距离法求语言特征主成分①

邵泽国

(上海师范大学 语言研究所, 上海 200234)
(上海电子信息职业技术学院, 上海 201411)

摘 要: 语言特征成分的数量决定了成分对语言特征的影响度. 作者把影响度视作距离, 建立了一个数学模型, 通过距离判断来求取语言特征主成分.

关键词: 语言特征; 主成分; 距离

Distance Method to Calculate the Principal Component of Linguistic Features

SHAO Ze-Guo

(Institute of Chinese Linguistics, Shanghai Normal University, Shagnhai 200234, China)
(Shanghai Technical Institute of Electronics & Information, Shagnhai 201411, China)

Abstract: The quantity of linguistic features component determines the component-linguistic feature influence degree. Considering influence degree as the distance, this paper establishes a math model to calculate the principal component by judging the distance.

Key words: linguistic features; principal component; distance

通过语言调查取得的语言基础材料, 都会详尽的描述语言的各个特征属性(如语音的声母、韵母、声调, 词汇的义项等), 每个特征都会由若干个成分(如声母 b、ph、m、f 等)来描述, 各个成分对特征的影响力是不同的. 当对语言做进一步研究时(如语言分类、语言比较), 就需要先求出各个特征的主成分, 而不是把所有成分都作为处理对象. 原因很简单, 在众多特征成分中有一部分成分对特征的描述或区分贡献甚微, 将其忽略掉几乎对整体特征的描述没有影响; 保留这些次要成分不仅会增加计算量, 有时反而会影响分析和研究结果.

目前, 常见有两种方法来求取语言特征主成分. 一是当语言材料数据量较大时会用数量值最大的成分来表征语言特征; 二是当语言材料数据量较小时通过人工观测来决定主成分. 用数量值最大的成分来表征特征, 显然对特征的描述力度不够, 数量值最大的成分成为主成分的成员是必要的, 但不是充分的; 人工观测能够保证主成分的充分、必要成员都出现(假设此人有较高的语言学水平), 但耗时较大, 主观性较强, 往往出现相同特征在不同语言点上的主成分确立的标准不一致.

作者设计的求语言特征主成分的距离法, 试图弥补以上两种方法的不足. 该算法通过程序实现, 可以高效处理大规模的语言数据, 同时作者在算法设计时致力追求算法的语言无关性.

1 语言特征成分分析

为了方便说明, 这里从钱乃荣先生的吴语基础语音材料中随机挑选江苏苏州和浙江杭州两个语言点的材料来示例说明. 作者经过简单加工得到表 1, 其中苏州现代韵母 50 个, 调查字 5575 个, 杭州现代声母 30 个, 调查字 3812 个.

在表 1 的“特征成分及数量”列, 每个小括号是一个数据对, 括号内逗号前是特征成分, 逗号后是该成分的数量. 例如(i,417)表示在苏州 5575 个调查字中有 417 个字的现代韵母读“i”, 同样(ɦ,364)表示在杭州 3812 个调查字中有 364 个字的现代声母读“ɦ”. 同时数据对按照成分数量降序自左向右排列.

① 收稿时间:2012-06-13;收到修改稿时间:2012-08-05

表 1 苏州现代韵母与杭州现代声母特征成分

语言点	特征	特征成分及数量
苏州	现代韵母	(i,417)(ɛ,408)(in,338)(ən,307)(iɪ,277)(iəʔ,258)(əu,254)(a,249)(æ,237)(ɔʔ,210)(ø,203)(iæ,198)(ɑ,196)(əʔ,192)(ʏ,189)(ɔ,141)(ɔŋ,140)(ɥ,126)(uɛ,121)(ɳ,112)(ia,104)(ɑʔ,92)(iɑ,77)(iʏ,70)(ʊ,67)(uɑ,59)(iø,58)(aʔ,53)(iɔŋ,43)(uø,43)(oŋ,42)(yəʔ,40)(iɔʔ,36)(iɑʔ,32)(uən,31)(y,30)(ʔn,24)(uəʔ,20)(u,18)(yn,12)(əl,10)(uaʔ,8)(iɔ,8)(ioŋ,6)(m,5)(ua,5)(uɑʔ,5)(n,2)(iaʔ,1)(yaʔ,1)
杭州	现代声母	(ɦ,364)(ts,247)(tɕ,232)(l,218)(s,216)(ʔ,215)(dz,175)(k,170)(ɕ,162)(tsh,155)(d,148)(dʑ,145)(m,141)(p,126)(z,124)(b,111)(tɕh,106)(t,106)(kh,97)(th,90)(h,78)(f,74)(ph,69)(v,67)(ŋ,53)(n,51)(ɹ,31)(g,20)(ŋ,16)(ɻ,5)

从表 1 中我们可以看出不管是苏州的现代韵母还是杭州的现代声母，其成分都比较多，而且成分数量不等，同时存在数量值差异很大的情况．显然，不需要太多语言学知识也能看出用“i”来表征苏州的现代韵母远比“yaʔ”的力度大．而且由于“yaʔ”的数量值过低，考虑到记音过程有可能存在误差，所以“yaʔ”通常会被舍弃．另外，虽然“i”在苏州的现代韵母中数量值最大，但若单独用“i”来表征苏州的现代韵母也不合适，毕竟它所表征的信息量偏少．杭州的现代声母也是同样情况．所以要找到一个恰当表征特征信息的成分集合——主成分，它要满足：用尽可能少的成分来尽可能多的表征特征信息．

2 距离法数学模型

根据上面的分析，可以看出成分数量值越大，它所表征的特征信息越多．若把一个成分数量值看做一个元素（记作 r_1、r_2…r_n, n 为成分的种类数），那么一个特征的所有成分数量值就构成了一个数学上的集合（记作 R），则有数学模型——集合 R{r_1, r_2…r_n}．

我们再为集合 R 添加一个元素“0”，这个“0”元素在语言学意思上表示对应特征成分的数量值为零，即无意义，对语言对象的研究和处理不产生影响，这样集合 R 变为{r_1, r_2…r_n, 0}．把元素放在一个直角坐标系上来观察，元素“0”放在坐标原点，其它元素放在距离原点为自身数量值的位置，如图 1．

如果我们把元素（如图中 r_1、r_2、r_3、r_4、0）当作一个圆的半径，那么由它构成的圆的面积的大小在意义上和语言特征成分对语言特征的影响力的大小是吻合的．即，面积越大影响力越大．简言之，我们可以通过判断元素到原点的距离来求取主成分．

由于集合 R 元素存在这样的规律 $r_1 \geqq r_2 \geqq \cdots \geqq r_n > 0$，便知取主成分元素可从由 $r_1 \rightarrow r_n$ 的方向取．关键问题在于取到哪个元素为止．

再看图 2，引入 $r_m=(r_1+r_2+r_3+r_4+0)/4$，$r_m$ 的引入把元素 r_1、r_2、r_3、r_4、0 分成了两组：r_1、r_2、r_3 一组，r_4、0 一组．r_m 成了分水岭，前一组的累计面积远大于后一组，且大于等于总面积的 50%．因此主成分必在前一组中且必不在后一组中．

是不是前一组中的所有元素都可以成为主成分呢？未必．原因是：前一组满足了表征信息量较大，但不一定满足成分数量较小．所以还要对前一组中的元素做进一步的距离判断．

如图 3，我们观察前一组相邻元素的间距，r_1 与 r_2 的间距记作 d_1，r_2 与 r_3 的间距记作 d_2．再取平均间距 d_m,$d_m=(d_1+d_2)/2$，如图 4．这时 d_m 的引入又将元素 r_1、r_2、r_3 分成两组，其中 r_1、r_2 为一组，r_3 为另一组．这时{ r_1, r_2}便是理想的主成分．依据是：首先 r1 是主成分必取的元素，接下来的 r_2、r_3 要不要取就看他们和 r_1 的紧密度，若间距小于平均间距，则紧密度较高，则取之．

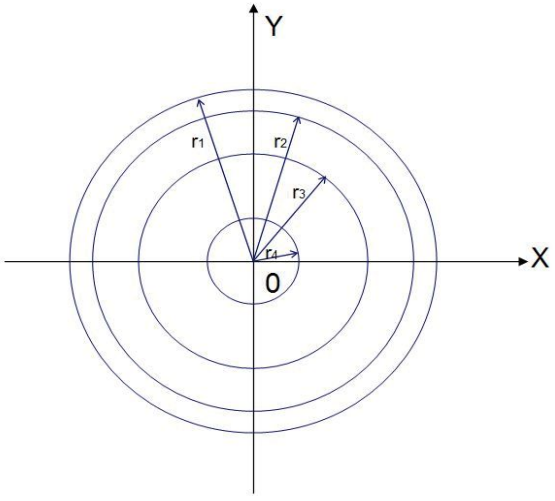

图 1 集合元素在直角坐标系上

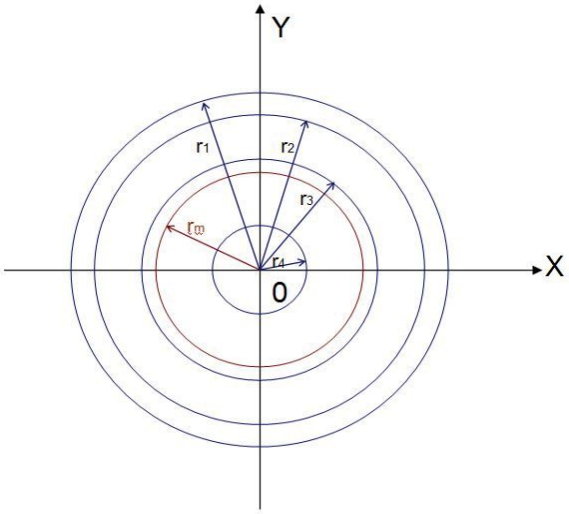

图 2　平均距离 r_m

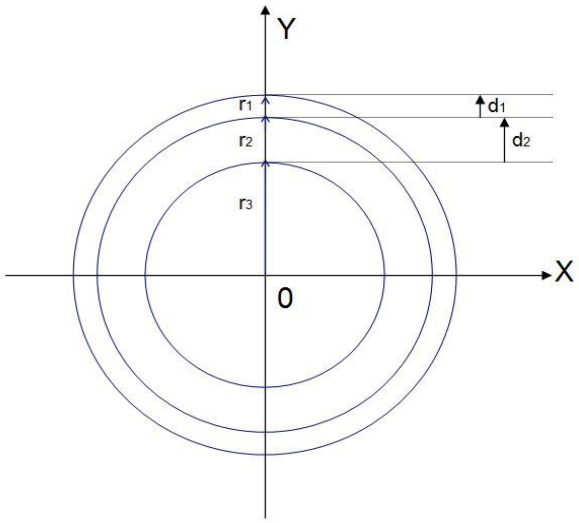

图 3　相邻元素间距

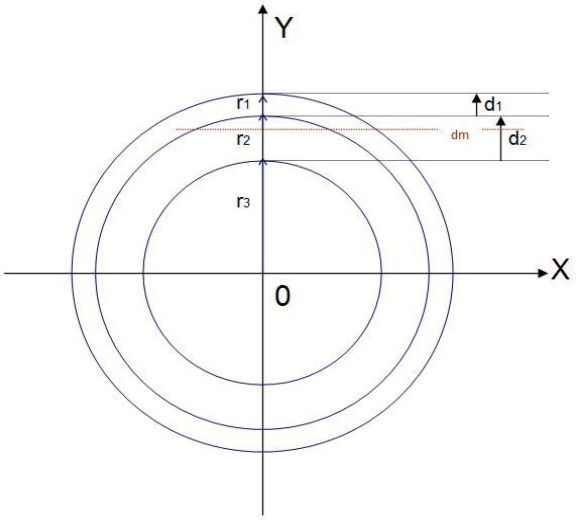

图 4　平均间距

3　距离法算法

已知集合 R{r_1, r_2…r_n}，且 $r_1 \geqq r_2 \geqq \cdots \geqq r_n$.

(1) 定义一个变量 r_m, $r_m=(r_1+r_2+\cdots+r_n)/(n+1)$;

(2) 求 R 的子集 R_m\{ r_1, r_2…r_k},要求满足 $r_k \geqq r_m$ 且 $r_{(k+1)}<r_m$;

(3) 求 R_m 元素间距集合{d_1,d_2…$d_{(k-1)}$}，$d_1=r_1-r_2$, $d_2=r_2-r_3\cdots d_{(k-1)}=r_{(k-1)}-r_{(k-2)}$;

(4) 求平均间距 d_m, $d_m=(d_1+d_2+\cdots+d_{(k-1)})/(k-1)$;

(5) 自 d_1 到 $d_{(k-1)}$依次与 d_m 比较，当第一个出现大于 d_m 时终止，生成集合 D\{ d_1,d_2…d_x},$d_{(x+1)}$是第一个大于 d_m 的;

(6) 将集合 D 中的元素对应集合 R 的元素(d_1 对应 r_2,d_2 对应 r_3，依此类推),再加上元素 r_1,生成集合 R'\{ r_1, r_2…r_x };

(7) 将 R'中的元素对应特征成分，即得到语言特征主成分.

4　距离法实例测试

这里用苏州的现代韵母数据来测试距离法算法.先生成两个集合:

特征成分集合{ i，ɛ，in，ən，iɪ，iəʔ，əu，a，æ，ɔʔ，ø，iæ，ɑ，əʔ，ʏ，ɔ，ɔŋ，ɥ，uɛ，ɿ，ia，ɑʔ，iɑ，iʏ，ʊ，uɑ，iø，aʔ，iɔŋ，uø，oŋ，yəʔ，iɔʔ，iɑʔ，uən，y，ɲn，uəʔ，u，yn，əl，uaʔ，iɔ，ioŋ，m，ua，uɑʔ，n，iaʔ，yaʔ}；

集合 R{417, 408, 338, 307, 277, 258, 254, 249, 237, 210, 203, 198, 196, 192, 189, 141, 140, 126, 121, 112, 104, 92, 77, 70, 67, 59, 58, 53, 43, 43, 42, 40, 36, 32, 31, 30, 24, 20, 18, 12, 10, 8, 8, 6, 5, 5, 5, 2, 1, 1}, n 为 50.

(1) 求 r_m

r_m=(417+408+338+307+277+258+254+249+237+210+203+198+196+192+189+141+140+126+121+112+104+92+77+70+67+59+58+53+43+43+42+40+36+32+31+30+24+20+18+12+10+8+8+6+5+5+5+2+1+1)/(50+1)≈109.31.

(2) 求 R_m

R_m\{417, 408, 338, 307, 277, 258, 254, 249, 237, 210, 203, 198, 196, 192, 189, 141, 140, 126, 121, 112}, 其中 k 为 20.

(3) 求 R_m 元素间距集合

结果为: {9, 70, 31, 30, 19, 4, 5, 12, 27, 7, 5, 2, 4, 3,

48, 1, 14, 5, 9}.

(4) 求平均间距 d_m

d_m=(9+70+31+30+19+4+5+12+27+7+5+2+4+3+48+1+14+5+9)/(20-1) ≈16.05.

(5) 生成集合 D

结果为: {9}.

(6) 生成集合 R’

结果为: {417, 408}.

(7) 得到语言特征主成分

结果为: i, ɛ.

5 结语

距离法求语言特征主成分的算法已经软件化, 是汉语方言地理信息系统（系潘悟云先生主持的教育部哲学社会科学研究重大课题攻关项目——汉语方言地理信息系统平台建设的成果）的一个重要的功能软件模块. 他在计算音系特征主成分及词汇义项主成分上得到了语言学研究者的普遍认可.

参考文献

1 Dasgupta S, Papadimitriou C, Vazirani U. 算法概论. 北京: 清华大学出版社,2008.

2 陆致极.汉语方言数量研究探索.北京:语文出版社,1992.

3 Hyvarinen A, Karhunen J, Oja E.独立成分分析.北京:电子工业出版社,2007.

4 赖国毅,陈超.SPSS 17.0 中文版常用功能与应用实例精讲.北京:电子工业出版社,2010.

5 马逢时,吴诚鸥,蔡霞.基于MINITAB的现代实用统计.北京:中国人民大学出版社,2009.

6 宗成庆.统计自然语言处理.北京:清华大学出版社, 2011. 77–90.

7 林海明,张文霖.主成分分析与因子分析的异同和 SPSS 软件--兼与刘玉玫、卢纹岱等同志商榷.统计研究,2005,(3):65 –69.

8 严慧,金忠,杨静宇.最小化相关性的二维主成分分析.模式识别与人工智能,2010,(1).

9 李靖华,郭耀煌.主成分分析用于多指标评价的方法研究——主成分评价.管理工程学报,2002,16(1).

10 郑伟.吴语太湖片果摄的演化模式与历史层次.语言科学, 2009,(4).

11 吴波.江淮官话语音研究[博士学位论文].上海:复旦大学, 2007.

12 钱乃荣.当代吴语研究.上海:上海教育出版社,1992.

(上接第 181 页)

的补偿, 仿真结果也证明了该装置对无功补偿的有效性和可靠性, 通过现场测试结果证明了本文提出的基于相控投切无功补偿装置的应用价值.

参考文献

1 朱晓清.基于 DSP 无功补偿控制器的研究.哈尔滨理工大学,2009.

2 丁富华.真空开关的选相控制及其应用研究.大连理工大学, 2006.

3 张庆杰,袁海文.配永磁机构的真空断路器同步分合闸控制系统设计与实现.电力自动化设备,2010.

4 方春恩,王佳颖,邹积岩.并联电容器组同步关合最佳目标相位的确定.电工技术学报,2006,21(1):24–27.

5 王奔,杨明,罗文.高压动态无功补偿智能选相投切控制器.四川省科学城久信科技有限公司,2011,2.

6 丁富华,段雄英,邹积岩.基于同步真空断路器的智能无功补偿装置.中国电机工程学报,2005,25(6).

7 方彦.基于永磁机构的智能选相真空断路器的研究和应用.天津大学,2005,6.

8 汪玉凤,刘芳芳,薛建清.MSC 动态无功补偿选相投切控制器的研究.电力电子技术,2011,45.