特征主成分算法再探①

邵泽国

(上海师范大学 语言研究所, 上海 200234)
(上海医疗器械高等专科学校, 上海 200093)

摘 要: 对作者之前设计的求语言特征主成分的“距离法”做了深度解析, 指出了“距离法”在应用过程中, 切分成两个集合的“断点”会影响计算结果的精度. 借助聚类分析的思想重新设计了语言特征主成分算法——“聚类法”来弥补“距离法”的缺陷. 最后指出两个方法的缺陷可以相互弥补, 在实际应用中应结合使用.

关键词: 语言特征主成分; 距离法; 聚类分析

Restudy of Features Principal Component Algorithm

SHAO Ze-Guo

(Institute of Language, Shanghai Normal University, Shanghai 200234, China)
(Shanghai Instrumentation Colledge, Shanghai 200940, China)

Abstract: This paper analyzes the “distance method” that is an algorithm of language features and points out the questions of “distance method”: “Breakpoint” that be used cut into two sets will affect the accuracy of calculation. It designs a new algorithm named “clustering method” with the idea of clustering analysis. The defects of two methods can complement each other and should be used together in practical application.

Key words: principal component of language features; distance method; cluster analysis

求取语言特征主成分是进行语言比较、语言分类、画语言地图等语言研究工作非常关键的前奏. 作者曾撰文《距离法求语言特征主成分》(计算机系统应用, 2012(1)), 文中作者建立了一个“距离”数学模型——特征成分的数量值落在直系坐标上的点到坐标原点的距离视作为特征成分对特征的影响度——来求取语言特征主成分. 该算法应用在了汉语方言地理信息系统平台上(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), 得到了用户(以语言学研究者为主)的普通认可. 但近期作者使用该算法对更多方言调查数据进行特征主成分计算时, 发现有小部分结果不尽人意. “距离法”通过两次取均值(本文将均值称之为切分集合的“断点”), 将语言特征成分集切分成两个不同的子集, 其中一个子集为主成分集, 另一个为非主成分集. 这一思想其实质就是聚类的思想, 是符合语言主成分应用需求的. 很显然, 切分两个子集的断点是整个计算的关键, 对结果的准确性和精度有直接的影响. “距离法”出现的“小部分结果不尽人意”, 会不会原因出现在断点上? 鉴此, 本文拟在“距离法”的基础上, 借鉴聚类分析的思想再探语言特征主成分的算法.

1 “距离法”介绍

这里仅介绍“距离分”法的基本思想和处理过程, 详情可参考《距离法求语言特征主成分》一文.

(1)“距离法”的基本思想

在语言调查材料中特征成分的主次体现在成分的数量上. 如下表:

表 1 是从上海师范大学语言研究所的方言调查材料数据库中挑选的山西沁县及山西怀仁两个语言点的“見母开口二等字”现代声母和现代韵母的种类及其数量. 我们认为成分的数量(如山西沁县的现代声母 tɕ 的

① 基金项目:教育部哲学社会科学研究重大课题攻关项目(09JZDH007)
收稿时间:2014-04-22;收到修改稿时间:2014-11-28

数量为 61 个)代表了成分对特征的影响度(声母 tɕ 对山西沁县的现代声母的影响度为 61 个单位).

表 1　“見母开口二等字”的现代声母及韵母

方言点	声母(数量)	韵母(数量)
山西沁县	tɕ(61)	iɔ(20)
	k(12)	ɪ(12)
	kh(1)	iɛ(11)
	ɕ(1)	iɑ(9)
		əŋ(6)
		ia(5)
		iʌʔ(5)
		ɔ(5)
		ʌʔ(2)
山西怀仁	tɕ(34)	iæ(12)
	k(15)	iɔu(9)
	tɕh(7)	əŋ(9)
	ɕ(3)	iɔ(6)
	kh(1)	iɑ(5)
		ia(5)
		iaʔ(3)
		yɑʔ(3)
		ɑʔ(3)
		ɔ(3)
		iəŋ(1)
		ɔu(1)

(2)“距离法”的处理过程

基于上述思想, 我们先为每个特征建立一个成分数量集合. 如山西沁县现代韵母数量集合:

I{20,12,11,9,6,5,2}

注, 为了方便处理我们在特征成分数量统计的同时按照数量值从大到小做了排序. 以下以求山西沁县现代韵母主成分为例说明:

①为集合 I 新增一个元素“0”, 得到结合{20,12, 11, 9,6,5,2,0};

②求集合{20,12,11,9,6,5,2,0}各元素到原点的距离均值 M1 为 8.13;

③舍去集合{20,12,11,9,6,5,2,0}中小于等于 M1(8.13)的元素, 得到新集合{20,12,11,9};

④求集合{20,12,11,9}中相邻元素的距离(即数量差)及均值, 元素“20”与“12”的距离为 8, “12”与“11”的距离为 1, “11”与“9”的距离为 2, 均值 M2 为(8+1+2)/3 =3.67;

⑤在集合{20,12,11,9}中从大到小取元素, 第一个取出的是“20”, 然后判断下一个要取出的元素和当前取出的元素的距离是否小于均值 M2, 若小则取出, 再继续; 若大于或等于则不取并操作结束; 该例结果为{20};

⑥数量值在结果集合{20}里的特征成分“iɔ”即为主成分.

通过该方法计算出来山西沁县的声母主成分为 tɕ, 韵母主成分为 iɔ; 山西怀仁的声母主成分为 tɕ, 韵母主成分为 iæ.

2　“距离法”的不足

作者经过分析认为问题出在 M_1 和 M_2 上, 这两个都是均值, 用均值作为切分集合的断点, 这个“断点”会被分到哪个集合里, 即计算式中是否要等于 $M_1(M_2)$?

(1)M_1 不合理

M_1 将集合元素分成了两部分(见上述第 3 步), 舍去了小的部分, 但有时主成分成员会在小的部分中而被舍掉. 如原集合 I 为{51,41,31}和{50,40,30}, 若语言特征成分数量集合是以上两个集合, 那么语言特征主成分都是全集才是合理的. 但用“距离法”得到的结果却是{51,41,31}和{50,40}. 集合{50,40,30}中的“30”被舍弃, 在语言学应用上, 语言学家认为是不合理的, 一般来讲“50”、“40”、“30”所对应的成分会认作为主成分. 这样看来, 否要等于 M_1 是不确定的.

(2)M_2 不合理

计算 M_2 意义上是计算最大元素同其它元素的聚合度, 但取均值也有待商榷, 所以这个 M 的精度存在问题. 基于这一点我们对山西怀仁的韵母主成分产生了质疑. 对 M_2 取值的质疑点在于是否取“=”值, 根据大量的实验结果数据表明, “=”值在不同案例里被语言学家判断为是主成分的情况是不一的. 所以, 否要等于 M_2 也是不确定的.

总结来说, “距离法”M_1、M_2 这两个断点在绝大多数情况下不会是待分集合中的元素, 这种情况对我们的计算结果不会产生丝毫的负面影响. 但, 一旦 M_1、M_2 有一个的值恰好为待分集合中的某个元素值得时候就有可能降低计算结果的精度. 显然, 在这种情况下断点是否要等于 $M_1(M_2)$应具有相应的灵活度.

3　聚类分析与主成分

“聚类分析指将物理或抽象对象的集合分组成为

由类似的对象组成的多个类的分析过程”, 这个思想同我们提取语言特征主成分的思想不谋而合: 所谓提取主成分, 就是将特征成分聚类成两类, 一类是主要成分集合, 一类是非主要成分集合, 舍去非主要的保留主要的.

有了这个理论做基础我们可以尝试用聚类分析的方法求语言特征主成分. 以下仍以山西沁县现代韵母为例.

(1)距离关系矩阵

计算出集合I{20,12,11,9,6,5,2}两两元素间的距离, 得到距离关系矩阵如下表:

表 2 距离关系矩阵 1

	20	12	11	9	6	5	2
20							
12	8						
11	9	1					
9	11	3	2				
6	14	6	5	3			
5	15	7	6	4	1		
2	18	10	9	7	4	3	

(2)聚类图

对表 2 按距离关系值从小到大用“平均连接法”画出聚类图, 如下:

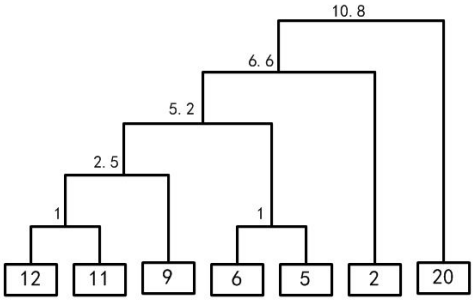

图 1 聚类图 1

解释: 图中元素“12”和“11”聚成了最小组, 它们之间的平均距离为 1; 元素“9”与“12”、“1”这个小组再聚合成一个组, 它们之间的平均距离(聚合度)为 2.5, 其余以此类推.

(3)判断主成分

聚类图 1 将集合I{20,12,11,9,6,5,2}的 7 个元素聚合成了不同的组. 对应语言特征成分, 只要聚集成两个组即可, 即主成分组和非主成分组. 那么只要断开聚类图 1 中平均距离最大的那条连接线就可以了, 如下图:

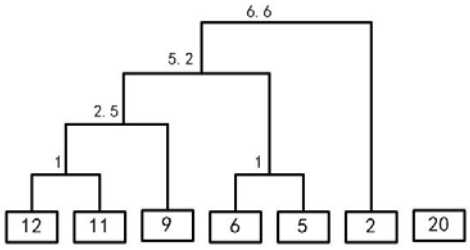

图 2 断开最大连线的聚类图

此时元素 20 为一组(主成分组), 元素 12, 11, 9, 6, 5, 2 非主成分组. 这个结果同“距离法”吻合.

同样方法算出的山西沁县的声母主成分、韵母主成分以及山西怀仁的声母主成分都与“距离法”计算的结果相同, 但山西怀仁的韵母主成分却不同.

我们看由山西怀仁的韵母主成分数量集合, 得到距离关系矩阵, 如下:

表 3 距离关系矩阵 2

	12	9	6	5	3	1
12						
9	3					
6	6	3				
5	7	4	1			
3	9	6	3	2		
1	11	8	5	4	2	

画出的聚类图, 如下:

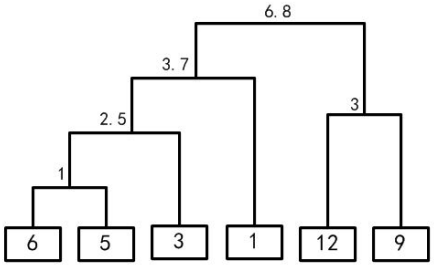

图 3 聚类图 2

断开最大距离连线后的距离图为:

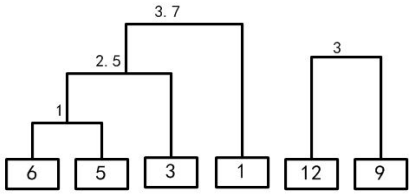

图 4 聚类图 3 断开最大距离连线后

由此图可看出元素“12”、“9”为主成分组, 所以山西怀仁的韵母主成分应该为 iæ、iɔu.

为何“距离法”的结果于此不同？是不是在“距离法”第 5 步判断元素间的距离是否小于均值 M_2 时而取了“小于的”, 若改成“小于或等于”呢？

测试发现, 若改成“小于或等于”得到的结果是“12”、“9”、“6”, 通用不符合预期. 原因是均值 M_2 是在原集合 I{12,9,6,5,3,1}经过一次由 M_1 排除“小成分”之后的子集{12,9,6}上计算出来的, 不是针对所有元素间距做计算, 所以 M_2 被放大. 另外, 元素的间距相同并不能表明元素与其他元素的聚合度相同.

例子中元素“12”和“9”的间距为 3, 元素“9”和“6”的间距也为 3. 如果不是“12”和“9”先聚合, 而是“9”和“6”(其实是“9”和“6”所在的组聚合, 因为“6”之前已与“5”聚合了, “6”和“5”的间距为 1, 优先度高)得到的聚类图就如下:

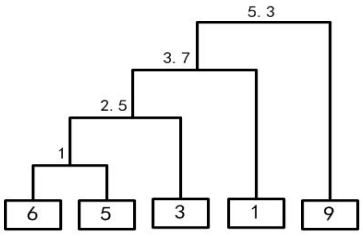

图 5　“9”和“6”聚合

由图知“9”和“6”聚合的平均聚合度为 5.3, 然而“12”和“9”聚合的平均聚合度为 3(见图 6), 比“9”和“6”的小, 所以选择“12”和“9”先聚合组成一个组, 再和“6”所在的组聚合.

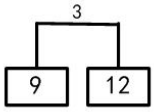

图 6　“12”和“9”聚合

4　简化基于聚类分析的主成分算法

上述基于聚类分析的主成分算法还是有些复杂的, 我们可以进一步简化.

观察发现, 在画距离图时如果不考虑计算聚合时的组(或元素)与组(或元素)之间的平均距离, 只需要两两元素间的距离值即可画出“聚类”图. 但前提是所有元素要排序(升序、降序均可). 这样聚类图 1、聚类图 2 就变化为:

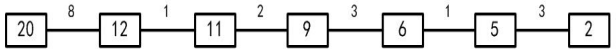

图 7　对聚类图 1 简化

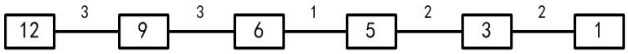

图 8　对聚类图 2 简化

要是简化聚类分为两组只要断开其中的一条连线即可, 当然针对求语言特征主成分任然要断开距离值最大的那条. 如若出现最大距离不唯一(如图 5), 则断开最坐边的那条(确切的讲应是: 当元素按降序(升序)排列时断开排序最小(大)的那条). 断开最大距离连线的简化聚类图如下:

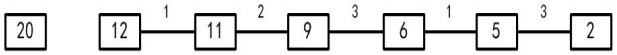

图 9　图 7 断开最大连线

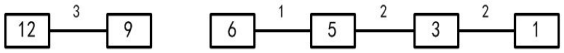

图 10　图 8 断开最大连线

5　总结

在特定情况下“距离法”断点等值问题降低了该算法的精度. 基于聚类分析的主成分算法避开了“等值”问题, 此外它与“距离法”相比具有另外两个优势:

一是语言特征成分的这种主次之分是语言的一种自然属性, 通过语言成分间的某种关系进行聚类能够区分哪些是主要成分哪些是次要成分. 基于聚类分析的主成分算法(以下简称为“聚类法”)比“距离法”思想上更具“自然性”, 更利于从学理上解释. 二是“聚类法”经过简化后, 算法的复杂度要小于“距离法”, 这将更利于计算机软件实现.

是不是“距离法”就一无是处, 可以被“聚类法”完全取代呢？事实不然, 也有“聚类法”不能应对的特殊情况. 比如, 当待分集合中的元素呈现为一个等差数列时, “聚类法”聚合得到的仍然是一个集合而不是两个.

这两个方法各有优势, 同时又各有短板. 可幸的是造成这两个方法“短板”的特殊情况 “距离法”的 M1(M2)是待分集合元素, “聚类法”可以解决; “聚类法”的待分集合元素为等差数列, “距离法”可以解决. 所有在实际应用时可以将这两个方法结合使用, 取长补短.

参考文献

1 邵泽国.距离法求语言特征主成分.计算机系统应用,2012,21(1):192−195.

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

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 高新波.模糊聚类分析及其应用.陕西西安:西安电子科技大学出版社,2004.

7 曲福恒,等.模糊聚类算法及应用.北京:国防工业出版社,2011.

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 陆致极.汉语方言数量研究探索.北京:语文出版社,1992.

11 冯志伟.自然语言处理的形式模型.合肥:中国科学技术大学出版社,2010.

12 宗成庆.统计自然语言处理.北京:清华大学出版社,2011.

13 江铭虎.自然语言处理.北京:高等教育出版社,2006.