Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

О.С. Чудинова

РАСЩЕПЛЕНИЕ СМЕСИ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ В ПАКЕТЕ STATISTICA И СРЕДЕ RSTUDIO

Методические указания

Рекомендовано к изданию редакционно-издательским советом федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет» для обучающихся по образовательным программам высшего образования по направлениям подготовки 01.03.04 Прикладная математика, 38.03.05 Бизнес-информатика, 38.04.01 Экономика

Оренбург
2018

УДК 519.22(076.5)
ББК 22.172я7
Ч 84

Рецензент – доцент, кандидат экономических наук О.Н. Яркова

Чудинова, О.С.

Ч84 Расщепление смеси вероятностных распределений в пакете Statistica и среде RStudio: методические указания / О.С. Чудинова; Оренбургский гос. ун-т.– Оренбург: ОГУ, 2018. – 38 с.

Методические указания содержат рекомендации к выполнению лабораторной работы по курсу «Анализ данных» на тему «Расщепление смеси вероятностных распределений». Могут использоваться для самостоятельной работы студентов, в том числе для выполнения индивидуальных заданий и курсовых работ, связанных решением задачи многомерной классификации.

Методические указания предназначены для обучающихся по направлениям подготовки 01.03.04 Прикладная математика, 38.03.05 Бизнес-информатика, 38.04.01 Экономика (профиль «Математические и инструментальные методы анализа социальных и экономических процессов») всех формы обучения.

УДК 519.22(076.5)
ББК 22.172я7

Содержание

Введение

Известным в литературе приложением задачи расщепления смеси вероятностных распределений в области анализа социально-экономических процессов является параметрическое структурное моделирование среднедушевого дохода населения, инновационного потенциала региона, интеграционной активности российских компаний различных секторов экономики [1, 2]. В этих примерах для описания рассматриваемых скалярных случайных величин используется плотность распределения смеси конечного числа логарифмически нормальных генеральных совокупностей. Обоснование используемых моделей дается исходя из свойств логнормального закона и геометрических соображений, а расщепление смеси осуществляется на основе экспертно-статистического подхода с использованием метода максимального правдоподобия для оценки параметров распределения компонент смеси. В работах не уделяется внимания инструментальным средствам решения задачи расщепления смеси вероятностных распределений, использование которых становится весьма актуальным в общей постановке задачи, когда рассматривается многомерный случай и требуется оценить количество компонент смеси.

В методических указаниях рассматриваются теоретические аспекты решения задачи расщепления смеси вероятностных распределений с помощью распространенного в анализе данных (Data Mining) EM-алгоритма, а также практические аспекты реализации EM-алгоритма в статистическом пакете прикладных программ Statistica и современной бесплатной интегрированной среде разработки для R. R – это язык программирования для статистического анализа, предсказания и визуализации данных; постоянно развивающаяся интерактивная программа с открытым кодом. Много полезной информации по работе с системой статистических вычислений R, примеры анализа и визуализации данных с помощью R, переводы документации по этой программе и новости из мира R можно найти на сайте http://r-analytics.blogspot.ru/.

Выполнение лабораторной работы на тему «Расщепление смеси вероятностных распределений» по дисциплине «Анализ данных», относящейся к обязательным дисциплинам (модулям) вариативной части блока 1 «Дисциплины (модули)», направлено на формирование у обучающихся по направлению подготовки 01.03.04 Прикладная математика следующих общепрофессиональных и профессиональных компетенций:

ОПК-1 – готовность к самостоятельной работе;

ОПК-2 – способность использовать современные математические методы и современные прикладные программные средства и осваивать современные технологии программирования;

ПК-1 – способность использовать стандартные пакеты прикладных программ для решения практических задач на электронных вычислительных машинах, отлаживать, тестировать прикладное программное обеспечение;

ПК-9 – способность выявить естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности, готовность использовать для их решения соответствующий естественнонаучный аппарат;

ПК-10 – готовность применять математический аппарат для решения поставленных задач, способность применить соответствующую процессу математическую модель и проверить ее адекватность, провести анализ результатов моделирования, принять решение на основе полученных результатов;

ПК-11 – готовность применять знания и навыки управления информацией;

ПК-12 – способность самостоятельно изучать новые разделы фундаментальных наук.

1 Постановка и решение задачи расщепления смеси вероятностных распределений

Пусть плотность распределения смеси имеет вид:

$$p_{\xi}(x)=\sum_{j=1}^{s}\pi_j p_{\xi_j}(x),\ \sum_{j=1}^{s}\pi_j=1,\ \pi_j\geq 0, \tag{1.1}$$

где $p_{\xi_j}(x)$ – плотность распределения *j*-ой компоненты смеси;

π_j – априорная вероятность *j*-ой компоненты смеси;

s – количество компонент смеси.

Будем считать, что плотности распределения компонент смеси принадлежат одному и тому же одномодальному параметрическому семейству распределения $\varphi(x;\theta)$ и отличаются только значениями параметра, т.е. $p_{\xi_j}(x)=\varphi(x;\theta_j)$, $j=\overline{1,s}$.

Задача расщепления (разделения) смеси заключается в том, чтобы, имея выборку *n* случайных и независимых наблюдений из смеси $p_{\xi}(x)$, зная число компонент смеси *s* и функцию φ, оценить вектор параметров $\Theta=(\pi_1,...,\pi_s,\theta_1,...,\theta_s)$ [3].

Попытка разделить смесь, непосредственно используя принцип максимума правдоподобия, приводит к громоздкой оптимизационной задаче. Обойти эту трудность позволяет EM-алгоритм (expectation-maximization, алгоритм максимизации ожидания). Идея алгоритма состоит в построении вспомогательной матрицы скрытых переменных *G*, обладающей двумя свойствами. Во-первых, она может быть вычислена, если известны значения вектора параметров Θ. Во-вторых, известные значения скрытых переменных значительно упрощают поиск максимума правдоподобия.

EM-алгоритм состоит из итерационного повторения двух шагов. На E-шаге вычисляется ожидаемое (expectation) значение матрицы скрытых переменных *G* по

текущему приближению вектора параметров Θ. На М-шаге решается задача максимизации (maximization) функции правдоподобия и находится следующее приближение вектора Θ по текущим значениям G и Θ.

Рассмотрим Е-шаг. Пусть $G = \{g_{ij}\}_{\substack{i=\overline{1,n} \\ j=1,s}}$ – матрица скрытых переменных. В качестве скрытых переменных g_{ij} выступают апостериорные вероятности $\pi_j(x^{(i)})$ того, что i-ый объект $x^{(i)}$ принадлежит j-ой компоненте смеси. Поскольку каждый объект обязательно принадлежит какой-то компоненте, то $\sum_{j=1}^{s} g_{ij} = 1 \ \ \forall i = \overline{1,n}$. Зная $\Theta = (\pi_1,...,\pi_s,\theta_1,...,\theta_s)$, переменные g_{ij} вычисляются по формуле Байеса:

$$g_{ij} = \frac{\pi_j p_{\xi_j}(x^{(i)})}{\sum_{l=1}^{s} \pi_l p_{\xi_l}(x^{(i)})}, \ i = \overline{1,n}, \ j = \overline{1,s}. \tag{1.2}$$

На М-шаге максимизируется логарифм функции правдоподобия

$$Q(\Theta) = \ln \prod_{i=1}^{n} p_{\xi}(x^{(i)}) = \sum_{i=1}^{n} \ln \sum_{j=1}^{s} \pi_j p_{\xi_j}(x^{(i)}) \to \max_{\Theta} \tag{1.3}$$

при ограничении $\sum_{j=1}^{s} \pi_j = 1$.

Решая оптимизационную задачу (1.3) методом множителей Лагранжа, получают:

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} g_{ij}, \ j = \overline{1,s}. \tag{1.4}$$

$$\hat{\theta}_j = \arg\max_{\theta} \sum_{i=1}^{n} g_{ij} \ln \varphi(x^{(i)};\theta_j), \ j = \overline{1,s}. \tag{1.5}$$

Таким образом, М-шаг сводится к вычислению оценок весов компонент смеси π_j как средних арифметических (1.4) и оцениванию параметров θ_j путем решения p

независимых оптимизационных задач (1.5). Итерации останавливаются, когда значения функционала $Q(\Theta)$ (1.3) или скрытых переменных G перестанут существенно изменяться. Удобнее контролировать скрытые переменные, так как они имеют смысл вероятностей и принимают значения из отрезка [0,1].

Качество решения и скорость сходимости алгоритма существенно зависят от начального приближения. Стандартный подход к заданию начальных параметров заключается в выборе значений случайным образом. Другой подход состоит в том, чтобы взять в качестве центров компонент p объектов, максимально удаленных друг от друга. Можно реализовать итерационный процесс несколько раз при различных начальных значениях параметров и затем выбрать наилучшее решение[1].

В Data Mining под EM-алгоритмом понимается итерационный самообучающийся метод кластеризации. Однако в отличие от кластерного анализа, решив задачу расщепления смеси вероятностных распределений, можно осуществлять классификацию новых данных. Базовая идея, лежащая в основе принятия решения, к какому из p классов отнести наблюдение, также как и в дискриминантном анализе, состоит в том, что наблюдение следует отнести к тому классу (к той генеральной совокупности), в рамках которого оно выглядит наиболее правдоподобно. Главное отличие от параметрического дискриминантного анализа состоит в способе оценивания неизвестных параметров распределения классов. Поскольку обучающих выборок нет, то оценивание осуществляется по классифицируемым наблюдениям, т.е. на основе исходной матрицы X типа «объект-свойство».

EM-алгоритм кластеризации может применяться к большим массивам данных (Big Data). К недостаткам EM-алгоритма относятся:

– с ростом количества итераций падает производительность алгоритма;

– алгоритм не всегда находит оптимальные параметры и может застрять в локальном оптимуме, так и не найдя глобальный.

[1] Воронцов К.В. Лекции по статистическим (байесовским) алгоритмам классификации. Режим доступа: http://www.machinelearning.ru/wiki/images/e/ed/Voron-ML-Bayes.pdf

2 Реализация EM-алгоритма в пакетах прикладных программ

Реализацию EM-алгоритма в пакетах прикладных программ рассмотрим на следующем примере. Имеются данные мониторинга социально-экономического положения и состояния здоровья населения Российской Федерации по работающим индивидам, проживающим в Москве и Московской области за 2013 год[2]. Фрагмент исходных данных представлен в таблице А.1. В таблице А.1 используются следующие обозначения:

Obr – профессиональное образование (1 – высшее профессиональное образование; 2 – среднее профессиональное образование; 3 – иначе);

Prof – профессиональная группа (1 – чиновник, руководитель высшего и среднего звена, специалист высшего уровня квалификации; 0 – иначе);

Vozrast – количество полных лет;

Zarplata – среднемесячная заработная плата, рублей.

С учетом рассматриваемых признаков провести анализ дифференциации населения Москвы и Московской области по уровню среднемесячной заработной платы.

2.1 Реализация EM-алгоритма в статистическом пакете прикладных программ Statistica

Для реализации EM-алгоритма в пакете Statistica после ввода исходных данных необходимо выбрать пункты меню Data Mining, Generalized EM & k-Means Cluster Analysis (рисунок 2.1).

[2] «Российский мониторинг экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE)», проводимый Национальным исследовательским университетом "Высшая школа экономики" и ЗАО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии РАН. (Сайты обследования RLMS-HSE: *http://www.cpc.unc.edu/projects/rlms* и *http://www.hse.ru/rlms*)».

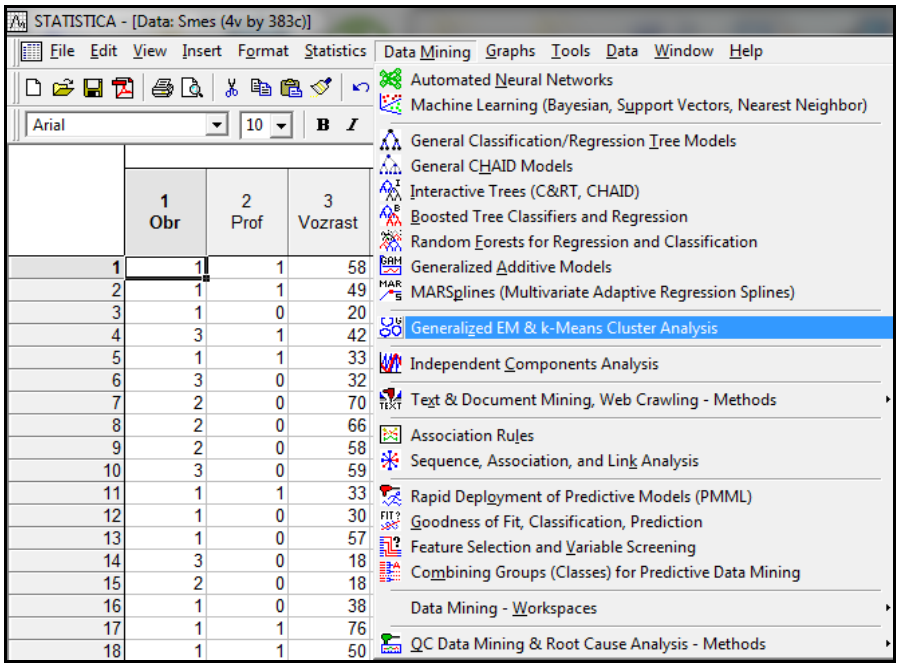

Рисунок 2.1 – Выбор пунктов меню для реализации EM-алгоритма в пакете Statistica

В появившейся форме Generalized Cluster Analysis: RLMS (рисунок 2.2) на странице Quick выбрать EM-алгоритм (EM) и с помощью кнопки Variables (рисунок 2.3) указать признаки для анализа. В левом окне выбираются качественные признаки, в правом – количественные.

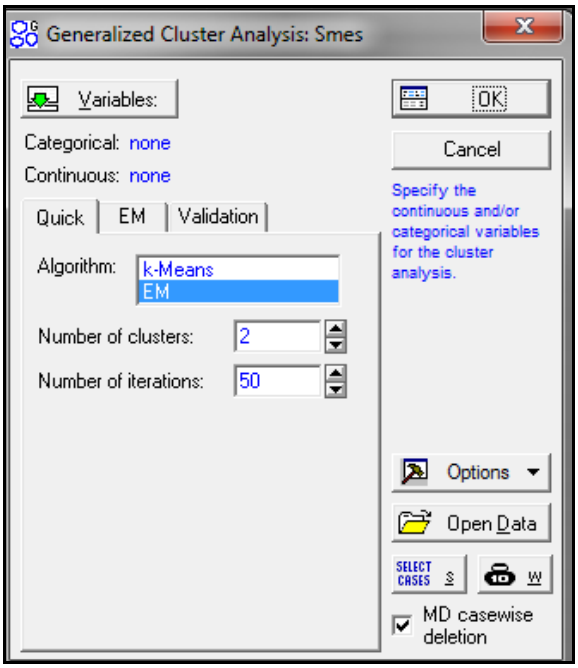

Рисунок 2.2 – Форма Generalized Cluster Analysis: RLMS

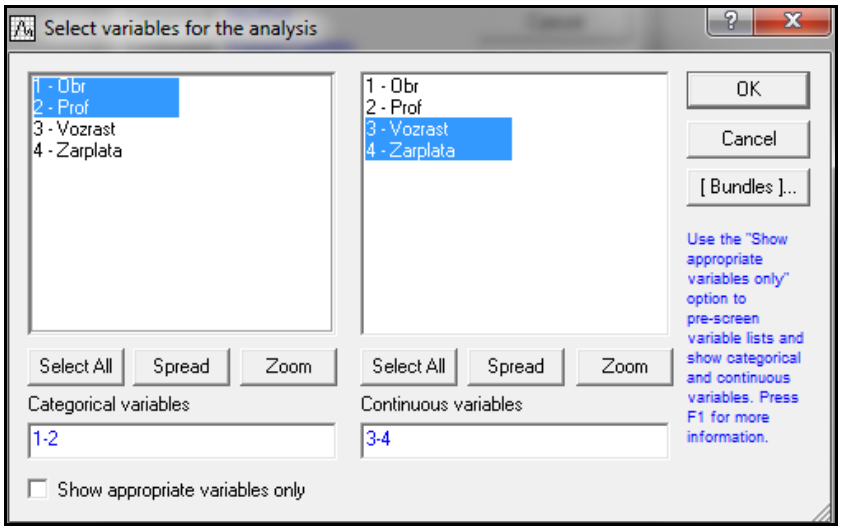

Рисунок 2.3 – Выбор признаков для анализа

На странице EM формы Generalized Cluster Analysis: RLMS (рисунок 2.4) с помощью кнопки Select distributions необходимо указать закон распределения количественных признаков.

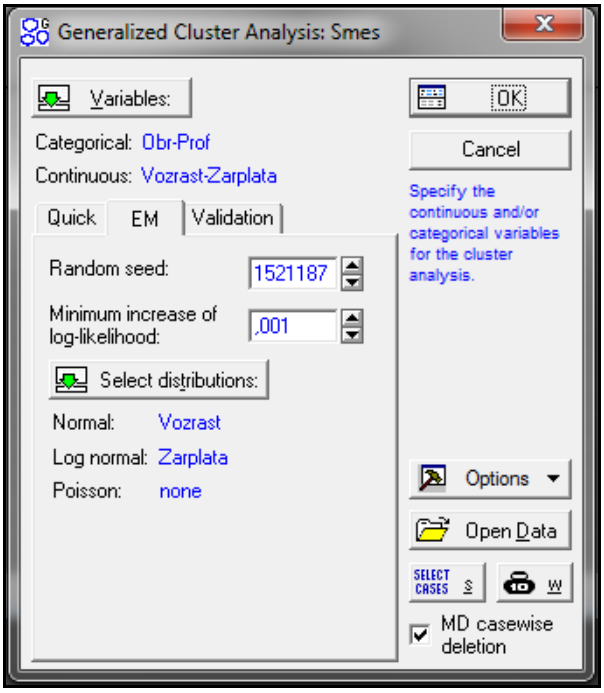

Рисунок 2.4 – Страница EM формы Generalized Cluster Analysis: RLMS

В пакете Statistica предусмотрены следующее законы распределения количественных признаков: нормальный, логнормальный, Пуассона. Учитывая результаты проверки гипотез о законе распределения (рисунки 2.5-2.6), для

признака «Возраст» укажем нормальный закон распределения, для признака «Заработная плата» – логнормальный (рисунок 2.7).

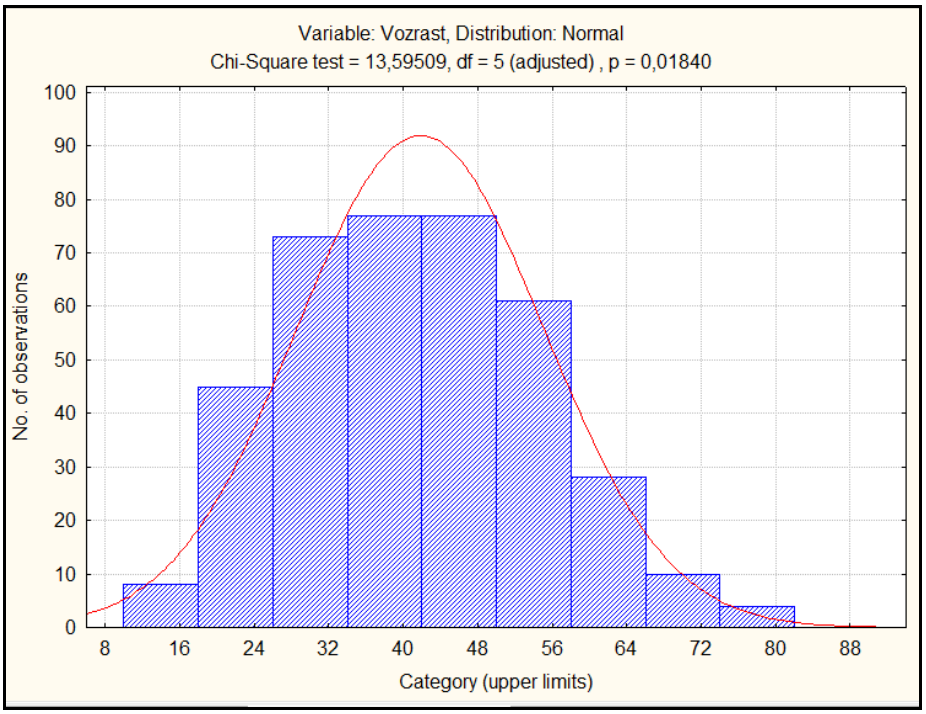

Рисунок 2.5 – Результаты проверки гипотезы о нормальном законе распределения признака «Возраст»

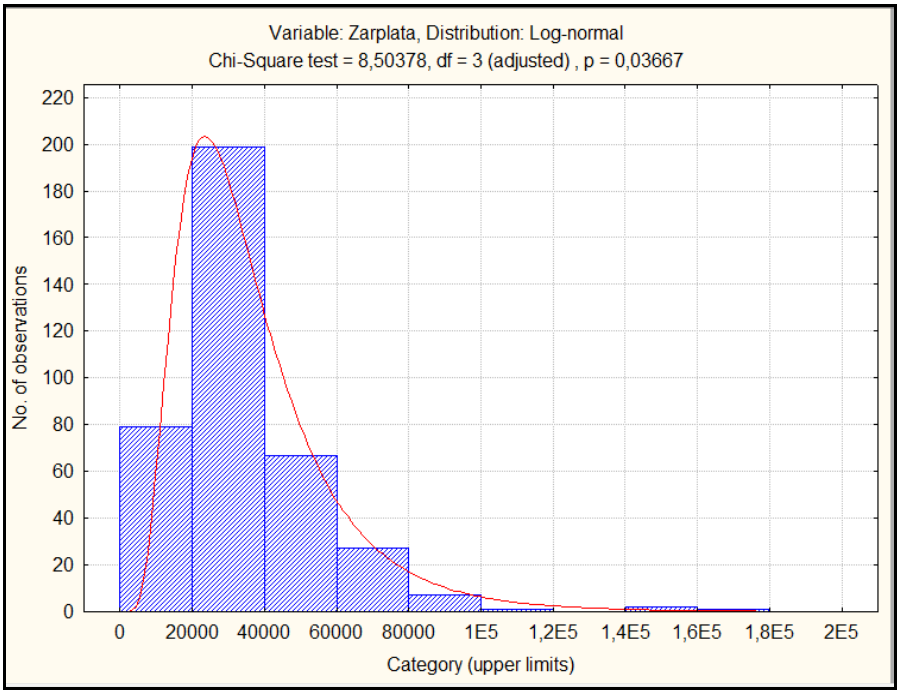

Рисунок 2.6 – Результаты проверки гипотезы о логнормальном законе распределения признака «Среднемесячная заработная плата»

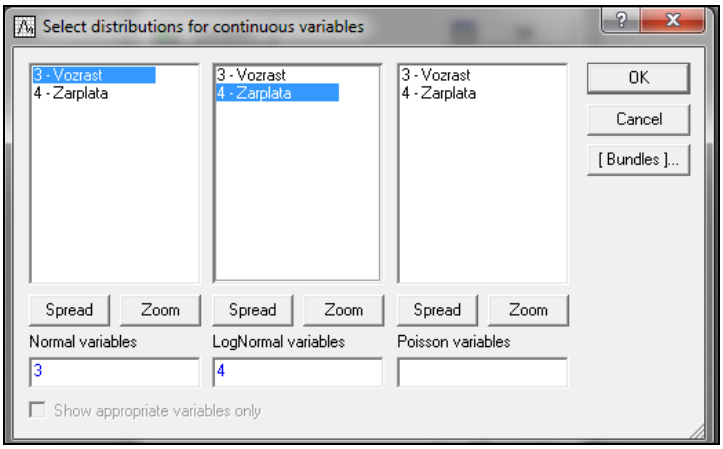

Рисунок 2.7 – Выбор закона распределения признаков

Если количество классов известно, то это значение указывается в поле Number of clasters на странице Quick, если не известно, то можно поставить галочку в поле V-fold cross-validation на странице Validation (рисунок 2.8), указав тем самым автоматический способ оценки количества классов на основе V-кратной кросс-проверки. Суть V-кратной кросс-проверки состоит в том, что исходные данные разбиваются на заданное пользователем количество V подвыборок (значение вводится в поле v value) примерно одинакового объема. EM-алгоритм реализуется V раз, причем каждый раз поочередно одна из подвыборок не участвует в анализе, а используется в качестве тестовой для кросс-проверки. Результаты EM-алгоритма, полученные на основе тренировочных данных, используются для вычисления логарифма функции правдоподобия для тестовых данных. В пакете Statistica для удобства логарифм функции правдоподобия умножается на -2, поэтому чем меньше полученное значение, тем лучше. Затем эти результаты агрегируются (усредняются) по всем V блокам и получают характеристику качества построенной модели. Сначала эта характеристика рассчитывается для наименьшего количества классов, введенного в поле Minimum number of clusters. Затем количество кластеров увеличивается на единицу и так далее продолжается до тех пор, пока процентное снижение рассматриваемой характеристики больше значения, введенного в поле Smallest percentage decrease, и количество классов не превосходит числа Maximum number of clusters.

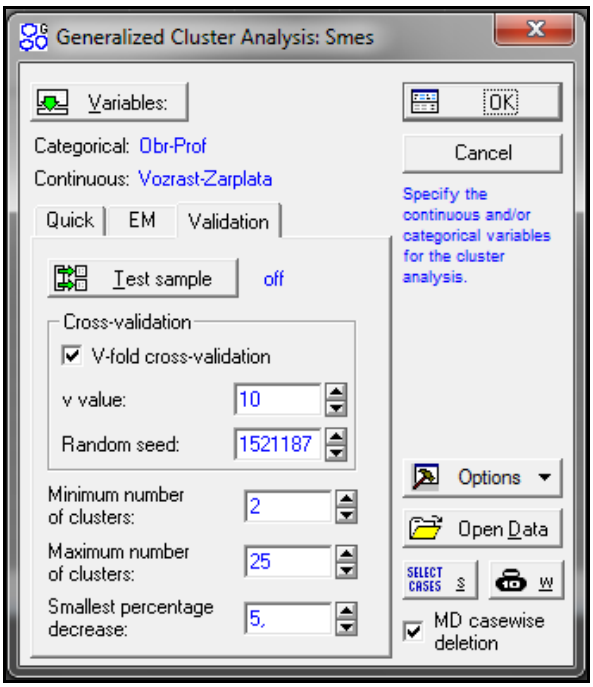

Рисунок 2.8 – Страница Validation

Поле Random seed на страницах EM и Validation заполняется автоматически и предназначено для инициализации генератора случайных чисел, с помощью которого задаются начальные значения параметров модели. Следует отметить, что алгоритм чувствителен к заданию начальных параметров, поэтом при разных запусках программы на одном и том же наборе данных могут получаться разные результаты. Рекомендуется решить задачу несколько раз и выбрать лучшее решение.

С помощью кнопки Test sample можно указать наблюдения, участвующие в анализе (рисунок 2.9). Остальные наблюдения (Test set) будут использоваться для тестирования модели.

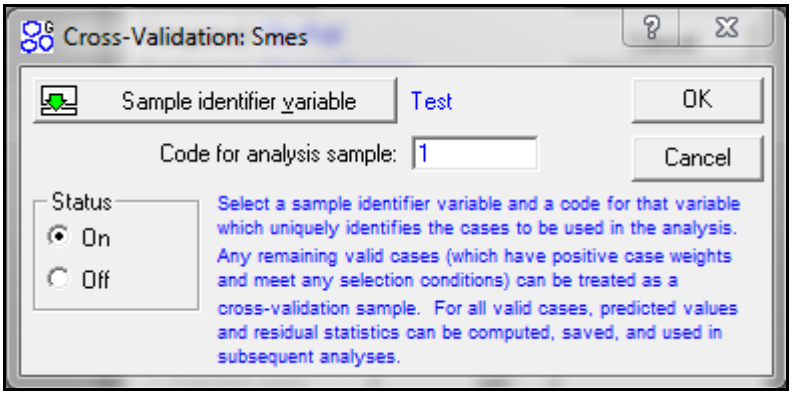

Рисунок 2.9 – Вид формы для формирования обучающей и тестовой выборок

Форма для вывода результатов EM-алгоритма получена после нажатия кнопки ОК и представлена на рисунке 2.10. В анализе участвовали все 383 наблюдения, логарифм функции правдоподобия составил -16,70, рекомендуемое количество классов равно двум.

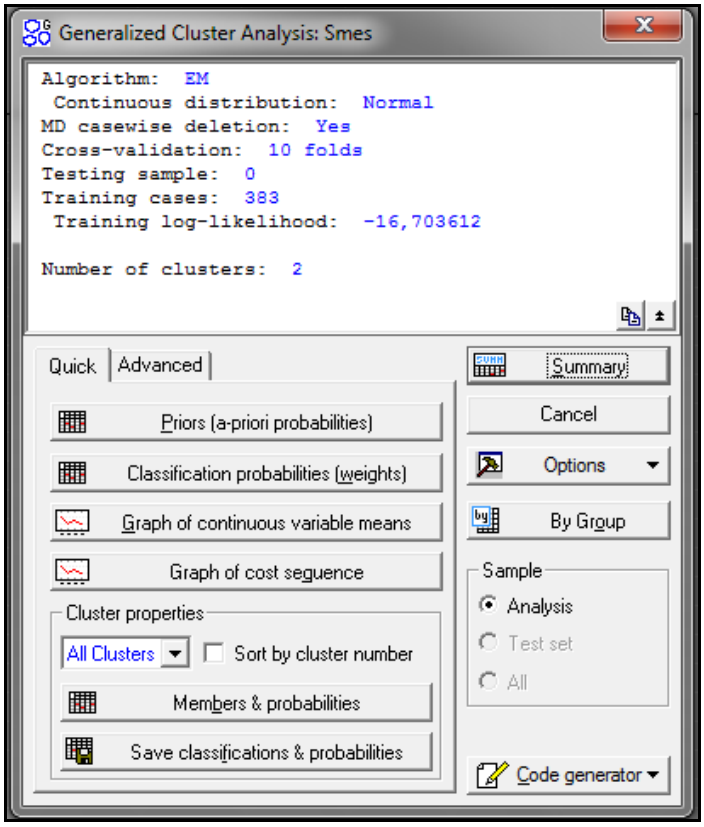

Рисунок 2.10 – Форма для вывода результатов реализации EM-алгоритма

Оценки априорных вероятностей (удельных весов классов) можно вывести на экран с помощью кнопки Priors (a-priori probabilities) (рисунок 2.11).

Priors (a-priori probabilities) for EM clustering (Smes)
Number of clusters: 2
Total number of training cases: 383

Cluster	PRIOR
1	0,448603
2	0,551397

Рисунок 2.11 – Результаты оценки удельных весов классов

Таким образом, оценки удельных весов классов составили: $\hat{\pi}_1 = 0{,}45$, $\hat{\pi}_2 = 0{,}55$.

Кнопка Classification probabilities (weights) предназначена для вывода оценок апостериорных вероятностей отнесения наблюдений к каждому из классов (рисунок 2.12).

	Classification probabilities (weights) fo Number of clusters: 2 Total number of training cases: 383		
	Cluster 1	Cluster 2	Final classification
1	0,966425	0,033575	1
2	0,977735	0,022265	1
3	0,303553	0,696447	2
4	0,654753	0,345247	1
5	0,992706	0,007294	1
6	0,011726	0,988274	2
7	0,044717	0,955283	2
8	0,046212	0,953788	2
9	0,042566	0,957434	2
10	0,021147	0,978853	2
11	0,974401	0,025599	1
12	0,237962	0,762038	2
13	0,402078	0,597922	2
14	0,007049	0,992951	2
15	0,015662	0,984338	2
16	0,329944	0,670056	2
17	0,977656	0,022344	1
18	0,984402	0,015598	1
19	0,665285	0,334715	1
20	0,911387	0,088613	1
21	0,995142	0,004858	1
22	0,021580	0,978420	2
23	0,968556	0,031444	1
24	0,965798	0,034202	1
25	0,012639	0,987361	2
26	0,013671	0,986329	2
27	0,509004	0,490996	1

Рисунок 2.12 – Фрагмент результатов оценки апостериорных вероятностей

С помощью кнопки Graph of continuous variables means на экран выводится график средних значений количественных признаков (рисунок 2.13). Кнопка Graph of cost sequence предназначена для анализа графика изменения логарифма функции правдоподобия при оценке количества классов на основе V-кратной кросс-проверки (рисунок 2.14). По рисунку видно, что значение функции для трех классов больше, чем для двух, поэтому в качестве оценки количества классов выбрано 2.

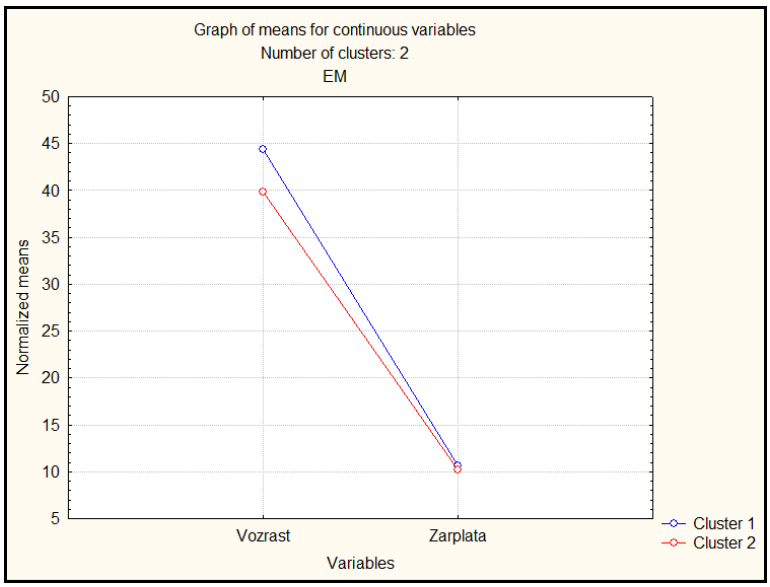

Рисунок 2.13 – График нормализованных средних значений возраста и зарплаты в классах

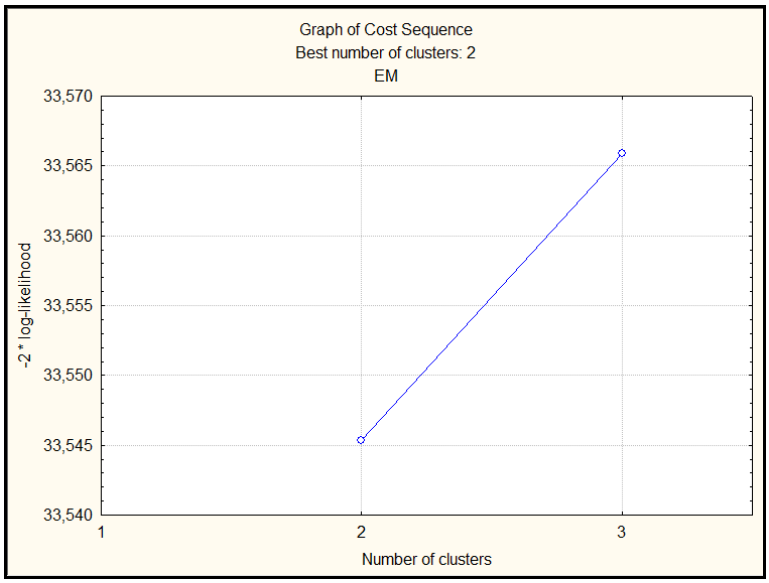

Рисунок 2.14 – График логарифма функции правдоподобия в зависимости от количества классов

С помощью кнопки Members & probabilities можно получить сводную таблицу результатов классификации объектов наблюдения (рисунок 2.15). В первом столбце (Final classification) выводится номер класса, к которому относится объект, в последнем столбце (Probability) – оценка вероятности отнесения объекта в этот класс. В остальных столбцах находятся значения признаков, характеризующих объекты наблюдения.

Case No.	Cluster Members (Smes) Number of clusters: 2 Total number of training cases: 383 **Final classification**	Obr	Prof	Vozrast	Zarplata	Probability
1	1	1	1	58,00000	10000,0	0,966425
2	1	1	1	49,00000	23000,0	0,977735
3	2	1	0	20,00000	25000,0	0,696447
4	1	3	1	42,00000	35000,0	0,654753
5	1	1	1	33,00000	60000,0	0,992706
6	2	3	0	32,00000	15000,0	0,988274
7	2	2	0	70,00000	7000,0	0,955283
8	2	2	0	66,00000	5000,0	0,953788
9	2	2	0	58,00000	12000,0	0,957434
10	2	3	0	59,00000	13600,0	0,978853
11	1	1	1	33,00000	29000,0	0,974401
12	2	1	0	30,00000	10000,0	0,762038
13	2	1	0	57,00000	12000,0	0,597922
14	2	3	0	18,00000	15000,0	0,992951
15	2	2	0	18,00000	15000,0	0,984338

Рисунок 2.15 – Фрагмент результатов классификации объектов наблюдения

Дополнительные характеристики классов можно получить с помощью кнопок на странице Advanced (рисунок 2.16).

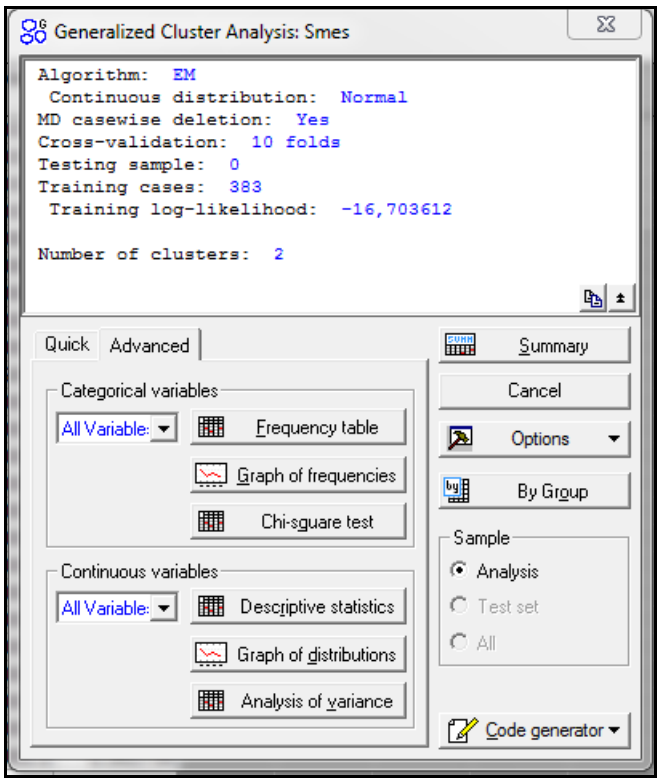

Рисунок 2.16 – Вид страницы Advanced формы с результатами реализации EM-алгоритма

Для количественных признаков (Continuous variables) с помощью кнопки Descriptive statistics осуществляется вывод оценок основных числовых характеристик признаков в классах (рисунки 2.17-2.18).

	Statistics for continuous variable: Vozrast (Smes) Number of clusters: 2 Total number of training cases: 383		
	Cluster 1	Cluster 2	Overall
Minimum	17,00000	17,00000	17,0000
Maximum	77,00000	80,00000	80,0000
Mean	44,60588	39,65728	41,8538
Standard deviation	12,48919	13,55144	176,9838

Рисунок 2.17 – Оценки основных числовых характеристик для признака «Возраст»

	Statistics for continuous variable: Zarplata (Smes) Number of clusters: 2 Total number of training cases: 383		
	Cluster 1	Cluster 2	Overall
Minimum	10000,0	4000,00	4000
Maximum	180000,0	80000,00	180000
Mean	47424,1	28577,46	36943
Standard deviation	26485,5	12919,79	490886434

Рисунок 2.18 – Оценки основных числовых характеристик для признака «Среднемесячная заработная плата»

Сравнивая оценки основных числовых характеристик количественных признаков в классах, можно дать интерпретацию классам. Средний возраст респондентов, отнесенных в первый класс, на 5 лет выше, чем средний возраст респондентов, отнесенных во второй класс, при этом выше и средняя заработная плата (на 18847 рублей). Максимальная же заработная плата в первом классе более чем в 2 раза выше, чем во втором классе. Разброс значений признаков относительно средних показателей для возраста в первом и втором классах почти совпадает, для заработной платы в первом классе выше, чем во втором.

Результаты анализа вариации количественных признаков, полученные с помощью кнопки Analysis of variance, позволяют сделать вывод о том, что каждый из признаков вносит значимый вклад в разделение объектов на классы (p value меньше 0,05) (рисунок 2.19).

	ANOVA for continuous variables (Smes) Number of clusters: 2 Total number of training cases: 383					
	Between SS	df	Within SS	df	F	p value
Vozrast	6,731896E+05	1	6,529258E+04	381	3928,245	0,00
Zarplata	5,562890E+11	1	1,539373E+11	381	1376,834	0,00

Рисунок 2.19 – Результаты анализа вариации количественных признаков в классах

С помощью кнопки Graph of distributions выводятся графики оценок плотности распределения признаков в классах (рисунки 2.20-2.21).

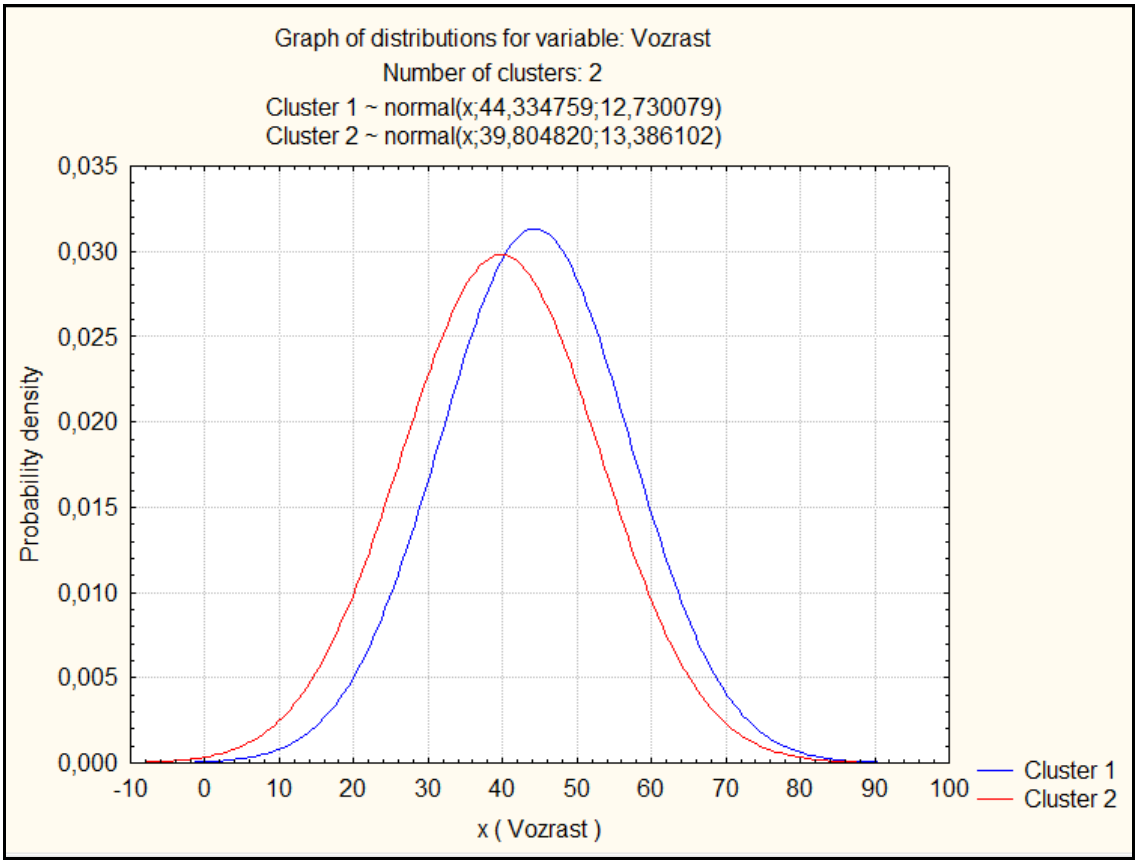

Рисунок 2.20 – Оценка плотности нормального закона распределения признака «Возраст» для первого и второго классов

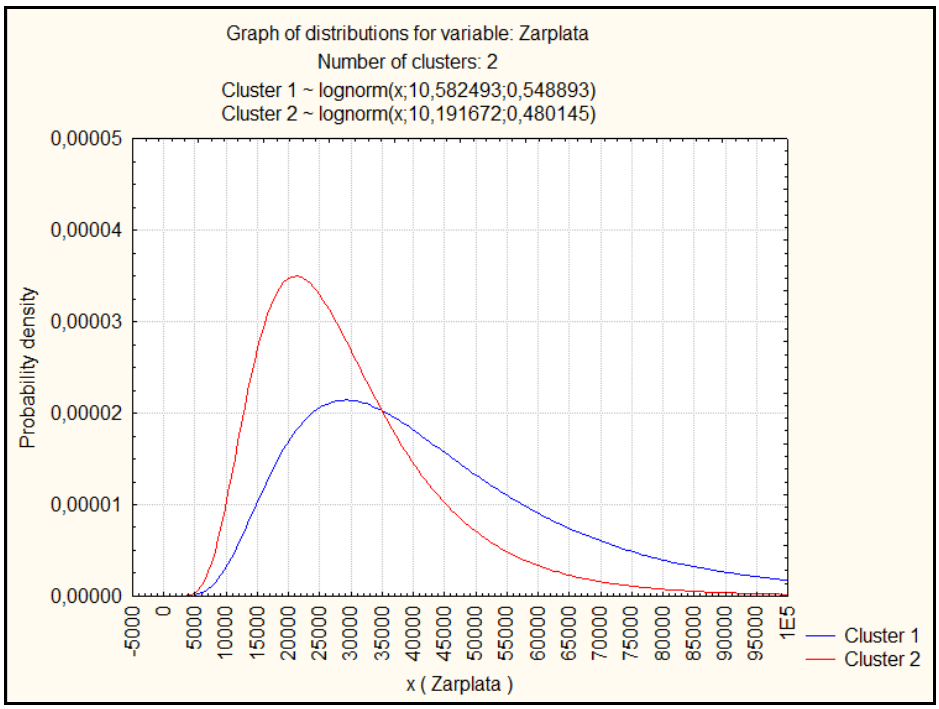

Рисунок 2.21 – Оценка плотности логнормального закона распределения признака «Среднемесячная заработная плата» для первого и второго классов

Для качественных признаков (Categorical variables) можно с помощью кнопки Frequency Table построить таблицы сопряженности с признаком «Номер класса» (рисунки 2.22-2.23).

Frequency table for categorical variable: Obr (Smes)
Number of clusters: 2
Total number of training cases: 383

	Cluster 1	Cluster 2	Total
1	147	33	180
2	12	68	80
3	11	112	123

Рисунок 2.22 – Таблица сопряженности признаков «Образование» и «Номер класса»

Frequency table for categorical variable: Prof (Smes)
Number of clusters: 2
Total number of training cases: 383

	Cluster 1	Cluster 2	Total
0	48	210	258
1	122	3	125

Рисунок 2.23 – Таблица сопряженности признаков «Профессия» и «Номер класса»

Графическое представление таблиц сопряженности можно получить с помощью кнопки Graph of frequencies. Результаты представлены на рисунках 2.24-2.25.

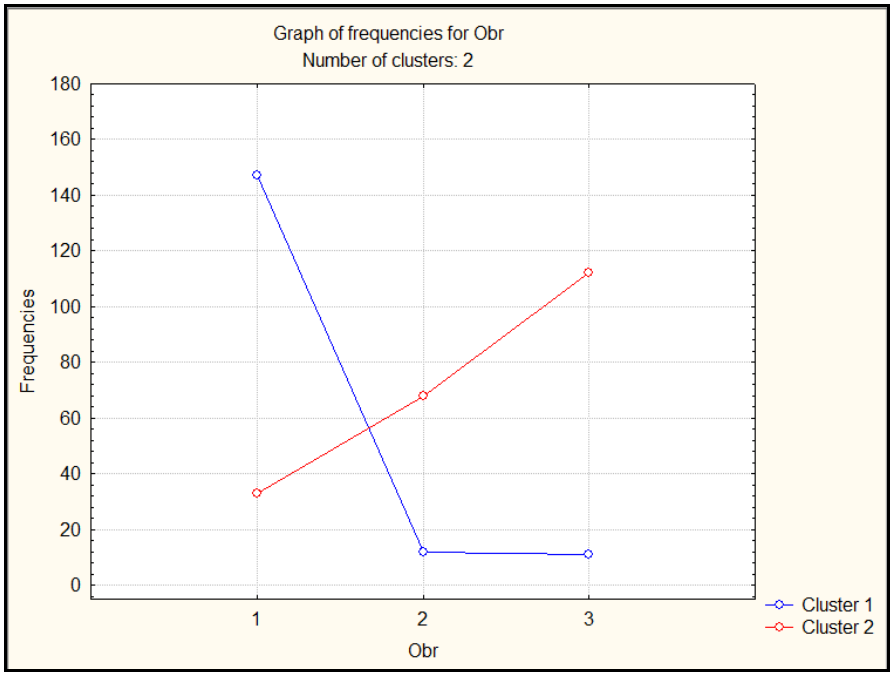

Рисунок 2.24 – Графическое представление таблицы сопряженности признаков «Образование» и «Номер класса»

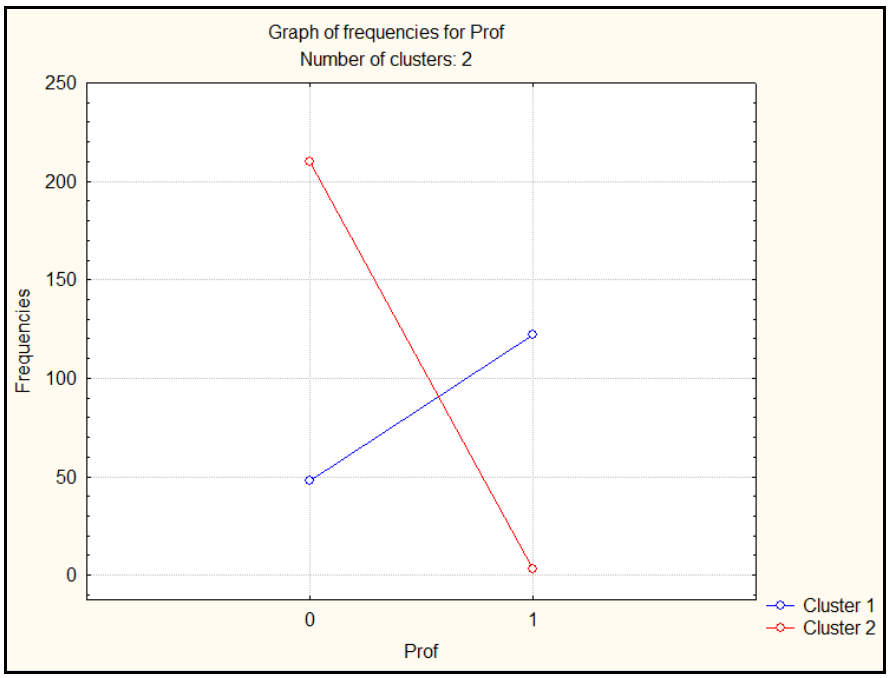

Рисунок 2.25 – Таблица сопряженности признаков «Профессия» и «Номер класса»

Информацию по построению, свойствам, интерпретации и анализу таблиц сопряженности можно найти в работах [4-6].

По таблицам сопряженности можно сделать следующие выводы:

1) из 383 респондентов в первый класс отнесены 170 человек (147+12+11 или 48+122), во второй – 213 человек;

2) среди 383 опрошенных 180 человек имеют высшее профессиональное образование, 80 человек – среднее профессиональное образование, остальные 123 человека имеют начальное профессиональное образование или не имеют профессионального образования;

3) в первом классе наибольший удельный вес принадлежит респондентам с высшим профессиональным образованием (86% или 147 человек). Это 82% среди всех респондентов с высшим образованием;

4) среди 383 опрошенных 125 человек, т.е. 33%, относятся к чиновникам, руководителям высшего и среднего звена, специалистам высшего уровня квалификации;

5) в первом классе наибольший удельный вес принадлежит респондентам, занимающим руководящие должности (72% или 122 человека). Это 98% всех респондентов этой профессиональной группы.

Учитывая интерпретацию классов, полученную на основе количественных признаков, можно сделать вывод, что первый класс респондентов характеризуется более высокой среднемесячной заработной платой, чем второй, и включает преимущественно высококвалифицированных специалистов, занимающих руководящие должности и имеющих большой опыт работы.

Результаты проверки независимости качественных признаков с помощью критерия Хи-квадрат, полученные с помощью кнопки Chi-square test и представленные на рисунке 2.26, доказывают, что между признаками «Образование» и «Номер класса», а также «Профессия» и «Номер класса» существует значимая связь, а, следовательно, оба признака вносят существенный вклад в разделение классов.

	Independence test for categorical variables (Smes) Number of clusters: 2 Total number of training cases: 383				
	df	Chi-square	p value	G-square	p value
Obr	2	191,9265	0,00	212,8717	0,00
Prof	1	212,8644	0,00	249,9006	0,00

Рисунок 2.26 – Результаты проверки гипотез о независимости признаков «Образование» и «Номер класса», «Профессия» и «Номер класса»

2.2 Реализация EM-алгоритма в среде RStudio

Бесплатно скачать инсталляции для установки программы R и среды разработки RStudio можно на следующих сайтах:

1) https://cran.rstudio.com/;
2) https://www.rstudio.com/products/rstudio/download/#download.

После запуска RStudio на экране появятся 4 рабочих окна (рисунок 2.27).

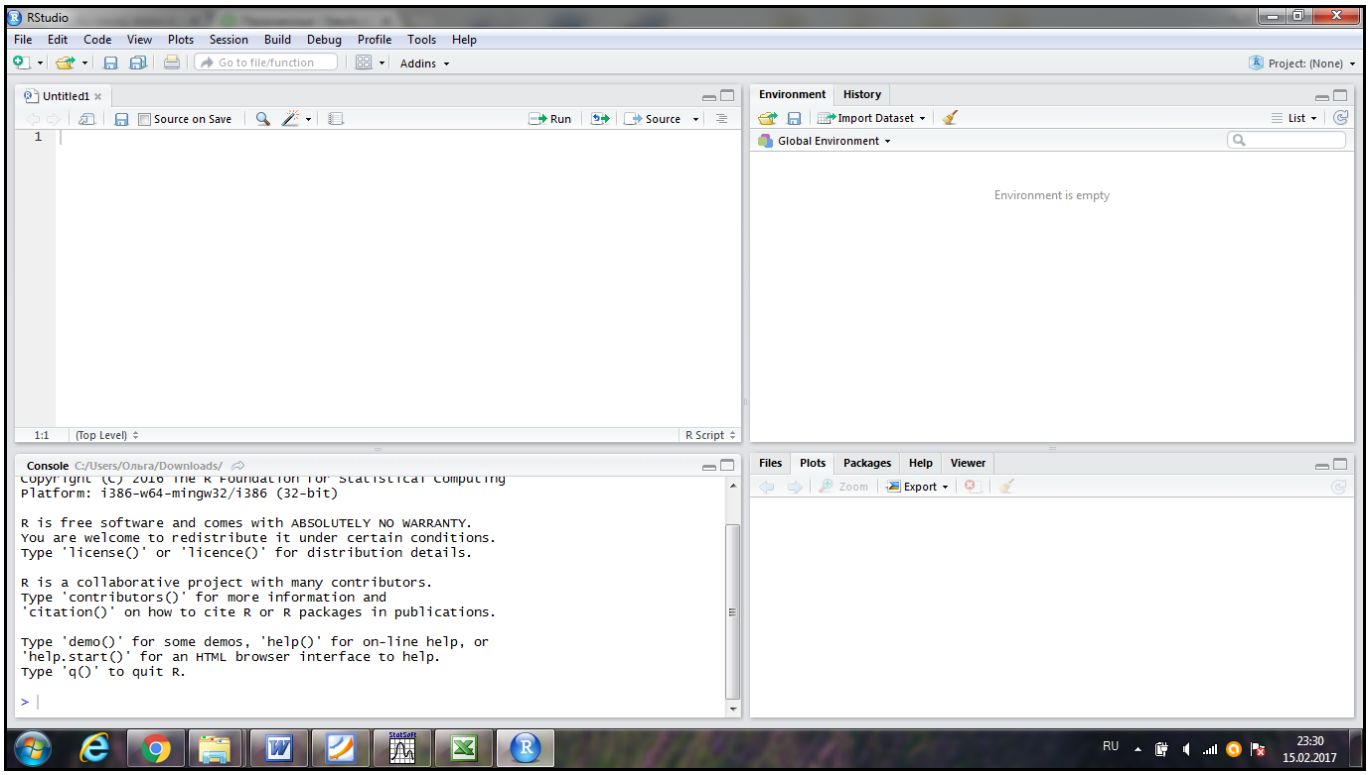

Рисунок 2.27 – Вид экрана после запуска RStudio

В нижнем левом углу находится окно Console, предназначенное для написания команд на языке R. После нажатия Enter в этом окне выдаются результаты выполнения введенной команды. Окно над консолью (редактор скриптов) предназначено для написания последовательности команд, которые можно выполнить с помощью кнопки Run. Окно справа от редактора скриптов состоит из двух вкладок (страниц). На странице Environment отображаются все объекты, созданные в процессе работы; на странице History отображается последовательность команд, выполненная на протяжении рабочей сессии. Нижнее правое окно состоит из пяти страниц. На странице Files отображается менеджер файлов, позволяющий передвигаться по папкам, не покидая RStudio. Страница Plots предназначена для отображения графиков, созданных в программе. На странице Packages отображаются библиотеки, используемые в процессе работы. Для вывода помощи предназначена страница Help.

Работу в среде RStudio начнем с установки рабочей директории. Для этого воспользуемся пунктами меню Session, Working Directory, Choose Directory (рисунок 2.28).

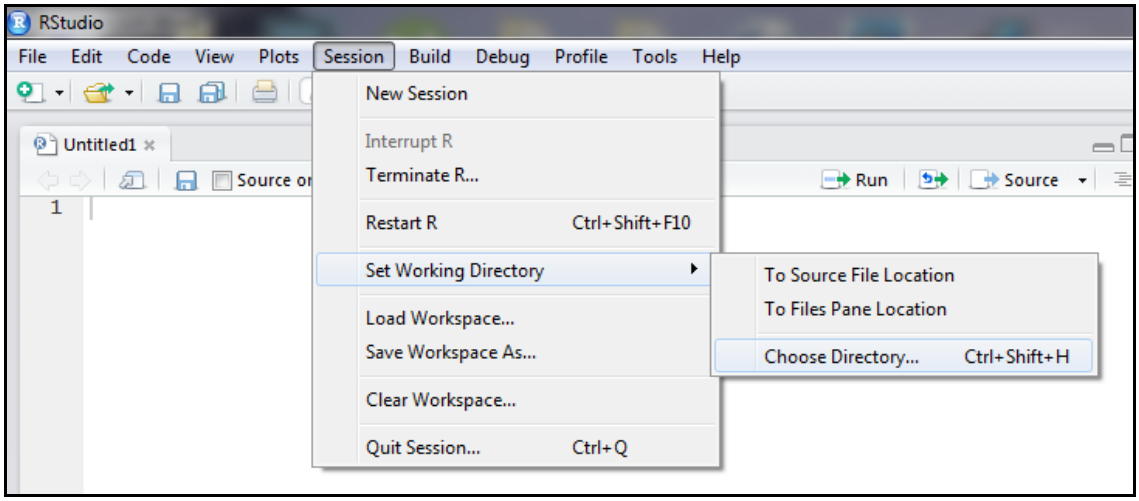

Рисунок 2.28 – Выбор пунктов меню для установки рабочей директории

Для загрузки данных воспользуемся командой read.csv2, позволяющей считать таблицу данных в RStudio из файла типа CSV. Предварительно сохраним файл с

исходными данными в Excel в указанном формате в выбранную рабочую директорию (рисунок 2.29).

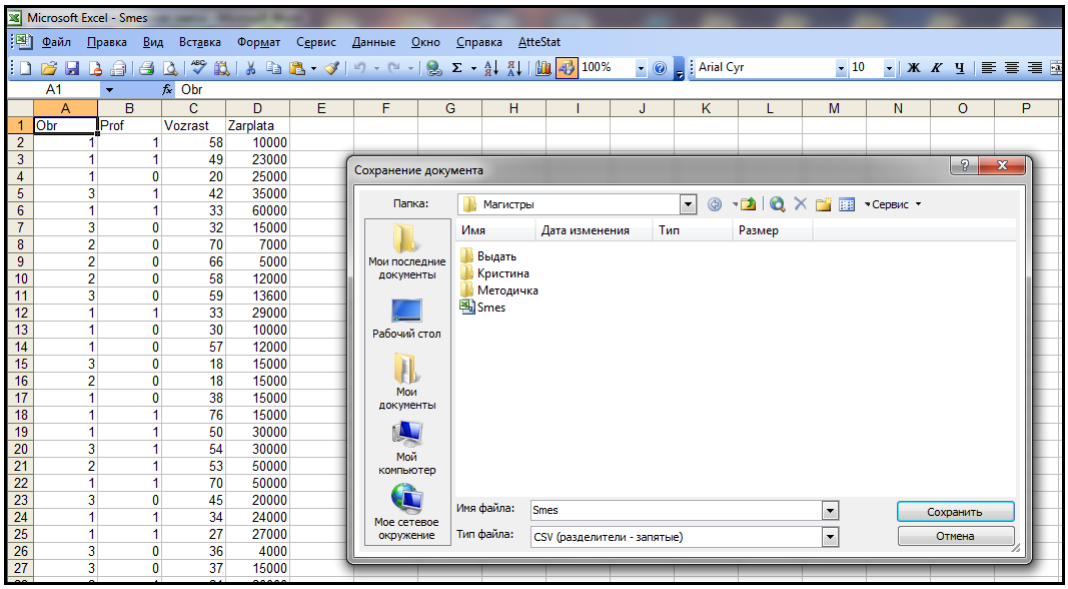

Рисунок 2.29 – Сохранение файла с исходными данными в формате CSV

В окне Console введем команду «x<-read.csv2('smes.csv')», с помощью которой создадим объект «x» типа таблица данных (data frame), в которую считаем данные из файла smes.csv. Созданный объект (x) будет размещен на странице Environment. Нажав кнопку мыши на этом объекте, таблица с данными отобразится в верхнем левом окне (рисунок 2.30).

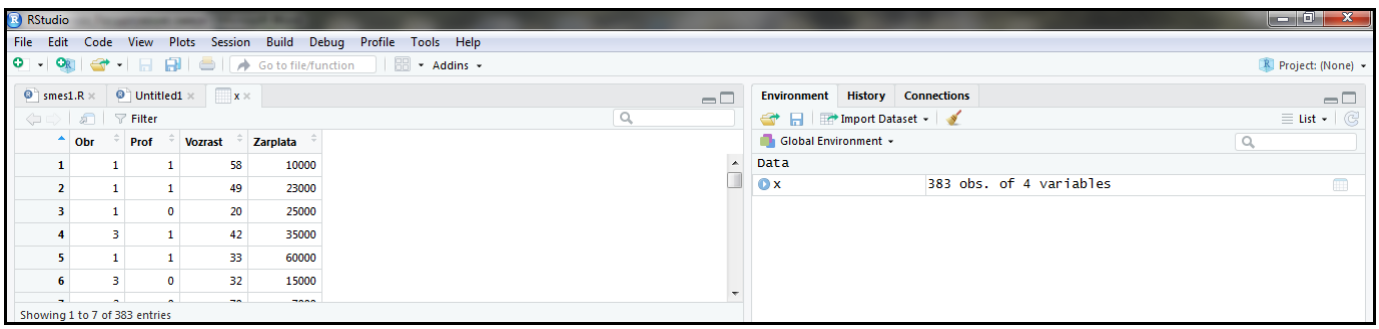

Рисунок 2.30 – Таблица с исходными данными в RStudio

Вид окна Console с выполненными командами представлен на рисунке 2.31.

```
> setwd("C:/Оля/Разное/Дисциплины/Анализ данных/Магистры")
> x<-read.csv2('smes.csv')
> View(x)
>
```

Рисунок 2.31 – Вид окна Console

Для реализации EM-алгоритма в R создана библиотека mclust. Если этот пакет не установлен (его нет в списке установленных пакетов на странице Packages), то необходимо ввести команду install.packages("mclust"). Если библиотека mclust установлена, то её необходимо выбрать (поставить галочку) на странице Packages (рисунок 2.32) или ввести команду library(mclust).

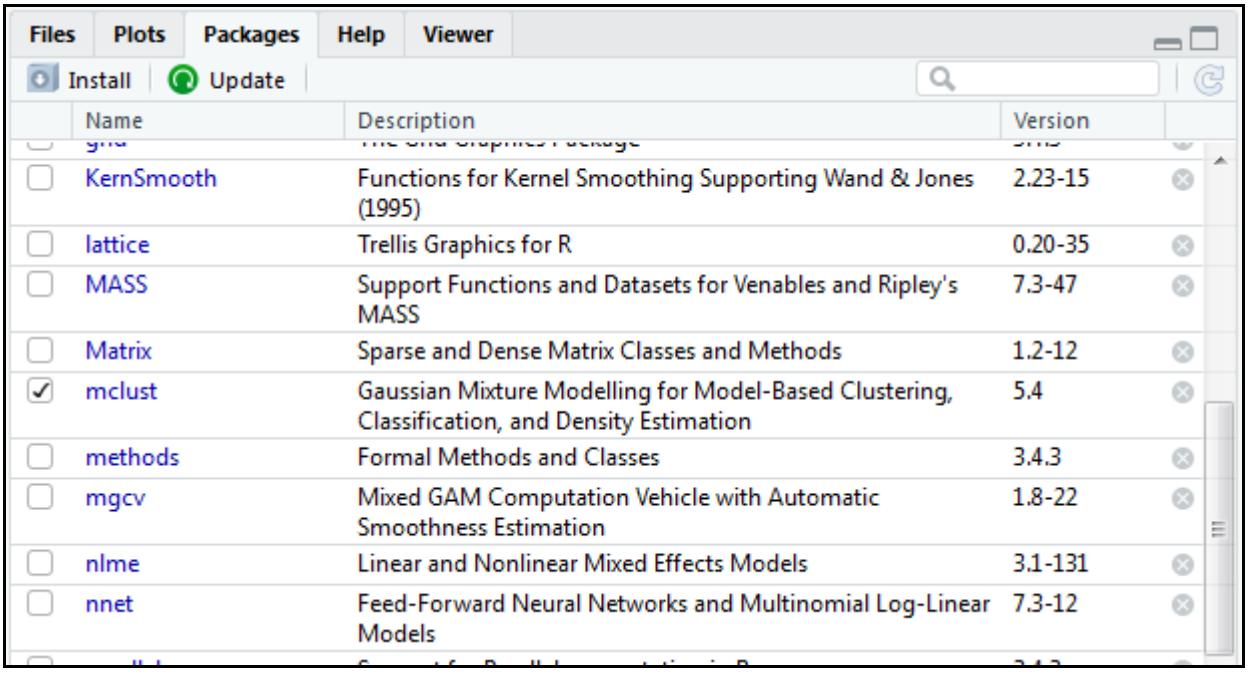

Рисунок 2.32 – Подключение библиотеки mclust в RStudio

Пакет mclust предназначен для решения задачи расщепления смеси конечного числа нормально распределенных генеральных совокупностей, поэтому EM-алгоритм реализуем только для количественных признаков «Возраст» и «Среднемесячная заработная плата». Описание пакета можно найти на официальном

сайте программы (Режим доступа: https://cran.r-project.org/web/packages/mclust/mclust.pdf).

С помощью команды model<-Mclust(x[,3:4]) создадим объект с именем model, в котором будут содержаться результаты реализации EM-алгоритма. Для вывода результатов на экран введем команду summary(model,parameter=TRUE). Результаты работы указанных команд приведены на рисунках 2.33-2.34.

```
> model<-Mclust(x[,3:4])
fitting ...
  |======================================================================| 100%
> summary(model,parameter=TRUE)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 3 components:

 log.likelihood   n df       BIC       ICL
      -5785.572 383 17 -11672.26 -11820.39

Clustering table:
  1   2   3
168 128  87

Mixing probabilities:
        1         2         3
0.4025487 0.3059394 0.2915119
```

Рисунок 2.33 – Результаты решения задачи расщепления смеси вероятностных распределений (начало)

```
Means:
                [,1]        [,2]        [,3]
Vozrast     49.92636    27.99814    45.24777
Zarplata 26303.50496 30614.51305 58276.15397

Variances:
[,,1]
             Vozrast    Zarplata
Vozrast     132.0421   -22920.16
Zarplata -22920.1638 77168594.42
[,,2]
             Vozrast     Zarplata
Vozrast     34.17416     34534.83
Zarplata 34534.83144 131204685.49
[,,3]
             Vozrast    Zarplata
Vozrast      84.3484    -26248.9
Zarplata -26248.9038 781824099.6
```

Рисунок 2.34 – Результаты решения задачи расщепления смеси вероятностных распределений (окончание)

Таким образом, получены следующие результаты решения задачи расщепления смеси вероятностных распределений:

1) оценка числа компонент смеси $\hat{s} = 3$;

2) оценки удельных весов классов $\hat{\pi}_1 = 0{,}40$, $\hat{\pi}_2 = 0{,}31$, $\hat{\pi}_3 = 0{,}29$;

3) оценки векторов математических ожиданий в классах $\bar{x}^{(1)} = (49{,}9;26303{,}5)^T$, $\bar{x}^{(2)} = (28{,}0;30614{,}5)^T$, $\bar{x}^{(3)} = (45{,}2;58276{,}1)^T$;

4) оценки ковариационных матриц в классах $\hat{\Sigma}^{(1)} = \begin{pmatrix} 132 & -22920 \\ -22920 & 77168594 \end{pmatrix}$, $\hat{\Sigma}^{(2)} = \begin{pmatrix} 34 & 34534 \\ 34534 & 131204685 \end{pmatrix}$, $\hat{\Sigma}^{(3)} = \begin{pmatrix} 84 & -26249 \\ -26249 & 781824100 \end{pmatrix}$.

Для этого решения значение логарифма функции правдоподобия составило $\ln L = -5785{,}6$, значение Байесовского информационного критерия равно $BIC = -11672$.

В итоге все 383 объекта разбиты на три класса. Графическую интерпретацию результатов кластеризации можно получить с помощью команды plot(model,what="classification") (рисунок 2.35)

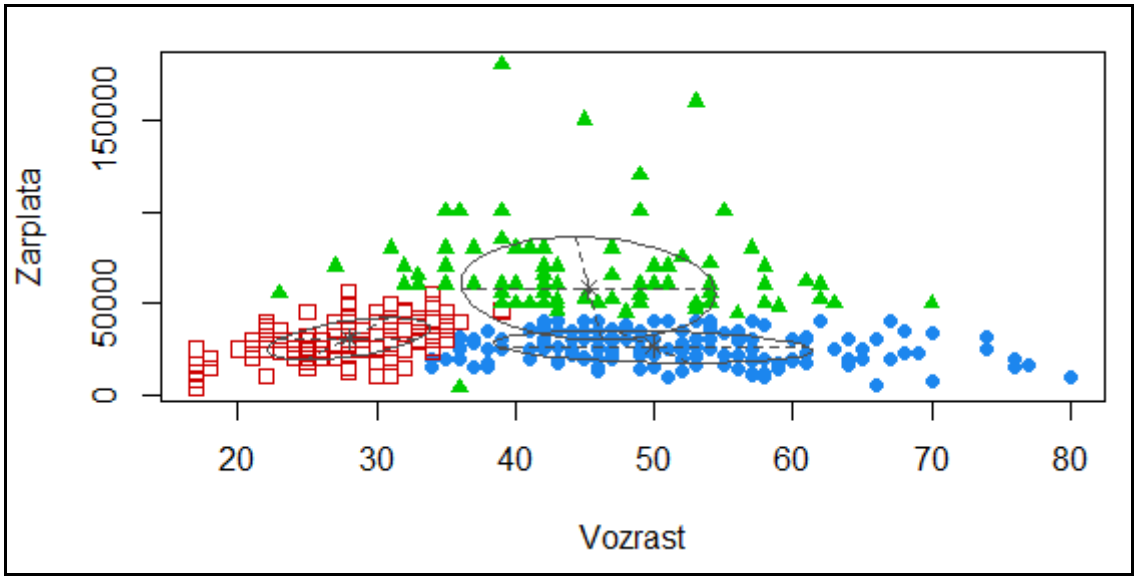

Рисунок 2.35 – Результаты разбиения объектов на три класса (обозначения: круглые – объекты первого класса, квадратные – второго, треугольные – третьего)

В первом классе респондентов, занимающим наибольший удельный вес, по сравнению со вторым и третьим классами наблюдается наибольшее среднее значение возраста и наименьшее среднее значение среднемесячной заработной платы. Наибольшее среднее значение среднемесячной заработной платы равное 58 тыс.руб. наблюдается в третьем классе респондентов, средний возраст которых составляет около 45 лет. У молодых респондентов, составивших преимущественно второй класс, среднее значение среднемесячной заработной платы равно около 31 тыс.руб.

Оценки апостериорных вероятностей отнесения объектов к классам сохраним в матрице pr с помощью команды pr<-model[["z"]], номер класса сохраним в массив klass с помощью команды klass<-model[["classification"]], оценку неопределенности (суммы двух наименьших оценок вероятностей) сохраним в массив neopr с помощью команды neopr<-model[["uncertainty"]]. Для объединения исходных данных, оценок вероятностей, номера класса и оценки неопределенности в одну таблицу с результатами воспользуемся командой rez<-cbind(x,r,klass,neopr). Содержимое созданной таблицы rez можно посмотреть, «кликнув» на объект rez в правом верхнем окне Environment или набрав команду View(rez)в окне Console. Фрагмент таблицы с результатами кластеризации представлен на рисунке 2.36.

	Obr	Prof	Vozrast	Zarplata	1	2	3	klass	neopr
1	1	1	58	10000	8.521732e-01	2.650517e-11	0.147826802	1	0.1478268020
2	1	1	49	23000	8.835626e-01	3.177717e-05	0.116405622	1	0.1164373990
3	1	0	20	25000	4.591555e-02	9.500092e-01	0.004075259	2	0.0499908095
4	3	1	42	35000	7.061631e-01	6.228863e-02	0.231548236	1	0.2938368669
5	1	1	33	60000	3.406462e-03	2.661601e-01	0.730433414	3	0.2695665856
6	3	0	32	15000	2.935887e-01	6.091240e-01	0.097287299	2	0.3908760110

Рисунок 2.36 – Фрагмент результатов кластеризации объектов

Экспортировать результаты в файл в формате CSV можно с помощью команды write.csv2(rez,file="rez.csv"), предварительно установив и подключив

библиотеку WriteXLS с помощью команд install.packages("WriteXLS") и library("WriteXLS").

Построить контурной график логарифма плотности распределения смеси вероятностных распределений можно с помощью команды plot(model,what="density"). Результаты представлены на рисунке 2.37.

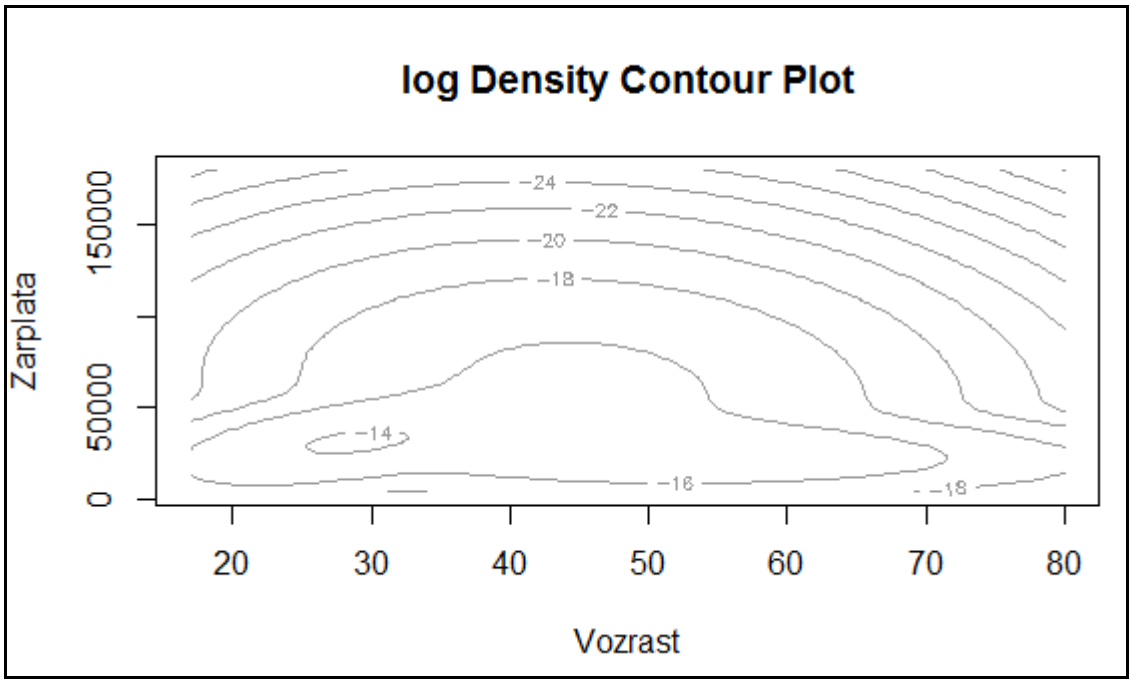

Рисунок 2.37 – Контурный график логарифма плотности распределения смеси трех нормально распределенных генеральных совокупностей

Текст программы на языке R, позволяющей решить рассмотренную задачу, приведен в приложении Б.

2.3 Сравнение результатов решения задачи расщепления смеси вероятностных распределений в пакете Statistica и среде RStudio

Результаты решения задачи расщепления смеси вероятностных распределений, полученные в пакете Statistica и в среде RStudio, различаются. Причина этого в том, что на одном и том же наборе данных решались две разные с математической точки зрения задачи. В пакете Statistica учитывались как

количественные, так и качественные признаки. Поскольку закон распределения задавался для каждого количественного признака в отдельности, следовательно, предполагалась независимость рассматриваемых показателей. В результате респонденты были разбиты на два класса, существенно различающиеся по всем четырем рассматриваемым признакам и имеющим примерно одинаковые удельные веса (0,45 и 0,55). При этом среднемесячная заработная плата респондентов первого класса почти на 19 тыс.руб. выше, чем второго класса.

В RStudio в пакете mclust реализован алгоритм расщепления смеси двумерных нормально распределенных классов, качественные признаки не учитывались. В результате респонденты были разбиты на три класса: с низкой, средней и высокой среднемесячной заработной платой; предпенсионного, молодого и среднего возраста соответственно. Результаты работы пакета mclust понятны, соответствуют теоретическим положениям, рассмотренным в главе 1, и могут использоваться для классификации новых респондентов.

3 Задание, требования к оформлению и защите отчета по лабораторной работе

Выполнение лабораторной работы по теме «Расщепление смеси вероятностных распределений» состоит из следующих этапов:

1) ознакомление с формулировкой задания к лабораторной работе и порядком её выполнения в пакетах прикладных программ;

2) выполнение расчетов по своим данным;

3) анализ полученных результатов;

4) подготовка письменного отчета по лабораторной работе;

5) защита отчета по лабораторной работе.

Задание к лабораторной работе:

1) выбрать предмет исследования и набор показателей, характеризующих данное явление или процесс;

2) собрать статистические данные по выбранным показателям, используя сайты Федеральная служба государственной статистики РФ (http://www.gks.ru), Российского мониторинга экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE) (http://www.cpc.unc.edu/projects/rlms и http://www.hse.ru/rlms), единого архива экономических и социальных данных Высшей Школы Экономики (http://sophist.hse.ru) и другие информационные ресурсы;

3) провести кластеризацию объектов с помощью EM-алгоритма, дать интерпретацию полученным результатам.

Отчет по лабораторной работе оформляется в соответствии с требованиями стандарта организации СТО 02069024.101 – 2015 [7]. Отчет должен сдержать титульный лист; задание к лабораторной работе; краткие теоретические сведения по теме лабораторной работы; результаты выполнения лабораторной работы, их анализ и интерпретацию; приложения с исходными данными и отчетами, полученными в пакетах прикладных программ.

Для защиты отчета по лабораторной работе необходимо подготовиться к ответу на вопросы и задания, приведенные ниже.

1) Дайте определения смеси вероятностных распределений.
2) Сформулируйте постановку задачи расщепления смеси вероятностных распределений.
3) В чем состоит идея EM-алгоритма?
4) Опишите E-шаг алгоритма.
5) Опишите M-шаг алгоритма.
6) Каким образом результаты решения задачи расщепления смеси вероятностных распределений могут быть использованы для классификации новых данных?
7) В чем заключаются достоинства и недостатки EM-алгоритма?
8) Каковы особенности реализации EM-алгоритма в пакете Statistica?
9) Каковы особенности реализации EM-алгоритма в среде RStudio?

Список использованных источников

1) Сиротин, В.П. Расщепление смеси вероятностных распределений в задачах моделирования социально-экономических процессов / В.П. Сиротин, М.Ю. Архипова: учебное пособие. – М: МЭСИ, 2007. – 64 с.

2) Поликарпова, М.Г. Математико-статистическое моделирование интеграционной активности российских компаний различных секторов экономики РФ / М.Г. Поликарпова // Экономика, статистика и информатика. – 2013. – № 6. – С. 181-184.

3) Айвазян, С.А. Прикладная статистика. Основы эконометрики: учебник для вузов: в 2 т. / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ-ДАНА, 2001. – Т. 1: Теория вероятностей и прикладная статистика. – 656 с.

4) Методы и модели эконометрики [Электронный ресурс] : учебное пособие для студентов, обучающихся по программам высшего образования по направлениям подготовки 01.03.04 Прикладная математика, 38.04.01 Экономика, 38.03.05 Бизнес-информатика / под ред. А. Г. Реннера; М-во образования и науки Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высш. проф. образования "Оренбург. гос. ун-т". - Ч. 1. Анализ данных. - Электрон. текстовые дан. (1 файл: 14.90 Мб). - Оренбург : ОГУ, 2015.

5) Методы и модели эконометрики. Том 1. Анализ данных: учебное пособие / О.И. Бантикова, В.И. Васянина, Ю.А. Жемчужникова, А.Г. Реннер, Е.Н. Седова, О.И. Стебунова, Л.М. Туктамышева, О.С. Чудинова / под ред. А.Г. Реннера; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2017. – 234 с.

6) Чудинова, О.С. Анализ таблиц сопряженности в пакетах Statistica, САНИ, Excel [Электронный ресурс] : методические указания к лабораторным работам, практическим занятиям и самостоятельной работе студентов / О. С. Чудинова; М-во образования и науки Рос. Федерации, Федер. гос. бюджет. образоват. учреждение высш. проф. образования "Оренбург. гос. ун-т", Каф. мат. методов и моделей в экономике. - Электрон. текстовые дан. (1 файл: 926.13 Kb). - Оренбург : ОГУ, 2014.

7) СТО 02069024.101–2015 РАБОТЫ СТУДЕНЧЕСКИЕ. Общие требования и правила оформления. – Оренбург: ОГУ, 2015. – Режим доступа: http://www.osu.ru/docs/official/standart/standart_101-2015_.pdf.

Приложение А

(обязательное)

Исходные данные для демонстрации реализации ЕМ-алгоритма в пакетах прикладных программ

Таблица А.1 – Фрагмент исходных данных

Номер п/п	Obr	Prof	Vozrast	Zarplata
1	1	1	58	10000
2	1	1	49	23000
3	1	0	20	25000
4	3	1	42	35000
5	1	1	33	60000
6	3	0	32	15000
7	2	0	70	7000
8	2	0	66	5000
9	2	0	58	12000
10	3	0	59	13600
11	1	1	33	29000
12	1	0	30	10000
13	1	0	57	12000
14	3	0	18	15000
15	2	0	18	15000
16	1	0	38	15000
17	1	1	76	15000
18	1	1	50	30000
19	3	1	54	30000
20	2	1	53	50000
21	1	1	70	50000
22	3	0	45	20000
23	1	1	34	24000
24	1	1	27	27000
25	3	0	36	4000
26	3	0	37	15000
27	2	1	21	20000
28	3	0	32	25000
…	…	…	…	…
383	2	0	24	25000

Приложение Б

(обязательное)

Текст программы на языке R для решения задачи расщепления смеси вероятностных распределений

```
setwd("C:/Анализ данных")
x<-read.csv2('smes.csv')
View(x)
install.packages("mclust")
library(mclust)
model<-Mclust(x[,3:4])
summary(model,parameter=TRUE)
plot(model,what="classification")
pr<-model[["z"]]
klass<-model[["classification"]]
neopr<-model[["uncertainty"]]
rez<-cbind(x,pr,klass,neopr)
View(rez)
install.packages("WriteXLS")
library("WriteXLS")
write.csv2(rez,file="rez.csv")
plot(model,what="density")
```