# 基于稀疏表示和特征加权的大数据挖掘方法的研究

蔡柳萍[1] 解 辉[2] 张福泉[3] 张龙飞[3]

(广东技术师范学院天河学院计算机科学与工程学院 广州 510540)[1]

(清华大学计算机科学与技术系 北京 100084)[2] (北京理工大学软件学院 北京 100081)[3]

**摘 要** 为了提高大数据挖掘的效率及准确度,文中将稀疏表示和特征加权运用于大数据处理过程中。首先,采用求解线性方程稀疏解的方式对大数据进行特征分类,在稀疏解的求解过程中利用向量的范数将此过程转化为最优化目标函数的求解。在完成特征分类后进行特征提取以降低数据维度,最后充分结合数据在类中的分布情况进行有效加权来实现大数据挖掘。实验结果表明,相比于常见的特征提取和特征加权算法,提出的算法在查全率和查准率方面均呈现出明显优势。

**关键词** 大数据,数据挖掘,特征加权,特征提取,稀疏表示

**中图法分类号** TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.11.040

## Study on Big Data Mining Method Based on Sparse Representation and Feature Weighting

CAI Liu-ping[1] XIE Hui[2] ZHANG Fu-quan[3] ZHANG Long-fei[3]

(School of Computer Science & Engineering, Tianhe College of Guangdong Polytechnic Normal University, Guangzhou 510540, China)[1]

(Department of Computer Sciences and Technology, Tsinghua University, Beijing 100084, China)[2]

(School of Software, Beijing Institute of Technology, Beijing 100081, China)[3]

**Abstract** In order to improve the efficiency and accuracy of big data mining, this paper applied the sparse representation and feature weighting into big data processing. At first, the features of big data are classified by solving the sparse mode of linear equation. In the process of solving the sparse solution, a vector norm is utilized to transform this process into the process of solving the optimization objective function. After feature classification, feature extraction is executed to reduce the dimensionality of data. Finally, the distribution of data in the class is combined sufficiently to conduct weighting effectively, thus realizing data mining. The experimental results suggest that the proposed algorithm is superior to the common feature extraction and feature weighting algorithms in the terms of recall and precision.

**Keywords** Big data, Data mining, Feature weighting, Feature extraction, Sparse representation

互联网的发展促使数据网络化管理进入新的阶段,网络中的大数据资源成为互联网跨行业发展的动力源,不同行业的大数据资源为企业的发展提供了有力帮助。由于网络平台的后台数据极其丰富,因此不论是对用户数据进行分析,还是对用户行为习惯进行分析,或者研究数据之间有价值的联系等,都意味着大数据的分析具有重要意义。大数据不仅数据量庞大,其数据结构更为复杂,如何充分运用大数据所包含的有效资源,对于大数据的管理及分析极其关键。

数据挖掘作为当前人工智能及数据分析的重要手段,能够从大量数据中完成具有规律性的有价值信息的检索与提取[1]。当前数据挖掘算法较多,大部分算法都是对数据进行分类整理,以寻求数据背后的联系与统计特征[2],或者是通过聚类与模糊处理等来寻找大数据中有规律、有价值的数据,其最终目的是根据数据挖掘结果来寻求决策支持。本文提出一种基于稀疏表示和特征加权的方法来完成大数据挖掘,该方法在数据的查全率和查准率方面具有明显优势。

## 1 大数据的稀疏表示

稀疏表示理论最早应用于通信领域,它通过稀疏系数来实现信号的还原,其优点是用尽可能少的数据表示出一整串

到稿日期:2018-02-14 返修日期:2018-04-23 本文受文化部国家科技支撑计划项目(2012BAH38F00),广东省本科高校应用型人才培养课程建设项目:能力培养导向的计算机类应用型课程建设(2017SZ03),广东省科技计划项目:基于医药电商大数据的服务系统研发(2016A010101029),广东技术师范学院天河学院计算机科学与技术重点学科建设项目(Xjt201702)资助。

蔡柳萍(1981—),女,硕士,讲师,高级工程师,主要研究方向为大数据、软件工程研究等,E-mail:4425335@qq.com(通信作者);解 辉(1981—),男,博士,副教授,主要研究方向为计算网络、大数据处理;张福泉(1975—),男,博士,副教授,CCF 会员,主要研究方向为创意计算、大数据;张龙飞(1977—),男,博士,副教授,主要研究方向为数字媒体科学、大数据。

信号。而目前稀疏表示理论被频繁应用于图像处理领域,该理论较好地解决了图像信号存储量大的问题。通过稀疏表示可以用较小的数据特征来实现整幅图像的存储,节约了存储空间。在稀疏表示描述过程中,通过离散傅立叶变换等计算方法完成大数据信号的变换[3]。

设测试样本集合为 $A_i=[v_{i,1},v_{i,2},\cdots,v_{i,n_i}]$,关于系数 $\alpha_{i,j}(j=1,2,\cdots,n_i)$ 的线性组合为:

$$y=\alpha_{i,1}v_{i,1}+\alpha_{i,2}v_{i,2}+\cdots+\alpha_{i,n_i}v_{i,n_i} \tag{1}$$

在式(1)的求解过程中,考虑到样本集合的类别 $i$ 未知,故定义新的矩阵 $A$:

$$A=[A_1,A_2,\cdots,A_k]=[v_{1,1},v_{1,2},\cdots,v_{k,n_k}] \tag{2}$$

新的矩阵 $A$ 是一个样本集合,样本集合的列向量有 $k$ 个,表示类别总数为 $k$;样本集合的行向量有 $n$ 个,表示集合有 $n$ 个数据样本。结合式(1)和式(2),可以用式(3)表示系数 $\alpha_{i,j}$ 的线性组合:

$$y=Ax_0 \tag{3}$$

其中,$x_0=[0,\cdots,0,\alpha_{i,1},\alpha_{i,2},\cdots,\alpha_{i,n_i},0,\cdots,0]^{\mathrm{T}}\in R^m$ 表示系数向量集合。

因此,样本集合 $A$ 的分类过程可以转变为求解方程 $y=Ax$ 稀疏解的过程,稀疏解的求解过程可以转变为 $x$ 的最优化求解[4],方法如下:

$$\begin{aligned}&(\ell^0):\hat{x}_0=\arg\min\|x\|_0\\&\text{s.t. } Ax=y\end{aligned} \tag{4}$$

其中,$\|x\|_0$ 表示向量的 $\ell^0$ 范数。在求解过程中,若式(4)的求解结果足够稀疏,则求解方法等价于求解样本集合的 $\ell^1$ 范数问题:

$$\begin{aligned}&(\ell^1):\hat{x}_1=\arg\min\|x\|_1\\&\text{s.t. } Ax=y\end{aligned} \tag{5}$$

上述方程的求解可以参考线性规划方法,在计算过程中,随着样本集合的增加,算法复杂度也随之线性增加。考虑到实际应用环境,噪声和干扰必不可免,这将对集合分类结果的准确性造成影响。因此,结合实际环境,重新修订式(3),得到式(6)[5]:

$$y=Ax_0+z \tag{6}$$

其中,$z\in R^m$ 为实际干扰项,且 $\|z\|_2<\varepsilon$。由此,式(5)变为:

$$\begin{aligned}&(\ell^1):\hat{x}_1=\arg\min\|x\|_1\\&\text{s.t. } \|Ax-y\|_2\leqslant\varepsilon\end{aligned} \tag{7}$$

## 2　基于稀疏表示的数据特征提取与加权

在对大数据进行稀疏表示之后,可以对大数据的存储空间进行有效压缩,将原样本集合数据特征转变为稀疏系数向量集合。接下来再对向量集合进行特征选择及特征提取,一方面是为了降低数据处理的复杂度,另外一方面是为高效率的数据分类做准备。

相比于传统的数据特征分类,基于稀疏表示的特征选择分类及提取算法的过滤性更高,具有较强的筛选性和无监督性。该方法是对 $d$ 维数据先进行稀疏表示,然后根据稀疏系数的大小升序排序,选择前 $k$ 个特征来实现数据的特征选择。具体实现方法如下[6]:

首先,利用式(7)求解稀疏系数。设大数据集合 $\{X_i\}_{i=1}^n$,$X_i\in R^d$,令数据矩阵 $X=[x_1,x_2,\cdots,x_n]\in R^{d\times n}$,利用式(8)求解稀疏系数 $s_i$:

$$\begin{aligned}&\min_{s_i}\|s_i\|_1\\&\text{s.t. } x_i=X's_i\end{aligned} \tag{8}$$

其中,$X'$ 为 $X$ 不包含第 $i$ 列 $x_i$ 的数据矩阵,$s_i=[s_{i1},\cdots,s_{ii-1},0,s_{ii+1},\cdots,s_{im}]^{\mathrm{T}}$ 是一个 $n$ 维系数向量。由于计算 $s_i$ 时,$x_i$ 包含在 $X$ 中,因此将 $s_i$ 中第 $i$ 个元素设置为 0,由此得到整个数据集 $\{X_i\}_{i=1}^n$ 稀疏表示的重构系数矩阵 $S=(\tilde{s_{ij}})_{n\times n}$,其中 $S=[\tilde{s}_1,\tilde{s}_2,\cdots,\tilde{s}_n]$。每个列向量 $\tilde{s}_i$ 是由数据 $x_i$[7-8] 组成,利用式(8)得到的稀疏系数来分析重构系数。

其次,考虑到稀疏矩阵还原原始数据特征可能出现干扰和误差,因此需要在还原数据过程中对重构系数进行修正。具体操作方法是:若当前数据特征与通过重构系数还原的特征相差较小,则认为该特征经过系数表示后还原能力强;若当前数据特征与通过重构系数还原的特征相差较大,则表示通过稀疏表示的特征不足以表达原始数据特征。因此,引入稀疏分数目标函数 $S(r)$[5],计算方法如下:

$$S(r)=\frac{\sum_{i=1}^{n}(x_{ir}-(X\tilde{s}_i)_r)^2}{Var(X(r))} \tag{9}$$

其中,$S(r)$ 的分子 $\sum_{i=1}^{n}(x_{ir}-(X\tilde{s}_i)_r)^2$ 表示样本集合 $x_{ir}$ 与重构还原集合 $(X\tilde{s}_i)_r$ 在第 $r$ 维特征之间的误差之和,$\sum_{i=1}^{n}(x_{ir}-(X\tilde{s}_i)_r)^2$ 越小,表明原始特征与还原特征越接近,算法的重构还原能力越强。$S(r)$ 的分母 $\mathrm{Var}(X(r))$ 表示集合第 $r$ 维特征的方差。

最后,根据式(9)求解 $S(r)$ 值,$S(r)$ 值表示算法通过重构系数还原原始特征的能力,$S(r)$ 的值与还原能力成反比。选取较小的 $S(r)$ 值作为特征选取的依据。

通过式(10)定义中心距离特征 CDF(Center Distance Feature):

$$CDF(t,c_i)=\frac{df_i(t)}{\sum_{i=1}^{m}df_i(t)}\times\frac{df_i(t)}{|c_i|}\times\frac{tf_i(t)}{df_i(t)} \tag{10}$$

其中,$df_i(t)$,$tf_i(t)$ 分别表示某一特征 $t$ 在样本集合的某一个类 $c_i$ 中出现的文档频数与文本词语频率,$\sum_{i=1}^{m}df_i(t)$ 表示集合文档的总数目,$|c_i|$ 表示包含 $c_i$ 类的所有文档数目[9]。式(10)表示了 3 部分乘积,$\frac{df_i(t)}{\sum_{i=1}^{m}df_i(t)}$,$\frac{df_i(t)}{|c_i|}$ 和 $\frac{tf_i(t)}{df_i(t)}$ 分别反映了特征 $t$ 对于 $c_i$ 类的集中程度、分布的均匀程度及平均程度[10]。化简式(10)可得:

$$CDF(t,c_i)=\frac{df_i(t)}{\sum_{i=1}^{m}df_i(t)}\times\frac{tf_i(t)}{|c_i|} \tag{11}$$

特征提取的实质是用某一个或者多个特征词来表示一类集合,该特征词对类别的辨识度要求高,若该特征词能够同时

表示很多类，则该特征词对类的分类贡献度小，并不适合作为区分类的特征词。正因为这个原因，需要对式(11)的求解方法采取更多策略。本文采用最大值的方法完成求解，如式(12)所示：

$$CDF_{\max}(t)=\max_{1\leqslant i\leqslant m}\{CDF(t,c_i)\} \tag{12}$$

*CDF* 值的大小是特征词选择的标准，系统首先设定阈值[11]，若 *CDF* 大于该阈值，则该词为特征提取的依据值。

TF-IDF 特征加权方法是由 Salton 教授提出的，该方法主要由特征项频度(TF)和反文档频度(IDF)两部分构成[12]。计算方法为：

$$Weight_{\text{TF-IDF}}(d_t)=k\times\log(\frac{N}{n}) \tag{13}$$

其中，$k$ 表示特征 $t$ 在本文 $d$ 中出现的次数，$N$ 为文本总数，$n$ 为包含特征 $t$ 的文本数目。

TF-IDF 运用特征词来区分文档类别[13]，却忽略了类的分布情况。本文将 TF-IDF 算法与式(11)和式(12)的特征提取算法相结合，得到改进的 TF-IDF 算法——TF-IDF-CD 算法，计算方法如下：

$$Weight(d_t)=k\times\log(\frac{N}{n})\times\max_{1\leqslant i\leqslant m}\{cd(t,i)\} \tag{14}$$

其中，$k$ 为特征 $t$ 在本文 $d$ 中出现的次数，$N$ 为所有文本数目，$n$ 为含有特征 $t$ 的文本数，$cd(t,i)$ 为：

$$cd(t,i)=\frac{df_i(t)}{\sum_{i=1}^{m}df_i(t)}\times\frac{tf_i(t)}{|c_i|} \tag{15}$$

## 3 实例仿真

### 3.1 实验介绍

为了验证本文算法在特征提取与特征加权中的性能，分别对 CDF 特征提取性能及 TF-IDF-CD 的特征加权性能进行对比实验。实验中，主要考虑算法的查全率、查准率、准确率等指标[14]。其中，特征提取实验将本文算法与常见的特征提取算法进行性能对比，重点对比 CDF 特征提取方法和传统的计算方法的指标值。特征加权方面，主要对比本文提出的 TF-IDF-CD 算法和传统的 TF-IDF 算法之间的性能差异。

### 3.2 实验数据来源

本文采用的资料库来自于某大型网络学习平台，从该平台中选取 6 个类别的文档，包括计算机、经济、交通、体育、环境、教育，样本集合文档有 1323 篇，测试集合文档有 655 篇。实验数据的具体分布情况如表 1 所列。

表 1 实验数据

Table 1 Experimental data

(单位：篇)

| 类别 | 计算机 | 经济 | 交通 | 体育 | 环境 | 教育 |
|---|---|---|---|---|---|---|
| 样本文档 | 196 | 261 | 178 | 216 | 232 | 240 |
| 测试文档 | 96 | 130 | 89 | 107 | 115 | 118 |

总样本集合与测试集合的比例约为 2∶1，表 1 中选取的各类别文档的比例也约为 2∶1。

### 3.3 特征提取实验

本实验选取当前常用的特征提取算法，即信息增益(IG)、期望交叉熵(CE)、CHI 和文本证据权(WET)等方法[15]，与本文提出的 CDF 算法进行提取实验性能对比，评价指标采用查全率(Recall)、查准率(Precision)和准确率[14]。经过实验计算，5 种算法的 Recall(R)与 Precision(P)如表 2 所列。

表 2 不同特征提取算法的查全率和查准率对比

Table 2 Comparison of recall and precision in different feature extraction algorithms

(单位：%)

| 特征提取方法 | 评价指标 | 计算机 | 经济 | 交通 | 体育 | 环境 | 教育 |
|---|---|---|---|---|---|---|---|
| IG | *R* | 97.22 | 93.11 | 95.23 | 96.35 | 67.23 | 87.76 |
| | *P* | 93.68 | 73.88 | 98.22 | 92.11 | 71.03 | 84.38 |
| CE | *R* | 98.02 | 92.52 | 93.28 | 97.34 | 63.46 | 90.32 |
| | *P* | 91.12 | 74.21 | 94.12 | 91.06 | 68.03 | 88.43 |
| WET | *R* | 95.67 | 91.65 | 95.18 | 93.24 | 72.99 | 83.78 |
| | *P* | 92.12 | 73.86 | 95.66 | 91.27 | 67.13 | 85.93 |
| CHI | *R* | 94.06 | 91.87 | 93.67 | 93.33 | 62.14 | 86.77 |
| | *P* | 92.16 | 72.63 | 94.26 | 91.78 | 61.99 | 87.92 |
| CDF | *R* | 98.23 | 94.12 | 95.81 | 96.23 | 79.72 | 86.87 |
| | *P* | 94.22 | 78.96 | 99.22 | 94.93 | 72.37 | 85.32 |

为了直观反映各种算法在查全率和查准率方面的性能，更便于对比算法的优劣，将表 2 中的数据用直方图进行表示，具体如图 1 和图 2 所示。

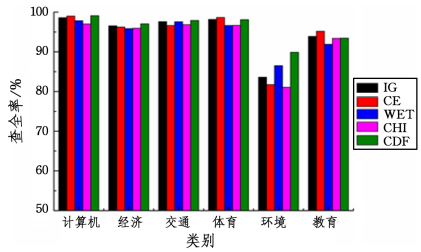


图 1 不同特征提取算法的查全率比较

Fig. 1 Comparison of recall of different feature extraction algorithms

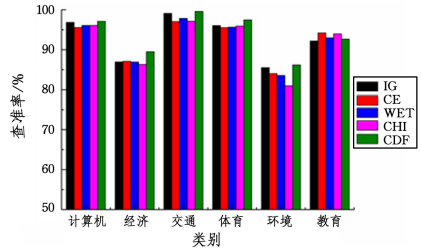


图 2 不同特征提取算法的查准率比较

Fig. 2 Comparison of precision of different feature extraction algorithms

对比表 2、图 1、图 2 可得，在查全率方面，CDF 算法在计算机、经济、交通和环境类别中，性能排在第一位。特别是在环境类别中，CDF 算法的查全率为 79.72%，IG，CE 和 CH 算法的查全率均低于 70%，本文算法明显优于 IG，CE 和 CHI 算法。在体育和教育方面，虽然 CDF 算法的性能并不在第一位，但是 CDF 算法和其他特征提取算法的差距非常小。综合而言，本文算法在查全率指标上表现出色。

对比表 2、图 1、图 2 可知，在查准率方面，CDF 算法在计算机、经济、交通和体育类别中，性能排在第一位，在交通类别

中，查准率达到 99.22%。相比于查全率，各种特征提取算法在查准率方面性能表现差别不是特别大，各种算法性能表现均衡。

综合比较，CDF 算法在计算机、经济和交通 3 个方面的查全率和查全率均取得了最优成绩，优势明显。

### 3.4　特征加权实验

本节实验主要验证 TF-IDF 算法和 TF-IDF-CD 算法的性能，算法评价指标主要采用 $F1$ 值进行对比。通过计算得到两种加权算法在不同特征提取算法下的 $F1$ 值，如表 3 所列。

表 3　不同特征加权法的 $F1$ 值

Table 3　$F1$ values of different feature weighting algorithms

| 特征提取方法 | 算法 | $F1$/% | | | | | |
|---|---|---|---|---|---|---|---|
| | | 计算机 | 经济 | 交通 | 体育 | 环境 | 教育 |
| IG | TF-IDF | 95.15 | 83.11 | 94.23 | 96.10 | 76.23 | 87.44 |
| | TF-IDF-CD | 97.28 | 87.55 | 96.72 | 97.11 | 81.07 | 87.78 |
| CE | TF-IDF | 97.02 | 82.52 | 93.37 | 96.74 | 77.46 | 86.23 |
| | TF-IDF-CD | 98.12 | 86.31 | 94.92 | 97.09 | 76.53 | 87.56 |
| WET | TF-IDF | 93.37 | 81.75 | 94.23 | 93.24 | 72.13 | 82.13 |
| | TF-IDF-CD | 94.29 | 89.76 | 95.66 | 93.76 | 77.79 | 83.62 |
| CHI | TF-IDF | 93.18 | 83.27 | 92.12 | 92.09 | 78.38 | 82.05 |
| | TF-IDF-CD | 94.29 | 85.18 | 93.79 | 94.85 | 82.09 | 83.31 |
| CDF | TF-IDF | 98.07 | 85.72 | 95.77 | 93.23 | 79.72 | 88.57 |
| | TF-IDF-CD | 98.99 | 89.67 | 97.02 | 94.95 | 81.37 | 90.11 |

将表 3 中的数据进行柱状图绘制，以清晰地反映不同特征提取算法下对特征加权后的 $F1$ 值的对比结果，如图 3—图 7 所示。

从图 3 可以得出，在所有类别中，相比于 TF-IDF 算法，本文算法的 $F1$ 值更高，特别是在环境类别中，本文算法比 TF-IDF 算法高出近 5 个百分点，因此本文改进算法在该性能方面表现更出色。从图 4 可以看出，除了环境类别之外，所有类别的分类中，本文算法的 $F1$ 值均超越了 TF-IDF 算法。

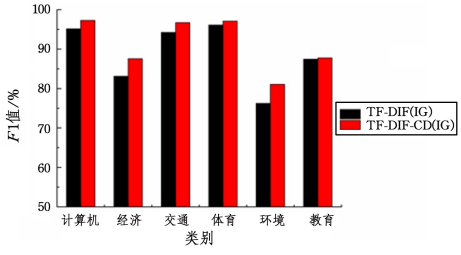


图 3　采用 IG 特征提取算法后不同加权算法的 $F1$ 值对比

Fig. 3　Comparison of $F1$ values of different weighting algorithms after using IG algorithm

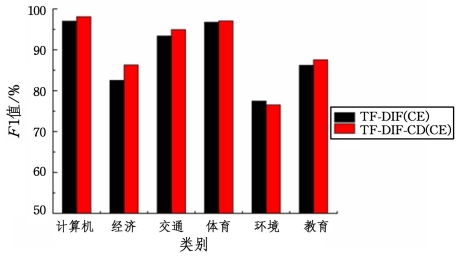


图 4　采用 CE 提取算法后不同加权算法的 $F1$ 值对比

Fig. 4　Comparison of $F1$ values of different weighting algorithms after using CE algorithm

结合图 5—图 7 可以看出，在 6 个类别的提取分类实验中，本文算法的 $F1$ 值均高于传统算法；本文提出的 TF-IDF-CD 加权算法的性能更优，特别在采用 CDF 特征提取算法之后，本文加权算法的 $F1$ 值相比于其他特征提取算法更高。

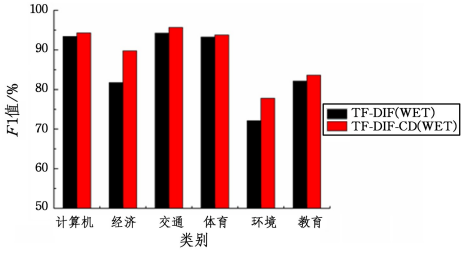


图 5　采用 WET 特征提取算法后不同加权算法的 $F1$ 值对比

Fig. 5　Comparison of $F1$ values of different weighting algorithms after using WET algorithm

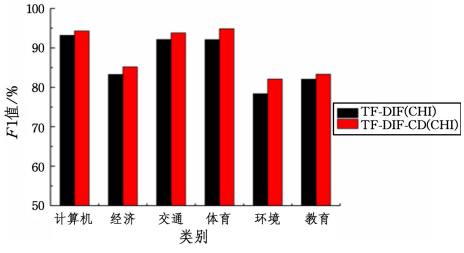


图 6　采用 CHI 特征提取算法后不同加权算法的 $F1$ 值对比

Fig. 6　Comparison of $F1$ values of different weighting algorithms after using CHI algorithm

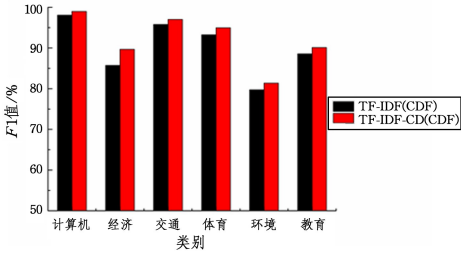


图 7　采用 CDF 特征提取算法后不同加权算法的 $F1$ 值对比

Fig. 7　Comparison of $F1$ values of different weighting algorithms after using CDF algorithm

**结束语**　本文将稀疏表示方法与特征加权相结合，在特征选择与特征提取方面，充分考虑了类与类之间、类自身内部的集中度和均匀度等因素，完成了特征词的提取。接着结合常用的 TF-IDF 算法提出 TF-IDF-CD 特征加权算法来实现特征加权，以完成数据挖掘。经特征提取与加权实验证明，本文算法优于常见的数据挖掘算法。

## 参 考 文 献

[1]　LIANG J Y. Challenges and Reflections on large data mining [J]. Computer Science, 2016, 43(7): 1-2. (in Chinese)
梁吉业. 大数据挖掘面临的挑战与思考[J]. 计算机科学, 2016, 43(7): 1-2.

[2]　FENG Z, ZHU Y. A Survey on Trajectory Data Mining: Techniques and Applications[J]. IEEE Access, 2017, 4: 2056-2067.

[3] ZHANG Z,XU Y,YANG J,et al. A Survey of Sparse Representation:Algorithms and Applications[J]. IEEE Access,2017,3:490-530.

[4] LIU L,TRAN T D,SANG P C. Partial face recognition:A sparse representation-based approach[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. IEEE,2016:2389-2393.

[5] QIU D,LIU Y. Improved image super-resolution via sparse representation[J]. Video Engineering,2016,12(8):100-104.

[6] BOLÓN-CANEDO V,SÁNCHEZ-MAROÑO N,ALONSO-BETANZOS A. Feature selection for high-dimensional data[J]. Progress in Artificial Intelligence,2016,5(2):65-75.

[7] LIU J H,LIN M L,ZHANG J,et al. A kind of heuristic local random feature selection algorithm[J]. Computer Engineering and Applications,2016,52(2):170-174. (in Chinese)
刘景华,林梦雷,张佳,等. 一种启发式的局部随机特征选择算法[J]. 计算机工程与应用,2016,52(2):170-174.

[8] ZHANG Z,HANCOCK E R. A Graph-Based Approach to Feature Selection[C]// International Conference on Graph-Based Representations in Pattern Recognition. Springer-Verlag,2017:205-214.

[9] RIVEROMORENO C J,BRES S. Texture Feature Extraction and Indexing by Hermite Filters[C]// International Conference on Pattern Recognition. IEEE,2017:684-687.

[10] JIANG F,LI G H,YUE X. Semantic-based Feature Extraction Method for Document[J]. Computer Science,2016,43(2):254-2589. (in Chinese)
姜芳,李国和,岳翔. 基于语义的文档特征提取研究方法[J]. 计算机科学,2016,43(2):254-258.

[11] ZHOU G,CICHOCKI A,ZHANG Y,et al. Group Component Analysis for Multiblock Data:Common and Individual Feature Extraction[J]. IEEE Transations Neural Networks Learning Systems,2016,27(11):2426-2439.

[12] XIAO L Y,CHEN X H,LIN X L. Feature Weighted and Improved Partition Fuzzy C-Means Cluster Algorithm[J]. Microelectronics &Computer,2016,33(10):143-146. (in Chinese)
肖林云,陈秀宏,林喜兰. 特征加权和优化划分的模糊 C 均值聚类算法[J]. 微电子学与计算机,2016,33(10):143-146.

[13] ZHANG L,JIANG L,LI C,et al. Two feature weighting approaches for naive Bayes text classifiers[J]. Knowledge-Based Systems,2016,100(C):137-144.

[14] LUO Y,ZHAO S L,LI X C,et al. Text keyword extraction method based on word frequency statistics[J]. Journal of Computer Applications,2016,36(3):718-725. (in Chinese)
罗燕,赵书良,李晓超,等. 基于词频统计的文本关键词提取方法[J]. 计算机应用,2016,36(3):718-725.

[15] CHEN Z,XIA J B,BAI J,et al. Feature Extraction Algorithm Based on Evolutionary Deep Learning[J]. Computer Science,2015,42(11):288-292. (in Chinese)
陈珍,夏靖波,柏骏,等. 基于进化深度学习的特征提取算法[J]. 计算机科学,2015,42(11):288-292.

[16] ZENG Q S,HUANG X Y. Fast Data Mining Algorithm Based on FP-tree[J]. Journal of Chongqing University of Technology (Natural Science),2009,23(10):72-76. (in Chinese)
曾庆森,黄贤英. 基于 FP-tree 的快速数据挖掘算法[J]. 重庆理工大学学报(自然科学),2009,23(10):72-76.

(上接第 225 页)

[13] YIN X X,HAN J W,PHILIP S Y. Object Distinction:Distinguishing Objects with Identical Names[C]// Proceedings of International Conference on Data Engineering(ICDE). 2007:1242-1246.

[14] XU R F,GUI L,LU Q,et al. Incorporating Multi-kernel function and Internet Verification for Chinese Person Name Disambiguation[J]. Frontiers of Computer Science,2016,10(6):1-13.

[15] HIEN T N,TRU H C. Named Entity Disambiguation:A Hybird Statistical and Rule-Based Incremental Approach[C]// Proceedings of the Semantic Web:the 3th Asian Semantic Web Conference(ASWC). 2008:420-433.

[16] FU J L,QIU J,GUO Y L,et al. Entity Linking and Name Disambiguation Using SVM in CHINESE Micro-blogs[C]// Proceedings of International Conference on Natural Computation. IEEE,2016:468-472.

[17] LI Y P. Bibliometric Analysis and Name Disambiguation Research Based on Knowledge Clustering[D]. Nanjing:Nanjing University of Posts and Telecommunications,2016. (in Chinese)
李永萍. 基于知识聚类的文献统计与重名消歧机制的研究[D]. 南京:南京邮电大学,2016.

[18] MIN S,ERIN H K,HA J K. Exploring author name disambiguation on PubMed-scale[J]. Journal of Informetrics,2015,9(4):924-941.

[19] MU L M. Research of the Nature & Operation of Finite Multi-Set[J]. Journal of Neijiang Normal University,2009,24(4):5-8. (in Chinese)
牟廉明. 有限多重集的运算及性质[J]. 内江师范学院学报,2009,24(4):5-8.

[20] TRAVERS J,MILGRAM S. An Experimental Study of the Small World Problem[J]. Sociometry,1969,32(4):425-443.

[21] MONGLI L,TOK W L,WAI L L. Intelliclean:A Knowledge-Based Intelligent Data Cleaner[C]// ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. 2000:290-294.