数据集分类可用性评估的置信区间方法

谈询滔 顾依依 阮 彤 袁玉波
(华东理工大学计算机科学与工程系 上海 200237)

摘 要 如何有效评价训练数据集的可用性,一直是困扰智能分类系统应用的难点问题。针对机器学习领域的数据分类问题,提出了一种基于区间分析和信息粒化的数据集分类可用性的评估方法,用于评价数据集的可分程度。该方法将待评估的数据集定义为分类信息系统,提出了分类置信区间的概念,通过区间分析进行信息粒化。在此信息粒化策略下,定义分类可用性的数学模型,并进一步给出单个属性以及整体数据集的分类可用性的计算方法。选择18个UCI标准数据集作为评估对象,给出了部分数据集分类可用性的评估结果,并且选取3种分类器对所选数据集进行分类实验,最终通过对上述实验结果的分析证明了该评估方法的有效性和可行性。

关键词 数据可用性,分类系统,区间分析,信息粒化,分类可用性
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.01.012

Confidence Interval Method for Classification Usability Evaluation of Data Sets

TAN Xun-tao GU Yi-yi RUAN Tong YUAN Yu-bo
(Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract It is always a difficult problem to evaluate the usability of training data sets effectively, which hinders the application of intelligent classification systems. Aiming at the issue of data classification in the field of machine learning, based on interval analysis and information granulation, this paper proposed an evaluation method of data classification usability to measure the separability of data sets. In this method, dataset is defined as the classification information system, and the concept of classification confidence interval is put forward, then the information granulation is carried out by interval analysis. Under this information granulation strategy, this paper defined the mathematical model of classification usability, and further gave the calculation method of the classification usability for single attribute and the total data set. In this paper, 18 UCI standard data sets were selected as evaluation objects, the evaluation results of classification usability were given, and 3 classifiers were selected to classify the above data sets. Finally, the effectiveness and feasibility of this evaluation method are verified by the analysis of experimental results.

Keywords Data usability, Classification system, Interval analysis, Information granulation, Classification usability

1 引言

早在20世纪60年代,数据分类问题就已成为有监督学习中的一个重要研究课题。数据分类可以被定义为根据原始数据中存在的规律对未知实例的类别进行预测的技术[1]。在机器学习领域,来自全世界的专家和学者提出了许多经典的分类算法,但是所提出的方法通常假定数据是高度可用的[2-3],很少考虑数据的可用性问题。然而,作为信息传递的载体,数据是所有科学研究的基础,如果原始数据中存在可用性问题,则将对分类算法的效果造成不良的影响。

近年来各行业的数据规模迅速扩大,在大数据背景下,不仅促进了数据驱动的智能分类系统的蓬勃发展,同时也带来了许多挑战[4]。随着现实生活中劣质数据的不断增加,数据的可用性问题严重制约着大数据的运用价值[5]。而且,由低可用性数据驱动的分类系统可能会产生错误的决策,以至造成严重的损失。根据欧洲中央银行于2014年的一项调查显示:"单一欧元支付区每年由于数据不可用而导致的信用卡欺诈失察的损失达13.3亿欧元"[6]。因此,数据的可用性评估已经成为当前亟需解决的关键问题。数据可用性的评价准则是所使用的数据必须符合特定的需求[7],具体可以总结为以下5条基本性质:1)一致性,数据集合中的数据之间没有互相冲突的情况发生;2)完整性,数据集合中的数据包含足够丰富的信息来进行相关的计算任务;3)精确性,数据集合中的数据粒度符合要求,并且数据中没有明显错误;4)时效性,数据集

到稿日期:2018-06-08 返修日期:2018-07-14 本文受国家自然科学基金项目(61772201),上海市科委基金项目(16511101000),上海市科委基金项目(17DZ11011003)资助。
谈询滔(1994-),男,硕士生,主要研究方向为数据质量评估、数据挖掘和机器学习;**顾依依**(1994-),女,硕士生,主要研究方向为数据质量评估、数据挖掘和机器学习;**阮 彤**(1973-),女,博士,教授,主要研究方向为自然语言处理、数据质量评估等;**袁玉波**(1976-),男,博士,副教授,主要研究方向为数据质量评估、数据挖掘和机器学习等,E-mail:ybyuan@ecust.edu.cn(通信作者)。

合中的数据不存在陈旧和过时，并且应当与实际情况相符；5)实体同一性，多个数据集合中存在的同一实体描述应当是一致的。

粒计算是一种以粒为单位解决实际问题的思想，它旨在以多层次的模型和多视角的理解来描述现实世界。1997 年，Zadeh[8]第一次提出了粒计算(Granular Computing)的概念。对于粒计算的重要性，Lin[9-11]和 Yao[12-13]又进行了说明和强调，引发了国际上对粒计算的研究热潮。此后，许多不同领域的学者都开始关注和研究这个问题，粒计算逐渐成为了人工智能研究中一种新的研究方向[14]。粒计算是计算和处理复杂数据的一种全新的方法[15]，本文基于粒计算的思想展开了数据分类可用性问题的研究。

为了避免由低可用性数据驱动的分类系统所造成的错误决策，提高分类系统的学习效率，本文针对数据可用性的完整性和精确性指标，提出了一种数据集分类可用性评估的置信区间方法。该方法通过区间分析进行信息粒化，在本文定义的分类可用性模型的基础上，给出了单个属性以及整体数据集的分类可用性的计算方法。文中给出了 18 个 UCI 标准数据集的分类可用性的评估结果，并且选择了 3 种分类器对这些数据集进行分类实验，通过分析证明了该评估方法的有效性和可行性。

本文第 2 节引用了目前国内外对于数据可用性和粒计算理论的相关研究工作，介绍了本文所提方法的思想来源；第 3 节介绍了数据集分类可用性的数学模型；第 4 节给出了该评估算法的具体实现步骤；第 5 节展示了部分数据集的分类可用性评估结果，并进行了分类实验，通过图表等方式呈现并分析了实验结果；最后总结全文。

2　相关工作

目前，在数据集合的可用性方面已经有许多学者开展了相关研究，并取得了大量的研究成果[16]。下文将介绍数据可用性的 5 个重要方面的研究进展，并从不同的研究方法中得到启发，考虑如何对分类系统中训练数据集的可用性进行评估。

基于统计原理，Korn 等[17]提出了数据一致性错误的表示模型，Xiong 等[18]描述了表示数据不一致性的方式，并且提出了一种数据一致性的增强算法。对于数据一致性的判定方法，Miao 等[19]进行了较全面的研究。

Ma 等[20]提出了一种用于描述数据的完整性的规则系统。对于数据完整性的判定方法，Emran 等在文献[21]中对多种相关的定义和计算模型进行了综述，同时在文献[22]中给出了某些特殊数据集合相应的数据完整性判定。

Cao 等[23]提出了一种基于规则的数据精确性表示机制。对于数据精确性的判定方法，Zhang 等[24]使用均方误差作为衡量参数，提出了一种多模态数据集合的精确性模型，该方法将数据分为可量度型、可比性型、可分类型[25]3 种类型，并且针对不同数据类型的特征分别建立了数学模型，最终组合得到了数据精确性的判定模型。

在同一个实体具有多个元组的前提下，Fan 等[26]提出了一种基于规则的数据时效性表示机制。对于数据时效性的判定方法，Li 等[27]进行了基于查询数据的相对时效性研究，提出了一种数据相对时效性的数学模型。

基于 EM 算法，Shen 等[28]实现了一种基于规则的实体同一性表示机制。对于数据同一性的判定方法，Li 等[29]研究了基于识别结果的实体同一性的判定。

上述研究成果都是基于数据集固有的可用性的。本文针对机器学习的分类问题这一应用场景，结合粒计算的思想提出了一种基于区间分析和信息粒化的数据集分类可用性的评估方法。该方法是一种多粒度的数据完整性和精确性的判定模型，可以用于评价数据集的可分程度。

粒计算是近年来的热门研究课题，钱宇华[30]对复杂数据的粒化机理进行了深入的研究，Skowron 等[31]给出了“粒”更广泛的定义。进行粒计算的第一步是确定采用哪种模型，然后根据相应的粒化策略来进行信息粒化。目前，国内外有 3 种经典的粒计算模型可供参考，即美国学者 Zadeh[8]提出的模糊集模型、波兰学者 Pawlak[32]提出的粗糙集模型和中国学者 Zhang 等[33]提出的商空间模型。

本文根据粗糙集模型中的信息粒化理论，提出了数据集分类可用性评估的置信区间方法。该方法构造了分类置信区间来描述分类信息粒，用分类置信区间引导的划分来表示粒空间，用粒度来描绘粒空间的单位大小。最终将数据集划分为不同粒度的粒空间，并在此基础上构造了分类可用性的评估模型。该方法对于样本选择和特征选择等数据预处理工作有着指导意义，并且其时间代价较低，可以提高分类系统的学习效率。

3　数据集分类可用性评估的模型

本节将介绍数据集分类可用性的数学模型及其相关的符号定义。本文定义了一种分类信息系统，提出了分类置信区间的概念，并且定义了一种基于分类置信区间的信息粒化策略。在此策略下，本文分别给出了单个属性和整体数据集的分类可用性的评估方法。

本文提出的分类可用性记为 CU，表示的含义是数据集对于分类系统的可用程度，相关定义如下。

定义 1(分类信息系统)　给定分类数据集 X，记为 $X=(U,A\cup C)$，其中 $U=\{x_1,x_2,\cdots,x_n\}$ 表示样本对象集合，$A=\{a_1,a_2,\cdots,a_m\}$ 表示信息属性集合，C 表示分类属性。可以用五元组 S 表示一个分类信息系统，记为 $S=(X,V_a,V_c,f,d)$。其中，V_a 是信息属性集合的值域；$f:U\times A\to V_a$ 是信息函数，给定 $a\in A$，对于 $\forall x\in U$，$f(x,a)\in V_a$ 成立；$V_c=\{c_1,c_2,\cdots,c_k\}$ 是分类标识的取值集合；$d:U\times A\to V_c$ 是分类函数，给定 $c\in V_c$，$a\in A$，对于 $\forall x\in U|c$，$d(x,a)=c$，$f(x,a)\in V_a(c)$ 成立，其中 $U|c$ 是样本集合 U 关于分类标识 c 的划分。

定义 2(分类置信区间)　给定 $c\in V_c$，$a\in A$，分类标识为 c 的样本在属性 a 上的分类置信区间，可记为 $I_a(c)=[x_a^-(c),x_a^+(c)]$。对于 $x\in U|c$，$\exists a\in A$，$f(x,a)\in I_a(c)$ 成立时，称样本 x 对分类系统是可用的。

若用 $O_a(c)$ 表示类标识为 c 的样本在属性 a 上的中心点，$O_a^-(c)$ 和 $O_a^+(c)$ 分别表示其相邻两个类别的样本中心点，其中 $O_a(c)$ 可由式(1)计算：

$$O_a(c)=\frac{1}{\|U|c\|}\sum_{x\in U|c}f(x,a) \tag{1}$$

则分类置信区间根据式(2)和式(3)计算：

$$x_a^-(c)=\begin{cases}m, & |O_a(c)-m|=\min\limits_{c\in V_c}\{|O_a(c)-m|\}\\ \dfrac{\beta*O_a^-(c)+\beta^-*O_a(c)}{\beta+\beta^-}+\varepsilon, & \text{else}\end{cases} \tag{2}$$

$$x_a^+(c)=\begin{cases}M, & |O_a(c)-M|=\min\limits_{c\in V_c}\{|O_a(c)-M|\}\\ \dfrac{\beta^+*O_a(c)+\beta^*O_a^+(c)}{\beta^++\beta}-\varepsilon, & \text{else}\end{cases} \tag{3}$$

其中，ε 代表类间间隔，β 代表类内散度。m 和 M 分别表示样本 x 在属性 a 上的最小值和最大值，如式(4)所示：

$$\begin{cases}m=\min\limits_{x\in U}f(x,a)\\ M=\max\limits_{x\in U}f(x,a)\end{cases} \tag{4}$$

则将属性 a 上所有分类置信区间的集合记为 I_a，如式(5)所示：

$$I_a=\bigcup_{c\in V_c}I_a(c)=\bigcup_{c\in V_c}[x_a^-(c),x_a^+(c)] \tag{5}$$

定义 3(基于分类置信区间的信息粒化策略) 对于给定 $a\in A$，可根据分类置信区间 $I_a(c)$ 将样本集合 U 划分为一系列分类置信粒 $G_a(c)=U|I_a(c)$，定义如下：

$$G_a(c)=\{x\in U|d(x,a)=c,f(x,a)\in I_a(c)\} \tag{6}$$

若考察对象为属性 a，则可以得到由分类置信区间集合 I_a 诱导的一个粒层 $G_a=U|I_a$，定义如下：

$$G_a=\bigcup_{c\in V_c}\{x\in U|d(x,a)=c,f(x,a)\in I_a(c)\} \tag{7}$$

G_a 由一系列分类置信粒构成，且有 $G_a=\{G_a(c)|c\in V_c\}$。将粒层 G_a 对样本空间 U 的覆盖率记为 $CU(a)$，称 $CU(a)$ 为属性 a 的分类可用度，可由式(8)计算：

$$CU(a)=\frac{|G_a|}{|U|}=\sum_{c\in V_c}\frac{|G_a(c)|}{|U|} \tag{8}$$

若考察对象为数据集 $X=(U,A\cup C)$，则可得到一个粗粒度的粒层 $G=\{G_a|a\in A\}$。将粒空间 G 的粒度记为 g，其与属性分类可用度 $CU(a)$ 的关系如下：

$$g=\frac{1}{|A|}\sum_{a\in A}\frac{|G_a|}{|U|}=\frac{1}{|A|}\sum_{a\in A}\sum_{c\in V_c}\frac{|G_a(c)|}{|U|}=\overline{CU(a)} \tag{9}$$

由此，评估任务可被转化为粒度为 g 的简单问题，本文将数据集 X 的分类可用性 $CU(X)$ 定义为：

$$CU(X)=\max_{a\in A}\{CU(a)\} \tag{10}$$

不难发现，$CU(X)$ 的取值范围为 $[0,1]$。当 $CU(X)=1$ 时，表示数据集 X 具有最大的可用性，用 X 作为训练集得到的分类模型具有最好的性能。

4 数据集分类可用性评估的算法

将本文提出的数据集分类可用性的评估算法记为 Evaluation，其主要包括以下 6 个步骤，对应的算法流程如图 1 所示。

1)输入原始数据集，并进行预处理；

2)统计数据的基本信息，构建分类信息系统；

3)基于类别划分初始的数据粒；

4)构造样本的分类置信区间；

5)基于分类置信区间，划分分类信息粒；

6)计算属性及数据集的分类可用性，输出结果。

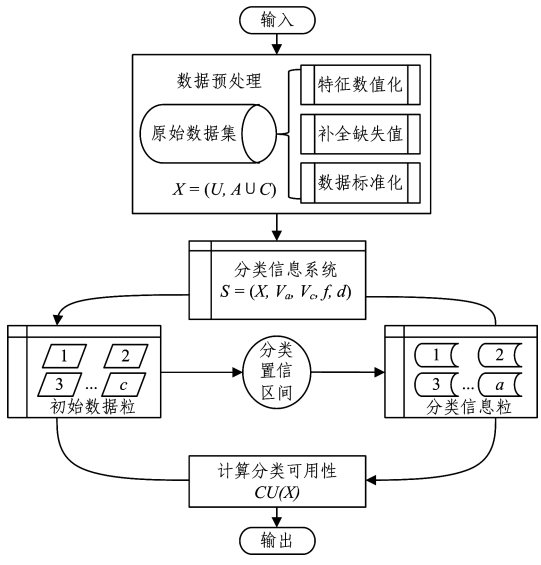


图 1 分类可用性评估的算法流程

Fig. 1 Algorithm flow of classification usability evaluation

Evaluation 算法可分为两个部分：1)评价每个属性的分类可用度；2)评价整个数据集的分类可用性。

4.1 属性的分类可用度评估

本节将详细描述上文中提到的 6 个步骤，并介绍如何评价单个属性的分类可用度。

步骤 1 对输入的原始数据集 X 进行预处理，具体包括以下内容：首先通过特征编码的技术将字符型特征数值化；然后利用均值插补法补全含有缺失值的记录；最后使用 robust_scale 的标准化方法使数据服从标准的正态分布。

步骤 2 统计预处理后数据集的基本信息，记样本个数为 n、特征维数为 m、类别数目为 k。将上述的分类数据集记为 $X=(U,A\cup C)$，其中 $U=\{x_1,x_2,\cdots,x_n\}$ 表示样本对象集合，$A=\{a_1,a_2,\cdots,a_m\}$ 表示信息属性集合，C 表示分类属性。则可得到一个分类信息系统，记为 $S=(X,V_a,V_c,f,d)$。

根据定义 1 可知，V_a 是信息属性集合的值域，$f:U\times A\rightarrow V_a$ 是信息函数，$V_c=\{c_1,c_2,\cdots,c_k\}$ 是分类标识集合，$d:U\times A\rightarrow V_c$ 是分类函数。

步骤 3 对于给定的 $c\in V_c$，可以得到样本集合 U 关于分类标识 c 的划分 $U|c$。称 $U|c$ 为一个数据粒，划分初始数据粒的方法可表示为：

$$U|c=\{x\in U|\forall a\in A,d(x,a)=c\wedge f(x,a)\in V_a(c)\} \tag{11}$$

已知 $V_c=\{c_1,c_2,\cdots,c_k\}$，则能得到基于类别划分的 k 个初始数据粒，它们之间互不相交且有如下关系：

$$U=\bigcup_{i=1,2,3,\cdots,k}(U|c_i) \tag{12}$$

步骤 4 对于给定的 $a\in A$，计算类标识为 c 的样本在每个属性 a 上的均值 $O_a(c)$ 和离差平方和 $SST_a(c)$，则可以得到初始数据粒的中心点和类内散度，其分别由式(13)和式(14)计算：

$$O_a(c)=\frac{1}{\|U|c\|}\sum_{x\in U|c}f(x,a) \tag{13}$$

$$SST_a(c)=\sum_{x\in U|c}[f(x,a)-O_a(c)]^2 \tag{14}$$

将各初始数据粒的中心点 $O_a(c)$ 升序排列，得到一个有

序集合 O_a，如式(15)所示：

$$O_a=\{O_a(c_i)\mid c_i\in V_c, i=1,2,3,\cdots,k\} \tag{15}$$

计算集合 O_a 中相邻元素间的加权平均值，将类内散度作为权重，则类间的置信边界如下：

$$O_a'(c_i)=(SST(c_{i+1})*O_a(c_i)+SST(c_i)*O_a(c_{i+1}))/(SST(c_{i+1})+SST(c_i)) \tag{16}$$

其中，$i=1,2,3,\cdots,k$，ε 为给定的极小数；计算样本集合 U 在属性 a 上取值的下确界和上确界，分别记为 $O_a'(c_0)$ 和 $O_a'(c_k)$，如式(17)和式(18)所示：

$$O_a'(c_0)=\min_{x\in U} f(x,a)-\varepsilon \tag{17}$$

$$O_a'(c_k)=\max_{x\in U} f(x,a)+\varepsilon \tag{18}$$

然后构造一个新的有序集合 O_a'，集合中的元素代表了各类别样本在属性 a 上的分类置信区间的上确界和下确界，O_a' 如式(19)所示：

$$O_a'=\{O_a'(c_i)\mid i=0,1,2,\cdots,k\} \tag{19}$$

因此根据定义 2，类标识为 c_i 的样本在属性 a 上的分类置信区间 $I_a(c_i)$ 等价于如式(20)所示的开区间。

$$I_a(c_i)=(O_a'(c_{i-1}),O_a'(c_i)),i=1,2,3,\cdots,k \tag{20}$$

我们认为当对于 $\forall x\in U\mid c, f(x,a)\in I_a(c)$ 成立时，样本具有最大的可分性。

步骤 5　对于给定的 $c\in V_c$，由给定属性 $a\in A$ 的分类置信区间 $I_a(c)$ 可以得到样本集合 U 的一个划分 $G_a(c)=U\mid I_a(c)$，称其为一个分类信息粒。已知 $V_c=\{c_1,c_2,\cdots,c_k\}$，则可得到 k 个信息粒，划分方法如式(21)所示：

$$G_a(c_i)=\{x\in U\mid d(x,a)=c, f(x,a)\in I_a(c_i)\} \tag{21}$$

若考查对象为属性 a，则根据分类置信区间的集合 I_a，可以得到由 I_a 诱导的一个粒层 $G_a=U\mid I_a$，定义如下：

$$G_a=\bigcup_{c\in V_c}\{x\in U\mid d(x,a)=c, f(x,a)\in I_a(c)\} \tag{22}$$

G_a 可以由 k 个分类信息粒构成，且有如下关系：

$$G_a=\{G_a(c_i)\mid c_i\in V_c, i=1,2,3,\cdots,k\} \tag{23}$$

步骤 6　对于分类信息系统 $S=(X,V_a,V_c,f,d)$ 中的任意样本 $x\in U$，属性 $a\in A$，定义函数 $g_a(x)$ 来说明 x 是否包含在基于属性 a 的粒层 G_a 中，如式(24)所示：

$$g_a(x)=\begin{cases}1, & \forall c\in V_c, x\in U\mid c\wedge f(x,a)\in I_a(c)\\ 0, & \text{else}\end{cases} \tag{24}$$

已知样本集合 U 的大小为 n，统计 U 中所包含于 G_a 的样本数目并记为 n_a，如式(25)所示：

$$n_a=\sum_{i=1}^{k}\sum_{x\in U\mid c_i} g_a(x) \tag{25}$$

属性 a 的分类可用度 $CU(a)$ 可由式(26)计算：

$$CU(a)=n_a/n \tag{26}$$

以上过程可以总结为属性的分类可用度评估算法 Evaluation-Attribute，具体伪代码如算法 1 所示。

算法 1　属性的分类可用度评估算法

```
输入：原始数据集 U
输出：属性对应的分类可用度列表 Usage_List
1. U=feature_encoder(U)
2. U=fill_null(U,mean(U))
3. U=robust_scaler(U)
4. X=(U,A∪C)
5. Subset_List=split_by(X,C)
6. Usage_List=empty()
7. FOR a IN A
8.     Usage=0
9.     O=empty(),SST=empty()
10.    FOR i,subset IN Subset_List:
11.        O(i)=mean(subset)
12.        SST(i)=var(subset) * count(subset)
13.    END FOR
14.    Interval=confidence_interval(O,SST)
15.    FOR sample IN X:
16.        IF sample∈Interval THEN:
17.            Usage++
18.        END IF
19.    END FOR
20.    Usage_List(a)=Usage / count(U)
21.  END FOR
22.  RETURN Usage_List
```

算法 1 为单个属性的分类可用度评估算法 Evaluation-Attribute 的伪代码。第 1－3 行对应了步骤 1 数据预处理的过程，函数 feature_encoder 对输入的原始数据集 U 进行特征编码，函数 fill_null 用均值填补了数据集的缺失值，函数 robust_scaler 对数据进行了去除异常值和标准化的处理。第 4 行对应了步骤 2 的统计数据集基本信息，其中，A 为属性集合，C 为分类标识。第 5－6 行对应了步骤 3 划分数据粒的过程，函数 split_by 通过类标识来划分子集。第 7－9 行循环处理数据集的每个属性，并返回其分类可用度。第 10－14 行对应了步骤 4 的计算过程，函数 confidence_interval 根据子集的中心点和类内散度来构造分类置信区间。第 15－21 行对应了步骤 5 和步骤 6，计算每个属性的分类可用度。

4.2　数据集的分类可用性评估

在 4.1 节中将每个属性作为独立的研究对象，本节将在此基础上对数据集整体的分类可用性进行评价。该方法存在两个重要的前提假设：1）属性之间没有强相关性；2）属性之间不具有几何拓补结构。

根据 3.1 节给出的分类信息粒的划分方法，对于任给属性 $a\in A$，已知 $V_c=\{c_1,c_2,\cdots,c_k\}$ 可以得到 k 个信息粒，并且求得属性 a 的分类可用度 $CU(a)$。对于数据集 $X=(U,A\cup C)$，将 A 中属性的最大分类可用度作为数据集 X 的分类可用性 $CU(X)$，如式(27)所示：

$$CU(X)=\max_{a\in A}\{CU(a)\} \tag{27}$$

其中，分类可用度 $CU(a)$ 代表了属性 a 对于分类模型的重要程度，数据集的分类可用性 $CU(X)$ 代表了数据集 X 对于分类模型的可分程度。本文设计了子算法 Evaluation-Dataset 来实现以上过程，具体伪代码如算法 2 所示。

算法 2　数据集的分类可用性评估算法

```
输入：属性对应的分类可用度列表 Usage_List
输出：原始数据集的分类可用性 CU
1. CU=0
2. FOR usage IN Usage_List:
3.     IF usage>CU THEN:
```

```
4.      CU=usage
5.  END IF
6.  END FOR
7. RETURN CU
```

算法 2 为数据集整体的分类可用性评估算法 Evaluation-Dataset 的伪代码。第 1 行定义了数据集的分类可用性 *CU*。第 2－6 行遍历了 Usage_List，找出了所有属性中最大的分类可用度。第 7 行输出数据集的分类可用性。

4.3 算法的复杂度分析

算法 1 在构造分类置信区间的过程中，分别循环遍历了 m 个属性和 k 个数据子集，代价为 $O(m\times k)$，其中 m 为数据集的属性数，k 为类别数。在属性的分类可用度的过程中，循环遍历了 n 个样本，计算代价为 $O(m\times n)$，其中 n 为数据集的样本数。算法 2 在计算数据集整体的分类可用性的过程中，循环遍历了 m 个属性的分类可用度，计算代价为 $O(m)$，其中 m 为数据集的属性数。

根据以上分析，算法 Evaluation 的总计算代价为 $O(m\times(k+n+1))$。通常情况下，属性数 m 和类别数 k 均为远小于样本数 n 的常数。因此，当数据规模较大时，算法的时间复杂度可归纳为 $O(n)$。

5 实验与结果

本节从两个方面进行分类可用性的评估实验。首先评价每个数据集每个属性的分类可用度，并且给出了数据集整体的分类可用性。然后选取 3 种分类器对这些数据集分别进行了分类实验，将数据集的分类可用性和分类模型的性能进行了对比，并分析了该方法的有效性和可行性。

5.1 数据集介绍

本研究所使用的数据均来源于 UCI 数据库，该平台积累了大量来自不同领域的标准数据集，覆盖了机器学习中的分类、聚类、回归等主要问题，是目前学术界广泛使用的数据库。本文从 UCI 机器学习库中选择了 18 个用于分类任务的训练数据集，在经过预处理后统计了它们的样本数(*Instance*)、属性数(*Attribute*)和类别数(*Class*)等信息，如表 1 所列。

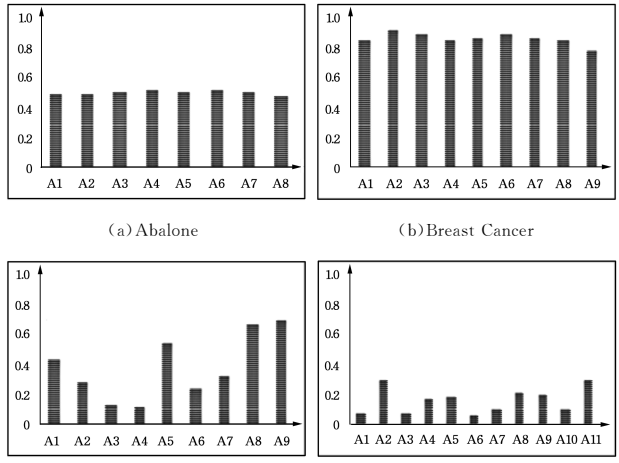
表 1 实验数据集的描述

Table 1 Description of experimental data sets

Data Set	*Instance*	*Attribute*	*Class*
Abalone	4177	8	3
Breast Cancer	699	9	2
Buzz Twitter	140707	77	2
Car Evaluation	1728	6	4
Contraceptive Method	1473	9	3
Heart Disease	270	13	2
Ionosphere	351	34	2
Iris	150	4	3
MAGIC Gamma	19020	10	2
Shuttle	14500	9	7
Spambase	4601	57	2
Teaching Assistan	151	5	3
User Knowledge	403	5	5
Wilt	4839	5	2
Wine Red	1599	11	6
Wine White	4898	11	7
Yeast	1484	8	10
Zoo	101	16	7

5.2 分类可用性评估结果

在实验中，我们对这些数据集的每个属性都评价了其分类可用度。由于结果较多，本文选取了 4 个具有代表性的数据集展示其评价结果，分别是 Abalone，Breast Cancer，Shuttle，Wine White，如图 2 所示。



(a) Abalone (b) Breast Cancer

(c) Shuttle (d) Wine White

图 2 部分数据集上的评估结果

Fig. 2 Evaluation results for partial data sets

图 2 展示了上述 4 个数据集中每个属性的分类可用度。图 2(a)所示的 Abalone 代表了分类可用性一般的数据集，每个属性的可用度都在 0.5 左右。图 2(b)所示的 Breast Cancer 代表了分类可用性较好的数据集，其中几乎所有属性的可用度都大于 0.7。图 2(c)所示的 Shuttle 代表了分类可用性差异较大的数据集。图 2(d)所示的 Wine White 代表了分类可用性较差的数据集，其中几乎所有属性的可用度都在 0.3 以下。

对于其余数据集，本文记录了其属性分类可用度的最大值(*max*)、最小值(*min*)和平均值(*avg*)，如表 2 所列。

表 2 其余数据集的分类可用度评估结果

Table 2 Classification availability evaluation results of other data sets

Data Set	*max*	*min*	*avg*
Abalone	0.5324	0.4846	0.5114
Breast Cancer	0.9270	0.7897	0.8728
Buzz Twitter	0.9219	0.2373	0.7525
Car Evaluation	0.5775	0.2975	0.4259
Contraceptive Method	0.4352	0.2695	0.3740
Heart Disease	0.7630	0.4667	0.6541
Ionosphere	0.7863	0.0000	0.5845
Iris	0.9600	0.5467	0.8000
MAGIC Gamma	0.7375	0.4725	0.5742
Shuttle	0.6921	0.1303	0.3865
Spambase	0.7768	0.3367	0.5751
Teaching Assistan	0.4106	0.3444	0.3881
User Knowledge	0.7829	0.2093	0.3891
Wilt	0.7628	0.5114	0.5993
Wine Red	0.3984	0.0775	0.2046
Wine White	0.3030	0.0635	0.1670
Yeast	0.5175	0.1119	0.2276
Zoo	1.0000	0.1089	0.6120

从以上的实验结果中可以发现，不同数据集的分类可用性存在很大的差别。事实上，用这些数据集训练出来的分类

模型的性能也存在巨大差异。

5.3　分类可用性与分类效果

为了验证本文提出的分类可用性评估方法用于评价训练数据集质量的可行性，本节选择了 3 种分类器来对文中的 18 个 UCI 数据集进行训练。在每个数据集中随机划分 80% 的样本作为训练集，剩余 20% 的样本作为测试集。在训练集上评估分类可用性以及构建分类模型，并在测试集上验证其分类模型的性能。本文将数据集分类可用性得分与对应分类模型的性能评分的均方误差（MSE），作为评判该方法效率的度量标准，MSE 的计算公式如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(CU_i-PI_i)^2 \tag{28}$$

其中，n 为实验数据集的个数，CU 为数据集的分类可用性，PI 为对应分类模型的性能指标。

本文的实验基于 Python 3.6 平台的 Scikit-Learn 机器学习库编程实现，选用的 3 种分类器分别是 GaussianNB[34]，CART[35] 和 LinearSVC[36]。GaussianNB 即 Gaussian Naive Bayes，是一种服从高斯分布的朴素贝叶斯算法。CART 即 Classification And Regression Tree，是一种经典的决策树算法，使用基尼系数来构造二叉树。LinearSVC 是基于台湾大学林智仁（Chih-Jen Lin）教授等提出的 LibSVM 而实现的支持向量机算法，LinearSVC 采用了线性核使其可以对大规模数据进行快速的分类或回归。

表 3 列出了 GaussianNB 分类器的实验结果。我们选择了准确率（Accuracy）、精度（Precision）、召回率（Recall）3 个指标来评价模型，在分类实验完成后，记录每个指标的平均值与标准差。同时，为了讨论数据集的分类可用性对所得分类模型性能的影响，表 3 也给出了每个数据集的分类可用性的评估结果。数据集的分类可用性及其 GaussianNB 分类模型的性能如图 3 所示。图 3 显示了在 GaussianNB 分类器上，数据集的分类可用性与其分类模型的性能之间的关系。其中，3 组柱体分别表示分类准确率、分类精度和召回率，而黑色曲线则表示相应数据集的分类可用性。

表 3　GaussianNB 分类器的实验结果

Table 3　Experimental results of GaussianNB classifier

（单位：%）

Data Set	*Accuracy*	*Precision*	*Recall*	*CU*
Abalone	51.88	50.54	53.63	52.53
Breast Cancer	95.86	95.03	96.15	92.42
Buzz Twitter	94.54	89.67	94.94	86.61
Car Evaluation	75.41	49.45	59.96	57.75
Contraceptive Method	47.11	48.68	49.51	43.52
Heart Disease	83.33	83.41	83.00	76.30
Ionosphere	86.60	87.45	83.43	74.93
Iris	95.33	95.84	95.33	94.00
MAGIC Gamma	72.69	72.05	64.68	73.65
Shuttle	85.84	47.90	73.78	71.61
Spambase	82.56	83.55	84.82	79.09
Teaching Assistan	48.95	48.67	49.05	41.06
User Knowledge	83.88	74.36	73.97	74.94
Wilt	88.10	63.34	71.15	74.60
Wine Red	52.73	31.52	29.87	45.03
Wine White	43.46	25.53	28.84	35.03
Yeast	14.82	35.39	40.01	25.74
Zoo	96.09	87.86	90.00	44.55

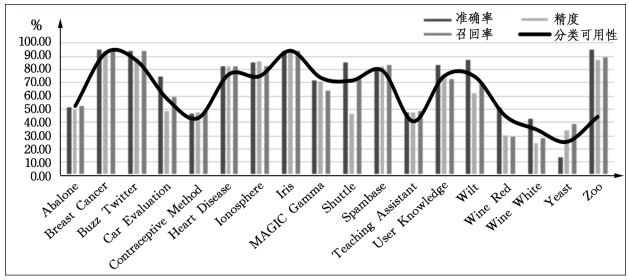


图 3　数据集的分类可用性及其 GaussianNB 分类模型的性能

Fig. 3　Classification usability of data sets and classification performance of GaussianNB model

从图中可以直观地看出，数据集的分类可用性与 GaussianNB 分类模型的性能在总体上呈现出了相同的趋势。通过计算得出，分类可用性与分类准确率、精度和召回率等指标的均方误差分别为 0.0803，0.0853 和 0.0754。这表明对于 GaussianNB 分类模型，本文提出的分类可用性评估方法可以有效地评价其训练数据集的质量。

表 4 列出了 CART 分类器的实验结果，包括准确率（Accuracy）、精度（Precision）、召回率（Recall）3 个指标的平均值与标准差，并且在表格的最后一列给出了数据集分类可用性 CU 的评估结果。同样地，图 4 显示了在 CART 分类器上，数据集的分类可用性与其分类模型的性能之间的关系。

表 4　CART 分类器的实验结果

Table 4　Experimental results of CART classifier

（单位：%）

Data Set	*Accuracy*	*Precision*	*Recall*	*CU*
Abalone	50.25	50.77	50.45	52.53
Breast Cancer	94.14	94.02	93.29	92.42
Buzz Twitter	95.02	92.06	92.26	86.61
Car Evaluation	79.40	68.37	70.34	57.75
Contraceptive Method	47.93	45.73	45.81	43.52
Heart Disease	72.96	74.22	72.67	76.30
Ionosphere	85.18	83.96	83.37	74.93
Iris	96.00	96.23	96.00	94.00
MAGIC Gamma	81.96	80.20	80.34	73.65
Shuttle	99.88	97.79	94.39	71.61
Spambase	88.63	88.34	88.22	79.09
Teaching Assistan	68.27	67.17	68.44	41.06
User Knowledge	86.34	78.18	77.33	74.94
Wilt	97.44	89.54	88.34	74.60
Wine Red	46.72	25.14	26.11	45.03
Wine White	41.08	21.74	21.93	35.03
Yeast	48.65	38.10	38.90	25.74
Zoo	95.24	86.19	90.00	44.55

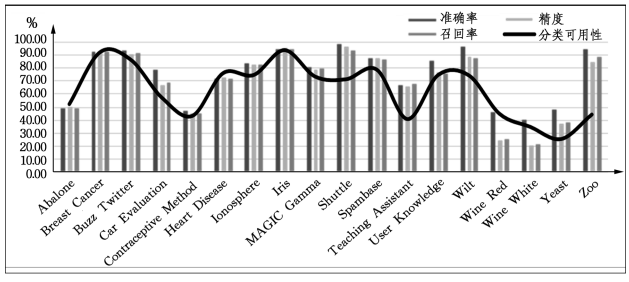


图 4　数据集的分类可用性及其 CART 分类模型的性能

Fig. 4　Classification usability of data sets and classification performance of CART model

不难发现，对于 CART 模型，数据集的分类可用性与分类模型的性能也存在着相似的规律，但在部分数据集上决策树模型的性能高于相应的分类可用性。计算得出，分类可用性与分类准确率、精度和召回率等指标的均方误差分别为 0.0939，0.0865 和 0.0840。表 5 列出了数据集在 LinearSVC 分类器上的准确率(Accuracy)、精度(Precision)、召回率(Recall)等实验结果，并附上其分类可用性 CU。根据实验结果可得一个组合图形，如图 5 所示，其显示了在 LinearSVC 分类器上，数据集的分类可用性与其分类模型的性能之间的关系。

表 5 LinearSVC 分类器的实验结果

Table 5 Experimental results of LinearSVC classifier

(单位：%)

Data Set	*Accuracy*	*Precision*	*Recall*	*CU*
Abalone	54.18	52.68	53.63	52.53
Breast Cancer	96.29	96.22	96.15	92.42
Buzz Twitter	90.86	90.42	94.94	86.61
Car Evaluation	77.07	43.88	59.96	57.75
Contraceptive Method	48.07	47.80	49.51	43.52
Heart Disease	75.93	81.56	83.00	76.30
Ionosphere	86.06	88.14	83.43	74.93
Iris	96.67	96.71	95.33	94.00
MAGIC Gamma	54.62	68.26	64.68	73.65
Shuttle	91.40	45.96	73.78	71.61
Spambase	78.72	83.49	84.82	79.09
Teaching Assistan	40.32	35.38	49.05	41.06
User Knowledge	77.33	51.53	73.97	74.94
Wilt	76.58	66.04	71.15	74.60
Wine Red	42.80	18.74	29.87	45.03
Wine White	35.55	19.56	28.84	35.03
Yeast	56.94	52.41	40.01	25.74
Zoo	96.09	87.76	90.00	44.55

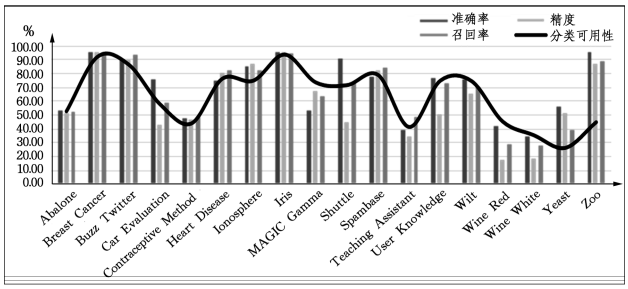


图 5 数据集的分类可用性及其 LinearSVC 分类模型的性能

Fig. 5 Classification usability of data sets and classification performance of LinearSVC model

从整体上看，数据集的分类可用性与其 LinearSVC 模型的性能表现出了高度的一致性。特别地，分类可用性与分类准确率、精度和召回率等指标的均方误差分别为 0.0776，0.0829 和 0.0754，比在其他模型上的表现更优。

通过对 GaussianNB，CART 和 LinearSVC 3 种分类器的实验结果进行分析，可以得出以下结论：

1)数据集的分类可用性与相应的分类模型的性能高度相关，特别是与召回率的均方误差最小。

2)数据集分类可用性对于支持向量机模型的指导意义最强，贝叶斯模型和决策树模型次之。

以上结果足以证明本文提出的数据集分类可用性评估方法的可行性，可将其作为评价分类训练数据集质量的一项指标，对于样本选择和特征选择等方面有着指导意义。

此外，对于属性间存在强相关性的数据集(如 Zoo)，本文提出的分类可用性无法准确评价其分类性能，具体表现为分类效果远高于分类可用性得分。对于这种情况，考虑如何消除属性的多重共线性，是我们未来的研究目标。

5.4 算法的运行时间

为了评价算法的效率，本文将分类可用性算法和以上 3 种分类算法的运行时间进行了对比，实验环境采用 3.3 GHz Inter Core i5 CPU 和 4 GB RAM 的主机上的 Python 3.6。图 6 给出了在 18 个 UCI 数据集上的实验结果，横坐标为数据集，纵坐标为算法的运行时间。

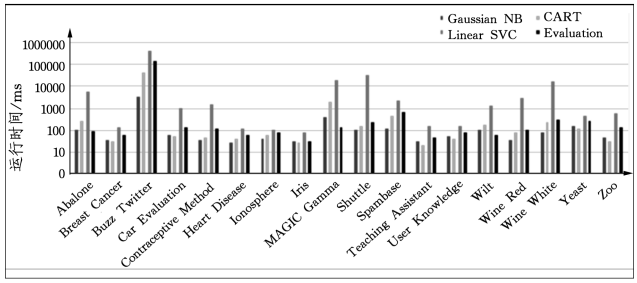


图 6 算法的运行时间

Fig. 6 Running time of algorithms

图 6 中，4 组柱体分别表示 GaussianNB 分类算法、CART 分类算法、LinearSVC 分类算法，以及分类可用性评估算法(Evaluation)的运行时间。从实验结果可以看出，本文提出的数据集分类可用性评估算法的运行时间与 GaussianNB 和 CART 的运行时间基本相当，但是远远少于 LinearSVC 的运行时间。因此，在构造复杂的分类模型之前，使用本文方法评估和选择训练数据集可以有效地缩减时间开销，进而提高分类系统的学习效率。

结束语 本文提出了数据集分类可用性评估的置信区间方法，针对机器学习的分类问题这一应用场景，评价训练数据集的可分程度。不同于以往宏观的数据可用性评价方法，本文着眼于分类训练数据集的可用性，结合粒计算的思想定义了分类可用性的评价指标。文中定义了分类信息系统以及分类置信区间的概念，并提出了一种基于区间分析的信息粒化策略，进而给出了分类可用性的评估模型。

本文根据单个属性以及整体数据集的分类可用性的计算方法设计了实验，选择了 18 个 UCI 标准数据集作为评估对象，给出了部分数据集分类可用性的评估结果。本文也选择了 3 种分类器对数据集进行分类实验，实验结果表明，数据集的分类可用性与相应的分类模型的性能高度相关，证明了本文所提出的分类可用性评估方法的有效性和可行性。此外，考虑属性间相关性对评价数据集分类可用性的影响，将是我们未来的研究方向。

参 考 文 献

[1] NOH Y K，ZHANG B T，LEE D D. Generative local metric learning for nearest neighbor classification[C]// Annual Conference on Neural Information Processing Systems. 2018：106-118.

[2] HOLLIFIELD T，SAILLET Y. Data quality assessment[J].

Communications of the Acm，2017，45(4)：211-218.

[3] CHEN Y C. Research on classification algorithm for weakly usable data[D]. Harbin：Harbin Institute of Technology，2014. (in Chinese)
陈懿诚. 弱可用数据上的分类算法研究[D]. 哈尔滨：哈尔滨工业大学，2014.

[4] LI J Z，WANG H Z，GAO H，et al. State-of-the-art of research on big data usability[J]. Journal of Software，2016，27(7)：1605-1625. (in Chinese)
李建中，王宏志，高宏，等. 大数据可用性的研究进展[J]. 软件学报，2016，27(7)：1605-1625.

[5] MERINO J，CABALLERO I，RIVAS B，et al. A data quality in use model for big data[J]. Future Generation Computer Systems，2016，63(C)：123-130.

[6] BAHNSEN A C，AOUADA D，STOJANOVICA. Feature engineering strategies for credit card fraud detection[J]. Expert Systems with Applications an International Journal，2016，51(C)：134-142.

[7] LI J，LIU X. An important aspect of big data：data usability[J]. Journal of Computer Research & Development，2013，50(6)：1147-1162.

[8] ZADEH L A. Toward a theory of fuzzy information granulation and itscentrality in human reasoning and fuzzy logic[J]. Fuzzy Sets & Systems，1997，90(90)：111-127.

[9] LIN T Y. Granular computing on binary relations I：data mining and neighborhood systems[J]. Rough Sets in Knowledge Discovery，1998，1(2)：165-166.

[10] LIN T Y. Granular computing on binary relations II：Rough set representations and belief functions[OL]. http://core. ac. uk/display/24652632.

[11] LIN T Y. Granular computing：Fuzzy logic and rough sets[M]// Computing with Words in Information/Intelligent Systems 1. Physica-Verlag HD，1999：183-200.

[12] YAO Y Y. Information granulation and rough set approximation [J]. International Journal of Intelligent Systems，2001，16(1)：87-104.

[13] YAO Y. Perspectives of granular computing[C]// IEEE International Conference on Granular Computing. IEEE，2005：85-90.

[14] YAO J T，VASILAKOS A V，PEDRYCZ W. Granular computing：Perspectives and challenges[J]. IEEE Transactions on Cybernetics，2013，43(6)：1977-1989.

[15] LI J，MEI C，XU W，et al. Concept learning via granular computing：A cognitive view point [J]. Information Sciences，2015，298(1)：447-467.

[16] BATINI C，CAPPIELLO C，FRANCALANCI C，et al. Methodologies for data quality assessment and improvement[J]. Acm Computing Surveys，2009，41(3)：16.

[17] KORN F，MUTHUKRISHNAN S，ZHU Y. Checks and balances：monitoring data quality problems in network traffic databases[C]// International Conference on Very Large Data Bases. VLDB Endowment，2003：536-547.

[18] XIONG H，PANDEY G，STEINBACH M，et al. Enhancing data analysis with noise removal[J]. IEEE Transactions on Knowledge & Data Engineering，2006，18(3)：304-319.

[19] MIAO D，LIU X，LI J. On the complexity of sampling query feed back restricted data base repair of functional dependency violations[J]. Theoretical Computer Science，2016，609：594-605.

[20] MA S，FAN W，BRAVO L. Extending inclusion dependencies with conditions[J]. Theoretical Computer Science，2014，515(1)：64-95.

[21] EMRAN N A. Data completeness measures[M]// Pattern Analysis，Intelligent Security and the Internet of Things. Springer International Publishing，2015：117-130.

[22] EMRAN N A，EMBURY S，MISSIER P. Measuring population-based completeness for single nucleotide polymorphism (SNP) databases[J]. Springer International Publishing，2014，551：173-182.

[23] CAO Y，FAN W，YU W. Determining the relative accuracy of attributes[C]// ACM SIGMOD International Conference on Management of Data. ACM，2013：565-576.

[24] ZHANG Y，WANG H，GAO H，et al. Efficient accuracy evaluation for multi-modal sensed data[J]. Journal of Combinatorial Optimization，2015，32(4)：1-21.

[25] ZHANG Y，WANG H，YANG Z，et al. Relative accuracy evaluation[J]. Plos One，2014，9(8)：e103853.

[26] FAN W，GEERTS F，WIJSEN J. Determining the currency of data[J]. Acm Transactions on Database Systems，2011，37(4)：1-46.

[27] LI M H，LI J Z，GAO H. Evaluation of data currency[J]. Chinese Journal of Computers，2012，35(11)：2348.

[28] SHEN W，LI X，DOAN A H. Constraint-based entity matching [C]// National Conference on Artificial Intelligence. AAAI Press，2005：862-867.

[29] LI L，LI J，GAO H. Evaluating entity-description conflict on duplicated data [M]. Springer-Verlag New York，Inc.，2016，31(2)：918-941.

[30] QIAN Y H. Granulating mechanism and data modeling of complex data[D]. Taiyuan：Shanxi University，2011. (in Chinese)
钱宇华. 复杂数据的粒化机理与数据建模[D]. 太原：山西大学，2011.

[31] SKOWRON A，WASILEWSKI P. Information systems in modeling inter active computation songranules [J]. Theoretical Computer Science，2010，412(42)：5939-5959.

[32] PAWLAK Z. Theoretical aspect of reasoning about data[M]// Rough Sets：Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers，1991.

[33] ZHANG Y P，ZHANG L，WU T. The representation of different granular worlds：A quotient space[J]. Chinese Journal of Computers，2004，27(3)：328-333.

[34] JIANG L，WANG S，LI C，et al. Structure extended multinomial naive bayes[J]. Information Sciences，2016，329(C)：346-356.

[35] SPEYBROECK N. Classification and regression trees[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery，2012，57(1)：243-246.

[36] FAN R E，CHANG K W，HSIEH C J，et al. LIBLINEAR：A library forlarge linear classification[J]. Journal of Machine Learning Research，2008，9(9)：1871-1874.