第 38 卷第 2 期　　声　学　学　报　　Vol. 38, No. 2
2013 年 3 月　　ACTA　ACUSTICA　　Mar., 2013

音子配列学语种识别系统中特征选择方法的研究*

梁春燕　杨　琳　汪俊杰　张建平　颜永红
(中国科学院声学研究所中国科学院语言声学与内容理解重点实验室　北京　100190)

2011 年 12 月 14 日收到
2012 年 6 月 13 日定稿

摘要　将信息增益和加权 log 似然比特征选择方法应用于音子配列学语种识别系统中进行特征降维。在美国国家标准技术研究院 2009 年语种识别评测数据集上进行实验，分别使用信息增益和加权 log 似然比准则以及传统的互信息，χ^2 统计量方法对数量巨大的 N-gram 进行特征选择，从中选出最具有鉴别性的部分组成特征向量，并用分类器进行分类。结果显示，当根据信息增益和加权 log 似然比准则选取一定数量的特征时，系统性能与使用全部特征的基线系统相比略好；当选取的特征数量很少时，信息增益和加权 log 似然比方法的性能要优于传统的互信息和 χ^2 统计量方法。实验表明，在音子配列学语种识别系统中，信息增益和加权 log 似然比方法均可以有效地去除冗余信息，降低特征向量的维数，并且能使系统性能得到一定的提高。

PACS 数：43.72

Feature selection in phonotactic language recognition system

LIANG Chunyan　YANG Lin　WANG Junjie　ZHANG Jianping　YAN Yonghong
(*Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences*　Beijing　100190)

Received Dec. 14, 2011
Revised Jun. 13, 2012

Abstract　Two feature selection methods of Information Gain (IG) and Weighted Log Likelihood Ratio (WLLR) are introduced into phonotactic language recognition to reduce the dimensions of feature vectors. Together with the traditional Mutual Information (MI) and χ^2-test (CHI), the proposed methods are compared on the NIST 2009 Language Recognition Evaluation (LRE) task. Different subsets of features are selected from the total N-gram, respectively according to the four criteria, as the input feature vectors of the classifier for language recognition. The experimental results show that IG and WLLR can obtain much lower dimensional feature vectors without affecting the language recognition performance even giving better performance than the system with all features. And when the number of selected features is very small, IG and WLLR achieve better performance than the existed MI and CHI criteria. The results indicate that IG and WLLR can effectively reduce the number of features and improve the system to some extent.

引言

自动语种识别，作为智能语音处理的一个重要方向，是指通过对语音信号的分析自动判断其语种类别的技术 [1]。语种识别的研究无论在民用领域还是国防安全领域，都受到了广泛的关注。

根据使用的特征不同，可以将当前主流的语种识别系统分为两类：基于声学频谱特征的语种识别系统 [2] 和基于音子配列学特征的语种识别系统。在音子配列学语种识别方法中，首先将一段语音经过一个或多个音子识别器解码为音子序列或者音子网格，然后利用不同语种间的音子搭配存在差异这一特点，利用解码后的音子序列或者音子网格训练各

* 国家自然科学基金 (10925419, 90920302, 61072124, 11074275, 11161140319, 91120001)、中国科学院战略性先导科技专项 (XDA06030100, XDA06030500)、国家 863 计划 (2012AA012503) 和中国科学院重点部署 (KGZD-EW-103-2) 资助项目。

个语种的模型[3]。基于音子配列学特征的语种识别方法以其性能优异稳定、推广性好等优点受到越来越多研究人员的重视。

基于音子配列学的系统主要采用的建模方式是 N-gram 语法模型建模和基于向量空间模型 (Vector Space Model, VSM) 的鉴别性建模方式[4−5]。在基于向量空间模型的建模方法中，首先提取具有鉴别性的 N-gram 音子串 “bag of N-gram” 作为特征项，然后将其作为分类器的输入特征向量进行分类[6−7]。在音子配列学语种识别系统中，一般来说，高阶的 N-gram 特征包含更丰富的语种信息。但随着 N 的增大，特征的维数呈指数增长，一方面会带来计算和存储的问题，另一方面，特征的维数和分类性能之间并不是成线性关系，当维数超过一定限度时，可能会导致分类性能下降，因此特征选择非常必要[9,13]。

特征选择可以认为是一个优化问题，关键步骤就是建立某种评估准则，以区分哪些特征组合对分类器起积极作用，哪些特征组合是冗余的，不同的评估准则往往会得到不同的结果。在音子配列学语种识别中，目前已使用的特征选择方法包括基于 SVM 分类间隔[8]、互信息[9] 和 χ^2 统计量[9] 等方法。在这些准则的基础上，一种递归的特征选择方法也广泛应用于语种识别中[8]。递归选择法的思想假设，如果高阶的 N-gram 特征存在相当的区分性，那么其对应的低阶 N-gram 特征也应该具有较好的区分性。因此递归选择法先从低阶 N-gram 中选择最具有区分性的关键特征，然后将其向高阶进行扩展，从而得到具有区分性的高阶特征。

本文将文本分类中所使用的信息增益和加权 log 似然比准则[9−10] 引入到音子配列学语种识别中进行特征选择，并与传统的互信息和 χ^2 统计量方法进行性能比较。

1　音子配列学语种识别系统框架

本文使用的语种识别系统是基于音子网格和向量空间模型建模方式，即 PR-Lattice-VSM 系统 (Phone Recognition-Lattice-VSM)，主要包括基于音子网格的特征提取和向量空间模型后端分类两部分。

1.1　基于音子网格的特征提取

Phone Recognition-Lattice-VSM 系统首先采用音子识别器对训练和测试数据进行解码。系统在解码过程中保留 HTK[12] 格式的音子网格结果，然后从音子网格结果中提取每个 N-gram 音子串的期望计数 (Expected Count)[8]，

$$\begin{aligned}\mathrm{count}(d_i|L) = E_W[\mathrm{count}(d_i|W)] = \\ \sum_{W\in L} p(W|L)\mathrm{count}(d_i|W),\end{aligned} \tag{1}$$

其中 L 代表解码得到的音子网格结果，它可以视为音子序列假设的集合；W 是网格中的音子序列候选路径；d_i 代表一个 N-gram 音子串。

预先根据训练数据选择一批具有鉴别性的 N-gram(bag of N-gram) 作为特征项，并对特征项中的每个 N-gram 对应的期望计数进行归一化处理，

$$p(d_i|L) = \frac{\mathrm{count}(d_i|L)}{\sum_j \mathrm{count}(d_j|L)}, \tag{2}$$

其中 $p(d_i|L)$ 即为 N-gram 音子串 d_i 对应的特征项数值。然后对特征项数值进行如下加权，

$$\phi(L)_i = D_i p(d_i|L), \tag{3}$$

取 D_i 为：

$$D_i = \min\left(C_i, g_i\left(\frac{1}{p(d_i|\mathrm{ALL})}\right)\right), \tag{4}$$

其中 $p(d_i|\mathrm{ALL})$ 表示在所有语种数据基础上计算得到的特征项归一化频率；$g_i(\cdot)$ 可以看作控制特征项数值动态范围的函数，通常取 $g_i(x) = \sqrt{x}$ 或者 $g_i(x) = \log(x)+1$; C_i 是一个经验常数。在本文中，我们设置 $g_i(x) = \sqrt{x}$, $C_i = \infty$。

这样，即可以将一段语音解码后的网格结果映射成输入分类器的特征向量，

$$\boldsymbol{\phi}(L) = [\phi(L)_1 \quad \phi(L)_2 \quad \cdots \quad \phi(L)_M], \tag{5}$$

其中 M 为选取的 bag of N-gram 的数目。

图 1 描绘了鉴别性特征向量的提取流程。

1.2　向量空间模型后端分类

鉴别性特征向量确定之后，需要进行向量空间模型的分类，可以采用多种分类策略，例如：人工神经网络 (Artificial Neural Network, ANN), 支持向量机 (Support Vector Machine, SVM), 逻辑回归等等。本文采用的建模方式是 2 级层次化结构，其中 1 级采用 SVM 分类，2 级采用 SVM 或者 LDA-Gaussian (Linear Discriminant Analysis-Gaussian) 方法。训练数据分为 1 级训练数据和 2 级训练数据两部分。

对于两个音子网格 L_1 和 L_2, 其核函数可以表示成如下的内积形式，

$$K(L_1, L_2) = \phi(L_1)^t \phi(L_2). \tag{6}$$

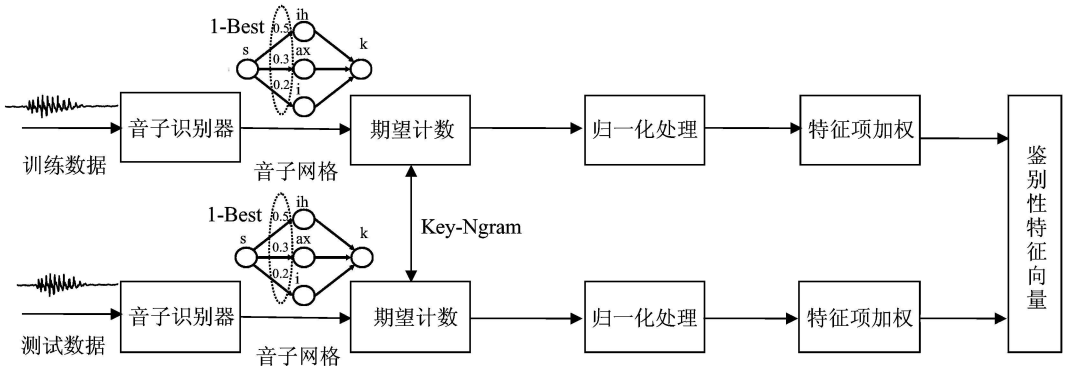

图 1 鉴别性特征向量提取流程

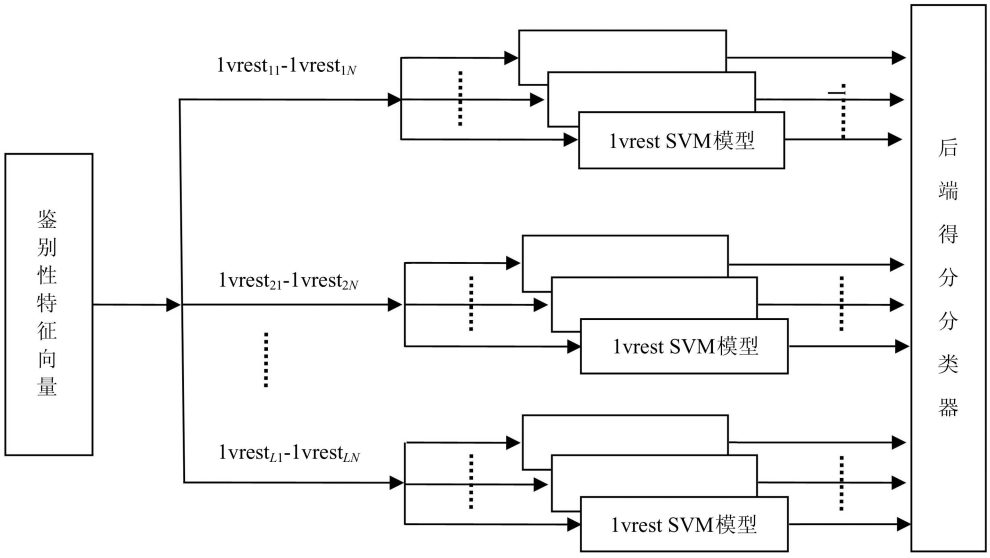

图 2 1 级 SVM 分类策略设计

SVM 本质上是两类分类器，而语种识别多数情况下是多类分类问题，因此采用 SVM 进行多类分类，需要对分类策略进行设计。常用的 SVM 分类方法有 1 对 1(1v1) 和 1 对多 (1vrest), 本文采用的 SVM 多类分类方法为 1vrest 方式，即把目标语种的训练数据看作正样本，所有其他语种的训练数据作为负样本。我们引入 anchor model 的思想 [18], 为每个语种训练多个 1vrest 模型，多个模型共同组成 anchor model 组。

首先使用 1 级训练数据来训练 1vrest 模型。我们把每个语种的 1 级训练数据分为 N 份。对于每个语种，我们将所有其他语种数据合并成 N 份 “rest” 训练数据, 因此也就可以训练 N 个 1vrest SVM 模型。这样假如共有 L 个语种，那么共有 $L*N$ 个 1vrest SVM 模型组成 anchor model 组。 2 级 SVM 训练数据经过 anchor model 组映射为得分向量，然后在得分向量上再进行分类建模。

图 2 描绘了整个 1 级 SVM 分类策略的设计。

后端得分分类器的设计，即 2 级分类建模，我们使用基于径向基核数 (Radial Basis Function, RBF)[17] 的 SVM 分类或者 LDA-Gaussian 方法。在 LDA-Gaussian 方法实现中，首先应用线性鉴别性分析 (LDA) 对输入得分进行降维，然后在降维后的得分上为每个语种建立 GMM 模型。为了使模型更加鲁棒，不同语种的 GMM 模型之间共享相同的全协方差矩阵。

我们在 1 级训练数据选择过程中，各个语种使用尽可能多的训练样本, 选择 2 级训练数据时, 要保证各个目标语种之间训练样本数量的均衡，以防止模型训练的过匹配。

2 特征选择方法

本节简要介绍各种不同的特征选择方法, 包括传统的互信息和 χ^2 统计量准则，以及本文提到的信息增益和加权 log 似然比方法。在特征选择过程中，根据不同的选择标准，每个特征项对应一个信息值，将所有特征项的信息值进行排序，选取其中值较高的特征项。

为了方便对特征选择方法进行描述，我们将训练数据的情况用如下双向列联表表示：

表 1　特征 t 与语种 l 之间的列联表

	l	非 l
t 存在	A_l	B_l
t 不存在	C_l	D_l

A_l: 属于语种 l 且包含特征 t 的训练数据个数；
B_l: 不属于语种 l 且包含特征 t 的训练数据个数；
C_l: 属于语种 l 且不包含特征 t 的训练数据个数；
D_l: 不属于语种 l 且不包含特征 t 的训练数据个数；
N_l: 属于语种 l 的训练数据个数，即 $A_l + C_l$；
N: 训练数据总数，即 $A_l + B_l + C_l + D_l$.

2.1 互信息

互信息 (mutual information, MI)[9,14] 是信息论中的概念，它用于度量一个消息中两个信号之间的相互依赖程度。在特征选择领域中经常用来计算特征和类别之间的依赖程度。特征 t 与语种类别 l 之间的互信息 $MI(t,l)$ 定义为：

$$\mathrm{MI}(t,l) = \log \frac{P(t|l)}{P(t)} \tag{7}$$

将上式中的概率用其相应的频率来代替，$\mathrm{MI}(t,l)$ 可以用下式来近似计算，

$$\mathrm{MI}(t,l) = \log \frac{A_l \times N_l}{(A_l + C_l) \times (A_l + B_l)}, \tag{8}$$

从式 (7) 和式 (8) 中可以看出，如果特征 t 和语种 l 之间是独立的，那么 $\mathrm{MI}(t,l)$ 等于 0; 互信息 $\mathrm{MI}(t,l)$ 越大，说明特征 t 中包含的与语种 l 有关的鉴别信息越多; 而且, 互信息方法会偏向选择出现频率较低的特征。

如果考虑特征 t 与所有语种类别之间的互信息，为了得到特征 t 的全局互信息值，常用的两种方法是：一种是取特征 t 与各个语种类别之间互信息的最大值，一种是特征 t 与所有语种类别之间互信息的平均值，即

$$\begin{cases} \mathrm{MI}_{\max}(t) = \max_{l=1}^{L}\{\mathrm{MI}(t,l)\}, \\ \mathrm{MI}_{\mathrm{avg}}(t) = \sum_{l=1}^{L} P(l) \cdot \mathrm{MI}(t,l), \end{cases} \tag{9}$$

文本分类中的实验表明选取最大值性能要优于平均值 [8], 所以在本文中我们选取最大值作为特征 t 最终的互信息值。在后面的 χ^2 统计量和加权 log 似然比中，我们也做同样的选择。

2.2 χ^2 统计量

χ^2 统计量 (The Chi-squared Statistic, CHI)[9] 又被称作为开方拟合检验，是数理统计中一种常用的检验两个变量独立性的方法，也是文本分类中常用的特征选择方法之一，用来衡量特征和类别之间的统计相关性。

特征 t 与语种类别 l 之间的 χ^2 统计量计算公式如下：

$$\chi^2(t,l) = \frac{N(A_l D_l - C_l B_l)^2}{(A_l + C_l)(B_l + D_l)(A_l + B_l)(C_l + D_l)}. \tag{10}$$

当特征 t 与语种类别 l 相互独立时，$\chi^2(t,l) = 0$, 此时特征 t 不包含任何与语种 l 有关的鉴别性信息。$\chi^2(t,l)$ 的值越大，则特征 t 包含的与语种 l 相关的鉴别性信息就越多。

2.3 信息增益

信息增益 (Information Gain, IG)[9,15] 是机器学习中的概念, 它是针对一个个的特征而言的, 就是看一个特征 t, 整个系统中在有它和没它的时候信息量各是多少，两者的差值就是这个特征 t 给系统带来的信息量, 即增益。在信息增益中, 重要性的衡量标准就是看一个特征能够为分类系统带来多少信息, 带来的信息越多，该特征越重要。

信息增益的计算公式如下：

$$\begin{aligned} \mathrm{IG}(t) = &-\sum_{l=1}^{L} P(l)\log P(l) + \\ &P(t)\left[\sum_{l=1}^{L} P(l|t)\log P(l|t)\right] + \\ &P(\overline{t})\left[\sum_{l=1}^{L} P(l|\overline{t})\log P(l|\overline{t})\right], \end{aligned} \tag{11}$$

将式 (11) 中各个事件的概率用其相应的频率来代替，那么信息增益可以用下式来近似计算，

$$\begin{aligned} \mathrm{IG}(t) = &\left\{-\sum_{l=1}^{L}\frac{N_l}{N}\log\frac{N_l}{N}\right\} + \\ &\left(\sum_{l=1}^{L}\frac{A_l}{N}\right)\left[\sum_{l=1}^{L}\frac{A_l}{A_l+B_l}\log\frac{A_l}{A_l+B_l}\right] + \\ &\left(\sum_{l=1}^{L}\frac{C_l}{N}\right)\left[\sum_{l=1}^{L}\frac{C_l}{C_l+D_l}\log\frac{C_l}{C_l+D_l}\right], \end{aligned} \tag{12}$$

由式 (11) 和式 (12) 可以看出，信息增益其实是互信息 $\mathrm{MI}(t,l)$ 和 $\mathrm{MI}(\overline{t},l)$ 的加权平均，其权重分别是联合概率 $P(t,l)$ 和 $P(\overline{t},l)$:

$$\mathrm{IG}(t) = \sum_{l=1}^{L} P(t,l)\mathrm{MI}(t,l) + \sum_{l=1}^{L} P(\overline{t},l)\mathrm{MI}(\overline{t},l). \tag{13}$$

信息增益考察特征对整个系统的贡献，而不是具体到某个类别上。特征 t 的信息增益越大，说明它对于整个系统来说包含的分类信息量越大。

2.4 加权 log 似然比

加权的 log 似然比 (Weighted Log Likelihood Ratio, WLLR)[8] 定义为：

$$\mathrm{WLLR}(t,l) = P(t|l)\log\frac{P(t|l)}{P(t|\overline{l})}, \tag{14}$$

将式 (14) 中的概率用其相应的频率来代替，$\mathrm{WLLR}(t,l)$ 可以用下式来近似计算：

$$\mathrm{WLLR}(t,l) = \frac{A_l}{N_l}\log\frac{A_l(N-N_l)}{B_l N_l}, \tag{15}$$

从式 (15) 中可以看出，对于那些出现频率高并且具有较多鉴别性信息的特征，其 WLLR 值会较大；并且特征的出现频率与其鉴别性相比决定性更大些。

3 实验

3.1 实验配置

本文使用 NIST LRE 2009 标准测试集 [16] 进行特征选择方法的比较。 NIST LRE 2009 的集内测试目标语种共有 23 种, 包括阿姆哈拉语、波斯尼亚语、广东话、克里奥尔语、克罗地亚语、达里语、英语、印度英语、波斯语、法语、乔治亚语、豪萨语、北印度语、韩语、汉语普通话、普什图语、葡萄牙语、俄语、西班牙语、土耳其语、乌尔都语、越南语和乌克兰语。我们使用的训练数据包括 2009 年 NIST 提供的 VOA2 和 VOA3 中提取的部分数据以及电话语音数据 (Conventional Telephone Speech, CTS), 其中 VOA 数据包括 22 个语种：阿姆哈拉语、波斯尼亚语、广东话、克里奥尔语、克罗地亚语、达里语、英语、印度英语、波斯语、法语、乔治亚语、豪萨语、北印度语、韩语、汉语普通话、普什图语、葡萄牙语、俄语、西班牙语、土耳其语、乌尔都语、越南语和乌克兰语；CTS 数据包括 12 个语种：广东话、英语、印度英语、波斯语、法语、北印度语、韩语、汉语普通话、俄语、西班牙语、乌尔都语和越南语。

训练数据中， VOA 数据与 CTS 数据信道差异较大。为了补偿不同信道带来的解码结果差异，我们将 VOA 数据与 CTS 数据分开训练 1 级 SVM 模型。我们分别对 VOA 数据和 CTS 数据类型下的每个语种训练数据进行聚类，每个语种聚类为 10 份，这样共有 $(22+12)*10=340$ 个模型。 2 级训练数据做交叉验证，从而用作后端得分分类器 (SVM 或者 LDA-Gaussian) 的模型训练和参数调整。

本文主要针对时长为 10 s 的测试情况。

本文实验基于第 1 节介绍的 PR-Lattice-VSM 系统框架，系统中使用的音子识别器是 BUT 公布的匈牙利语音子识别器 (BUT_HU), 包含 62 个音子。我们使用第 2 节中提到的 4 种方法分别进行特征选择，并用选出的特征项作为分类器的输入向量。

实验中所使用的性能指标是等错率 (Equal Error Rate, EER)[16] 。

3.2 实验结果

3.2.1 特征选择方法的比较

理论上，对于一个包含有 62 个音子的音子识别器来说，其解码后得到的 3-gram 数应该为 $62*62*62=238328$ 。但在训练数据中，实际存在的 3-gram 数为 184360 。我们去掉期望计数比较小的 (小于 0.1) 的特征项，最后得到的特征总数为 107906 。在本文的特征选择实验中，我们将这 107906 个特征作为特征全集，使用特征全集的系统作为实验的基线系统。

图 3 比较了不同的特征选择方法在 LRE 2009 10 s 测试集上的性能，其中 2 级分类器使用的是基于径向基核数的 SVM 分类。从图 3 可以看出，在 10 s 测试集上，当信息增益和加权 log 似然比方法选择的特征维数为 30 k 时，其性能即可与使用全部特征 (100 k) 时的基线系统性能相当；当选择的特征数为 60 k, 其性能要略好于基线系统。我们还可以看出，当选择的特征数较低时 (5 k, 10 k 和 20 k), 信息增益和加权 log 似然比方法要明显优于传统的互信息和 χ^2 统计量方法，其中加权 log 似然比方法相对更好一些。互信息方法性能较差主要因为它只侧重于特征的鉴别性而没有考虑特征项的代表性，基于这样的标准选择的特征项有很大一部分在训练数据中出现频率较低, 从而会影响系统性能。相比之下，

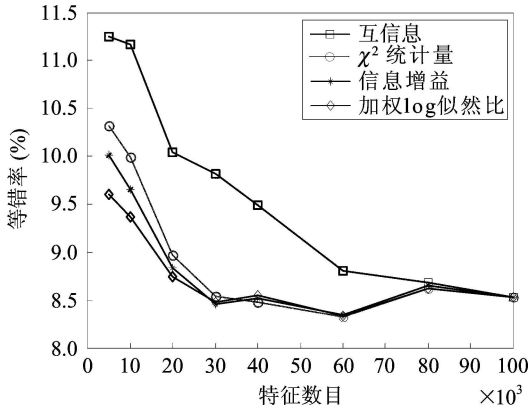

图 3 根据不同特征选择方法选择不同维数特征在 LRE 2009 10 s 上等错率比较图 (2 级分类器为 SVM)

信息增益、加权 log 似然比与 χ^2 统计量一样，会同时考虑特征的出现频率和鉴别性，只是因侧重的程度不同从而产生差异。当选择的特征数量较少时，虽然性能比基线系统差，但因为特征数较低，计算量和内存需求量会相应少很多，在实际应用中，如果考虑速度的因素，并且允许以牺牲部分系统性能为代价，可以选择较低数量的特征，此时几种特征选择的性能比较就有了参考意义。

图 4 比较了不同的特征选择方法在 LRE 2009 10 s 测试集上的性能，其中 2 级分类器使用的 LDA-Gaussian 方法。从图 4 中我们可以看出，此时的性能整体要好于 2 级分类器使用 SVM 时的性能，不同特征选择方法之间的性能趋势跟后者大致相同。信息增益和加权 log 似然比方法都可以在基本不影响识别性能的前提下有效地降低特征数量，甚至可以一定程度地提高系统性能 (选择特征数目为 60 k) 。当选取的特征维数较低时 (5 k, 10 k 和 20 k), 信息增益和加权 log 似然比方法要优于 χ2 统计量和互信息方法。

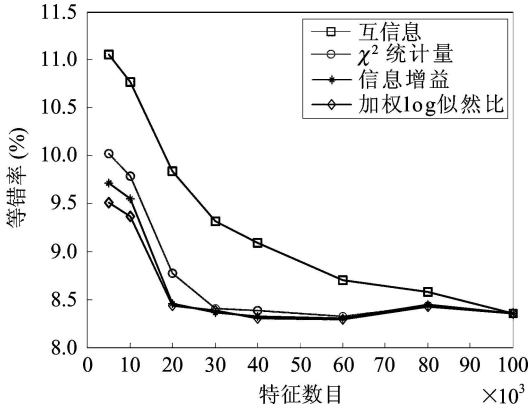

图 4　根据不同特征选择方法选择不同维数特征在 LRE 2009 10 s 上等错率比较图 (2 级分类器为 LDA-Gaussian)

3.2.2 特征扩展实验

按照递归特征选择法 [8]，将 3-gram 特征扩展为 4-gram 特征，即如果根据特征选择准则选取了音子串 $x_1x_2x_3$ 作为 3-gram 特征，那么则将其对应的 4-gram 音子串 $x_1x_2x_3x_i$ 和 $x_ix_1x_2x_3$ 加入到特征向量中，其中 x_i 为音子库中的任意一个音子。本文将 3.2.1 节中选出的 30 k 个 3-gram 特征用于扩展，2 级分类器使用的是 SVM 。实验结果如表 2 所示。

由表 2 可以看到，不管是信息增益法还是加权 log 似然比准则，扩展后得到的 4-gram 特征较 3-gram 特征在性能上都有了较明显的提高，均相对提升了 10% 左右。

表 2　LRE 2009 10 s 测试集上不同 N-gram 阶数 EER 及特征数目的比较

N-gram 阶数	EER(%)	特征数目
基线 3-gram	8.53	107906
信息增益 3-gram	8.46	30000
信息增益 4-gram	7.59	213945
加权 log 似然比 3-gram	8.48	30000
加权 log 似然比 4-gram	7.56	219143

4 分析和总结

在音子配列学语种识别系统中，特征维数太高会带来计算和存储等问题。针对这一问题，将文本分类中信息增益和加权 log 似然比两种特征选择方法引入到语种识别系统中，选取部分子集，从而得到较低维数的特征向量。实验表明，信息增益和加权 log 似然比方法都可以在基本不影响识别性能的前提下有效地降低特征数量，甚至可以一定程度地提高系统性能。当选取的特征维数较低时，信息增益和加权 log 似然比方法要优于已有的 χ^2 统计量和互信息方法。

在接下来的工作中，我们将根据选择出 4-gram 向更高阶的 5-gram 、6-gram 扩展，从而进一步提高语种识别系统的性能，并研究不同特征选择方法的互补性。

致谢

本论文工作由国家自然科学基金 (10925419, 90920302, 61072124, 11074275, 11161140319, 91120001) 、中国科学院战略性先导科技专项 (面向感知中国的新一代信息技术研究，XDA06030100, XDA06030500) 、国家 863 计划 (2012AA012503) 和中国科学院重点部署项目 (KGZD-EW-103-2) 经费资助。

参　考　文　献

1　Zhao L, Zou C, Wu Z. Integration of speech and language processing in Chinese continuous speech recognition. *Chinese Journal of Acoustics*, 2002; **21**(4): 343—351

2　Liang C, Zhang X, Yang L, Zhang J, Yan Y. Perceptual MVDR-based cepstral coefficients (PMCCs) for speaker recognition. *Chinese Journal of Acoustics*, 2012; **31**(4): 489—498

3　Petr Schwarz, Matejka Pavel, Jan Cernocky. Hierarchical structure of neural networks for phoneme recognition. In: Proceedings of ICASSP, 2006: 325—328

4 Zissman M. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. Speech and Audio Processing*, 1996; **4**(1): 31—44
5 Li h, Ma B, Lee C H. A vector space modeling approach to spoken language identification. *IEEE Trans. Audio, Speech and Language Processing*, 2007; **15**(1): 271—284
6 Joachims T. Learning to classify text using support vector machine. *Kluwer Academic Publishers*, 2002
7 Campbell W M, Richardson F, Reynolds D A. Language recognition with word lattices and support vector machine. In: Proceeding of ICASSP, 2007
8 Richardson F S, Campbell W M. Language recognition with discriminative keyword selection. In: Proceeding of ICASSP, 2008: 4145—4148
9 Yang Y, Pedersen J. A comparative study on feature selection in text categorization. In: The 14th International Conference on Machine Learning, 1997: 412—420
10 Nigam K, McCallum A, Thrun S, Mitchell T. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 2000; **39**(2/3): 103—134
11 Tong R, Ma B, Li H, Chng E S. Selecting phonotactic features for language recognition. In: Proceeding of ICASSP, 2010: 737—740
12 Young S, Gunnar Evermann, Thomas Hain, Kershaw D, Gareth Moore, Odell J, Ollason D, Valtchev D, Woodland P. The HTK book, Entropic, Ltd. Cambridge, UK, 2002
13 Tong R, Ma B, Li H, Chng E S. A target-oriented phonotactic front-end for spoken language recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 2009; **17**: 1335—1347
14 Kenneth Ward Church, Patric Hanks. Word association norms, mutual information and lexicography. In: Proceedings of ACL 27, Vancouver, Canada, 1989: 76—83
15 Tom Mitchell. Machine Learning. McCraw Hill, 1996
16 NIST language recognition evaluation plan 2009. http://www.itl.nist.gov/iad/mig/tests/lre/2009/
17 Buhmann M D. Radial basis functions. *Acta Numerica*, 2001; **9**: 1—38
18 Noor E, Aronowitz H. Efficient language identification using anchor models and support vector machines. IEEE Odyssey: The Speaker and Language Recognition Workshop, 2006: 1—6