Модификация метода k-средних с неизвестным числом классов

Исакин Максим Александрович

Государственный университет – Высшая школа экономики

В работе осуществляется построение и исследование модификации метода k-средних с неизвестным числом классов. В процедуре используется адаптивная метрика Махаланобиса и динамически рассчитываемые меры аномальности наблюдения и однородности классов. С помощью предложенного метода осуществляется классификация регионов РФ с целью определения кластеров регионов, однородных по уровню благосостояния населения, качеству населения и качеству социальной сферы.

Modification of K-Means method with unknown number of classes

Isakin Maksim

State University – Higher School of Economics

The classification procedure based on K-Means method with unknown number of classes is designed and investigated in the paper. An adaptive Mahalanobis' metrics and dynamically computed measures of observation anomality and class homogeneity are used in the method. The classification of regions of the Russian Federation by quality of life integral characteristics is obtained with the proposed method. A brief interpretation of the results is proposed.

При проведении различных социально-экономических исследований нередко возникает задача разбиения рассматриваемой совокупности регионов на однородные[1] классы. В частности результаты классификации могут использоваться для нахождения групп регионов со схожим уровнем социально-экономического развития при определении приоритетов социально-экономического развития региона на основе анализа интегрального индикатора качества жизни населения.

Этот механизм определения приоритетов, подробно описанный в [Айвазян, Исакин (2006)], основывается на выявлении так называемых проблемных областей общественной жизни и областей, наиболее значимо влияющих на качество жизни населения региона. Таким образом, каждый фактор (показатель) качества жизни населения должен быть рассмотрен с точки зрения его проблемности, а также его значимости. При этом

приоритетными направлениями в социально-экономической политике региона являются направления, характеризующиеся одновременно проблемностью и высокой значимостью факторов (показателей). Определение проблемных областей общественной жизни анализируемого региона основывается на анализе, с одной стороны, динамики социально-экономических показателей области, а с другой – положения региона относительно других субъектов Федерации. В то же время следует отметить, что ранжирование регионов Российской Федерации по интегральным индикаторам качества жизни может не дать адекватной оценки в связи со существенными различиями в социально-экономическом положении регионов. Поэтому ранжирование следует производить в первую очередь, в рамках группы регионов, схожих с анализируемым регионом по социально-экономическим, географическим и другим базовым признакам.

В связи с тем, что обучающая выборка отсутствует, определить такие группы возможно с помощью процедуры кластерного анализа. Одним из наиболее известных методов кластерного анализа является процедура последовательной кластеризации – метод k-средних при неизвестном числе классов [MacQueen (1967)]. В этой процедуре выделяют несколько очевидных недостатков, таких как: игнорирование различий в масштабе шкал; игнорирование корреляционных зависимостей между анализируемыми признаками, что обуславливает использование только евклидовой метрики в соответствующем признаковом пространстве[2]; чисто эвристический характер выбора значений свободных параметров процедуры - так называемых "мер точности и грубости". И если проблема различия масштабов шкал признаков традиционно решается с помощью подходящего унифицирующего преобразования, приводящего все признаки единой шкале измерений, то два других недостатка существенно снижают эффективность процедуры: если мы работаем с наблюдениями, извлеченными, например, из многомерных

[1] Вид типологической однородности зависит от решаемой задачи.
[2] Этот недостаток характерен для многих других методов кластерного анализа, использующих расстояние Евклида.

нормальных совокупностей, различающимися только своими средними значениями, то обычная евклидова метрика должна быть заменена на так называемую "махаланобисскую" (см. [Айвазян, Мхитарян (2001)], с. 498); в то же время, априорное фиксирование мер точности и грубости, с одной стороны, связано с определенными трудностями выбора численных значений этих параметров, с другой – некорректно, так как эти величины могут меняться в процессе классификации.

В данной работе осуществляется построение и исследование модификации метода k-средних при неизвестном числе классов, общая схема которой предложена проф. С.А. Айвазяном. Этот метод не обладает описанными выше недостатками, присущими «классическому» методу, потому что, во-первых, в модифицированном методе используется адаптивная метрика Махаланобиса, во-вторых, меры точности и грубости (аномальности наблюдения и однородности классов) рассчитываются в алгоритме динамически[3]. Модифицированный метод апробируется на выборках, сгенерированных Монте-Карло. С помощью предложенного подхода осуществляется классификация регионов РФ на основе трех систем показателей, отражающих интегральные категории качества жизни населения в модели интегральных индикаторов качества жизни, разработанной проф. С.А. Айвазяном: уровень благосостояния населения, качество населения и качество социальной сферы. (см. [Айвазян (2003)]). Результаты, представленные в данной статье использовались в научно-исследовательской работе «Внедрение системы целеполагания и оценки деятельности администрации и органов исполнительной власти Пермской области», которая проводилась по заказу Аппарата администрации Пермской области в 2004-2005гг.

[3] Использование метрической функции с изменяющимися параметрами в некотором смысле опасно (как и вообще игра с изменяющимися в процессе правилами), однако в данном случае продуктивно.

Описание метода

В рассматриваемой задаче предполагается, что генеральная совокупность, из которой извлечены объекты, представляет собой смесь многомерных нормальных распределений с различными математическими ожиданиями и одинаковыми ковариационными матрицами. В процедуре кластеризации используется адаптивная метрика Махаланобиса. Классы, которые образуются в результате выполнения процедуры, обозначаются $\{S_i\}$, $i = 1,2,...,k$, где k-число классов. Класс S_i характеризуется двумя величинами: центром (вектор размерности p) $e_i = (e_{i,1}, e_{i,2},..., e_{i,p})^T$ и весом w_i (скаляр). Центр класса характеризует «центр тяжести» наблюдений, отнесенных к соответствующему кластеру, а вес класса определяет количество таких наблюдений. Алгоритм кластеризации описывается следующей последовательностью действий.

1. Задается некое число классов k_0. Из выборочной совокупности извлекается k_0 точек[4], и эти точки объявляются центрами классов. Вес каждого класса устанавливается равным единице:

$$\begin{cases} e_i = X_i \\ w_i = 1, \ i = \overline{1,k}. \end{cases}$$

2. Выполняется процедура проверки однородности классов. Для этого рассчитываются расстояния между каждой парой классов по формуле

$$d^2(S_i, S_j) = \frac{n_i n_j}{n_i + n_j}(e_i - e_j)^T \Sigma^{-1}(i,j)(e_i - e_j), \qquad (1)$$

где n_k, e_k и $\Sigma(i,j)$ - соответственно, количество точек попавших в класс S_k, центр тяжести этого класса и совместная ковариационная матрица классов S_i и S_j. Величины $\{n_k\}$ и $\{\Sigma(i,j)\}$ рассчитываются на основе «минимального дистанционного разбиения»[5].

---

[4] При этом могут быть извлечены первые k_0 точек выборки, либо из выборки случайным образом могут быть извлечены k_0 точек.

[5] См. ниже.

После этого расстояние между двумя ближайшими классами (смысле указанной метрики) сравнивается с мерой однородности классов

$$c_0(S_i, S_j) = \frac{(n_i + n_j - 2)p}{n_i + n_j - p - 1} F_\alpha(p; n_i + n_j - p - 1), \qquad (2)$$

где $F_\alpha(\nu_1, \nu_2)$ – 100α-процентная точка F-распределения с ν_1 и ν_2 степенями свободы числителя и знаменателя соответственно.

Если расстояние между ближайшими классами оказывается меньше $c_0(S_i, S_j)$, то соответствующие классы объединяются, т.е. эти два класса S_i и S_j удаляются и формируется новый класс S_k со следующими характеристиками

$$\begin{cases} e_k = \dfrac{w_i e_i + w_j e_j}{w_i + w_j} \\ w_k = w_i + w_j \end{cases}.$$

Процедура объединения классов повторяется до тех пор, пока расстояние между любыми двумя классами не будет превышать меру однородности классов $c_0(S_i, S_j)$. После объединения классов мы имеем $k_0' < k_0$ классов.

3. Из выборочной совокупности извлекается очередная точка X и вычисляется расстояние от этой точки до каждого из классов по формуле

$$d^2(X, S_j) = \frac{n_j}{n_j + 1}(X - e_j)^T \Sigma^{-1}(j)(X - e_j), \qquad (3)$$

и определяется ближайший до рассматриваемой точки в смысле указанной метрики класс S_i. После этого расстояние до ближайшего класса сравнивается с мерой аномальности наблюдения

$$c_1(S_i) = \frac{(n_i - 1)p}{n_i - p} F_\alpha(p; n_i - p). \qquad (4)$$

Если расстояние до ближайшего класса превышает величину $c_1(S_i)$, то рассматриваемая точка объявляется центром нового класса с весом равным единице. Если же минимальное расстояние не превышает меру аномальности, то рассматриваемая точка присоединяется к ближайшему

классу S_i, при этом характеристики этого класса пересчитываются по следующим формулам

$$\begin{cases} e_i = \dfrac{1}{w_i + 1}\left(w_i e_i + X\right) \\ w_i = w_i + 1. \end{cases}.$$

Параметры других классов при этом не изменяются.

После этого алгоритм повторяется с шага 2, то есть производится процедура объединения близких классов с использованием меры однородности классов. Объекты поочередно рассматриваются до тех пор, пока не закончится выборка.

Для определения величин $\{n_k\}$ и $\{\Sigma(i)\}$, используемых при расчете расстояний между парами классов, классом и точкой, а также при вычислении мер однородности классов и аномальности наблюдения, осуществляется процедура минимального дистанционного разбиения. Эта процедура заключается в разбиении всей выборочной совокупности наблюдений на классы с центрами определенными на момент выполнения процедуры. При этом подсчитываются количества наблюдений n_i, попавших в класс S_i, определяются выборочные средние наблюдений, отнесенных к классу S_i:

$$\overline{X}(i) = \frac{1}{n_i} \sum_{X_k \in S_i} X_k \ ,$$

и ковариационные матрицы наблюдений, попавших в класс S_i:

$$\Sigma(i) = \frac{1}{n_i - 1} \sum_{X_k \in S_j} \left(X_k - \overline{X}(i)\right)\left(X_k - \overline{X}(i)\right)^T \ .$$

При вычислении меры однородности классов S_i и S_j используется совместная ковариационная матрица этих классов

$$\Sigma = \frac{1}{n_i + n_j - 2}\left[\sum_{X_k \in S_i} \left(X_k - \overline{X}(i)\right)\left(X_k - \overline{X}(i)\right)^T + \sum_{X_k \in S_j} \left(X_k - \overline{X}(j)\right)\left(X_k - \overline{X}(j)\right)^T \right].$$

В процедуре минимального дистанционного разбиения используется метрика (3) с ковариационной матрицей, определенн ой в предыдущем

минимальном дистанционном разбиении. При первом выполнении процедуры в качестве ковариационных матриц классов могут использоваться единичные ковариационные матрицы.

При недостатке информации для расчета выборочной ковариационной матрицы класса $\Sigma(i)$ (например, когда к классу отнесено лишь одно наблюдение) при вычислении расстояний может использоваться априорно заданная ковариационная матрица. Аналогично, в случае, когда в классах недостаточно наблюдений для того, чтобы использовать формулы (2) и (4) для вычисления мер грубости и точности, используются априорно заданные характеристики.

В виду того, что выборочная совокупность может иметь небольшой объем, возможна ситуация, при которой не достигается устойчивое разбиение на классы к моменту, когда закончится выборка. В этом случае после последнего наблюдения выборочной совокупности снова извлекается первое, второе наблюдения и т.д., таким образом, выборочная совокупность зацикливается. В качестве критерия устойчивости разбиения на классы (и, следовательно, остановки алгоритма) предлагается следующее естественное условие. Процесс классификации останавливают, когда расстояния каждого из нескольких последних перемещений центров каждого из классов не превысили априорно заданную величину, т.е. центры всех классов стабилизировались[6]. В этом случае считается, что удалось получить устойчивое разбиение выборочной совокупности на классы, и задача кластеризации решена.

Полученное разбиение зависит от уровня значимости, используемого для расчета мер однородности классов и аномальности наблюдения, априорно заданного начального числа классов, а также от выбора первых k_0 точек выборки, которые объявляются начальными центрами классов. Уровень значимости, как правило, задается в промежутке от 1% до 10%. Это

значение определяет меру однородности классов и меру аномальности наблюдений (как доверительный интервал F-распределения), при этом чем меньше уровень значимости, тем больше указанные меры, и, вероятно, меньшее число классов получится в результате классификации. Определенное влияние на результаты классификации и на скорость сходимости алгоритма оказывает начальный выбор ковариационной матрицы. Если отсутствует информация о дисперсии признаков внутри классов имеет смысл определить диагональные элементы ковариационной матрицы в зависимости от общего разброса значений признаков, по которым осуществляется классификация, априорно заданного начального числа классов k_0 и предполагаемой структуры размещения классов. В большинстве случаев отсутствует априорная информация, позволяющая определить коэффициенты ковариации между признаками внутри классов, в следствие чего первоначальная ковариационная матрица определяется как диагональная матрица. В общем случае первоначальная ковариационная матрица должна определятся исследователем как априорный параметр метода.

Предложенный алгоритм классификации апробирован на выборках, сгенерированных Монте-Карло, из различных смесей нормальных распределений с коррелированными признаками (корреляционные матрицы признаков всех классов равны по предположению предлагаемого метода) и расстоянием Махаланобиса между классами от 7 до 10 (см. [Апраушева (1981)]). При классификации использовалось начальное число классов близкое к истинному количеству компонентов смеси (в частности, для пяти классов начальное число классов выбиралось от 3 до 7). В результате получена правильная классификация, и результаты классификации совпадают для уровней значимости 1, 5 и 10% и не зависят от порядка выбора наблюдений из выборки. Классификация с помощью «классического»

[6] Число учитываемых последних итераций задается до начала процедуры классификации. При этом предполагается, что на этих последних итерация не происходило образование новых классов или объединения уже существующих.

метода *k*-средних приводит к неверным результатам (формируются «лишние» классы либо сливаются априорно различные классы в зависимости от выбора мер точности и грубости), в связи с тем, что в алгоритме используется «сферическое» евклидово расстояние (на рис. 1 отражены результаты классификации классическим методом *k*-средних Монте-Карло сгенерированной выборки с тремя классами и высокой степенью корреляции между факторами: большие точки – центры полученных классов).

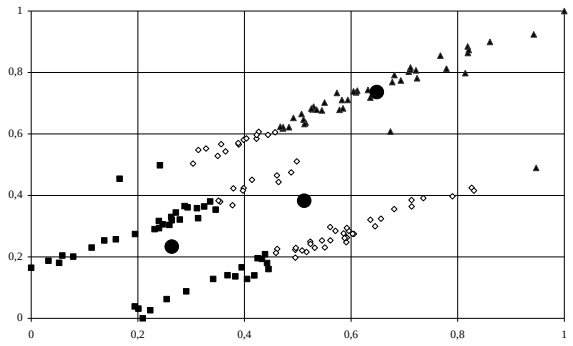

Рис. 1. Результаты классификации

Классификация регионов

С помощью предложенного метода получены три классификационных разбиения регионов РФ по показателям интегральных категорий качества жизни населения: уровень благосостояния населения, качество населения и качество социальной сферы [Айвазян (2003)]. В табл. 1 отражены три соответствующих набора показателей, определяющих пространства классификации. Для проведения исследования использовались данные Госкомстата [Регионы России (2005)] за 2004 год.

Классификационная процедура имеет ряд априорно задаваемых параметров, таких как начальное число классов, уровень значимости, начальная ковариационная матрица, начальные центры классов, порядок извлечения объектов выборки и др. В практике кластерного анализа существуют инструменты, позволяющие обоснованно осуществить выбор каждого их перечисленных параметров. При реализации алгоритма были

использованы методы, автоматизирующие выбор начальной ковариационной матрицы и начальных центров классов. В то же время начальное число классов и уровень значимости определяется экспертным путем. Порядок извлечения объектов выборки не оказывает существенного влияния на результаты классификации. Классификация регионов осуществлялась на уровне значимости 5%, при этом начальное число классов определялось в интервале от 2 до 15. В результате классификации регионов с различными начальными параметрами получается, как правило, 2 -3 различных разбиения, устойчивых к небольшим изменениям начального количества классов. Наиболее устойчивыми разбиениями при проведении классификации с различным начальным числом классов по каждой из категорий являются классификации с четырьмя классами.

Таблица 1. Центры классов полученных классификационных разбиений

Уровень благосостояния населения	Класс 1	Класс 2	Класс 3	Класс 4
1. ВРП на душу населения	315,67	48,65	51,125	97,9
2. Покупательская способность среднедушевых денежных доходов по отношению к наборам прожиточного минимума	312,5	159,21	176,45	247,22
3. Доля численности населения с денежными доходами ниже величины прожиточного минимума	20,533	34,363	30,507	21,346
4. Отношение совокупных доходов 20% самых богатых и 20% самых бедных	18,633	10,452	10,367	13,646
5. Обеспеченность населения собственными легковыми автомобилями	21,033	18,333	21,435	20,537
6. Доля семей, состоящих на учете на получение жилья	9,4333	7,5479	8,5531	10,145
7. Приходится общей площади жилищного фонда на одного жителя	160,73	146,77	131,49	154,6
8. Доля ветхого и аварийного жилья	234,37	87,181	138,84	127,43
9. Плотность автомобильных дорог общего пользования (в расчете на 1 000 населения)	3,2667	6,2042	2,7227	4,075
Качество населения				
1. Ожидаемая продолжительность жизни при рождении: все население - оба пола	61,892	64,35	66,175	63,545
2. Число умерших детей в возрасте до 1 года на 1000 населения	15,831	16,975	12,231	12,061
3. Коэффициент естественного прироста	-7,1979	0,8125	-3,694	-8,4818
4. Число умерших от инфекционных и паразитарных болезней и туберкулеза на 100	62,723	28,562	46,149	52,109

000 населения				
5. Число умерших от новообразований на 100 000 населения	198,02	162,02	181,89	203,06
6. Число умерших от болезней системы кровообращения на 100 000 населения	1034,9	549,05	815,01	965,83
7. Число умерших от болезней органов дыхания на 100 000 населения	83,379	52,625	61,885	98,376
8. Число умерших от болезней органов пищеварения на 100 000 населения	61,292	41,075	51,282	69,727
9. Число умерших от несчастных случаев, травм и отравлений на 100 000 населения	326,99	239,86	192,07	289,16
10. Число инвалидов на 1 000 населения	78,183	53,263	66,66	80,745
11. Зарегистрировано случаев заболевания врожденными аномалиями на 1 000 населения	1,8188	1,4125	1,7104	1,6364
12. Доля специалистов с высшим образованием среди занятых в экономике	20,254	26,787	23,21	20,124
13. Приведенная производительность труда (ВРП на среднегодовую численность занятых в экономике)	110,64	550,01	137,96	130,98
14. Доля учащихся средних специальных и высших учебных заведений среди лиц, не достигших 23 лет	18,877	17,437	19,612	19,615
Качество социальной сферы				
1. Уровень безработицы	8,625	10,247	7,4595	13,073
2. Доля работников занятых во вредных и опасных условиях труда в среднегодовой численности занятых в экономике	27,85	28,065	27,576	24,414
3. Численность пострадавших на производстве со смертельным исходом или с утратой трудоспособности на 1 рабочий день и более в расчете на 1 000 работающих	0,26071	0,3551	0,35952	0,25405
4. Коэффициент миграционного прироста на 10 000 населения	12,546	-22,386	-12,105	-10,319
5. Число зарегистрированных умышленных убийств и покушений на убийство на 10 0 000 населения	31,836	26,733	21,181	18,549
6. Число зарегистрированных фактов умышленного причинения тяжкого вреда здоровью на 100 000 населения	76,671	48,786	44,586	31,397
7. Число зарегистрированных изнасилований и покушений на изнасилование на 100 000 населения	9,2536	8,7102	5,1643	4,8054
8. Число разбоев, грабежей, краж из квартир граждан в расчете на 100 000 населения	406,4	361,19	358,45	269,18
9. Число зарегистрированных незаконных присвоений или растрат в расчете на 100 000 населения	33,043	50,751	31,514	28,624
10. Число больных, состоящих на учете с диагнозом наркомания и токсикомания	524,06	210,92	190,15	459,79
11. Число больных, состоящих на учете с	2110,3	2070,7	3156,6	1768,2

диагнозом алкоголизм				
12. Число самоубийств на 100 000 населения	44,914	60,294	40,2	29,738
13. Число больных, инфицированных ВИЧ	428,3	62,671	63,752	83,095

Результаты классификаций по показателям категорий «Уровень благосостояния населения», «Качество населения» и «Качество социальной сферы» представлены в табл. 2 и на рис. 1.

Таблица 2. Результаты кластеризации субъектов РФ

Регион РФ	**Уровень благосостояния населения**	**Качество населения**	**Качество социальной сферы**
Алтайский край	2	3	4
Амурская обл.	2	1	2
Архангельская обл.	3	1	2
Астраханская обл.	2	3	4
Белгородская обл.	3	3	3
Брянская обл.	3	1	3
Владимирская обл.	3	1	3
Волгоградская обл.	2	3	2
Вологодская обл.	3	4	2
Воронежская обл.	3	4	3
г. Москва	1	2	3
г. Санкт-Петербург	4	3	1
Еврейская авт.обл.	3	1	2
Ивановская обл.	3	1	3
Иркутская обл.	3	4	1
Кабардино-Балкарская Респ.	2	3	4
Калининградская обл.	2	4	1
Калужская обл.	3	3	3
Камчатская обл.	2	3	2
Карачаево-Черкесская Респ.	2	3	4
Кемеровская обл.	3	1	4
Кировская обл.	3	1	2
Костромская обл.	3	1	3
Краснодарский край	2	3	4
Красноярский край	4	4	4
Курганская обл.	2	4	2
Курская обл.	3	3	2
Ленинградская обл.	3	4	1
Липецкая обл.	4	3	3

Регион РФ	Уровень благосостояния населения	Качество населения	Качество социальной сферы
Магаданская обл.	4	3	3
Московская обл.	3	3	1
Мурманская обл.	4	3	4
Нижегородская обл.	3	1	3
Новгородская обл.	3	1	3
Новосибирская обл.	3	3	4
Омская обл.	3	3	2
Оренбургская обл.	2	1	1
Орловская обл.	3	1	3
Пензенская обл.	3	1	3
Пермская обл.	4	4	3
Приморский край	2	3	4
Псковская обл.	3	1	3
Респ. Адыгея	3	3	4
Респ. Алтай	2	1	2
Респ. Башкортостан	3	3	2
Респ. Бурятия	2	1	2
Респ. Дагестан	2	3	4
Респ. Калмыкия	3	3	2
Респ. Карелия	3	1	3
Респ. Коми	4	4	2
Респ. Марий Эл	3	4	2
Респ. Мордовия	3	3	2
Респ. Саха (Якутия)	4	3	3
Респ. Северная Осетия-Алания	3	3	4
Респ. Татарстан	4	3	2
Респ. Тыва	2	1	1
Респ. Хакасия	2	1	2
Ростовская обл.	2	3	4
Рязанская обл.	3	4	3
Самарская обл.	4	3	1
Саратовская обл.	3	1	4
Сахалинская обл.	4	3	3
Свердловская обл.	3	4	1
Смоленская обл.	3	1	3
Ставропольский край	2	3	4
Тамбовская обл.	3	4	3
Тверская обл.	3	1	3
Томская обл.	3	3	4
Тульская обл.	3	4	3

Регион РФ	Уровень благосостояния населения	Качество населения	Качество социальной сферы
Тюменская обл.	1	2	1
Удмуртская Респ.	2	4	2
Ульяновская обл.	3	1	1
Хабаровский край	3	1	2
Челябинская обл.	2	3	1
Читинская обл.	2	1	2
Чувашская Респ.	3	4	2
Чукотский авт.окр.	1	2	3
Ярославская обл.	4	4	3

В категории качества жизни «Уровень благосостояния населения» образовался самостоятельный класс из следующих объектов: г. Москва, Тюменская область, Чукотский автономный округ. Выделение указанных субъектов среди прочих регионов России в первую очередь связано с высокими значениями по показателю ВРП на душу населения. Во второй класс сгруппировались регионы транссибирской транспортной магистрали, регионы Южного федерального округа (за исключением республик Калмыкии, Адыгеи и Северной Осетии), также в этот класс вошли регионы, лежащие на основных направлениях российско-казахских транспортных путей. В третьем классе сосредоточились регионы средней полосы и северы европейской части России, при этом аналогичной структурой благосостояния населения обладают регионы – центры образования Сибири (Омская, Томская и Новосибирская области). Четвертый класс регионов образуют субъекты с относительной низкой плотностью заселения территории, где, при этом, большая часть населения сосредоточена в крупных населенных пунктах.

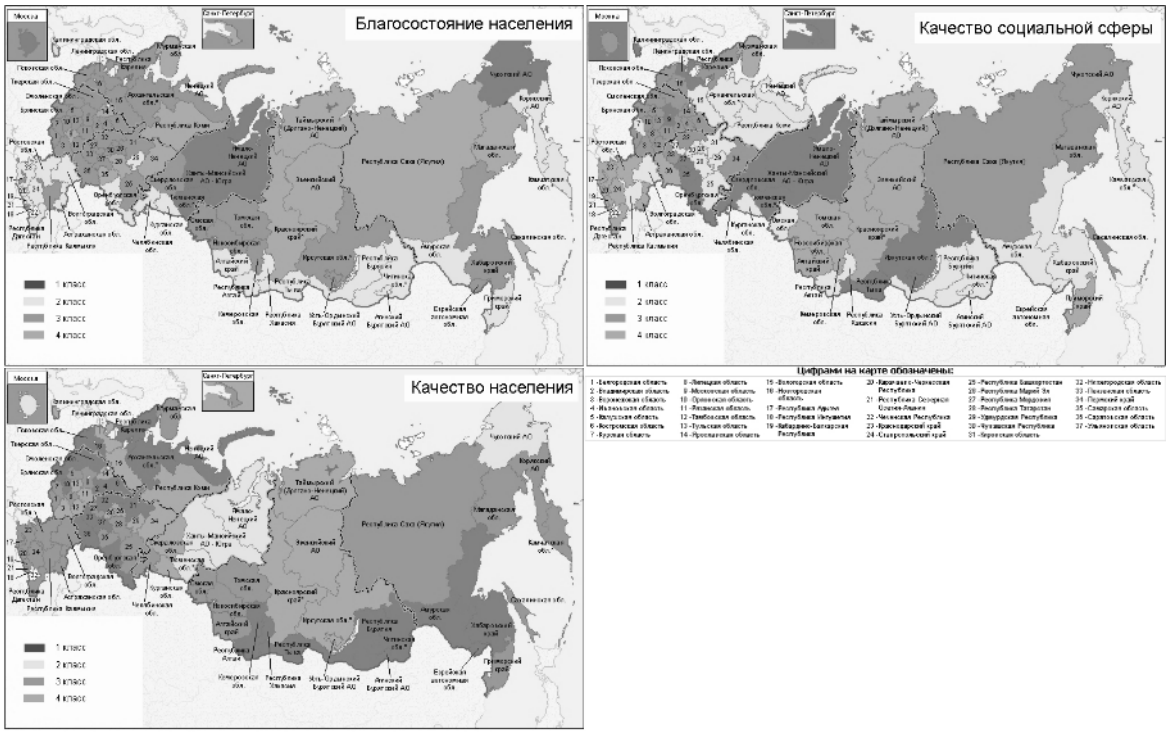

Рис.1. Результаты классификации регионов РФ.

Классификация в категории «Качество населения» обладает сильной территориальной разнородностью и не позволяет сделать однозначные выводы об образованных классах. Однако, в некоторых классах наблюдаются сравнительно однородные группы регионов. Так в первом классе присутствуют регионы, расположенные вдоль транссибирской магистрали, при этом в данный класс качества населения вошла почти половина регионов средней полосы европейской части России, а также Республика Карелия и Архангельская область. Второй класс образован тремя субъектами: г. Москвой, Тюменской областью и Чукотским ОА, как и первый класс категории «Уровень благосостояния». В третий класс вошли все регионы Южного федерального округа, группа субъектов, состоящая из Нижегородской, Челябинской областей, республик Татарстана, Башкортостана, и группа регионов из Омской, Томской, Новосибирской областей и Республики Алтай. Четвертый класс сформирован следующими группами регионов: в западно-европейской части России – Ленинградская, Вологодская Ярославская, Тульская, Тамбовская, Рязанская и Вор онежская

области. В Приуралье – Республика Коми, Удмурдская республика, Пермская и Свердловская области. В данный (четвертый) класс в азиатской части вошли Красноярский край и Иркутская область.

Объекты классов категории «Качество социальной сферы» также имеют сильный пространственный разброс. Однако внутри классов выделяются гомогенные структуры, территориально совпадающие в административными границами федеральных округов и экономических районов, что может быть косвенным признаком локальной непрерывности социогенеза вокруг основных центров развития территорий. В первом классе данной категории не удалось выдвинуть гипотезу о возможных причинах однородной, в смысле классификации, структуры качества социальной сферы населения. Тем не менее в данную классификацию практически полностью вошел Уральский федеральный округ (исключение составила Курганская обл.). Так, например, второй класс включает в себя центральные субъекты Приволжского федерального округа, восточную часть Северо-Западного федерального округа и основные горнодобывающие регионы Забайкалья, включая Амурскую область. В третий класс вошли локация регионов Центрального федерального округа. Аналогичной структурой «Качества социальной сферы» обладает «пояс» из Республики Саха (Якутия), Магаданской области и Чукотского АО. Четвертый класс составляют субъекты Южного федерального округа и западные регионы Сибирского ФО.

Выводы

В работе рассмотрена модификация метода k-средних с неизвестным числом классов, общая схема которой предложена С.А. Айвазяном. Рассматриваемая процедура позволяет устранить ряд недостатков присущих «классическому» методу k-средних, связанных в первую очередь с игнорированием зависимостей между показателями, по которым осуществляется классификация. Модифицированный метод успешно

классифицирует Монте-Карло сгенерированные выборки и демонстрирует устойчивость к различному выбору начальных центров классов и определенную чувствительность к выбору начальной ковариационной матрицы. С помощью предложенной процедуры осуществлена классификация регионов РФ с целью нахождения кластеров регионов, однородных по уровню благосостояния, качеству населения и качеству социальной сферы. На основе полученной классификации в научно-исследовательской работе «Внедрение системы целеполагания и оценки деятельности администрации и органов исполнительной власти Пермской области» осуществлен анализ факторов развития и определение приоритетов социально-экономической политики Пермской области.

Литература

1. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. Основы эконометрики. - М.: ЮНИТИ, 2001.
2. Айвазян С.А. К методологии измерения синтетических категорий качества жизни населения // Экономика и математические методы, том 39, №2, 2003, с. 33-53.
3. Айвазян С.А., Исакин М.А. Интегральные индикаторы качества жизни населения региона как критерии эффективности социально-экономической политики, проводимой органами региональной власти. // Прикладная эконометрика, 2006, №1.
4. Апраушева Н.Н. Определение числа классов в задачах классификации. // Известия АН СССР, серия "Техническая кибернетика", 1981, №3, с. 71-77 и №5, с. 153-160.
5. Регионы России. Социально-экономические показатели. – Госкомстат, 2005.

6. MacQueen, J. Some Methods for Classification and Analysis of Multivariate Observations // Proc. Fifth Berkeley Symp. Math. Stat. and Probab., 1967, vol. 1, pp. 281-297.