文章编号:1671-4598(2021)10-0164-06 DOI:10.16526/j.cnki.11—4762/tp.2021.10.030 中图分类号:TP212.1 文献标识码:A

异常信息的智能分类算法研究

马宗方,马祥双,宋 琳,罗 婵
(西安建筑科技大学 信息与控制工程学院,西安 710055)

摘要:信息处理过程中对异常信息的智能化处理是一个前沿的且富有挑战性的研究方向;针对所获取的信息由于噪声干扰等因素存在缺失这一异常现象,提出了一种不完整(缺失)数据的智能分类算法;对于某一个不完整样本,该方法首先根据找到的近邻类别信息得到单个或多个版本的估计样本,这样在保证插补的准确性的同时能够有效地表征由于缺失引起的不精确性,然后用分类器分类带有估计值的样本;最后,在证据推理框架下提出一种新的信任分类方法,将难以划分类别的样本分配到对应的复合类来描述由于缺失值引起的样本类别的不确定性,同时降低错误分类的风险;用 UCI 数据库的真实数据集来验证算法的有效性,实验结果表明该算法能够有效地处理不完整数据分类问题。

关键词:智能化处理;异常信息;不完整数据;分类

Incomplete Data Belief Classification Algorithm Based on Adaptive KNN Imputation

MA Zongfang, MA Xiangshuang, SONG Lin, LUO Chan
(School of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China)

Abstract: The intelligent processing of abnormal information in the process of information processing is a frontier and challenging research direction. In this paper, an intelligent classification algorithm for missing data is proposed for the abnormal phenomenon of missing information due to noise interference and other factors. This method yield one or more versions of estimations for a specific incomplete sample, which can not only ensures the accuracy of estimation, but also capture its imprecision. Then, the basic classifier that can implement in complete data well is employed here to classify the edited data with estimations. A new belief classification approach is developed in this paper to assign the indistinguishable object to the proper meta—classes so as to characterize the uncertainty of classification caused by missing values and decrease the risk of misclassification. The real data sets in the UCI are employed to verify the effectiveness of the proposed method, and the result represents that the proposed method can effectively handle the problem of incomplete data classification.

Keywords: intelligent processing; abnormal information; missing data; classification

0 引言

智能信息处理是信号与信息技术领域一个前沿的富有挑战性的研究方向,它以人工智能理论为基础,侧重于信息处理的智能化,包括计算机智能化(文字、图象、语音等信息智能处理)、通信智能化以及控制信息智能化[1-3]。然而在实际应用中,各种复杂因素可能会导致采集到的数据信息是不完整的。例如,在气象数据的采集过程中,由于传感器故障或者在数据传输过程中信号受到噪声干扰,就会造成某一段时间内的数据缺失;在填写调查问卷时,调查者不愿意回答那些涉及到个人隐私的问题,这样就导致无法获取这部分信息。由于大多数传统的分类器不能直接处理含有缺失值的数据集,因此有大批学者研究并提出了适用于不完整数据的分类算法[4-5]。其中,最简单的就是删除含有缺失值的样本或者删除缺失值所在的属性项,然后再用传统的分类器分类[5]。但是删除法仅适用于缺失率不到 5%的数据集,并且删除属性项可能会改变数据分布,进而会影响算法分类性能。最为常见并且有效的处理不完整数据分类问题的方法是插补法,就是通过合理的估值来填补缺失数据,这样就能用基础分类器对带有估计值的完整数据分类[5]。

1987 年,Little 和 Rubin[7-8]针对数据缺失机制提出了 3 种不同类型的缺失情况:完全随机缺失(MCAR, missing complete at random)、随机缺失(MAR, missing at random)和非随机缺失(MNAR, missing not at random)。目前对缺失数据的研究主要集中在 MAR 和 MCAR 上。在众多的缺失值插补方法中,主要分为单值插补和多值插补。常用的单值插补方法是均值插补(MI, mean imputation)[9],其主要思想是根据已观测属性值的平均值代替缺失值,但是均值插补没有考虑到样本不同属性之间的联系

收稿日期:2021-03-01; **修回日期**:2021-03-18。
基金项目:陕西省工业攻关项目(Z20200051)。
作者简介:马宗方(1980-),男,安徽临泉人,博士,副教授,硕士生导师,主要从事智能信息处理、机器视觉工业应用方向的研究。
引用格式:马宗方,马祥双,宋 琳,等. 异常信息的智能分类算法研究[J]. 计算机测量与控制,2021,29(10):164-169.

并且可能会改变数据的分布。K 近邻插补（K－nearest neighbors imputation，KNNI）[10]，主要根据数据中缺失样本的完整变量找出与其距离最近的 K 个样本，然后利用距离函数分别计算这 K 个样本与该样本的距离加权这 K 个样本对应不完整样本的缺失项得出估计值。FCM 插补（FCMI，fuzzy c－means imputation，）[11] 是首先对数据集用 FCM 聚类，然后用聚类后的类中心和隶属度来估计不完整样本的缺失值。多值插补方法，也就是为缺失值提供多个版本的估计值来表征估计值的不精确性。最早提出的多重插补方法[12]是对缺失数据集插补 m 次（$m>1$），每次插补后得到一个完整的数据集，最终可以得到 m 个完整的数据集，接着对这 m 个完整数据集进行分析，综合这 m 次插补结果，做出统计推断。一种基于随机森林的多值插补方法[13]是通过构造大量的回归树与随机树来给缺失值提供多个版本的估计值。

传统的基于插补的不完整数据分类方法会将待测样本分配给确切的类别。然而由于数据的缺失，样本的类别可能变得很模糊，而这些方法并没有考虑到这种数据缺失对分类的影响。在这种情况下如果强硬的划分样本给某一单类，这会增加错误分类的风险。在这种情况下，如何去表征数据缺失引起的不确定性是亟待解决的问题。

证据推理[14-19]，是由 Dempster 在 1967 年最先提出的，后来由 Shafer 推广并形成的理论，所以也称为 Dempster－Shafer 理论。因为证据理论具有处理不精确和不确定信息的优势，因此广泛应用于数据分类、专家系统和信息融合等领域。文献［20］提出了一种基于证据推理的不完整数据分类方法（PCC，Pototype－based credal classification）。PCC 首先用不同类别的中心分别估计缺失值来表征估计值的不精确性，然后对带有不同版本估计值的不完整样本的分类结果折扣融合，最后将那些难以确切划分类别的样本分配到合适的复合类来表征由于缺失值引起的类别的不确定性。

本文提出了一种不完整数据智能分类方法，该方法依据不完整样本近邻中的类别信息自适应的估计缺失值，也即采取单值插补与多值插补相结合的混合插补策略，并且在证据推理框架下，提出一种新的信任分类方法来合理的表征不完整样本类别的不确定性，并有效地避免错误分类的风险。

1　背景知识

1.1　K 近邻插补

K 近邻插补是用不完整数据的已知属性在完整数据中找到 K 个近邻，然后用它们估计不完整数据对应的缺失值。假设测试样本集 $X=\{x_1, x_2, \ldots, x_n\}$ 利用训练样本集 $Y=\{y_1, y_2, \ldots, y_n\}$ 在类别识别框架 $\Omega=\{w_1, w_2 \ldots, w_c\}$ 下进行分类，其中 x_i 是测试样本集中的第 i 个样本。目标 x_i 已知 T 个属性值。首先，利用 x_i 已知属性计算它与训练集 Y 中的每一个样本 y_i（1，2，…，n）之间的欧氏距离，距离公式如下：

$$D(x_i,y_i)=\|x_i-y_i\|=\sqrt{\sum_{a=1}^{T}(x_{ia}-y_{ia})^2} \tag{1}$$

其中：x_{ia} 和 y_{ia} 分别表示测试样本 x_i 和训练样本 y_i 的第 a 个已知属性。然后，对这 n 个距离从小到大的排序，K 个最小距离对应的样本即为 x_i 的 K 个近邻。因为样本 x_i 与 K 个近邻训练样本 y_i 的距离不同，因此 K 个近邻估计值的权重也就不同。估计值的权重 $W_k^i(k=1,2,\ldots,K)$ 计算公式如下：

$$W_k^i=\frac{e^{-\|x_i-y_k^i\|}}{\sum_{k=1}^{K}e^{-\|x_i-y_k^i\|}} \tag{2}$$

由此，我们可以得到，距离 x_i 越近的近邻所占的权重越大。最后对 x_i 缺失值对应的 K 近邻的已知属性值加权求和，得到的结果即为 x_i 的估计值。

1.2　信任函数的理论基础

信任函数是由 Shafer 在其独创的数学证据理论中引入的，该理论也被称为证据理论（evidence theory），简称 DST。该理论已经在信息融合、模式识别以及决策分析等领域得到了广泛应用。

假设 Ω 是一个识别框架，或称为假设空间。在证据推理中，识别框架从 $\Omega=\{\omega_1,\omega_2,\ldots\omega_c\}$ 扩展到幂集 2^{Ω}，其包含了 Ω 所有的子集。一个证据的基本信任分配（BBA，basic belief assignment），指从幂集 2^{Ω} 到［0，1］上的一个映射函数 $m(.):2^{\Omega}\rightarrow[0,1]$（又称为基本信任函数或者 mass 函数），并满足以下条件：

$$\begin{cases}\sum_{A\in 2^{\Omega}}m(A)=1\\ m(\varphi)=0\end{cases} \tag{3}$$

如果 $m(A)>0$，则称 A 为焦元。如果 $m(A)=\max(m(\cdot))$，则称 A 为主焦元。

设在识别框架上有两个独立证据 B 和 C，它们的 mass 函数分别为 m_1，m_2 的 Dempster 合成规则为：

$$m_1\oplus m_2(A)=\frac{1}{K}\sum_{B\cap C=A}m_1(B)m_2(C) \tag{4}$$

其中：K 为归一化常数即矛盾因子：

$$K=\sum_{B\cap C\neq\varphi}m_1(B)m_2(C)=1-\sum_{B\cap C=\varphi}m_1(B)m_2(C) \tag{5}$$

由于满足交换律和结合律，即可推广到 n 个互相独立的证据，Dempster 合成结果为：

$$(m_1\oplus m_2\oplus\ldots\oplus m_n)(A)=$$
$$\frac{1}{K}\sum_{A_1\cap A_2\cap\ldots\cap A_n=A}m_1(A_1).m_2(A_2)\ldots m_n(A_n) \tag{6}$$

其中：

$$K=\sum_{A_1\cap\ldots\cap A_n\neq\varphi}m_1(A_1).m_2(A_2)\ldots m_n(A_n)=$$
$$1-\sum_{A_1\cap\ldots\cap A_n=\varphi}m_1(A_1).m_2(A_2)\ldots m_n(A_n) \tag{7}$$

尽管证据理论在解决不确定性的问题方面具有一定优势，但是在实际应用中，Dempster 组合规则不适用于高冲突证据，在这种情况下常常会得出与常识相悖的结论。为

此，许多学者提出了改进的组合规则。大致分为两种：1）改进组合规则；2）融合前对证据进行修改。在改进规则中，一般认为不合理融合结果主要由于 DS 规则对高冲突信息的分配不当造成的，针对此一些学者提出了改变冲突信息分配方式来改进规则。在证据修改方法中，一般认为高冲突证据在融合过程中每个证据的权重是不一样的，要先对证据加权处理，然后再融合。

2 不完整数据智能分类算法

针对不完整数据分类问题，本文提出了一种不完整数据智能分类方法。它主要包含 3 个步骤：1）自适应插补；2）折扣分类结果；3）全局融合。

2.1 自适应插补

假定训练集为 $Y=\{y_1,\ldots,y_m\}$，共包含了 c 个类别，并且训练样本都是完整的；测试集 $X=\{x_1,\ldots,x_n\}$，并且测试样本都存在缺失值。用缺失样本 $x_i(i=1,\ldots,n)$ 的 T 个已知属性在训练集中寻找 K 近邻。首先计算 x_i 与每一个测试样本 $y_j(j=1,\ldots,m)$ 之间的欧式距离，如下所示：

$$D(x_i,y_j)=\|x_i-y_j\|=\sqrt{\sum_{a=1}^{T}(x_{ia}-y_{ja})^2} \tag{8}$$

其中：x_{ia} 和 y_{ja} 和分别表示测试样本 x_i 和训练样本 y_i 的第 a 个属性。计算得出 m 个距离值，对这 m 个距离从小到大排序，其中最小的 K 个距离对应的 K 个样本 $\{y_1,\ldots,y_k,\ldots,y_K\}$ 即为 x_i 的 K 近邻。这 K 个近邻可能来自于 p $(1\leqslant p\leqslant c)$ 个类别 $\{\omega_1,\ldots,\omega_g,\ldots,\omega_p\}$。当 $p=1$ 时，也即近邻都来自于同一个类，在这种情况下样本 x_i 就有很大的可能属于这个类，那么就用这些近邻插补 x_i 得到一个精确的估计值。当 $p>1$ 时，也即近邻都来自于多个类，说明样本的数据缺失导致它的类别变得很模糊，为了降低这种数据缺失带来的不确定性的影响，在这种情况下我们采取多值估计策略，用这不同类别的近邻分别估计 x_i 的缺失值。

此外，由于每个近邻与不完整样本间的距离不同，因此对估计缺失值的贡献也就不同，也即距离不完整样本越近的近邻在估计缺失值时的比重应该越大。对于不完整样本 x_i 来说，我们用它的 K 近邻中属于类的近邻来估计缺失值，那么用这些近邻估计缺失值 $\omega_g(g=1,\ldots,p)$ 时的权重计算如下所示：

$$\alpha_h^g=\frac{e^{-\|x_i-y_h^g\|}}{\sum_{y_h^g\in\omega_g}e^{-\|x_i-y_h^g\|}} \tag{9}$$

其中：α_h^g 表示 K 近邻中属性 ω_g 类的第 h 个近邻 y_h^g 在估计 x_i 缺失值时的权重。目标样本 x_i 的缺失值估计如下：

$$x_{io}^g=\sum_{y_h^g\in\omega_g}\alpha_h^g\cdot y_{ho}^g \tag{10}$$

其中：x_{io}^g 表示 x_i 的第 o 个缺失值的估计值，对应于近邻 y_h^g 的第 o 个完整的属性。因此，我们可以得到不完整样本 x_i 的 p 个完整样本 $\{x_i^p\ldots x_i^p\}$。

对样本 x_i 估计的 p 个版本用用标准分类器分类。然而，不同版本估计值的准确性是不同的，这会导致其分类结果的可靠性不同。

2.2 折扣分类结果

对于样本的第 g 个版本 x_i^g 用训练集 Y 所训练得到的标准分类器 $\Gamma(.)$ 分类，得到的不完整样本 x_i^g 的分类结果定义为：

$$P_i^g=\Gamma(x_i^g|Y) \tag{11}$$

如果 $p=1$，也即样本 x_i 的 K 近邻均来自同一个类别，那么带有唯一估计值的不完整样本就会得到特定的分类结果，在这种情况下，样本 x_i 分配给结果中支持所属类别概率最大的那一类。

如果 $1<p\leqslant c$，也即样本 x_i 的 K 近邻来自多个类别，那么就会得到多个版本插补后样本的多个分类结果。在这种情况下，需要对多个版本的分类结果进行融合，可以选择简单有效的 DS 融合规则去处理这多个分类结果用于决策。然而，由于这不同的分类结果之间可能会存在冲突，DS 融合在处理高冲突信息时会得到不合理的结果。因此我们考虑使用折扣融合的方法，它将证据的一部分信息分配给完全未知类来减小证据间的冲突。

这多个版本的分类结果的可靠性是不同的，我们认为不完整样本与某一类的近邻越近，那么用这一类的近邻得到的估计值越可靠，从而得到分类结果越可靠，相对应的折扣系数也越小。因此对目标样本 x_i 的不同 p 个版本分类结果的折扣系数 β_i^g 定义为：

$$\beta_i^g=\frac{\delta_i^g}{\delta_{\max}^g} \tag{12}$$

其中：

$$\delta_i^g=e^{-\|x_i-\tilde{c}_g\|} \tag{13}$$

$$\tilde{c}_g=\frac{1}{N_g}\sum_{y_h^g\in\omega_g}y_h^g \tag{14}$$

$\tilde{c}$ 表示每个类的模拟的类中心，N_g 表示 K 近邻中属于类 ω_g 的近邻个数，也即用属于同一类的近邻求均值得到模拟类中心，这个模拟类中心 $\tilde{c}_g$ 与样本 x_i 之间的距离来表示这个类 ω_g 与该样本之间的距离，其中 $\delta_{\max}^g=max\{\delta_i^1,\ldots,\delta_i^p\}$。

我们用经典的证据折扣法根据折扣系数 β_i^g 来对每个分类结果进行折扣处理，定义如下：

$$\begin{cases}m_i^g(A)=\beta_i^gP_i^g(A),A\subset\Omega\\ m_i^g(\Omega)=1-\beta_i^g\sum_{A\subset\Omega}P_i^g(A)\end{cases} \tag{15}$$

基于可靠性的折扣规则通过修改证据能够有效地降低证据间的冲突，在这种情况下就可以用 DS 直接融合这些折扣后的证据。然而，由于数据缺失会带来不完整样本类别的不精确性，为了表征这种不精确性并且降低错误分类的风险，本文提出一种新的全局融合策略。

2.3 全局融合

这 p 个分类结果可能将样本目标分给不同的类别。由于折扣后的分类结果只有单类和由整个辨识框架表示的完全未知类，也即每个表示分类结果的 BBA 中只有单焦元和完全未知焦元。为此我们要确定样本目标 x_i 最有可能属于

的复合类。假设这 p 个分类结果支持样本 x_i 属于 q 个不同的类 $\{\omega_1,\ldots,\omega_r,\ldots,\omega_q\}$，那么根据这些分类结果所属类别将它们划分成 q 个不同的群组。对于样本目标 x_i 来说，假定有 s 个分类结果支持它属于 ω_r，然后定义如下函数来计算支持样本 x_i 分别属于这 q 个类最大的 $m(.)$ 值。

$$m_i(\omega_r)=\max\{m_{i1}(\omega_r),\ldots,m_{is}(\omega_r)\},\ r=1,\ldots,q \tag{16}$$

然后通过下式计算得到拥有最大信任值支持样本 x_i 属于的那个类 ω_{max}。

$$\omega_{max}=\arg\max_{\omega_r} m_i(\omega_r),r=1,\ldots,q \tag{17}$$

接着通过下式找到样本 x_i 最有可能属于的复合类中的单类。

$$\Lambda_i=\{\omega_{\max}\cup\omega_r\mid m_i(\omega_{\max})-m_i(\omega_r)\leqslant\alpha\} \tag{18}$$

最后，在 DS 融合的基础上，定义如下规则融合这多个版本的分类结果，如下所示：

$$m_i(A)=\begin{cases}\dfrac{1}{K}\displaystyle\sum_{\bigcap_{x=1}^{p}B_x=A}[m_i^1(B_g)\ldots m_i^p(B_p)], \text{ for } A\in\Omega \text{ with } |A|=1\\ \dfrac{1}{K}\displaystyle\prod_{\omega_r\in\Lambda_i} m_i(\omega_r), \text{ for } A=\Lambda_i \text{ with } |\Lambda_i|\geqslant 2\end{cases} \tag{19}$$

其中：

$$K=\sum_{\bigcap_{x=1}^{p}B_x=A}[m_i^1(B_g)\ldots m_i^p(B_p)]+\prod_{\omega_r\in\Lambda}m_i(\omega_r) \tag{20}$$

其中：α 为控制不精确率的参数，α 越大，就会计算越多的复合类，从而会增加分类结果中的不精确性，但同时也会有效降低错误分类的风险。在实际应用中，α 可以根据可接受的不精确率来选择，本文默认为 $\alpha=0.3$。在实际应用中，对于分配到复合类中的样本，可以根据其他的信息来进一步的准确划分它们的类别。

为了清楚直观地表示算法的基本原理，我们绘制了算法的流程，如图 1 所示。

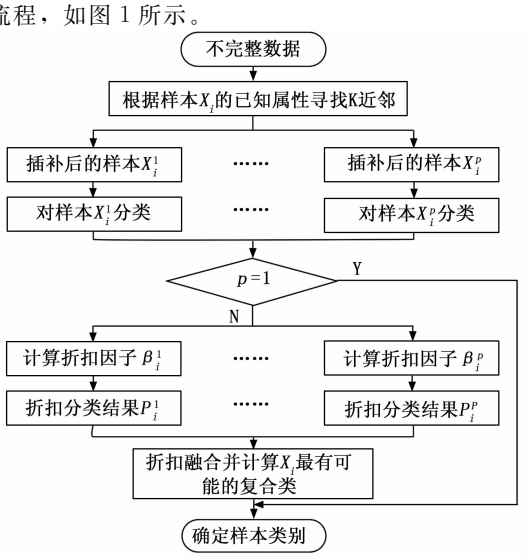

图 1　本文方法算法流程图

3　实验及结果分析

3.1　数据集

本文从 UCI 数据库(http://archive.ics.uci.edu/ml)选取了 7 个标准数据集进行测试。Ecoli（Ec）是关于蛋白质定位的数据集，Yeast（Ye）为预测蛋白质细胞定位位置数据集，Vehicle（Ve）是关于车辆轮廓特征的数据集，Wifi（Wi）是无线数据定位数据集，Satimage（Sa）数据集是关于卫星图像的像素值，Segment（Se）是一个图像分割数据集，Connectionist（Co）是一个关于英式元音识别数据集。这些数据的类别在 3～11 类之间，属性个数在 7～36 之间，样本数在 255～6 435 之间，因此这些数据是 UCI 数据库中比较有代表性的数据，这样也能够更加全面并充分地验证不同算法在处理不同类型数据集性能。这些数据的基本信息，包括数据集的名称和简写、类别数、属性数以及样本数，如表 1 所示。

表 1　数据集基本信息

名称(简写)	类别数	属性数	样本数
Ecoli(Ec)	3	7	255
Yeast(Ye)	3	8	1 136
Vehicle(Ve)	4	18	846
Wifi(Wi)	4	7	2 000
Satimage(Sa)	6	36	6 435
Segment(Se)	7	19	2 310
Connectionist(Co)	11	10	528

3.2　对比实验

本文方法用以上的真实数据集分别与 KNNI[10]、FCMI[11]、LLA[18]、PCC[19]算法进行对比分析。同时采用了 K－NN、EK－NN 和决策树 DT 三种标准分类器进行实验。在本文方法算法中设置不精确率的阈值 α=0.3，近邻数 K=11。

实验中，我们随机选取每个数据集的一半数据作为训练集，另一半数据作为测试集，然后对测试集进行随机缺失处理（MCAR）。每个测试样本有 γ 个缺失属性值，随机分布在样本的各个属性上。实验以分类器在测试集上的最终分类误分率 Re 和不精确率 Ri 作为评价标准，$Re=N_e/N$，$Ri=N_i/N$，其中 N 表示样本数量，N_e 表示错误分类的样本数量，N_i 表示分配到复合类的样本数量。Re 是用于评估分类结果中错误分类样本所占比重，Re 值越小说明误分的样本越少，算法性能越好。Ri 是用于评估分类结果中划分到复合类的样本所占比重，该值越大，说明划分到复合类中样本越多，这样并不利于决策，因此根据具体实际应用的要求，该值应在一个可接受范围内。

不同方法用 K－NN 分类器分类后的结果如表 2 所示。由于传统的概率框架下的方法 KNNI、FCMI 和 LLA 得到的是确切的类别输出，因此只有错误率，而 PCC 和本文方法在证据框架下能得到不精确输出表征不确定性，因此存在有不精确率。从实验结果可以看出，本文方法得到了比 KNNI、FCMI、LLA 和 PCC 更低的误分率。当每个测试样

表 2 不同算法用 K－NN 的分类结果

数据	γ	KNNI	FCMI	LLA	PCC	本文方法
		Re	*Re*	*Re*	{*Re*,*Ri*}	{*Re*,*Ri*}
Ec	1	10.94	10.94	10.94	{9.38,7.03}	{9.38,6.25}
	2	10.94	10.94	9.38	{10.16,7.03}	{9.38,10.94}
	3	16.41	18.75	17.19	{16.41,8.59}	{10.16,21.88}
Ye	1	41.90	42.43	42.08	{39.61,6.34}	{34.68,16.73}
	2	45.25	43.66	42.25	{38.73,10.74}	{31.51,28.35}
	3	45.42	44.54	44.54	{39.08,16.55}	{30.28,36.44}
Ve	2	42.79	44.21	41.61	{42.55,5.67}	{40.43,5.20}
	4	41.13	43.26	39.95	{39.95,12.29}	{37.83,7.33}
	6	40.90	47.75	43.03	{44.92,13.71}	{40.43,6.38}
Wi	1	2.50	4.70	2.50	{2.10,1.70}	{2.10,0.70}
	2	4.20	12.80	5.00	{3.60,3.60}	{3.00,3.20}
	3	7.30	21.30	8.40	{8.20,7.70}	{4.80,5.80}
Sa	4	11.37	11.62	11.31	{10.04,3.08}	{11.00,1.65}
	8	11.44	14.51	11.47	{10.13,4.16}	{10.57,2.11}
	12	11.68	19.17	12.06	{11.47,4.94}	{10.53,2.39}
Se	2	11.00	15.93	10.65	{10.74,2.08}	{10.39,2.51}
	4	12.90	25.02	12.03	{10.91,5.97}	{10.13,4.24}
	6	14.72	38.10	14.29	{14.72,6.41}	{11.77,4.16}
Co	1	32.12	36.97	30.91	{31.92,13.94}	{24.24,15.35}
	2	32.12	41.82	32.32	{34.34,20.61}	{24.85,20.00}
	3	33.74	50.51	34.55	{38.38,23.03}	{26.06,25.86}

表 3 不同算法用 EK－NN 的分类结果

数据	γ	KNNI	FCMI	LLA	PCC	本文方法
		Re	*Re*	*Re*	{*Re*,*Ri*}	{*Re*,*Ri*}
Ec	1	8.59	10.94	9.38	{11.72,3.13}	{6.25,7.03}
	2	10.94	10.94	13.28	{12.50,7.81}	{6.25,17.97}
	3	14.84	15.63	12.50	{16.41,9.38}	{6.25,28.91}
Ye	1	43.66	43.84	44.37	{41.55,2.82}	{40.85,6.34}
	2	46.65	48.77	45.25	{44.19,4.05}	{40.14,11.97}
	3	47.89	48.77	46.83	{46.65,6.69}	{40.32,17.43}
Ve	2	38.77	40.90	38.77	{39.72,0.95}	{38.53,3.78}
	4	37.35	40.90	37.59	{42.55,0.71}	{35.46,9.22}
	6	39.72	47.99	43.50	{46.10,0.24}	{34.52,14.42}
Wi	1	2.60	3.70	2.40	{2.10,2.60}	{2.10,1.00}
	2	4.00	11.80	4.40	{4.10,4.30}	{3.20,3.00}
	3	7.10	19.10	8.40	{8.90,8.20}	{4.90,6.30}
Sa	4	10.53	11.16	10.53	{10.01,0.96}	{10.10,1.52}
	8	10.60	14.48	10.78	{10.47,0.99}	{9.07,2.45}
	12	11.62	19.08	11.62	{10.85,1.93}	{9.20,2.83}
Se	2	7.88	13.68	7.45	{7.97,1.56}	{5.89,4.59}
	4	9.18	26.06	9.09	{11.52,4.24}	{6.41,7.10}
	6	12.12	38.53	12.03	{14.55,5.02}	{6.23,8.66}
Co	1	13.33	19.39	12.93	{14.14,4.04}	{10.51,3.84}
	2	19.60	32.93	19.80	{20.40,11.72}	{12.53,11.72}
	3	22.42	40.40	23.23	{26.67,14.75}	{11.72,24.85}

本中缺失值的个数（γ）增加时，所有分类器的错误率也会增加，这是由于数据缺失会影响分类的性能，缺失的数据越多，分类性能也就越差。在本文方法中，一部分难以划分类别的样本被分配到复合类中，这能够表征缺失值引起的类别的不确定性，同时这种谨慎的决策方式能够有效降低错误分类的风险。

由于本文方法以及对比方法都是基于插补的不完整数据分类方法，也即先补全缺失数据，再用能够分类完整样本的基础分类器分类。为了研究使用不同分类器情况下不同方法的性能，这里分别使用 EK－NN 和 DT 作为基础分类器分类不同数据集，实验结果如表 3 和表 4 所示。从实验结果可以看出，使用 EK－NN 分类器相较于使用其他两种分类器的分类结果更优，当然这主要由分类器本身的分类性能决定的。当然，虽然使用不同分类器得到的分类结果不同，但是本文方法在使用不同分类器情况下的整体分类性能要优于其他方法。根据实验结果，在实际应用中，本文方法在执行过程中推荐使用 EK－NN 或者 K－NN 分类器。

3.3 算法参数对分类性能的影响

本文算法主要有两个参数：K 和 α。其中 K 用于表示用于估计缺失值的近邻个数；α 用于判别本文算法中待测不完整样本是否会被划分到复合类的阈值，因此它也能够用于调节本文算法中错误率和不精确率。本小节主要是通过实验研究这两个参数对本文算法以及其他方法的影响。

3.3.1 K 值的影响

由于这些对比方法中只有 KNNI、LLA 以及本文方法

表 4 不同算法用 DT 的分类结果

数据	γ	KNNI	FCMI	LLA	PCC	本文方法
		Re	*Re*	*Re*	{*Re*,*Ri*}	{*Re*,*Ri*}
Ec	1	17.97	17.19	16.41	{16,41,7.03}	{14.06,7.81}
	2	17.97	21.88	21.88	{14.06,9.38}	{9.38,22.66}
	3	22.66	20.31	22.66	{17.19,13.28}	{14.06,29.69}
Ye	1	46.13	48.24	47.36	{41.73,8.80}	{40.14,14.79}
	2	48.59	50.00	47.89	{38.91,14.26}	{35.21,28.17}
	3	52.11	54.23	53.52	{41.20,18.49}	{35.56,36.27}
Ve	2	30.73	34.52	31.68	{34.28,12.06}	{23.64,14.66}
	4	33.57	38.77	34.99	{36.64,14.89}	{23.64,20.57}
	6	34.99	44.92	35.70	{38.30,19.15}	{22.46,23.88}
Wi	1	3.80	16.20	4.00	{3.40,2.40}	{3.60,0.70}
	2	5.50	31.80	7.00	{5.20,4.80}	{4.20,2.70}
	3	7.90	43.00	11.20	{8.90,8.50}	{5.90,5.40}
Sa	4	15.16	24.30	15.04	{16.38,4.75}	{13.42,2.39}
	8	14.54	33.81	14.54	{18.99,7.86}	{12.49,3.76}
	12	15.07	41.92	15.35	{23.21,9.48}	{13.11,3.70}
Se	2	7.71	13.77	7.62	{8.14,2.16}	{5.54,3.81}
	4	10.65	21.73	10.48	{12.12,6.32}	{6.06,5.97}
	6	12.47	32.21	13.07	{13.77,7.97}	{5.80,6.49}
Co	1	37.58	42.42	36.36	{38.38,8.69}	{31.92,7.68}
	2	40.20	50.91	41.01	{42.02,16.97}	{30.30,14.95}
	3	45.05	57.17	48.48	{44.65,23.43}	{31.31,20.40}

存在该参数，因此这里选择部分有代表性的数据集来测试 K 值对这些算法性能的影响。如图 2 所示，X 轴表示分类

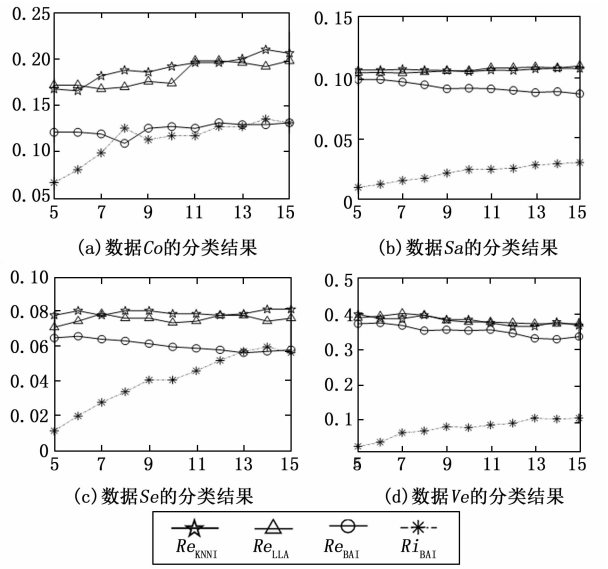

图 2　选择不同的 K 值时不同方法的分类结果

器 K 的个数从 5～15，Y 轴表示 KNNI、LLA 和本文方法算法不同分类器错误率和不精确率。从图中可以看到，与其他方法相比，本文方法的错误率更低，并且不精确率也在可接受的范围内。可以看到，随着 K 值的改变，不同方法的分类结果有所变化。本文方法受 K 值的影响比较小，这也证实了本文方法对于 K 值的选择具有一定的鲁棒性。

3.3.2　阈值 α 的影响

由于阈值 α 只存在于本文方法中，因此这里选择两个代表性的数据来研究阈值 α 对本文方法的影响，图 3 显示的是阈值 α 取不同值时本文方法的分类正确率和不精确率，其中 X 轴表示 α 的不同取值，Y 轴表示本文方法的错误率和不精确率。从图中可以看到随着 α 值的增大，错误率在降低的同时不精确率在逐渐升高。α 的值过小会导致很多样本被错误分类。当然，α 的值并不是越大越好，因为这样会使大量样本分配到复合类，这并不利于决策分析。在实际应用中，α 可以根据可接受的不精确率取值。

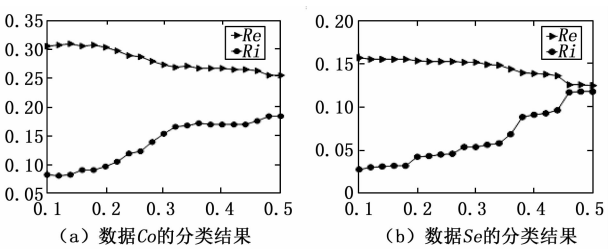

图 3　选择不同阈值 α 时本文方法的分类结果

4　结束语

本文提出了一种不完整数据智能分类算法，该方法通过自适应插补来提高估计精度，同时在证据推理框架下将难以划分类别的不完整样本分配到相应的复合类，这样能够有效地表征缺失值带来的不确定性，同时降低错误分类的风险。虽然实验结果证实了本文方法分类不完整数据的有效性，但是本文方法的插补是基于 K 近邻的，这需要大量的计算，在未来的工作中会考虑研究一种快速的 K 近邻插补技术，在保证估计精度的同时降低算法计算 K 近邻所产生的计算量。

参考文献：

[1] 魏茂胜. 数据挖掘中的分类算法综述 [J]. 网络安全技术与应用，2017 (6)：65－66.
[2] 周　頔. 基于人工智能技术的不完备信息系统智能诊断方法研究 [J]. 计算机测量与控制，2018，26 (9)：5－8.
[3] 王惠宇，顾苏杭. 挖掘数据模式结构信息的混合数据分类方法 [J]. 计算机测量与控制，2019，27 (4)：190－197，217.
[4] 宗　威，吴　锋. 大数据时代下数据质量的挑战 [J]. 西安交通大学学报（社会科学版），2013，33 (5)：38－43.
[5] 李　欢，王士同. 基于 SVM 和多观测样本的相似不完整数据分类 [J]. 控制与决策，2015，30 (7)：1207－1213.
[6] 赵　亮，陈志奎，张清辰. 基于分布式减法聚类的不完整数据填充算法 [J]. 小型微型计算机系统，2015，36 (7)：1409－1414.
[7] RUBIN D B. Inference and missing data [J]. Biome－trika，1976，63：581－592.
[8] RODERICKJ A. Test of missing completely at random for multivariate data with missing values [J]. Journal of the American Statistical Association，1988，83 (404)：1198－1202.
[9] 张松兰，王　鹏，徐子伟. 基于统计相关的缺失值数据处理研究 [J]. 统计与决策，2016 (12)：13－16.
[10] 于力超，金勇进，王　俊. 缺失数据插补方法探讨——基于最近邻插补法和关联规则法 [J]. 统计与信息论坛，2015，30 (1)：35－40.
[11] HATHAWAY R J，BEZDEK J C. Fuzzy c－means clustering of incomplete data [J]. IEEE Trans. Systems，Man，and Cybernetics，Part B，2001，31 (5)：735－744.
[12] LITTLE R J，RUBIN D B，Statistical analysis with missing data [M]. Hoboken，NJ，USA：Wiley，2014.
[13] STEKHOVEN D J，BüHLMANN P. Miss forest － nonparametric missing value imputation for mixed－type data [J]. Bioinformatics，2012，28：112－118.
[14] 张山鹰，潘　泉，张洪才. 证据推理冲突问题研究 [J]. 航空学报，2001 (4)：369－372.
[15] 尹慧琳，王　磊. D－S 证据推理改进方法综述 [J]. 计算机工程与应用，2005 (27)：22－24.
[16] SHAFER G. A mathematical theory of evidence [M]. Princeton Univ. Press，1976.
[17] Denoeux T，Logistic regression，neural networks and Dempster－Shafer theory：a new perspective [J]. Knowl. －Based Syst.，2019 (176)：54－67.
[18] 段中兴，梅思雨. 基于数据挖掘的建筑能耗异常检测研究 [J]. 计算机测量与控制，2020，28 (7)：253－259.
[19] LIU Z G，PAN Q，MERCIER G，et al. A new incomplete pattern classification method based on evidential reasoning [J]. IEEE Transactions on Cybernetics，2015，45 (4)：635－646.
[20] DAI J，HU H，HU Q. Locally linear approximation approach for incomplete data [J]. IEEE Transactions on Cybernetics，2018，48 (6)：1720－1732.