面向序数回归的组合特征提取方法

曾庆田[1,2] 刘晨征[1] 倪维健[1] 段 华[3]
(山东科技大学计算机科学与工程学院 山东 青岛 266590)[1]
(山东科技大学电子信息工程学院 山东 青岛 266590)[2]
(山东科技大学数学与系统科学学院 山东 青岛 266590)[3]

摘 要 序数回归(也称序数分类)是一种监督学习任务,即使用具有自然顺序的标签对数据项进行分类。序数回归与诸多实际问题密切相关,近几年关于序数回归的研究受到越来越多的关注。序数回归与其他监督学习任务(分类、回归等)一样,需要通过特征提取来提高模型的效率和准确性。虽然特征提取被广泛研究并用于分类学习任务中,但是在序数回归中的研究较少。众所周知,相比单特征,组合特征可以表达更多的数据底层语义,但是加入一般的组合特征很难提高模型的准确性。文中基于频繁模式挖掘,借助K-L散度值来选取最有区分能力的频繁模式进行特征组合,提出了一种新的序数回归组合特征提取方法,并在公开数据集和自有数据集上使用多个序数回归模型进行实验。结果表明,使用最有区分能力的频繁模式组合特征,能够有效提升大多数序数回归模型的训练效果。

关键词 序数回归,频繁模式,特征组合,特征选择
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.009

Combined Feature Extraction Method for Ordinal Regression

ZENG Qing-tian[1,2] LIU Chen-zheng[1] NI Wei-jian[1] DUAN Hua[3]
(College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)[1]
(College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)[2]
(College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)[3]

Abstract Ordinal regression, also known as ordinal classification, is a supervised learning task that uses the labels with a natural order to classify data items. Ordinal regression is closely related to many practical problems. In recent years, the research on ordinal regression has attracted more and more attention. Ordinal regression, like other supervised learning tasks (classification, regression, etc.), requires feature extraction to improve the efficiency and accuracy of the model. However, while feature extraction has been extensively studied for other classification tasks, there are few researches in ordinal regression. It is well known that the combined features could capture more underlying data semantics than single features, but it is difficult to improve the accuracy of the model by adding general combined features. Based on the frequent mining patterns, this paper used the K-L divergence value to select the most discriminative frequent patterns for feature combination, and proposed a new ordinal regression combination feature extraction method. Multiple ordinal regression models are used for validation on both the public and our own datasets. The experimental results show that using the most distinguishing frequent pattern combination features can effectively improve the training effect of most ordinal regression models.

Keywords Ordinal regression, Frequent pattern, Feature combination, Feature selection

1 引言

近年来,关于序数回归的研究受到越来越广泛的关注[1]。序数回归(也称序数分类)是一种用于预测序数变量的回归分析,其中目标变量标签呈现出自然排序。它也可以被认为是介于回归和分类之间的一类问题。例如,在信贷客户信用评估中,客户的信用等级可以分为{优,良,中,差};在医学研究方面,病人疾病发展阶段可以分为{无病,早期,中期,晚期}。

到稿日期:2018-06-20 返修日期:2018-08-26 本文受国家自然科学基金(61472229,61702306,61602278,61602279),山东省科技发展项目(2016ZDJS02A11,ZR2017BF015,ZR2017MF027),山东省泰山学者攀登计划专项和山东科技大学科研创新团队支持计划项目基金(2015TDJH102)资助。

曾庆田(1976-),男,教授,博士生导师,CCF会员,主要研究方向为过程挖掘、智能信息处理、个性化推荐、人工智能,E-mail:qtzeng@163.com;刘晨征(1994-),男,硕士生,主要研究方向为数据挖掘、人工智能;倪维健(1981-),男,博士,副教授,主要研究方向为机器学习、数据挖掘、信息检索,E-mail:niweijian@gmail.com(通信作者);段 华(1976-),女,博士,副教授,主要研究方向为机器学习、优化算法。

序数回归不同于标准回归和分类：1)序数回归中目标变量值是有限的，且不同等级之间的度量距离也不尽相同，而标准回归中目标变量是连续值；2)序数回归的目标变量之间是有序的，而分类问题中目标标签没有序数关系。

目前关于序数回归的研究主要集中在学习算法方面，按照解决思路的不同，序数回归模型可以分为以下 3 类：1)通过一些简单的假设，直接把序数回归当作多分类或回归问题来求解，如使用面向回归的支持向量机[2]和面向多分类的支持向量机[3]等方法来解决序数回归问题。这类方法不考虑标签之间的序数关系。2)将序数回归分解为多个二分类子问题，通过训练单个或多个模型解决序数回归问题，如 Cheng 等提出了使用神经网络的方法解决序数回归问题[4]，Deng 等将极限学习机算法应用在序数回归问题中[5]等。这类方法把标签的序关系融入到普通多分类模型中，保留了原始标签的序数信息。3)基于阈值的模型，寻找函数 $f(x)$ 把输入 n 维空间样本 x 映射为一维实数，求解一组阈值，把 $f(x)$ 分成几个区间，每个区间对应一个序数等级，如 Mccullagh 提出的 POM 算法[6]，Mathieson 使用神经网络实现的 POM 的非线性方法[7]，以及 Wei 等提出的面向序数回归的支持向量机[8]等。

特征是模型学习的基础，特征选取的好坏直接影响模型训练的结果。虽然特征提取方法被广泛应用到分类问题中，但是在序数回归中的研究很少。目前关于序数回归中的特征研究主要集中在特征选择——从原始特征中抽取一部分特征用于训练模型，如 Mukras 等提出的概率再分配过程(PRP)[9]，Baccianella 等提出最小方差法和循环最小方差法[10]，随后 Baccianella 等对这两种方法进行改进，又提出了 6 种方法来解决文本序数回归中的特征选择问题[11-12]。但是这些方法都是从原始数据中选择部分特征用于模型训练，并没有生成新的有用特征，而本文的主要研究工作是从原始数据中提取新的有效组合特征，用来提升模型的训练效果。

相比单个特征，组合特征可以表达更多的底层含义，某些特征经过组合之后，与预测标签之间的相关性将会提高。例如，在信贷客户信用评估中，将年龄与收入进行关联可以得到“青年收入低”与“中年收入低”两种新的特征。由生活常识可知，一个低收入的中年人的违约风险要大于一个低收入的青年人的违约风险。因为青年人刚步入社会，收入低是一种正常现象，而中年人收入低则反映了一种不正常现象。通过特征组合可以挖掘这种隐含信息。

在给定的数据集合 D 中有 n 个单特征，可以列举全部的组合特征(2^n 个)，并把其用于模型训练。但是这样存在两个明显的问题：1)组合特征的数量与单特征的数量呈指数关系，当单特征的数量较多时，很难生成全部的组合特征；2)生成的组合特征中大多数都不具有代表性，使用这些特征会导致模型的准确性下降，此外还会降低模型的学习效率。因此，在生成组合特征时，必须采用合理的策略，选择有用的组合特征，这样才能提升模型学习的效率和准确性。

自 Agrawal 提出频繁模式挖掘[13]后，相关研究人员提出了许多可扩展的方法来挖掘频繁模式，如 Apriori 算法[14]、FP-growth 算法[15]、CHARM 算法[16]等。此外，频繁模式在关联规则挖掘、分类和聚类等领域有着广泛的应用[17-19]。频繁模式反映不同属性之间的强关联关系，并且具有数据的可解释性。研究人员使用频繁模式与类之间的强关联关系，提出了关联分类。在关联分类中，分类器基于高支持度和高置信度的关联规则构建，利用频繁模式与类之间的关联规则进行分类[20-22]。此外，频繁模式也是进行特征组合的一种方法。因为频繁模式在数据集合中是频繁的，具有统计意义，所以使用频繁模式能够有效提升模型训练的效果。

在目前序数回归的研究中，还未出现有关组合特征提取的研究。序数回归与分类不同，不同类别之间是有序的。在处理序数回归特征提取时，需要考虑类别之间的序数关系。例如在信贷客户评估中，一个信用等级为“差”的客户被评为“良”级与被评为“优”级，这两种评估与实际都是不符合的，但是前者要比后者更能让人接受。因此，在提取序数回归中的组合特征时需要考虑不同类别之间的序数关系。

本文提出一种序数回归中组合特征提取的方法，主要工作包括以下几点。1)把序数回归有序分解为多个二元子问题，在每个二元子问题上挖掘带有类别属性的频繁模式；2)根据每个频繁模式的 K-L 散度值，在每个二元子问题上使用循环选择方法，选择最具有区分能力的频繁模式；3)基于选择的频繁模式进行特征组合，并在多个数据集和多个模型上验证方法的有效性。本文所提方法解决了序数回归中的频繁模式挖掘，以及挖掘后如何选择具有区分能力的频繁模式进行特征组合的问题；此外，提出一种循环选择频繁模式的方法，平衡选择区分不同等级的频繁模式。实验结果表明，所提方法能够有效提高序数回归模型的训练效果。

2 基本概念

本节将详细阐述对序数回归和特征组合的基本概念。

2.1 序数回归

给定训练数据集合 $D=\{(\boldsymbol{x}_i,y_i),\ i=1,\cdots,N\}$，序数回归的学习任务是找到一个函数 $f:X\rightarrow Y$ 来预测新输入模式的等级。其中，$\boldsymbol{x}\in X\subseteq\mathbb{R}^m$，$y\in Y=\{C_1,C_2,\cdots,C_q\}$，即 $\boldsymbol{x}$ 在 m 维输入空间中，y 在 q 维输出空间中。序数回归标签之间具有自然序数关系，即 $C_1\prec C_2\prec\cdots\prec C_q$，其中 $\prec$ 表示序数关系。在使用序数回归算法进行训练时，需要把序数回归标签转化为序数值，如使用函数 $O(C_r)=r(r=1,\cdots,q)$。但需要注意，序数回归标签与标准回归标签的含义是不同的。在标准回归中标签 $y\in\mathbb{R}$ 是定量属性，可以比较大小；而序数回归标签 $y\in Y$，是定性属性，标签之间只有序数关系，不可以比较大小。此外，序数回归标签与分类标签也不同，在分类中各标签之间相互独立，没有序关系。

2.2 特征组合

设数据集 $D\in\{0,1\}^{N\times m}$，$\boldsymbol{\alpha}\in\{0,1\}^N\subseteq D$。$\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_m$ 是一组要组合的单特征向量，其中 $\boldsymbol{\alpha}_j=(a_{1j},a_{2j},\cdots,a_{Nj})^{\mathrm{T}}$。组合特征向量 $\boldsymbol{\beta}=(b_1,b_2,\cdots,b_N)^{\mathrm{T}}$，当且仅当 $\forall a_{ij}=1$ 时，$b_i=1$；否则 $b_i=0$。其中 $0\leqslant i\leqslant N$，$0\leqslant j\leqslant m$。

3 面向序数回归的组合特征提取方法

本节将详细介绍序数回归中组合特征提取的方法。该方法主要分为 3 个步骤。

1）频繁模式挖掘：在数据集 D 上提取满足用户指定最小支持度 min_sup 的频繁项集 $\mathscr{L}$。

2）序数回归的有序二元分解：把序数回归分解成多个二元子问题，并生成二元频繁项集合。

3）组合特征提取：在二元频繁项集合中应用组合特征提取算法，提取最有区分能力的频繁模式集合 F_s，根据频繁模式 F_s 进行特征组合，生成组合特征数据集 D_s。最后，结合单特征数据集 D 和组合特征数据集 D_s 建立序数回归模型。

3.1　频繁模式挖掘

设序数数据集 D 有 k 个分类属性 $A_1,A_2,\cdots,A_k$。对连续属性进行离散化，转化为分类属性。形如 (att,val) 的（属性，值）对表示为项 p，项集 $I=\{p_1,p_2,\cdots,p_m,C_1,C_2,\cdots,C_q\}$，其中 m 表示数据中（属性，值）对的总数，q 表示序数等级数。令 $\boldsymbol{x}=(x_1,x_2,\cdots,x_m)$ 表示数据集中一条记录 s 的特征向量。若（属性，值）对 $p_j(s)$ 存在，则 $x_j=1$；若 $p_j(s)$ 不存在，则 $x_j=0$。因此数据集表示为 $D=\{\boldsymbol{x}_i,y_i\}_{i=1}^n$，其中 n 为数据集的大小，$\boldsymbol{x}_i\in\{0,1\}^m$，且 $x_{ij}\in\{0,1\}$，$i\in[1,n]$，$j\in[1,m]$。

在挖掘用于组合特征的频繁模式时，我们只对形如 $p_1\wedge p_2\wedge p_3\wedge\cdots\wedge p_l\Rightarrow y_i$（$l\leqslant m$）的关联规则感兴趣，其中规则后件是序数等级，规则前件是特征项的合取。本文定义这样的关联规则为有关类别的关联规则，其中规则前件称为有关类别的频繁模式，记作 $FP=\{p_1,p_2,\cdots,p_l\}$，则规则 $FP\Rightarrow y_i$ 的支持度如下：

$$support(P\Rightarrow y_i)=\frac{support_{count(FP\cup y_i)}}{N} \tag{1}$$

其中，$support_{count(FP\cup y_i)}$ 是项集 $FP\cup\{y_i\}$ 的支持度计数，表示项集 $FP\cup\{y_i\}$ 在 D 中出现的次数，N 表示数据集 D 中数据的总条数。令 θ_0 表示最小支持度阈值，$0\leqslant\theta_0\leqslant1$。满足最小支持度 min_sup 的频繁项集记为 $\mathscr{L}$。

在数据集 D 中，频繁模式的确定与最小支持度 min_sup 有关。我们知道，一个模式具有很高的支持度，说明它在数据集中的覆盖率比较高。然而，如果该频繁模式在不同类别中出现的比率相同，说明该频繁模式在数据中具有普遍性，那么该模式对于提高模型的准确率没有帮助。与之相反，低支持度的模式如果只与某一类别有关，在其余类别中没有出现，那么该模式具有很高的区分能力。然而若支持度选择得过低，会出现两个问题：1）若原始数据的单特征较多，频繁模式会随着最小支持度的下降而爆炸性增长；2）过低的支持度会产生不具有代表性的频繁模式。如果选择这些频繁模式进行特征组合，会降低模型的鲁棒性。因此，在生成频繁模式集时，最小支持度应该根据数据大小和特征数量来确定。本文实验中，最小支持度设置为 2%～10%之间。

FP-growth 算法在挖掘频繁模式时具有有效性和扩展性，并且比 Aprori 算法快一个数量级，因此使用 FP-growth 算法来生成满足最小支持度 min_sup 的频繁项集合 $\mathscr{L}$。

3.2　序数回归的有序二元分解

由于序数回归的标签是有序的，在提取序数回归组合特征时需要考虑不同类别之间的序数关系。我们对序数回归进行有序的二元分解，把序数回归分解为多个二元子问题，再进行组合特征提取，通过这种方式，我们把序数信息融入到频繁模式提取中。

假设序数回归具有 q 个等级标签 $y=\{C_1,C_2,\cdots,C_q\}$，在进行有序二元分解时，会被分解为 $q-1$ 个二元子问题。设 y_r 为分解后二元子问题的标签，数据中每个记录标签 $y_{ri}=\{+1,-1\}$，$r=1,2,\cdots,q-1$，$i=1,\cdots,N$，如果 $y_i>C_q$，那么 $y_{ri}=+1$；如果 $y_i\leqslant C_q$，那么 $y_{ri}=-1$。图 1 给出具有 5 个等级的序数回归有序二元分解，其中每列表示分解后的一个二元子问题，每行表示对应二元子问题中每个类的符号，－表示负类，＋表示正类。

使用有序二元分解方法对频繁项集 $\mathscr{L}$ 进行分解，得到 $q-1$ 个二元频繁项集，记为 $\mathscr{L}_1,\mathscr{L}_2,\cdots,\mathscr{L}_{q-1}$。

$$\begin{pmatrix}-,-,-,-\\+,-,-,-\\+,+,-,-\\+,+,+,-\\+,+,+,+\end{pmatrix}$$

图 1　序数回归的有序分解

Fig. 1　Ordered decomposition of ordinal regression

3.3　组合特征提取

通过上文的分析可知，虽然频繁模式能够提升模型的准确性，但并不是所有的频繁模式都对类别区分有帮助。因此，需要一个指标对频繁项集进行筛选，从而得到最有区分能力的频繁模式。本文使用 K-L 散度来度量频繁模式的区分能力。设 $C\in\{-1,1\}$ 是序数回归分解后二元子问题的类别。$P(C)$ 表示二元问题中 C 的二元概率分布。设 FP 是二元频繁项集中的一个有关类别的频繁模式，$Q(C)$ 表示在频繁模式 FP 条件下 C 的二元概率分布，则概率分布 P 对 Q 的 K-L 散度，即频繁模式 FP 的区分能力为：

$$D_{KL}=\sum_{c_i\in\{-1,1\}}P(c_i)\cdot\log\frac{P(c_i)}{Q(c_i)} \tag{2}$$

由 K-L 散度性质可得，D_{KL} 越大表明概率分布 P 和 Q 的差异越大，即在频繁模式 FP 下，正负类的记录条数的比例变化越大。这表明频繁模式 FP 对类别的区分能力越大，因此频繁模式 FP 具有区分能力。反之，D_{KL} 的值越小，频繁模式 FP 的区分能力越小。

在每个有序分解后生成的二元频繁项集中，只计算有关类别的频繁模式的K-L 散度值。使用式(2)对每个二元频繁项集计算有关类别频繁模式的 K-L 散度值，生成候选频繁模式集，记为 $S_1,S_2,\cdots,S_{q-1}$。

对候选频繁模式集进行筛选时，根据每个频繁模式的 K-L 散度值选择最有区分能力的模式进行特征组合。由于在计算模式的 K-L 散度值时，每个二元频繁项集独立计算，因此同一个频繁模式可能出现在多个候选频繁模式集中，且 K-L 散度值也不尽相同。此外，不同频繁模式集中频繁模式的K-L 散度值的大小也有区别。如果将不同等级的频繁模式集合并，选择最有区分能力的频繁模式，会造成某个等级选择的频繁模式过多，而有关其他等级的频繁模式的选择过少或者没有，这样不利于其他等级的区分。例如，图 2 是 SWD 数据集[23]（儿童受虐风险数据集）上不同二元候选频繁模式集上频繁模式 K-L 散度值的分布。由图 2 可知，二元频繁模式

集 S_1 中频繁模式 K-L 散度值的上限比其余两个频繁模式集的要小。如果只按照 K-L 散度值的大小来选择频繁模式，那么选择的频繁模式多数属于 S_2 与 S_3，对第一个等级有区分力的频繁模式的数量会很少，甚至没有。

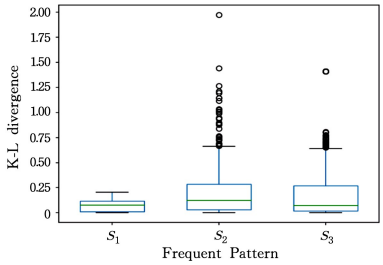

图 2 SWD 数据频繁模式 K-L 散度值分布

Fig. 2 Distribution of K-L divergence values of frequent patterns in SWD

因此，我们使用一种循环策略，均衡考虑区分不同等级的频繁模式，从每个候选频繁模式集中选择出最有区分能力的频繁模式。设最终用于特征组合的频繁模式集为 F_s。其基本思想是，根据 K-L 散度值对候选频繁模式集中的频繁模式进行排序，从第一个候选频繁模式集 S_1 开始，选择最顶层 K-L 散度值最大的频繁模式 α 添加到 F_s 中，并将该模式从 S_1 中去除。然后从第二个候选频繁模式集 S_2 中选择最有区分能力的频繁模式 β，判断 β 是否存在于 F_s 中，如果不存在，则把 β 添加到 F_s 中，如果存在则把 β 从 S_2 中去除，并对下一个等级进行频繁模式选择。循环执行该选择策略，直到选择出用户指定个数 m 的频繁模式。综上，提取最有区分能力的频繁模式集的算法如算法 1 所示。

算法 1 最有区分能力的频繁模式集的提取算法

输入：二元频繁项集 $\mathcal{L}_1,\mathcal{L}_2,\cdots,\mathcal{L}_{q-1}$；选择组合特征个数 m

输出：m 个最有区分能力的频繁模式集

步骤：

```
1. Si={ };/* 候选频繁模式集 i=1,2,…,q-1 */
2. for(i=1,i<q,i++)/* 求候选频繁模式集 S1,S2,…,Sq-1 */
3.    for p in Li
4.       if p 是有关类别的频繁模式 AND p 不在候选频繁模式集合 Si 中
5.          计算频繁模式 p 的 K-L 散度值，并将其添加至候选频繁模
            式集合 Si 中；
6. Fs={ };
7. while(length(Fs)<m)/* 选择前 m 个最有区分能力的频繁模式集
   Fs */
8.    for(j=1,j<q,j++)
9.       按 K-L 散度值对 Sj 排序；
10.      从 Sj 中选择具有最大 K-L 散度值的频繁模式 α；
11.      if α 不在 Fs 中
12.         把 α 添加到 Fs 中，并从 Sj 中去除；
13.         if length(Fs)==m
14.            break；
15.      else
16.         把 α 从 Sj 中去除；
17. return Fs.
```

最后，使用频繁模式集 F_s 对单特征进行特征组合，得到组合特征数据集 D_s。结合单特征数据集 D，对序数回归模型进行训练。

4 实验与分析

4.1 实验设置

本文在 5 个数据集上进行实验。其中 PCD 是我们自有的银行信贷客户信用数据集，其余 4 个数据集来自公开数据集 mldata.org[23] 与 UCI[24]。表 1 给出了这些数据集的详细信息，包括记录条数、属性个数、等级数、项数和每个等级包含的记录条数。对数据中所有的标称属性和序数属性进行 One-Hot 编码，对于连续属性则先进行离散化，再进行 One-Hot 编码。表 1 中的项数表示数据预处理后的单特征维度。

表 1 序数回归数据集详情

Table 1 Details of ordinal regression datasets

数据集	记录数	属性数	项数	等级数	等级分布
PCD(PC)	6714	19	102	3	(5053,995,666)
Balance-scale(BS)	325	4	20	3	(288,49,288)
Car(CA)	1728	6	21	4	(1210,384,69,65)
SWD(SW)	1000	10	31	4	(32,352,399,217)
LEV(LE)	1000	4	20	5	(93,280,403,197,27)

经过预处理后，在各个数据集上使用 FP-growth 算法挖掘频繁模式时，统一设置支持度为 5%。使用本文所提方法在各个数据集上选取一组区分能力最强的频繁模式集，使用这些频繁模式在单特征数据集(Item)上产生组合特征(FS)，然后分别在单特征数据集和单特征加组合特征数据集(Item&FS)上进行模型训练。在本文实验中，频繁模式选择的数量分别为 10，50，100 和 150。最终实验结果是这 4 个不同频繁模式数量下的最优值。

为了减小由数据选择造成的随机误差，对每个数据集进行 5 次等比抽样，每次抽取 80%的数据作为训练集，20%的数据作为测试集，最终结果为全部数据集上的平均值。所有模型的超参数通过在训练集上的五折交叉验证来确定，选择最好的模型在测试集上进行验证。

选择当前具有代表性的序数回归模型进行实验，如表 2 所列。其中 SVM 有关的序数回归模型基于 libsvm 库(3.0 版本)运行实现，其他模型使用其作者提供的开源程序来实现。

表 2 序数回归模型的选择

Table 2 Selection of ordinal regression models

缩写	全称
SVC1VA	Support Vector Classifier with OneVsAll[3]
SVR	Support Vector Machines for regression[2]
CSSVC	Cost-Sensitive Support Vector Classifier[3]
NNOP	Neural Network with Ordered Partitions[4]
ELMOP	Extreme Learning Machine with Ordered Partitions[5]
POM	Proportional Odds Model[6]
NNPOM	Neural Network based on Proportional Odd Model[7]
SVOREX	Support Vector Ordinal Regression with Explicit Constraints[8]
SVORIM	Support Vector Ordinal Regression with Implicit Constraints[8]
SVORLin	SVORIM using a linear kernel[8]

4.2 评测指标和模型选择

选用平均零一误差(Mean Zero-one Error，MZE)和平均绝对误差(Mean Absolute Eror，MAE)来评估序数回归模型。

MZE 表示模型的错误率，其计算式如下：

$$MZE=\frac{1}{N}\sum_{i=1}^{N}[\![y_i^*\neq y_i]\!]=1-Acc \tag{3}$$

其中，y_i 是真实标签，y_i^* 是预测标签，Acc 是模型的准确率。*MZE* 的取值区间为[0，1]，代表的是模型整体的准确性，但是没有考虑标签的序数关系。

MAE 表示预测等级($O(y_i^*)$)与真实等级($O(y_i)$)的绝对平均误差，计算式如下：

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|O(y_i)-O(y_i^*)| \tag{4}$$

MAE 的取值范围为[0，$q-1$]。这两个指标的通俗理解是，*MZE* 考虑的是错误分类 0-1 损失，*MAE* 考查的是绝对损失。

4.3　实验结果与分析

表 3 给出了在不同数据集上使用组合特征前后模型 *MAE* 的实验结果，表 4 给出了在不同数据集上使用组合特征前后模型 *MZE* 的实验结果。其中，加粗的数字表示相同模型下最优的实验结果。从表 3 和表 4 可以看出，使用组合特征时，多数模型效果都得到了提升，如 SVR，CSSVC，EL-MOP，POM，NNPOM，SVOREX，SVORLin 模型在 5 个数据集上的 *MZE* 和 *MAE* 都有提升，只有 SVORIM 模型在 PC 数据集上的效果没有提升，但与最优结果相差很小。这一实验结果证实了挖掘有区分能力的频繁模式进行特征组合的有效性，另外也证实了使用频繁模式产生的组合特征的判别力高于单特征。

表 3　使用组合特征前后 *MAE* 的变化情况

Table 3　Variation of *MAE* before and after using combination features

Model	PC		BS		CA		SW		LE	
	Item	Item&FS	Item	Item&FS	Item	Item&FS	Item	Item&FS	Item	Item&FS
SVC1VA	0.32566	**0.31419**	0.07619	**0.07301**	**0.01156**	0.01387	0.48824	**0.45326**	0.41800	**0.39900**
SVR	0.34360	**0.33709**	0.38095	**0.26349**	0.03699	**0.01965**	0.45327	**0.43316**	0.40900	**0.38200**
CSSVC	0.32343	**0.31829**	0.08095	**0.07301**	0.03237	**0.01387**	0.48623	**0.44924**	0.41100	**0.39200**
NNOP	0.32701	**0.30578**	**0.01111**	**0.01111**	0.03468	**0.00404**	0.45513	**0.42211**	0.41300	**0.39000**
ELMOP	0.34755	**0.32666**	0.23809	**0.16667**	0.18150	**0.12080**	0.44723	**0.43316**	0.40300	**0.37900**
POM	0.33817	**0.33169**	0.06825	**0.04444**	0.07457	**0.04740**	0.45527	**0.41608**	0.40400	**0.38900**
NNPOM	0.30311	**0.27652**	0.04127	**0.02063**	0.00520	**0.00404**	0.47417	**0.42713**	0.40200	**0.37100**
SVOREX	0.29754	**0.24838**	0.03175	**0.02222**	0.02312	**0.01560**	0.44718	**0.42814**	0.41900	**0.38000**
SVORIM	**0.31293**	0.31968	0.03175	**0.02222**	0.02312	**0.01560**	0.44121	**0.42512**	0.42200	**0.37900**
SVORLin	0.34666	**0.33666**	**0.00634**	**0.00476**	0.08150	**0.04624**	0.44723	**0.43718**	0.41800	**0.40300**

表 4　使用组合特征前后 *MZE* 的变化情况

Table 4　Variation of *MZE* before and after using combination features

Model	PC		BS		CA		SW		LE	
	Item	Item&FS	Item	Item&FS	Item	Item&FS	Item	Item&FS	Item	Item&FS
SVC1VA	0.23476	**0.22162**	**0.06825**	0.06984	**0.01156**	0.01387	0.43608	**0.41708**	0.37600	**0.36700**
SVR	0.23208	**0.22805**	0.38095	**0.25396**	0.03699	**0.01965**	0.43618	**0.41809**	0.37800	**0.35400**
CSSVC	0.23408	**0.22073**	0.07460	**0.06984**	0.02543	**0.01271**	0.45904	**0.41708**	0.37800	**0.35900**
NNOP	0.24368	**0.23832**	**0.01111**	**0.01111**	0.03410	**0.00231**	0.44105	**0.40804**	0.38100	**0.35600**
ELMOP	0.24949	**0.22748**	0.18095	**0.14285**	0.15318	**0.11156**	0.43015	**0.40415**	0.36900	**0.35100**
POM	0.25307	**0.24793**	0.06667	**0.04444**	0.07341	**0.04682**	0.43819	**0.40301**	0.36900	**0.35300**
NNPOM	0.22649	**0.21040**	0.04127	**0.02063**	0.00520	**0.00404**	0.45608	**0.40502**	0.36600	**0.34300**
SVOREX	0.24413	**0.24368**	0.03175	**0.02222**	0.02312	**0.01560**	0.43809	**0.41105**	0.38600	**0.35400**
SVORIM	**0.23207**	0.23422	0.03175	**0.02222**	0.02312	**0.01560**	0.42814	**0.42110**	0.38900	**0.35200**
SVORLin	0.24748	**0.23946**	0.00634	**0.00476**	0.07919	**0.04450**	0.43115	**0.41909**	0.38500	**0.37200**

接下来，分析选择组合特征的数量对模型结果的影响。以 SVOREX 模型为例，验证使用不同数量的组合特征时模型的变化情况。表 5 给出了 5 个数据集上使用不同组合特征数量时 SVOREX 模型的 *MAE* 值。由实验结果可以得到，使用组合特征的 *MAE* 值都比单特征的要小，这说明了添加有区分能力的组合特征会提升模型预测的准确性。同时可以看到，随着组合特征数量的不断增加，*MAE* 值先变小后稍微增大，由此可以得出，不是所有的频繁模式都有区分能力，使用区分能力弱的频繁模式进行组合的特征不会提高模型的准确性，反而会因为特征空间过大使得模型训练的效果减弱。因此，在使用频繁模式提取组合特征时，应该选择具有区分能力的频繁模式进行特征组合。

表 5　使用 SVOREX 模型时不同组合特征数量的 *MAE* 值

Table 5　*MAE* of different number of combination features when using SVOREX model

FS_Num	PC	BS	CA	SW	LE
0	0.29754	0.03175	0.02312	0.44718	0.41900
10	0.27652	0.02964	0.01965	0.43221	**0.38000**
50	**0.24838**	**0.02222**	0.01734	**0.42814**	0.38700
100	0.25987	0.25874	**0.01560**	0.43125	0.39200
150	0.25374	0.26586	0.01697	0.43284	0.39800

结束语　本文提出一种新的面向序数回归的组合特征提取方法，首先将序数回归有序分解为多个二元子问题，在每个二元子问题上，挖掘有关类别的频繁模式，并计算每个频繁模式的 K-L 散度值，选择最具有区分能力的频繁模式作为组合

特征依据，最后进行特征组合。与现有序数回归中特征选择的方法不同，本文通过挖掘新的有效特征来提升模型的训练效果。本文在公开数据和自有数据 5 个数据集上使用多个序数回归模型进行验证。实验结果表明，使用最有区分能力的频繁模式组合特征，能够有效提升序数回归模型的准确性。此外，分析了组合特征选择数量与模型准确性提升的关系。当使用有区分能力的频繁模式构成的组合特征时，模型效果会得到提升，但是当区分能力较弱的频繁模式加入时，模型的提升效果减弱。

参 考 文 献

[1] GUTIÉRREZ P A, PÉREZ-ORTIZ M, SÁNCHEZ-MONEDERO J, et al. Ordinal Regression Methods: Survey and Experimental Study[J]. IEEE Transactions on Knowledge & Data Engineering, 2016, 28(1): 127-146.

[2] SMOLA A J, SCHOELKOPF B. A tutorial on support vector regression[J]. Statistics and Computing, 2004, 14(3): 199-222.

[3] HSU C W, LIN C J. A Comparison of Methods for Multiclass Support Vector Machines[C]// IEEE TRANS. on Neural Networks. 2002: 415-425.

[4] CHENG J, WANG Z, POLLASTRI G. A neural network approach to ordinal regression[C]// IEEE International Joint Conference on Neural Networks. IEEE, 2008: 1279-1284.

[5] DENG W Y, ZHENG Q H, LIAN S, et al. Ordinal extreme learning machine[J]. Neurocomputing, 2010, 74(1-3): 447-456.

[6] MCCULLAGH P. Regression Models for Ordinal Data[J]. Journal of the Royal Statistical Society, 1980, 42(2): 109-142.

[7] MATHIESON M. Ordinal Models for Neural Networks [C]// Neural Networks in Finanicial Engineering. 1996: 523-536.

[8] WEI C, KEERTHI S S. Support Vector Ordinal Regression [M]. MIT Press, 2007.

[9] MUKRAS R, WIRATUNGA N, LOTHIAN R, et al. Information Gain Feature Selection for Ordinal Text Classification using Probability Redistribution [C] // Proceedings of the Textlink Workshop at IJCAI 2007. Hyderabad, 2007: 1-10.

[10] BACCIANELLA S, ESULI A, SEBASTIANI F. Multi-facet Rating of Product Reviews[C]// European Conference on Ir Research on Advances in Information Retrieval. Springer-Verlag, 2009: 461-472.

[11] BACCIANELLA S, ESULI A, SEBASTIANI F. Feature selection for ordinal regression[C]// ACM Symposium on Applied Computing. DBLP, 2010: 1748-1754.

[12] BACCIANELLA S, ESULI A, SEBASTIANI F. Feature Selection for Ordinal Text Classification[J]. Neural Computation, 2014, 26(3): 557-591.

[13] AGRAWAL R, IMIELIŃSKI T, SWAMI A. Mining association rules between sets of items in large databases[C]// ACM SIGMOD International Conference on Management of Data. ACM, 1993: 207-216.

[14] AGRAWAL R. Fast algorithms for mining association rules [C] // Proc. VLDB Conference. 1994: 487-499.

[15] HAN J, PEI J, YIN Y. Mining frequent patterns without candidate generation[C]// ACM SIGMOD International Conference on Management of Data. ACM, 2000: 1-12.

[16] ZAKI M J. Hsiao: CHARM: An efficient algorithm for closed itemset mining[C]// Proceedings of the Second SIAM International Conference on Data Mining. Arlington, 2002: 457-473.

[17] YAN X, YU P S, HAN J. Graph indexing: a frequent structure-based approach [C] // SIGMOD' 04 Proceedings of the 2004 ACM SIGMOD international conference on Management of data. New York: ACM, 2004: 335-346.

[18] WANG K, XU C, LIU B. Clustering transactions using large items[C]// International Conference on Information and Knowledge Management, Proceedings. Staff Publications, 1999: 483-490.

[19] LIUB, HSUW, MAY M. Integrating classification and association rule mining[C]// 4th International Conference on Knowledge Discovery and Data Mining. 1998: 80-86.

[20] LODHI H, SAUNDERS C, SHAWE-TAYLOR J, et al. Text classification using string kernels[J]. Journal of Machine Learning Research, 2002, 2(3): 419-444.

[21] DEHKORDI M N, SHENASSA M H. CLoPAR: Classification based on Predictive Association Rules[C]// 2006 3rd International IEEE Conference Intelligent Systems. IEEE, 2007: 483-487.

[22] WANG J, KARYPIS G. HARMON Y : Efficiently Mining the Best Rules for Classification [C] // Siam Conference on Data Mining. 2005: 205-216.

[23] PASCAL. Pascal (Pattern Analysis, Statistical Modelling and Computational Learning) machine learning benchmarks repository[EB/OL]. Available: http://mldata.org.

[24] ASUNCIONA, NEWMAND. UCI Machine Learning Repository[EB/OL]. http://www.ics.uci.edu/~mlearn/MLRepository.html.