东北师范大学

硕士学位论文

相似性传播算法及其在中国经济区域划分中的应用

姓名：高艳敏

申请学位级别：硕士

专业：概率论与数理统计

指导教师：张宝学

20090501

摘 要

通过识别一组代表点来聚类数据对于探测数据模式是非常重要的. 随机抽取数据点集然后反复修正则可以找到这些代表点, 但只有当初始的选择非常好的时候这种方法才是有效的. 2007 年 *Frey* 与 *Dueck* 给出了一种新的方法, 称为 "相似性传播". 相似性传播需要应用每组数据点之间的相似性测度. 实值信息在数据点之间相互交换直到找到一组高质量代表点, 其相应的分类是逐渐呈现的. 应用相似性传播对数据聚类的误差要小于其他方法, 并且它所用的时间少于其他方法的百分之一. 在这篇论文中, 我们使用相似性传播算法处理当前热点问题 -- 中国各省市自治区经济区域划分问题.

关键词: 相似性传播, 相似性测度, 代表点, 责任性, 可用性

Abstract

Clustering data by identifying a subset of representative examples is important for processing sensory signals and detecting patterns in data. Such "exemplars" can be found by randomly choosing an initial subset of data points and then iteratively refining it, but this works well only if that initial choice is close to a good solution. Frey and Duck devised a method called "affinity propagation," which takes ad input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of ecemplars and corresponding clusters gradually emerges. In this paper, we use affinity propagation to deal with a current hot issue–the division of the economic region in China.

Keywords: Affinity propagation, similarity, exemplar, responsibility, availability

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果. 据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东北师范大学或其他教育机构的学位或证书而使用过的材料. 与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意.

学位论文作者签名 高艳敏 日期：2009.05

学位论文版权使用授权书

本学位论文作者完全了解东北师范大学有关保留、使用学位论文的规定，即：东北师范大学有权保留并向国家有关部门或机构送交学位论文的复印件和磁盘，允许论文被查阅和借阅. 本人授权东北师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存、汇编学位论文.

(保密的学位论文在解密后适用本授权书)

学位论文作者签名：高艳敏 指导教师签名：张玉学

日 期：2009.05 日 期：2009.05

学位论文作者毕业后去向：

工作单位：_____ 电话：_____

通讯地址：_____ 邮编：_____

引　言

聚类是人类最原始的精神活动，用于处理他们每天接收到的大量信息，将每个信息片段作为一个单独实体进行处理是不可能的．因此，人类试图将实体（如对象，个人，事件）分类，每一类由它包含的实体的共同特征来标识．

聚类分析的基本思想非常朴素，直观和简单．它是根据各个待分类的模式特征相似程度进行分类的，相似的归为一类，不相似的归为一类，简单地说，相似就是两个特征矢量各分量之间分别较接近．聚类分析包括两个基本内容：模式相似性的度量和聚类算法．

为了能分化模式的类别，必须首先定义模式相似性测度，以此来描述各模式之间特征的相似程度．其中最常见的相似性测度为距离测度（差值测度）．这类测度是以两个矢量矢端得距离作为考虑的基础，因此距离测度值是两矢量各相应分量之间的函数．两矢量之间距离测度的具体算法有许多种，一般的讲，两矢量的距离定义应满足下面的公理．设矢量 x 和 y 的距离记为 $d(x,y)$.

1.$d(x,y)\geq 0$, 当且仅当 $x=y$ 时等号成立，即 $d(x,y)=0\Leftrightarrow x=y$,

2.$d(x,y)=d(y,x)$,

3.$d(x,y)\leq d(x,z)+d(z,y)$.

需要指出的是，模式识别中定义的某些距离测度不满足公理 3, 只是在广义意义上称之为距离．下面给出一些距离测度的具体形式：

1. 欧氏（ Euclidean ）距离：

$$d(x,y)=\|x-y\|=\Big[\sum_{i=1}^{n}(x_i-y_i)^2\Big]^{1/2},$$

2. 绝对值距离（街坊距离或 Manhattan ）：

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i|,$$

3. 切氏（ Chebyshev ）距离：

$$d(x,y)=\max_i|x_i-y_i|,$$

4. 明氏（ Minkowski ）距离：

$$d(x,y)=\Big[\sum_{i=1}^{n}|x_i-y_i|^m\Big]^{\frac{1}{m}},$$

5. 马氏（ Mahalanobis ）距离：

$$d^2(x,y) = (x-y)'\Sigma^{-1}(x-y).$$

其中 $\Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})', \overline{x} = \sum_{i=1}^{n} x_i$.

在实际应用中，较多地使用欧氏距离. 以上定义的各种相似性测度算式属距离测度，两模式越相似，其测度值越小. 在正文部分我们将看到相似性传播算法中用到的相似性测度是非常一般化的，因为它可以不满足以上三条公理中的任何一条.

聚类算法可以看做是通过考虑包含在样本点中的所有可能划分集合的一小部分就可以得到可以判断聚类的方案，这个结果依赖于使用的算法和准则. 因此，聚类算法是一个试图识别数据集合聚类的特征性质的学习过程. 当洪水般的数据量涌入科学与医学实验，研究者们渴望更多更有效的方法来组织与分析数据. 例如，基因是由数以万计的序列表达的，在这种情形下，每个数据类是具有相似表达模式的基因. 2007 年 11 月， *Frey* 和 *Dueck* 给出了一种新的方法 [1], 以找到最优数据类，这种算法探测一些特殊的数据点，我们称之为代表点，并把每个数据点与能最好表达它的代表点联系起来. 原则上，找到一组最优代表点集是很困难的问题，但是这种方法可以迅速有效地处理一些非常大的问题（例如，把 75000 个 DNA 片断归为 2000 类），曾经需要数百个小时计算时间的分析应用这种新方法只需几分钟.

当代表点本身储藏了复合信息时，探索代表点将超越简单的聚类. 一个有广泛应用的例子是语言的统计分析. 例如，我们考虑一篇科学论文，把其中每个句子当成一个数据点. 任何两个句子之间的相似性可由标准信息理论方法计算（也就是说当两个句子含有相同单词的时候，相似性会增加）. 知道了相似性，则可以检测到论文中的代表语句，这些代表语句提供了最优的压缩后的描述. 一个速读者则可以直接阅读这些代表语句以了解文章的大概内容 [2].

这个新算法与通常所用的聚类方法是不同的，这种算法始于一个相似矩阵的构造，它实质是一个数据表，建立了每两个数据点之间的关系. 经过数据点之间信息的相互传递最后选定最优代表点集，最优代表点集是每个数据点与其相应代表点之间的相似测度和最大的数据点集.

相似性传播算法相对于其他算法拥有许多优点，K- 均值算法 [3] 、K- 中心聚类算法 [3] 及 EM 算法 [4] 在每一步都储存较少的聚类中心，如果运气不够好则会得到质量很差的结果. 若初始时选择较多的聚类中心，然后在细化这些聚类中心可以提高这些算法的效率 [5], 但是它们始终依赖随机抽样，并且细化判决准则非常难得到. 相比之下，相似性传播算法可以避免许多低质量的聚类中心，因为此算法同时考虑所有的数据点为潜在的聚类中心 - 代表点. $MCMC$ 技术 [6] 可以随机选择一些好的结果，但却不具备相似性传播算法的其他优点. 分层聚类算法 [7] 与谱聚类算法 [8] 不要求每一类的所有数据点与唯一的一个聚类中心相似，因此，在许多实际问题中，这两种算法都是不适用的.

国内现在有很多人也在做经济区域划分问题 [9][10], 但结果都不是很理想，要么分类太多没有实际意义，要么分类过于粗糙忽视区域内的差异性，给国家经济宏观调控造成困扰.

本文将主要介绍“相似性传播算法”，并介绍这一新算法在中国经济区域划分中的应用与分类结果的经济意义.

§1. 相似性传播算法

§1.1. 算法的直观解释

在使用严谨的数学语言表述相似性传播算法之前，我们可以先通过一种拟人的观点来理解. 假设你是一个数据点，你想找到一个代表点，原则是这个代表点与自己是最相似的，但是你的选择是受限制的：如果你选择其他某个点 A 作为代表点，则 A 必须成为自己的代表点. 这对每一个数据点制造了一个限制条件，进而建立了一个限制网络使得所有的数据点都必须满足这一条件. 当这个相似性网络被最大化并且所有的限制条件均被满足，则可找到最优代表点集.

现在假设每个数据点旁站着一个引导天使，这个天使告诉我们是否有其他数据点已经选择了这个数据点作为代表点，通过相互告诉这些特征，形成了一个包含所有约束条件的复杂网络. 在一个给定时刻，所有的天使发送信息给所有的数据点，同时，所有的数据点反馈信息给所有的天使，一个数据点告诉其他每一个数据点的天使它所喜欢的代表点的顺序列表. 天使告诉其他每一个数据点：在天使们的限制下，它自己成为代表点的能力值. 以所收到的信息与相似矩阵为基础，每一个被传播的信息可以通过简单的计算来估计. 经过几轮信息传递后，所有的特征达到了一个平衡，每一个数据点知道了它的代表点. 在实际应用中，这个算法的运行时间与相似性测度的值呈线性关系.

§1.2. 算法的基本概念

基于相似性测度聚类数据在科学数据分析及工程系统研究中是非常重要一步，与通常方法不同， *Frey* 与 *Dueck* 在 2007 年给出了一种全新的聚类方法 [1], 它通过数据点之间相互传播信息聚类，我们称之为 "相似性传播算法". 它把所有的数据点当作潜在的代表点，即选自实际数据点的聚类中心，并使每两个数据点之间循环相互传递信息，直到产生一组好的代表点及其相应的分类.

(1) **相似性测度** (*similarity*)

为了能分化模式的类别，必须首先定义模式相似性测度，以此来描述各模式之间特征的相似程度. 在相似性传播算法中，数据点之间传递的信息是以相似性测度为基础的. 所谓相似性测度是指数据点 k 与点 i 之间的相似度，用 $s(i,k)$ 表示，它是在应用相似性传播算法时需要输入的实值信息，说明了在不考虑其他任何点的情况下点 k 作为点 i 的代表点的优劣程度. 它是一个有向测度，也就是说 $s(i,k)$ 与 $s(k,i)$ 是可以不相同的，它不满足距离测度的公理 2, 在以后的解释中我们会发现这个相似性测度可以不满足距离测度公理的任何一条，因此说它是非常一般化的，是一种广义意义上的测度. 当我们的目标是最小化均方误差时，每个相似性

测度可以设为负的均方差（负的欧氏距离）：对于点 i 与点 k, $s(i,k) = -\|x_i - x_k\|^2$. 事实上，当最优化准则更一般的情况下，这里所描述的方法也是可以适用的. 当依赖代表点的概率模型已知时， $s(i,k)$ 可以设为：在给定点 k 为代表点时数据点 i 的对数似然，这与它的条件概率是紧密相关的 ($\log p(x_i, x_k) = \log p(x_i \mid x_k) + \log p(x_k)$).

我们说相似性测度是可以不满足测度公理那三条的，它是非常一般化的，是一种广义意义上的测度，这种一般性在实际应用中是十分有效的. 如在识别空中航线 [1] 的例子中， $s(i,k)$ 是城市 i 到城市 k 所需时间的相反数，它不满足距离测度公理. 首先，时间是正数，所以 $s(i,k)$ 是小于等于零的，不满足公理 1；第二，由于地势、风向等客观原因城市 i 到城市 k 所需时间与城市 k 到城市 i 所需时间不一定相同，这不满足公理 2；最后，很明显可以看出 $s(i,k)$ 所定义的相似性测度是不满足三角不等式的，也就是说不满足公理 3.

$s(k,k)$ 表示数据点 k 被选为代表点的可能性，由于 $s(i,k)$ 是需要输入的实值信息，所以我们认为 $s(k,k)$ 所代表的可能性是一种先验知识，与经过信息传递之后判断出的成为代表点的可能性是不同的. $s(k,k)$ 越大，点 k 被选为代表点的可能性越大， $s(k,k)$ 的取值影响最终的分类结果：

(a) 当没有任何先验信息时，我们认为各数据点成为代表点的可能性相同，这种情况下，对于所有数据点 k, $s(k,k)$ 被赋予相同的实数值，其取值越大，分类数就越多.

若 $s(k,k) = median_{i \neq k} s(i,k)$, 则可得到适中的分类数；

若 $s(k,k) = minmum_{i \neq k} s(i,k)$, 则可得到较少的分类数.

(b) 当存在先验信息时——数据点 k_0 一定是代表点，则指定 $s(k_0, k_0) = \infty$, 在应用相似性传播算法之后，数据点 k_0 如所期望地被判断为代表点. 这在探测基因的例子中有实例 [1].

以上分析是非常有道理的，直观上我们可以通过两个极端的例子理解，当 $s(k,k)$ 非常大时，那么 $s(i,k)_{i \neq k}$ 相对较小，数据点 k 被选为代表点的可能性大，它与其他数据点的相似程度小，这时每个数据点倾向于自己独立成类，所以最终分类数较多；而当 $s(k,k)$ 很小时，那么 $s(i,k)_{i \neq k}$ 相对较大，数据点 k 被选为代表点的可能性小，它与其他数据点的相似程度大，这时不同的数据点容易被分为一类，所以最终分类较少. 在实际问题的处理中，我们通常取

$$s(k,k) = median_{i \neq k} s(i,k).$$

(2) **责任性** ($responsibility$)

数据点之间相互传播的信息包括两种，每种信息中包含了不同的竞争机制，在任何一个阶段结合这些信息可以判断出哪些数据点是代表点以及每一个数据点所属于的数据类. 第一种信息被称为“责任性” ($responsibility$), 简写为 r. $r(i,k)$ 是数据点 i 向其备选代表点 k 发出的信

息，它反映了在考虑数据点 i 的其他备选代表点的情况下，点 k 作为 i 的代表点有多好，也就是说与其他备选代表点相比，数据点 k 作为 i 的代表点的适宜度.

(3) **可用性** (*availability*)

数据点之间相互传播的第二种信息被称为“可用性” (*availability*), 简写为 a. $a(i,k)$ 是备选代表点 k 向数据点 i 发出的信息，它反映了在考虑了其他数据点是否选择点 k 作为自己的代表点的情况下，点 i 选择点 k 作为代表点的适宜度，也就是说在考虑了点 k 成为代表点的自身“能力”的基础上，告诉点 i 选择点 k 作为其代表点的适宜度.

§1.3. 算法的基本过程

(1) **迭代过程**

首先，令所有的“可用性” (*availability*) 为零，即 $a(i,k)=0$,

然后，通过证明与计算（参看下一节）我们得到“责任性” (*responsibility*) 的计算公式：

$$r(i,k) \leftarrow s(i,k) - \max_{k's.tk' \neq k} \{a(i,k') + s(i,k')\}.$$

在首次迭代过程中，由于所有的“可用性” (*availability*) 为零， $r(i,k)$ 为点 i 与其代表点 k 之间的相似性测度减去数据点 i 与其他备选代表点 k' 之间相似性测度的最大值，这一竞争性更新没有考虑到备选代表点 k' 作为代表点的自身能力. 在之后的迭代过程中，当数据点被有效地分配给一些代表点之后，根据“可用性” (*availability*) 的更新准则，代表点的“可用性” (*availability*) 将降为负值，这样一来，若部分数据点的“可用性” (*availability*) 太小将不再被视为备选代表点.

“责任性” (*responsibility*) 的更新使得所有备选代表点为了某一个数据点的所有权而竞争，相比之下，“可用性” (*availability*) 则是从各数据点那里搜集信息以视是否每一个备选代表点都会成为一个好的代表点，通过证明与计算（参看下一节）我们得到其计算公式

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i's.t.i' \notin \{i,k\}} \max\{0, r(i',k)\}\}, \quad i \neq k,$$

$$a(i,k) \leftarrow \sum_{i's.ti' \neq k} \max\{0, r(i',k)\}, \quad i = k.$$

我们可以这样理解这个公式，若点 k 作为代表点的能力较强，即有较多数据点以点 k 为代表点，则 $a(i,k)$ 的值较大，但最大也不会大于零，这主要是为了限制“责任性” (*responsibility*) 对“可用性” (*availability*) 的正面影响. 另外，在迭代过程中，为避免数值灾难，如数值摇摆形成死循环或者模糊判断某一数据所属的代表点，我们引入抑制因子 $\lambda(<0\lambda<1)$, 于是我们

得到迭代公式：

$$r^{n+1} \leftarrow \lambda r^{n+1}(i,k) + (1-\lambda) r^{n}(i,k),$$

$$a^{n+1} \leftarrow \lambda a^{n+1}(i,k) + (1-\lambda) a^{n}(i,k).$$

其中，抑制因子 λ 与 $(1-\lambda)$ 可视为权重，当 λ 越大时，新一轮迭代产生的数值对结果的影响就越大.

(2) **判决过程**

以上更新过程只需要简单局部的计算，其信息的传播也只需在已知相似性测度的一组数据点之间传递即可. 在相似性传播过程中的任意时刻，我们都可以结合“责任性”与“可用性”来识别代表点或某一数据点所属的代表点，这就需要引入判决函数 [11].

首先我们引入判决矩阵 E, 使得

$$e(i,k) = r(i,k) + a(i,k).$$

对于给定的数据点 i：

若 $e(i,i) = \max_k\{e(i,k)\}$, 则数据点 i 即为代表点；

若 $\exists k_0 \neq i \quad \text{s.t.} \quad e(i,k_0) = \max_k\{e(i,k)\}$, 则数据点 k_0 为点 i 的代表点.

证明过程参看下一节.

(3) **终止迭代**

对于任何一个程序，无论如何我们都不能让它陷入死循环或者无限期的运行下去，那么如何避免这一问题呢，这就需要事先设定一些终止迭代的准则. 对于相似性传播算法，终止信息传递过程可以有一下三种方法 [12]：

1. 在一定次数的迭代之后终止信息的传递，即预先指定迭代次数；
2. 当信息的改变低于某一阈值时终止信息的传递，即预先设定阈值；
3. 当局部判决函数在几次迭代过程中均为常数时，停止信息的传递.

§1.4. 算法的理论证明

(1) **责任性与可用性公式的证明**

$$r(i,k) = s(i,k) - \max_{k' s.t k' \neq k} \{a(i,k') + s(i,k')\},$$

$$a(i,k)=\begin{cases}\sum_{i':i'\neq k}\max(0,r(i',k)), & k=i,\\ \min\left[0,r(k,k)+\sum_{i':i'\notin\{i,k\}}\max(0,r(i',k))\right], & k\neq i.\end{cases}$$

证明：设样本为 $\{x_1,x_2,\cdots,x_N\}$，其相应的代表点下标为 $C=\{c_1,c_2,\cdots,c_N\}$. 很显然，若 x_i 以 x_k 为代表点则 x_k 必须以自己为代表点. 取目标函数 [13]

$$S(C)=\sum_{i=1}^{N}S(i,c_i)+\sum_{k=1}^{N}\delta_k(C).$$

其中

$$\delta_k=\begin{cases}-\infty, & \text{若 } c_k\neq k, \text{但 } \exists i s.t. c_i=k,\\ 0, & \text{其他情形.}\end{cases}$$

能够使得 $S(C)$ 最大化的那一组代表点下标 $C=\{c_1,c_2,\cdots,c_N\}$ 是解. 所以我们只需解：

$$\max_{c_1,\cdots,c_N}\Big\{\sum_i S(i,c_i)+\sum_k\delta_k(C)\Big\}.$$

我们把所有的 c_i 设为变量节点， $\delta_k(C)$ 为 c_i 的函数，设它们为因子节点，两种节点之间通过“桥”连接，我们称这样的图为因子图 [14]. 在因子图中，桥上同时传递两种信息：

1. 由变量节点 c_i 传向因子节点 $\delta_k(C)$ 传递的信息 $\rho_{i\to k}(c_i)$ ：

$$\rho_{i\to k}(c_i)=(\rho_{i\to k}(1,\cdots,\rho_{i\to k}(j),\cdots,\rho_{i\to k}(N)),$$

2. 由因子节点 $\delta_k(C)$ 传向变量节点 c_i 传递的信息 $\alpha_{i\leftarrow k}(c_i)$ ；

$$\alpha_{i\leftarrow k}(c_i)=(\alpha_{i\leftarrow k}(1),\cdots,\alpha_{i\leftarrow k}(j),\cdots,\alpha_{i\leftarrow k}(N)).$$

根据因子图写出 ρ,α 的表达式：

$$\rho_{i\to k}(c_i)=S(i,c_i)+\sum_{k'\neq k}\alpha_{i\leftarrow k'}(c_i),\tag{a1}$$

$$\alpha_{i\leftarrow k}(c_i)=\max_{-c_i}\Big\{\delta_k(c)+\sum_{i'\neq i}\rho_{i'\to k}(c_{i'})\Big\}.\tag{b1}$$

所以

$$\alpha_{i\leftarrow k}(c_i)=\overbrace{\max_{j_1,\cdots,j_{i-1},j_{i+1},\cdots,j_N}\Big[\delta_k(j_1,\cdots,j_{i-1},c_i,j_{i+1},\cdots,j_N)+\sum_{i'}\rho_{i'\to k}(j_{i'})\Big]}.$$

为消除 max[.] 中 $\delta_k(.)$ 的影响，以下我们分四种情况讨论：

$$= \begin{cases} \sum_{i':i'\neq k} \max_{j'} \rho_{i'\to k}(j'), & c_i = k = i, \\ \sum_{i':i'\neq k} \max_{j':j'\neq k} \rho_{i'\to k}(j'), & c_i \neq k = i, \\ \rho_{k\to k}(k) + \sum_{i':i'\notin\{i,k\}} \max_{j'} \rho_{i'\to k}(j'), & c_i = k \neq i, \\ \max\left[\max_{j':j'\neq k} \rho_{k\to k}(j') + \sum_{i':i'\notin\{i,k\}} \max_{j':j'\neq k} \rho_{i'\to k}(j'), \right. & \\ \qquad \left. \rho_{k\to k}(k) + \sum_{i':i'\notin\{i,k\}} \max_{j'} \rho_{i'\to k}(j') \right], & c_i \neq k \neq i. \end{cases} \tag{$b1'$}$$

因为传递的信息中包含无用信息，所以对其提取或删除，这些向量信息可以看成是一个常量与一个变量成分（与 c_i 有关）的和，即

$$\rho_{i\to k}(c_i) = \tilde{\rho}_{i\to k}(c_i) + \bar{\rho}_{i\to k},$$

$$\alpha_{i\leftarrow k}(c_i) = \tilde{\alpha}_{i\leftarrow k}(c_i) + \bar{\alpha}_{i\leftarrow k}.$$

于是，我们得到方程：

$$\rho_{i\to k}(c_i) = S(i,c_i) + \sum_{k'\neq k} \tilde{\alpha}_{i\leftarrow k'}(c_i) + \sum_{k'\neq k} \bar{\alpha}_{i\leftarrow k'}(c_i), \tag{$a2$}$$

$$\alpha_{i\leftarrow k}(c_i) = \begin{cases} \sum_{i':i'\neq k} \max_{j'} \tilde{\rho}_{i'\to k}(j') + \sum_{i':i'\neq k} \bar{\rho}_{i'\to k}, & c_i = k = i, \\ \sum_{i':i'\neq k} \max_{j':j'\neq k} \tilde{\rho}_{i'\to k}(j') + \sum_{i':i'\neq k} \bar{\rho}_{i'\to k}, & c_i \neq k = i, \\ \tilde{\rho}_{k\to k}(k) + \sum_{i':i'\notin\{i,k\}} \max_{j'} \tilde{\rho}_{i'\to k}(j') + \sum_{i':i'\neq i} \bar{\rho}_{i'\to k}, & c_i = k \neq i, \\ \max\left[\max_{j'\neq k} \tilde{\rho}_{k\to k}(j') + \sum_{i'\notin\{i,k\}} \max_{j'\neq k} \tilde{\rho}_{i'\to k}(j') + \sum_{i'\neq i} \bar{\rho}_{i'\to k}, \right. & \\ \qquad \left. \tilde{\rho}_{k\to k}(k) + \sum_{i':i'\notin\{i,k\}} \max_{j'} \tilde{\rho}_{i'\to k}(j') + \sum_{i':i'\neq i} \bar{\rho}_{i'\to k} \right], & c_i \neq k \neq i. \end{cases} \tag{$b2$}$$

对于 $\rho_{i\to k}(c_i)$, 如果令

$$\bar{\rho}_{i\to k} = \max_{j:j\neq k} \rho_{i\to k}(j),$$

则

$$\max_{j':j'\neq k} \tilde{\rho}_{i\to k} = 0,$$

$$\max_{j'} \tilde{\rho}_{i\to k}(j') = \max(0, \tilde{\rho}_{i\to k}(k)).$$

对于 $\alpha_{i\leftarrow k}(c_i)$, 如果令

$$\bar{\alpha}_{i\leftarrow k} = \alpha_{i\leftarrow k}(c_i : c_i \neq k),$$

则

$$\tilde{\alpha}_{i\leftarrow k}(c_i) = 0 \Rightarrow \begin{cases} \sum_{k':k'\neq k} \tilde{\alpha}_{i\leftarrow k'}(c_i) = \alpha_{i\leftarrow c_i}(c_i), & c_i \neq k, \\ \sum_{k':k'\neq k} \tilde{\alpha}_{i\leftarrow k'}(c_i) = 0, & c_i = k. \end{cases}$$

于是简化的方程:

$$\rho_{i\to k}(c_i) = \begin{cases} s(i,k) + \sum_{k':k'\neq k} \bar{\alpha}_{i\leftarrow k'}, & c_i = k, \\ s(i,c_i) + \tilde{\alpha}_{i\leftarrow c_i}(c_i) + \sum_{k':k'\neq k} \bar{\alpha}_{i\leftarrow k'}, & c_i \neq k, \end{cases} \tag{a3}$$

$$\alpha_{i\leftarrow k}(c_i) = \begin{cases} \sum_{i':i'\neq k} \max(0, \tilde{\rho}_{i'\to k}(k)) + \sum_{i':i'\neq k} \bar{\rho}_{i'\to k}, & c_i = k = i, \\ \sum_{i':i'\neq k} \bar{\rho}_{i'\to k}, & c_i \neq k = i, \\ \tilde{\rho}_{k\to k}(k) + \sum_{i':i'\neq k} \max(0, \tilde{\rho}_{i'\to k}(k)) + \sum_{i':i'\neq i} \bar{\rho}_{i'\to k}, & c_i = k \neq i, \\ \max\left[0, \tilde{\rho}_{k\to k}(k) + \sum_{i':i'\notin\{i,k\}} \max(0, \tilde{\rho}_{i'\to k}(k))\right] + \sum_{i':i'\neq i} \bar{\rho}_{i'\to k}, & c_i \neq k \neq i. \end{cases} \tag{b3}$$

因为

$$\tilde{\rho}_{i\to k}(c_i=k)=\rho_{i\to k}(c_i=k)-\bar{\rho}_{i\to k},$$

$$\tilde{\alpha}_{i\leftarrow k}(c_i=k)=\alpha_{i\leftarrow k}(c_i=k)-\bar{\alpha}_{i\leftarrow k}.$$

所以得到更新方程：

$$\begin{aligned}\tilde{\rho}_{i\to k}(c_i=k) &= \rho_{i\to k}(c_i=k)-\bar{\rho}_{i\to k}=\rho_{i\to k}(k)-\max_{j:j\neq k}\rho_{i\to k}(j)\\ &= s(i,k)+\sum_{k':k'\neq k}\bar{\alpha}_{i\leftarrow k'}-\max_{j:j\neq k}\Big[s(i,j)+\tilde{\alpha}_{i\leftarrow j}(j)+\sum_{k':k'\neq k}\bar{\alpha}_{i\leftarrow k'}\Big],\end{aligned}$$

$$\begin{aligned}\tilde{\alpha}_{i\leftarrow k}(c_i=k) &= \alpha_{i\leftarrow k}(c_i=k)-\bar{\alpha}_{i\leftarrow k}=\alpha_{i\leftarrow k}(k)-\alpha_{i\leftarrow k}(j:j\neq k)\\ &= \begin{cases}\sum_{i':i'\neq k}\max(0,\tilde{\rho}_{i'\to k}(k))+\sum_{i':i'\neq k}\bar{\rho}_{i'\to k}-\sum_{i':i'\neq k}\bar{\rho}_{i'\to k}, & k=i,\\ \tilde{\rho}_{k\to k}(k)+\sum_{i':i'\neq k}\max(0,\tilde{\rho}_{i'\to k}(k))+\sum_{i':i'\neq i}\bar{\rho}_{i'\to k} & \\ \quad-\max\Big[0,\tilde{\rho}_{k\to k}(k)+\sum_{i':i'\notin\{i,k\}}\max(0,\tilde{\rho}_{i'\to k}(k))\Big]-\sum_{i':i'\neq i}\bar{\rho}_{i'\to k}, & k\neq i.\end{cases}\end{aligned}$$

因为 $\tilde{\alpha}_{i\leftarrow k}(c_i\neq k)=0$, 所以 $\tilde{\rho}_{i\to k}(c_i\neq k)$ 和 $\tilde{\alpha}_{i\leftarrow k}(c_i\neq k)$ 在更新信息的过程中并不被使用，因此可以考虑所传递的信息与责任性 (*responsibility*) 和可用性 (*availability*) 成比例，其中责任性 (*responsibility*) 定义为 $r(i,k)$:

$$r(i,k)=\tilde{\rho}_{i\to k}(c_i=k)=s(i,k)-\max_{j:j\neq k}[s(i,j)+a(i,j)],$$

可用性 (*availability*) 定义为 $a(i,k)$:

$$a(i,k)=\tilde{\alpha}_{i\leftarrow k}(c_i=k)=\begin{cases}\sum_{i':i'\neq k}\max(0,r(i',k)), & k=i,\\ \min\Big[0,r(k,k)+\sum_{i':i'\notin\{i,k\}}\max(0,r(i',k))\Big], & k\neq i.\end{cases}$$

所以公式得证.

(2) **判决公式的证明**

相似性传播算法的目标就是估计一组 c_i, 使得目标函数最大化. 为了在任意一步迭代过程中估计 c_i, 我们把所有流入 c_i 的信息加和，取 $\hat{c}_i$ 为解：

$$\hat{c}_i=\arg\max_j\Big[a(i,j)+s(i,j)\Big].$$

证明：对 $c_i, i=1,2,\cdots,N$ 进行估计，取目标为

$$\begin{aligned}
&\max_{c_1,c_2,\cdots,c_N}\left[\sum_{i=1}^{N}S(i,c_i)+\sum_{k=1}^{N}\delta_k(C)\right]\\
=&\max_{c_1,c_2,\cdots,c_N}\left[\sum_{i=1}^{N}S(i,c_i)+\sum_{k=1}^{N}\sum_{i=1}^{N}\alpha_{i\leftarrow k}(c_i)\right]\\
=&\max_{c_1,c_2,\cdots,c_{N-1}}\left[\max_{c_N}\left\{\sum_{i=1}^{N}S(i,c_i)+\sum_{k=1}^{N}\sum_{i=1}^{N}\alpha_{i\leftarrow k}(c_i)\right\}\right]\\
=&\max_{c_1,c_2,\cdots,c_{N-1}}\left[\sum_{i=1}^{N-1}S(i,c_i)+\sum_{k=1}^{N}\sum_{i=1}^{N-1}\alpha_{i\leftarrow k}(c_i)+\max_{c_N}\left\{S(N,c_N)+\sum_{k=1}^{N}\alpha_{N\leftarrow k}(c_N)\right\}\right],
\end{aligned}$$

所以 $\hat{c}_N=\arg\max_j\left[S(N,j)+\sum_{k=1}^{N}\alpha_{N\leftarrow k}(j)\right]$，

因为 $\bar{\alpha}_{i\leftarrow k}$ 为常数，且

$$\tilde{\alpha}_{i\leftarrow k}(j)=\begin{cases}a(i,j), & j=k,\\ 0, & j\neq k.\end{cases}$$

所以，对 $\forall i, i=1,2,\cdots,N$ 有

$$\begin{aligned}
\hat{c}_i&=\arg\max_j\left[S(i,j)+\sum_{k=1}^{N}\alpha_{i\leftarrow k}(j)\right]\\
&=\arg\max_j\left[S(i,j)+\sum_{k=1}^{N}\tilde{\alpha}_{i\leftarrow k}(j)+\sum_{k=1}^{N}\bar{\alpha}_{i\leftarrow k}\right]\\
&=\arg\max_j\left[S(i,j)+a(i,j)\right].
\end{aligned}$$

为了体现 $\hat{c}_i$ 与 $f(i,j), a(i,j)$ 的关系，我们选择另外一种表达形式：

$$\begin{aligned}\hat{c}_i &= \arg\max_j \left[a(i,j) + S(i,j) - \max_{j':j'\neq j}\left\{s(i,j') + a(i,j')\right\}\right] \\ &= \arg\max_j \left[a(i,j) + r(i,j)\right] \\ &= \arg\max_j \left[e(i,j)\right].\end{aligned}$$

证明：虽然 $\hat{c}_i$ 的表达形式改变，但是它的估计值不会变.

$\max_{j':j'\neq j}[s(i,j') + a(i,j')]$ 仅有两种取值： 1. 最大值， 2. 次大值.

1. 若

$$\hat{c}_i = k_0 = \arg\max_j [a(i,j) + s(i,j)],$$

则

$$\max_{j':j'\neq k_0} [s(i,j') + a(i,j')] < a(i,k_0) + s(i,k_0) \Rightarrow a(i,k_0) + r(i,k_0) > 0.$$

2. 若

$$\hat{c}_i = k_1 \neq \arg\max_j [a(i,j) + s(i,j)],$$

则

$$\max_{j':j'\neq k_1} [s(i,j') + a(i,j')] > a(i,k_1) + s(i,k_1) \Rightarrow a(i,k_1) + r(i,k_1) < 0.$$

由 1 、 2 可知，做变换后 $\hat{c}_i$ 的估计不会改变，且，若 $e(i,i) > 0$, 则样本点 x_i 为代表点，否则不是代表点.

§2. 相似性传播算法与 K- 均值算法比较

§2.1. K- 均值算法

(1) **算法的条件及约定**

设定分类的模式特征矢量集为 $\{x_1, x_2, \cdots, x_N\}$, 类的数目 K 是事先取定的.

(2) **算法的基本思想**

该方法取定 K 类和选取 K 个初始聚类中心，按最小距离原则将各模式分配到 K 类中的某一类，之后不断的计算和调整各模式的类别，最终使各模式到其判属类别中的距离平方和最小 [15].

(3) **算法的步骤**

1. 任选 K 个模式特征矢量作为初始聚类中心： $z_1^{(0)}, z_2^{(0)}, \cdots, z_K^{(0)}$, 令 $c = 0$.

2. 将待分类的模式特征矢量集 $\{x_i\}$ 中的模式逐个按最小距离原则分划给 K 类中的某一类，即：

如果

$$d_{il}^{(c)} = \min_j [d_{ij}^{(c)}], i = 1, 2, \cdots, N,$$

则判

$$x_i \in \omega_l^{(c+1)}.$$

式中 $d_{ij}^{(c)}$ 表示 x_i 和 $\omega_j^{(c)}$ 的中心 $z_j^{(c)}$ 的距离，上角标表示迭代次数. 于是产生新的聚类 $\omega_j^{(c+1)}$ $(j = 1, 2, \cdots, K)$.

3. 计算重新分类后的各类心

$$z_j^{(c+1)} = \frac{1}{n_j^{(c+1)}} \sum_{x_i \in \omega_j^{(c+1)}} x_i \ , \ j = 1, 2, \cdots, K,$$

式中 $n_j^{(c+1)}$ 为 $\omega_j^{(c+1)}$ 类中所含模式的个数.

4. 如果 $z_j^{(c+1)} = z_j^{(c)} (j = 1, 2, \cdots, K)$, 则结束；否则， $c = c + 1$, 转至 2.

§2.2 两种算法的比较

(1) **分析 K- 均值算法**

K- 均值算法是以确定的类数及选定的初始聚类中心为前提，是各模式到其判属类别中心距离（平方）之和最小的最佳聚类. 显然，此算法的分类结果受到取定的类别的数目及聚类中

心的初始位置的影响，同时还与样本的几何分布有关，所以结果只是局部最优的. 但其方法简单，结果尚令人满意，故应用较多. 在实际应用中，需要试探不同的 K 值和选择不同的聚类中心初始值，以进一步达到更大范围的最优结果. 此算法存在一个明显的不足，即只用一个聚类中心点作为一类的代表，一个点往往不能充分地反映该类的模式分布结构，从而损失很多有用的信息. 只有当聚类数非常小并且至少有一组初始代表点非常接近一个好的解的时候，这种方法才是有效的 [16].

(2) **分析相似性传播算法**

基于相似性测度聚类数据在科学数据分析与工程系统中是非常关键的一步. 通常的方法是找到一组聚类中心使得数据点与它最近的中心点的平方误差和最小. 当中心点选自实际数据点时，被称为“代表点”，这篇文章里我们研究的相似性传播算法是在工作网络的桥上循环传递信息直到找到一组好的代表点以及与其相应的聚类，这种算法具备以下优点：

1. 代表点的个数不需要事先指定，相反，适宜代表点的个数有信息传递方法逐渐确定，它依赖于输入的代表点的优先权 (*exemplar preference*). 相比而言，同时考虑所有的数据点为潜在聚类中心并逐渐识别类别的相似性传播算法可以避免许多低质量的结果，这些低质量的结果通常是由初始值与复杂的判别准则引起的. 相似性传播算法是一种自动模型选择，它以每个数据点成为代表点的可能性即代表点的优先权为基础，这种可能性是一种先验信息，即之前讲过的 $s(k,k)$.

2. 在计算时间相同的情况下，运行 K- 均值算法 100 次中最好的一次与相似性传播算法相比，后者的均方误差比前者低很多，即使运行 K- 均值算法无穷多次，相似性传播算法也会在较短的时间内一致地得到均方误差较小的结果.

3. 相似性传播算法的运行可以以非标准化的最优准则为基础，这使得此算法适用于使用非标准的相似性测度的探索数据分析，不像正则空间的聚类技术，如 K- 均值算法，相似性传播算法可以适用于非连续空间数据点的聚类. 事实上，这种算法可以适用于相似性测度非对称，即 $s(i,k) \neq s(k,i)$ 的问题或者相似性测度不满足三角不等式的问题，即 $s(i,k) \leq s(i,j)+s(j,k)$.

§2.3 K- 均值算法在经济分区问题中的应用

聚类算法有很多种，其中较为简单实用的称之为 K- 均值算法，本节我们应用 K- 均值算法处理经济分区问题. 2007 年研究者们搜集了全国三十一个省市近十年的人口、GDP 和人均 GDP 数据，以这十年数据的平均值作为每个模式（省市）的特征值，应用 K- 均值算法对其分类. 在整个分类过程中我们主要对人均 GDP 进行研究.

(1) **确定分类数**

K- 均值算法是以确定的类数及选定的初始聚类中心为前提，在类别数目未知的情况下运用 K- 均值算法，可让类数 K 从较小逐渐增加，在这个过程中，对每一个选定的 K 分别使用此算法. 显然，准则函数 J 是随 K 的增加而单调减小. 以人均 GDP 为例，作一条 $J-K$ 曲线可以直观地看出这一结论，如下图.

在 K 增加的过程中，总会出现使本来比较密集的一些模式点在被分划开的情况，此时 J 虽减小，但减小速度将变缓. 如果作一条 $J-K$ 曲线，当其曲率变化趋于平缓时对应的类数是比较接近从模式几何分布上看最优的类数. 然而，在多数情况下曲线中并无这样明显的点，但幸运的是，对于我们讨论的这个问题，这样的点是十分明显的.

由下图，可明显看出 $K=3$ 为人均 GDP 的最佳分类数.

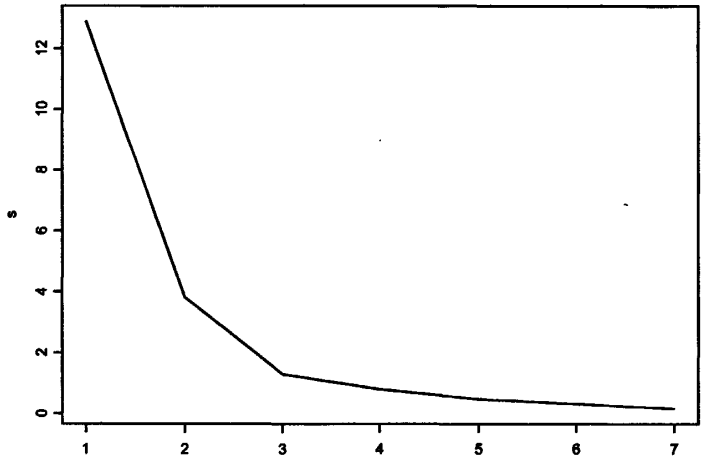


图 2.1

(2)K- **均值算法的结果**

第一类：北京　上海

第二类：天津　辽宁　江苏　浙江　福建　山东　广东

第三类：河北　山西　内蒙古　吉林　黑龙江　安徽　江西　河南　湖北　湖南　广西　海南　重庆　四川　贵州　云南　西藏　陕西　甘肃　青海　宁夏　新疆

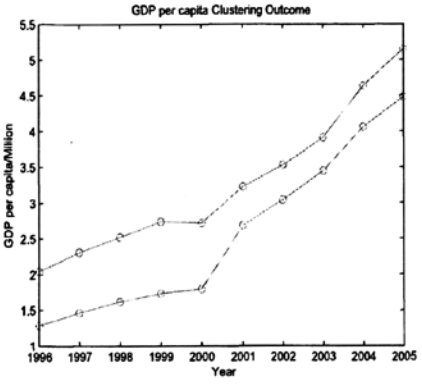


图 2.2A: 第一类

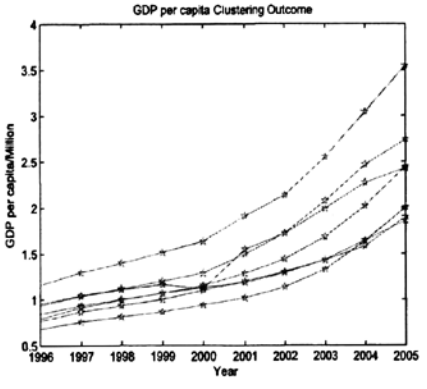


图 2.2B: 第二类

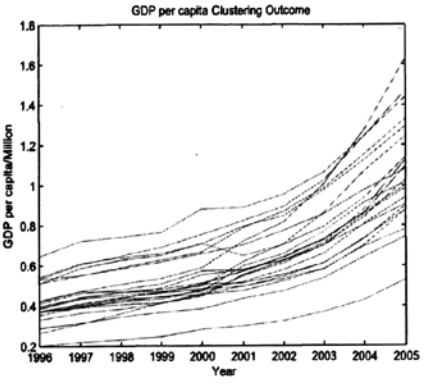


图 2.2C: 第三类

§3. 相似性传播算法在经济分区问题中的应用

§3.1. 提出问题

经济区域的划分关系到区域经济发展布局以及区域的未来发展能力和走势，对国家指定区域经济政策具有重要影响. 我国地大人多，内部区域差距十分突出. 然而我国以往的经济区域的划分往往带有一定的主观性和政策依赖性，而不是完全意义上的区域经济性质的研究. 以中国如此广袤的国土，如果经济区域划分过粗，就会忽视区域内的差异性，实施真正的区域倾斜政策或区域产业政策，将可能成为一句空话.

从建国以来对于区域经济发展的管理和调控过程来看，改革开放前更多地侧重于国防和军事考虑，注重公平而忽视对于效率的追求，依赖的是计划经济和财政手段. 从 " 六五 " 至 " 十五 ", 开始强调全国范围内的区域分工，逐步发挥市场机制的资源配置作用，但是分工的地域单元要么是沿海地区与内陆地区，要么是东部、中部与西部. 沿海地区与内陆地区，或东部、中部与西部这种区域划分，作为分析地区间的发展差距或确定国家的区域发展优先顺序是有重要意义的，但作为区域分工的区划方案，显然是不合适的. 因为东部、中部、西部的各省区市间经济发展的基础条件相差很大，硬要将东部、中部或西部作为一个统一体来考虑承担某些分工，实践中是难以操作的. 长期以来不断变化的区域划分和管理，以及经济区划的不合理和不科学性，使得我国的区域政策难以有效实施，也是难以形成合理的区域经济体系的重要原因.

为此国务院发展研究中心 2004 年 6 月发表报告指出，中国所沿袭的东、中、西的区域划分方法已经不合时宜，拟提出 "十一五" 期间将内地划分为东部、中部、西部、东北部四大板块，并将四个板块划分为八大综合经济区的具体构想. 本文针对这一思路，合理选取反应区域经济社会发展程度的指标——人均 -GDP, 本文单纯考虑各个区域的这一经济存量，试图对我国经济区域做出合理、科学的划分. 我国大陆地区，不包括台湾、香港、澳门三地，共计 31 个行政区域，其中四个直辖市， 27 个省和自治区，本文搜集整理 1996-2005 共计 10 年 31 个省市自治区的人均 -GDP 数据 (见附录 II).

由于单纯从经济意义角度出发，往往忽略了地理位置因素以及自然条件因素，因此，在实际划分类别的时候要加入这些因素以助于分类的合理和准确. 以下是应用相似性传播算法处理 10 年 31 个省市自治区的人均 -GDP 数据，得出的我国经济区域的划分结果.

§3.2 聚类结果

之前在介绍相似性传播算法时，我们有结论： $s(k,k)$ 的不同取值会直接影响分类结果. 我们将通过经济分区这个实际问题来验证这个结论. 根据相似性传播算法的理论知识，它在处理实际问题时通常使 $s(k,k)$ 取中位数，以下详细分析当 $s(k,k) = median_{i \neq k} s(i,k)$ 时的分类结果，并讨论此结果的合理性，其他不同的取值所产生的分类结果（如均值，最小值，最大值）请参看附录 I（分类结果中括号内的为聚类中心 — 代表点）：

相似矩阵的对角线位置取值为中位数， $s(k,k) = median_{i \neq k} s(i,k)$ ：

应用相似性传播算法时，取 $s(k,k) = median_{i \neq k} s(i,k)$，三十一个省市被分为七个经济区域，结果如下：

第一类（北京）：北京

第二类（上海）：上海

第三类（天津）：天津

第四类（广东）：江苏　浙江　广东

第五类（辽宁）：辽宁　福建　山东

第六类（吉林）：河北　山西　内蒙古　吉林　黑龙江　湖北　海南　新疆

第七类（四川）：安徽　江西　河南　湖南　广西　重庆　四川　贵州　云南

西藏　陕西　甘肃　青海　宁夏

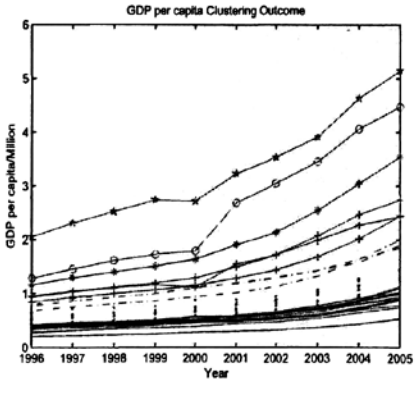


图 3.1：分类结果

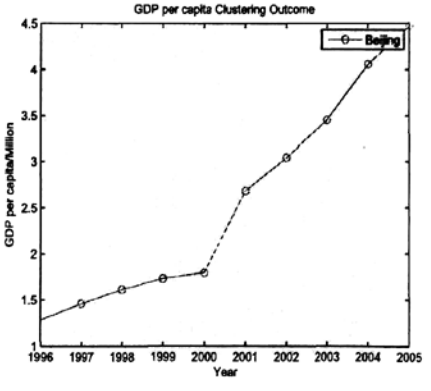


图 3.1A：北京类

图 3.1B：上海类

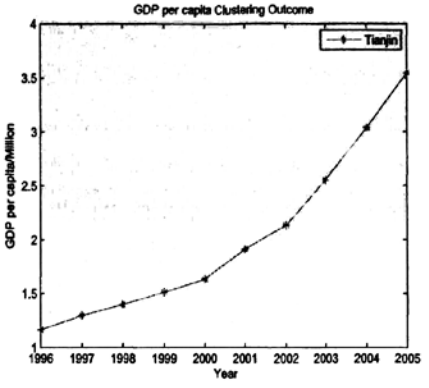


图 3.1C：天津类

图 3.1D：广东类

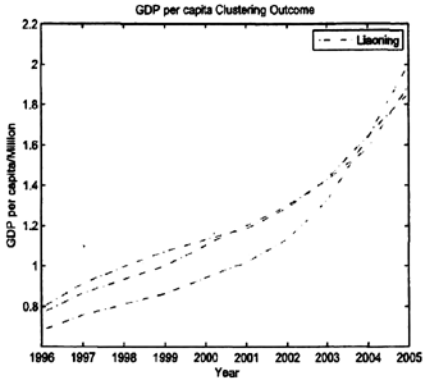


图 3.1E：辽宁类

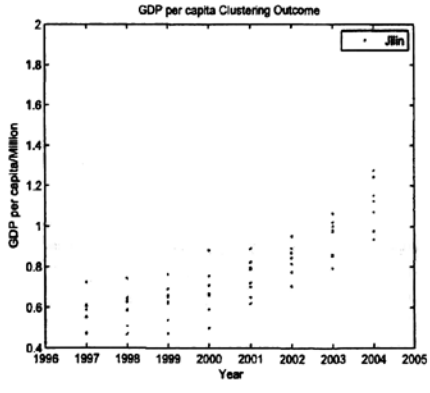


图 3.1F：吉林类

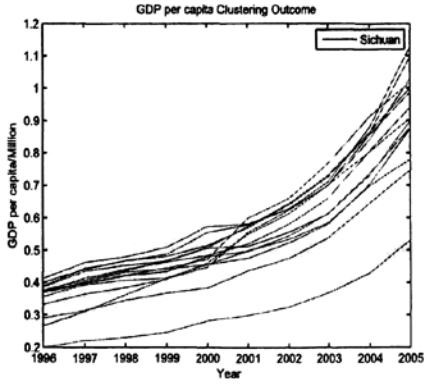


图 3.1G：四川类

§3.3 经济意义分析

1. 第一类：北京是我国的政治、文化与国际交往中心，是综合性产业城市. 北京是中国重要的金融中心和商业中心之一，国家金融宏观调控制部门、中国人民银行、银监会、证监会、保监会均在北京. 包括中国工商银行、中国建设银行、中国银行、中国农业银行四大国有商业银行在内的中国主要商业银行、国家开发银行、中国农业发展银行等政策性银行. 中国人寿、泰康人寿等全国性保险公司总部均设在北京，北京同时还聚集了大部分国有大型企业总部，其中包括中国石化、中国石油、国家电力、中国电信、中国移动通信、中国联通等企业. 大量境外跨国公司在北京建立地区总部. 作为首都，北京占有优越的地理优势和政策优势，经济发展必然快于其他地区.

2. 第二类：上海是中国大陆经济最发达的城市之一. 上海工业发达，国民党政府统治时期和计划经济时期，上海的工业发展水平就已全国领先. 改革开放初期，由于中央政策原因，中国东南地区飞速发展，一度使上海的工业面临边缘化的危机，但自 90 年来中期以来，随着浦东新区的开发，以及财政转移支付比重减少等多方面的原因，上海的工业又重新焕发了新的生机. 上海工业总产值占全国的十分之一，主要以轻纺、重工业、冶金、石油化工、机械、电子工业为主，其他还有汽车、航空、航天等工业. 张江高科汇集了大量的高端制造业. 农业占总体经济的比例较小，大约在 1.7

3. 第三类：天津地理区位优势明显，地处中国北方黄金海岸的中部，不仅毗邻首都，还是华北、西北广大地区的出海口，是中国北方对内外开放两个扇面的轴心，是亚欧大陆桥中国境内距离最短的东部起点. 天津港是中国北方最大的综合性贸易港口. 天津铁路枢纽是京山、京沪两大铁路干线的交汇处. 天津公路四通八达，交通基础设施建设有了长足发展. 目前，天津已形成以港口为中心的海陆空相结合、立体式的综合性的现代化的运输网络.

4. 第四类：江苏、浙江、广东这三个地区在 90 年代后一直发展较为快速，现实中，三地都属于沿海地区. 广东是最早的沿海开放地区，发展速度相当快，在改革开放初期，一直得益于国家的政策，因此珠三角成了劳动密集型企业的大工厂，再加上香港向广东的输血，广东一时间成了世界工厂. 江苏也最早发展成为制造中心，改革开放前期主要靠集体经济，发展了大量的乡镇企业， 90 年代后期以来，主要靠外资. 其中台商占较大一部分，昆山的发展主要依赖于台商的贡献. 浙江的民营企业迅速发展，作为中国民营经济的领跑者，浙江与全国第一批个体工商户、第一批私营企业、第一批专业市场、第一座中国农民城、第一个股份合作制企业等诸多“第一”联系在一起，是国民营经济的领跑者. 三地拥有相似的自然条件并且地理位置接近，划归为一类相当合理，称为沿海开放的发达地区.

5. 第五类：辽宁、福建、山东三个省都有发展十分突出的港口城市，辽宁有大连、沈阳，

福建有厦门、福州，山东有青岛、济南．辽宁地区是发达老工业基地，属于国有企业改制并未彻底进行，市场经济发展速度较为不稳定的区域．虽然它属于东北老工业基地，与黑龙江、吉林发展状态相似，但是辽宁具有先天的自然优势，同时环渤海湾，交通运输较为便利，省内的产业结构调整较快，经济得到快速发展．辽宁在大力发展沿海城市的同时，以良好的工业基础借助“振兴东北老工业基地”的契机，招商引资、广纳人才，使辽宁省经济出现了蓬勃发展的良好态势．福建有厦门这座人杰地灵的好城市，又与台湾隔海相望，加上福州的经济发展使得福建经济处于全国的中上水平．山东作为中国沿海经济大省，近年来经济发展势头强劲，它大力发展沿海城市并以此来带动本省其他城市的发展，综合经济实力不断攀升，使山东经济水平成为少数能与南部沿海城市匹敌的省份．这三个省份自然条件相似，划归为一类十分合理．

6. 第六类：我国经济发展中不发达或者落后的经济区域，其中包括东北地区的黑龙江和吉林，这两个省份的地域相近，产业结构也大致相同，经济发展水平和所处状态也趋于一致，称为东北老工业基地．河北又与毗邻北京，受集聚效应的影响，其发展程度受到了限制．山西省是我国能源重化工业基地，在长期的经济发展中，取得了收益但也付出了沉重的资源和环境的代价．内蒙古跨越我国的东北、华北、西北地区，该地区的经济发展具备了这些地区的特点，并且独具特色的畜牧业促使内蒙古大力发展畜牧业即畜牧副产品加工业，成立了如蒙牛集团等享誉国内外的知名企业．新疆虽是西北地区但是它的人民生活水平、商业发展水平相对西北其他地区都较高．海南虽是沿海城市，但是由于开发较晚，与内陆交通往来不便，并且仅以旅游业为支柱产业，因此发展相对其他是沿海城市较为落后．湖北处于中部地区，但经济体制相对滞后，经济总体实力不强．这八个省份各有优势但又各有弊端，致使其经济发展受到限制．因此单从经济角度出发，这八个省份归为一类是十分合理的．

7. 第七类：安徽、云南、贵州经济大部分以农业为主，省内城市之间经济发展不平衡，使得它们的经济水平也较为底下．宁夏、青海、甘肃、四川、重庆同属于西部地区，经济发展较为缓慢，而湖南虽地处中部地区，但经济体制相对滞后，总体上经济实力还不够强．广西、江西都是经济相对落后的省份，但在“十一五”规划中提出了富裕广西、文化广西、生态广西和平安广西的建设计划，使得广西加快了经济发展的脚步，江西也是在扶持革命老区的政策下，大力发展了本地特色的旅游业．陕西属西北地区，但是它的人民生活水平、商业发展水平相对西北其他地区都较高．西藏的地理位置特殊，交通运输极为不便，为西藏与国内其他地区的交流带来巨大困难，形成了现在西藏经济较为落后的局面．由于在地理位置的负面限制下，这些省市的经济水平处于全国的最底层．

通过以上研究分析，可以看出相似性传播算法在处理我国经济区域划分问题上还是很有效的．把我国三十一个省市自治区划分为 7 个经济区也算是符合实际情况的．虽然有些被划分

为同类经济区的地区在地理位置上相距其远，但是它们在经济发展上的确有相似之处. 通过这一分类结果，我们可以发现，沿海发达地区和沿海开放地区属于最为发达且目前经济水平最高的地区，其他沿海地区和发达工业基地的发展水平处于中等偏上的状态，而中心欠发达地区和东南沿海地区以及老工业基地处于中等偏下的发展状态之中，西部落后地区和西部开发地区处于落后状态.

参考文献

[1] Frey B J,and Dueck D. Clustering by Passing Messages between Data Points[J].*Science* ,2007,**315**:972-976.

[2] Frey B J,and Dueck D. Supporting material[J/OL]. http://www.psi.toronto.edu/affinitypropagation/FreyDueckScience07.pdf, 2007-01-11.

[3] Macqueen,Cam L L,Neyman J. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability[M]. *Univ. of California Press, Berkeley*,1967,**1**:281-297.

[4] Dempster A P,Laird N M, Rubin D B.Maximum Likelihood for incomplete data Via the EM Algrithm[J].*Journal of the Royal Statistical Society B*,1997,**39**:1-38.

[5] Dasgupta S,Schulman L J. A Two-Round Variant of EM for Gaussian Mixtures[J]. *UAI* 2000, 152-159.

[6] Jain S,Neal R M.A split-merge Markov chain Monte Carlo procedurefor the Dirichlet process mixture model[J].*Journal of Computational and GraphicalStatistics*,2004,(13):158-182.

[7] Sokal R R,Michener C D.A statistical method for evaluating systematic relationships[J]. *Univ. Kans. Sci. Bull*,1958,**38**:1409-1438.

[8] Shi J, Malik J.Normalized cuts and image segmentation[J].*IEEE Trans. Pattern Anal. Mach, Intell*,2000,**22(8)**:888-905.

[9] 刘芳，刘日昊，我国经济区域划分及区域发展状态的实证分析 [D].长白学刊,2007,**3**:84-86.

[10] 王佼佼，对我国经济区域划分的探讨 [D],长白学刊,2008,**2**:20-21.

[11] M. Mézard.Passing Messages Between Disciplines[J].*Science*,2003,**301**:1685-1686.

[12] Software implementations of affinity propagation,along with the data sets and similarities used to obtain the redults described in this manuscript are available. http://www.psi.toronto.edu/affinity propagation.

[13] Hopfield J J.Neural networks and physical systems with emergent collective computational abilities[J].*Proc. Natl. Acad. Sci. U. S. A.*,1982,**79**:2554-2558.

[14] Kschischang F R,Frey B J,Loeliger H A.Factor graphs and the sum-product algorithm[J].*IEEE Trans. Inform. Theory*,2001,**47**:498-519.

[15] 孙即祥. 现代模式识别 [M]. 国防科技大学出版社,2002,112-137.

[16] 舒宁，马洪超，孙和利. 模式识别的理论与方法 [M]. 武汉大学出版社, 2004,13-46.

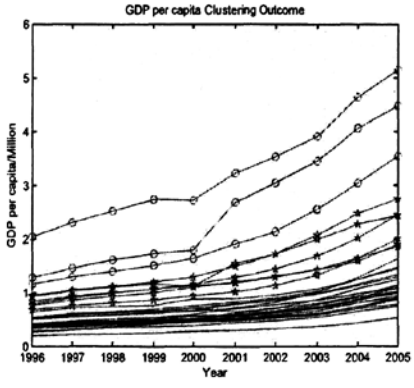
附　录 I

A.$s(k,k) = mean_{i \neq k} s(i,k)$ **时分类图解详情**

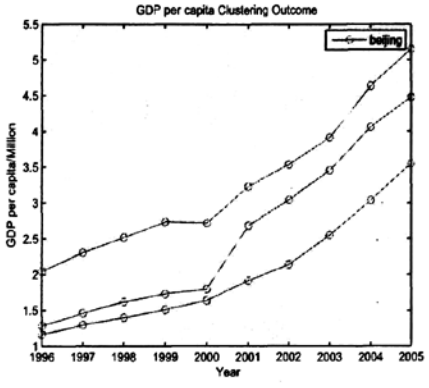
第一类（北京）：北京　天津　上海

第二类（江苏）：辽宁　江苏　浙江　福建　山东　广东

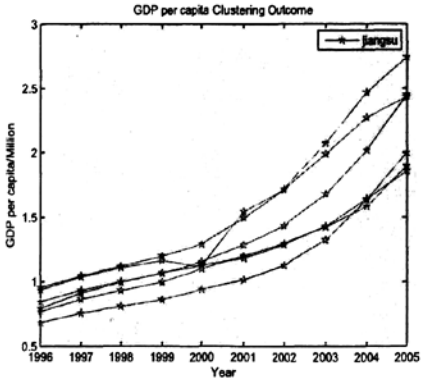
第三类（重庆）：河北　山西　内蒙古　吉林　黑龙江　安徽　江西　河南　湖北　湖南　广西　海南　重庆　四川　贵州　云南　西藏　陕西　甘肃　青海　宁夏　新疆



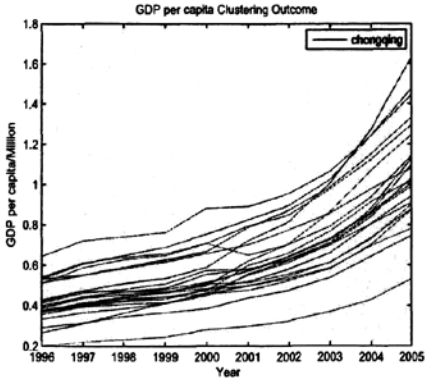
(分类结果)



(北京类)



(江苏类)

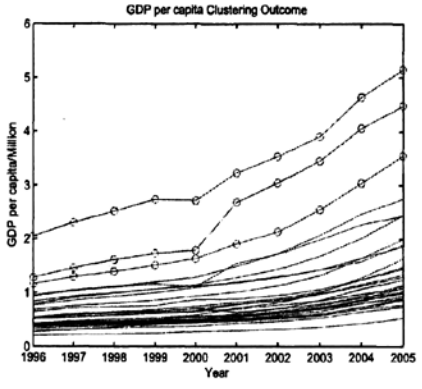


(重庆类)

B.$s(k,k) = min_{i \neq k} s(i,k)$ **时分类图解详情**

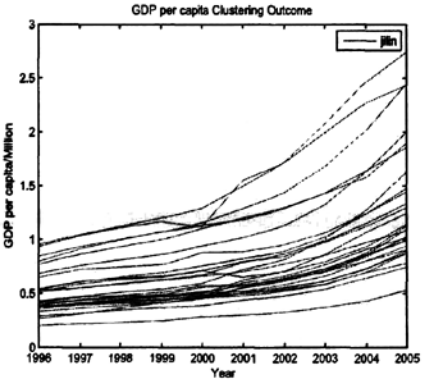
第一类（北京）：北京　天津　上海

第二类（吉林）：河北　山西　内蒙古　辽宁　吉林　黑龙江　江苏　浙江　安徽
福建　江西　山东　河南　湖北　湖南　广东　广西　海南　重庆　四川
贵州　云南　西藏　陕西　甘肃　青海　宁夏　新疆



(分类结果)

(北京类)



(吉林类)

C.$s(k,k) = max_{i \neq k} s(i,k)$ **时分类结果**

应用相似性传播算法时，如果取 $s(k,k) = max_{i \neq k} s(i,k)$, 则三十一个省市被分为三十一个经济区域，结果为每个省市自治区各成一类.

这是符合相似性传播算法理论的. 直观上我们可以这样理解，每个样本点与自己是最相似的，而与其他样本点的相似性就会相对很小，那么在自动分类过程中，每个样本点都不会与其他样本点聚为一类，所以出现了现在的结果.

附　录 II

	1996 年	1997 年	1998 年	1999 年	2000 年
北京	1.283344	1.45975	1.614213	1.729881	1.793603
天津	1.162869	1.296201	1.396426	1.512054	1.637722
河北	0.532537	0.605943	0.647893	0.690836	0.754591
山西	0.420717	0.471229	0.468499	0.470281	0.498577
内蒙古	0.426866	0.472816	0.508439	0.536918	0.589651
辽宁	0.767174	0.865747	0.933782	1.000165	1.101713
吉林	0.512322	0.550575	0.589175	0.628126	0.667592
黑龙江	0.644469	0.722063	0.741821	0.764085	0.881811
上海	2.045243	2.306253	2.519262	2.737422	2.718728
江苏	0.844474	0.934575	1.002499	1.067215	1.153903
浙江	0.954653	1.045826	1.119277	1.198858	1.290644
安徽	0.385379	0.435768	0.453663	0.466344	0.507558
福建	0.792343	0.914186	0.996229	1.070639	1.129378
江西	0.369613	0.413296	0.441895	0.463952	0.483833
山东	0.682126	0.756974	0.810387	0.862558	0.940901
河南	0.399169	0.441335	0.467697	0.487493	0.555063
湖北	0.509906	0.587475	0.627088	0.649712	0.709409
湖南	0.411817	0.462954	0.479559	0.509300	0.573273
广东	0.936523	1.037514	1.108655	1.164279	1.118055
广西	0.369993	0.39224	0.407067	0.414443	0.456703
海南	0.530695	0.551629	0.582895	0.618412	0.658806
重庆	0.388114	0.44382	0.467078	0.481207	0.51435
四川	0.355714	0.393845	0.421554	0.434106	0.48148
贵州	0.200759	0.219906	0.230148	0.245784	0.281852
云南	0.36903	0.401619	0.432891	0.442686	0.455944
西藏	0.26541	0.310403	0.361825	0.412539	0.448321
陕西	0.331747	0.364154	0.384185	0.411169	0.460727
甘肃	0.289493	0.313288	0.345276	0.366488	0.383825
青海	0.376168	0.407359	0.437694	0.467431	0.508861
宁夏	0.371631	0.397962	0.422788	0.444733	0.472544
新疆	0.540053	0.611257	0.639193	0.658709	0.708758

	2001 年	2002 年	2003 年	2004 年	2005 年
北京	2.68295	3.043148	3.450391	4.059129	4.477445
天津	1.911444	2.135809	2.54998	3.038057	3.545177
河北	0.82352	0.893583	1.022498	1.245062	1.47367
山西	0.620272	0.705768	0.861566	1.070876	1.245759
内蒙古	0.720997	0.815864	1.003521	1.275617	1.63267
辽宁	1.200067	1.298649	1.425781	1.582167	1.89742
吉林	0.787941	0.870152	0.984497	1.152458	1.332942
黑龙江	0.889564	0.953895	1.063539	1.24459	1.442801
上海	3.228079	3.532942	3.912466	4.634231	5.148583
江苏	1.28577	1.437048	1.680107	2.018512	2.448918
浙江	1.495413	1.722331	2.073722	2.467945	2.743538
安徽	0.51307	0.555336	0.612028	0.736623	0.878288
福建	1.183968	1.288964	1.428804	1.641512	1.858255
江西	0.519752	0.580407	0.659946	0.806886	0.941025
山东	1.017038	1.131414	1.323633	1.636366	2.002257
河南	0.57907	0.627846	0.710427	0.880291	1.128723
湖北	0.649461	0.703544	0.792644	0.936376	1.141881
湖南	0.580943	0.626269	0.699383	0.842332	1.029298
广东	1.546865	1.718084	1.992034	2.271751	2.432732
广西	0.476053	0.523378	0.580834	0.702291	0.874624
海南	0.70152	0.774558	0.854747	0.97665	1.080399
重庆	0.570126	0.640492	0.726141	0.862527	1.097387
四川	0.496932	0.544795	0.612999	0.73119	0.899307
贵州	0.298307	0.324063	0.368563	0.429764	0.530579
云南	0.498789	0.533769	0.5841	0.698054	0.780425
西藏	0.555285	0.62382	0.700333	0.804161	0.906895
陕西	0.5495	0.613334	0.701279	0.857107	0.988081
甘肃	0.437037	0.475137	0.537776	0.644708	0.745559
青海	0.573862	0.643951	0.730712	0.86475	1.000589
宁夏	0.599361	0.659371	0.767862	0.913537	1.016946
新疆	0.795096	0.846535	0.975362	1.125364	1.295617

致　谢

首先，我要感谢我的导师张宝学教授．两年来，张老师在我的学习和生活中倾注了无限的关怀．他广博扎实的专业知识，富有启发性的思维方法，孜孜不倦的言传身教和非凡的人格魅力使我获益匪浅，终生难忘．攻读硕士两年期间，还得到了东北师范大学数学系其他各位老师的悉心指导和耐心帮助，他们在我有疑难问题的时候提出了很好的意见和建议．在论文准备阶段，我要特别感谢陶剑教授，感谢陶老师在百忙之中给我提出宝贵的修改意见，使我的论文能够顺利完成；同时也要感谢我的师兄师姐在编程和论文写作等很多方面给予我的热心指导和帮助．在此特向以上各位致以最诚挚的谢意！由于本人能力有限，文中难免有不足之处，还请各位专家不吝赐教．

相似性传播算法及其在中国经济区域划分中的应用

作者：高艳敏
学位授予单位：东北师范大学