邻相对保持稳定，因此性能随着样本数目的变化保持稳定。

2) 对比多值多类标决策树算法 SSC，MK1 和 MK2 的性能都比 SSC 高的原因：决策树算法在于生成有用的规则而不在于完全分类；对于多值的处理，SSC 采用分支的策略没有考虑多值属性下多个取值之间的相关关系，而 ML-KNN 能够利用训练集合中类标的先验统计信息，因此能够提高预测的性能。

3) MK1 比 MK2 性能高的原因：在 MK1 中，一条多值样本经过 MD 的分解之后，变成了多条只在多值属性下取值不同的单值样本，在进行预测的时候，一个测试样本的 K 近邻中将有多个来源于同一个多值样本分解而来，而这些样本的类标集合是一样的，那么根据 ML-KNN 的判定原理，测试样本能够得到更加一致和准确的预测结果；而对于 MK2 由于 RD、ND 和 CD 三种策略都将整个训练样本看作一个整体对待，因此 K 近邻的判定的误差要大于 MK1，所以 MK1 的性能要比 MK2 好。

表 5 和表 6 分别给出了 BR 和 LP 下不同组合的算法的结果，粗体表示当前性能最好的组合算法，斜体表示当前性能最差的组合算法。

表 5　基于 BR 的组合算法实验结果

评价指标	算法	N=2000	N=4000	N=6000	N=8000
Hoss	MD+BR	0.0653	0.0534	0.0478	0.0381
	RD+BR	0.0442	0.0370	0.0320	0.0324
	ND+BR	0.0457	0.0426	0.0361	0.0352
	CD+BR	0.0596	0.0513	0.0432	0.0373
Acc	MD+BR	0.8807	0.8914	0.9144	0.9113
	RD+BR	0.9121	0.9283	0.9334	0.9368
	ND+BR	0.9082	0.9178	0.9295	0.9306
	CD+BR	0.8925	0.9084	0.9202	0.9246
F_1	MD+BR	0.9114	0.9256	0.9267	0.9292
	RD+BR	0.9356	0.9461	0.9567	0.9634
	ND+BR	0.9253	0.9358	0.9362	0.9366
	CD+BR	0.9192	0.9273	0.9287	0.9342

表 6　基于 LP 的组合算法实验结果

评价指标	算法	N=2000	N=4000	N=6000	N=8000
Hoss	MD+LP	0.0576	0.0465	0.0382	0.0335
	RD+LP	0.0316	0.0292	0.0273	0.0262
	ND+LP	0.0386	0.0334	0.0292	0.0287
	CD+LP	0.0627	0.0572	0.0447	0.0463
Acc	MD+LP	0.9068	0.9273	0.9006	0.9334
	RD+LP	0.9321	0.9433	0.9440	0.9481
	ND+LP	0.9185	0.9235	0.9105	0.9292
	CD+LP	0.8981	0.9152	0.9187	0.9137
F_1	MD+LP	0.9483	0.9512	0.9563	0.9665
	RD+LP	0.9654	0.9679	0.9701	0.9713
	ND+LP	0.9534	0.9618	0.9674	0.9702
	CD+LP	0.9248	0.9322	0.9432	0.9431

如表 5 所示，RD+BR 的性能是最好的，而 MD+BR 的性能是四种组合中最差的，分析原因如下：在 RD+BR 组合算法下，多值属性的多个取值之间的关系被弱化；多值属性的多个取值之间的关系比较紧密，因此随机选择的那个值保留了大部分的重要信息；而对于 MD+BR，忽略类标之间的相关信息，并且通过 BR 进行学习的单个类标的数据集合存在大量的重合样本，因此导致了单个分类器弱分类性能。ND 考虑了类标之间的相关信息，在保持数据维度的前提下，取得了与 RD 相近的分类性能，而 CD 增大了数据的维度，加入的一些空值可能被识别为噪声，因此分类性能不及 RD 和 ND。

由表 6 可知 4 种组合算法的分类性能从优到差依次为：RD+LP，ND+LP，MD+LP，CD+LP。由于类标总数 K=5，而 $N>>2^K-1$，数据集合为 LP 策略下每个类标组合提供了足够的样本信息进行学习，因此 4 种组合算法的整体分类性能比 BR 下的要好。其中 ND 和 RD 的性能很接近，而由于是对整个类标组合进行分类器学习，MD 在 LP 下也取得了较高的性能，CD 在 LP 下性能比其他 3 个要差的原因是编码造成的稀疏矩阵影响了分类器的学习。

(下转第 22 页)

(上接第 190 页)

4 总结

本文提出结合 4 种多值属性分解方法和 3 种经典的多类标算法进行多值多类标分类的学习框架，实验结果表明：对比已有的多值多类标决策树算法，大部分的组合算法都显著地提高了多值多类标数据的分类性能，而且 RD+LP 适用于数据维度低、类标数目少的数据集合，在实际的数据量大、维度高、类标数目多的多值多类标数据集合上，ND+LP 的性能将优于 RD+LP 成为最优或者近优的学习算法。

参考文献

1 赵蕊,李宏.一种多值属性多类标数据的决策树算法.计算机工程, 2007,33(13):87－89.

2 Chou S, Hsu C. MMDT: a multi-valued and multi-labeled decision tree classifier for data mining. Expert Systems with Applications, 2005,28(2):799－812.

3 Tsoumakas G, Katakis I. Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, 2007,3(3):1－13.

4 Clare A, King RD. Knowledge Discovery in Multi-label Phenotype Data. Lecture Notes in Computer Science Vol. 2168, Springer, Berlin 2001.

5 Eisseeff A, Weston J. A kernel method for multi-labelled classification. Dietterich TG, Becker S, Ghahramani Z, Editors, Advances in Neural Information Processing Systems 14, MIT Press, Cambridge, MA, 2002:681－687.

6 Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 2007,40(7):2038－2048.

7 Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2005.