

文章编号:1001-9081(2014)09-2608-04

doi:10.11772/j.issn.1001-9081.2014.09.2608

## 互信息与模糊 C 均值聚类集成的特征优选方法

朱接文\*, 肖军

(江西工业工程职业技术学院 计算机工程系, 江西萍乡 337000)

(\*通信作者电子邮箱 15029463@qq.com)

**摘要:**针对大型数据中大量冗余特征的存在可能降低数据分类性能的问题,提出了一种基于互信息(MI)与模糊 C 均值(FCM)聚类集成的特征自动优选方法 FCC-MI。首先分析了互信息特征及其相关度函数,根据相关度对特征进行排序;然后按照最大相关度对应的特征对数据进行分组,采用 FCM 聚类方法自动确定最优特征数目;最后基于相关度对特征进行了优选。在 UCI 机器学习数据库的 7 个数据集上进行实验,并与相关文献中提出的基于类内方差与相关度结合的特征选择方法(WCMFS)、基于近似 Markov blanket 和动态互信息的特征选择算法(B-AMBDMI)及基于互信息和遗传算法的两阶段特征选择方法(T-MI-GA)进行对比。理论分析和实验结果表明,FCC-MI 不但提高了数据分类的效率,而且在有效保证分类精度的同时能自动确定最优特征子集,减少了数据集的特征数目,适用于海量、数据特征相关性大的特征约简及数据分析。

**关键词:**互信息;特征优选;模糊 C 均值聚类;数据分组

**中图分类号:** TP391    **文献标志码:**A

### Feature selection method based on integration of mutual information and fuzzy C-means clustering

ZHU Jiewen\*, XIAO Jun

(Department of Computer Engineering, Jiangxi Polytechnic College, Pingxiang Jiangxi 337000, China)

**Abstract:** Plenty of redundant features may reduce the performance of data classification in massive dataset, so a new method of automatic feature selection based on the integration of Mutual Information and Fuzzy C-Means (FCM) clustering, named FCC-MI, was proposed to resolve this problem. Firstly, MI and its correlation function were analyzed, then the features were sorted according to the correlation value. Secondly, the data was grouped according to the feature with the maximum correlation, and the number of the optimal features were determined automatically by FCM clustering method. At last, the optimization selection of the features was performed using correlation value. Experiments on seven datasets of UCI machine learning database were conducted to compare FCC-MI with three methods come from the literatures, including WCMFS (Within class variance and Correlation Measure Feature Selection), B-AMBDMI (Based on Approximating Markov Blank and Dynamic Mutual Information), and T-MI-GA (Two-stage feature selection algorithm based on MI and GA). The theoretical analysis and experimental results show that the proposed method not only improves the efficiency of data classification, but also ensures the classification accuracy and automatically determine the optimal feature subset, which reduces the number of the features of the dataset, thus it is suitable for feature reduction and analysis of mass data with large correlation features.

**Key words:** Mutual Information (MI); feature selection; Fuzzy C-Means (FCM) clustering; data grouping

## 0 引言

数据挖掘与模式识别领域常常需要处理包含高维特征的海量数据集。组成这些数据集的部分特征是冗余的,甚至是不相关的。冗余或不相关特征(噪声特征)的存在不但会降低数据处理算法的效率,有损学习算法的性能,甚至会造成维数灾难<sup>[1]</sup>;但也并不是特征数越少越好,太少的特征反映不出数据样本的固有信息。因此,在分析处理这类数据集时,有必要对其进行预处理,去除冗余和噪声特征,选择出最能反映数据集固有信息的相关特征,以提高数据分类的设计和处理效率<sup>[2]</sup>。

特征选择是模式识别过程中的一个重要步骤。目前特征

选择方法大致可分为封装型方法(wrapper)和过滤型方法(filter)。封装型方法,如以遗传算法(Genetic Algorithm, GA)为基础的特征选择算法<sup>[3-4]</sup>,由于该算法将分类器作为其在特征选择过程中的评价标准,虽然其准确率较高,但由于要训练一个分类器,其计算复杂度高,不适合大规模的数据集。而过滤方法,如互信息(Mutual Information, MI)方法<sup>[5-7]</sup>等主要是依据训练样本集中数据特征选择出一个特征子集,其独立于具体的学习算法,运行效率较高,因此适用于大规模数据集;但由于特征子集选择的非典型性,其数据处理精度有限,所得到的特征子集也不是最优的。

因此,人们试图将过滤方法与封装方法组合起来进行特征选择,如文献[8]指出了仅仅根据互信息选择出来的特征

收稿日期:2014-03-11;修回日期:2014-05-13。

作者简介:朱接文(1976-),男,江西抚州人,副教授,硕士,主要研究方向:智能信息处理、数据库;肖军(1975-),男,江西泰和人,副教授,硕士,主要研究方向:电子与通信。

不一定能构成最优特征子集,必须结合其他方法才能达到此目的;文献[9]基于互信息及遗传算法,提出了一种两阶段的、混合式的特征选择方法 T-MI-GA (Two-stage feature selection algorithm based on MI and GA),在对数据学习分类器的设计中取得了较好的应用效果;文献[5]提出一种基于近似 Markov blanket 和动态互信息集成的特征选择算法 B-AMBDMI (Based on Approximating Markov Blank and Dynamic Mutual Information),以支持向量机 (Support Vector Machine, SVM)为分类器,在 UCI 数据集上进行实验,获得了较高的分类精度。这些组合方法能够使封装方法充分利用过滤方法得到的结果,加快封装算法的收敛,并能产生具有更高分类性能的典型特征子集,然而这些组合方法在实际应用中仍会存在这样或那样的不足。如文献[8]中算法涉及的具体参数选择标准还有待进一步研究;文献[9~10]算法所针对的是离散性的特征变量,对于连续特征变量没有给出相关处理方式;文献[5]所提出的方法过于复杂,且必须与 Bagging 方法结合使用。另外,所有上述方法基本上属于有监督学习方法,不能自动确定最优特征的维数。

本文在分析数据集构成特征的基础上,运用模糊 C 均值 (Fuzzy C-Means, FCM) 聚类方法,以互信息为特征分组度量标准,提出了一种基于互信息分组聚类的特征自动优选方法 FCC-MI (Fuzzy C-Means Clustering based on MI)。即首先根据互信息的度量规则,按最大相关度所对应的特征对数据样本进行分组;然后基于各个分组,采用 FCM 聚类方法对特征进行优选,并自动确定最优特征的维数,以达到数据自动降维之目的。同时为了证明算法的有效性,以 Naive Bayes (NB) 为分类器,通过实验等实际应用,验证了基于互信息与 FCM 聚类集成在对样本降维的特征优选过程中有较高的效率。

## 1 基于互信息的特征排序及数据分组

### 1.1 互信息及特征相关性度量

互信息是从信息熵的概念引申出来的,熵是对随机变量不确定性的一种度量,因此,要定义特征互信息,首先必须给出信息熵的度量。

设数据集  $D$  由  $n$  个样本实例组成,每一个样本包含  $d$  维特征  $(f_1, f_2, \dots, f_d)$ ,  $P(f_i)$  是特征  $f_i$  ( $i < d$ ) 的概率分布密度函数,这里为  $f_i$  取不同可能值的概率,则特征  $f_i$  的信息熵  $H(f_i)$  定义如下:

$$H(f_i) = - \sum_{f_i} P(f_i) \ln P(f_i) \quad (1)$$

若已知另一个特征  $f_j$  的取值之后,  $f_i$  取值的不确定性可用条件熵来度量,即

$$H(f_i | f_j) = - \sum_{f_j} P(f_j) \sum_{f_i} P(f_i | f_j) \ln(f_i | f_j) \quad (2)$$

其中:  $P(f_i | f_j)$  表示已知  $f_j$  值时  $f_i$  的概率(条件概率),在通常情况下其值小于或等于初始不确定性信息熵,在两个特征变量相互独立时取值相等,即  $H(f_i) = H(f_i | f_j)$ ;另外,式(2)中的所有特征为离散性特征变量,如果特征为连续性变量,则上式中的求和替换为求积分,其余不变。

有了熵的定义后,两个特征  $f_i$  和  $f_j$  之间的互信息可以定义如下:

$$I(f_i; f_j) = I(f_j; f_i) = H(f_i) - H(f_i | f_j) =$$

$$\sum_{f_i} \sum_{f_j} P(f_i, f_j) \ln \frac{P(f_i, f_j)}{P(f_i)P(f_j)} \quad (3)$$

其中  $P(f_i, f_j)$  表示它们的联合概率密度。从式(3)中可以看出,互信息可以看作是已知特征  $f_j$  或  $f_i$  的信息后对于特征  $f_i$  或  $f_j$  不确定性的减小量,即二者共有的信息量,这个值越大,意味着所包含的相同信息越多,说明两个变量之间的相关程度越高;当两变量完全无关或互相独立时,它们的互信息量为 0,说明两者之间不存在相同的信息。另外从信息论的角度来看,特征选择的目标就是寻找一个能包含原特征集合中绝大部分或全部信息的特征子集,该特征子集的存在可以最大限度地降低其他未选特征的不确定性。因此,特征选择算法应该选择那些与其他特征具有最大互信息的典型特征,因为它们可以最大限度地降低其他特征的不确定性,也就是说其他特征由于这些特征的存在将会提供微不足道的信息。

基于以上分析,特征  $f_i$  的相关度  $Rel(f_i)$  定义如下:

$$Rel(f_i) = \frac{1}{d} \sum_{j=1}^d I(f_i; f_j) = \frac{1}{d} \left( H(f_i) + \sum_{j=1, j \neq i}^d I(f_i; f_j) \right) \quad (4)$$

其中:  $f_i$  的相关度就是它与整个特征集合的平均互信息;  $H(f_i)$  表示特征  $f_i$  包含的信息熵,其值越大,表明  $f_i$  能够给相关算法提供更多的信息;  $\sum_{j=1, j \neq i}^d I(f_i; f_j)$  表示特征  $f_i$  和其他特征共同含有的信息量,其值越大表示其共有信息越多,因而其他特征能够提供的“新”信息越少。如果选择具有最大  $Rel(f_i)$  值的特征,那么数据就可以最低限度地丢失信息,该特征即为数据集中最“典型”的特征。

### 1.2 数据分组方法

有了上述特征相关度的定义,本文对数据集进行分组处理,从而为下一步的模糊聚类选择最优特征子集及自动确定最优特征数目打下基础,具体分组实现步骤如下:

步骤 1 对于给定的训练集  $D$ ,其所构成的  $d$  维特征变量为  $(f_1, f_2, \dots, f_d)$ ,用式(4)计算出特征  $f_i$  ( $i = 1, 2, \dots, d$ ) 的相关度  $Rel(f_i)$ 。

步骤 2 对各特征所对应的相关度值,按由大到小的顺序进行排序。

步骤 3 取第 1 个相关度(最大相关度)值所对应的特征,根据该特征的特征值将数据集  $D$  分成多个分组,分组数目及每一分组中特征值的取值范围可根据实际情况设定。

步骤 4 考察每一分组中的数据子集  $D_i$  ( $i = 1, 2, \dots, c$ ),设  $c$  为分组数目,如果组成  $D_i$  内的数据样本相差甚远,说明该组内的数据样本可能不属于“同一类别”,还可以再分,则取第 2 个相关度(次大相关度)值所对应的特征,返回步骤 3 再次对该分组进行划分。如此循环,直到每一个分组中的样本分布基本合理为止。

经过上述处理后,训练数据集  $D$  按特征相关度值分成了  $c$  个子集  $D_i$  ( $i = 1, 2, \dots, c$ ),由于本文选择了最大、次大相关度值所对应的特征进行分组,这些特征最能反映样本信息,属于典型特征,因此,每一子集  $D_i$  中的样本分布较为合理有效,为后续的特征自动优选打下了基础。

## 2 由分组引导的特征自动优选算法

从第 1 章中可知互信息可以用来度量两个特征之间的相

互关联程度,但是,互信息只评价了单个特征与目标变量的相关性与典型性,不能对一个可能的特征子集优劣进行评价和度量;而且用互信息确定特征是否冗余也存在一定的困难,例如当两个特征彼此不完全相关时,很难判断哪一个是冗余的。基于此,本文在前述数据分组的基础上,采用 FCM 聚类方法来优选一个特征子集,并自动确定特征子集的数目。

## 2.1 相关概念定义

上一章将数据集分成  $c$  个分组后,设每个分组内有  $l$  个样本,则对于第  $k$  个分组内的样本  $x_i^{(k)}$  ( $i = 1, 2, \dots, l$ ;  $k = 1, 2, \dots, c$ ) 只有相关度大的特征才对分组有利,相关度愈小的特征对分组愈不利。因此,  $l$  个样本在每一维特征值上都应近似呈高斯分布,每个样本的各维特征值都聚集在各自均值附近的一定范围内,与均值偏差越小,该特征概率密度函数值越大,说明该特征相关度值越大,对分组越有利;反之亦然。因此,组内各维特征的均值矢量可以初步确定为组内样本的“核心”,即代表性样本的特征值,如式(5) :

$$m_j^{(k)} = \frac{1}{l} \sum_{i=1}^l x_i^{(k)}; k = 1, 2, \dots, c, j = 1, 2, \dots, d \quad (5)$$

其中:  $m_j^{(k)}$  为第  $k$  组各样本的第  $j$  维特征的均值,  $c$  为分组数目,  $d$  为特征维数。则  $m^{(k)} = \bigcup_{j=1}^d m_j^{(k)}$  初步确定为第  $k$  组内所有特征的均值矢量集。

## 2.2 基于 FCM 聚类自动确定优选特征数目

有了上述概念后,本文采用 FCM 聚类方法来实现最优特征数目的自动确定,首先需用  $m^{(k)}$  的均值矢量构造各分组间的特征值集  $\mathbf{y}_j$ :

$$\mathbf{y}_j = (m_j^{(1)}, m_j^{(2)}, \dots, m_j^{(c)}); j = 1, 2, \dots, d \quad (6)$$

其中  $c, d$  分别为上面所述的分组数目及特征维数。这样就得到了  $d$  个  $c$  维各分组的特征矢量,同时为了便于下一步操作,还需对  $\mathbf{y}_j$  作归一化处理:

$$\mathbf{y}'_j = \mathbf{y}_j / \max_{k=1}^c \{m_j^{(k)}\} \quad (7)$$

归一化后,组间相关性强的特征矢量变为相似矢量,这样可用相似矢量之间的距离定义其相似性:

$$S(\mathbf{y}'_i, \mathbf{y}'_j) = [D^2(\mathbf{y}'_i, \mathbf{y}'_j) + 1]^{-1} \quad (8)$$

其中  $D$  为 Euclid 距离,可见两个矢量间的距离越小,其相似性越大。取一限值  $S_T$  ( $S_T = S^*(\mathbf{y}'_i, \mathbf{y}'_j)$ ),当相似性大于  $S_T$ ,则需要压缩,否则保留,那么就有了基于 FCM 自动确定典型特征维数的方法,即:

首先取  $q = 0.5d$ ,即将上面所述的  $d$  个  $c$  维的特征矢量分成  $q$  个聚类,并随机选  $q$  个初始聚类原型,按式(9)计算每个聚类的类内均方差:

$$\varepsilon(\mathbf{y}', \beta_i) = \sum_{j=1}^d \mu_j^m D^2(\mathbf{y}'_j, \beta_i); i = 1, 2, \dots, q \quad (9)$$

其中:  $\beta_i$  为每个聚类的原型模式,  $\mathbf{y}'$  为归一化后某一聚类中特征均值矢量。得到每个聚类的类内方差后,判断  $\max_{i=1}^q \{\varepsilon(\mathbf{y}', \beta_i)\}$  是否小于  $(S_T - 1)^{-1}$ ,不满足则增大  $q$ ,否则减小  $q$ ,再转到式(10)进行聚类分析,直到达到临界点为止,即:

$$\max_{i=1}^q \{\varepsilon(\mathbf{y}', \beta_i)\} \leq (1 - S_T)/S_T \leq \max_{i=1}^{q-1} \{\varepsilon(\mathbf{y}', \beta_i)\} \quad (10)$$

则此时的  $q$  即为符合条件的特征维数。

## 2.3 调用 FCM 聚类算法对特征分类

要把  $d$  维特征划分到  $q$  个子集中,相似特征被划分到同一子集,并得到特征数据的隶属度函数  $\mathbf{u} = [\mu_i(\mathbf{y}_j')]$ ,见式(11),可直接调用 FCM 算法,FCM 算法中,聚类数为  $q$ ,有关该算法的应用请参见文献[11],限于篇幅,本文不再一一详述。

$$u_i(\mathbf{y}_j') = FCM(\{\mathbf{y}_j' | j = 1, 2, \dots, d\}); i = 1, 2, \dots, q \quad (11)$$

## 2.4 基于相关度进行特征优选

在获得  $q$  个聚类后,由于每一个类内的特征相似度大,因此,需要去掉冗余特征,按前面相关度定义,计算每一聚类中各特征与其聚类原型的相关度(聚类原型在 FCM 算法中会给出),取每一类中最大相关度对应的特征,即可得到  $q$  个最优特征。

算法具体描述如下:

输入 数据集  $D$  及其  $d$  维特征  $(f_1, f_2, \dots, f_d)$ ;

输出  $q$  个最优特征 ( $q < d$ )。

根据式(4)计算  $f_i$  的相关度  $Rel(f_i)$ ;

$sort\_desc Rel(f_1), Rel(f_2), \dots, Rel(f_d);$

$h = 1; While(h \leq d)$

根据  $Rel(f_h)$  grouping  $D$  as  $c$  个分组  $D_i$ ;

$While(D_i < \varepsilon) \varepsilon$  为设定的组内样本数目

$h = h + 1;$

按  $Rel(f_h)$  grouping  $D_i$ ;

End While

按式(5)计算分组  $D_k$  中代表性样本的特征值  $m_j^{(k)}$  ( $j$  为组内样本的第  $j$  维特征)

按式(6)、(7)计算归一化后的各分组间第  $j$  维特征值集  $\mathbf{y}_j'$ ;

取  $q = d/2$ , 设定  $S_T$ ;

$While$ (不满足式(10)中的条件)

调整  $q$ ;

End While

按式(11)对特征聚类;

对聚类结果进行整理,取每一类中  $Rel(f_i)$  最大的特征对应的属性,即得  $q$  个最优特征。通过以上过程,本文方法不仅压缩了特征维数(由原来的  $d$  维压缩成  $q$  维),而且消除了冗余特征,实现了海量数据中的特征优选。

## 3 实验结果与分析

### 3.1 实验样本数据集

为了验证本文所提出的特征优选方法的有效性,从 UCI 机器学习数据库<sup>[12]</sup>中选取 7 个数据集进行实验,如表 1 所示。

表 1 实验数据集

数据集	特征数	实例样本数	类别数
Soybean-large	35	307	19
Anneal	38	898	6
Sonar	60	208	2
Ionosphere	33	351	2
Zoo	17	101	7
Glass	9	214	7
Lung-cancer	56	32	2

### 3.2 实验方法

为了验证本文所提出的基于互信息分组聚类特征优选方法的有效性,本文选取了文献[2,5,9]中所提出的方法与本文算法作比较,其中文献[2]提出的基于类内方差与相关度结合的特征选择方法简称为 WCMFS (Within class variance and Correlation Measure Feature Selection), 文献[5]提出的基于近似 Markov blanket 和动态互信息的特征选择算法简称为 B-AMBDMI, 文献[9]中提出的基于互信息和遗传算法的两阶段特征选择方法简称为 T-MI-GA, 这三种方法反映了目前特征选择算法发展的前沿, 具有较好的代表性。实验对于表中的每一个数据集, 分别运用上述四种算法进行特征子集优选, 其得到最优特征子集的大小(特征数目)如表2所示。另外, 根据各自得到的特征子集, 采用 Naive Bayes (NB) 为分类器对这7个数据集进行分类, 并计算出相应的分类准确率和运行时间(单位为s), 如表3与表4所示。在实验中, 本文 FCC-MI 算法中阈值  $S_T$  取 0.9, FCM 算法的迭代次数为 10, 其他三种算法各相关参数的选取及设计见各自文献。实验环境为: CPU P4 3.1 GHz, 内存 2 GB, 算法基于 Matlab 7.10 (R2010a) 实现。

表2 四种算法得到的特征子集大小

数据集	所有特征	WCMFS	T-MI-GA	B-AMBDMI	FCC-MI
Soybean-large	35	10	13	23	8
Anneal	38	18	15	21	12
Sonar	60	22	6	10	5
Ionosphere	33	10	12	9	7
Zoo	17	12	8	8	8
Glass	9	6	3	4	3
Lung-cancer	56	5	4	13	4
均值	35.43	11.86	8.71	12.57	6.71

表3 基于各特征子集的数据样本聚类准确率对比 %

数据集	WCMFS	T-MI-GA	B-AMBDMI	FCC-MI
Soybean-large	93.44	92.84	94.31	92.36
Anneal	98.85	98.78	98.18	98.95
Sonar	80.72	77.40	83.68	90.16
Ionosphere	94.68	93.62	93.41	95.73
Zoo	77.53	82.57	83.65	82.78
Glass	90.73	91.66	90.73	95.87
Lung-cancer	97.66	96.77	97.71	98.39
均值	90.52	90.52	91.67	93.46

表4 基于各特征子集的数据样本聚类运行时间对比 s

数据集	WCMFS	T-MI-GA	B-AMBDMI	FCC-MI
Soybean-large	0.006	0.038	0.019	0.061
Anneal	0.005	0.036	0.012	0.048
Sonar	0.004	0.009	0.014	0.015
Ionosphere	0.004	0.008	0.013	0.020
Zoo	0.003	0.008	0.007	0.009
Glass	0.003	0.007	0.006	0.013
Lung-cancer	0.003	0.005	0.005	0.004
均值	0.004	0.016	0.011	0.024

### 3.3 实验结果分析

从表2中可以得到, 相对其他三种方法, 本文提出的

FCC-MI 方法由于采用自动寻优策略, 因此得到了更优的特征子集, 其所得的各数据集特征子集的特征数目也相对较少, 从而为提高数据分类效率打下基础(特征越少, 分类效率越高); 另外, 从表3中得到本文所提出的算法在7个数据集上的平均准确率也高于其他三种算法, 虽然在数据集 Soybean-large 上比其他三种算法稍低, 在 Zoo 上也低于B-AMBDMI, 但在其余数据集上均高于其他三种算法。然而, 由于 WCMFS、T-MI-GA、B-AMBDMI 算法进行特征子集选择的时间复杂度仅与特征维数( $d$ )及数据集大小( $n$ )有关, 且都为  $O(d^2n)$  量级, 而本文提出的 FCC-MI 算法是一种爬山迭代法, 其时间复杂度不但与特征维数及数据集的大小有关, 而且算法的收敛性与迭代的次数及设计的终止阈值有关, 因此其计算的时间远远多于其他三种算法, 如表4所示。

因此, 本文所提出的特征选择算法能够有效保证分类的精度, 而且能够自动确定最优特征子集, 较大幅度降低了数据集的特征数目, 对海量的、数据特征相关性大的特征约简及数据分析处理有一定的应用价值, 虽然其时间复杂度较大, 但运用现代高性能的计算机, 运行时间问题可以很好地解决。

### 4 结语

本文提出一种由互信息引导的数据分组聚类特征子集优选方法, 采用最大相关度所对应的特征对数据样本进行分组后, 通过分析数据特征值之间的关系, 采用 FCM 聚类方法对数据特征进行优选, 并自动确定最优特征子集的数目, 从而实现了特征子集的自动优选。与其他方法相比, 本文所提出的算法存在两大优势: 一是最优特征子集能自动获取, 在优选过程中不需要人工干预; 二是本文所提出方法的有效性较高, 它既较大幅度地降低数据集的特征维数, 又能保证基于该特征子集的分类精度。当然算法中有关参数的选择, 如前面所述的数据分组数目的确定、矢量相关性门限值  $S_T$  的选取、FCM 聚类算法迭代次数等都需靠人工事先确定, 这些参数的具体选择标准还有待进一步研究。

### 参考文献:

- [1] WU S, ZHANG W, HUANG H, et al. FD-CABOSFV interval variable high dimensional data clustering [J]. China Journal of Information Systems, 2011, 5(2): 77–87. (武森, 张文丽, 黄慧敏, 等. FD-CABOSFV 区间变量高维数据聚类 [J]. 信息系统学报, 2011, 5(2): 77–87)
- [2] ZHANG X, SUN Z, XU G, et al. A feature selection algorithm combining within-class variance with correlation measure [J]. Journal of Harbin Institute of Technology, 2011, 43(2): 133–136. (张晓光, 孙正, 徐桂云, 等. 一种类内方差与相关度结合的特征选择算法 [J]. 哈尔滨工业大学学报, 2011, 43(2): 133~136.)
- [3] RATA G A, VEGA J, MURARI A, et al. Improved feature selection based on genetic algorithm for real time disruption prediction on JET [J]. Fusion Engineering and Design, 2012, 87(9): 1670–1678.
- [4] ZHANG Y, YAN Y. A feature selection method based on adaptive genetic strategy [J]. Journal of Changchun University of Technology, 2010, 31(2): 126–130. (张云鹏, 尹一功. 一种基于自适应遗传策略的特征选择算法 [J]. 长春工业大学学报, 2010, 31(2): 126–130.)

(下转第 2649 页)

## 参考文献:

- [1] LI J, XU J, LOU Q. The research of real protection management system for the wide-band net information security [J]. Microcomputer Information, 2004, 20(2): 97–98. (李锦伟,徐进,楼巧萍.宽带网络信息安全实时保护管理系统的研究[J].微计算机信息,2004,20(2):97–98.)
- [2] LIN Y. Design of peer-to-peer communication network security model based on IP technology [J]. Microcomputer Information, 2005, 21(10X): 1–2. (林永和.基于IP技术的端对端通信网络安全模型分析设计[J].微计算机信息,2005,21(10X): 1–2.)
- [3] WANG Y. Internet communication security based on IPSec protocol [J]. Microcomputer Information, 2003, 19(12): 123–128. (王艳芳.基于IPSec安全协议的Internet通信安全[J].微计算机信息,2003,19(12): 100–102.)
- [4] CHEN J. Theory and method for research on covert secrecy communication based on steganography [D]. Zhengzhou: Information Engineering University, 2012. (陈嘉勇.基于隐写术的隐蔽保密通信理论与方法研究[D].郑州:信息工程大学,2012.)
- [5] CHEN J, ZHANG W, HAN T, et al. An efficient adaptive image steganographic method for  $\pm k$  embedding [J]. Acta Automatica Sinica, 2013, 39(10): 1594–1601. (陈嘉勇,张卫明,韩涛,等.高效 $\pm k$ 自适应图像隐写术[J].自动化学报,2013,39(10): 1594–1601.)
- [6] ZHANG H, PING X. Spatial steganography detection using merged DCT features [J]. Journal of Information Engineering University, 2012, 13(6): 646–649. (张昊,平西建.使用混合DCT特征检测空域隐写术[J].信息工程大学学报,2012,13(6): 646–649.)
- [7] BENDER W, GRUHL D, MORIMOTO N, et al. Techniques for data hiding [J]. IBM System Journal, 1996, 35(3/4): 313–336.
- [8] WU D-C, TSAI W-H. A steganographic method for images by pixel-value differencing [J]. Pattern Recognition Letters, 2003, 24(9/10): 1613–1626.
- [9] LIU J, HAN T, ZHANG W, et al. A histogram-preserving steganography based on wet paper coding and graphic matching theory [J]. Journal of Electronics & Information Technology, 2011, 33(3): 592–596. (刘九芬,韩涛,张卫明,等.一种基于湿纸编码和图匹配理论的直方图保持隐写算法[J].电子与信息学报,2011,33(3): 592–596.)
- [10] CHEN J, LIU J, ZHU Y, et al. Cryptographic secrecy of steganographic matrix encoding [J]. Journal on Communications, 2012, 33(6): 174–179. (陈嘉勇,刘九芬,祝跃飞,等.隐写术中矩阵编码的保密安全性[J].通信学报,2012,33(6): 174–179.)
- [11] CHEN J, WANG C, ZHANG W, et al. A secure image steganographic method in encrypted domain [J]. Journal of Electronics & Information Technology, 2012, 34(7): 1721–1726. (陈嘉勇,王超,张卫明,等.安全的密文域图像隐写术[J].电子与信息学报,2012,34(7): 1721–1726.)
- [12] TAN Y. Comparative study of information hiding method based on text typesetting format [J]. Computer and Modernization, 2013(6): 52–56. (谭瑛.基于文本排版格式的信息隐藏方法比较研究[J].计算机与现代化,2013(6): 52–56.)
- [13] HUANG S. Research on information hiding based on PDF file [D]. Changsha: Hunan University, 2011. (黄思敏.基于PDF文件的信息隐藏技术研究[D].长沙:湖南大学,2011.)
- [14] LIU Y, SUN X, LUO G. A novel information hiding algorithm based on structure of PDF document [J]. Computer Engineering, 2006, 32(17): 230–232. (刘友继,孙星明,罗纲.一种新的基于PDF文档结构的信息隐藏算法[J].计算机工程,2006,32(17): 230–232.)
- [15] TIAN G. Digital watermarking algorithm for PDF document [J]. Computer Engineering and Applications, 2012, 48(32): 85–88. (谭国律.PDF文档中的一种数字水印算法[J].计算机工程与应用,2012,48(32): 85–88.)
- [16] ZHAO L, GU Z, FANG Z, et al. An anti-fake method based on visual characteristic and morphology screen coding [J]. Journal of Optoelectronics · laser, 2008, 19(11): 1526–1529. (赵立龙,顾泽仓,方志良,等.一种基于视觉特性及形态网屏编码的纸质信息防伪方法[J].光电子·激光,2008,19(11): 1526–1529.)
- [17] GUO W, LIU Y, YANG B, et al. Research on information hiding technology in paper based document [J]. China Printing and Packaging Study, 2013(2): 30–34. (国伟,刘宇鑫,杨斌,等.印刷纸介质文档中的信息隐藏技术研究[J].中国印刷与包装研究,2013(2): 30–34.)
- [18] SU Y. The random problem on the FM screening [J]. Journal of Institute of Surveying and Mapping, 2001, 18(3): 226–228. (苏永宪.调频加网中的随机问题[J].测绘学院学报,2001,18(3): 226–228.)
- [19] ZHOU X, SHI R, AN J, et al. Study of FM screening efficiency based on multiplicative congruence pseudo-random algorithm [J]. China Printing and Packaging Study, 2011(3): 15–20. (周啸,史瑞芝,安敬,等.基于乘同余伪随机算法的调频加网效率研究[J].中国印刷与包装研究,2011(3): 15–20.)

(上接第2611页)

- [5] YAO X, WANG X, ZHANG Y, et al. Ensemble feature selection algorithm based on Markov blanket and mutual information [J]. Journal of Systems Engineering and Electronics, 2012, 34(5): 1046–1050. (姚旭,王晓丹,张玉玺,等.基于Markov blanket和互信息的集成特征选择算法[J].系统工程与电子技术,2012,34(5): 1046–1050.)
- [6] SYLVAIN V, TEODOR T, ABDESSAMAD K. Fault detection and identification with a new feature selection based on mutual information [J]. Journal of Press Control, 2008, 18(5): 479–490.
- [7] GUO B F, MARK S N. Gait feature subset selection by mutual information [J]. IEEE Transactions on Systems, Man and Cybernetics — Part A: System and Humans, 2009, 39(1): 36–46.
- [8] HSU H H, HSIEH C W, LU M. Hybrid feature selection by combining filters and wrappers [J]. Expert Systems with Applications, 2011, 38(7): 8144–8150.
- [9] QIU G, WANG N, WANG W. Two-stage feature selection algorithm based on mutual information and genetic algorithm [J]. Application Research of Computers, 2012, 29(8): 2903–2905. (裘国永,王娜,汪万紫.基于互信息和遗传算法的两阶段特征选择方法[J].计算机应用研究,2012,29(8): 2903–2905.)
- [10] ESTEVEZ P A, MICHEL T, PEREZ C A, et al. Normalized mutual information feature selection [J]. IEEE Transactions on Neural Networks, 2009, 20(2): 189–201.
- [11] XIAO M, LIU Y, ZHOU X. A property optimization method in support of approximately duplicated records detecting [C]// Proceedings of the 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems. Piscataway: IEEE, 2009, 3: 118–122.
- [12] BLAKE C, MERZ C. UCI repository of machine learning database [EB/OL]. [2013-03-15]. [http://www.ics.uci.edu/~mlearn/MLR\\_repository.html](http://www.ics.uci.edu/~mlearn/MLR_repository.html).