基于模糊C均值和改进的LSA的文档聚类研究

胡永丽，龚沛曾
（同济大学 电子与信息工程学院 计算机科学与技术系，上海 201804）

摘 要：文中研究的是文档聚类的方法，即将给定文档集合中的文档进行分类，以达到准确聚类的目的。提出了一种将模糊C均值(FCM)和改进的LSA(Latent Semantic Analysis)相结合进行文档聚类的方法。采用改进的词语特征提取方法构建词－文档矩阵，对该词－文档矩阵进行奇异值分解，从传统的VSM向量空间中提取文本的潜在语义空间，进而将高维的文档向量映射为低维空间的语义向量，文档之间相似度的计算采用文档语义向量的余弦表示。然后采用模糊C均值根据上述计算文档相似度的结果对文档进行聚类。针对校园论坛中的文档数据进行聚类，该方法降低了处理的复杂度同时提高了相似度计算的准确性。实验结果表明该方法对目标文档的聚类有较好的效果，聚类准确性较高。

关键词：模糊C均值；LSA；文档聚类
中图分类号：TP391.1 **文献标识码**：A **文章编号**：1673－629X(2010)12－0126－04

Document Clustering Research Based on Fuzzy C－Means and Improved Latent Semantic Analysis

HU Yong-li, GONG Pei-zeng
(Dept. of Computer Science and Technology, Electronics and Information Engineering College, Tongji University, Shanghai 201804, China)

Abstract: This paper is focused on the methods of document clustering, that is to classify the documents in the document set so as to achieve the aim of accurate clustering. Proposed a method which combines the Fuzzy C－means with improved LSA to do document clustering. A new method of feature extraction was used to construct term－document matrix. Do singular value decomposition for the matrix, extract the document's latent semantic space from the traditional VSM vector space so as to change the document vector of high dimension to semantic vector of low dimension. Use cosine between the documents semantic vectors to present the similarity between documents. Then use Fuzzy C－means to do document clustering based on the results of similarity calculation above. Do the experiment on the documents data of campus forum, this method reduces the computer processing complexity and improves the veracity of similarity calculation. Experimental result shows that the proposed method can get better document clustering effect and the accuracy of clustering is high.

Key words: fuzzy c－means; LSA; document clustering

0 引 言

随着网络的发展与普及，互联网上的海量信息都以静态或动态网页的形式存在，通常是无固定格式的、非结构的。从这些海量信息中准确、快速地获取用户感兴趣的、有用的信息难度是非常大的。文档聚类技术是搜索引擎、信息检索、文本挖掘、自动问答、数字化图书馆等领域的技术基础，在多文档自动文摘的预处理、对搜索引擎返回的结果进行聚类、改善文本分类的结果、文档集合的自动整理等领域有着非常重要的应用。

文档聚类指的是将文档集合中的文档分为更小的簇，要求同一簇内文档之间相似性尽可能大[1]。文档聚类主要是依据著名的聚类假设：同类的文档相似度较大，而不同类的文档相似度较小。文档聚类是一种无监督的机器学习方法，聚类不需要训练过程，以及不需要预先对文档进行手工标记类别，因此具有一定的灵活性和较高的自动化处理能力。

文档聚类的方法有很多，常用的文档聚类方法是基于向量空间模型(VSM)的方法。VSM本质上将文档中的词语看作空间的维度，每个词语对应着一个权值，将包含词语的文档表示为一个词语向量并对应为空间中的一个点，文档之间的相似度便是向量空间中对应两个点之间的距离，将相似度大，即两点之间距离

收稿日期：2010－04－11；**修回日期**：2010－07－25
作者简介：胡永丽(1984－)，女，内蒙古呼和浩特人，硕士生，研究方向为图像处理、模式识别；龚沛曾，教授，上海市名师，硕士生导师，研究方向为模式识别、智能系统。

近的文档分为一类。采用传统的基于 VSM 的方法进行文档聚类有两大缺陷:首先,文档所包含的词汇量巨大,故生成的文档向量通常是高维的,造成许多传统的算法难以处理[2];另外,基于 VSM 的方法是一种基于统计的方法,考虑的是文档中关键词语的词频信息,并没有考虑词语在文档上下文中的语义信息,加上文档中同义词与多义词的干扰,造成聚类的不准确。文中采用了改进的 LSA 模型来代替传统的 VSM。LSA 是一种用于知识获取和展示的计算理论和方法,它使用统计计算的方法对大量的文本集进行分析,从而表示及提取出词的语义,这种潜在语义是词语所在的上下文语境信息的总和。在 LSA 中对文档的分类和检索充分考虑到了词语之间的相关性,同时可以消除同义词与多义词的影响,提高文档表示的准确性。潜在语义分析的关键就是特征的提取,文中采用的改进 LSA 模型在原有 TF - IDF 特征选择算法中加入了文档长度和熵权重两个因素,更好地进行特征的提取以构造词 - 文档矩阵。

模糊 C 均值(FCM)聚类算法作为一种无监督聚类算法已成功应用在数据分类等领域。它的思想是使得被划分到同一簇的对象之间相似度较大,而不同簇之间的相似度较小。传统的聚类分析是一种硬划分,它把每个待识别的对象严格地划分到某个类中,具有非此即彼的性质,这种分类的类别界限是十分明确的,但是实际上大多数待分类的对象并没有严格的特征界限,它们在各种属性上存在一定的中介性,FCM 就是一种柔性的模糊划分,以一种模糊的方法来处理聚类问题,其算法简单、收敛速度快,且能处理大数据集、解决问题范围广、易于计算机实现,所以被应用于很多领域。它是一种非常有效的模糊聚类算法,使用每个样本隶属于某个聚类的隶属度,即使对于很难分类的变量,FCM 也能够得到比较满意的聚类效果[3]。

1　改进的 LSA

1.1　LSA 基本思想

潜在语义分析(LSA)是由 Scoot Deerwester, Thomas K. Landauer 等五位学者于 1990 年提出的一种自然语言处理方法[4]。是一种用于知识获取和展示的计算理论和方法,它使用统计计算的方法对大量的文本集进行分析,从而提取和表示出词的语义,这种潜在语义是词语所有的上下文语境信息的总和。这是因为,上下文环境对其中的事物提供了一组相互联系和制约,在很大程度上决定了词语之间语义上的相关性[5]。

LSA 的出发点就是认为文档中词与词之间存在某种潜在的语义结构,这种语义结构隐藏在词语使用的上下文中,并且认为同义词之间具有相同的语义结构,而多义词具有不同的语义结构,故采用统计的方法提取特征词语并量化其中潜在的语义结构,从而用语义结构来表示词语和文本。在潜在语义空间中,文档和特征词根据语义上的相关程度被分散在不同的位置,处于不同文档的同义词空间位置相邻,包含不同特征词但语义相近的文档空间位置相邻,包含相同特征词但语义不同的文档空间位置也比较远,从而可以消除同义词与多义词的影响,提高文档表示的准确性。

实现 LSA 首先是构造一个词 - 文档矩阵。假设文本集中包含 n 个文档,用到了 m 个词语,构造一个 $m \times n$ 的词语文档矩阵 $\boldsymbol{A}_{m\times n} = [\boldsymbol{A}_{ij}] = (\mathrm{doc}_1, \mathrm{doc}_2, \cdots, \mathrm{doc}_n) = (\mathrm{term}_1, \mathrm{term}_2, \cdots, \mathrm{term}_m)^T$,其中 $\boldsymbol{A}_{ij}$ 表示词语 i 在文档 j 中出现的频率,term_i 和 doc_j 分别代表词语和文档的列向量。利用数学方法奇异值分解(SVD)理论对矩阵 $\boldsymbol{A}$ 进行分解得 $\boldsymbol{A} = \boldsymbol{USV}^T$,其中 $\boldsymbol{U}$ 是 $m \times m$ 的正交矩阵,$\boldsymbol{U}$ 的列称为矩阵 $\boldsymbol{A}$ 的左奇异值向量;$\boldsymbol{V}$ 是 $n \times n$ 的正交矩阵,$\boldsymbol{V}$ 的列称为 $\boldsymbol{A}$ 的右奇异值向量;$\boldsymbol{S}$ 为对角矩阵,对角元素称为 $\boldsymbol{A}$ 的奇异值。设 $\boldsymbol{A}$ 的秩为 r,取 $\boldsymbol{U}$ 的前 k 列,$\boldsymbol{V}$ 的前 k 行,以及 $\boldsymbol{S}$ 中的前 k 个奇异值得 $\boldsymbol{A}_k = \boldsymbol{U}_k\boldsymbol{S}_k\boldsymbol{V}_k^T$,这种方法称之为"截断",从而实现特征空间的降维[6]。与其他降维方法相比,$\boldsymbol{A}_k$ 是原矩阵 $\boldsymbol{A}$ 在 k 维子空间上最小二乘意义上的最佳近似[7]。LSA 中降维的过程可以视为去除了语义空间中代表低信息量的自由度,而保留了代表语义空间中主要信息的自由度。

1.2　特征选择算法改进

在构造词 - 文档矩阵时往往要考虑使用某种特征选择算法,实际上就是对 $\boldsymbol{A}_{ij}$ 进行加权,原因是每个词语对文档的语义贡献度不同,对文档语义贡献度高的词语权重应该高于对文档语义贡献度低的词语权重[8]。

传统的特征选择算法为 TF - IDF,其中 TF(Term Frequency)表示一个词语在一个文档中出现的次数,一个词语在文档中出现的次数越多,说明它的重要性越高,出现的次数越少,说明它的重要性越低;DF(Document Frequency)表示包含某一词语的文档数目,包含这个词语的文档数越多,这个词语的重要性越小,相反,如果包含这个词语的文档数很少,那么这个词语的重要性就比较大。IDF 是与 DF 相反的函数。文中对传统的 DF - IDF 特征选择算法进行了改进,增加了文档长度和词语熵权重这两个因素。

TF - IDF 描述为:

$(\mathrm{tf} - \mathrm{idf})_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i =$

$$\frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{d:t_i \in d\}|} \tag{1}$$

其中 $n_{i,j}$ 为词语 i 在文档 j 中出现的频率，$\sum_k n_{k,j}$ 为文档 j 中的所有词语数，$|D|$ 为文档集中的所有文档数，$|\{d:t_i \in d\}|$ 为包含词语 i 的文档数。为了避免分母为零，即词语 i 在任何文档中都没有出现，通常将 $|\{d:t_i \in d\}|+1$ 作为分母。

文档长度即文档中包含的词语数，假设词语 i 在一个长度为 1000 的文档中出现 10 次和在一个长度为 500 的文档中出现了 5 次，则该词语在很大程度上对两个文档的语义贡献度相差不大，故在此引入文档长度这一因素，权重计算公式变为：

$$W_{ij} = \frac{\mathrm{tf}_{i,j} \times \mathrm{idf}_i}{l_j} \tag{2}$$

其中 l_j 为文档 j 的长度。

熵权重是在信息理论的基础上提出来的，它是最复杂的权重计算方法，但非常有效。熵权重计算方法为：

$$E_{i,j} = \log(\mathrm{tf}_{i,j} + 1.0) \times$$
$$(1 + \frac{1}{\log(n)} \sum [\frac{\mathrm{tf}_{i,k}}{\mathrm{gf}_i} \log(\frac{\mathrm{tf}_{i,k}}{\mathrm{gf}_i})]) \tag{3}$$

其中 $\frac{1}{\log(n)} \sum_{k=1}^{n} [\frac{\mathrm{tf}_{i,k}}{\mathrm{gf}_i} \log(\frac{\mathrm{tf}_{i,k}}{\mathrm{gf}_i})]$ 是特征 i 的平均熵。当该特征在所有的文档中是均匀分布时，这个值为 -1；若特征只在一篇文档中出现，则其值为 0。

最后的权重计算公式为两部分因素的结合，公式如下：

$$W_{ij} = \frac{\mathrm{tf}_{i,j} \times \mathrm{idf}_i}{l_j} \times E_{i,j} \tag{4}$$

2 模糊 C 均值

模糊 C 均值算法（FCM）首先由 Dunn[9] 提出，经过 Bezdek 等[10] 完善和推广。在 C 均值算法中，把 N 个样本 $\{x_1, x_2, \cdots, x_N\}$ 划分成 R 个子类 $G_1, G_2, \cdots, G_R$，使得所有样本到聚类中心的距离平方和最小，即使下面的准则函数达到最小。

$$J = \sum_{j=1}^{R} \sum_{x_i \in G_j} \| x_i - m_j \|^2 \tag{5}$$

其中 m_j 为第 j 个子类 G_j 的聚类中心；x_i 表示分到 G_j 的所有样本，$j = 1,2,\cdots,R$。模糊 C 均值算法就是将 C 均值算法中的硬分类变为模糊分类[11]。

设 $\mu_j(x_i)$ 是第 i 个样本 x_i 属于第 j 类 G_j 的隶属度，利用隶属度定义的聚类损失函数为：

$$J_f = \sum_{j=1}^{R} \sum_{i=1}^{N} [\mu_j(x_i)]^b \| x_i - m_j \|^2 \tag{6}$$

其中，$b > 1$ 是一个可以控制聚类结果的模糊程度的常数，称之为模糊加权指数。模糊加权指数 b 在模糊聚类中是一个非常重要的参数，它控制模糊聚类中模糊程度的量级。不同的 b 会对模糊聚类的精度和速度产生不同的影响。但对于 b 的选择没有理论和经验性的公式，McBrateney 等[12] 通过试验发现 b 的取值接近 2 时 FCM 的聚类效果较好。

要求各个样本属于各个聚类的隶属度之和为 1，即：

$$\sum_{j=1}^{R} \mu_j(x_i) = 1 \ (i = 1,2,\cdots,N) \tag{7}$$

FCM 实现的具体步骤如下：

1）用随机数生成法产生隶属度初始矩阵 $\boldsymbol{U}^{(1)}$，$\boldsymbol{U}^{(1)}$ 是 $r \times n$ 维的矩阵，且 $\boldsymbol{U}^{(1)}$ 要满足上述约束条件；设定聚类数目 r 和加权指数 m，迭代截止误差 $\varepsilon > 0$。

2）计算聚类中心 V。

$$v_i = \frac{\sum_{k=1}^{n} (\mu_j(x_k))^m x_k}{\sum_{k=1}^{n} (\mu_j(x_k))^m}, (j = 1,2,\cdots,r) \tag{8}$$

3）更新隶属度值 $\mu_j(x_k)$。

$$\mu_j(x_k = 1 / \sum_{j=1}^{r} (\| x_k - v_i \| / \| x_k - v_j \|)^{2/m-1}$$
$$i = 1,2,\cdots,r; k = 1,2,\cdots,n \tag{9}$$

4）重复步骤 2）和 3），直到 $\| \boldsymbol{U}^{(l)} - \boldsymbol{U}^{(l-1)} \| < \varepsilon$（$\varepsilon$ 为预定的误差值）停止计算。

当该算法收敛时，就得到了各类的聚类中心和各个样本属于各类的隶属度，从而完成了模糊聚类。然后可将模糊聚类结果去模糊化，把模糊聚类变为确定性分类。

3 文中算法

输入值：欲分类文档集，语义空间维数 K，模糊加权指数 b。

输出值：文档聚类结果 $\mathrm{DC} = \{\mathrm{DC}_1, \mathrm{DC}_2, \cdots, \mathrm{DC}_R\}$ 和相应的聚类中心 $\mathrm{Cid} = \{\mathrm{Cid}_1, \mathrm{Cid}_2, \cdots, \mathrm{Cid}_R\}$，$R$ 表示聚类数。

步骤 1：对文档进行分词和停词预处理，构造文档集中每个文档的文档向量；

分词和停词预处理目的是去掉一些冗余的信息，包括去除多余的空格、特殊符号的转换等等。

步骤 2：利用改进的特征选择算法构造词 - 文档矩阵 $\boldsymbol{A}$；

步骤 3：利用 LSA 对矩阵 $\boldsymbol{A}$ 进行奇异值分解，得 K 维语义空间 $\boldsymbol{A}_K$；

步骤 4：将文档向量映射为语义空间 $\boldsymbol{A}_K$ 中的语义向量，并计算文档语义向量之间的相似度。相似度计算

采用两向量之间夹角的余弦值表示:

$$\mathrm{Sim}(D_i \cdot D_j) = \frac{\sum_{m=1}^{K} D_{im} D_{jm}}{\sqrt{\sum_{m=1}^{K} D_{im}^2} \cdot \sqrt{\sum_{m=1}^{K} D_{jm}^2}} \qquad (10)$$

其中$(D_{i1}, D_{i2}, \cdots, D_{iK})$和$(D_{j1}, D_{j2}, \cdots, D_{jK})$分别是文档 i 和 j 在语义空间中的语义向量;

步骤 5:利用上述 FCM 算法迭代求解得各类聚类中心,以及各个样本属于各类的隶属度;

步骤 6:将模糊聚类结果去模糊化,选取样本各类隶属度最大的一个作为该样本的分类结果。

4　实验结果分析

文本实验文档集采用自同济大学论坛下载的 700 多个页面,从中提取出 700 篇文档,按主题分为:高考、选课、考试、考研、留学、就业、兼职七类,每类包含 100 篇文档。

图 1～图 3 为 b 和 K 分别取不同值时的分类结果。试验中采用分类常用的评价标准:准确率(Precision)、查全率(Recall)[13]。

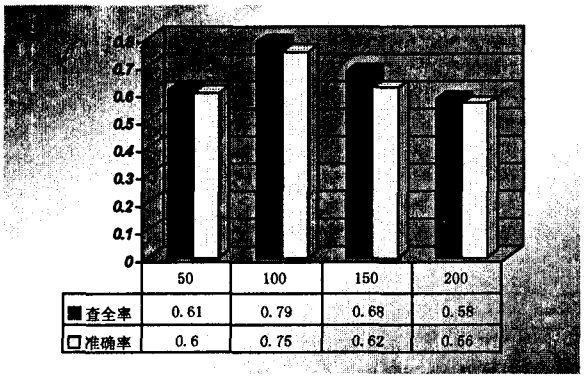

图 1　不同语义空间维数 K 及 b=1.5 下的准确率与查全率

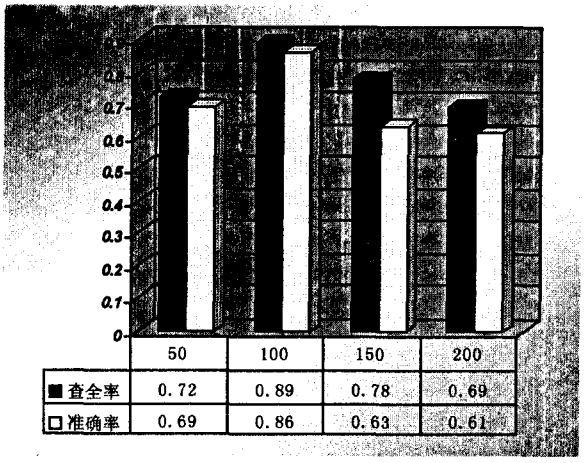

图 2　不同语义空间维数 K 及 b=2 下的准确率与查全率

由图可以看出,语义空间维数 K 的选取对聚类的准确度有一定的影响,K 值选取太小,文档语义空间中保留下来的文档语义太少,导致分辨文档的能力不足;如果 K 值选取太大,则生成的语义空间接近于向量空间模型,失去其降噪的能力,而且计算量比较大。因此 K 值的选取应根据具体实验结果,选取使文档聚类准确度最高的 K 值。同时对模糊加权指数 b 进行了不同的取值,从实验数据可以看出,b 的取值越接近 2 时分类效果越好,证实了 McBrateney 等的实验发现。从实验结果来看,采用改进的 LSA 和模糊 C 均值来进行文档聚类,聚类结果在准确率和查全率这两个指标上取得了比较好的值,聚类结果的准确度比较高。

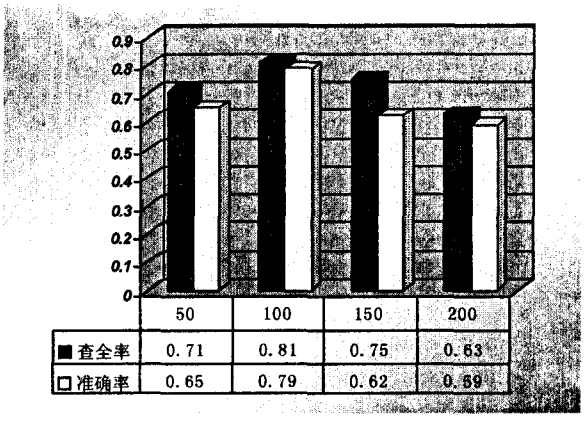

图 3　不同语义空间维数 K 及 b=2.2 下的准确率与查全率

5　结束语

文中提出了一种将模糊 C 均值和改进的 LSA 相结合的方法进行文档聚类。改进的 LSA 方法将高维的文档向量映射为低维的语义向量,提取了文档的潜在语义,降低了处理的复杂度。采用算法简单、收敛速度快且易于计算机实现的模糊 C 均值对文档进行聚类,实验结果证明了该方法的有效性。但 LSA 的基础是基于词频统计的方法,它并没有考虑句子结构、句式、时态,以及词语在句中充当的句子成分等因素,这些因素也对文档语义有着非常重要的作用,因此该方法还有很大的局限性。因此,考虑如何让计算机自动识别文档中句子的结构、句式、时态以及词语在句中充当的句子成分是接下来研究的重点,解决了此难题便可将这些因素的影响考虑进来,从而可以将聚类的准确度进一步提高。

参考文献:

[1]　俞　辉.基于改进 LSA 的文档聚类算法[J].小型微型计算

(下转第 136 页)

参考文献:

[1] Cody R L, Cosmas J, Tsekleves E. Open - standards rich media mobile platform & rapid service creation tool[C]// Global Mobile Congress. [s.l.]: IEEE, 2009: 12 - 14.

[2] 程其江,吕述望,刘越男. WAP 增值业务中终端适配的研究[J]. 计算机应用,2006,26(12):345 - 346.

[3] 纪合宝,扬明极,刘萌萌. WAP 增值业务中图片处理的研究[J]. 哈尔滨理工大学学报,2004,9(5):74 - 75.

[4] 喻 斌. 内容管理系统中模板技术的研究与应用[D]. 南昌:南昌大学,2007.

[5] Mukherjee D, Delfosse E, Kim J G, et al. Optimal adaptation decision - taking for terminal and network quality - of - service[J]. IEEE Transactions on Multimedia, 2005, 7(3): 454 - 462.

[6] 童名文, 杨宗凯, 张景中. 面向服务的内容适配框架研究[J]. 计算机应用研究, 2008(3):749 - 751.

[7] 刘瑞祥,方 济. 自适应移动终端框架的研究与开发[J]. 计算机工程,2009,35(18):266 - 268.

[8] XSL style sheets[EB/OL]. 2009 - 11 - 12. http://www.w3.org/Style/XSL/.

[9] Kurt Cagle. XSL 高级编程[M]. 北京:机械工业出版社, 2002:4 - 100.

[10] 李 江,张 威. 实例解析 XML/XSL/Java 网络编程[M]. 北京:希望电子出版社,2002.

[11] B'Far R. Mobile Computing Principles: Designing and Developing Mobile Applications with UML and XML[M]. Cambrige: Syndicate of the University of Cambrige Press, 2005: 12 - 56.

[12] 陈海山. 深入 Java servlet 网络编程[M]. 北京: 清华大学出版社,2002:56 - 161.

(上接第 129 页)

机系统,2009,30(5):963 - 966.

[2] 王剑锋,乔 冬,麻丽娜,等. 基于潜在语义分析的网页文本分类研究[J]. 应用能源技术,2009(11):41 - 44.

[3] 李 雷,罗红旗,丁亚丽. 一种改进的模糊 C 均值聚类算法[J]. 计算机技术与发展,2009,19(12):71 - 73.

[4] Deerwester S, Dumais S T, Furnas G W, et al. Indexing by latent semantic analysis[J]. J. Amer. Soc. Info. Sci, 1990, 41:391 - 407.

[5] Landauer T K, Foltz P W, Laham D. Introduction to latent semantic analysis[J]. Discourse Processes, 1998, 27(25):259 - 284.

[6] 乌庆敏,杨思春. 基于潜在语义分析的智能答疑系统研究与实现[J]. 计算机技术与发展,2008,18(9):251 - 253.

[7] Golub G H, Van loan C F. Matrix computations[M]. 2nd ed. Baltimore: John - Hopkins, 1986: 56 - 60.

[8] 刘云峰,齐 欢. 潜在语义分析权重计算的改进[J]. 中文信息学报,2005,19(6):64 - 69.

[9] Dunn J C. Well - separated clusters and the optimal fuzzy partition[J]. Journal of Cybernetic, 1974, 4:95 - 104.

[10] Bezdek J C. Pattern recognition with fuzzy objection function algorithms[M]. New York: Plenum Press, 1981.

[11] 吴 瑛. 模糊 C 均值聚类算法在 Web 使用挖掘上的应用研究[J]. 计算机技术与发展,2008,18(6):32 - 35.

[12] McBrantney A B, Moor A W. Application of fuzzy set to climatic classification[J]. Agricultural and Forest Meteorology, 1985, 35:165 - 185.

[13] Sebastiani F. Machine learning in automated text categorization[J]. ACM Computing Surveys, 2002, 34(1):1 - 47.

(上接第 132 页)

Learning for Cascade Face Detection[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 30(3): 369 - 382.

[2] Li Zhifeng, Lin Dahua, Tang Xiaoou. Nonparametric Discriminant Analysis for Face Recognition[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 31(4):755 - 761.

[3] Pentland A, Starner T, Etcoff N, et al. Experiments with eigenfaces[J]. IEEE Trans Pattern Anal Machine Intell, 2004, 26(5):572 - 581.

[4] Shi J Z, Reichenbach S E. Image interpolation by two - dimensional parametric cubic convolution[J]. IEEE Transactions on Image Processing, 2006, 15(7):1857 - 1870.

[5] 孙 亚. 基于粒子群 BP 神经网络人脸识别算法[J]. 计算机仿真,2008,25(8):201 - 204.

[6] 袁 健,姚明海. 基于简化局部二元法的人脸特征提取[J]. 计算机技术与发展,2009,19(6):84 - 90.

[7] 张 熠,张桂林. 基于总变分模型的光照不变人脸识别算法[J]. 中国图象图形学报,2009,14(2):208 - 213.

[8] 姚同庆,房 斌,尚赵伟. 基于 CSVD - NMF 的人脸识别算法[J]. 计算机工程,2009,35(3):214 - 216.

[9] 王李冬. 一种新的人脸识别算法[J]. 计算机技术与发展,2009,19(5):147 - 149.

[10] 贾淑华,李星野,姜共乾. 基于小波分解和分类的人脸识别[J]. 计算机测量与控制,2009(1):167 - 169.

[11] 徐 勇,张 海,周森鑫,等. 基于统计学习理论的人脸识别方法研究[J]. 计算机技术与发展,2007,17(11):118 - 124.

[12] 王 辉. 主成分分析及支持向量机在人脸识别中的应用[J]. 计算机技术与发展,2006,16(8):24 - 26.