

面向主属性值的类标特征分析

张明卫¹, 张小旭², 刘莹¹, 韩春燕¹

(1. 东北大学 软件学院, 辽宁 沈阳 110169; 2. 浙江大学 计算机科学与技术学院, 浙江 杭州 310058)

摘 要: 为了提取一个类标区别于其他类标的本质特征, 增强类标数据集的可解释性, 提出了一种面向主属性值的类标特征分析方法. 该方法首先建立了一种直观的面向主属性值的类标特征模型, 然后设计了对应的类标特征抽取算法, 最后给出了一种基于类标特征分析的分类算法. 实验结果表明, 所建立的类标特征模型能够直观、有效地描述类标数据集中各类标的特征, 给出的类标特征抽取算法有较高的执行性能, 提出的分类算法在针对类标较少的数据集时有较高的分类准确率.

关 键 词: 数据挖掘; 分类; 聚类; 类标特征; 主属性值

中图分类号: TP 311

文献标志码: A

文章编号: 1005 - 3026(2016)10 - 1388 - 05

Primary Value Oriented Class Label Characteristic Analysis

ZHANG Ming-wei¹, ZHANG Xiao-xu², LIU Ying¹, HAN Chun-yan¹

(1. School of Software, Northeastern University, Shenyang 110169, China; 2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China. Corresponding author: ZHANG Ming-wei, E-mail: zhangmw@swc.neu.edu.cn)

Abstract: A primary value oriented class label characteristic analyzing approach was proposed to extract the essential characteristics of one class label distinguishing with the others. In addition, the interpretability of label datasets could also be improved by this proposed method. Firstly, an intuitive primary value oriented class label characteristic model was built. Then, the corresponding class label characteristic extracting algorithm was designed. Finally, a classification algorithm was presented based on class label characteristic analysis. Experimental results demonstrated that the class label characteristic model can describe the characteristics of each class label for label datasets intuitively and effectively, and the given class label characteristic extracting algorithm has high execution performance. What's more, the proposed classification algorithm has relatively high accuracy for datasets with fewer class labels.

Key words: data mining; classification; clustering; class label characteristic; primary value

数据挖掘是知识发现的一种手段, 能够从大量数据中抽取隐含的、先前未知的、对决策有潜在价值的规则^[1]. 数据分类和聚类均是数据挖掘的主要研究内容之一. 前者主要是通过分析训练数据样本, 构造一个分类模型, 能够把新采集的数据映射到给定类别中的某一个. 后者则是按照相似程度把大量的数据样本聚成几个类, 使得同一类内样本的相似性较大, 而不同类间样本的相似性较小. 无论是分类所使用的训练数据集, 还是聚类的结果数据集, 它们所包含的每个元组都具有类

标, 即属于某一个类别, 称此类数据集为类标数据集. 针对类标数据集, 如何建立一种知识模型, 并设计其对应的挖掘方法, 能够准确地反映各类标区别于其他类标的本质特征, 帮助领域专家更好地认识各个类标, 即类标特征分析将成为数据挖掘领域一个有实际应用需求和价值的问题.

以分类为例, 应用不同的分类模型, 比如决策树模型^[2]、贝叶斯模型^[3]、基于规则模型^[4]、神经网络模型^[5]、K - 近邻模型^[6]和关联分类模型^[7-8]等可以对类标数据集进行较高效、较准确

地分类.然而,这些模型并不以各类标的解释和帮助领域专家对各类标的认知为目标,甚至建立的某些种类的分类模型很难被专家理解.以聚类为例,其是一种在没有预先指定类别的前提下,用“物以类聚”的思想来分析数据的常用方法.然而聚类所产生的结果——类标数据集的可解释性是聚类分析方法在应用层面上取得成功的关键.只有领域专家充分理解聚出来的各个类标,才能更好地应用该类知识进行给定领域的分析.现有的聚类模型有很多^[9-12],但缺乏对聚类知识描述和聚类结果解释的研究.因而,无论是针对分类的训练集、还是聚类的结果集或者其他类型的类标数据集,类标特征分析将促进相关知识模型在各领域的应用,成为一项有价值的数据分析方法,但现在尚缺乏针对性研究.

为了增强类标数据集面向领域专家的可解释性,本文从统计学角度提出一种面向主属性值的类标特征分析方法.主属性值即为一个类标区别于其他类标出现的主要属性值,它反映了同一类标内数据的同质性和类标间数据的差异性,同时考虑了所建模型的鲁棒性.本文方法的大致过程包括:给出主属性值的定义,并建立面向主属性值的类标特征模型;设计面向主属性值的类标特征抽取算法;将建立的类标特征模型应用于分类,给出一种基于主属性值的数据分类方法.主属性值直观地表述了各类标数据的特征,便于领域专家理解,是一种简单有效的类标特征分析方法.另外,将面向主属性值的类标特征应用于分类中,发现针对类标较少的数据集,可以提升分类准确率,可为以后分类研究提供新的思路.

1 面向主属性值的类标特征建模

本文以增强类标数据集的可解释性为目标,建立了一种面向主属性值的类标特征模型.下面给出相关定义.

设 D 是一个类标数据集,具有类标号属性 $C = \{c_1, c_2, \dots, c_k\}$, 即有 k 个类标号. 给定 D 的某一特征属性 $A = \{a_1, a_2, \dots, a_l\}$, 属性值 v 是一个值对,即为属性 A 及在该属性上的取值 a_p ($1 \leq p \leq l$) 的组合,记作 $v = A \mu_p$.

定义 1 覆盖度(coverage, 简记为 cov): 给定一类标数据集 D , 则属性值 v 在数据集 D 上针对类标 c_q ($1 \leq q \leq k$) 的覆盖度为

$$\text{cov}(v, c_q) = P(v \cup c_q) / P(c_q) = |v \cup c_q| / |c_q|.$$

(1)

其中 $P(v \cup c_q)$ 和 $P(c_q)$ 分别为 $v \cup c_q$ 和 c_q 在 D 上出现的概率; $|v \cup c_q|$ 和 $|c_q|$ 分别为 $v \cup c_q$ 和 c_q 在 D 上出现的次数. 由定义可知, $\text{cov}(v, c_q) \in [0, 1]$, 描述了属性值 v 在类标为 c_q 样本中的覆盖程度. 给定最小覆盖度阈值 min_cov , 如果 $\text{cov}(v, c_q) \geq \text{min_cov}$, 则称 v 为 D 中针对类标 c_q 的频繁属性值.

定义 2 特异度(specificity, 简称 spec): 给定一类标数据集 D , 则属性值 $v = A \mu_p$ 在数据集 D 上针对类标 c_q ($1 \leq q \leq k$) 的特异度为

$$\text{spec}(v, c_q) = 1 - \frac{P(v \cup \overline{c_q})}{P(\overline{c_q})} = 1 - \frac{|v| - |v \cap c_q|}{|D| - |c_q|}. \quad (2)$$

其中 $P(\overline{c_q})$ 为在类标数据集 D 中非 c_q 类标出现的概率; $P(v \cup \overline{c_q})$ 为属性值 v 和非 c_q 类标在 D 中共同出现的概率. 由定义可知, $\text{spec}(v, c_q) \in [0, 1]$, 描述了属性值 v 对类标 c_q 的特属程度. 当属性值 v 对类标 c_q 的特异度越高, 则 v 在非 c_q 类标中出现的概率越小, 从而 $\text{spec}(v, c_q)$ 值越大. 给定最小特异度阈值 min_spec , 如果 $\text{spec}(v, c_q) \geq \text{min_spec}$, 则称 v 为 D 中针对类标 c_q 的特异属性值.

定义 3 主属性值(primary value, 简记为 pv): 给定类标数据集 D , 类标 c_q 及属性值 v , 如果 v 在 D 中针对类标 c_q 既是频繁属性值, 又是特异属性值, 则称属性值 v 为 D 中类标 c_q 的一个主属性值.

从直观上理解, 一个类标的主属性值就是在该类标中频繁出现、而在其他类标中出现较少的属性值. 根据类标数据集中的数据分布情况和阈值设定的大小, 一个类标 c_q 在给定特征属性 A 上可能有零到多个主属性值, 由这些所有的主属性值组成的集合 S_{pv}^A 称之为类标 c_q 在特征属性 A 上的主属性值全集.

定义 4 类标特征(class label characteristic, 简记为 lc): 给定类标数据集 D , 其由 m 个特征属性 $\{A_1, A_2, \dots, A_m\}$ 和一个类标号属性 $C = \{c_1, c_2, \dots, c_k\}$ 描述. c_q ($1 \leq q \leq k$) 是 D 中的某个类标. 则定义 c_q 的类标特征 $\text{lc}(c_q) = S_{pv}^1, S_{pv}^2, \dots, S_{pv}^m$. 其中 S_{pv}^i 是类标 c_q 在特征属性 A_i 上的主属性值全集, 即一个类标的类标特征是由该类标在各个特征属性上的主属性值全集构成的向量.

以上所建立的面向主属性值的类标特征模型将各个特征属性独立考虑, 主要有两个出发点: 一是在实际应用中, 特别是针对大数据环境, 类标数据集所包含的各特征属性可能有不同的数据来源和存储方式, 将各特征属性分开处理可以提升类

标特征分析的效率和可行性 ;二是该模型较为直观 ,便于领域专家理解和对相关知识模型的应用.

2 面向主属性值的类标特征抽取

本文所设计的类标特征抽取算法的大致思路如下.

1) 扫描类标数据集 D 一次 ,统计各特征属性 $A_i(1 \leq i \leq m)$ 的所有属性值在各类标 $c_q(1 \leq q \leq k)$ 中出现的频率 ,得数据集 D 的属性值类标分布矩阵 M .

2) 基于矩阵 M ,针对各类标 c_q ,计算各属性值 v 在 c_q 中的覆盖度 ,筛选大于最小覆盖度阈值 \min_cov 的各项 ,得各类标的频繁属性值集.

3) 基于矩阵 M ,针对各类标 c_q 所对应的频繁属性值集中的各项 ,计算其特异度 ,筛选其中大于最小特异度阈值 \min_spec 的各项 ,即可得到各类标的主属性值集 S_{pv} .

本文面向主属性值的类标特征抽取算法 (primary value objected class label characteristic extraction algorithm , PVOCLCE)具体描述如下.

算法 1 PVOCLCE

输入 类标数据集 D ,最小覆盖度阈值 \min_cov ,最小特异度阈值 \min_spec

输出 各类标的主属性值集 S_{pv}

Count[$k + 1$] = Initiate(); //将 D 的元组总数及各类标个数写入数组 Count 中

$M[k][n]$ = Initiate(); //初始化存放各属性值在各类标出现频数的矩阵 M

FOR($t_j \in D$) //对于 D 中的任一元组 t_j

FOR($v \in t_j$) //对于 t_j 所包含的任一属性值 v

Count[q][v] + + ; // t_j 所对应的类标为 c_q

FOR ($c_q \in C$) //对于任一类标 c_q

FOR ($v \in D$) //对于任一属性值 v

IF($cov(v , c_q) \geq \min_cov$) // $cov(v , c_q)$ 为属性值 v 在类标 c_q 中的覆盖度

$S_q = \text{Insert}(v) ; // S_q$ 为类标 c_q 所

对应的全局频繁属性值集

FOR ($c_q \in C$)

FOR ($v \in S_q$)

IF($\text{spec}(v , c_q) < \min_spec$) // $\text{spec}(v , c_q)$ 为值 v 在类标 c_q 中的特异度

$S_q = S_q - v ;$

RETURN $S_{pv} = \bigcup_q S_q ;$

以上描述了在类标数据集 D 中抽取类标特征的算法 PVOCLCE. 设 D 中属性值个数为 m ,元组个数为 n ,则算法 PVOCLCE 的时间复杂度为 $O(m \times n)$. 因为 m 一般数值较小 ,且该算法仅需扫描一遍数据库即可求解 ,因而有着较高的运行效率和可伸缩性.

3 面向主属性值的类标特征分类

基于上节所抽取的面向主属性值的类标特征 ,可以构建一类面向主属性值的类标特征分类方法. 此类方法的大致思路比较简单 ,即分析新采集的未知类标的数据中包含的主要是哪个类标的主属性值 ,就应该属于该类标. 本文所设计的面向主属性值的类标特征分类算法 (primary value objected class label characteristic classification algorithm , PVOCLCC)的具体思路如下.

1) 扫描抽取出的类标数据集中各类标的类标特征 ,提取待分类元组 t 所包含的各类标 c_q 的主属性值集 $S_q = \{ pv_1 , pv_2 , \dots , pv_x \}$.

2) 依次基于以下规则来判断新元组 t 所属的类标.

2.1) 如果存在类标 $c_q \in C(1 \leq q \leq k)$,使得元组 t 所包含 c_q 的主属性值个数大于其他类标 ,则 t 的类标为 c_q .

2.2) 如果存在类标 $c_q \in C(1 \leq q \leq k)$,使得元组 t 对 c_q 的归属度最大 ,则 t 的类标为 c_q . 其中 ,归属度定义为

$$be(t , c_q) = \begin{cases} \max_{i=1}^x \left(\frac{|c_q| \times cov(pv_i , c_q)}{|c_q| \times cov(pv_i , c_q) + (n - |c_q|) \times (1 - spec(pv_i , c_q))} \right) , & S_q \neq \emptyset ; \\ 0 & S_q = \emptyset . \end{cases} \quad (3)$$

式中 n 为类标数据集中的样本总数. 当 $S_q = \emptyset$,即元组 t 中不包含类标 c_q 的主属性值时 ,则 t 对 c_q 的归属度为 0. 否则 ,对于 t 所包含 c_q 的任一主属性值 $pv_i \in S_q$,可按式 (3)求得当 pv_i 发生时类标 c_q 也发生的条件概率 ,此时 t 对 c_q 的归属度 $be(t , c_q)$ 定义为这些概率中的最大值.

2.3) 如果存在类标 $c_q \in C(1 \leq q \leq k)$,使得元组 t 到达类标 c_q 的欧几里得距离最小 ,则 t 的类标为 c_q .

以上大致描述了本文面向主属性值的类标特征分类算法 PVOCLCC 的思路. 因为算法过程较简单 ,这里不再详细展开.

4 实 验

实验将着重从以下 2 个方面分析文中方法的有效性。一是分析所建立的类标特征模型中两个重要阈值——最小覆盖度和最小特异度对类标主属性值个数的影响；二是分析所提出的面向主属性值的类标特征分类方法的性能。

4.1 阈值对类标主属性值个数的影响

最小覆盖度和最小特异度是本文所建立的类标特征模型中 2 个重要阈值,也是模型的核心,因而实验首先探索这两个阈值对主属性值个数的影响。为了阐述实验结果,定义 num_{avg} 为一个类标数据集所属各类标的主属性值个数的平均值。实验时采用 4 个典型的 UCI 数据集 letter-recognition、glass、iris 和 wine 进行分析。

1) 最小覆盖度阈值对主属性值个数的影响：设定最小特异度阈值 $min_spec = 0$,图 1 给出了各数据集的 num_{avg} 值随最小覆盖度阈值 min_cov 的变化趋势。

由图 1 可知,随着 min_cov 的增大,各数据集的 num_{avg} 值均逐渐减少。当 min_cov 依次取值 0.2、0.3、0.4、0.5 和 0.6 时,这 4 个数据集的 num_{avg} 值分别为 18.25、12.25、8.75、6.5 和 3.5。而当 min_cov 的值增加到 0.9 和 0.99,平均

主属性值个数则降到 0.5 和 0。

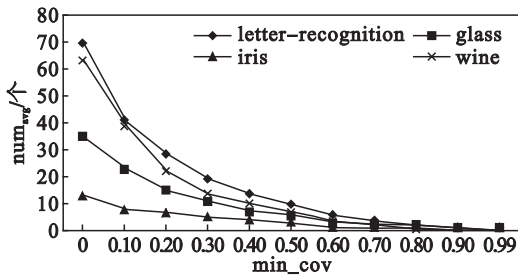


图 1 最小覆盖度阈值对主属性值个数的影响
Fig. 1 Influence of minimum coverage threshold on number of primary values

因为该实验分析时设定 $min_spec = 0$,所以实际上统计出来的是各数据集在各类标上的平均频繁属性值个数。最终产生的主属性值个数要比图 1 中的数据小,因为还需要满足最小特异度阈值。在设定 min_cov 时,不宜将其设置过小而违背主属性值的语义,同时也不宜将其设置过大而导致抽取出的主属性值个数太少。由图 1 中 4 个数据集的结果可以看出, min_cov 设定值不宜超过 0.6。

2) 最小特异度阈值对主属性值个数的影响：当 min_cov 分别取值为 0.0、0.4、0.5 和 0.55 时,各数据集的 num_{avg} 值随最小特异度阈值 min_spec 的变化曲线如图 2 所示。

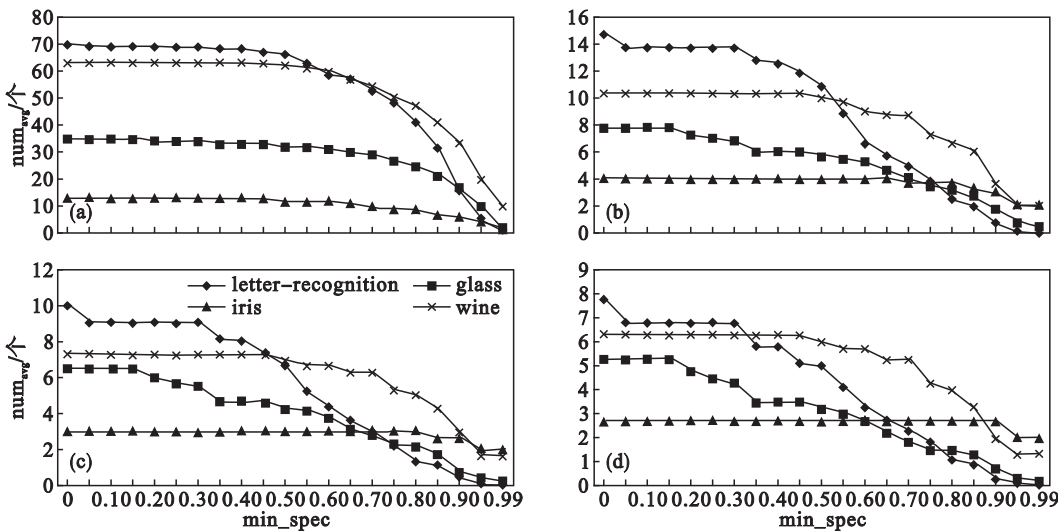


图 2 最小特异度阈值对主属性值个数的影响
Fig. 2 Influence of minimum specificity threshold on number of primary values
(a)— $min_cov = 0$; (b)— $min_cov = 0.4$; (c)— $min_cov = 0.5$; (d)— $min_cov = 0.55$.

由图 2 可知,对于大多数类标数据集,最小覆盖度阈值设定在 0.2 至 0.5 之间较为合理。而最小特异度阈值设定时应满足 $min_spec + min_cov$

> 1 的条件,一般在 0.7 至 0.95 的区间较为合理。在设定阈值时,为了使抽取出的主属性值个数合理,当 min_cov 较小时, min_spec 需适度增大;

而当 min_cov 较大时,min_spec 可适度减小.

4.2 算法 PVOCLCC 的分类性能

主要测试算法 PVOCLCC 的分类准确率,并与经典分类算法进行对比分析.表 1 分别列出了本文分类算法 PVOCLCC 和 C4.5,CBA^[7],CMAR^[8]这三个经典分类算法在各 UCI 数据集上的分类准确率.

表 1 分类准确率比较

Table 1 Comparison of classification accuracy %

数据集	属性数	类标数	实例数	C4.5	CBA	CMAR	PVOCLCC
Austral	14	2	690	84.7	84.9	86.1	88.2
Auto	25	7	205	80.1	78.3	78.1	75.0
Breast	10	2	699	95.0	96.3	96.4	96.8
Cleve	13	2	303	78.2	82.8	82.2	82.8
Crx	15	2	690	84.9	84.7	84.9	85.4
Diabete	8	2	768	74.2	74.5	75.8	76.7
German	20	2	1 000	72.3	73.4	74.9	75.0
Glass	9	7	214	68.7	73.9	70.1	65.0
Iris	4	3	150	95.3	94.7	94.0	97.4
Wine	13	3	178	92.7	95.0	95.0	98.0

由表 1 可知,本文面向主属性值的类标特征分类算法 PVOCLCC 虽然模型简单,但与经典决策树分类算法 C4.5 及经典关联分类算法 CBA 和 CMAR 相比,在以上 10 个数据集中,7 个分类准确率最优,2 个最差.由该实验可知,本文所给出的面向主属性值的分类算法 PVOCLCC 在针对类标较少的数据集时有较高的分类准确率.

5 结 论

- 1) 针对类标数据集,提出了有着广泛应用背景
- 的类标特征分析问题,可为以后此类研究提供原型参考.
- 2) 建立了面向主属性值的类标特征模型,能够直观、有效地描述类标数据集中的各个类标的特征,增强类标数据集的可解释性.
- 3) 设计了一种面向主属性值的类标特征抽取算法 PVOCLCE,有着较高的执行性能和可伸缩性,能够针对大多数类标数据集有效地抽取各

类标的特征.

4) 提出了一种基于类标特征分析的面向主属性值的分类算法 PVOCLCC,其有着较高的分类准确率,是一类值得深入研究的分类方法.

参考文献:

[1] Khamisi K ,Kazuto S ,Hideyuki T ,et al. Four decades of data mining in network and systems management[J]. *IEEE Transactions on Knowledge and Data Engineering* ,2015 27 (10) :2700 – 2716.

[2] Prabhu Y ,Varma M. FastXML :a fast ,accurate and stable tree-classifier for eXtreme multi-label learning [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York :ACM ,2014 263 – 272.

[3] Bi W ,Kwok J T. Bayes-optimal hierarchical multilabel classification[J]. *IEEE Transactions on Knowledge and Data Engineering* ,2015 27(11) :2907 – 2918.

[4] Wang J Y , Karypis G. On mining instance-centric classification rules[J]. *IEEE Transactions on Knowledge and Data Engineering* ,2006 ,18(8) :1497 – 1511.

[5] Tian J ,Li M ,Chen F. Learning subspace-based RBFNN using coevolutionary algorithm for complex classification tasks[J]. *IEEE Transactions on Neural Networks and Learning Systems* ,2016 27(1) :47 – 61.

[6] Samantha B K ,Elmehdwi Y ,Jiang W. K-nearest neighbor classification over semantically secure encrypted relational data[J]. *IEEE Transactions on Knowledge and Data Engineering* ,2015 27(6) :1261 – 1273.

[7] Liu B ,Hsu W , Ma Y. Integrating classification and association rule mining[C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York :ACM ,1998 80 – 86.

[8] Li W M ,Han J W ,Pei J. CMAR :accurate and efficient classification based on multiple class-association rules [C]// IEEE International Conference on Data Mining. Piscataway :IEEE ,2001 369 – 376.

[9] 张明卫 ,刘莹 ,张斌 ,等. 一种基于概念的数据聚类模型 [J]. *软件学报* ,2009 20(9) :2387 – 2396.
(Zhang Ming-wei ,Liu Ying ,Zhang Bin ,et al. Concept based data clustering mode[J]. *Journal of Software* ,2009 20(9) :2387 – 2396.)

[10] Hou Y ,Whang J J ,Gleich D F ,et al. Non-exhaustive ,overlapping clustering via low-rank semidefinite programming [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York :ACM ,2015 427 – 436.

[11] Mashayekhi H ,Habibi J ,Khalafbeigi T ,et al. GDCluster :a general decentralized clustering algorithm [J]. *IEEE Transactions on Knowledge and Data Engineering* ,2015 27 (7) :1892 – 1905.

[12] Liu H ,Liu T ,Wu J ,et al. Spectral ensemble clustering [C]// ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York :ACM ,2015 715 – 724.