Computer Science and Application 计算机科学与应用, 2018, 8(5), 573-581
Published Online May 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.85065

Hans汉斯

Cluster Classification Algorithm Based on Feature Entropy Weight

Dong Chen[1], Yongbin Yu[1], Chenxi Yang[1], Yindong Chen[1], Nyima Tashi[2]

[1]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu Sichuan
[2]College of Information and Technology, Tibet University, Lhasa Tibet
Email: 1040155788@qq.com

Received: Apr. 12th, 2018; accepted: Apr. 26th, 2018; published: May 3rd, 2018

Abstract

Classification is a significant method in data mining. Most existing classification algorithms fail to make full use of the distribution information of various types of data. This paper proposes a cluster classification algorithm based on feature entropy weight. During the training, the information entropy weights are used to represent the data distribution of different categories and characteristics, and feature entropy weight vectors which can represent different data clusters can be obtained. During the test, if the test set is not divided into clusters of different types, it will be clustered according to the number of training clusters first. Then the feature entropy weight vectors of the test clusters are calculated, and the categories to which the test clusters belong are found by the cosine similarity. The experimental results show that the algorithm has high classification accuracy for data sets with distinct differences in the distribution of different types of features, and is more insensitive to abnormal data than existing classification algorithms. It can solve the problem that man-made labels are prone to errors to certain extent.

Keywords

Entropy Weight, Feature Distribution, Cluster, Classifier

基于特征熵权的集群分类算法

陈　董[1]，于永斌[1]，杨晨曦[1]，陈音东[1]，Nyima Tashi[2]

[1]电子科技大学信息与软件工程学院，四川 成都
[2]西藏大学，信息科学与技术学院，西藏 拉萨
Email: 1040155788@qq.com

收稿日期：2018年4月12日；录用日期：2018年4月26日；发布日期：2018年5月3日

*通讯作者。

文章引用: 陈董, 于永斌, 杨晨曦, 陈音东, Nyima Tashi. 基于特征熵权的集群分类算法[J]. 计算机科学与应用, 2018, 8(5): 573-581. DOI: 10.12677/csa.2018.85065

摘 要

分类是数据挖掘中一种重要方法，现有分类算法大多未能充分利用不同类别数据分布信息。本文提出一种基于特征熵权的集群分类算法。训练时，通过信息熵权重表示不同类别、特征的数据分布情况，得出能代表不同数据集群的特征熵权向量。测试时，若测试集未分成不同类别的集群，则先根据训练集群数进行聚类，后算出各测试集群的特征熵权向量，通过余弦相似找出各测试集群所属类别。经实验验证，本算法对不同类别特征分布差异明显的数据集具有较高分类精度，且比起现有分类算法对异常数据更为不敏感，一定程度上解决了人为标签容易出错的问题。

关键词

熵权，特征分布，集群，分类器

1. 引言

分类是一种重要的数据挖掘技术，其目的为根据数据集特点构造相应的分类模型，过程一般分为训练和测试两个阶段[1]。现有的成熟分类算法主要有：以实例为基础的归纳学习算法决策树[2] [3] [4]；基于概率论统计的贝叶斯分类算法[5] [6] [7] [8]；以实例为基础的懒惰学习算法 k-Nearest Neighbor (KNN) [9] [10] [11]；训练分界函数过程较为简单的逻辑回归[12] [13] [14]；训练分界函数过程较为复杂的神经网络[15] [16] [17]等。现阶段分类算法研究方面，杜景林提出了一种基于距离权值的 C4.5 组合决策树算法[18]；童先群提出了一种基于属性值对类别重要性的改进算法 Entropy-KNN [19]；这些改进算法都在一定程度上考虑了数据集的分布情况，但仍未能充分利用不同类别不同特征的数据集的分布信息。

本文提出一种基于特征熵权的集群分类算法，它不再以单个样本为分类研究对象，而以不同类别的数据集群为分类研究对象。训练时，根据标签将数据分为若干集群，后使用信息熵权重表示不同类别数据的特征分布情况[20] [21] [22]。测试时，若数据已经分为若干集群，则直接通过余弦相似找出训练集群中特征分布最为接近的集群，从而判断类别。若数据未分为若干集群，则先根据训练集群数进行聚类，再按余弦相似找出各个测试集群所属类别。实验证明本算法在不同类型数据分布差异较为明显时，有着较高的准确率，且比起现有分类器，本文所提算法对错误数据不敏感，一定程度上解决了人为标签容易出错的问题。

2. 基于信息熵的特征分布向量表示

熵的概念源于热力学，是对系统状态不确定性的一种度量。1948 年 Shannon 提出“信息熵”的概念，解决了对信息的量化度量问题。信息熵(entropy)是一种描述随机变量分散程度的统计量，信息熵越大，表示变量的离散程度越高[23]。

本文利用信息熵对分类问题中，不同类别的数据分布进行度量，从而将各类数据的分布情况量化。

2.1. 信息熵

信息熵是系统无序程度的一种度量，信息熵越小，无序度越低。设 X 为服从某种分布的随机变量，

其信息熵为：

$$H(X)=-\sum_{i=1}^{n}p(a_i)\log_2 p(a_i) \tag{1}$$

其中 X 取值 $\{a_1,a_2,\cdots,a_n\}$， $p(a_i)$ 为 X 取值 a_i 的概率[24]。

2.2. 熵权与特征分布

基于信息熵的特征分布表示，以最常见的高斯分布为例。设一可分为 m 类的数据集中，每个样本存在 k 个特征，且任意特征 X_i 均服从位置参数为 μ_i 、尺度参数为 δ_i 的概率分布，则其概率密度函数为：

$$f(X_i)=\frac{1}{\sqrt{2\pi}\delta_i}\mathrm{e}^{-\frac{(X_i-\mu_i)^2}{2\delta_i^2}} \tag{2}$$

特征 $X_i=\{x_{i1},x_{i2},\cdots,x_{in}\}$ 的信息熵为：

$$H(X_i)=-\frac{1}{\ln(n)}\sum_{j=1}^{n}p_{ji}\ln(p_{ji}) \tag{3}$$

其中 $p_{ij}=\dfrac{x_{ij}}{\sum_{j=1}^{n}x_{ij}}$，$n$ 为任一类数据的样本数。

若 $p_{ij}=0$，则有：$\lim\limits_{p\to 0}p_{ji}\ln(p_{ji})=0$

为增加不同类别样本间各个特征熵值可比较性，本文使用熵权表示数据变化剧烈程度的差异。对于任一类别，特征 X_i 的权重为：

$$W_i=\frac{1-H(x_i)}{k-\sum_{i=1}^{k}H(x_i)} \tag{4}$$

计算不同类别数据所有特征的熵权得特征熵权向量：

$$A_l=\{W_1,W_2,\cdots,W_k\} \tag{5}$$

其中 $l=0,1,2,\cdots,m$；$W_1+W_2+\cdots+W_k=10$。

2.3. 基于熵权的特征分布合理性仿真实验

分类器应尽可能的区分不同类别样本间数据的差异性，对于本文所采用的权重向量表示不同类别样本，应当使不同特征间权值差异尽可能的大。

设 $\{X_1,X_2\}$ 为同一类样本的特征集合，且 $\{X_1,X_2\}$ 服从高斯分布(2)。

令 $\mu_1=\mu_2=50,\delta_1=\delta_2=15$，生成样点数 num = 4000。采用控制变量法，分别研究三种参数变化对熵权变化影响情况。

2.3.1. 位置参数 μ_i 对权重分布的影响

令 $\mu_2=\mu_1+v,\mu_1=50$，v 从 0~100 进行迭代，$\delta=15$ 固定不变，按 2.2 中过程生成特征熵权，如图 1。

由图 1 可得，当 δ 固定时，不同特征间相差较小的 μ 可以获得较大的熵权差异。

2.3.2. 范围参数 δ 对权重分布的影响

令 $\delta_2=\delta_1+v,\delta_1=1$，$v$ 从 0~100 进行迭代，$\mu=80$，按 2.2 中过程生成特征熵权，如图 2。

由图 2 可得，当 μ 固定时，不同特征间相差较小的 δ 可以获得显著的熵权差异。

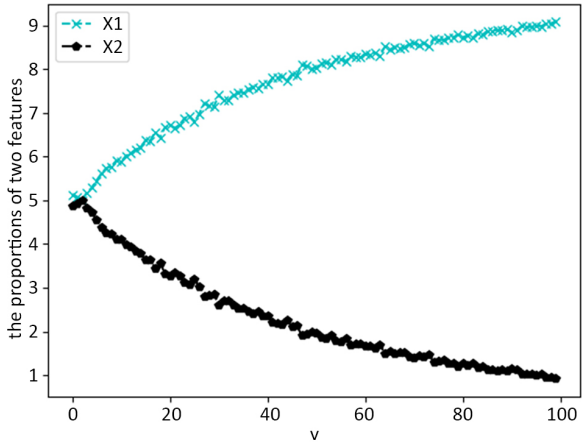

Figure 1. Weight changes with location parameters
图 1. 权重随位置参数变化情况

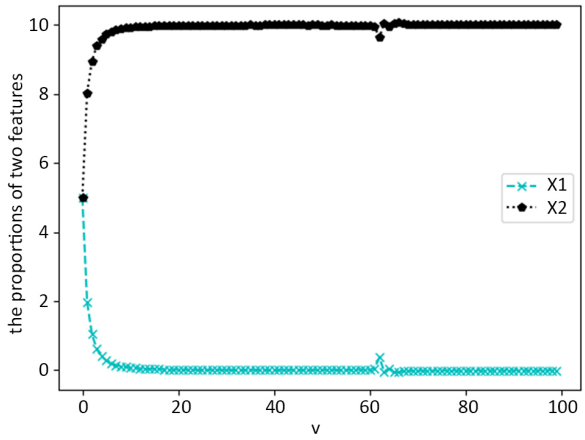

Figure 2. Weight changes with range parameters
图 2. 权重随范围参数变化情况

2.3.3. 范围参数 δ 对权重分布的影响

固定 $\mu_1 = 50$、$\delta_1 = 5$，$\mu_2 = 90$、$\delta_2 = 15$。随机生成样本数 num = 500，且 num = num + 10 * v，v 从 0~2000 进行迭代，并按 2.2 中过程生成特征熵权，如图 3。

可以看出，当随机生成样点数较少时(500 左右)权重有较大随机性，相对容易出现误判，但不同特征之间权值依然有较大差别。对于不同类别的样本集群，仍将具有良好的分类能力。且随着数据量增加这种随机误差将不断减小。

从以上的仿真实验可以看出，用熵权表示不同类别数据之间的分布差异是合理的。

3. 集群分类器

现有分类器多利用单个数据的信息进行训练与测试，而忽略了不同类别、特征数据集之间的分布信息。本文所提出集群分类器以不同类别、特征之间熵权差异为基础，进行数据集的分类。

集群分类算法，首先根据标签将训练集划分为不同类型的集群，再按 2 中所述计算出各类样本特征熵权向量。测试时，若数据已经分类，则直接计算测试集群特征熵权向量；若未分类，则根据训练集标签类数进行聚类。随后将训练集群的特征熵权向量与测试集群的特征熵权向量进行余弦相似匹配。整体算法流程如图 4。

3.1. 高斯混合聚类

当测试集未能分为集群给出时，首先需要对测试集进行聚类。由不同类别的特征分布之间的差异性可知，集群分类的主要误差产生于聚类过程，本文中采用高斯混合模型(GMM)完成聚类。对 GMM 进行简要介绍：

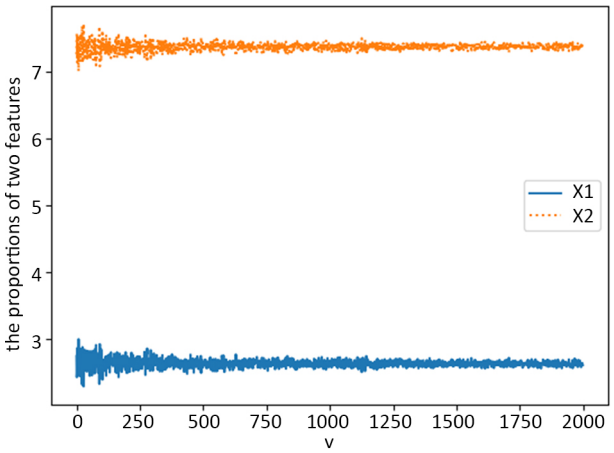

Figure 3. Weight changes with the number of samples
图 3. 权重随样点个数变化情况

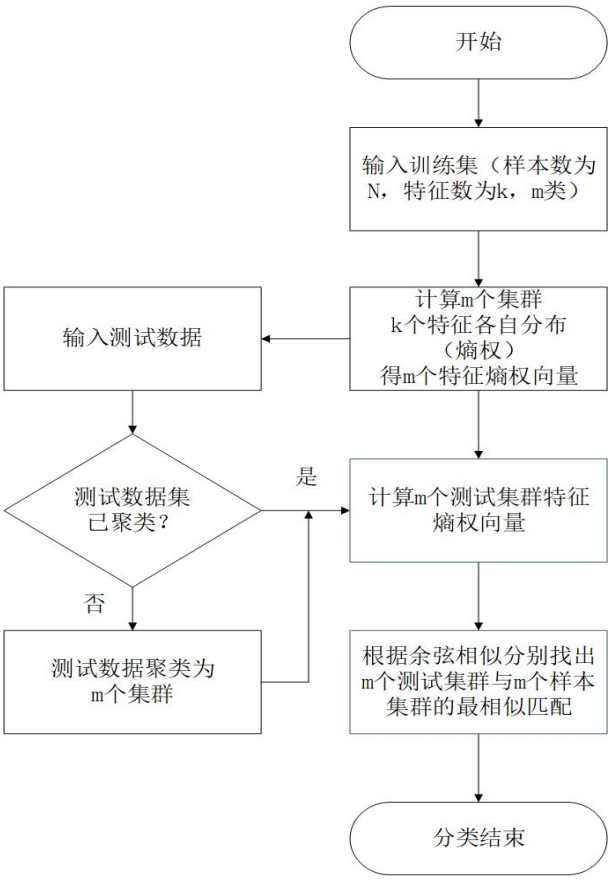

Figure 4. The Algorithm flow of cluster Classification Algorithm Based on Feature Entropy Weight
图 4. 基于特征熵权的集群分类算法

设随机变量 X 服从 $\mathrm{N}(x|\mu_l,\Sigma_l)$，则混合高斯模型可以用下式表示：

$$p(x|\theta)=\sum_{l=1}^{m}\alpha_k\mathrm{N}(x|\theta_l) \tag{6}$$

其中，α_l 是系数，$\alpha_l\geq 0$；$\mathrm{N}(x|\theta_l)$ 是高斯分布密度，$\theta_l=(\mu_l,\sigma_l^2)$ 且：

$$\mathrm{N}(x|\theta_l)=\frac{1}{\sqrt{2\pi}\sigma_l}\mathrm{e}^{-\frac{(x-\mu_l)^2}{2\delta_l^2}} \tag{7}$$

3.2 集群间的余弦相似匹配

假设根据标签将训练集数据分为 m 个集群，且特征数为 k。对训练集群求特征熵权得矩阵 $A_{m\times k}$：

$$A=\begin{bmatrix} w_{11}^A & \cdots & w_{1k}^A \\ \vdots & \ddots & \vdots \\ w_{m1}^A & \cdots & w_{mk}^A \end{bmatrix}=\begin{bmatrix} A_1 \\ \vdots \\ A_m \end{bmatrix} \tag{8}$$

其中行为不同的集群，每列为不同的特征熵权；w_{pq}^A 为第 p 个集群，第 q 个特征的熵权；$\boldsymbol{A}_p$ 为第 p 个集群的特征熵权向量。

求 m 个集群特征熵权向量的模向量得：

$$a=\begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}=\begin{bmatrix} |A_1| \\ \vdots \\ |A_m| \end{bmatrix} \tag{9}$$

同理，求测试集群的特征熵权矩阵 $B_{m\times k}$：

$$B=\begin{bmatrix} w_{11}^B & \cdots & w_{1k}^B \\ \vdots & \ddots & \vdots \\ w_{m1}^B & \cdots & w_{mk}^B \end{bmatrix}=\begin{bmatrix} B_1 \\ \vdots \\ B_m \end{bmatrix} \tag{10}$$

特征熵权向量的模向量 $b\,b$：

$$b=\begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}=\begin{bmatrix} |B_1| \\ \vdots \\ |B_m| \end{bmatrix} \tag{11}$$

定义向量 A_c 和 B_d 间的余弦相似度为：

$$cos_{cd}=\frac{\sum_{q=1}^{k}w_{cq}^A w_{dq}^B}{\sqrt{\sum_{q=1}^{k}\left(w_{cq}^A\right)^2}\sqrt{\sum_{q=1}^{k}\left(w_{dq}^B\right)^2}}=\frac{A_cB_d^{\mathrm{T}}}{a_cb_d} \tag{12}$$

从而可得训练集群与测试集群的余弦相似矩阵：

$$cos=\frac{AB^{\mathrm{T}}}{ab^{\mathrm{T}}} \tag{13}$$

其中 cos 为一个 $m\times m$ 的矩阵。每一行代表一个训练集群，每一列代表一个测试集群。要求测试集群与训练集群的余弦匹配只需求 cos 中每一列的最大值对应的训练集群即可。

3.3. 算法步骤

基于特征熵权的集群分类算法步骤如下：

Step 1：将数据分为训练集与测试集，假设数据类别数为 m，特征数为 k。训练集中根据标签，将不同类别的样本划分为 m 个集群；

Step 2：根据 $H\left(X_i\right)=-\frac{1}{\ln(n)}\sum_{j=1}^{n}p_{ji}\ln\left(p_{ji}\right)$ 计算 m 个集群 k 个特征的信息熵。对 m 个集群求各特征熵占比 $W_i=\frac{1-H\left(x_i\right)}{k-\sum_{i=1}^{k}H\left(x_i\right)}$，得 m 个特征熵权向量 $A_l=\left\{W_1^A,W_2^A,\cdots,W_k^A\right\}$。并以行为集群，列为特征，由 A_l 组合成特征熵权矩阵 $A_{m\times k}$，并计算每个集群特征熵权向量的模，放入特征熵权向量的模向量 $a_{m\times 1}$；

Step 3：输入测试数据，并判断测试数据是否已经分成不同集群，而只需要判断集群所属类别，若是转 Step 5；若否，则转 Step 4；

Step 4：使用聚类算法，将测试数据聚成 m 个集群。

Step 5：根据熵权计算测试数据 m 个集群 k 个特征各自分布，得 m 个特征熵权向量 $B_l=\left\{W_1^B,W_2^B,\cdots,W_k^B\right\}$。并以行为集群，列为特征，由 B_l 组合成特征熵权矩阵 $B_{m\times k}$，并计算每个集群特征熵权向量的模，放入特征熵权向量的模向量 $b_{m\times 1}$；

Step 6：由 $cos=\frac{AB^{\mathrm{T}}}{ab^{\mathrm{T}}}$ 得训练集群与测试集群的余弦相关矩阵。找出每一列中余弦相似的最大值，即得该列所代表测试集群对应的训练集群类别。

4. 仿真结果及分析

4.1. 实验设计及参数设置

为了增强与其它分类器的可对比性，在此仿真实验中不考虑测试数据在采样时已经分为不同集群的情况。以二维特征为例，按正态分布生成两类数据集：

第一类 $\mu_1=50$、$\delta_1=15$；$\mu_2=40$、$\delta_2=8$；第二类 $\mu_1=40$、$\delta_1=8$；$\mu_2=280$、$\delta_2=80$。

随机生成两类数据各 500 个，如图 5。

按上述分布再次生成两类测试数据集各 500 个，如图 6。

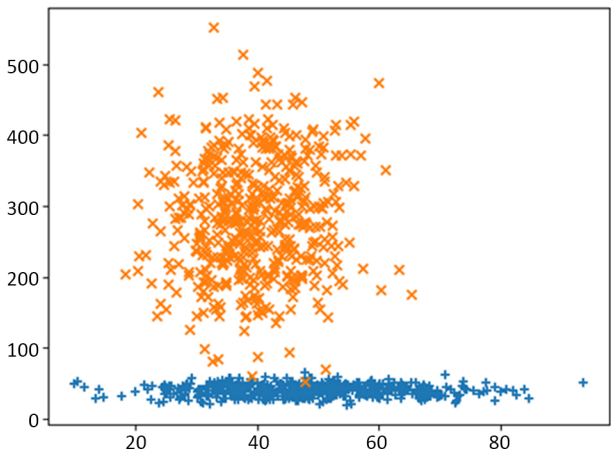

Figure 5. Two types of training set data
图 5. 两类训练集数据

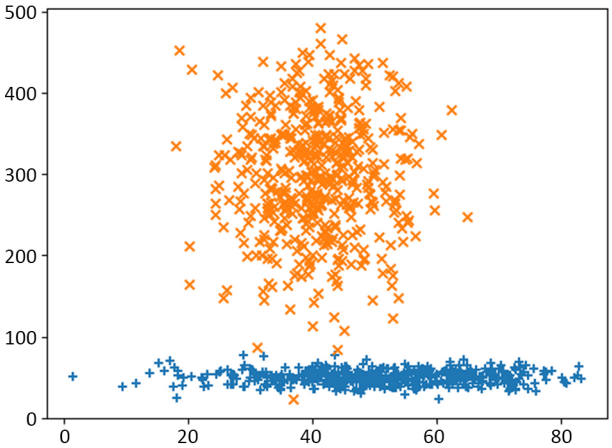

Figure 6. Two types of test set data
图 6. 两类测试集数据

Table 1. The comparison of different algorithm
表 1. 不同分类算法对比结果

噪声数量	分类精度			
	集群	决策树	贝叶斯	KNN
0	0.998	0.998	0.998	0.997
10	0.998	0.978	0.990	0.998
20	0.997	0.951	0.981	0.996
30	0.998	0.929	0.98	0.995
40	0.998	0.92	0.988	0.983
50	1.000	0.898	0.989	0.994

本实验基于 PyCharm Community Edition 2017.1，且各种成熟算法均直接调用自 sklearn 库。

4.2. 实验结果与分析

由于本算法基于集群的大量数据之间的分布特征进行分类，故本算法对异常数据不敏感。为了体现本算法的优势，在测试数据中引入错误标签数据，并与现有成熟分类器：决策树、贝叶斯、KNN 进行分类精度对比，结果见表 1。

从实验结果可见本文所提出的算法性能更为稳定。面对数据分类中，经常出现的人为标签错误问题拥有更优越的分类能力。

5. 结束语

本文首先论证了基于熵权的特征向量表示不同类别数据的可行性。在此基础上提出了基于特征熵权的集群分类算法。经实验验证，该算法对特征分布差异明显的数据集具有相当高的分类精度，且比起现有分类算法对异常数据更为不敏感，一定程度上解决了人为标签容易出错的问题。

基金项目

中国国家自然科学基金国际青年科学家基金(NSFC Grant No.61550110248)；西藏自治区重点科研项目(批准号：Z2014A18G2-13)。

致 谢

这项工作由中国国家自然科学基金国际青年科学家基金(NSFC Grant No.61550110248)和西藏自治区重点科研项目(批准号:Z2014A18G2-13）资助。同时，作者要感谢编辑和审稿人的认真负责。

参考文献

[1] 罗可, 林睦纲, 郗东妹. 数据挖掘中分类算法综述[J]. 计算机工程, 2005, 31(1).

[2] Quinlan, J.R. (1979) Discovering Rules by Induction from Large Collections of Examples. In: *Expert System in the Micro Electronic Age*, 26-37.

[3] Trendowciz, A. and Jeffery, R. (2014) Classification and Regression Trees. *Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery*, **1**, 14-23.

[4] Quinlan, J.R. (1993) C4.5:Programs for Machine Leaning. *Morgan Kauffman*, 23-30.

[5] Flach, P.A. and Lachiche, N. (2004) Naive Bayesian Classification of Structured Data. *Machine Learning*, **57**, 233-269. https://doi.org/10.1023/B:MACH.0000039778.69032.ab

[6] Hsieh, H.Y. and Chen, N. (2012) Recognising Daytime and Nighttime Driving Images Using Bayes Classifier. *IET Intelligent Transport Systems*, **6**, 482-493. https://doi.org/10.1049/iet-its.2010.0153

[7] Feng, X.D., Li, S.C., Yuan, C., Zeng, P. and Sun, Y. (2018) Prediction of Slope Stability Using Naive Bayes Classifier. *KSCE Journal of Civil Engineering*, **22**, 941-950. https://doi.org/10.1007/s12205-018-1337-3

[8] Zhang, N.N., Wu, L.F., Yang, J. and Guan, Y. (2018) Naive Bayes Bearing Fault Diagnosis Based on Enhanced Independence of Data. *Sensors* (Basel, Switzerland), **18**.

[9] Zhang, M.L. and Zhou, Z.H. (2007) ML-KNN: A Lazy Learning Approach to Multi-Label Leaming. *Pattern Recognition*, **40**, 2038-2048. https://doi.org/10.1016/j.patcog.2006.12.019

[10] Peterson, L. (2009) K-Nearest Neighbor. *Scholarpedia*, **4**, 156. https://doi.org/10.4249/scholarpedia.1883

[11] Denoeux, T. (2008) A k-Nearest Neighbor Classification Rule Based on Dempster-Shafer Theory. *IEEE Transactions on Systems Man & Cybernetics*, **25**, 804-813. https://doi.org/10.1109/21.376493

[12] Lemeshow, S. (2004) Applied Logistic Regression. *Journal of the American Statistical Association*, **85**, 121-136.

[13] Keating, K.A. and Cherry, S. (2009) Use and Interpretation of Logistic Regression in Habitat-Selection Studies. *Journal of Wildlife Management*, **68**, 774-789. https://doi.org/10.2193/0022-541X(2004)068[0774:UAIOLR]2.0.CO;2

[14] Wooff, D. (2004) Logistic Regression: A Self-Learning Text, 2nd. *Journal of the Royal Statistical Society*: *Series A* (*Statistics in Society*), **91**, 433. https://doi.org/10.1111/j.1467-985X.2004.298_12.x

[15] Hong, S., You, T., Kwak, S., *et al*. (2015) Online Tracking by Learning Discriminative Saliency Map with Convolutional Neural Network. *International Conference on Machine Learning*, Lille, 6-11 July 2015, 597-606.

[16] Zhang, M.M., Li, W. and Du, Q. (2018) Diverse Region-Based CNN for Hyperspectral Image Classification. *IEEE Transactions on Image Processing*, **27**, 2623-2634. https://doi.org/10.1109/TIP.2018.2809606

[17] Wang, J.Y. and Zhang, C. (2018) Software Reliability Prediction using a Deep Learning Model Based on the RNN Encoder-Decoder. *Reliability Engineering & System Safety*, **170**, 73-82. https://doi.org/10.1016/j.ress.2017.10.019

[18] 杜景林, 严蔚岚. 基于距离权值的 C4.5 组合决策树算法[J]. 计算机工程与设计, 2018(1): 96-102.

[19] 童先群, 周忠眉. 基于属性值信息熵的 KNN 改进算法[J]. 计算机工程与应用, 2010, 46(3): 115-117.

[20] Shannon, C.E. (1948) A Mathematical Theory of Communication. *Bell System Technical Journal*, **27**, 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[21] Wang, Z. and Zhan, W. (2012) Dynamic Engineering Multi-Criteria Decision Making Model Optimized by Entropy Weight for Evaluating Bid. *Systems Engineering Procedia*, **5**, 49-54. https://doi.org/10.1016/j.sepro.2012.04.008

[22] Simic, S. (2009) Jensen's Inequality and New Entropy Bounds. *Applied Mathematics Letters*, **22**, 1262-1265. https://doi.org/10.1016/j.aml.2009.01.040

[23] Zheng, K.F. and Wang, X.J. (2018) Feature Selection Method with Joint Maximal Information Entropy between Features and Class. *Pattern Recognition*, **77**, 20-29. https://doi.org/10.1016/j.patcog.2017.12.008

[24] 陈运, 周亮, 陈新, 陈建伟. 信息论与编码[M]. 北京: 电子工业出版社, 2015: 13-19.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org