基于组合关联分析的符号数据分类方法

崔丽娜[1]
CUI Lina

摘 要 分类是数据挖掘中一项非常重要的任务，目前大多分类算法处理的都是数值数据，少数处理符号数据的算法泛化性能不佳。针对这个问题，提出了一种基于组合关联分析的符号数据分类方法（categorical data classification approach based on combinatorial association analysis，CD_CAA）。通过引入提升度，将符号数据的属性与标签关联起来转化成数值数据后训练模型；在预测过程中，将测试数据与所有可能的类标签进行组合关联分析，将一个实际的测试样本转化为多个虚拟的测试样本，综合分析虚拟样本的假设类别标签，最终得到真实的预测标签。通过组合关联分析，将符号数据转换为数值数据，而转化过程所采用的提升度恰好反映了同一属性对不同类别的不同贡献，提高了模型泛化性能。实验结果表明，与传统机器学习方法相比，所提出的 CD_CAA 方法能更好地处理符号数据分类任务。

关键词 符号数据分类；CD_CAA 方法；提升度；组合关联分析；虚拟样本

doi：10.3969/j.issn.1672-9528.2024.06.017

0 引言

在数据挖掘领域，分类是一项关键任务[1]。对于数值属性数据的分类，已经提出了许多成功的算法，例如支持向量机（SVM）[2]、K 最近邻（K-nearest neighbor，KNN）[3-4] 分类算法以及人工神经网络（artificial neural network）[5] 等。然而，在实际应用中，符号数据（categorical data）[6-8] 普遍存在，例如医学诊断[9] 中，患者数据通常包含许多描述患者姓名、性别的标称属性，以及测量其生理参数的序数指标。一般标称属性数据指无序的符号属性数据。与数值数据相比，符号数据具有离散、非线性、非数值等特点，并且符号数据往往具有复杂的结构和语义，因此这类数据的分类任务比处理数值数据更具挑战性。

目前关于符号数据的分析主要集中在聚类分析方面。对于符号数据的分类，经典的算法有决策树[10]、朴素贝叶斯（naive bayes，NB）[11] 和基于距离的方法，如 KNN 和基于原型的分类器[12-13]。这些线性方法在处理具有多种属性类型的数据并进行分类时表现出了灵活性[14]。但是这些方法要么假定符号属性之间是相互独立的，而这与实际情况不符；要么计算符号样本的距离，而这种方法计算复杂度较高；要么通过简单的 0-1 相异性度量或者其扩展版本，但这种度量方法并不能很好地揭示符号数据的内在组成结构，这些问题都会影响最终的分类结果。因而，研究符号数据的分类方法，有效地提高符号数据的分类性能，仍然有很大的应用价值。

为深入挖掘符号数据不同属性值与标签之间存在的关系，解决目前方法计算复杂度高、泛化性能不佳等问题，本文提出了一种基于组合关联分析的符号数据分类方法（categorical data classification approach based on combinatorial association analysis，CD_CAA）。在模型的训练阶段，该方法借助提升度将符号数据的属性值和类别标签关联起来，将符号数据转换成数值数据，且转换后的属性值恰好包含了部分类别标签信息，反映了同一属性对不同类的不同贡献，更好地完成了符号数据到数值数据的转化，以便训练分类模型。在测试阶段，每一真实的测试样本均结合所有可能的类标签进行组合关联分析，将其转化为多个虚拟测试样本，用分类模型预测每个虚拟测试样本的类别标签，以预测该样本的最终测试标签。

本文借助提升度进行关联分析，完成数据类型的转化，提高了模型的泛化性能；通过组合关联分析，将一个真实的测试样本转化为多个虚拟的测试样本，解决了模型测试问题；提出的方法充分考虑了样本属性和类别标签的关系，为符号数据的分类提供了一种新的思路。

1 相关工作

目前已经提出了许多对符号数据分类的方法，包括基于决策树归纳的 C4.5[15] 和基于概率分类的朴素贝叶斯（NB）[16]。从技术上讲，决策树是一个类似于流程图的结构，其中内部节点的每个分支都代表一个对属性的“测试”，因此它特别

1. 长治幼儿师范高等专科学校信息技术部 山西长治 046000

适合分类数据。然而，当数据集中包含大量属性时，该方法会遇到困难，因为此时生成的树将变得极其复杂。朴素贝叶斯（NB）基于这样的假设提出了一个简单的解决方案，即给定类别属性，预测属性是条件独立的。当应用于分类数据时，NB使用频率估计器或拉普拉斯修正来计算响应的后验概率。因此，如果数据集不够大，估计器通常会导致较大的估计方差，分类决策会存在一定的错误率。

基于距离的分类器因其固有的简单性和有效性而备受关注，包括KNN和基于原型的分类（prototype-based classification，PBC）。对于此类方法，通常必须解决两个关键问题：一是用于与测试样本进行距离比较的代表实例的选择，二是用于计算的距离度量。关于第一个问题，KNN方法选择k个最近邻进行类别预测，而在基于原型的分类器中，学习一个质心向量（即原型）来表示每个训练类别。基于距离的分类器的性能在很大程度上取决于距离度量，对于符号数据，常用的度量包括卡方距离或简单匹配系数（也称为重叠度量）[17-18]。从本质上讲，这些度量假设所有属性都同等重要，这在许多实际应用场景中很难成立。例如，在高维数据中，通常有许多对类别预测没有贡献的噪声属性。一些方法被提出以应对这一挑战，例如自适应距离度量学习[19-20]以及使用启发式加权方案的属性加权方法。然而，大多数工作仅专注于数值数据的分类任务。

本文提出的算法借助提升度将属性标签关联在一起，进行数值转化，这样不仅能避免上述算法的弊端，还能充分反映同一属性对不同类的不同贡献。另外，本文提出的方法对二分类和多分类任务均适用。

2 基于组合关联分析的符号数据分类方法

2.1 关联分析

关联分析是本文所提出方法的关键前期准备阶段，它旨在选定合适的关联指标，明确数据的表示方法，并加载所需的数据集。在符号数据分类任务中，关联分析不仅有助于人们理解数据内部的关联关系，更能通过特定的关联指标，将符号数据转化为数值数据，为后续的模型训练提供基础。

2.1.1 关联指标

关联规则的分析和度量通常依赖于一系列的指标，其中最为典型的是支持度、置信度和提升度。这些指标为研究者提供了不同的视角来观察和理解数据中的关联模式。

支持度：它反映了某项集在数据集中出现的频率，即数据集中包含该项集的记录所占的比例，见式（1）。它提供了关联规则出现的频率信息，但是可能会忽略支持度较低但具有实际意义的模式。

$$Support(X)=\frac{P(X)}{P(I)}=\frac{Number(X)}{Number(I)} \tag{1}$$

式中：I代表总数据集，$P(X)$表示项集X出现的次数，$Number(X)$表示求含有项集X的数据个数。

置信度：它衡量了一个项集出现时另一个项集也出现的概率，评估了关联规则的强度，见式（2）。然而，它有时会忽略规则后件中项集的支持度，可能导致对规则强度的误判。

$$Confidence(X\rightarrow Y)=P(Y\mid X)=\frac{P(X\cup Y)}{P(X)} \tag{2}$$

提升度：它综合考虑了支持度和置信度的信息，表示在含有某一项集X的条件下，含有另一项集Y的比例。它不仅反映了关联规则的强度，还能避免支持度和置信度单独使用时可能存在的问题。

$$Lift(X\rightarrow Y)=\frac{Confidence(X\rightarrow Y)}{Support(Y)} \tag{3}$$

使用支持度度量关联规则可以衡量规则出现的频率，但是许多潜在有意义的模式会由于含有支持度计数较小的项而被删去。使用置信度可以衡量规则的强度，但有时会忽略规则后件中项集的支持度，而且二者在不平衡数据集上表现均不佳。提升度将二者综合起来，可以避免以上弊端。因此，本文选择提升度作为主要的关联指标，以更全面地度量训练样本属性和类别标签的关联强度，从而将符号型属性数据转化为更为合理的数值型数据。

2.1.2 数据表示

本文记训练集为$T=\{(x_i,y_i)\}_{i=1}^{|T|}$，测试集为$S=\{(m_j,n_j)\}_{j=1}^{|S|}$，训练样本属性为$x_i^p,p=1,\cdots,d$，转化为数值属性表示为$\widetilde{x_i^p}$，测试样本属性为$m_j^p,p=1,\cdots,d$，转化为数值属性为$\widetilde{m_j^p}$，对第$p$个属性而言，所有可能的取值为$v_1^p,v_2^p,\cdots,v_{|v^p|}^p$，训练样本标签为$y_j$，测试样本标签为$n_j$。

在以上约定下，可得属性提升度的计算方式为：

$$Lift(v_i^p\rightarrow y)=\frac{P(v_i^p,y)}{P(v_i^p)\cdot P(y)} \tag{4}$$

它衡量了第p个属性的取值v_i^p与标签y之间的关联性。

2.1.3 数据集

表1是原始的符号数据集，如果是二分类问题，则它的标签为label1；如果是多分类问题，则它的标签是label2。

表1 原始符号数据集

a1 (Age)	a2 (Prescription)	a3 (Astigmatic)	a4 (Tear)	label1 (Class)	label2 (Class)
young	myope	no	reduced	T	not
young	myope	no	normal	F	soft
middle	myope	yes	reduced	T	not
middle	myope	yes	normal	F	hard
middle	hypermetrope	no	reduced	T	not
middle	hypermetrope	no	normal	F	soft

表 1（续）

a1 (Age)	a2 (Prescription)	a3 (Astigmatic)	a4 (Tear)	label1 (Class)	label2 (Class)
old	myope	yes	reduced	T	not
old	myope	yes	normal	F	hard
old	hypermetrope	yes	reduced	T	not

2.2 概述

为了挖掘符号数据不同属性值与标签之间存在的关系，本文提出了一种基于组合关联分析的符号数据分类方法（CD_CAA）。该方法通过提升度将符号数据的特征值与标签关联起来，转化为数值数据，以便训练分类模型。利用关联分析，符号数据在转换成数值数据前包含了部分标签信息，反映了同一属性对不同类的不同贡献，提高了泛化性能。

（1）在训练阶段

利用提升度将符号数据的特征值与标签关联起来，转化为数值数据，用→表示；使用转化后的数值数据训练分类器，用→表示。

（2）在预测阶段

将无标签的符号数据转化为多个虚拟带标签的符号数据，用→表示；将虚拟的符号数据转化为数值数据，用→表示；分类器根据数值数据预测标签，综合所有虚拟标签得到最终的预测标签，用→表示。

具体流程如图 1 所示。

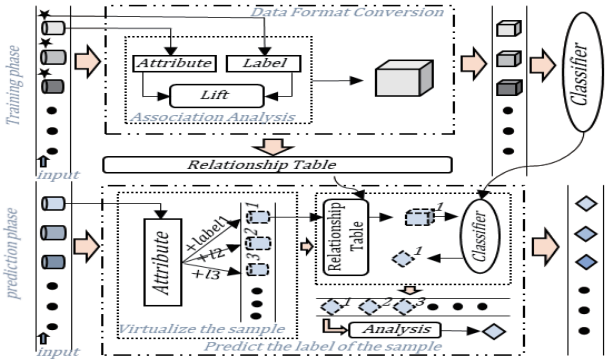

图 1 CD_CAA 框架图

本算法适用于二分类及多分任务，由于二者在处理方式上略有不同，所以 2.3 节针对二分类问题搭建模型，2.4 节针对多分类问题搭建模型。

2.3 二分类模型

2.1 节中提出的公式（4）可以将属性与标签进行关联分析，从而完成数据类型的转化，然而转化后的数据并没有体现同一属性对不同标签的不同贡献。为了解决上述问题，本文创造性地对公式（4）进行了改进，提出了新的属性相对提升度：

$$Lift(v_i^p \to y) = y \cdot \frac{P(v_i^p, y)}{P(v_i^p) \cdot P(y)} \tag{5}$$

该方法在原来的基础上增加了标签本身取值这个因素，对于二分类问题而言，标签的取值可为 {+1，-1}，这样同一个属性的同一个取值会因为标签差异而转化成不同的数值。

根据式（5），可以将表 1 中的符号数据转化成如表 2 所示的数值数据。

表 2 经过关联分析后的二分类数据集

a1 (Age)	a2 (Prescription)	a3 (Astigmatic)	a4 (Tear)	label1 (Class)
0.900	0.900	0.900	1.080	+1
-1.125	-1.125	-1.125	-2.250	-1
0.900	0.900	1.080	1.800	+1
-1.125	-1.125	-0.900	-2.250	-1
0.900	1.200	0.900	1.800	+1
-1.125	-0.750	-1.125	-2.250	-1
1.200	0.900	1.080	1.800	+1
-0.750	-1.125	-0.900	-2.250	-1
1.200	1.200	1.080	1.800	+1

在训练阶段，将用转化后的数值数据训练分类器。在预测阶段，将用分类器对预测样本进行标签预测。分类器是在数值数据上训练出来的，而预测样本只有属性没有标签，无法直接通过关联分析转化成数值数据，所以无法直接进行预测。针对这个问题，本文采用了“虚拟化样本”的方法来转化数据，即结合所有可能的标签 {+1，-1} 将一个实际样本m_j转化为两个不同的虚拟样本m_j^+和m_j^-。

根据表 2 将虚拟的符号数据样本m_j^+和m_j^-转化成数值数据$\widetilde{m_j^+}$和$\widetilde{m_j^-}$，并对它们进行分类，得到各自的预测标签$\widetilde{n_j^+}$和$\widetilde{n_j^-}$。

假设当前分类面为$f = w \cdot x + b$，则预测标签的确定方法如下。

定义 1 *distance* 表示虚拟化样本到超平面的距离。

$$distance\left(\widetilde{m_j^+}, f\right) = \widetilde{n_j^+} \cdot \frac{w \cdot \widetilde{m_j^+} + b}{\|w\|}$$
$$distance\left(\widetilde{m_j^-}, f\right) = \widetilde{n_j^-} \cdot \frac{w \cdot \widetilde{m_j^-} + b}{\|w\|} \tag{6}$$

定义 2 *unity* 表示预测标签的统一性。

$$unity\left(\widetilde{n_j^+}, \widetilde{n_j^-}\right) = \widetilde{n_j^+} \cdot \widetilde{n_j^-} \tag{7}$$

unity=1 表示不论预测标签是什么，分类器总是倾向于将样本分为特定的一类。

定义 3 *max* 表示 *distance* 大的样本的预测标签。

$$max = \begin{cases} +1, distance\left(\widetilde{m_j^+}, f\right) \geq distance\left(\widetilde{m_j^-}, f\right) \\ -1, distance\left(\widetilde{m_j^+}, f\right) < distance\left(\widetilde{m_j^-}, f\right) \end{cases} \tag{8}$$

最终预测标签 n_j^* 表示如下：

$$n_j^* = unity\left(\widetilde{n_j^+}, \widetilde{n_j^-}\right)\cdot \widetilde{n_j^+} + \left(1 - unity(\widetilde{n_j^+}, \widetilde{n_j^-})\right)\cdot max \quad (9)$$

下面展示了基于关联分析的符号数据二分类算法：

输入：训练集 $T = \{(x_i, y_i)\}_{i=1}^{|T|}$，测试集 $S = \{(m_j, n_j)\}_{j=1}^{|S|}$，训练样本属性 $x_i^p, p = 1, \cdots, d$，测试样本属性 $m_j^p, p = 1, \cdots, d$，训练样本标签为 y_j，测试样本标签为 n_j；

输出：测试集的精度 T_s

1. 根据式（5）将符号型训练样本转化为数值型样本；

2. 在数值型的训练集上训练构建分类器 f；

3. 构造虚拟的测试样本集；

4. 根据式（9）得到测试样本的标签；

5. 根据测试样本标签计算测试精度。

2.4 多分类模型的搭建

现实生活中的事物往往被分为许多类，因此需要在二分类的基础上进行拓展。处理多类任务的方法有多种，通常选择将多分类任务进行拆分，对拆分出的每一个二分类任务训练一个分类器，集成多个分类器的结果得到最终分类结果。拆分的方式有三种：OVO、OVR、MVM。

OVO 是将 N 个类别两两配对，得到 $N(N\text{-}1)/2$ 个二分类任务，从而得到 $N(N\text{-}1)/2$ 个分类器。

OVR 是将 N 个类别中的一个作为正类，其余作为负类，得到 N 个分类任务，从而得到 N 个分类器。

MVM 是将 N 个类别中的一部分作为正类，一部分作为负类，进行 M 次划分，得到 M 个分类任务，从而得到 M 个分类器。

OVR 在类别多时需要的训练开销较大，MVM 类别的选取必须有特殊的设计。综合考虑，本文采用 OVO 的方式处理任务。根据 OVO 方式以及公式 5，可以将多分类任务的符号数据转化成对应的二分类任务下的数值数据，例如对表 1 进行改进得到表 3、表 4、表 5，这里定义标签数量少的为正类。

表 3 转化后的二分类子问题（not 和 soft 类）

a1 (Age)	a2 (Prescription)	a3 (Astigmatic)	a4 (Tear)	label1 (Class)
-0.700	-1.050	-0.700	-1.400	-1(not)
1.750	0.875	1.750	3.500	+1(soft)
-0.933	-1.050	-1.400	-1.400	-1(not)
-0.933	-1.400	-0.700	-1.400	-1(not)
1.167	1.167	1.750	3.500	+1(soft)
-1.400	-1.050	-1.400	-1.400	-1(not)
-1.400	-1.400	-1.400	-1.400	-1(not)

表 4 转化后的二分类子问题（not 和 hard 类）

a1 (Age)	a2 (Prescription)	a3 (Astigmatic)	a4 (Tear)	label1 (Class)
-1.400	-0.840	-1.400	-1.400	-1(not)
-0.933	-0.840	-0.840	-1.400	-1(not)
1.167	0.560	0.560	3.500	+1(hard)
-0.933	-1.400	-1.400	-1.400	-1(not)
-0.933	-0.840	-0.840	-1.400	-1(not)
1.167	0.560	0.560	3.500	+1(hard)
-0.933	-1.400	-0.840	-1.400	-1(not)

表 5 转化后的二分类子问题（hard 和 soft 类）

a1 (Age)	a2 (Prescription)	a3 (Astigmatic)	a4 (Tear)	label1 (Class)
-1.400	-1.167	-3.500	-1.750	-1(soft)
1.750	2.333	3.500	1.750	+1(hard)
-1.750	-3.500	-3.500	-1.750	-1(soft)
3.500	2.333	3.500	1.750	+1(hard)

在训练阶段，对每一组数值数据进行训练，最终得到 $N(N\text{-}1)/2$ 个分类器（如上例为 3 个）。

在预测阶段，仍然使用虚拟化样本方法：不同于二分类问题，多分类问题中的标签为 {1, 2, ⋯, c}，结合所有可能的标签，一个实际的测试样本 m_j 将会转化为多个虚拟测试样本 $\widetilde{m_j^1}$，$\widetilde{m_j^2}$，⋯，$\widetilde{m_j^c}$。

根据之前得到的关联分析表（即表 3、表 4、表 5），可以将这 c 个虚拟的符号数据样本转化为数值数据样本，将每一个数据样本放入 c-1 个分类器中进行训练，并统计这 c-1 个分类器中对样本打结合的标签的数量来作为这 c-1 个分类器的综合预测结果。综合所有虚拟数值数据样本的预测结果，得到最后的预测结果。

可以把虚拟样本 $\widetilde{m_j^c}$ 的预测结果记为 $\widetilde{n_j^c}$，第 j 个样本的最终预测结果记为 n_j^*，则确定最终标签的框架如下。

定义 1 表示标签 c 作为负类时，打出 -1 的分类器的个数。

$$\overline{n_j^c} = Count\left(\left\{f_{i,c} \mid c == f_{i,c}\left(-\widetilde{m_j^c}\right) and\ i < c\right\}\right) \quad (10)$$

定义 2 表示标签 c 作为正类时，打出 +1 的分类器的个数。

$$\underline{n_j^s} = Count\left(\left\{f_{s,k} \mid s == f_{s,k}\left(-m_j^s\right) and\ k > s\right\}\right) \quad (11)$$

最后的预测标签的计算如下：

$$n_j^* = arg \max_{s\in\{1,2,\cdots,c\}}\left(\overline{n_j^s} + \underline{n_j^s}\right) \quad (12)$$

本文将多分类问题采用 OVR 的思想转化为多个二分类的子问题，对于每个二分类的子问题，训练和测试方式同算法 1 所述。

3 实验与性能分析

为验证所提算法在符号数据分类方向的性能，本文分别在 10 个标准数据集上进行实验。实验平台为 Win-

dows10，CPU 为 Ryzen 5 4600U，内存为 24 G，语言为 Matlab2020b。实验中，将本文所提出的算法 CD_CAA 与 C4.5、KNN、NBC 进行对比来验证。SVM 在小规模样本上具有良好的泛化性能，因此本文选择 SVM 作为基础分类器来构建模型。

3.1 实验框架

实验中对于每一个数据集的处理过程如图 2 所示。首先，数据集 D 被随机地划分成训练集 D_A 和测试集 D_B 两部分。这两部分的规模比为 4:1，即训练集占原始数据集的 80%，训练集占原始数据集的 20%。划分后的 D_A 将被用来训练分类模型，由于本文提出的算法需要与其他分类算法进行比较，所以会训练 KNN 分类模型 M_1，NBC 分类模型 M_2，C4.5 分类模型 M_3，SVM 分类模型 M_4，最后这些模型将在 D_B 上进行测试。

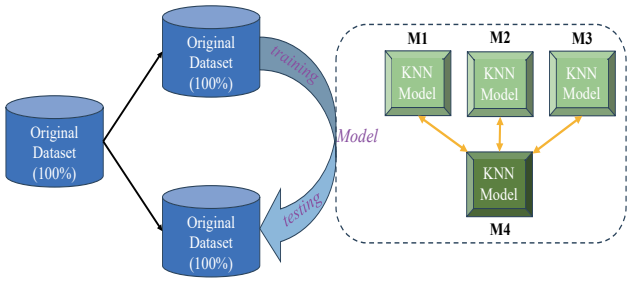

图 2 实验设计图

3.2 实验设置

数据集设置：本文采用 UCI 数据集，UCI 是用于机器学习的标准数据库。为验证所提出算法的性能，本文在 10 个标准数据集上进行实验，具体见表 6。

参数设置：为了能够训练出较好的模型，本文采用不同的参数进行训练。C4.5 中参数可选 {3、5、7}，KNN 中参数可选 k={3、5、7}，NBC 没有参数，SVM 中参数可选 g={ 默认值、1、2、5}，c={0.01、0.1、1、10、100、1000}。

表 6 实验数据集

数据集	属性维数	样本类别数	样本数量
breast_cancer	9	2	683
chess	36	2	3196
promote	57	2	106
spect	21	2	267
Tic-Toc-Toe	9	2	958
vote	16	2	435
balance123	4	3	625
HR	4	3	132
splice	60	3	3190
car	6	4	1728

测评指标：

①类别准确率 Acc（Accuracy），定义如下：

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{13}$$

式中：TP（true positive）表示将正类预测为正类的数目，TN（true negative）表示将负类预测为负类的数目，FP（false positive）表示将负类预测为正类的数目，FN（false negative）表示将正类预测为负类的数目。

②标准差 STDEV（Standard Deviation），在概率统计中最常使用作为统计分布程度（statistical dispersion）上的测量。它是总体各单位标准值与其平均数离差平方的算术平均数的平方根，反映组内个体间的离散程度。

3.3 实验结果与分析

为有效衡量本文所提出的 CD_CAA 算法在符号数据分类方面的性能，本文将该算法与 KNN、C4.5、NBC 进行比较，由于 KNN、C4.5、SVM 有多个参数，所以先进行参数选择，将最佳模型进行比较。

3.3.1 最佳 C4.5 模型

实验采用了 Acc 指标来评测分类性能，每个数据集进行五次交叉验证。图 3 展示了在 C4.5 上进行分类得到的 Acc 值，表 7 展示了 Acc 的平均值以及排名。从实验结果可看出，大多数情况下参数为 3 时准确率较高，并且综合结果显示参数为 3 时平均 Acc 值较高，方差较小，较稳定。故选最佳参数为 3。

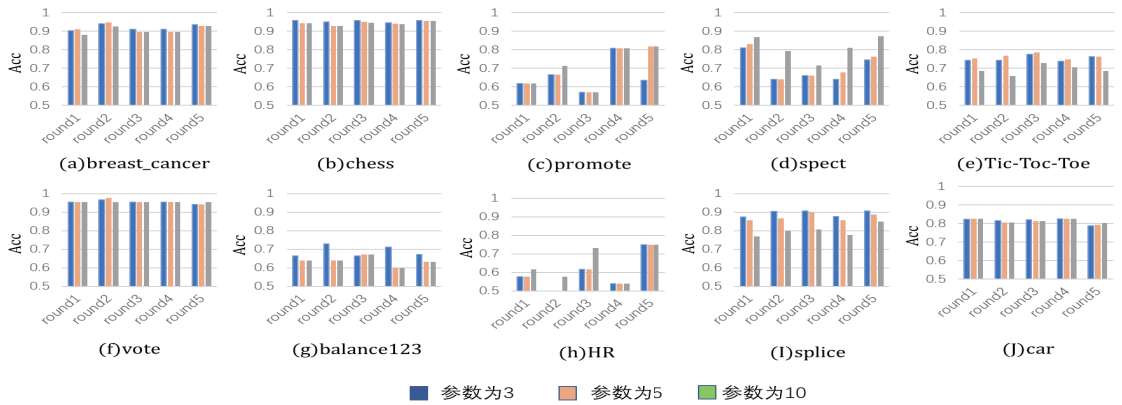

图 3 C4.5 作为基准模型的累积 Acc 值

表 7 C.5 作为基准模型的测试精度及排名

Date set	Parameter						
	Option	3		5		10	
		figures	rank	figures	rank	figures	rank
breast_cancer	Average Acc	0.921	1	0.916	2	0.906	3
	STDEV	0.016	1	0.022	3	0.020	2
chess	Average Acc	0.955	1	0.943	2	0.942	3
	STDEV	0.005	1	0.011	3	0.010	2
promote	Average Acc	0.661	3	0.697	2	0.706	1
	STDEV	0.090	1	0.112	3	0.111	2
spect	Average Acc	0.700	3	0.715	2	0.812	1
	STDEV	0.076	2	0.080	3	0.064	1
Tic-Toc-Toe	Average Acc	0.753	2	0.761	1	0.693	3
	STDEV	0.016	2	0.014	1	0.026	3
vote	Average Acc	0.954	2	0.956	1	0.954	2
	STDEV	0.008	1	0.012	2	0.000	1
balance123	Average Acc	0.688	1	0.634	2	0.634	2
	STDEV	0.023	1	0.026	2	0.026	2
HR	Average Acc	0.596	2	0.596	2	0.642	1
	STDEV	0.096	2	0.096	2	0.093	1
splice	Average Acc	0.893	1	0.873	2	0.799	3
	STDEV	0.016	1	0.019	2	0.032	3
car	Average Acc	0.815	1	0.813	3	0.814	2
	STDEV	0.016	3	0.014	2	0.011	1
Average rank	Average Acc	1.7		1.9		2.1	
	STDEV	1.5		2.3		1.8	

3.3.2 最佳 KNN 模型

同 C4.5 的选取方式一致，图 4 展示了在 KNN 上进行分类得到的 Acc 值，表 8 展示了 Acc 的平均值以及排名。从实验结果可看出，大多数情况下参数为 7 时准确率较高，并且综合结果显示参数为 7 时平均结果较高，虽然参数为 7 时的方差不是最小的，但是各种参数下的方差均相近，所以最终选定最佳参数为 k=7。

表 8 KNN 作为基准模型的测试精度及排名

Date set	Parameter						
	Option	3		5		10	
		figures	rank	figures	rank	figures	rank
breast_ cancer	Average Acc	0.956	3	0.958	2	0.959	1
	STDEV	0.019	1	0.020	2	0.023	3
chess	Average Acc	0.955	3	0.959	1	0.957	2
	STDEV	0.011	3	0.007	2	0.005	1
promote	Average Acc	0.821	2	0.820	1	0.821	2
	STDEV	0.062	3	0.053	1	0.059	2
spect	Average Acc	0.745	3	0.831	1	0.824	2
	STDEV	0.057	3	0.038	1	0.049	2
Tic-Toc-Toe	Average Acc	0.875	3	0.924	2	0.947	1
	STDEV	0.023	1	0.026	2	0.029	3
vote	Average Acc	0.924	3	0.926	2	0.931	1
	STDEV	0.026	1	0.026	1	0.028	2
balance123	Average Acc	0.709	3	0.731	2	0.798	1
	STDEV	0.027	1	0.036	2	0.036	2
HR	Average Acc	0.629	1	0.581	3	0.620	2
	STDEV	0.170	3	0.149	2	0.108	1
splice	Average Acc	0.769	3	0.774	2	0.798	1
	STDEV	0.017	2	0.009	1	0.017	2
car	Average Acc	0.843	3	0.860	2	0.878	1
	STDEV	0.011	1	0.012	2	0.011	1
Average rank	Average Acc	2.4		1.8		1.4	
	STDEV	1.9		1.6		1.9	

3.3.3 最佳 SVM 模型

图 5、图 6、图 7、图 8 分别为展示了当 g= 默认值、g=1、g=2、g=5 时的 SVM 上的 Acc 值，图 9 是所有参数组合下的平均 Acc 值，图 10 是所有参数组合下的标准差。可以看出，在固定 g= 默认值的时候，c=0.1 在个别数据集上表现最佳，在大多数数据集上表现良好；固定 g=1、g=2、g=5 时，c=1、c=10、c=100、c=1000 的效果相一致且比 c=0.1、c=0.01 时的效果好一些。从平均 Acc 值上看，固定 g=1、g=2、g=5 时，在部分数据集上 c=1 时平均 Acc 值比 c=10、c=100、c=1000 略低一些，而 c=10、c=100、c=1000 的效果完全一致，故 g=1、g=2、g=5 时选定一个参数 c=100 即可。

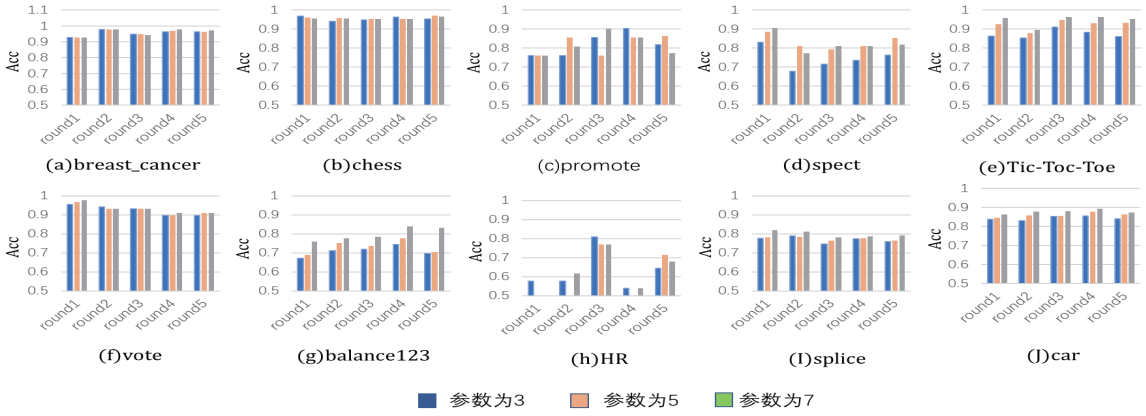

图 4 KNN 作为基准模型的累积 Acc 值

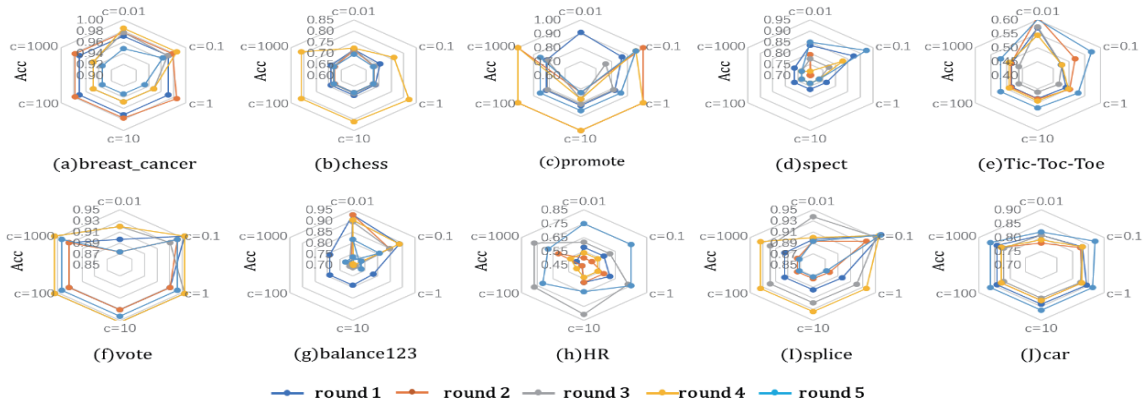

图 5 SVM 作为基准模型的测试精度 (g = 默认值)

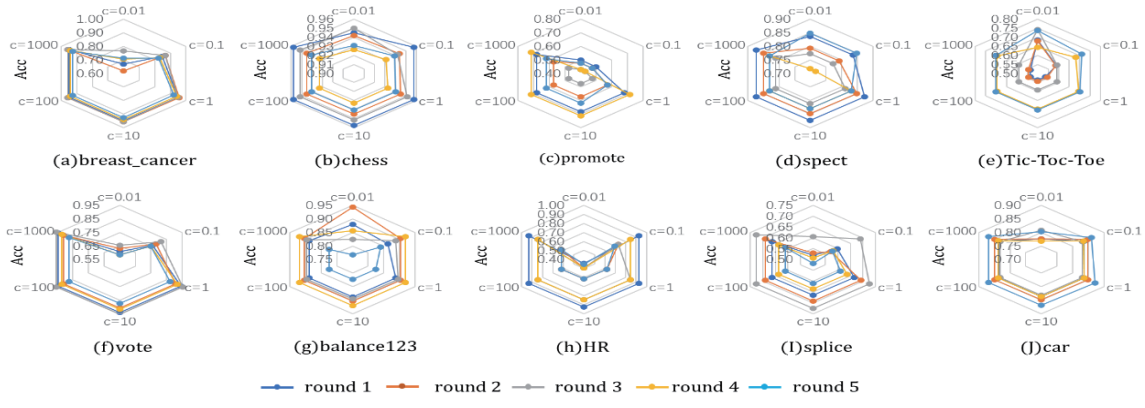

图 6 SVM 作为基准模型的测试精度 (g=1)

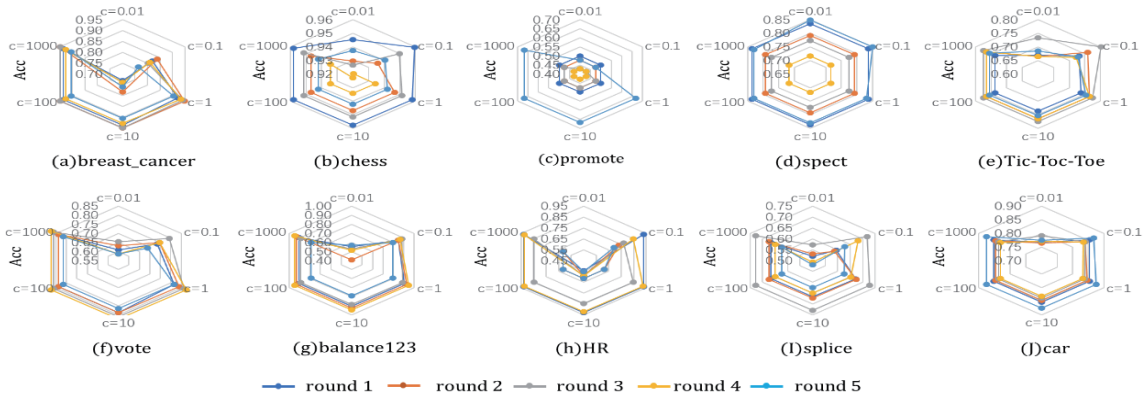

图 7 SVM 作为基准模型的测试精度 (g=2)

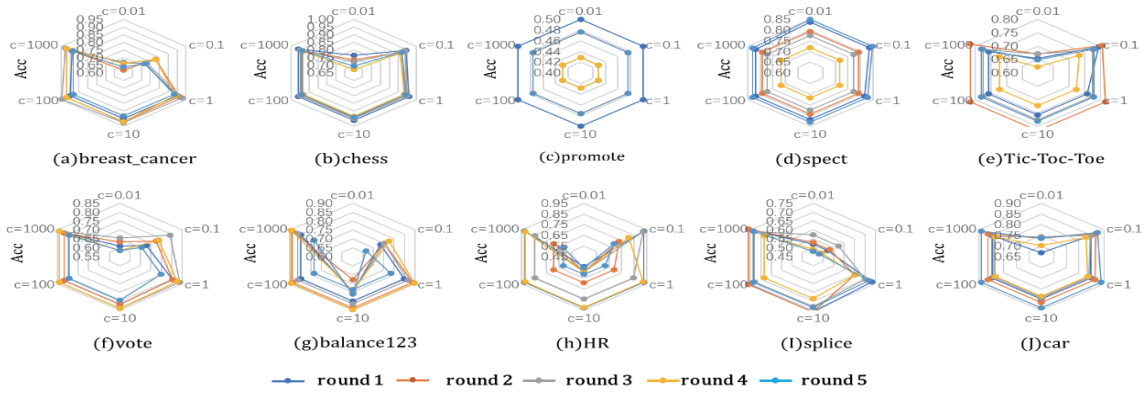

图 8 SVM 作为基准模型的测试精度 (g=5)

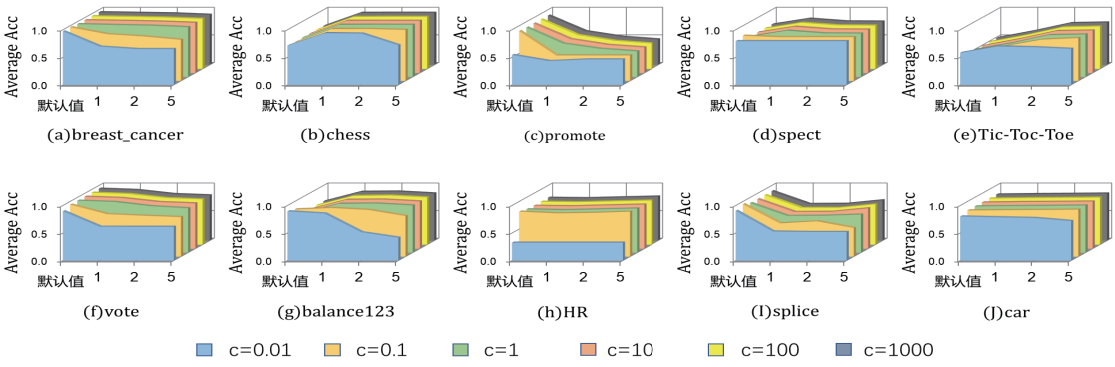

图 9 SVM 作为基准模型不同参数下的平均测试精度

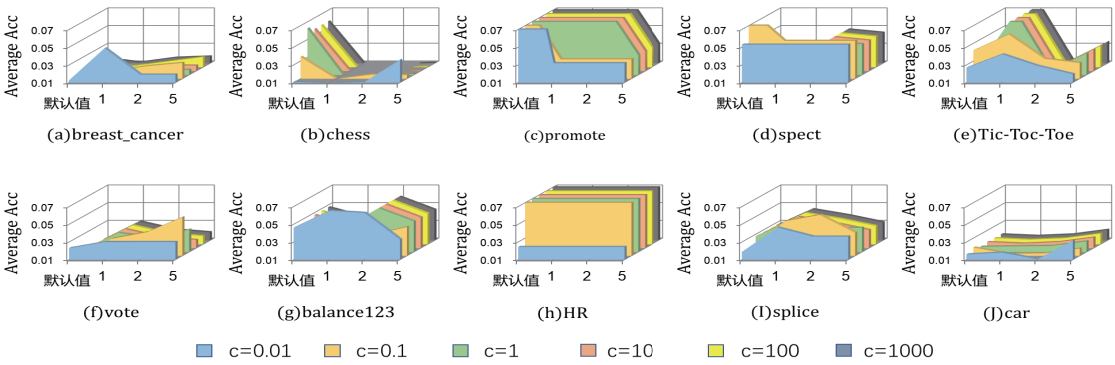

图 10 SVM 作为基准模型不同参数下的测试方差

从标准差上来看，g=1 的标准差总是比 g 的其余取值略大些，故最终选定 3 组参数组合 g= 默认值，c=0.1、g=2，c=100、g=5，c=100 进行下一步比较。

表 9 展示了上一步筛选出的 3 组参数组合在 10 个数据集的平均 Acc 值的平均值及方差。可以看出，g= 默认值、c=0.1 时的平均准确率较高，方差较其余两组参数较高，说明其在极个别数据集上表现效果差异较大，但是整体来看，g= 默认值、c=0.1 能取得较好效果。

表 9 不同参数组合的平均 Acc 值（SVM）

Data set	Parameter		
	g= 默认值 c=0.1	g=2 c=100	g=5 c=100
	Average Acc	Average Acc	Average Acc
breast_cancer	0.973 6	0.932 6	0.909 3
chess	0.701 5	0.946 5	0.945 9
promote	0.906 1	0.509 5	0.461 9
spect	0.819 8	0.789 9	0.786 3
Tic-Toc-Toe	0.503 2	0.758 9	0.770 4
vote	0.944 8	0.846 0	0.820 7
balance123	0.856 0	0.891 2	0.856 0
HR	0.816 5	0.785 7	0.801 1
splice	0.952 0	0.668 7	0.733 9
car	0.838 0	0.849 0	0.855 3
Average	0.831 2	0.797 8	0.794 1
STDEV	0.133 7	0.124 7	0.126 2

3.3.4 算法比较

图 11 展示了各个数据集下每种模型的类别准确率。可以看出，在二分类问题上，本文所提出的算法整体上与传统算法的效果一致。当涉及多分类问题时，本文所提出的算法的优势便凸显了出来，基本上一直比传统算法的准确率要高。可见，本文所提出的算法在二分类问题上可以与传统算法媲美，在多分类问题上性能要高于传统算法。综合来看，本文所提出的算法能够更好地处理符号数据分类任务。

4 结语

本文介绍了一种名为 CD_CAA 的基于组合关联分析的符号数据分类方法，通过引入提升度，使得研究者能够更全面地评估关联规则的强度和有效性，从而更准确地捕捉数据中的关联关系。通过将符号数据转化为数值数据，研究者能够充分利用现有的机器学习方法进行模型训练，并进一步提高分类准确率。此外，CD_CAA 方法还展现出较强的泛化能力，能够适应不同数据集和分类任务的需求。通过实验验证了其在提升分类性能方面的有效性。总体而言，CD_CAA 方法为符号数据分类问题提供了一种新的解决方案，并通过实验证明了其有效性。该方法不仅提高了分类准确率，还为解决符号数据分类任务提供了新的思路和方法。未来的研究将继续探索更多有效的符号数据处理方法，并将 CD_CAA 方法推广应用于更广泛的领域，以推动数据挖掘和机器学习领域的发展。

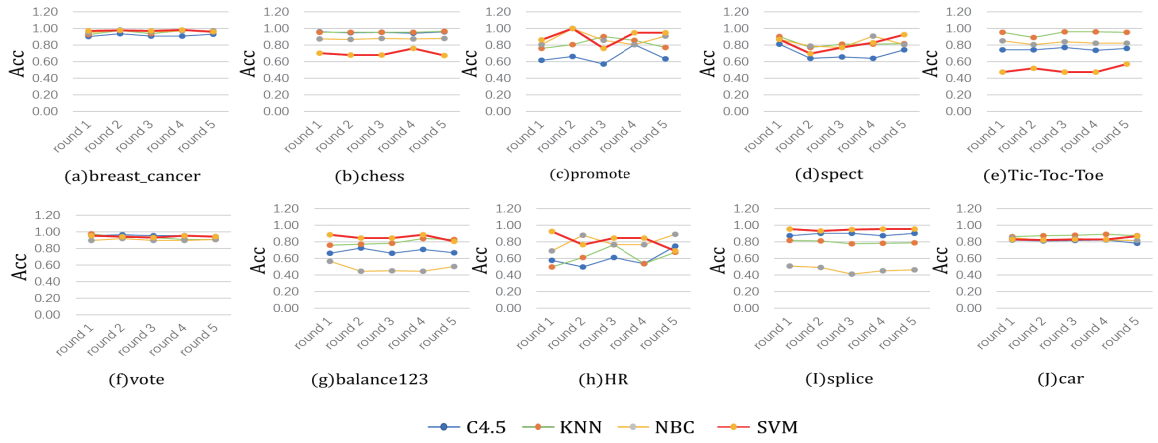

图 11 不同模型测试结果比较

参考文献：

[1]DUDA R O, HART P E.Pattern classification and scene analysis[J].IEEE transactions on automatic control, 2003, 19(4): 462-463.

[2]CRISTIANINI N, SCHOLKOPF B.Support vector machines and kernel methods:the new generation of learning machines[J]. AI magazine,2002,23(3):31-38+40-41.

[3]COVER T, HART P.Nearest neighbor pattern classification[J]. IEEE transactions on information theory,1967,13(1):21-27.

[4]GUO G, WANG H, BELL D, et al.Using KNN model for automatic text categorization[J].Soft computing, 2006,10:423-430.

[5]WU Y, FENG J.Development and application of artificial neural network[J].Wireless personal communications, 2018, 102(2):1645-1656.

[6]HAN J, KAMBER M, PEI J.Data mining:concepts and techniques[M].San Francisco,CA,United States:Morgan Kaufmann Publishers, 2006.

[7] 周志华 . 机器学习 [M]. 北京 : 机械工业出版社 , 2021.

[8]FU K, WANG W, GUO H.Categorical Data classification method based on spatial correlation analysis[J].Computer science and exploration,2019,13(7):1165-1173.

[9]AKAY M F.Support vector machines combined with feature selection for breast cancer diagnosis[J].Expert systems with applications, 2009,36(2):3240-3247.

[10]QUINLAN J R.Induction of decision trees[J].Machine learning, 1986,1(1):81-106.

[11]LEWIS D D.Naive(bayes) at forty:the independence assumption in information retrieval[C]//Machine learning: ECML-98. Berlin:Springer-Verlag,1998:4-15.

[12]HAN E, KARYPIS G.Centroid-based document classification:analysis and experimental results[C]//Principles of Data Mining and Knowledge Discovery.Berlin:Springer, 2000:424-431.

[13]ZHANG J, CHEN L, GUO G.Projected-prototype based classifier for text categorization[J]. Knowledge-based systems, 2013,49:179-189.

[14]BORIAH S, CHANDOLA V, KUMAR V.Similarity measures for categorical data:a comparative evaluation[C]//8th SIAM International Conference on Data Mining.[s.l]:Society for Industrial and Applied Mathmatics(SIAM),2008:334-345.

[15]HALL M, FRANK E, HOLMES G, et al.The WEKA data mining software:an update[J].SIGKDD explorations, 2009, 11(1): 10-18.

[16]SEEGER M.Bayesian modelling in machine learning:a tutorial review[EB/OL].(2006-01-01)[2024-02-20].https://api.semanticscholar. org/CorpusID:3236390.

[17]JIANG L, CAI Z, WANG D, et al.Bayesian citation-KNN with distance weighting[J].International journal of machine learning and cybernetics,2014,5(2):193-199.

[18]SEN P K.Gini diversity index,hamming distance,and curse of dimensionality[J].Metron-international journal of statistics, 2005, 63(3):329-349.

[19]XIONG T, WANG S, MAYERS A, et al.DHCC:divisive hierarchical clustering of categorical data[J].Data mining and knowledge discovery, 2012,24(1):103-135.

[20]PAREDES R, VIDAL E.Learning weighted metrics to minimize nearest-neighbor classification error[J].IEEE transactions on pattern analysis and machine intelligence, 2006, 28(7):1100-1110.

【作者简介】

崔丽娜（1981—），女，山西长治人，硕士，讲师，研究方向：数据挖掘、智能信息处理。

（收稿日期：2024-04-06）