МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

Белорусский национальный технический университет

Кафедра «Системы автоматизированного проектирования»

МЕТОД КОРРЕЛЯЦИОННЫХ ПЛЕЯД

Методические указания

Минск
БНТУ
2014

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
Белорусский национальный технический университет

Кафедра «Системы автоматизированного проектирования»

МЕТОД КОРРЕЛЯЦИОННЫХ ПЛЕЯД

Методические указания
к лабораторной работе для студентов специальности
1-40 01 02 «Информационные системы и технологии»

Минск
БНТУ
2014

УДК 004.932.2(076.5)(075.8)
ББК 32.97я7
М54

Составители:
И. Л. Ковалёва, *Л. В. Федосова*

Рецензенты:
Н. Н. Гурский, *Т. А. Долгова*

Методические указания посвящены одному из разделов теории распознавания образов – анализу и оптимизации размерности пространства параметров. В теоретической части вводятся основные понятия и обсуждаются особенности корреляционного анализа. Основное внимание уделено методу корреляционных плеяд.

Изучение теоретического материала, изложенного в указаниях, и выполнение лабораторной работы дает возможность студентам не только понять основные правила построения корреляционных плеяд, но и позволяет проанализировать особенности различных вариантов реализации метода корреляционных плеяд.

Цель работы: выполнить программную реализацию одного из алгоритмов группировки параметров по степени их связи, построенного на основании метода корреляционных плеяд.

Основные понятия и определения

В процессе проектирования разработчик часто имеет дело с некоторыми значимыми, но непосредственно неизмеримыми характеристиками, определяющими разные стороны предмета проектирования, и большим (во много раз большим, чем число значимых характеристик) числом его измеримых параметров. В этом случае сокращение количества информации, достаточного для адекватного описания предмета проектирования, является немаловажной проблемой.

При решении этой проблемы перед разработчиком, прежде всего, встает задача группировки измеримых параметров, т. е. разбиения всей совокупности параметров на группы таким образом, чтобы параметры каждой из групп в основном отражали изменения какой-либо одной из значимых характеристик предмета проектирования и мало бы зависели от изменений других его значимых характеристик.

Если такая группировка выполнена, то во многих случаях оказывается возможным оценить значение каждой значимой характеристики только по одному или нескольким параметрам соответствующей группы, т. е. уменьшить количество информации, необходимой для достаточно полного описания предмета проектирования.

Благодаря наличию взаимосвязи каждого параметра из группы с соответствующей значимой характеристикой, параметры многомерного объекта (предмета проектирования) в группе в некотором смысле связаны между собой. Кроме того, связи могут присутствовать и между параметрами из различных групп. Все эти связи могут быть более или менее тесными. Как известно, наиболее тесной связью между двумя пара-

метрами является функциональная зависимость. Она имеет место, если каждому значению одного из параметров соответствует определенное значение другого параметра. Однако чаще связь между параметрами не является столь явной и присутствует лишь в виде тенденции. В этом случае можно говорить, например, только о том, что увеличение одного из параметров в среднем соответствует увеличению (или уменьшению) другого, однозначное же соответствие между параметрами отсутствует. Связь такого рода называют корреляционной связью. О параметрах, связанных корреляционной связью, говорят, что они взаимно коррелированны.

Критерии количественной оценки взаимосвязи (зависимости) между параметрами называются коэффициентами корреляции или мерами связности. Два параметра коррелируют между собой положительно, если между ними существует прямое, однонаправленное соотношение. При однонаправленном соотношении малые значения одного параметра соответствуют малым значениям другого параметра, большие значения – большим. Два параметра коррелируют между собой отрицательно, если между ними существует обратное, разнонаправленное соотношение. При разнонаправленном соотношении малые значения одного параметра соответствуют большим значениям другого параметра и наоборот. Значения коэффициентов корреляции всегда лежат в диапазоне от –1 до +1. Сила связи характеризуется абсолютной величиной коэффициента корреляции. Для словесного описания величины коэффициента корреляции можно использовать следующие градации: до 0,2 – очень слабая корреляция, от 0,2 до 0,5 – слабая корреляция, от 0,5 до 0,7 – средняя корреляция, от 0,7 до 0,9 – «высокая корреляция, свыше 0,9 – очень высокая корреляция.

Метод вычисления коэффициента корреляции зависит от вида шкалы, к которой относятся параметры. Параметры с интервальной и номинальной шкалой описываются коэффициентом корреляции Пирсона (корреляция моментов произведе-

ний). Если, по крайней мере, один из двух параметров имеет порядковую шкалу либо не является нормально распределенным, применяется ранговая корреляция по Спирману или Кендалу. В случае, когда один из двух параметров является дихотомическим, используется точечная двухрядная корреляция. При этом следует учесть, что каждый дихотомический параметр можно рассматривать как порядковый.

Коэффициенты корреляции могут быть вычислены для каждой пары измеримых параметров, определяющих предмет проектирования. Совокупность взаимных коэффициентов корреляции принято представлять в виде так называемой корреляционной матрицы.

В качестве примера рассмотрим корреляционную матрицу для 5 параметров ($v1$; $v2$; ...; $v5$; P = 5), измеренных на выборке количеством N = 30. На рис. 1 приведена таблица исходных данных, а на рис. 2 – корреляционная матрица.

№	v1	v2	v3	v4	v5
1	10	23	4	111	56
2	13	26	6	98	52
3	8	12	2	105	58
4	9	25	7	100	49
5	11	16	3	101	65
...	...	...	...	...	...
30	7	19	6	94	41

Рис. 1. Таблица исходных данных

	v1	v2	v3	v4	v5
v1	1	0,52	–0,11	–0,29	–0,38
v2	0,52	1	0,28	0,32	–0,34
v3	–0,11	0,28	1	0,48	0,42
v4	–0,29	0,32	0,48	1	0,38
v5	–0,38	–0,34	0,42	0,38	1

Рис. 2. Корреляционная матрица

Корреляционная матрица является квадратной, так как количество строк и столбцов равно числу параметров. Она симметрична относительно главной диагонали, так как корреляция параметра *x* с параметром *y* равна корреляции параметра *y* с параметром *x*. На ее главной диагонали располагаются единицы, так как корреляция параметра с самим собой равна единице. Поэтому корреляционную матрицу можно не приводить целиком, ограничиваясь ее частью, лежащей, например, ниже главной диагонали.

Анализ корреляционной матрицы позволяет выявить структуру взаимосвязей множества параметров и уменьшить количество измеримых параметров, необходимых для достаточно полного описания предмета проектирования. При этом возможен визуальный анализ корреляционных плеяд – графического изображения структуры статистически значимых связей.

Использовать корреляционные плеяды предложил в конце 50-х годов прошлого века ленинградский зоолог, профессор П. В. Терентьев. Занимаясь изучением корреляций между множеством различных признаков озерной лягушки, он объединил их в группы по абсолютной величине коэффициентов корреляции, тем самым получив два распределения: признаки с малой и большой по величине корреляцией. Терентьев назвал эти группы корреляционными плеядами и опубликовал несколько способов их анализа.

Вслед за биологами новый метод быстро переняли психологи университета, и метод корреляционных плеяд стал одним из основных в дифференциальной психологии и многих других психологических науках. Методика корреляционных плеяд в дальнейшем была расширена Н. С. Ростовой.

Метод корреляционных плеяд предназначен для нахождения таких групп параметров или объектов – «плеяд», когда корреляционная связь, т. е. сумма модулей коэффициентов корреляции между параметрами одной группы (внутриплеядная связь) достаточно велика, а связь между параметрами из разных групп (межплеядная) – мала. По определенному правилу по корреляционной матрице объектов образуют чертеж – граф, который затем с помощью различных приемов разбивают на подграфы. Элементы, соответствующие каждому из подграфов, и образуют плеяду.

Таким образом, корреляционная плеяда – это фигура, состоящая из вершин и соединяющих их линий. Вершины соответствуют параметрам и обозначаются обычно цифрами – номерами параметров. Линии соответствуют статистически достоверным связям и графически выражают знак, а иногда – и уровень значимости связи. Корреляционная плеяда может отражать все статистически значимые связи корреляционной матрицы (иногда называется корреляционным графом) или только их содержательно выделенную часть (например, соответствующую одному фактору по результатам факторного анализа).

Правила построения корреляционных плеяд

При построении корреляционной плеяды параметры изображаются при помощи геометрических фигур (например, кругов или прямоугольников), внутри которых записывается название параметра, а связи между параметрами – при помощи соединительных линий.

Существует несколько правил построения корреляционных плеяд.

1. Поскольку в корреляционных исследованиях наличие взаимосвязи говорит лишь о сопряженности изучаемых параметров, но никак не о причинно-следственной зависимости, при построении корреляционных плеяд не рекомендуется использовать односторонние стрелочки, показывающие направление взаимосвязи. Используются либо двусторонние стрелки, либо простые соединительные линии.

2. Прямые и обратные взаимосвязи обозначаются (маркируются) посредством разных графических характеристик линий: например, прямые – сплошной линией, обратные – пунктирной.

3. Корреляционный анализ может включать достаточно большое количество параметров, между которыми могут быть получены самые разнообразные взаимосвязи. Чтобы рисунок, отражающий все эти взаимосвязи, был читаемым, важно удачно расположить элементы корреляционной плеяды относительно друг друга.

Обычно в центре плеяды размещают тот параметр, у которого обнаружено наибольшее количество значимых взаимосвязей, а на периферию выносятся параметры, имеющие единичные связи. Иногда исследователю важно сконцентрироваться на взаимосвязях только какого-то одного параметра. Тогда его помещают в центр рисунка, а параметры, связанные с ним, располагают вокруг (при этом имеющиеся связи между остальными параметрами игнорируют). Во многих случаях (и особенно в сравнительных исследованиях) информацию несут факты не только о наличии связей, но и об их отсутствии. Наиболее ясной становится картина, когда на рисунке обозначаются все параметры, занятые в исследовании, хотя при этом между многими из них может не быть соединительных линий.

Варианты построения корреляционных плеяд

Существующие различные варианты метода корреляционных плеяд являются в действительности несколько упрощенными эвристическими версиями более совершенных в математическом плане алгоритмов исследования структуры связей между компонентами многомерного параметра, использующими графы-деревья. Рассмотрим некоторые из них.

Вариант 1

Рассмотрим корреляционную матрицу n параметров. Нарисуем n кружков, в каждом из них напишем номер одного из параметров. Соединим каждый кружок линиями с каждым из остальных n–1 кружков и обозначим над линиями значения коэффициентов корреляции. Получим некоторый исходный граф. Зададимся пороговым значением коэффициента корреляции t и исключим из графа все линии, которые соответствуют коэффициентам корреляции, меньшим пороговой величины. Будем постепенно увеличивать t и выбрасывать из графа связи по тому же правилу. При некотором достаточно большом t граф распадется на несколько подграфов – изолированных друг от друга групп кружков.

Полученной группировке кружков будет соответствовать группировка параметров, характерная тем, что коэффициенты корреляции между параметрами каждой группы больше, а между параметрами разных групп – меньше пороговой величины.

Данный вариант неудобен тем, что приходится рассматривать большое число (около n^2) связей.

На рис. 3 приведен пример корреляционной матрицы, построенной для 6 параметров, и соответствующего ей корреляционного графа, а на рис. 4 – корреляционных плеяд, полученных при пороге t = 0,35.

	X1	X2	X3	X4	X5	X6
X1	1	0,5	0,5	-0,7	0,3	0
X2	0,5	1	0,6	0,3	-0,2	-0,1
X3	0,5	0,6	1	0,2	-0,5	-0,1
X4	-0,7	0,3	0,2	1	0,5	0,4
X5	0,3	-0,2	-0,5	0,5	1	0,4
X6	0	-0,1	-0,1	0,4	0,4	1

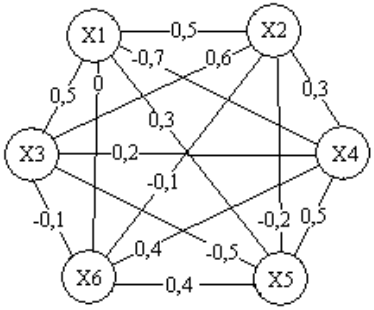

Рис. 3. Корреляционная матрица и корреляционный граф

	X1	X2	X3	X4	X5	X6
X1	1	0,5	0,5	-0,7	0,3	0
X2	0,5	1	0,6	0,3	-0,2	-0,1
X3	0,5	0,6	1	0,2	-0,5	-0,1
X4	-0,7	0,3	0,2	1	0,5	0,4
X5	0,3	-0,2	-0,5	0,5	1	0,4
X6	0	-0,1	-0,1	0,4	0,4	1

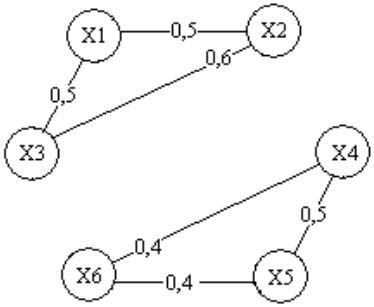

Рис. 4. Корреляционные плеяды при пороге $t = 0{,}35$

Вариант 2

Рассмотрим корреляционную матрицу $R = (r_{ij})$, $i,j = 1,2, ..., p$ исходных параметров. В данном варианте корреляционных плеяд предполагается упорядочивать параметры и рассматривать только те коэффициенты корреляции, которые соответствуют связям между параметрами в упорядоченной системе. Упорядочение производится на основании принципа макси-

мального корреляционного пути: все p параметры связываются при помощи (p–1) линий (ребер) так, чтобы сумма модулей коэффициентов корреляции была максимальной. Это достигается следующим образом: в корреляционной матрице находят наибольший по абсолютной величине коэффициент корреляции, например $|r(l,m)| = r(1)$ (коэффициенты на главной диагонали матрицы, равные единице, не рассматриваются).

Рисуем кружки, соответствующие параметрам $x(l)$ и $x(m)$, и над связью между ними пишем значение $r(1)$. Затем, исключив $r(1)$, находим наибольший коэффициент в m-м столбце матрицы (это соответствует нахождению параметра, который наиболее сильно после $x(l)$ «связан» с $x(m)$), и наибольший коэффициент в l-й строке матрицы (это соответствует нахождению параметра, наиболее сильно после $x(m)$ «связанного» с x(l)). Из найденных таким образом двух коэффициентов корреляции выбирается наибольший – пусть это будет $|r(l,j)| = r(2)$. Рисуем кружок $x(j)$, соединяем его с кружком $x(l)$ и проставляем значение $r(2)$. Затем находим параметры, наиболее связанные с $x(l)$, $x(m)$ и $x(j)$, и выбираем из найденных коэффициентов корреляции наибольший. Пусть это будет $|r(j,q)| = r(3)$. Требуем, чтобы на каждом шаге появлялся новый параметр, поэтому параметры, уже изображенные на чертеже, исключаются, следовательно, q не равно l, q не равно m, q не равно j.

Далее рисуем кружок, соответствующий $x(q)$, и соединяем его с $x(j)$ и т.д. На каждом шаге находятся параметры, наиболее сильно связанные с двумя последними рассмотренными параметрами, а затем выбирается один из них, соответствующий большему коэффициенту корреляции. Процедура заканчивается после (p–1)-го шага; граф оказывается состоящим из p кружков, соединенных (p–1) ребром. Затем задается пороговое значение t, а все ребра, соответствующие меньшим чем t, коэффициентам корреляции, исключаются из графа.

Незамкнутым графом называется такой граф, для которого для любых двух кружков существует единственная траекто-

рия, составленная из линий связи, соединяющая эти два кружка. Очевидно, что в данном варианте метода корреляционных плеяд допускается построение только незамкнутых графов.

Вариант 3

Возможен еще более простой вариант, который тем не менее дает хорошие результаты. Он отличается от предыдущего тем, что все кружки выстраиваются в одну линию и на каждом шаге отыскиваются два наибольших коэффициента корреляции, связывающих крайние параметры графа, полученного на предыдущем шаге, с параметрами, еще не занесенными в граф. Из этих двух коэффициентов выбирается больший, и граф дополняется параметром, соответствующим этому наибольшему коэффициенту.

Затем задается пороговое значение t, а все ребра, соответствующие меньшим чем t коэффициентам корреляции, исключаются из графа.

Вариант 4

Построение корреляционной плеяды можно начать с выделения в корреляционной матрице статистически значимых корреляций (иногда разным цветом – в зависимости от уровня значимости). Затем для строк (столбцов) матрицы, содержащих статистически значимые корреляции, подсчитывается их количество. Построение плеяды начинают с параметра, имеющего наибольшее число значимых связей, постепенно добавляя в рисунок другие параметры – по мере убывания числа связей – и связывая их линиями, соответствующими связям между ними.

На рис. 5 приведен пример корреляционной матрицы, а на рис. 6 – соответствующей ей корреляционной плеяды, построенной по 4 варианту. В данной корреляционной плеяде для обозначения разного характера взаимодействий между параметрами используются линии разных типов.

	v1	v2	v3	v4	v5
v1	1	0,52	–0,11	–0,29	–0,38
v2	0,52	1	0,28	0,32	–0,34
v3	–0,11	0,28	1	0,48	0,42
v4	–0,29	0,32	0,48	1	0,38
v5	–0,38	–0,34	0,42	0,38	1

Рис. 5. Корреляционная матрица для 4 варианта

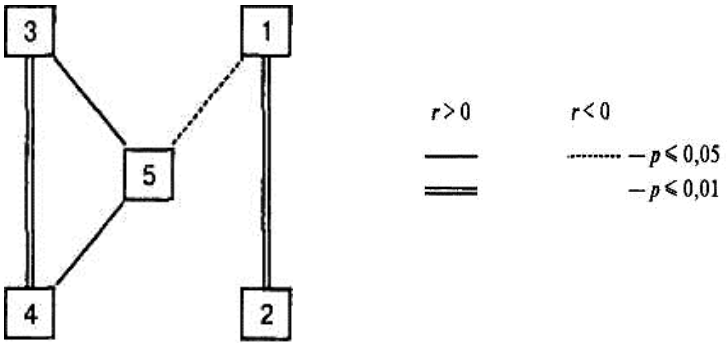

Рис. 6. Корреляционная плеяда и условные обозначения для 4 варианта

Основные этапы и требования к выполняемой работе

В ходе выполнения лабораторной работы необходимо разработать программный модуль, реализующий один из вариантов метода корреляционных плеяд. Рассмотрим подробнее требования к функциональности программного модуля.

1. На первом этапе должна быть сформирована корреляционная матрица.

Как уже говорилось выше, в качестве коэффициентов корреляции в матрице могут использоваться коэффициенты корреляции Пирсона, Спирмана и Кендала.

Коэффициент Пирсона вычисляется по следующей формуле:

$$r = r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \overline{y})}{(n-1) \cdot s_x \cdot s_y},$$

где x_i и y_i – значения двух параметров;

$\bar{x}$ и $\overline{y}$ – средние значения параметров;

s_x и s_y – стандартные отклонения параметров;

n – количество пар значений.

Если корреляционная матрица заполняется по результатам проведенных исследований, то должна быть предусмотрена возможность ее импорта в программный модуль. Также необходимо предусмотреть возможность редактирования корреляционной матрицы. Для отладки алгоритма корреляционных плеяд требуется реализовать функцию автоматического заполнения матрицы любого размера, обеспечивая при этом согласованность и взаимосвязь всех коэффициентов корреляции.

Пример диалогового окна приведен на рис. 7.

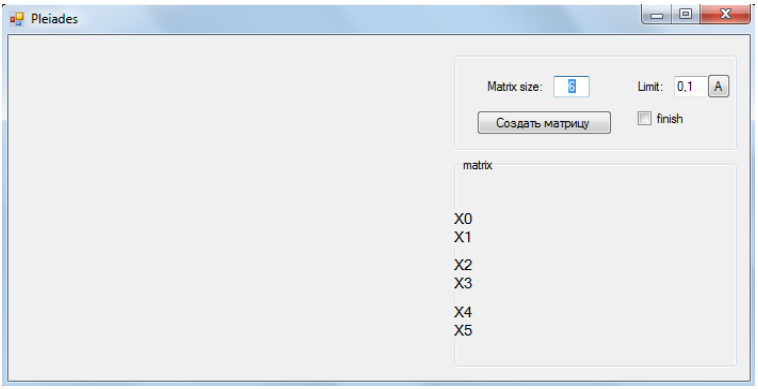

Рис. 7. Диалоговое окно программного модуля

В диалоговом окне предусмотрена возможность задания размера матрицы («Matrix size») и порога («Limit») для деле-

ния корреляционного графа на плеяды. В левой части окна располагается область для отрисовки графа. Переменные X0, X1, …, X5 обозначают параметры.

Пусть корреляционная матрица была сгенерирована автоматически (рис. 8).

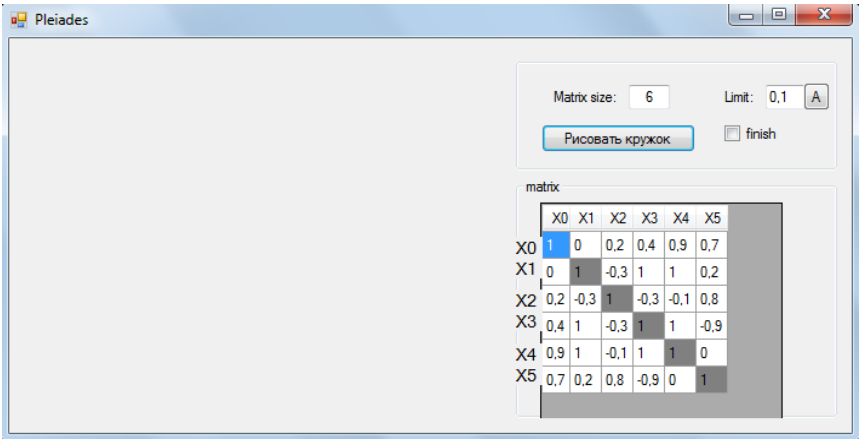

Рис. 8. Заполнение корреляционной матрицы

На этом этапе пользователь должен иметь возможность изменять значения коэффициентов матрицы.

2. Рассмотрим реализацию второго варианта метода корреляционных плеяд. Запуск алгоритма производится по нажатию на кнопку «Рисовать кружок». В корреляционной матрице в правой части диалогового окна автоматически выбираются две ячейки: ячейка с наибольшим коэффициентом корреляции и ячейка со следующим по величине значением в этой же строке и этом же столбце. Выбор должен быть отражен визуально в матрице, например, с помощью изменения цвета ячеек. На рис. 9 красным цветом отмечается ячейка с наибольшим коэффициентом корреляции (строка X1, столбец X3), а зеленым цветом – ячейка со следующим по величине значением коэффициента корреляции (строка X1, столбец X4).

Для красной ячейки отрисовываются два кружка и ребро. Кружкам присваиваются номера строки 1 и столбца 3 красной ячейки (ячейки с наибольшим коэффициентом корреляции), которые соответствуют номерам двух параметров X1 и X3. На ребре указывается значение коэффициента корреляции из выбранной ячейки.

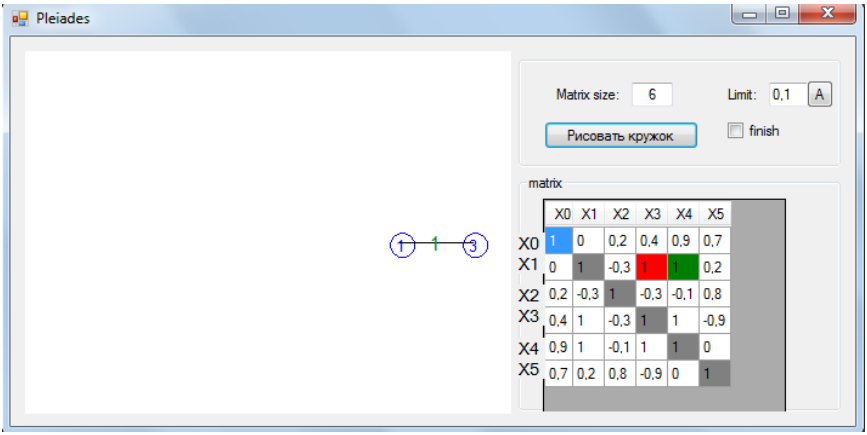

Рис. 9. Начало работы алгоритма

3. На следующем этапе ячейка, помеченная на предыдущем этапе красным цветом, закрашивается в желтый цвет и не участвует в сравнениях коэффициентов. А из текущего столбца X1 и текущей строки X3 выбирается ячейка с наибольшим коэффициентом корреляции. Ею оказывается ячейка, расположенная в строке X1 и столбце X4. Эта ячейка закрашивается в красный цвет, для нее создается объект и ребро, а в строке X1 и столбце X4 выбирается новая ячейка (строка X3, столбец X4), которая закрашивается в зеленый цвет (рис. 10).

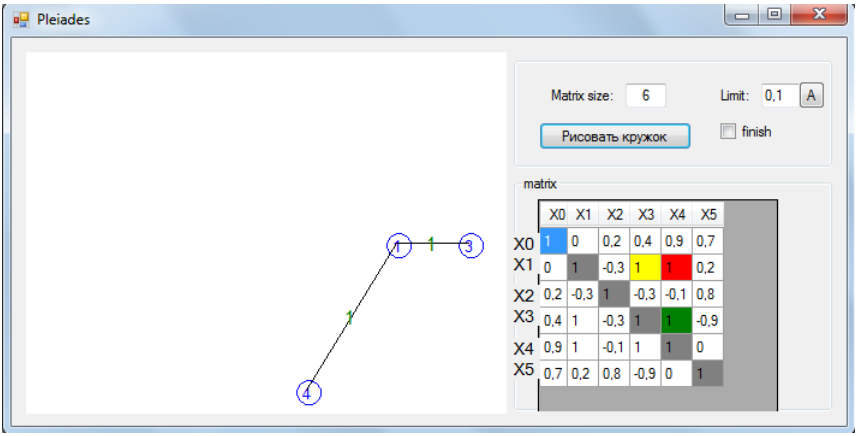

Рис. 10. Этап 3

4. Объекты (кружки) 3 и 4 уже были созданы на предыдущих этапах, поэтому в результате выполнения этого шага создалось лишь одно ребро между этими объектами (рис. 11).

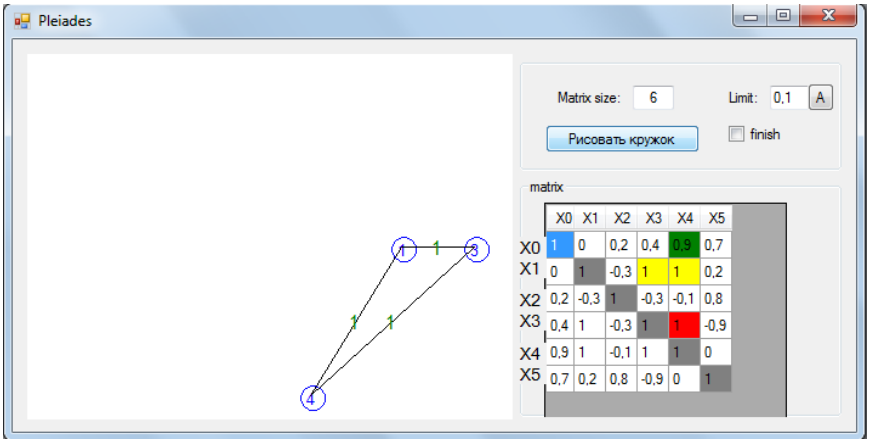

Рис. 11. Создание нового ребра

5. Аналогично выполняются следующие шаги. Обработка корреляционной матрицы ведется до тех пор, пока весь корре-

ляционный граф не будет построен. То есть пока не будет сформировано столько кружков, сколько параметров задано в корреляционной матрице. На рис. 12 показан полученный корреляционный граф.

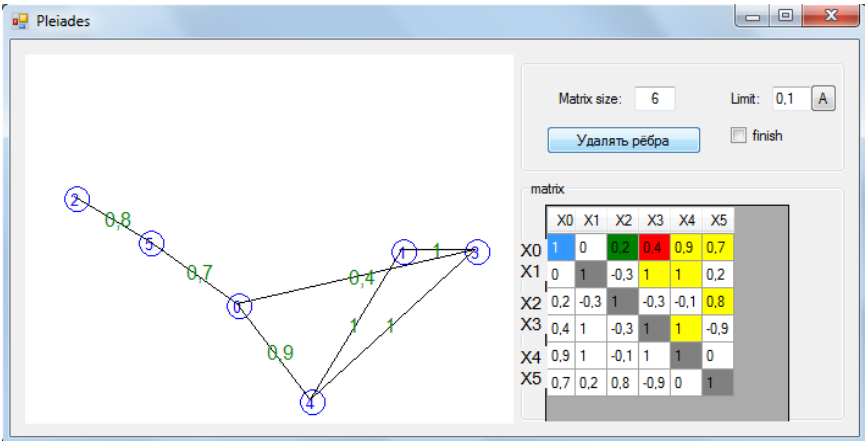

Рис. 12. Корреляционный граф

6. На этом этапе начинается формирование корреляционных плеяд. Для этого пользователю должна быть предоставлена возможность задавать различные значения порога. В рассматриваемом примере эта возможность реализована с помощью диалогового окна «Limit». Согласно алгоритму, ребра со значением коэффициента корреляции меньшим чем порог Limit, должны быть удалены из графа. Например, если порог равен 0,8, то граф распадается на две плеяды (рис. 13).

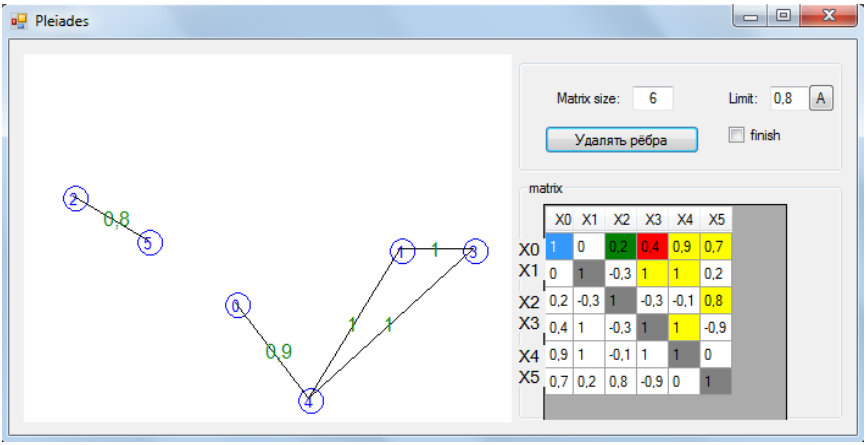

Рис. 13. Корреляционные плеяды

На этом работа алгоритма заканчивается.

Литература

1. Горелик, А. Л. Методы распознавания : учеб. пособие для вузов / А. Л. Горелик, В. А. Скрипкин. – 3-е изд., перераб. и доп. – М. : Высшая школа, 1989. – 234 с.

2. Гренандер, У. Лекции по теории образов : в 3 т. / У. Гренандер ; под ред. Ю. И. Журавлева. – М. : Мир, 1979–1983. – 1267 с.

3. Ростова, Н. С. Корреляции: структура и изменчивость / Н. С. Ростова. – СПб. : Изд-во С-Петерб. ун-та, 2002. – 308 с.

4. Терентьев, П. В. Метод корреляционных плеяд / П. В. Терентьев // Вестник ЛГУ. – 195. – № 9.

5. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей : пер. с нем. / Ахим Бююль, Петер Цефель. – СПб. : ООО «ДиаСофтЮП», 2005. – 608 с.

6. STATIATICA. Искусство анализа данных на компьютере: Для профессионалов. – 2-е изд. – СПб. : Питер, 2003. – 688 с.

Учебное издание

МЕТОД КОРРЕЛЯЦИОННЫХ ПЛЕЯД

Методические указания к лабораторной работе
для студентов специальности
1-40 01 02 «Информационные системы и технологии»

Составители:
КОВАЛЁВА Ирина Львовна
ФЕДОСОВА Людмила Владимировна

Редактор *Л. Н. Шалаева*
Компьютерная верстка *А. Г. Занкевич*

Подписано в печать 23.01.2014. Формат 60×84 $^{1}/_{16}$. Бумага офсетная. Ризография.
Усл. печ. л. 1,22. Уч.-изд. л. 0,96. Тираж 100. Заказ 1015.

Издатель и полиграфическое исполнение: Белорусский национальный технический университет.
Свидетельство о государственной регистрации издателя, изготовителя, распространителя
печатных изданий № 1/173 от 12.02.2014. Пр. Независимости, 65.220013, г. Минск.