基于聚类分析的数据挖掘方法在中国边疆研究中的实践与应用

中国社会科学院中国边疆研究所

常永宽

changyk@cass.org.cn

2016年4月20日

目　录

一、引言

边疆研究是一个集历史、民族、政治、经济、地理等多学科的综合研究，可以多视角、多层次的进行研究。而边疆口岸作为国家对外开放的重要门户，是发展边境贸易和跨国合作的重要平台，也是“一带一路”的重要组成部分。

本文主要以边疆省区(吉林、辽宁、黑龙江、内蒙古、新疆、西藏、云南、广西)的陆路口岸与所在城市为研究对象，利用聚类分析的数据分析挖掘方法，并用R语言编程实现对边疆口岸的岸城关系的分析。

由于边疆地区各口岸和城市形成于不同历史时期，经历了各自的发展过程，其地理区位、发展背景和制约条件不同。一方面，口岸和城市关系（以下简称“岸城关系”）错综复杂，发展相对不平衡。另一方面，口岸功能和城市功能相互影响、相互促进。口岸城市具有强大的口岸功能，城市功能相对较弱，则会限制口岸的发展；反之，口岸功能较弱，城市功能较强，则不能充分发挥口岸对城市的带动作用，城市发展缓慢。

因此，本文在分析岸城关系的基础之上，探讨不同类型岸城关系下的驱动模式，为岸城关系的优化和可持续发展提供理论借鉴。

二、岸城关系的聚类分析及实现

2.1 数据来源

截止2015年底，共有288个一类口岸，其中水运口岸135个，空运口岸66个，陆运口岸87个（铁路口岸20个，公路口岸67个）。剔除未开通的口岸5个，剔除外贸货物吞吐量及出入境人员为0的口岸6个，合并满洲里、阿拉山口等8个铁路、公路口岸并举的数据，最终得到53个口岸的有效数据。

考虑到数据的连续性和可获得性，选取口岸的货物吞吐量作为衡量口岸运输状况的指标，选取城市 GDP 作为衡量城市发展状况的指标。吞吐量数据来源于国家口岸管理办公室、中国口岸协会主编的《中国口岸年鉴》；GDP数据来源于新疆、内蒙古、黑龙江、吉林、辽宁、西藏、云南、广西等各省区2006-2015年的统计年鉴。

2.2聚类分析的概念

关于聚类分析的具体内涵可以用下面这个例子进行阐述，如要想把中国城市分成若干类，有很多分类法：可以按照自然条件来分；可以按照降水、土地、日照、湿度等来划分；也可以依据收入、教育水准、医疗条件、基础设施等指标来划分；既可以用某一项来分类，也可以同时考虑多项指标来分类。

而聚类分析则是按照一定的规律和要求对事物进行区分和分类的过程，在这一过程中没有任何关于分类的先验知识，没有指导，仅靠事物间的相似性作为类属划分的准则。因此，聚类分析是一种无监督分类，它没有预定义的类。

因此，聚类分析就是按照事物的某些属性，把事物聚集成类，使同一聚类内对象的相似性尽可能最大，而不同聚类内的对象相似性尽量达到最小。即形成聚类之后，同一个聚类内的对象具有很高的相似性，而且与不属于该聚类的对象有迥然的差异,从而便于了解数据的分布情况。

随着数据聚类分析的不断发展，目前已经提出有多种不同的聚类算法，主要有：基于划分的方法、层次的方法、基于密度的方法，基于网格的方法和基于模型的方法，这些方法根据自身的特点分别应用于不同的领域中。

在各聚类算法中，而基于划分聚类的算法中最经典的是 K-means算法，本文亦采用此种算法。K-means算法具有算法简单和收敛速度快的特点，下面就K-means算法的思想和具体步骤进行简要的阐述和分析。

2.3 K-means算法

K-means算法是聚类算法中最常用的一种算法。通常K-means算法是从N个数据中随机选择K个对象作为初始聚类中心，计算每个对象与这些中心对象的距离,依最小距离对相应对象进行划分。其中，距离的计算方式采用欧式距离。

向量$X(x_1,x_2,\ldots,x_p)$与 $Y(y_1,y_2,\ldots,y_n)$间的欧氏距离d定义为：

$$d = \sqrt{\sum_i (x_i - y_i)^2}$$

K-means算法步骤具体如下：

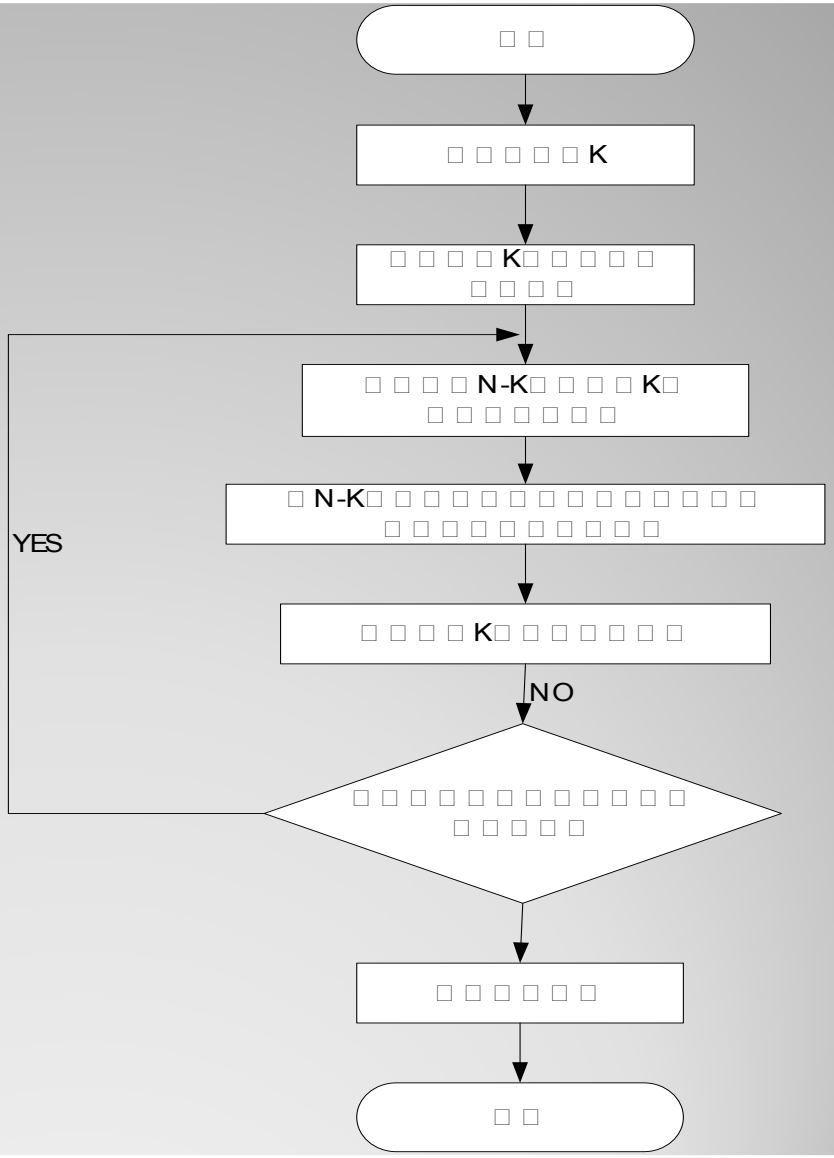
(l)从n个数据对象任意选择k个对象作为初始聚类中心；

(2)根据每个聚类对象的均值(中心对象)，计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分，将每个对象(重新)赋给最相近的类；

(3)重新计算每个(有变化)聚类的均值(中心对象)；

(4)重复(2)和(3)直到每个聚类不再发生变化为止。

算法的基本流程如图所示：



2.4 轮廓系数及K的取值

K-means 算法虽具有算法简单和收敛速度快的特点，但 K 值的预先未知性及初始中心位置选择的随机性都将有可能产生不同的聚类结果，进而导致结果簇集的不确定性。

此外，K-means 算法聚类的个数K需要手动输入，一般情况下，值是未知的，因此需要对聚类个数进行猜测和预判，而预判的结果往往是不准确的，得到的结果也很不理想。因此，K值的确定对聚类的效果起着重要的作用，K值的设置影响到聚类分析效果。

轮廓系数作为聚类效果好坏的一种评价方式，它结合内聚度和分离度两种因素。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。

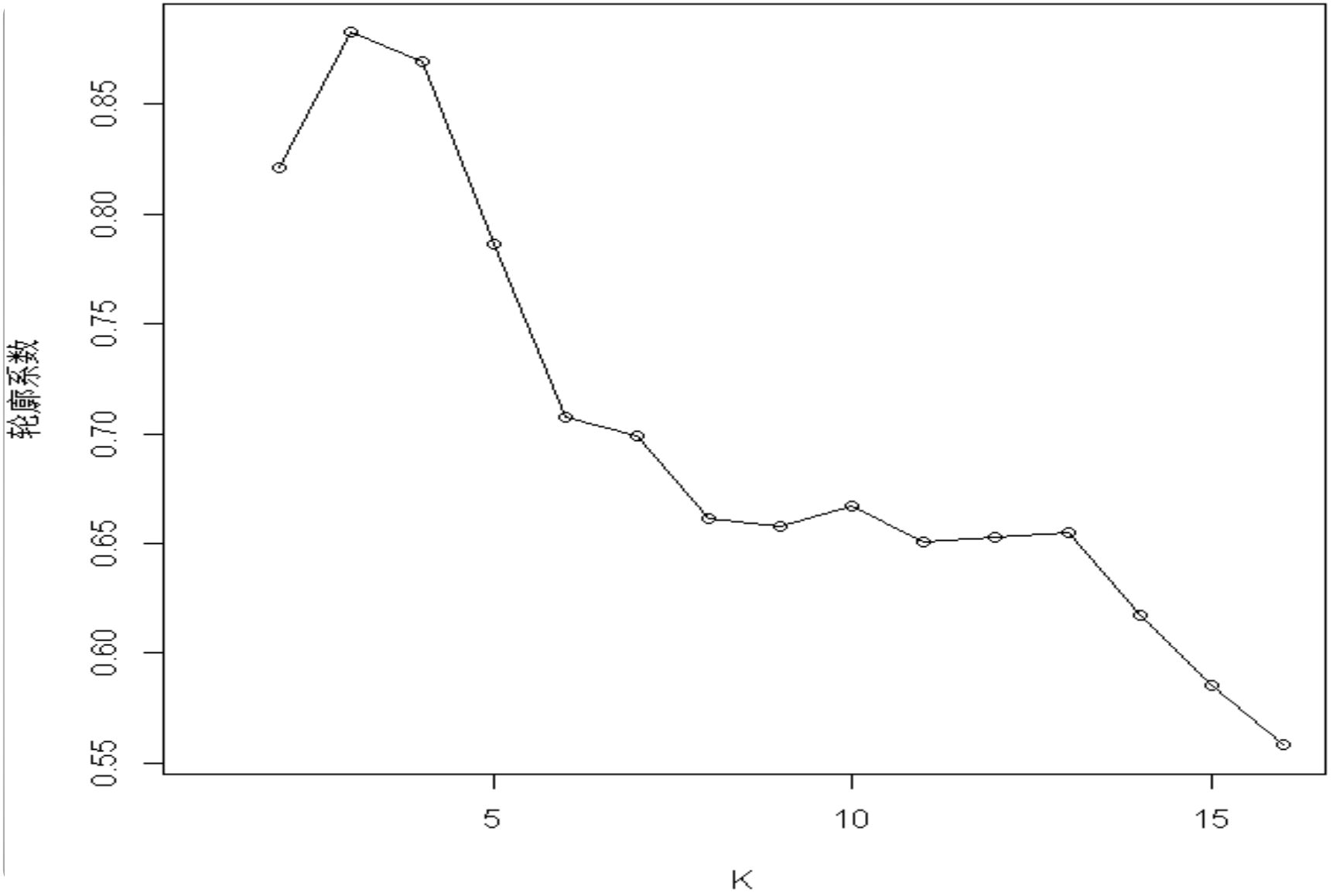
轮廓系数被定义为：

假设样本d_i被聚类到簇A,其轮廓系数s_i定义如下:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

其中 a_i是对象i到本簇中其他对象的平均距离，b_i是对象i到其他簇中对象平均距离的最小值。

轮廓系数可以用来判断聚类的优良性，其值在-1到+1之间取值，值越大表示聚类效果越好。


轮廓系数
0.55
0.60
0.65
0.70
0.75
0.80
0.85
5
10
15
K

上图为通过对边疆口岸各组数据进行计算轮廓系数的示意图。从中不难发现，当 K=3时，轮廓系数最大。根据轮廓系数的定义，值较大时的K较优，所以估算最优K=3为最优点。

2.5 聚类的程序运行结果

R是一个免费的开源软件，它提供了首屈一指的统计计算和绘图功能，尤其是大量的统计分析、数据挖掘方面的算法功能，是一个功能强大的数据挖掘和分析工具。本文主要是基于R实现的数据挖掘和分析。下图为K=3时用R进行聚类分析的运行结果。

```
> sort(KM$cluster)
```

丹东	珲春	珠恩嘎达布其	黑山头	满都拉
1	1	1	1	1
卡拉苏	伊尔克什坦	吐尔尕特	巴克图	塔克什肯
1	1	1	1	1
乌拉斯台	磨憨	腾冲猴桥	打洛	龙邦
1	1	1	1	1
绥芬河	满洲里	二连浩特	甘其毛都	策克
2	2	2	2	2
阿拉山口	霍尔果斯	老爷庙	樟木	瑞丽
2	2	2	2	2
河口	东兴	友谊关	图们	南坪
2	2	2	3	3
古城里	临江	长白	三合	圈河
3	3	3	3	3
沙坨子	开山屯	东宁	密山	虎林
3	3	3	3	3
室韦	额布都格	阿日哈沙特	阿尔山	都拉塔
3	3	3	3	3
吉木乃	红其拉甫	红山嘴	畹町	金水河
3	3	3	3	3
天保	孟定清水河	水口		
3	3	3		

以上聚类结果，具体详细如下：

1、（13个）绥芬河、满洲里、二连浩特、甘其毛都、策克、阿拉山口、霍尔果斯、老爷庙、樟木、瑞丽、河口、东兴、友谊关

2、（25个）图们、南坪、古城里、临江、长白、三合、圈河、沙坨子、开山屯、东宁、密山、虎林、室韦、额布都格、阿日哈沙特、阿尔山、都拉塔、吉木乃、红其拉甫、红山嘴、畹町、金水河、天保、孟定清水河、水口

3、（15个）丹东、珲春、珠恩嘎达布其、黑山头、满都拉、卡拉苏、伊尔克什坦、吐尔尕特、巴克图、塔克什肯、乌拉斯台、磨憨、腾冲猴桥、打洛、龙邦

三、边疆口岸的聚类分析的实际意义

通过上节利用K-means聚类算法对各口岸进行聚类分析可知，将边疆口岸聚类为三大类。下面结合岸城发展的弹性系数和岸城增长的相对集中指数来分析边疆口岸的聚类分析的实际意义。

根据郭建科等人对港口与城市的关系研究，本文将其提出的DCI（Dynamic Concentration Index）引入口岸与城市的关系之中进行分析。DCI主要由岸城发展的弹性系数和岸城增长的相对集中指数构成。其中，岸城发展的弹性系数强调口岸相对于城市发展的增长率，岸城增长的相对集中指数强调口岸相对于城市的增长量比重，两者分别从增长速度和增长规模两个角度诠释口岸与其所在城市的发展活力和相对关系。具体定义如下：

（1）岸城发展的弹性系数(D_e)，指在一定研究周期内，口岸运输的平均增长率与所在城市经济发展的平均增长率的比值，计算公式为：

$$D_e = \left(\sqrt[n-1]{\frac{T_n}{T_1}} - 1 \right) \Big/ \left(\sqrt[n-1]{\frac{C_n}{C_1}} - 1 \right)$$

T_n为研究期内第n年的口岸的吞吐量， C_n为研究期内第n年的城市GDP。

（2）岸城增量的相对集中指数(D_i)，指在一定研究周期内，口岸吞吐量平均增长量比重与其所在城市经济的平均增长量比重的比值，计算公式为：

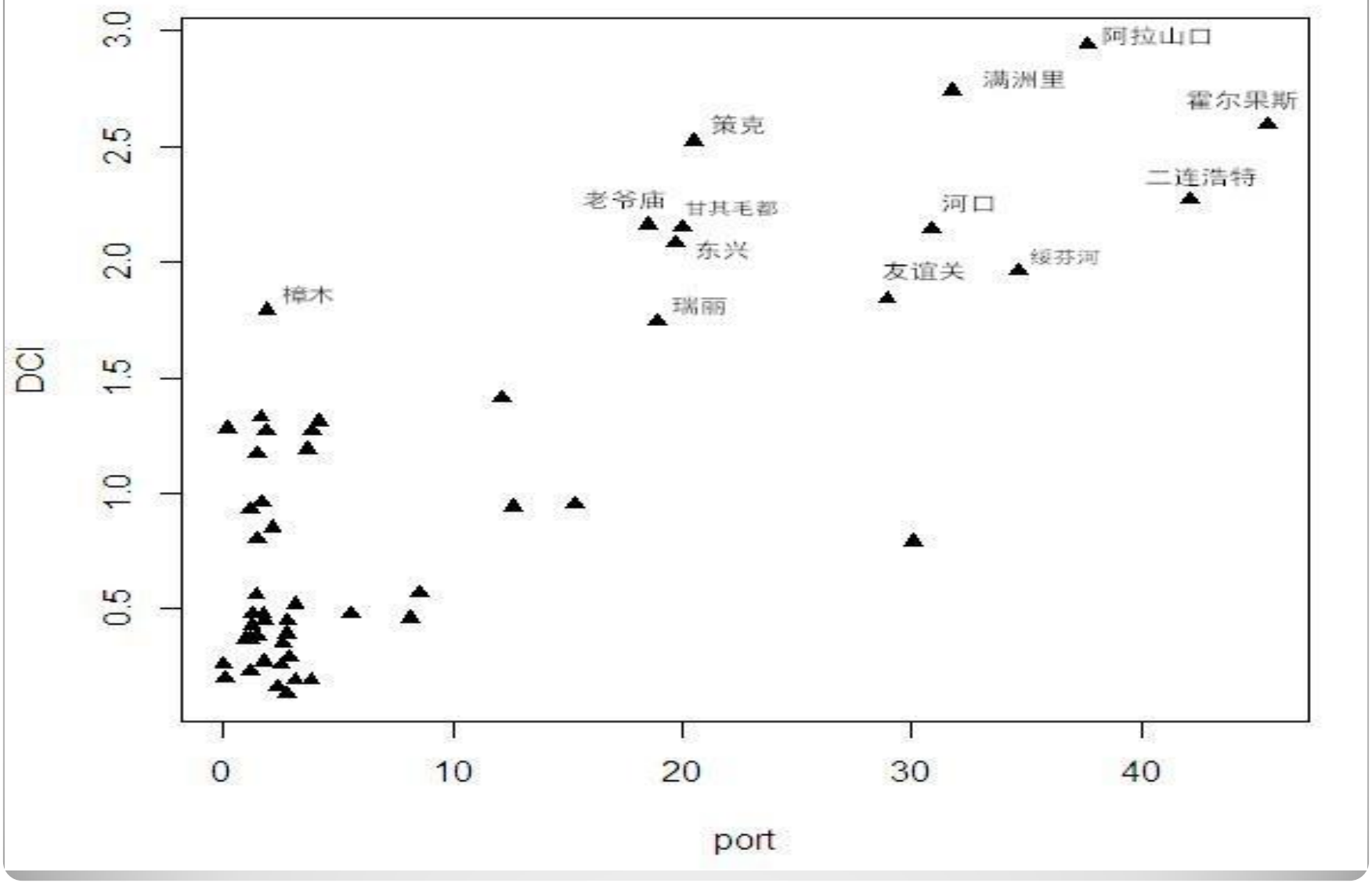
$$D_i = \left(\frac{T_n - T_1}{(n-1) \times \sum_{i=1}^{n} T_i} \right) \Big/ \left(\frac{C_n - C_1}{(n-1) \times \sum_{i=1}^{n} C_i} \right)$$

表征岸城相对速度的D_e相对于某一个口岸城市来讲是确定的，是基础性指标；表征岸城相对比重的D_i根据研究区域不同具有可变性，是补充性指标；在此基础上，确定D_e权重为0.6，D_i权重为0.4，从而最终确定DCI值。

一般来说DCI值越大表明岸城关系中口岸功能明显，反之，城市功能明显。

3.1 各口岸类型DCI分布与分析

根据各口岸的吞吐量、GDP等基础数据，计算各口岸的De、Di、DCI值。其示意图如下图所示。


阿拉山口
满洲里
霍尔果斯
策克
二连浩特
老爷庙
甘其毛都
河口
东兴
绥芬河
友谊关
樟木
瑞丽
DCI
port
0
10
20
30
40
0.5
1.0
1.5
2.0
2.5
3.0

可以得出，通过聚类分析的三类口岸均分布在三个区间，即DCI>1.5、0.75≤DCI≤1.5、DCI<0.75。我们将其定义为三类，分别为口岸驱动型、互动型口岸、城市驱动型。具体如下:

(1) 口岸驱动型(DCI>1.5)。主要指绥芬河、二连浩特、甘其毛都、策克、满洲里、阿拉山口、霍尔果斯、河口、东兴、友谊关、瑞丽、樟木、老爷庙等13个口岸。


满洲里
绥芬河
霍尔果斯
老爷庙
二连浩特
甘其毛都
策克
樟木
瑞丽
河口
友谊关
东兴
南海诸岛

此类城市的岸城关系特点是口岸发展快于城市，口岸职能明显高于城市职能，口岸成长相对处于活跃期，口岸对城市发展具有拉动作用，口岸的流通地位显著。

(2) 互动型口岸(0.75≤DCI≤1.5)。主要指珠恩嘎达布其、卡拉苏、伊尔克什坦、吐尔尕特、吉木乃、巴克图、乌拉斯台、龙邦、丹东、珲春、满都拉、塔克什肯、磨憨、腾冲猴桥、打洛等15个口岸。


吉木乃
塔克什肯
巴克图
乌拉斯台
吐尔尕特
伊尔克什坦
卡拉苏
珠恩嘎达布其
满都拉
珲春
丹东
腾冲猴桥
打洛
磨憨
龙邦
南海诸岛

此类口岸与城市发展速度几乎相同，其口岸功能和城市功能处于较平衡的状态，口岸与城市互为依托，岸城之间达到了较为理想的互动状态。

(3)城市驱动型(DCI<0.75)。主要指图们、南坪、古城里、临江、长白、三合、圈河、沙坨子、开山屯、东宁、密山、虎林、黑山头、室韦、额布都格、阿日哈沙特、阿尔山、都拉塔、红其拉甫、红山嘴、畹町、金水河、天保、孟定清水河、水口等25个口岸。


室韦
黑山头
阿日哈沙特
额布都格
阿尔山
密山
虎林
沙坨子
东宁
三合
开山屯
临江
圈河
南坪
长白
开山屯
古城里
红山嘴
都拉塔
红其拉甫
畹町
天保
水口
孟定清水河
金水河
南海诸岛

此类岸城关系为城市发展快于口岸，城市功能略强于口岸功能，城市自身发展能力较强，对口岸的依赖程度较小，城市发展对口岸的带动作用较明显。城市发展相对处于活跃期。

由以上可见，有一半以上的边疆口岸发展较差，口岸运行不畅，其发展对城市的带动作用不强，没有充分发挥口岸的优势作用。城市职能明显高于口岸职能，岸城关系较弱，城市发展不以口岸为依托，与内陆地区城市发展轨迹相似。

3.2 部分岸城关系分析

序号	口岸	DCI	De	Di	序号	口岸	DCI	De	Di
1	霍尔果斯	2.026	1.471	2.859	7	巴克图	1.375	1.541	1.127
2	老爷庙	1.962	2.776	0.741	8	塔城	1.043	1.319	0.628
3	瑞丽	1.892	1.324	2.743	9	图们	0.699	0.596	0.853
4	二连浩特	1.816	0.824	3.305	10	室韦	0.603	0.663	0.514
5	丹东	1.444	1.548	1.287	11	畹町	0.583	0.829	0.214
6	珲春	1.404	1.394	1.419	12	水口	0.478	0.447	0.525

在口岸驱动型关系中，从吞吐量角度来说，霍尔果斯吞吐量超过3000万吨，二连浩特吞吐量超过1000万吨，老爷庙、瑞丽吞吐量亦超过200吨，口岸的流通地位显著，，但城市只是边境贸易的进出口转换点、“过货化”特征明显。另一方面，口岸发展水平远高于城市发展，口岸城市无法为口岸进一步发展提供足够的支撑作用，进而在一定程度上制约了口岸的发展。

从中我们也可以看出老爷庙的D_e值最大，为2.776，但D_i仅为0.741，表明虽然口岸规模与城市经济规模相比仍较小，但增长速度远远快于城市，口岸是现阶段整个城市发展的源动力，今后该口岸将逐步取得与其城市经济规模相符的枢纽地位。二连浩特的D_i值最大，为3.305，但D_e仅为0.824，表明其口岸相对规模远远超过城市经济，口岸相关产业仍是城市的主要经济支柱，但口岸增速小于城市，岸城开始转向融合。

2、在互动型口岸关系中，从吞吐量角度来说，丹东、磨憨的吞吐量均超过100万吨，属大型口岸；珲春、巴克图吞吐量为25万吨左右，属于小型口岸。

这四个4个口岸在增长速度和增长规模上均与城市趋于平衡，其口岸功能和城市功能处于较平衡的状态，口岸与城市互为依托，两者之间达到了较为理想的互动状态。

磨憨是DCI最接近1的城市，其口岸运输和城市经济在区域中的地位相当，今后磨憨岸城关系将更倾向于互动、融合发展。

3、在城市驱动型关系中，图们、室韦、畹町和水口的吞吐量均小于20万吨，这些口岸的城市职能明显高于口岸职能，岸城关系较弱。

其中畹町的Di值最小，说明相对于畹町的城市经济而言，其口岸规模较小，在现阶段，该城市自身经济发展仍能拉动口岸运输保持较快增长势头。

四、总结

本文通过利用聚类分析的算法进行数据挖掘分析，并利用DCI来进一步探讨聚类分析的实际意义。由以上分析可知，仍有将近一半口岸属于城市驱动型城关系，口岸运行不畅甚至基本没有发展，口岸没有真正发挥作用，对口岸城市带动作用较弱，“过货化”现象比较明显。

随着“一带一路”战略的推进，加大内陆开放步伐，口岸发展取得了明显进展，口岸城市的发展也逐渐向好。在未来发展中，如何借助“一带一路”战略推动口岸城市的进一步发展，加强岸城关系的协调发展，发挥其对边疆地区的带动作用，值得进一步探讨和分析。

谢谢！