УДК 330.1
doi:10.18720/SPBPU/2/id20-254

Фам Тхи Ми Зунг[1],
студент Института компьютерных наук и технологий;
Черненькая Людмила Васильевна[2],
д-р техн. наук, профессор,
профессор Института компьютерных наук и технологий

КЛАССИФИКАЦИЯ РАЙОНОВ ПО УРОВНЮ БЕЗРАБОТИЦЫ ВО ВЬЕТНАМЕ МЕТОДОМ КЛАСТЕРНОГО АНАЛИЗА

[1, 2] Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Россия,
[1] phammydung200695@gmail.com, [2] ludmila@qmd.spbstu.ru

Аннотация. Безработица является неизбежным явлением, характерным для рыночного капиталистического хозяйства. Для решения задачи мониторинга уровня безработицы во Вьетнаме целесообразно применение математических методов многомерной статистики. Классификация районов по уровню безработицы во Вьетнаме методом кластерного анализа очень важна для социально-экономического развития Вьетнама. Данная разработка позволит принимать решения относительно дальнейших действий с выбором приоритетных направлений развития экономики Вьетнаме.

Ключевые слова: классификация районов, уровень безработицы, метод кластерного анализа, Вьетнам.

Pham Thi My Dung[1],
Master Student, Institute of Computer Science and Technology;
Liudmila V. Chernenkaya[2],
Doctor of Technical Sciences, Professor,
Professor of Institute of Computer Science and Technology

CLASSIFICATION OF DISTRICTS BY THE UNEMPLOYMENT RATE IN VIETNAM USING THE CLUSTER ANALYSIS TECHNIQUE

[1, 2] Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russia,
[1] phammydung200695@gmail.com, [2] ludmila@qmd.spbstu.ru

Abstract. Unemployment is an inevitable phenomenon characteristic of a market capitalist economy. To solve the problem of monitoring of the unemployment rate in Vietnam, it is advisable to use mathematical methods of multidimensional statistics.

Classification of districts by the unemployment rate in Vietnam using the cluster analysis technique is very important for the socio-economic development of Vietnam. The development of this task will allow to make decisions regarding further actions with the selection of priority areas for the development of the Vietnam economy.

Keywords: areas classification, unemployment rate, cluster analysis technique, Vietnam.

Введение

Вьетнам – это страна с молодым населением. Однако, в настоящее время темпы роста трудоспособного населения выше, чем темпы естественного прироста всего населения.

В период стабилизации населения темпы общего прироста и прироста трудоспособного населения будут сближаться. Вследствие этого актуальной становится задача разработки планов и политики в области труда и занятости. Поэтому классификация районов по уровню безработицы очень важна для социально-экономического развития Вьетнама.

В работе рассмотрены методы статистического анализа. Метод кластерного анализа разработан в 1939 году исследователем Трионом (Tryon) и является многомерным статистическим методом [1]. Цель метода – разбиения совокупности объектов на однородные группы или кластеры. Основное преимущество метода кластерного анализа в том, что он позволяет проводить разделение объектов не только по одному параметру, но и по полному комплексу признаков. Кроме того, метод позволяет анализировать множество исходных данных фактически произвольной природы и большого объема [2].

Поэтому метод кластерного анализа целесообразно использовать при проведении классификации районов по уровню безработицы во Вьетнаме.

1. Статистические данные для исследования задачи определения уровня безработицы во Вьетнаме

Вьетнам разделяется на следующие районы: район равнинной Красной реки, район равнинной реки Меконг, район северо-запада, южные и северные районы центрального Вьетнама, район юго-востока страны, плато Тэйнгуен.

Статистические данные для исследования задачи определения уровня безработицы для разных возрастных групп населения по районам, городу и деревне во Вьетнаме показаны в таблице 1 [3].

Таблица 1

Статистические данные для исследования уровня безработицы во Вьетнаме

Год, район	**Уровень безработицы, (%)**		
2017	всего	город	деревня
Вьетнам	2,24	3,18	1,78
район равнинной Красной реки	2,20	3,19	1,64
район северо-запада	1,01	2,71	0,68
южные и северный районы центрального Вьетнама	2,54	4,00	1,98
плато Тэйнгуен	1,05	1,98	0,70
район юго-востока страны	2,68	2,83	2,43
район равнинной реки Меконг	2,88	3,63	2,64
2016			
Вьетнам	2,30	3,23	1,84
район равнинной Красной реки	2,24	3,23	1,73
район северо-запада	1,17	3,20	0,77
южные и северный районы центрального Ветнама	2,78	4,30	2,17
плато Тэйнгуен	1,24	2,19	0,88
район юго-востока страны	2,46	2,61	2,19
район равнинной реки Меконг	2,89	3,73	2,62
2015			
Вьетнам	2,33	3,37	1,82
район равнинной Красной реки	2,42	3,42	1,94
район северо-запада	1,10	3,11	0,72
южные и северный районы центрального Вьетнами	2,71	4,51	2,05
плато Тэйнгуен	1,03	2,27	0,57
район юго-востока страны	2,74	3,05	2,17
район равнинной реки Меконг	2,77	3,22	2,63

2. Математические описания кластерного анализа

Задача кластерного анализа заключается в следующем. Для совокупности n объектов каждый объект характеризуется k признаками. С помощью метода кластерного анализа необходимо разбить эту совокупность на сходные по некоторым признакам группы, которые называются кластерами (таксонами). Кластеризация представляет собой разбиение множества объектов на сходные группы (кластеры).

Для приведения диапазона изменения значений признаков к некоторым требуемым границам выполняется процедура нормирования. Существуют разные способы обработки исходных данных задачи [2]:

$$z = \frac{x - \overline{x}}{\sigma}, z = \frac{x}{\overline{x}}, z = \frac{x}{x_{\max}}, z = \frac{x - \overline{x}}{x_{\max} - x_{\min}}$$

где $x, \sigma -$ среднее и среднеквадратическое отклонение x;

$x_{\max}, x_{\min} -$ соответственно наибольшее и наименьшее значение x.

Первым этапом решения задачи кластеризации является выбор способа вычислений расстояний между признаками или объектами.

Расстояние (*метрика*) между объектами в пространстве параметров d_{ab} удовлетворяет следующим условиям:

$$d_{ab} \geq 0; d_{ab} = d_{ba}; d_{ab} + d_{bc} \geq d_{ac}$$

Мерой близости (*сходства*) μ_{ab} имеет предел и возрастает с увеличением близости объектов: μ_{ab} непрерывна, $\mu_{ab} = \mu_{ba}; 0 \leq \mu_{ab} \leq 1$

Процедура кластерного анализа предполагает объединение в группы объектов, наиболее схожих между собой, то есть тех, расстояние между которыми является наименьшим [5].

Пусть K_i – группа (кластер), которая состоит из n объектов.

$x_i -$ среднее арифметическое векторного наблюдения K_i группы, т.е. «центр тяжести» $i -$ й группы;

$\rho(K_i, K_j) = \rho_{ij}$ – расстояние между кластерами K_i и K_j.

Существуют различные способы вычисления расстояния между кластерами:

1) Метод ближнего соседа или метод одиночной связи, когда расстоянием между двумя кластерами является наименьшим:

$$\rho_{\min}(K_i, K_j) = \min_{x_i \in K_i, x_{ji} \in K_j} \rho(x_i, x_j).$$

2) Метод дальнего соседа или метод полной связи, когда расстояние между кластерами определяется как расстояние между самыми удаленными объектами:

$$\rho_{\min}(K_i, K_j) = \max_{x_i \in K_i, x_{ji} \in K_j} \rho(x_i, x_j).$$

3. Алгоритмы решения задачи классификации районов по уровню безработицы во Вьетнаме методом кластерного анализа

На рисунке 1 показан алгоритм решения задачи классификации районов по уровню безработицы.

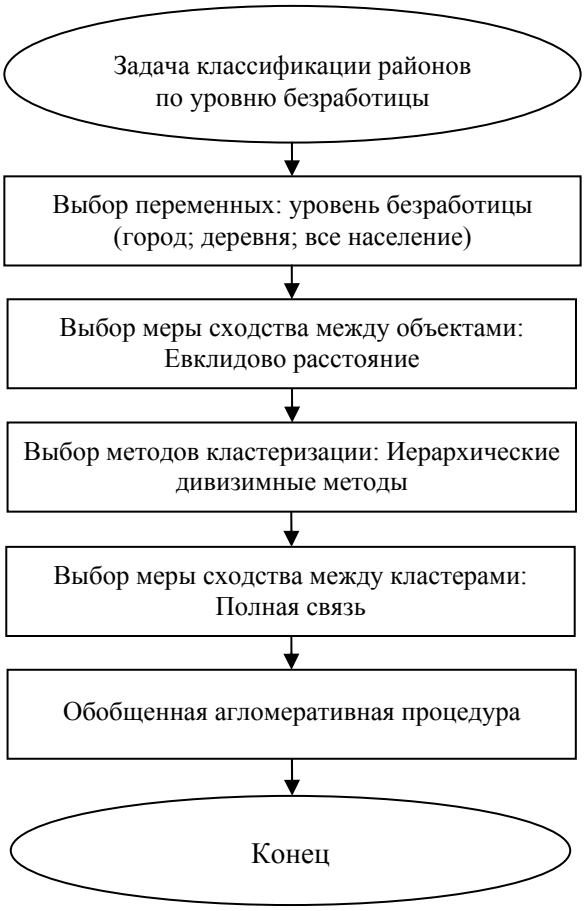

Рис. 1. Алгоритмы решения задачи классификации методом кластерного анализа

На первом шаге каждый объект определяется к отдельному классу. На следующем шаге объединяются два самых близких объекта, которые составляют новый кластер, рассчитываются расстояния от этого кластера до всех остальных объектов, размерность матрицы расстояний D уменьшается.

На p-м шаге повторяется та же процедура на матрице $D_{(n-p)(n-p)}$.

4. Классификация районов по уровню безработицы во Вьетнаме методом кластерного анализа с использованием программного обеспечения R-studio

Программное обеспечение (ПО) R-studio позволяет последовательно анализировать несколько наборов данных. Модуль Hierarchical clustering программы R-studio является наиболее подходящим для выполнения задачи классификации [4].

Рассмотрим процедуру решения методом кластерного анализа в системе.

На рисунке 2 представлены данные по уровню безработицы во Вьетнаме: общей, в городе и деревне за три года (2015, 2016 и 2017). Задача состоит в том, чтобы распределить объекты по однородным группам и установить качественные взаимосвязи между группами с близкими значениями показателей.

Районы	Всего в 2017	Город в 2017	Деревня в 2017	Всего в 2016	Город в 2016	Деревня в 2016	Всего в 2015	Город в 2015	Деревня в 2015
район равниной Красной реки	2.2	3.19	1.64	2.24	3.23	1.73	2.42	3.42	1.94
район северо-запада	1.01	2.71	0.68	1.17	3.2	0.77	1.1	3.11	0.72
южные и северный районы центрального	2.54	4	1.98	2.78	4.3	2.17	2.71	4.51	2.05
плато ТэйНгуен	1.05	1.98	0.7	1.24	2.19	0.88	1.03	2.27	0.57
район юго-востока страны	2.68	2.83	2.43	2.46	2.61	2.19	2.74	3.05	2.17
район равниной реки Меконг	2.88	3.63	2.64	2.89	3.73	2.62	2.77	3.22	2.63

Рис. 2. Данные по уровню безработицы во Вьетнаме за 2015, 2016 и 2017 годы

После процедуры нормирования данные выглядят следующим образом (см. рис. 3):

```
> data_1.scaled <- scale(data_1)
> data_1.scaled
           V1         V2          V3         V4          V5           V6         V7          V8          V9
[1,]  0.1690456  0.1860799 -0.04558273  0.1460886  0.02644707  0.004421836  0.3499566  0.21575529  0.3107239
[2,] -1.2678423 -0.4838078 -1.18713287 -1.2749555 -0.01322353 -1.269067000 -1.2338470 -0.21116475 -1.1472884
[3,]  0.5795850  1.3165155  0.35871627  0.8632511  1.44136524  0.588104219  0.6979135  1.71686121  0.4421841
[4,] -1.2195435 -1.5025954 -1.16335057 -1.1819900 -1.34880050 -1.123146404 -1.3178366 -1.36798032 -1.3265522
[5,]  0.7486307 -0.3163359  0.89381790  0.4382659 -0.79341206  0.614635237  0.7339090 -0.29379443  0.5855951
[6,]  0.9901244  0.8001437  1.14353199  1.0093398  0.68762378  1.185052111  0.7699045 -0.05967699  1.1353374
attr(,"scaled:center")
      V1       V2       V3       V4       V5       V6       V7       V8       V9
2.060000 3.056667 1.678333 2.130000 3.210000 1.726667 2.128333 3.263333 1.680000
attr(,"scaled:scale")
       V1        V2        V3        V4        V5        V6        V7        V8        V9
0.8281787 0.7165380 0.8409618 0.7529675 0.7562275 0.7538346 0.8334367 0.7261313 0.8367556
```

Рис. 3. Нормированные данные

Построим вертикальную древовидную дендрограмму, фрагмент программного кода представлен на рисунке 4.

```
data_1·<-·read.delim("text1.txt",·header·=·F)¶

data_1.scaled·<-·scale(data_1)¶

data_1.scaled¶

hclust.data_1·<-·hclust(dist(data_1.scaled),·method·=·"complete")¶

plot(hclust.data_1,·main·=·"Дендрограмма·делимых·классов·методом·«полных·связей»")¶

abline(h·=·3,·col·=·"red")¶

cutree(hclust.data_1,·h·=·3)¶
```

Рис. 4. Фрагмент программного кода

На рисунке 5 показана дендрограмма, по горизонтальной оси представлены наблюдения, по вертикальной – расстояния объединения. На дендрограмме 1, 2, 3, 4, 5, 6 соответствуют районам равнины Красной реки, южным и северным районам центрального Вьетнама, плато Тэйнгуен; району юго-востока страны и району равниной реки Меконг.

Рис. 5. Дендрограмма делимых классов методом полных связей

На первом шаге были объединены объекты 2 и 4, как имеющие минимальное расстояние, а на последнем – ранее объединенные в какие-либо кластеры. Далее определяем, сколько этапов следует выполнить, чтобы, исходя из анализа дендрограммы, считать полученную классификацию окончательной. В результате получили число кластеров $K = 3$. Первый кластер будет состоять из 2 и 4, второй кластер объединяет 1 и 5, третий – 3 и 6. Результаты объединения представлены в таблице 2.

Таблица 2

Объединение классов методом полной связи

Номер класса	Кол-во объектов в классе	Состав классов
C_{11}	2	район северо-запада, плато Тэйнгуен
C_{12}	2	район равнины Красной реки, район юго-востока страны
C_{13}	2	южные и северный районы центрального Вьетнама, район равниной реки Меконг

С точки зрения рынка труда первый кластер является наиболее стабильным, экономически активным и производящим больше продукции. В этот класс вошли следующие районы: плато Тэйнгуен, район северо-запада.

Второй кластер с точки зрения рынка труда является относительно стабильным. Этому кластеру принадлежат район равниной Красной реки и район юго-востока страны.

Третий кластер для рынка труда является наименее развитым. В этот кластер вошли следующие районы: южные и северный районы центрального Вьетнама, район равниной реки Меконг.

Опираясь на полученные результаты, можно рекомендовать правительству Вьетнама проводить политику инвестирования в районах 3-го и 2-го кластеров для снижения безработицы, ускорения экономического развития и улучшения качества жизни населения.

Заключение

В данной работе была исследована возможность применения метода кластерного анализа для проведения исследования уровня безработицы во Вьетнаме.

Показано, что задача классификации районов по уровню безработицы во Вьетнаме хорошо решается методом кластерного анализа с применением программного обеспечения “R-studio”. Полученные результаты имеют важное значение для развития рынка труда во Вьетнаме.

Список литературы

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. 607 с.

2. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. С. 176–180.

3. Статистические данные для исследования уровня безработицы во Вьетнаме: сайт. URL: https://www.gso.gov.vn/default.aspx?tabid=714 (дата обращения: 13.05.2020).

4. Савельев А.А., Мухарамова С.С., Пилюгин А.Г. Учебно-методическое пособие: Использование языка R для статистической обработки данных. М.–Казань, 2007. 10 с.

5. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: Высшая школа, 2000. 450 с.