ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 4,Jul. 2016,pp. 838－844
DOI:10.16337/j.1004-9037.2016.04.024

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

# 不平衡数据集上的 Relief 特征选择算法

菅小艳　韩素青　崔彩霞

(太原师范学院计算机系,晋中,030619)

**摘　要**: Relief 算法为系列特征选择方法,包括最早提出的 Relief 算法和后来拓展的 ReliefF 算法,核心思想是对分类贡献大的特征赋予较大的权值;特点是算法简单,运行效率高,因此有着广泛的应用。但直接将 Relief 算法应用于有干扰的数据集或不平衡数据集,效果并不理想。基于 Relief 算法,提出一种干扰数据特征选择算法,称为阈值-Relief 算法,有效消除了干扰数据对分类结果的影响。结合 K-means 算法,提出两种不平衡数据集特征选择算法,分别称为 K-means-ReliefF 算法和 K-means-Relief 抽样算法,有效弥补了 Relief 算法在不平衡数据集上表现出的不足。实验证明了本文算法的有效性。

**关键词**: 特征选择;Relief 算法;ReliefF 算法;不平衡数据集

**中图分类号**: TP18　　**文献标志码**:A

## Relief Feature Selection Algorithm on Unbalanced Datasets

Jian Xiaoyan, Han Suqing, Cui Caixia

(Department of Computer Science, Taiyuan Normal University, Jinzhong, 030619, China)

**Abstract**: Relief algorithm is a series of feature selection method. It includes the basic principle of Relief algorithm and its later extensions reliefF algotithm. Its core concept is to weight more on features that have essential contributions to classification. Relief algorithm is simple and efficient, thus being widely used. However, algorithm performance is not satisfied when applying the algorithm to noisy and unbalanced datasets. In this paper, based on the Relief algorithm, a feature selection method is proposed, called threshold-Relief algorithm, which eliminates the influence of noisy data on classification results. Combining with the K-means algorithm, two unbalanced datasets feature selection methods are proposed, called K-means-ReliefF algorithm and K-means-relief sampling algorithm, respectively, which can compensate for the poor performance of Relief algorithm in unbalanced datasets. Experiments show the effectiveness of the proposed algorithms.

**Key words**: feature selection; Relief algorithm; ReliefF algorithm; unbalanced datasets

## 引　　言

分类系统的好坏取决于所利用的特征是否能够很好地反映所要研究的分类问题。特征选择即从输入的 $p$ 个特征中选择 $d<p$ 个使某种评估最优的特征[1]。一般而言,特征集中或多或少都包含着一些对

**基金项目**:国家自然科学基金(61273294)资助项目;山西省科技基础条件平台(2014091004-0104)资助项目。
**收稿日期**:2015-05-26;**修订日期**:2015-07-22

分类结果无贡献的特征,这些特征也称为冗余特征,对学习问题有很大的负面影响[2]。已有的研究结果表明,大多数分类系统设计所需的训练样本数随无贡献特征的增多成指数性增长。因此,特征选择对不同情况下的分类系统的设计都有着不可忽视的作用,选择好的特征不仅能降低特征空间的维数,加快算法的运行效率,而且选择合适的特征能获得更好的分类结果[3,4]。因此如何设计和获取特征在分类系统的设计中扮演着非常重要的角色,是机器学习的热点之一。

不平衡数据集是指一个数据集中,某一类的样本数目明显大于另一类的样本数目。由于在不平衡数据集上采样的悬殊,传统方法往往得不到很好的分类效果。目前用于处理不平衡数据集分类问题的方法可以分为两类:(1)从数据集入手,通过改变数据的分布,将不平衡数据变为平衡数据;或者通过特征选择,选出更能表达不平衡数据集的特征。(2)从算法入手,根据算法应用在不平衡数据上所体现的缺点,改进算法,提高正确率[5,6]。本文从数据集入手,通过改变随机选择样本的策略,利用 Relief 算法,研究有干扰数据集的分类问题,以及通过将 K-means 算法和抽样机制与 ReliefF 和 Relief 算法巧妙结合起来,利用 KNN 分类算法,研究不平衡数据集的分类问题。

## 1 Relief 和 ReliefF 算法

Relief 算法由 Kira 和 Rendell 于 1992 年提出,是一种针对二分类问题通过计算特征权重对特征进行选择的方法。它的基本思想是根据各个特征和类别的相关性赋予特征不同的权重,权重小于某个阈值的特征将被移除。算法从训练集 $D=\{(x_n,y_n)\}_{n=1}^{N}$ 中随机选择一个样本 $x_i$,然后从与 $x_i$ 同类的样本中寻找最近邻样本 $NH_i$,从与 $x_i$ 不同类的样本中寻找最近邻样本 $NM_i$,最后根据以下规则更新每个特征的权重[7,8],即

$$w(j)=w(j)+\frac{d(x_i(j),NM_i(j))}{m}-\frac{d(x_i(j),NH_i(j))}{m} \tag{1}$$

式中:$x_i(j)$表示样本 $x_i$ 关于第 $j$ 个特征的值;$d(\cdot)$表示距离函数,用于计算两个样本关于某个特征的距离;$m$ 是随机抽取样本的次数。

(1)如果 $x_i$ 和 $NM_i$ 关于某个特征的距离大于 $x_i$ 和 $NH_i$ 关于该特征的距离,说明该特征对区分同类和不同类的最近邻是有益的,则根据式(1)该特征的权重增加,反之,该特征的权重减少。

(2)距离函数定义

当特征为非数值型特征时,定义

$$d(x_i(j),NM_i(j))=\begin{cases}0 & x_i(j)\neq NM_i(j)\\ 1 & x_i(j)=NM_i(j)\end{cases} \tag{2}$$

当特征为数值型特征时,定义

$$d(x_i(j),NM_i(j))=\left|\frac{x_i(j)-NM_i(j)}{\max(j)-\min(j)}\right| \tag{3}$$

其中 $\max(j)$,$\min(j)$分别表示第 $j$ 个特征所取值中的最大值和最小值。

ReliefF 算法是 Knonenko 在 1994 年对 Relief 作的扩展,可以用于处理多类别问题[9]。ReliefF 算法每次从训练样本集中随机取出一个样本 $x_i$,然后从与 $x_i$ 同类的样本中找出 $x_i$ 的 $k$ 个近邻样本 $NH_i$,同时从每个与 $x_i$ 不同类的样本中也找出 $k$ 个近邻样本 $NM_i$,然后根据以下规则更新每个特征的权重[9],有

$$w(j)=w(j)+\sum_{c\neq \mathrm{class}(x_i)}\frac{\left(\frac{p(c)}{1-p(\mathrm{class}(x_i))}\sum_{j=1}^{k}d(x_i(j),NM_i(j))\right)}{mk}-\sum_{j=1}^{k}\frac{d(x_i(j),NH_i(j))}{mk} \tag{4}$$

式中:$\mathrm{class}(x_i)$表示样本 $x_i$ 所属的类别;$c$ 表示某个类别,$p(c)$表示类别 $c$ 的先验概率[10,11]。

## 2 干扰数据和不平衡数据分类

Relief 算法不仅算法原理比较简单,运行效率高,而且对数据类型没有限制,因此获得了广泛的应用。然而在实际应用中,该算法却存在一些局限性,比如不适合处理有干扰的数据,也不适合处理不平衡数据。

### 2.1 干扰数据分类问题

Relief 算法首先需要从样本集中随机选择一个样本作为训练样本,这种随机选择样本的方式,很可能取到不具有代表性的样本,或干扰样本,这意味着训练得到的特征可能不具有代表性,从而影响分类结果。针对上述问题,本文提出一种新的样本选取方法。首先计算样本集中每个样本与各自类中心的距离,将距离小于某一阈值的样本生成一个样本集 $D'$,然后再从 $D'$ 中随机选择一个样本,这样得到的样本可以比较好地排除干扰样本和不具有代表性的样本。

**算法 1**(阈值-Relief 算法)

输入:样本集 $D$,特征集 $F$,类别集 $C$,取样次数 $m$,阈值 $\delta$

输出:特征权重向量 $\boldsymbol{W}$

过程:

(a)特征权值初始化:$W(i)=0$;

(b)计算与每个类中心距离小于某一阈值 $\delta$ 的样本集合 $D'$;

(c)从过滤后的样本集 $D$ 中随机选取一个样本 $x_i$,同时从 $D$ 与 $x_i$ 同类的样本集中选取最近邻的样本 $NH_i$,从异类样本集中选取最近邻的样本 $NM_i$;

(d)利用式(1)更新特征权重;

(e)重复(c,d)$m$ 次;

(f)输出特征权重向量 $\boldsymbol{W}$。

### 2.2 不平衡数据分类问题

在不平衡数据集上,利用 Relief 算法选择的特征有可能出现权重值伪偏大。因为在随机选择样本时,样本数目较多的类别中样本被选中的概率要大,而样本较少的类别中样本被选中的概率要小,因而影响分类的效果。本文首先利用 K-means 算法对大类样本集中的样本进行聚类,然后在多类数据集上,分别基于 K-means-ReliefF 算法和 K-means-Relief 抽样算法解决数据不平衡带来的问题[12,13]。

#### 2.2.1 K-means-ReliefF 算法

首先将大类集聚类得到一些新的类别集,从而使得样本类别集基本平衡;然后采用 ReliefF 算法进行特征选择。

**算法 2**(K-means-ReliefF 算法)

输入:大类集 $L$,小类集 $S$($L$ 和 $S$ 同时表示类标签)和聚类数 $q$($q$ 为大类集与小类集中的样本数之比取整)

输出:特征权重向量 $\boldsymbol{W}$

过程:

(a)利用 K-means 算法对样本集 $L$ 聚类,得到 $q$ 个新类集 $S_1, S_2, \cdots, S_q$,与 $S$ 一起构成 $q+1$ 个基本平衡的数据类集;

(b)利用 ReliefF 算法在 $q+1$ 个类集上计算特征权重;

(c)输出特征权重向量 $\boldsymbol{W}$。

#### 2.2.2 K-means-Relief 抽样算法

首先将大类集聚类,得到一些新类,在每个类内抽取样本,组成新的集合,代表大类集,然后采用

Relief 算法计算特征权重[14]。

**算法 3**(K-means-Relief 抽样算法)

输入:大样本集 $L$,小样本集 $S$($L$ 和 $S$ 同时表示类标签)和聚类数 $q$($q$ 为大类集与小类集中的样本数之比取整)

输出:特征权重向量 $\boldsymbol{W}$

过程:

(a) 利用 K-mean 算法对样本集 $L$ 聚类,得到 $q$ 个新的类集 $S_1, S_2, \cdots, S_q$;

(b)在集合 $S_1, S_2, \cdots, S_q$ 中按集合大小比例抽取样本组成新类集 $L'$,使得 $|L'| = |S|$;

(c)利用 Relief 算法在 $L'$ 和 $S$ 上计算特征权重;

(d)输出特征权重向量 $\boldsymbol{W}$。

## 3 实验结果

### 3.1 数据集

本文采用的数据集均来自 UCI 数据集。为了实验的需要,在原始数据集上做了人为的修改。首先,为了得到有干扰样本的数据集,修改了原始数据集中一些样本的类别标签;其次,为了得到不平衡数据集,删除了原始数据集中的一些样本。具体数据见表 **1**。

**表 1 3 个数据集**

**Tab. 1 Three data sets**

| 数据集名称 | 原始数据 | | | 干扰点数 | 不平衡数据 | | |
|---|---|---|---|---|---|---|---|
| | 样本数 | 特征数 | 类别数 | | 样本数 | 大类样本数 | 小类样本数 |
| ionosphere | 351 | 34 | 2 | 10 | 275 | 225 | 50 |
| wdbc | 421 | 22 | 2 | 15 | 257 | 206 | 51 |
| Breast | 699 | 10 | 2 | 20 | 608 | 458 | 150 |

### 3.2 实验与结果分析

实验利用 Relief 算法或 ReliefF 算法进行特征权重计算,以 KNN 做分类器($K=3$),为了得到更可靠的实验结果,采用 5 折交叉验证法,即将数据集平均分成 5 份,随机选择其中 1 份作为测试数据,其余的作为训练数据,并且取 5 次的平均值作为最终的结果。结果表明,选择原始数据特征数的 20% 进行分类效果最佳(以 ionosphere 为例,选择权重较大的 7 个特征进行分类效果最佳)。

2.2.1 实验 1

实验中,迭代次数 $m$ 取训练集中的样本数。表 2 为在 ionosphere 数据集上随机选择样本和阈值选择得到的权重较大的 7 个特征以及它们的权重;图 1 为在 ionosphere 数据集上两种方法选择样本得到各特征权重值。表 3 为 3 个数据集上的分类正确率。

**表 2 Relief 算法与阈值-Relief 算法在 ionosphere 数据集上获得的前 7 个特征的权重**

**Tab. 2 The first 7 feature weights of Relief algorithm and the threshold -Relief algorithm in ionosphere**

| 算 法 | | 特征选择 | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Relif 算法 | 特征排序 | 27 | 29 | 8 | 24 | 33 | 21 | 25 |
| | 权重 | 0.093 4 | 0.091 1 | 0.087 0 | 0.080 5 | 0.078 5 | 0.073 3 | 0.071 1 |
| 阈值-Relief 算法 | 特征排序 | 8 | 29 | 24 | 34 | 27 | 28 | 19 |
| | 权重 | 0.116 0 | 0.095 8 | 0.093 4 | 0.087 4 | 0.085 9 | 0.081 7 | 0.078 9 |

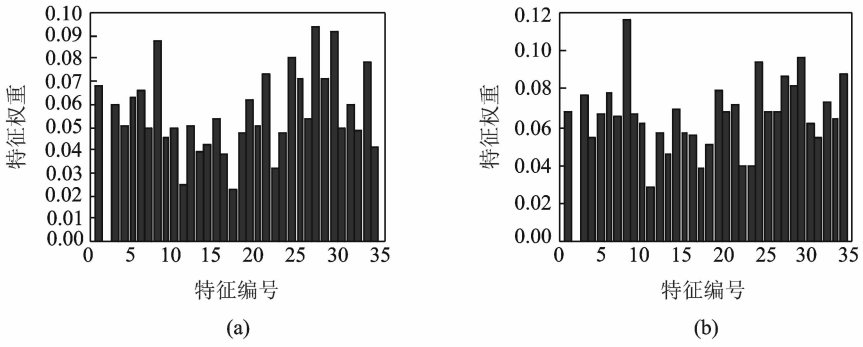



图 1 Relief 算法与阈值-Relief 算法在 ionosphere 数据集上各特征权重值比较

Fig. 1 Comparison of feature weights of Relief algorithm and Threshold-Relief algorithm in ionosphere

**表 3 Relief 算法与阈值-Relief 算法在 3 个数据集上的分类正确率** %

**Tab. 3 The classification accuracy of Relief algorithm and threshold -Relief algorithm in three data sets**

| 算 法 | ionosphere | wdbc | Breast |
|---|---|---|---|
| Relif 算法 | 88.3 | 90.3 | 89.5 |
| 阈值-Relief 算法 | 90 | 92 | 93.3 |

从表 2 和图 1 可以看出，在 ionosphere 数据集上，Relief 算法与阈值-Relief 算法得到的特征排序和权重并不相同。由表 3 可知，在 3 个数据集上，阈值-Relief 算法的分类正确率均比 Relief 算法的分类正确率要高，说明干扰点确实是 Relief 算法的消极因素，而改进后的阈值-Relief 算法，可以有效地避开干扰点，选出更能代表类别的特征。该方法同样也适用于不平衡数据。

3.2.2 实验 2

表 4 为在 ionosphere 的不平衡数据集上使用 Relief 算法、K-means-ReliefF 算法（$K=10$）和 Kmeans-Relief 抽样算法得到的权重较大的 7 个特征的排序、特征权重以及小类别的分类正确率和总的分类正确率。图 2 为在 Ionosphere 数据集上 3 种方法得到的各特征权重。

**表 4 3 种抽样算法在 Ionosphere 数据集上的前 7 个特征权重及分类正确率**

**Tab. 4 The first 7 feature weights and classification accuracy of three kinds of algorithm in ionosphere**

| 算 法 | | 特征选择 | | | | | | | 小类别正确率/% | 正确率/% |
|---|---|---|---|---|---|---|---|---|---|---|
| Relief 算法 | 特征排序 | 8 | 14 | 34 | 24 | 32 | 22 | 10 | 50 | 85.1 |
| | 权重 | 0.186 6 | 0.160 4 | 0.145 6 | 0.134 6 | 0.124 5 | 0.1185 | 0.1179 | | |
| K-means-ReliefF 算法 | 特征排序 | 15 | 19 | 17 | 13 | 21 | 23 | 11 | 70 | 94.5 |
| | 权重 | 0.214 5 | 0.099 5 | 0.199 6 | 0.194 9 | 0.179 9 | 0.157 8 | 0.152 2 | | |
| K-means-Relief 抽样算法 | 特征排序 | 8 | 24 | 32 | 14 | 6 | 1 | 27 | 65 | 89.1 |
| | 权重 | 0.148 5 | 0.141 5 | 0.129 9 | 0.127 5 | 0.112 4 | 0.112 4 | 0.109 7 | | |

通过表 4 和图 2 可以看出，Relief，K-means-ReliefF 算法和 K-means-Relief 抽样算法在 ionosphere 数据集上得到的前 7 个权重较大的特征及其特征权重，同时可以看到，在小类别的分类正确率上有了明显的提高。表 5 给出了 Relief 算法、K-means-ReliefF 算法和 K-means-Relief 抽样算法在不同数据集上分类正确率的比较。实验结果显示，大类集聚类后，无论是利用 ReliefF 算法在多类上进行特征选择，还

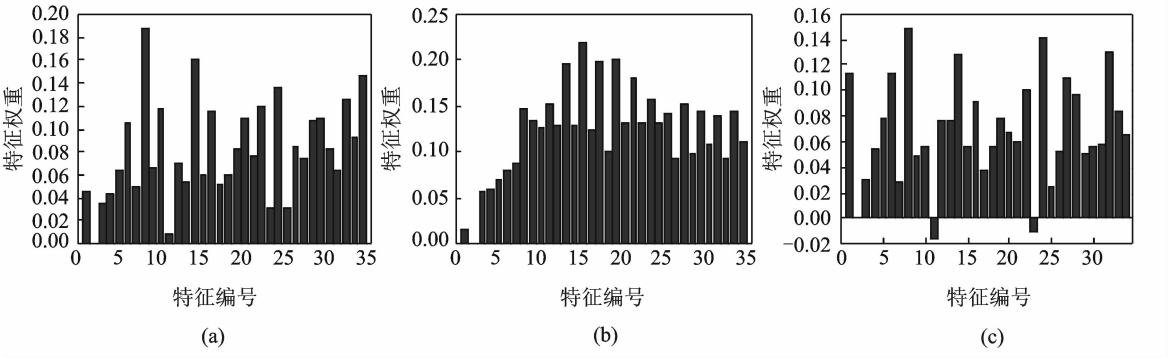



图 2 Relief 算法、Kmeans-ReliefF 算法和 Kmeans-Relief 抽样算法在 ionosphere 数据集上各特征权重

Fig. 2 Feature weights of Relief algorithm, Kmeans-ReliefF algorithm and Kmeans-Relief sampling algorithm in ionosphere

是利用 Relief 算法通过从多个类别中抽取样本进行特征选择,都可以有效弥补不平衡数据带来的不足。在 wdbc 数据集上表现不明显的原因是由于数据集的特征较少,抽取的特征有限,可能丢掉了一些相对重要的特征。

**表 5 Relief 算法、K-means-ReliefF 算法和 K-means-Relief 抽样算法在 3 个数据集上的分类正确率** %

**Tab. 5 The classification accuracy of Relief algorithm, K-means-ReliefF algorithm and K-means-Relief sampling algorithm in three data sets**

| 数据集 | ionosphere | wdbc | Breast |
|---|---|---|---|
| Relief 算法 | 86.3 | 90.5 | 89.2 |
| Kmeans-ReliefF 算法 | 89.8 | 89.8 | 91 |
| Kmeans-Relief 抽样算法 | 90.3 | 90.3 | 92.3 |

## 4 结束语

本文提出的阈值 Relief 算法,可以有效消除干扰数据集中干扰点对分类准确率的消极影响,提高分类精度。而结合 K-means 聚类算法提出的 K-means-ReliefF 算法和 K-means-Relief 抽样算法,可以有效避开 Relief 算法在随机选择样本时在不平衡数据集上表现出来的不足。实验结果表明,在不降低大类样本的正确率的基础上,有效地提高了小类样本的正确率。本文还存在一定的不足之处,希望能在下面几方面加以改进:(1)首先干扰点和不平衡样本集是人工构造,存在一定的随机性,下一步的工作希望能在真实的不平衡数据集上去验证算法的有效性;(2)数据集中特征偏少,这样可能在少量的特征上,每个特征都是比较重要的特征,故实验结果不是很明显。今后的工作重点是将该算法应用在不同的更大规模的真实地数据集上测试,针对数据集的不同改进算法。

**参考文献:**

[1] 张学工. 模式识别(第三版)[M]. 北京:清华大学出版社,2010.
Zhang Xuegong. Pattern recognition (Third Edition) [M]. Beijing: Tsinghua University Press, 2010.

[2] 钱宇华,梁吉业,王锋. 面向非完备决策表的正向近似特征选择加速算法[J]. 计算机学报,2011,31(3):435-442.
Qian Yuhua, Liang Jiye, Wang Feng. A positive-approximation based accelerated algorithm to feature selection from incomplete decision tables[J]. Chinese Journal of Computers, 2011,31(3):435-442.

[3] 刘全金,赵志敏,李颖新. 基于特征间距的二次规划特征选取算法[J]. 数据采集与处理,2015,30(1):126-136.
Liu Jinquan, Zhao Zhimin, Li Yingxin. Feature selection algorithm based on quadratic programming with margin between fea-

tures[J]. Journal of Data Acquisition and Processing, 2015, 30(1): 126-136.

[4] 李嘉. 语音情感的维度特征提取与识别[J]. 数据采集与处理, 2012, 27(3): 389-393.
Li Jia. Dimensional feature extraction and recognition of speech emotion[J]. Journal of Data Acquisition and Processing, 2012, 27(3): 389-393.

[5] Chawla NV, Japkowicz N, Kotcz A. Editorial: Special issue onlearning from imbalanced data sets[J]. SIGKDD Explorations Newsletters, 2004, 6(1): 1-6.

[6] Abe N, Kudn M. Non-parametric classifier-independent feature selection[J]. Pattern Recognition, 2006, 39(5): 737-746.

[7] Kira K, Rendell L. A practical approach to feature selection[C]//Proc 9th International Workshop on Machine Learning. San Francisco: Morgan Kaufmann, 1992: 249-256.

[8] Kira K, Rendell L. The feature selection problem: Traditional methods and new algorithm[J]. Proc AAAI'92. San Jose, CA: [s. n.], 1992: 129-134.

[9] Knonenko I. Estimation attributes: Analysis and extensions of Relief [C]//Euopean Conference on Machine Learning. Catania: Springer Verlag, 1994: 171-182.

[10] Sun Yijun. Iterative Relief for feature weighting: Algorithms, theories, and applications[J]. IEEE Trans on Pattern Anslysis and Machine Intelligence, 2007, 29(6): 1035-1051.

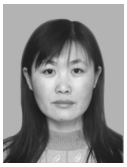
[11] Sun Y, Wu D. A Relief based feature extraction algorithm [C]//Proceedings of the 8th SIAM International Conference on Data Mining. Atlanta, GA, USA: [s. n.], 2008: 188-195.

[12] Liang Jiye, Bai Liang, Dang Chuangyin, et al. The K-means-type algorithms versus imbalanced data distributions[J]. IEEE Transactions on Fuzzy Systems, 2012, 20(4): 728-745.

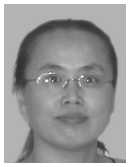
[13] 黄莉莉. 基于多标签 ReliefF 的特征选择算法[J]. 计算机应用, 2012, 32(10): 2888- 2890, 2898.
Huang Lili. Feature selection algorithm based on multi-label ReliefF[J]. Journal of Computer Applications, 2012, 32(10): 2888-2890, 2898.

[14] 林舒杨, 李翠华. 不平衡数据的降采样方法研究[J]. 计算机研究与发展, 2011, 48(S): 47-53.
Lin Shuchang, Li Cuihua. Under-sampling method research in class-imbalanced data[J]. Journal of Computer Research and Development, 2011, 48(S): 47-53.

**作者简介:**



**菅小艳**(1975-),女,讲师,研究方向:数据挖掘,机器学习,E-mail: jianxiaoyan@tynu.edu.cn。



**韩素青**(1964-),女,副教授,研究方向:数据挖掘,机器学习,E-mail: hansuqing@tynu.edu.cn。

**崔彩霞**(1974-),女,副教授,研究方向:数据挖掘,机器学习,E-mail: cuicaixia@tynu.edu.cn。