

融合相对熵与自适应 LLE 的两阶段文本降维方法

李卫平^{1,2}, 杨 杰¹, 王 钢²

(1 武汉理工大学 信息工程学院, 湖北 武汉 430070; 2 铁道警察学院 公安技术系, 河南 郑州 450053)

摘 要: 多数基于贪婪策略的特征选择往往只能得到次优解. 对此提出了一种两阶段特征降维方法, 首先设计条件乘积相对熵算法以选择文档的特征子集, 然后在文档特征子集中使用提出的自适应 LLE 算法进行特征抽取以进一步降低文档特征维度. 实验结果显示, 两阶段降维方法可显著降低维数并提高文本挖掘性能.

关键词: 文本挖掘; 特征降维; 乘积条件相对熵; 自适应 LLE 算法

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2015)04-0056-05

Two-Stage Integration of Text Relative Entropy and Adaptive LLE Dimensionality Eeduction Method

LI Wei-ping^{1,2}, YANG-jie¹, WANG-gang²

(1 Department of Information Engineering, School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China; 2 Department of Police Technology, Railway Police College Zhengzhou, Zhengzhou 450053, China)

Abstract: The results of most feature selection based on greedy strategies are suboptimal solutions. So, a two-stage dimension reduction method is proposed in this paper. A conditional product Kullback-Leibler divergence algorithm for feature selection is designed to get the feature subset at the first stage and then the adaptive Local Linear Embedding (ALLE) for feature extraction in the feature subset is given out. Experimental results show the novel method can significantly reduce the dimension of features and improve performance of text mining.

Key words: text mining; dimension reduction; product conditional kullback-leibler divergence; adaptive LLE

1 引言

在信息爆炸式增长的大数据时代, 文本挖掘则是一种信息处理的重要手段. 要使计算机高效地处理真实文本, 必须找到一种形式化表示方法, 例如用文档的字、词、短语, 甚至是从文中凝练的概念等特征来表示文档^[1], 形式化表示文档存在维数灾难与特征稀疏问题. 一个文档集的特征维度可能达到数万甚至十万, 并极大地增加分类算法的计算开销; 而向量空间中一个具体文档的大部分特征是零向量, 这使得相似性度量失去意义^[2]. 因此, 文档的特征降维就成为影响文本挖掘效果的关键因素.

现有的文本降维技术可分为两大类: 特征选择

(FS)和特征抽取(FE). 特征选择是从文档的特征集中选择合适的特征构成特征子集. 这一方面比较成功的应用一般基于文档频率(DF)、卡方统计和互信息(MI)方法^[3]. 特征选择虽然具有较高的效率, 但个体最优的特征组成的集合未必是最优特征集^[4]. 特征提取是从文档的原始特征中生成一些新特征, 如概念与核心思想等. 特征提取的主要方法有主成分分析法(PCA)、概率语义分析法(pLSA)、海量特征关联方法^[5]和局部线性嵌入法(LLE)^[6]等. 虽然理论上特征提取能够求得最优解, 但巨大的计算开销影响了其在文本分类中的使用.

TOFA 算法^[7]将特征选择与特征抽取相结合以发挥 FE 求最优解的能力并在一定程度上限制了

收稿日期: 2014-05-30; 修回日期: 2014-07-28

基金项目: 公安部重大项目基金 (201202ZDYJ017)

其计算复杂度.受此启发,相继有学者提出FS与FE结合的方法.然而,这些算法存在实现复杂、计算开销大或者降维效果不佳等局限.

本文提出了一种两阶段文本特征降维方法:在降维过程第一阶段提出一种高效的乘积相对熵度量方法并用其选择文档的特征子集;在第二阶段用改进的LLE算法在特征子集中进行特征提取,从而进一步精简特征集.

2 乘积相对熵与特征选择

相对熵(Kullback-Leibler divergence)是一种能从本质上衡量特征相似度的指标,并且计算相对熵的时间复杂度不高,可作为两阶段降维方法中的特征选择算法基础.

2.1 Kullback-Leibler divergence 与相似度量

相对熵可定义为

$$D_{LK}(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (1)$$

相对熵采用了 $0 \log(0/k) = 0$ 以及 $k \log(k/0) = \infty$ 的规约.显然,相对熵具有非负性,它的值越大两个随机变量分布的差别就越大.然而,相对熵还具有显著的非对称性:

$$D_{LK}(p \parallel q) \neq D_{LK}(q \parallel p) \quad (2)$$

2.2 乘积条件相对熵

相对熵非对称性引起了用 $D_{LK}(p \parallel q)$ 还是 $D_{LK}(q \parallel p)$ 评价差异度的问题.有学者提出用均值来度量相似度,然而这种方法是不精确的,可能会影响到信息运算的效果^[8].最大相对熵 $\max [D_{LK}(p \parallel q), D_{LK}(q \parallel p)]$ 有比均值相对熵更高的精确性,但其忽略了相对熵中的较小者,对度量信息的利用并不充分.在此可以提出一种乘积相对熵的判别标准:

$$\bar{D} = D_{LK}(p \parallel q) \cdot D_{LK}(q \parallel p) \quad (3)$$

用上式作为衡量特征相似度的最终差异能以很少的计算开销代价来提高衡量标准的精确性并充分利用已知信息.

相对熵只是简单刻画了两个分布之间的距离,没有考虑两个分布的差异性随条件的变化.在文本特征选择中,衡量两个特征子集在包含或排除某些特征项的条件下的差异性是非常有意义的工作.因此,在特征选择中可以使用如下条件相对熵定义:

$$D_C(p(y|x) \parallel q(y|x)) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \quad (4)$$

因此,可得出如下乘积条件相对熵 $\bar{D}_C$ 的定义:

$$\bar{D}_C = D_C(p(y|x) \parallel q(y|x)) \times D_C(q(y|x) \parallel p(y|x)) \quad (5)$$

在文本分类中,使用乘积条件相对熵作为差异性度量标准以进行特征选择可以在限制计算复杂度的同时综合各种因素充分考虑特征之间的相似度.

2.3 PCRE 特征选择算法

本文提出一种基于乘积相对熵的特征选择算法,称为 Product Conditional Relative Entropy-based (PCRE) Feature Selection 算法.与 MI 和 IG 等类似,PCRE 算法的核心思想是筛选相对熵更大的特征构成特征子集.具体实现上,首先排除停用词,然后计算特征的初步相对熵.相对熵明显低于或者高于阈值的特征将被直接排除.位于临界值附近的特征将被计算条件相对熵.其存在性对条件相对熵没有显著影响的特征将被排除,剩余特征构成最终的特征子集.在算法中两种阈值被设定,分别是用来排除低区分度以及干扰词的相对熵阈值 $T_{\min}$ 和 $T_{\max}$ 以及判断是否选择相对熵在临界域内的条件相对熵阈值 T_c . 算法的流程如图 1 所示.

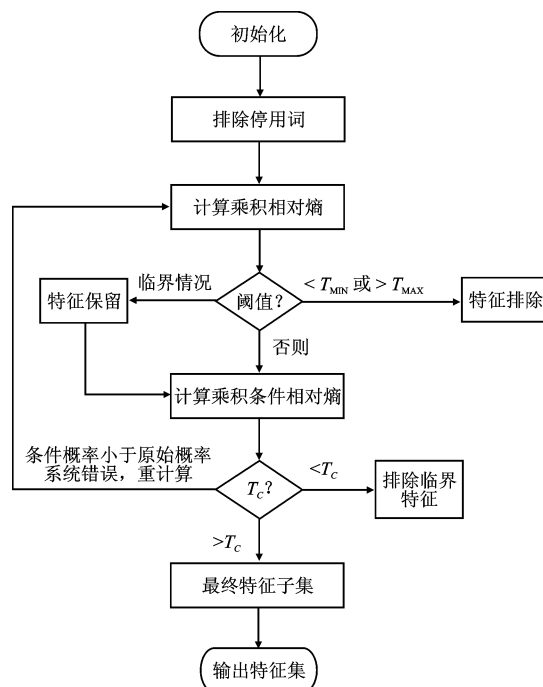


图1 基于乘积条件相对熵的特征选择算法流程

按图 1 所示的流程,系统可以用较小的计算复杂度和时间开销选择最合适的特征构成特征子集.

3 基于改进 LLE 的特征提取

属性子集选择通常使用压缩搜索空间的启发式

算法. 通常这些算法都是贪婪算法, 在搜索特征空间时, 总是做在当时看上去最佳的选择. 所以特征选择方法无法保证解的最优性.

特征提取算法中主成分分析法使用最为广泛. 当样本点具有一些非线性性质时, 采用 PCA 得到的降维结果无法反映出样本点之间所隐藏的非线性性质. 而非线性是大量文本对象的一个主要特征. 尽管有针对非线性问题的 Kernel PCA 算法被提出, 但是这类算法在处理文本分类问题时计算代价过大. 此外, 对 PCA 所要保持的主分量的个数估计比较困难. 另一类热点概率主题模型如 pLSA、LDA 等算法复杂度高, 时间开销大, 如果不经过并行化, 无法适用于轻量级的在线文本实时处理.

3.1 LLE 特征抽取方法

LLE 算法自 2000 年被 Roweis 等学者提出以来^[9], 十余年来在模式识别与机器学习等 AI 相关领域得到了越来越多的关注. 作为降维领域一种较新的高效算法, LLE 的基本思路: 整体来看高维数据应该呈非线性分布; 从给定点与其近邻点之间的局部考虑高维数据应呈线性分布. 也就是每一个点均可由其近邻点线性组合而成, 在高维空间中数据的整体分布可以看作一系列的线性超平面小块连接而成. 对于给定输入特征集 $f_i \in R^D$ 和输出特征集 $f_i' \in R^d$, 标准的 LLE 算法步骤如下:

步骤 1: 在欧氏空间中寻找 f_i 的若干邻近特征.

步骤 2: 通过最小化重建误差计算重建权重. 若 f_i 和 f_j 为邻近点, 则其误差计算为

$$\epsilon(W) = \sum_{i=1}^N \left\| f_i - \sum_{j=1}^N W_{ij} f_j \right\|^2 \quad (6)$$

步骤 3: 通过下式搜索低维嵌入

$$\varphi(F) = \sum_{i=1}^N \left\| f_i' - \sum_{j=1}^N W_{ij} f_j' \right\|^2 \quad (7)$$

上式有两个约束条件:

$$\begin{cases} \sum_{i=1}^N f_i = 0 \\ 1/N \sum_{i=1}^N f_i^T f_i = I \end{cases} \quad (8)$$

式中, I 是 $d \times d$ 阶识别矩阵.

步骤 4: 为了在以上约束条件下寻找矩阵 F , 需要基于下式新建一个矩阵 W :

$$W; M = (I - W)^T (I - W) \quad (9)$$

通过与矩阵 M 中 d 个最小的非零特征值相关的特征向量 d 就可产生最终的嵌入矩阵 F . 其直观的示意图如图 2 所示.

通过以上步骤可以看出, LLE 算法待定的参数

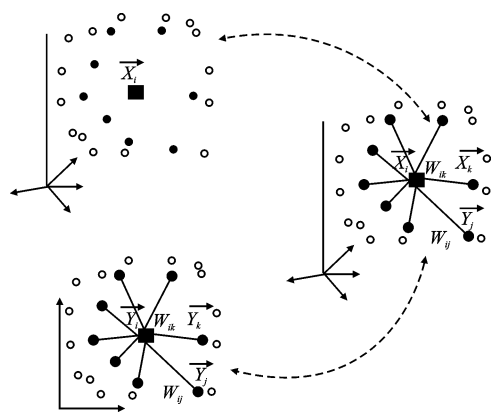


图 2 LLE 特征抽取原理示意图

很少, 它存在整体的最优解, 没有局部最小问题, 不需要迭代. 而且该算法归结为稀疏矩阵特征值计算, 计算复杂度相对较小, 容易执行. 然而该算法也具有有一些显著的缺陷. 首先, 在特征空间稀疏到一定程度时, 对特征相似度的识别能力骤减. 其次, 若 K 取值过大则不能体现局部线性特性. 此外, 若 K 取值过小则无法体现邻近节点之间的拓扑关系. 为了克服如上问题, 一些 LLE 算法的变种相继被提出^[10]等. 这些算法在一定程度上解决了原始 LLE 算法的局限性, 但降维效果仍有提升空间. 此外, 有些算法变种还引入了新的问题, 比如为了增加精度而提高了计算开销或者为了提高速度而影响了算法的鲁棒性.

3.2 自适应 LLE 特征提取算法

由于欧氏距离不能很好的区分非线性的文档特征, 为了解决以往 LLE 算法族的稀疏特征相似性度量问题, 本文在经典 LLE 算法中引入了马氏距离 (Mahalanobis distance) 作为特征距离的度量. 对于一个均值为 $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$, 协方差矩阵为 Σ 的多变量特征矢量 $x = (x_1, x_2, \dots, x_n)^T$, 其马氏距离为:

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} \quad (10)$$

与欧氏距离不同的是它能体现各种特性之间的联系, 而且注意到了有些特征是有内部关联并且是尺度无关的, 即独立于测量尺度的. 马氏距离的这些度量特性非常契合文档数据自身的特点以及文本特征选取的要求.

参数选择的经验依赖性制约 LLE 算法族准确率的另一个重要问题. 由于采用了两阶段特征降维策略, 本文的特征提取工作在已被选择的特征子集中进行. 在这种前提下, 不依赖预设的 LLE 的参

数 K , 确定应该通过简单的自适应算法实现, 以降低算法的计算复杂度. 邻居特征选择的目的是在保证误差足够小的前提下尽可能限制特征数目. 首先设置一个时间间隔 S_T , 通过式(5), 可以设置两个误差上下限 ϵ_{MAX} 和 ϵ_{MIN} , 每过一个时间间隔, 算法自动检测误差与上下界的关系. 高于上界说明参考特征太少以至于影响算法的准确性, 低于下界则可减少邻居特征以提高效率. 同理, 位于两个阈值之间则保持 K 值不变.

结合传统 LLE 算法的原理和流程, 引入以上两点重要改进可构成自适应局部线性嵌入 (Adaptive Local Linear Embedding, ALLE) 算法. 算法流程和细节如图 3 所示.

```

Algorithm: 自适应局部线性嵌入算法

Input: 文档集  $S=\{s_1, s_2, \dots, s_m\}$ 
      误差阈值  $\epsilon_{\text{MAX}}, \epsilon_{\text{MIN}}$ 
      时间间隔  $S_T$ 
Output: 降维之后的特征集合  $F=\{f_1, f_2, \dots, f_n\}$ 

1 begin
2   随机选择邻居特征点的数目  $K(0)$ 
3   计算误差  $\epsilon$ 
4   if ( $\epsilon < \epsilon_{\text{MIN}}$ )
5      $K(t+1)=K(t)-1$ 
6   else if ( $\epsilon_{\text{MIN}} \leq \epsilon \leq \epsilon_{\text{MAX}}$ )
7      $K(t+1)=K(t)$ 
8   else
9      $K(t+1)=K(t)+1$ 
10  局部特征线性转换
11  if ( $t \geq S_T$ )
12     $t=0$ 
13    go to 4
14  else
15    continue
16  求解重新建立的矩阵  $W$ , 输出降维结果
17 end
  
```

图3 自适应局部线性嵌入算法伪代码

传统 LLE 算法的收敛性已经由 Roweis 等学者^[9]证明. 由于 ALLE 算法没有分支缺失情况, 所以算法是自闭和的, 且误差等各种参数均是有界的, 因此 ALLE 算法也是收敛的.

4 两阶段文本特征降维方法

通过前两节的工作, 新的文本特征选择与特征提取算法均已实现. 将这两种算法有机地结合起来, 可以构成两阶段文本特征降维方法. 算法流程如下所示.

(1) 预处理. 文档集分割、分词、生成停用词表和去除停用词等.

(2) 计算文档特征的乘积相对熵, 初步判断文档集中不同文档的特征分布和各特征之间的信息差异度, 在此基础上进行特征初筛.

(3) 通过计算条件相对熵最终确定特征项的信息距离, 根据特征的最终差异度进行取舍, 输出最终的特征子集.

(4) 在步骤(3)所输出的特征子集中计算各特征项的马氏距离.

(5) 随机初始化邻居特征数目 K .

(6) 通过邻近特征来重建特征, 用马氏距离衡量重建特征的误差值.

(7) 通过定期更新重建特征误差自适应调节 K 值.

(8) 特征矩阵求解, 产生相对于文档原始特征数目大大减少的最终文本特征项.

(9) 用新的文本特征项表示文档集中的所有文档, 完成文档特征降维.

通过以上步骤, 基于乘积条件相对熵的特征选择算法和自适应 LLE 特征提取方法有机地结合在一起, 构成了两阶段文档特征降维模型. 该模型兼有特征选择的高效率和特征抽取的高效果, 并通过引入新的距离度量增加了算法的区分能力和鲁棒性.

5 实验及结果分析

为了充分考察和对比两阶段文本特征降维方法的效率和效果, 在多个语料库中用大量中、英文文档测试了算法的时间开销, 参照以往研究者的实验结果比较了几个经典分类算法在不同特征降维技术基础上的分类准确率. 实验所涉及到的语料库为 Reuters-21578 和宾州树库的英文部分以及北京大学现代汉语语料库. 为了增强实验结果的可比性, 文档均被分割为 10 kB 大小的 .txt 文档.

实验结果表明:

降维效率是衡量降维算法的重要指标. 本文从四大语料库中分别选取 100、300、1 000、3 000 和 10 000 个文档测试其处理速度. 并与几种经典的文档特征降维算法做对比, 包括词频-文档 (TF) 法、MI、PCA、标准 LLE 和 RKLLE. 为了便于对比, 绘图中 X 轴采用了对数坐标, 实验结果如图 4 所示.

通过图 4 可知, 大体上特征选择算法的时间开销要小于特征提取算法. 特别是基于 TF 的算法, 由于实现简单, 所以效率明显高于其他算法. 本文所提出的两阶段文本特征降维方法在处理速度上优于其他特征提取方法, 与基于互信息的特征选

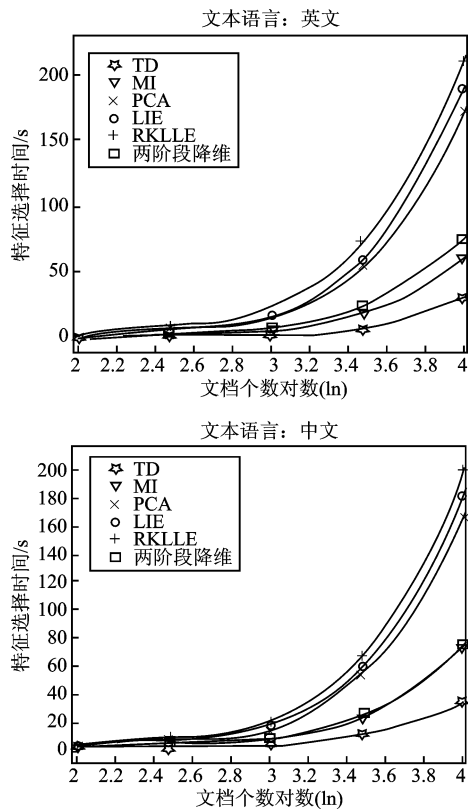


图 4 各种降维算法运行时间对比

择算法大致相当. 可见, 先用特征选择算法对特征提取对象进行预处理, 然后用自适应 LLE 在特征子集上进行提取的策略, 可以显著减小提取的时间开销, 使 ALLE 比同类算法高效. 此外, 乘积条件相对熵在实现上有比互信息更小的计算开销, 所以 ALLE 的效率与基于 MI 降维方法相比也是具有竞争性的.

准确率是文本分类最重要的指标. 我们从前述的四大语料库中个选择了 20 000 个文本, 其中 10 000 个作为训练文本, 其余作为测试文本, 测试了几种文本分类经典算法: Naïve Bayes、Rocchio、SVM、AdaBoost 和 kNN 在两阶段降维方法提供的精简特征集上进行分类的准确率. 通过和基于其他特征降维方法的分类准确率^[11-13]相对比, 可得出表 1.

表 1 基于各种降维算法的分类准确率 %

	NB	SVM	Rocchio	AdBst	kNN
DF	70.3	78.8	75.4	78.1	79.2
MI	74.4	81.5	79.9	80.2	82.3
PCA	78.7	86.9	81.5	86.6	86.8
LLE	78.3	86.2	82.3	85.3	87.7
两阶段法	81.1	89.0	85.5	88.4	91.5

由表 1 可见, 基于两阶段文本特征降维方法的分类精确率要明显高于基于特征选择的分类精确率. 此外, 基于本文方法的分类准确率也高于传统特征提取方法的分类结果. 综合考虑效率因素, 可以说两阶段文本特征降维方法显著提升了降维性能, 是文本挖掘领域一个非常有用的工具.

6 结束语

本文提出了一种两阶段文本特征降维方法. 在第一阶段用乘积条件相对熵算法从原始特征中选择特征子集, 第二阶段用一种改进的自适应局部线性嵌入算法在第一阶段提供的特征子集里进行特征提取. 实验结果显示本文提出的方法具有较高的降维效率和很好的准确性.

本文所提出的两阶段降维策略在具体实现上选择了相对熵与 LLE. 实际上 FE 算法族中还有一些优秀的算法如 MMC 和 OCA, 它们与相对熵结合的两阶段特征提取的效果还有待研究, 这可以作为我们今后的研究课题.

参考文献:

- [1] Ronen Feldman, James Sanger. The text mining handbook—advanced approaches in analyzing unstructured data [M]. Beijing: Post & Telecom Press, 2009: 4-7.
- [2] Cheng Qiang, Zhou Hongbo, Chengjie. Selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 8 (2): 1217-1233.
- [3] Sombut Foithong, Ouen Pinngern, Boonwat Attachoo. Feature subset selection wrapper based on mutual information and rough sets [J]. Expert Systems with Applications, 2012, 39(2): 574-584.
- [4] 王博. 文本分类中特征选择技术的研究[D]. 长沙: 国防科学技术大学, 2009.
- [5] 彭其华. 关联挖掘下的海量文本信息深入挖掘实现[J]. 微电子学与计算机, 2013, 30(10): 157-164.
- [6] 文贵华, 江丽君, 文军. 邻域参数动态变化的局部线性嵌入[J]. 软件学报, 2008, 19(7): 1666-1673.
- [7] Yan Jun, Liu Ning, Yan Shuicheng, et al. Trace-oriented feature analysis for large-scale text data dimension reduction [J]. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(7): 1103-1117.
- [8] 张丽青, 寿永熙, 马志强. 最大熵算法在汉语拼音标注中的研究与实现[J]. 微电子学与计算机, 2012, 29 (8): 120-126.

(下转第 64 页)

照正常的顺序对 memory 进行读写控制,然后经过算法加密乱序处理之后,addr_r,w_data_r,r_data_r 为实际读写 memory 的地址和数据信号.

wr_r 为读写控制信号,当 wr_r 为 1 时,表示当前为写操作,否则为读操作.

仿真过程中,首先写入数据到 memory 中,然后依次将数据读出,从仿真图中可以看出,其数据结果与理论分析的数据完全相同,从而证明了此加密方法的可行性.

此外,从图 3 中看书,在给出读信号的同时,r_data_nor_r 有一个无效的返回值,这是因为此时地址还没被更新为有效地址,因此读出一个无效的值,但是由于读周期需要一个时钟周期,所以在给出读数据地址一个周期后才能得到正确的值,正如波形图中所示,给出地址 0x00000 一个周期后,得到正确的数值 0x0420. 所以该无效值不影响数据的正确读写.

4 结束语

本文利用 memory 自身的特性,通过对地址作用一定的加密算法,实现了对数据的加密,提高了数据存储的安全性.

参考文献:

- [1] 李聪. 数据加密技术在网络安全中的重要性[J]. 企业导报, 2011, 14(10): 296.

- [2] Gao Lan, Yang Jun, Chrobak Marek, et al. A low-cost memory remapping scheme for address bus protection[C] // Proceedings of the 15th international conference on Parallel architectures and compilation techniques. USA:Seattle, Washington, 2006: 74-83.
- [3] Xu Huachun, Li Zhengsheng, Wang Zenghong. Study on hardware protection technique of flash disk[C] // The 8th International Conference on Electronic Measurement and Instruments. Xi'an, 2007: 55-59.
- [4] 吴延军. 基于 FLASH 芯片的加密存储技术研究[D]. 广州: 暨南大学, 2006.
- [5] 苏长龄, 庄晋林. 基于序列变换的加密方法[J]. 微电子学与计算机, 1998, 15 (5): 46-47.
- [6] Feuser Markus, Sommer Sabine. Address encryption method for flash memories; USA, US7640437B2[P]. 2009-12-29.
- [7] Mei Hong, Hui Guo, Parameswaran S. Dynamic encryption key design and management for memory data encryption in embedded systems[C] // IEEE Computer Society Annual Symposium on VLSI (ISVLSI). Brazil, Natal, 2013: 70-75.

作者简介:

张金旻 男,(1990-),硕士研究生. 研究方向为专用集成电路与系统设计. E-mail:gzyd2008@yeah.net.

谢小东 男,(1977-),博士,副教授,硕士生导师. 研究方向为超大规模集成电路设计、专用集成电路设计、微系统与 SOC 设计等.

(上接第 60 页)

- [9] Li Bo, Zheng Chunhou, Huang Deshuang. Locally linear discriminant embedding: an efficient method for face recognition [J]. Pattern Recognition, 2008, 41 (12): 3813-3821.
- [10] Sam T Roweis, Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding [J]. Science, 2000, 290(5500): 2323-2326.
- [11] Yamada K, Komine H, Kinukawa H, et al. Abstract of abstract: a new summarizing method based on document frequency and clause length [C] // 8th World Multi-Conference on Systemics, Cybernetics and Informatics. Finland Orlando, 2004: 56-61.
- [12] Chen Chih Ming, Lee Hahn Ming, Chang Yu jung. Two novel feature selection approaches for web page

classification [J]. Expert systems with Applications, 2009, 36 (4): 260-272.

- [13] Ferjani Fethi, Elloumi Samir, Jaoua Ali, et al. Formal context coverage based on isolated labels: an efficient solution for text feature extraction [J]. Information Sciences, 2012, 188(4): 198-214.

作者简介:

李卫平 男,(1974-),博士研究生,副教授. 研究方向为计算机软件与理论、数据挖掘. E-mail:liwp360@163.com.

杨杰 女,(1960-),博士生导师,教授. 研究方向为信息传输理论及应用、模式识别.

王钢 男,(1958-),博士,教授. 研究方向为云计算、计算机犯罪侦查.