**МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
И ПРОДОВОЛЬСТВИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

ГЛАВНОЕ УПРАВЛЕНИЕ ОБРАЗОВАНИЯ, НАУКИ И КАДРОВ

**УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКАЯ ГОСУДАРСТВЕННАЯ
СЕЛЬСКОХОЗЯЙСТВЕННАЯ АКАДЕМИЯ»**

**Кафедра математического моделирования
экономических систем АПК**

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

**Практикум для магистрантов
специальности 1-25 80 01 «Экономика»**

Горки 2019

МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
И ПРОДОВОЛЬСТВИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

ГЛАВНОЕ УПРАВЛЕНИЕ ОБРАЗОВАНИЯ, НАУКИ И КАДРОВ

УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКАЯ ГОСУДАРСТВЕННАЯ
СЕЛЬСКОХОЗЯЙСТВЕННАЯ АКАДЕМИЯ»

Кафедра математического моделирования
экономических систем АПК

ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Практикум для магистрантов
специальности 1-25 80 01 «Экономика»

Горки 2019

УДК 338.27
ББК 65.054в6
Б 946
Рекомендовано методической комиссией экономического факультета 2019 (протокол №) и научно-методическим советом 2019 (протокол №).

Составил: В.И. БУЦЬ.

П 69 **Технологии интеллектуального анализа данных**/ Белорусская государственная сельскохозяйственная академия; Сост. В. И. Б у ц ь . Горки, 2019. 64 с.

Изложена методика выполнения лабораторных работ по учебному курсу «Технологии интеллектуального анализа данных».

Для магистрантов специальности 1-25 80 01 «Экономика».

Таблиц 15, рисунков 10.

Рецензенты: и.о.директора Республиканского научного унитарного предприятия «Институт системных исследований в АПК Национальной академии наук Беларуси, д.э.н., доц. А. В. ПИЛИПУК; профессор кафедры математических методов в экономике УО «БГЭУ», д.э.н., доц. Э. М. АКСЕНЬ

УДК 338.27
ББК 65.054в6
Б 946

ВВЕДЕНИЕ

Методы интеллектуального анализа данных и процессов являются актуальными и востребованными как методы прикладной информатики, нацеленные на решение сложных задач анализа и исследования закономерностей в информационных процессах и информационных системах. Такие методы позволяют извлекать из большого объема данных полезную информацию, необходимую для принятия обоснованных решений в бизнесе и в управлении предприятием и организацией. Практикум нацелен на поддержку магистров при выполнении лабораторных работ по дисциплине «Интеллектуальный анализ данных и процессов», позволяющих получить компетенции в области применения методов интеллектуального анализа данных и процессов, познакомиться с опытом зарубежных исследователей и разработчиков, выполнить исследование применимости интеллектуальных методов к решению поставленных задач. Учитывая, что в настоящее время разработано и реализовано достаточно большое количество методов интеллектуального анализа данных, которые рассматриваются на начальном этапе в бакалавриате, в методических рекомендациях к выполнению лабораторных работ для магистров сделан акцент на исследовательский аспект и на методы интеллектуального анализа данных и процессов, недостаточно представленных в отечественной литературе. Это позволит расширить компетенции магистров на различных уровнях освоения. Необходимые начальные навыки для выполнения лабораторных работ включают умение искать и применять новую информацию, программировать и знание английского языка.

1. ПРОЦЕСС ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ: ВЫБОР ДАННЫХ

Методические указания: В общем случае процесс интеллектуального анализа (т.е. поиска нового знания) состоит из следующих этапов:

1) отбор данных (выбор признаков, которые, по предположению исследователей, являются значимыми для конкретного исследования, в нашем случае для анализа ценностных ориентаций современного студенчества);

2) предобработка данных или очистка (устранение неточностей, принятие решения о работе с неответами и т.д.);

3) трансформация (преобразование шкал для применения выбранных методов);

4) собственно извлечение знаний (Data Mining);

5) интерпретация результатов в контексте содержательных гипотез.

Можно утверждать, что Data Mining обозначает такой подход к анализу эмпирических данных, при котором исследователь готов к тому, что анализируемый феномен может оказаться слишком запутанным и не поддающимся точному анализу с помощью традиционных методов. Перед началом интеллектуального анализа данных, вы можете задать вопрос «сколько данных требуется?» или «Существуют ли специальные требования, которые мне нужно знать при очистке или форматировании данных?». Пользователи, ранее не знакомые с интеллектуальным анализом данных, часто сталкиваются с проблемами в Excel, например, с необходимостью согласованного форматирования данных в столбцах, очисткой отсутствующих значений или группированием чисел.

Выбор данных, используемых для анализа, возможно, является наиболее важной частью процесса интеллектуального анализа данных — даже более важной, чем выбор алгоритма. Это связано с тем, что интеллектуальный анализ основывается не на гипотезах, а на данных. Вместо того чтобы выбрать и проверить переменные заранее, как это делается в традиционных статистических моделях, интеллектуальный анализ может получать данные и открывать новые связи (или совсем не обнаружить никаких закономерностей). Качество и объем данных могут оказать значительное влияние на результат. В целом придерживайтесь следующих правил:

1) Получите как можно более чистые данные.

2) Выполните профилирование данных перед использованием моделей. Необходимо понимать, с какими данными вы работаете. Как минимум:

3) Используйте инструменты в надстройках для поиска минимальных и максимальных значений, наиболее распространенных и средних значений.

4) Заполните все отсутствующие значения. Надстройки (а также некоторые алгоритмы) реализуют средства для ввода отсутствующих значений.

5) Исправьте недопустимые данные, если это возможно. Проекты интеллектуального анализа данных часто используются в качестве стимула для новых по обработке данных.

6) Попробуйте создать тестовую модель и найти проблемы с данными с ее помощью. Проверяя результаты, вы можете обнаружить, например, что прогнозы продаж основаны на аномальных данных из-за ошибки конвертации валют.

7) Попробуйте преобразовать данные в различные форматы или сегментировать значения. Шаблоны часто выявляются после преобразования данных.

8) Поместите числа в соответствующие сегменты, чтобы уменьшить число возможных значений для анализа.

9) Создайте несколько версий данных и постройте несколько моделей.

10) Старайтесь придерживаться следующего правила: для самых простых типов моделей и сценариев должно иметься не менее 50–100 строк данных. Например, если осуществляется прогноз одного атрибута с помощью модели упрощенного алгоритма Байеса и набор данных правильного формата, то можно будет получить довольно точные прогнозы с использованием 50–100 строк данных. Для моделей взаимосвязей обычно требуется больше данных — тысячи строк может оказаться недостаточно, если анализируются множество атрибутов. Если набор данных слишком велик или мал, иногда можно улучшить результаты с помощью сжатия строк по категориям. Например, вместо того чтобы анализировать взаимосвязи между отдельными продуктами агропромышленного производства, можно классифицировать продукты по категориям. Если набор данных имеет разумный размер, сосредоточьтесь на качестве данных, а не на добавлении новой информации. В какой-то момент, когда будут найдены все статистически действительные шаблоны, добавление данных перестанет повышать качество шаблонов. И наоборот, по мере добавления данных иногда могут появиться случайные корреляции.

Задание: сгенерировать данные для 150 наблюдений для построения линии тренда по каждому виду сельскохозяйственной продукции (табл.1).

Порядок выполнения задания: На основании показателей таблицы 1 сгенерируем в EXCEL массив данных для целей простой линейной регрессии. Линейный тренд зададим уравнением: $Y=k\cdot X+m$. Для академических целей бывает полезно сгенерировать в MS EXCEL значения двух переменных, которые демонстрируют приблизительно линейную взаимосвязь. Для этого нам понадобятся следующие исходные данные: наклон k (для зерна $k=-1.07$) и сдвиг m (для зерна $m=804$) для задания линии тренда $Y=k\cdot x+m$; начальное значение X (для зерна $X_{2005}=664$) и шаг изменения X (шаг будет равномерный; для зерна 5% от начального значения); величина разброса (в процентах от среднего значения Y). Сначала сгенерируем массив значений X и $Y=k*X+m$ (столбец: $Y_{трено}$). Линию тренда построим на диаграмме типа «Точечная». Генерацию случайного разброса точек будем производить по нормальному закону (столбец $Y_{нормальный\ разброс}$). Для этого необходимо использовать функцию НОРМ.ОБР() или генератор случайных чисел: =ЕСЛИ(И($B$15>0; СРЗНАЧ ($D$23: $D$83)); НОРМ.ОБР(СЛЧИС (); D23; $B$15*ABS (СРЗНАЧ ($D$23: $D$83))); D23). В качестве среднего значения (второй аргумент функции НОРМ.ОБР()) будем брать текущее значение Y тренда (D23). Для наглядности также построим линию регрессии (с помощью инструмента «Линия тренда» и с помощью формул, предварительно вычислив параметры модели).

Таблица 1. **Производство сельскохозяйственной продукции на душу населения, кг/чел.**

Виды продукции	Годы																В ср-м 2005-2018 **m**	Прирост в ср-м за год, **k**
	1995	2000	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018		
Зерно	540	487	664	617	755	946	895	736	873	975	803	1 009	912	785	842	649	804	-1,07
Картофель	932	874	847	867	915	918	749	825	755	730	624	663	632	630	675	618	746	-16,36
Овощи	101	138	208	226	225	241	243	246	192	167	172	183	178	199	206	184	205	-1,71
Скот и птица (в убойном весе)	64	60	72	80	85	88	97	102	108	115	124	113	121	123	127	129	106	4,07
Молоко	497	450	587	614	617	653	692	698	686	715	701	707	743	751	771	775	694	13,43
Яйца, шт.	331	329	321	347	338	348	361	373	386	399	407	407	395	380	370	355	371	2,43

Источник: данные Национального статистического комитета Республики Беларусь.

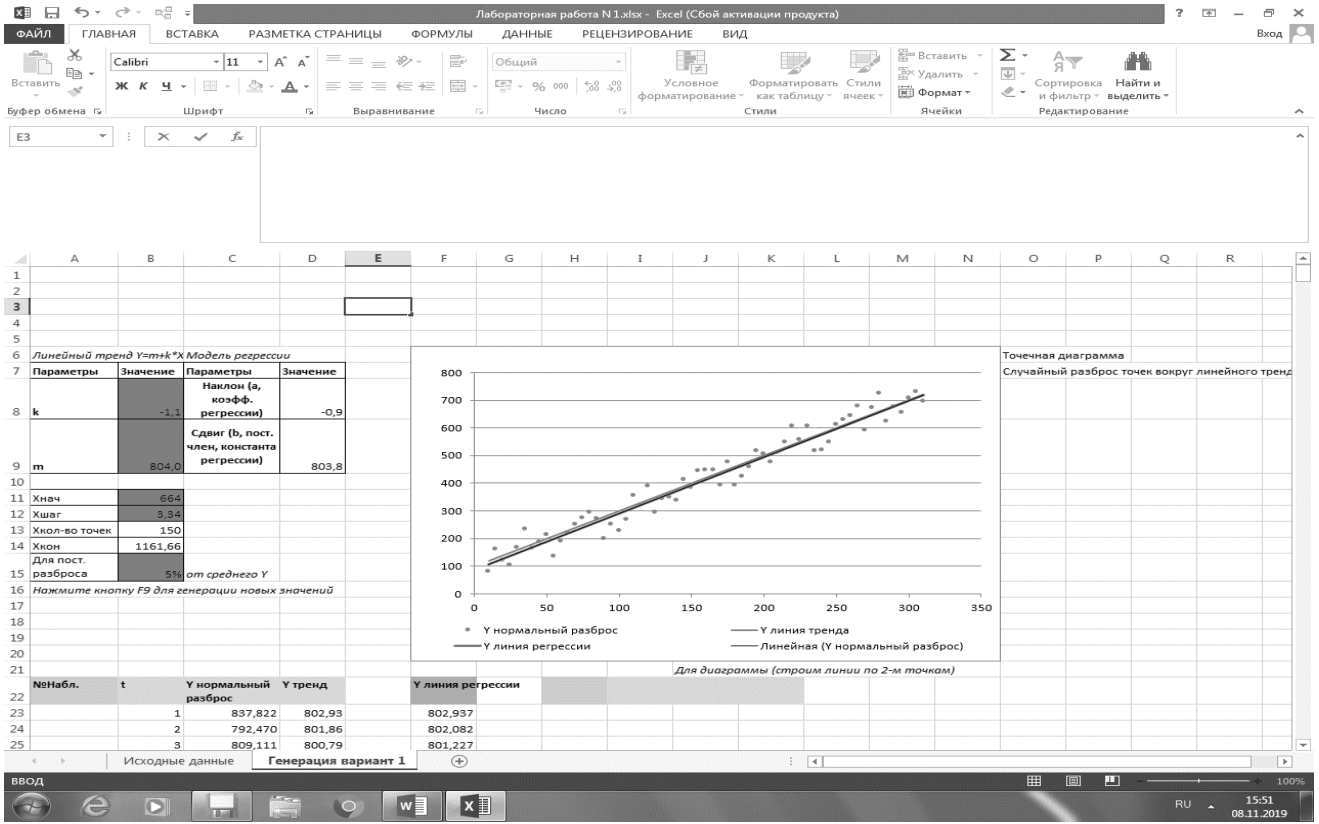


Рис 1. **Фрагмент расчетов по лабораторной работе № 1 в EXCEL**

2. ТЕХНОЛОГИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Методические указания: Логистическая регрессия является одним из статистических методов классификации, как технологии интеллектуального анализа данных с использованием линейного дискриминанта Фишера. Также она входит в топ часто используемых алгоритмов в науке о данных.

В отличие от обычной регрессии, в методе логистической регрессии не производится предсказание значения числовой переменной исходя из выборки исходных значений. Вместо этого, значением функции является вероятность того, что данное исходное значение принадлежит к определенному классу. Для простоты, давайте предположим, что у нас есть только два класса (множественная логистическая регрессия для задач с большим количеством классов) и вероятность, которую мы будем определять, P+ вероятности того, что некоторое значение принадлежит классу "+". И конечно P–=1–P+. Таким образом, результат логистической регрессии всегда находится в интервале [0, 1].

Основная идея логистической регрессии заключается в том, что пространство исходных значений может быть разделено линейной границей (т.е. прямой) на две соответствующих классам области. Итак, что же имеется ввиду под линейной границей? В случае двух измерений — это просто прямая линия без изгибов. В случае трех — плоскость, и так далее. Эта граница задается в зависимости от имеющихся исходных данных и обучающего алгоритма. Чтобы все работало, точки исходных данных должны разделяться линейной границей на две вышеупомянутых области. Если точки исходных данных удовлетворяют этому требованию, то их можно назвать линейно разделяемыми. Посмотрите на изображение (рис. 2).

Указанная разделяющая плоскость называется линейным дискриминантом, так как она является линейной с точки зрения своей функции, и позволяет модели производить разделение, дискриминацию точек на различные классы.

Если невозможно произвести линейное разделение точек в исходном пространстве, стоит попробовать преобразовать векторы признаков в пространство с большим количеством измерений, добавив дополнительные эффекты взаимодействия, члены более высокой степени и т.д.

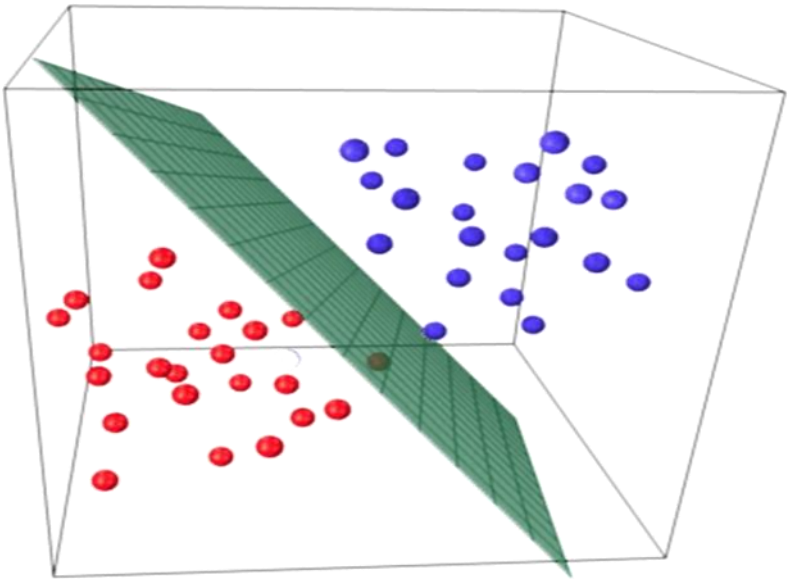


Рис 2. **Области логистической регрессии**

Использование линейного алгоритма в таком пространстве дает определенные преимущества для обучения нелинейной функции, поскольку граница становится нелинейной при возврате в исходное пространство.

Главное отличие логистической регрессии от линейной является то, что в логистической регрессии ответ является конкретной константой. Тогда как в линейной регрессии ответ предсказания имеет непрерывный вид. В качестве примера к линейной регрессии можно отнести решение задачи покупки товара клиентом, а в логистической регрессии примером может послужить предсказание цены на товар, акцию, дом. Зависимую переменную в логистической регрессии традиционно кодируют как 0–1, где 0 обозначает отсутствие какой-то характеристики, а 1 – ее наличие.

Логистическая регрессия в расчетах использует метод максимального правдоподобия (maximum likelihood estimation), тогда как линейная регрессия использует метод наименьших квадратов. Данный метод является весьма эффективным средством решения задачи оценивания параметров случайного процесса по данным наблюдений при известном виде законов распределения соответствующих случайных выбо-

рок, зависящих от оцениваемых параметров. Метод был проанализирован, рекомендован и значительно популяризирован Р. Фишером.

Задание: На основании исходных агрегированных данных (таблица 2) для изучения взаимосвязи прибыли (*Y*) и ресурсных факторов (*X*) по районам Могилевской области необходимо провести следующие расчеты:

1. С использованием методики, изученной при выполнении лабораторной работы № 1, осуществить генерацию данных для 100-150 наблюдений.

2. С помощью пакета анализа данных в EXCEL провести расчеты стандартной линейной регрессии $y=\sum k_i x_i$.

3. По критерию коэффициента эффективности использования ресурсов, как отношения фактического значения Y к расчетному ($Y_{факт}/Y_{расч.}$) провести классификацию совокупности на две группы.

4. Построить косвенную группировку по результатам корреляционно-регрессионного анализа и определить принадлежность данных по отдельным наблюдений к конкретному классу эффективности.

5. С использованием логистической регрессии провести эмулирование нейронной сети и определить вероятность принадлежности данных по отдельным наблюдениям к выделенному классу эффективности.

Порядок выполнения задания: Имеется набор переменных *X* (столбцы B-L) и одно выходное значение (столбец *Y*). Выход – это результат расчета нейросети в одном из нейросетевых пакетов. Проблема состоит в том, что нет возможности постоянного в него загонять и перерасчитывать данные – потоки динамические. В связи с этим сэмулирован «черный ящик» стандартной линейной регрессией (функция «Регрессия» пакета анализа) Excel (см. рис.3). То есть результат работы реальной нейросети с логистической функцией $y=1/(1+exp(-\sum k_i x_i)$ можно в первом приближении описать линейной функцией $y=\sum k_i x_i$ с учетом того, что точные *y* уже рассчитаны (см. рис.4).

Расчетные значения y получим в ходе регрессионных расчетов (рис.3). Коэффициент эффективности согласно предложенному фрагменту расчетов определен с помощью формулы EXCEL: =L2/M2 (рис.4), скопированной для всех 100 наблюдений. Эмулирование нейронной сети для расчета вероятности достижения коэффициентом эффективности соответствующих значений осуществлено с использованием функции EXCEL: =1/(1+EXP((-1)*ЛИНЕЙН ($N$2: $N$56; $A$2: $K$56;ИСТИНА;ИСТИНА))).

Таблица 2. **Исходные данные к лабораторной работе № 2**

Районы	X_1 Амортизация, тыс.руб	X_2 Оборотные активы, тыс.руб.	X_3 Долгосрочные кредиты, тыс.руб.	X_4 Энергетические мощности, тыс.л.с.	X_5 Субсидии и дотации, тыс.руб.
Белыничский	2144,06	2253,81	1896,49	104	19549
Бобруйский	3379,79	1419,01	1635,29	106	25202
Быховский	2052,07	1110,72	966,58	80	22353
Горецкий	3866,44	2100,35	2243,26	136,5	29303
Глусский	2120,14	659,23	928,73	70	12750
Дрибинский	1447,81	779,75	1219,53	63	16768
Кировский	3171,83	2233,9	2137,44	84	15187
Климовичский	1192,64	643,87	687,31	39	10684
Кличевский	2522,17	1414,85	945,69	79	15765
Костюковичский	1132,98	749,75	918,06	56	14951
Краснопольский	783,57	329,81	520,49	36	11626
Кричевский	1459,44	789,01	925,39	61	13080
Круглянский	1137,42	923,43	1623,64	70	14067
Могилевский	6509,15	3258,51	2787,91	161	46773
Мстиславльский	2827,16	1840,13	2237,55	155	29035
Осиповичский	2246,35	814,66	824,47	64	11288
Славгородский	1949,65	815,67	1093,74	87,5	25861
Хотимский	2054,84	1107,6	1330,97	61	16348
Шкловский	3926,61	4374,03	6331,85	150	27016
Чаусский	2584,35	1266,96	1354,08	100	27956
Чериковский	210,83	266,03	395,71	17,9	6038

Источник: данные Национального статистического комитета Республики Беларусь.

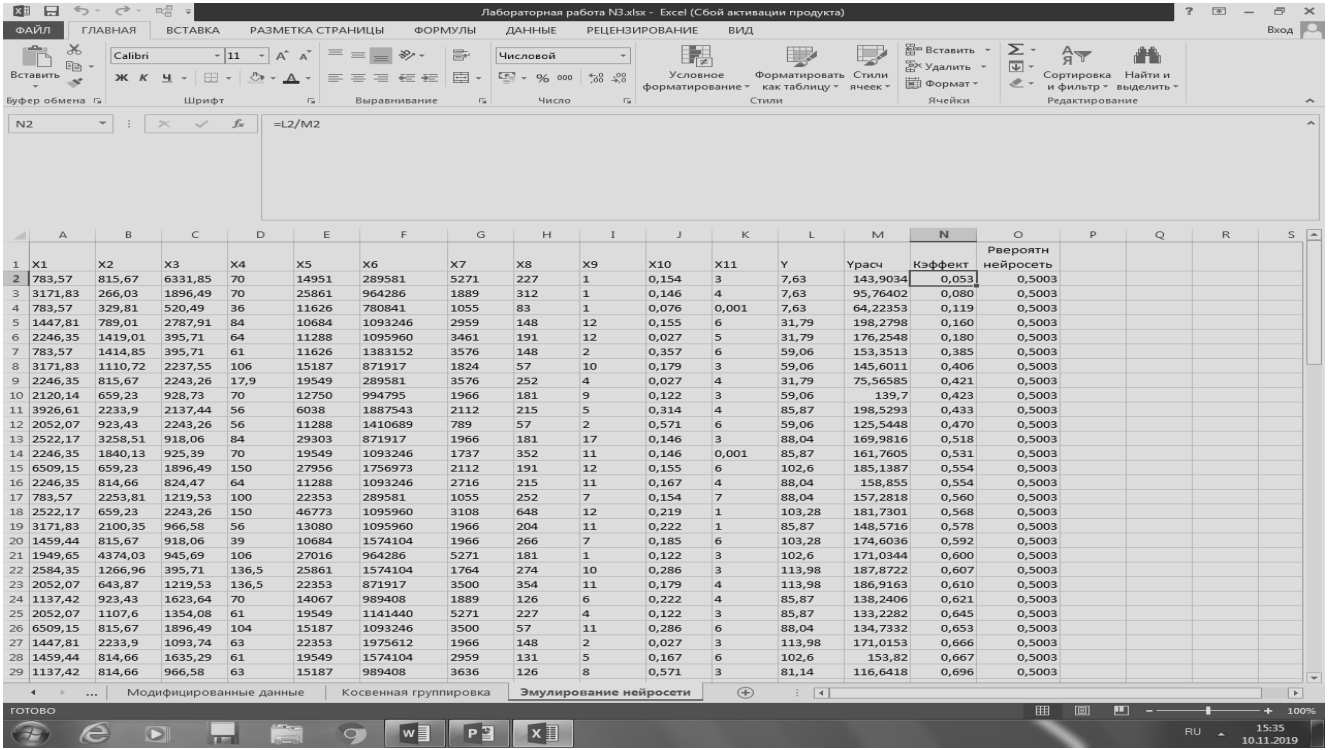
Таблица 3. **Исходные данные к лабораторной работе № 2 (продолжение)**

Районы	X6 Земельные ресурсы, б-га	X7 Затраты труда, тыс. чел. -ч	X8 Персонал управления	X9 Доступ в Интернет	X10 Удельный вес специалистов с высшим образованием	X11 Наличие программных продуктов	$У_{прибыль}$
Белыничский	1383152	3500	252	2	0,179	5	286,03
Бобруйский	1574104	3576	274	11	0,151	5	202,99
Быховский	1410689	3108	227	4	0,146	4	146,84
Горецкий	1975612	3636	352	11	0,571	6	241,03
Глусский	994795	1966	181	9	0,122	3	59,06
Дрибинский	964286	2112	148	7	0,412	3	102,6
Кировский	1193161	3461	312	8	0,286	6	191,84
Климовичский	871917	1737	131	5	0,179	3	81,14
Кличевский	1095960	2959	249	10	0,155	5	259,14
Костюковичский	970061	1764	130	5	0,185	1	137,47
Краснопольский	780841	1055	83	1	0,076	0,001	7,63
Кричевский	923064	1824	176	4	0,314	1	113,98
Круглянский	989408	1889	126	6	0,222	4	85,87
Могилевский	2315634	6215	648	17	0,363	10	317,73
Мстиславльский	1906365	5271	354	12	0,219	11	301,32
Осиповичский	1093246	2716	215	11	0,167	4	88,04
Славгородский	1141440	2544	204	10	0,06	7	154,67
Хотимский	1056316	2539	191	2	0,027	3	103,28
Шкловский	1887543	4420	425	16	0,357	6	362,78
Чаусский	1756973	3753	266	6	0,14	3	190,59
Чериковский	289581	789	57	1	0,154	0,001	31,79

Источник: данные Национального статистического комитета Республики Беларусь.

ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,591664					
R-квадрат	0,350066					
Нормированный R-квадрат	0,268825					
Стандартная ошибка	76,7912					
Наблюдения	100					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	11	279503,4048	25409,40044	4,308950008	3,92463E-05	
Остаток	88	518926,2429	5896,889124			
Итого	99	798429,6477				
	Коэффициен	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	15,29753	32,63871008	0,468692957	0,640448106	-49,56504422	80,16011128
X1	-0,00796	0,006566412	-1,212875471	0,228423288	-0,021013604	0,005085124
X2	0,016437	0,010373145	1,584554072	0,116654962	-0,004177637	0,037051255
X3	0,008324	0,005902362	1,410284915	0,161980517	-0,003405692	0,020053715
X4	0,595072	0,281290581	2,115505223	0,037212725	0,036065773	1,154077613
X5	-0,0022	0,001056023	-2,081353316	0,040306958	-0,004296581	-9,93332E-05
X6	4,78E-05	1,99446E-05	2,394395957	0,018770078	8,11955E-06	8,73908E-05
X7	0,001769	0,006806018	0,259901663	0,79554674	-0,011756636	0,015294427
X8	0,121216	0,079691529	1,521062367	0,131829392	-0,037154382	0,279585953
X9	3,209975	1,897236204	1,691921492	0,09419924	-0,560383571	6,980332988
X10	-37,9792	59,35880823	-0,639824399	0,523949054	-155,9423718	79,98394416
X11	3,969884	3,191383408	1,243938139	0,216825679	-2,372320363	10,31208744

Рис 3. **Фрагмент расчетов по лабораторной работе № 2 в EXCEL**



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Y	Yрасч	Кэффект	Вероятн нейросеть				
2	783,57	815,67	6331,85	70	14951	289581	5271	227	1	0,154	3	7,63	143,9034	0,053	0,5003				
3	3171,83	266,03	1896,49	70	25861	964286	1889	312	1	0,146	4	7,63	95,76402	0,080	0,5003				
4	783,57	329,81	520,49	36	11626	780841	1055	83	1	0,076	0,001	7,63	64,22353	0,119	0,5003				
5	1447,81	789,01	2787,91	84	10684	1093246	2959	148	12	0,155	6	31,79	198,2798	0,160	0,5003				
6	2246,35	1419,01	395,71	64	11288	1095960	3461	191	12	0,027	5	31,79	176,2548	0,180	0,5003				
7	783,57	1414,85	395,71	61	11626	1383152	3576	148	2	0,357	6	59,06	153,3513	0,385	0,5003				
8	3171,83	1110,72	2237,55	106	15187	871917	1824	57	10	0,179	3	59,06	145,6011	0,406	0,5003				
9	2246,35	815,67	2243,26	17,9	19549	289581	3576	252	4	0,027	4	31,79	75,56585	0,421	0,5003				
10	2120,14	659,23	928,73	70	12750	994795	1966	181	9	0,122	3	59,06	139,7	0,423	0,5003				
11	3926,61	2233,9	2137,44	56	6038	1887543	2112	215	5	0,314	4	85,87	198,5293	0,433	0,5003				
12	2052,07	923,43	2243,26	56	11288	1410689	789	57	2	0,571	6	59,06	125,5448	0,470	0,5003				
13	2522,17	3258,51	918,06	84	29303	871917	1966	181	17	0,146	3	88,04	169,9816	0,518	0,5003				
14	2246,35	1840,13	925,39	70	19549	1093246	1737	352	11	0,146	0,001	85,87	161,7605	0,531	0,5003				
15	6509,15	659,23	1896,49	150	27956	1756973	2112	191	12	0,155	6	102,6	185,1387	0,554	0,5003				
16	2246,35	814,66	824,47	64	11288	1093246	2716	215	11	0,167	4	88,04	158,855	0,554	0,5003				
17	783,57	2253,81	1219,53	100	22353	289581	1055	252	7	0,154	7	88,04	157,2818	0,560	0,5003				
18	2522,17	659,23	2243,26	150	46773	1095960	3108	648	12	0,219	1	103,28	181,7301	0,568	0,5003				
19	3171,83	2100,35	966,58	56	13080	1095960	1966	204	11	0,222	1	85,87	148,5716	0,578	0,5003				
20	1459,44	815,67	918,06	39	10684	1574104	1966	266	7	0,185	6	103,28	174,6036	0,592	0,5003				
21	1949,65	4374,03	945,69	106	27016	964286	5271	181	1	0,122	3	102,6	171,0344	0,600	0,5003				
22	2584,35	1266,96	395,71	136,5	25861	1574104	1764	274	10	0,286	3	113,98	187,8722	0,607	0,5003				
23	2052,07	643,87	1219,53	136,5	22353	871917	3500	354	11	0,179	4	113,98	186,9163	0,610	0,5003				
24	1137,42	923,43	1623,64	70	14067	989408	1889	126	6	0,222	4	85,87	138,2406	0,621	0,5003				
25	2052,07	1107,6	1354,08	61	19549	1141440	5271	227	4	0,122	3	85,87	133,2282	0,645	0,5003				
26	6509,15	815,67	1896,49	104	15187	1093246	3500	57	11	0,286	6	88,04	134,7332	0,653	0,5003				
27	1447,81	2233,9	1093,74	63	22353	1975612	1966	148	2	0,027	3	113,98	171,0153	0,666	0,5003				
28	1459,44	814,66	1635,29	61	19549	1574104	2959	131	5	0,167	6	102,6	153,82	0,667	0,5003				
29	1137,42	814,66	966,58	63	15187	989408	3636	126	8	0,571	3	81,14	116,6418	0,696	0,5003				

Рис 4. **Эмулирование нейронной сети**

3. СТАТИСТИЧЕСКИЕ МЕТОДЫ

Методические указания: Статистические методы в интеллектуальном анализе данных представляют собой четыре взаимосвязанных раздела:

предварительный анализ природы статистических данных (проверка гипотез стационарности, нормальности, независимости, однородности, оценка вида функции распределения, ее параметров и т.п.);

выявление связей и закономерностей (линейный и нелинейный регрессионный анализ, корреляционный анализ и др.);

многомерный статистический анализ (линейный и нелинейный дискриминантный анализ, кластерный анализ, компонентный анализ, факторный анализ и др.);

динамические модели и прогноз на основе временных рядов.

Задание: По данным таблиц 2 и 3 с помощью инструментария пакета анализа EXCEL провести следующие расчеты:

1. Показателей описательной статистики (дескриптивный анализ): среднее, медиана, мода, дисперсия, стандартное отклонение, коэффициенты асимметрии и эксцесса. Проверить гипотезу нормальности с использованием правила трех сигм и расчета ошибок асимметрии и эксцесса.

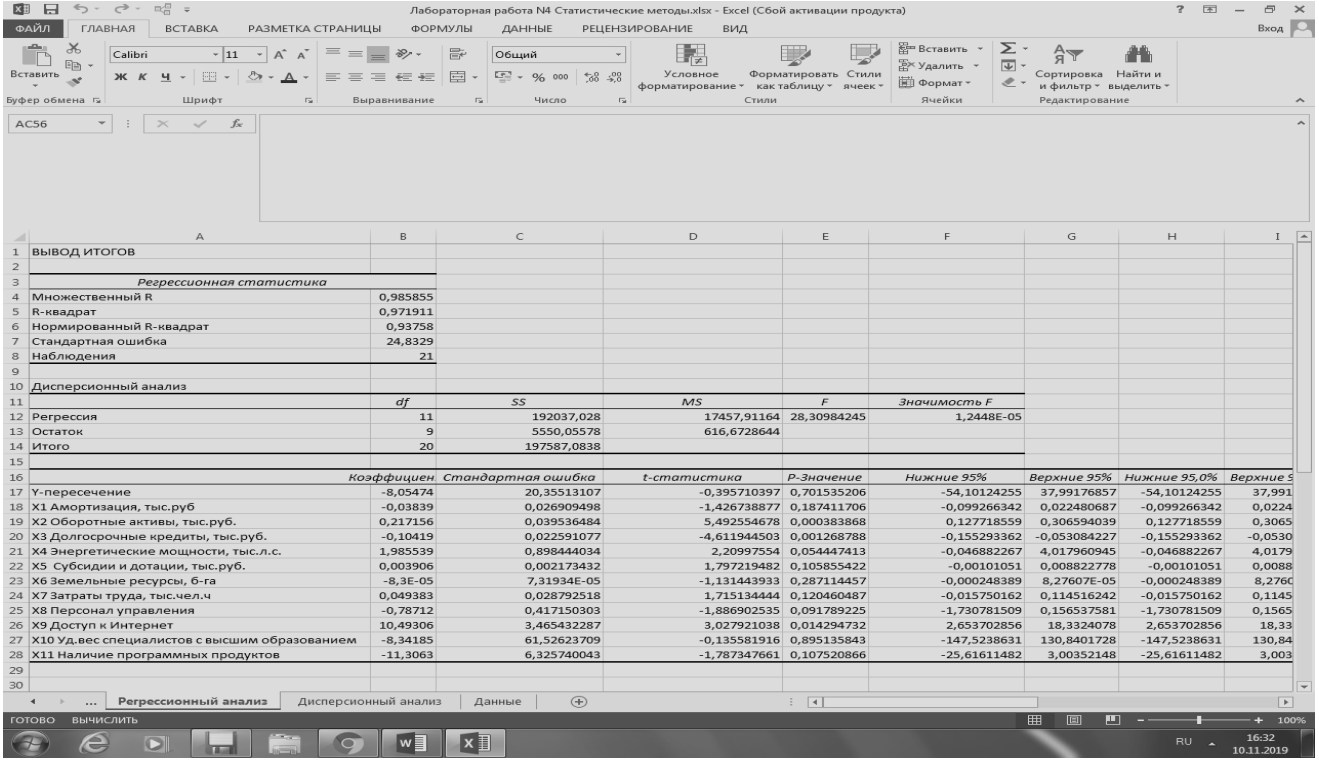
2. Построить матрицу коэффициентов корреляции. Проверить мультиколлинеарности коэффициентов.

3. Провести регрессионный анализ по агрегированным данным. Сравнить результаты регрессионных расчетов (рис. 5) с аналогичными расчетами в теме 2 (рис.3).

4. Провести однофакторный дисперсионный анализ.

Порядок выполнения задания: Экспортировать агрегированные данные (см. табл. 2 и 3) в EXCEL. Открыть надстройку «Пакет анализа». В надстройке EXCEL использовать функции: «Описательная статистика», «Корреляция», «Регрессия», «Однофакторный дисперсионный анализ». Провести расчету согласно заданию.

Сделать интерпретацию результатов расчетов с позиций интеллектуального анализа данных и роли статистических методов. Обсудить достоинства и недостатки статистических методов в интеллектуальном анализе данных.



	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,985855							
5	R-квадрат	0,971911							
6	Нормированный R-квадрат	0,93758							
7	Стандартная ошибка	24,8329							
8	Наблюдения	21							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	11	192037,028	17457,91164	28,30984245	1,2448E-05			
13	Остаток	9	5550,05578	616,6728644					
14	Итого	20	197587,0838						
15									
16		*Коэффициен*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 9*
17	Y-пересечение	-8,05474	20,35513107	-0,395710397	0,701535206	-54,10124255	37,99176857	-54,10124255	37,991
18	X1 Амортизация, тыс.руб	-0,03839	0,026909498	-1,426738877	0,187411706	-0,099266342	0,022480687	-0,099266342	0,0224
19	X2 Оборотные активы, тыс.руб.	0,217156	0,039536484	5,492554678	0,000383868	0,127718559	0,306594039	0,127718559	0,3065
20	X3 Долгосрочные кредиты, тыс.руб.	-0,10419	0,022591077	-4,611944503	0,001268788	-0,155293362	-0,053084227	-0,155293362	-0,0530
21	X4 Энергетические мощности, тыс.л.с.	1,985539	0,898444034	2,20997554	0,054447413	-0,046882267	4,017960945	-0,046882267	4,0179
22	X5 Субсидии и дотации, тыс.руб.	0,003906	0,002173432	1,797219482	0,105855422	-0,00101051	0,008822778	-0,00101051	0,0088
23	X6 Земельные ресурсы, б-га	-8,3E-05	7,31934E-05	-1,131443933	0,287114457	-0,000248389	8,27607E-05	-0,000248389	8,2760
24	X7 Затраты труда, тыс.чел.ч	0,049383	0,028792518	1,715134444	0,120460487	-0,015750162	0,114516242	-0,015750162	0,1145
25	X8 Персонал управления	-0,78712	0,417150303	-1,886902535	0,091789225	-1,730781509	0,156537581	-1,730781509	0,1565
26	X9 Доступ к Интернет	10,49306	3,465432287	3,027921038	0,014294732	2,653702856	18,3324078	2,653702856	18,33
27	X10 Уд.вес специалистов с высшим образованием	-8,34185	61,52623709	-0,135581916	0,895135843	-147,5238631	130,8401728	-147,5238631	130,84
28	X11 Наличие программных продуктов	-11,3063	6,325740043	-1,787347661	0,107520866	-25,61611482	3,00352148	-25,61611482	3,003
29									
30									

Рис 5. **Фрагмент расчетов по лабораторной работе № 3 в EXCEL**

4. НЕЙРОСЕТЕВЫЕ МОДЕЛИ

Методические указания: Развитие технологии баз данных и систем управления базами данных, способствует росту объема данных, хранящихся в базе. Эти данные содержат в себе много важной информации, которая имеет большой потенциал для агропромышленного производства. Ввиду этого многие агропромышленные организации используют технологию интеллектуального анализа данных (Data Mining), которая позволяет обрабатывать массивные базы данных и извлекать из них полезную информацию.

Задачей интеллектуального анализа данных является выявление латентных правил и закономерностей в наборах данных. Продолжительное время основным инструментом интеллектуального анализа данных была традиционная математическая статистика, но и она зачастую не в состоянии решить задачи из реальной жизни. Математическая статистика (тема 3) в основном полезна при проверке заранее сформулированных гипотез (verification-driven data mining).

Нейронным сетям в интеллектуальном анализе данных присущи следующие недостатки: сложная структура, плохая интерпретируемость и долгое время обучения. Однако их преимущества, такие как, высокая допустимость к зашумленным данным и низкий коэффициент ошибок, непрерывное усовершенствование и оптимизация различных алгоритмов обучения сетей, алгоритма извлечения правил, алгоритма упрощения сетей, делают нейронные сети все более и более перспективным направлением в Data Mining.

Задание: Построить и обучить нейронную сеть для экспертного выбора перспективного сорта сельскохозяйственных культур из 4 вариантов для четырех производственных ситуаций (таблица 4):

Таблица 4. **Исходные данные для лабораторной работы № 4**

№	In1	In2	In3	In4	Out
1	1	2	3	4	1
2	2	1	3	4	2
3	3	2	1	4	3
4	4	3	2	1	4

Порядок выполнения задания: В данном задании необходимо

построить простейшую трёхслойная нейронную сеть. Первый слой сети должен состоять из 4-х псевдонейронов (входы выделим синими стрелками) с весами *W11, W12, W13, W14*. Второй слой из 4-х псевдонейронов (входы выделим зеленым цветом) с весами *W21, W22, W23, W24*. Третий слой будет состоять из 1 псевдонейрона с весом *W13* (входы выделим малиновым цветом). Данные подаваемые в сеть и сравниваемые с выходом сети подвергаются нормированию.

Обучение сети производится с помощью надстройки Excel "Поиск решения...". Изначально веса можно взять случайным образом. Целевой ячейкой выбирается ошибка сети в нормированном виде EOutN. Целевое значение выбирается примерно 0,2 (требуемая ошибка распознавания данных). Изменяемые ячейки – это таблица весов W (табл.5):

Таблица 5. **Веса нейронной сети**

W				
W11	0,76	-2,26	-5,11	-4,85
W12	2,33	0,90	-2,01	-0,08
W13	-0,58	-2,29	-3,01	-1,01
W14	-1,69	-3,03	-3,93	-2,42
W21	-0,85	2,12	3,16	1,20
W22	-0,87	-2,01	-1,29	-0,63
W23	-0,73	0,65	1,78	0,91
W24	-8,46	-14,48	-9,10	-9,45
W31	1,63	-1,26	0,54	-33,07

Обучение производится последовательно:

а) производится "Поиск решения...";

б) индекс в ячейке A4 увеличивается на 1 (от 1 до 4 и снова 1);

в) на вход поступает новый вектор, возврат к пункту а).

Выход сети (обозначается красными стрелками) и ошибка сети (обозначается оранжевыми стрелками) представляются как в нормированном, так и в абсолютном виде. Построение уже обученной нейронной сети в EXCEL включает ряд последовательных шагов:

1) В новой книге EXCEL на отдельном листе создать закладку «Данные» и в область ячеек A1:F5 экспортировать таблицу 4 (без названия).

2) В другом листе создать закладку «Нейронная сеть». В ячейки K11:O20 экспортируем таблицу весов после обучения (см. табл.5).

3) В области ячеек A1:F2 создать таблицу входных данных. Вос-

пользовавшись функцией =МАКС(Данные!B:B) в ячейке C1 вывести максимальное значение по столбцу исходных данных (закладка данные). Так max “B”{1;2;3;4}=4. Скопировать формулу для столбцов данных C, D, E для ячеек D1, E1, F1. Воспользовавшись функцией =МИН(Данные!B:B) в ячейке C2 вывести минимальное значение по столбцу исходных данных (закладка данные). Так min “B”{1;2;3;4}=1. Скопировать формулу для столбцов данных C, D, E для ячеек D2, E2, F2. В ячейках A1 и A2 воспользовавшись выше приведенными формулами найдем максимальное и минимальное значение в столбце A исходных данных.

4) В ячейки C4:F4 последовательно занесем формулы: =ВПР($A4;Данные!$A:$F;2;1);ВПР($A4;Данные!$A:$F;3;1); ВПР($A4; Данные!$A:$F;4;1); =ВПР($A4;Данные!$A:$F;5;1). ФУНКЦИЯ ВПР используется, если нужно найти элементы в таблице или диапазоне по строкам.

5) В ячейки C5:F5 последовательно занесем формулу =(C4-C2)/(C1-C2). Она отражает отношение разницы между выбранным нейронной сетью с помощью функции ВПР числом и введенным числом к разнице между максимальным и минимальным значением на входе нейросети. Эти значения дублируются в строке C7:F7 построенного нейрона «Neuron 11».

6) Веса нейрона «11» дублируются из таблицы весов. По ячейкам C9: F9 заносим формулу произведения элементов строки C7:F7 на веса нейрона. Net 11 (ячейка C10) получен, как сумма выше приведенных произведений. Выход нейрона Out 11 формируется с использованием логистической функции (см. тему 2): =1/(1+EXP(-C10)).

7) Аналогичным образом строятся нейроны для остальных слоев. Причем выходная информация с нейронов первого слоя служит для входной информации для нейронов второго слоя и т.д.

8) В ячейках H1:I8 закладки «Нейронная сеть» формируется входящая и исходящая информация для ситуаций 1,2,3,4 (закладка «Данные», ячейки F2:F5. Также, как для таблиц слева используются функции EXCEL: МИН, МАКС, ВПР.

9) Out NN (0,11) дублирует Out Net с третьего слоя нейрона. F(Out NN) находится с помощью функции: =I7*(I1-I2)+I2.

10) Ошибка нейронной сети находится по формуле: =ABS(O1-I7). При этом O1=0. При этом EFOutN находится, как: =I4-I8.

Результаты построения нейронной сети представлены на рисунке 6:

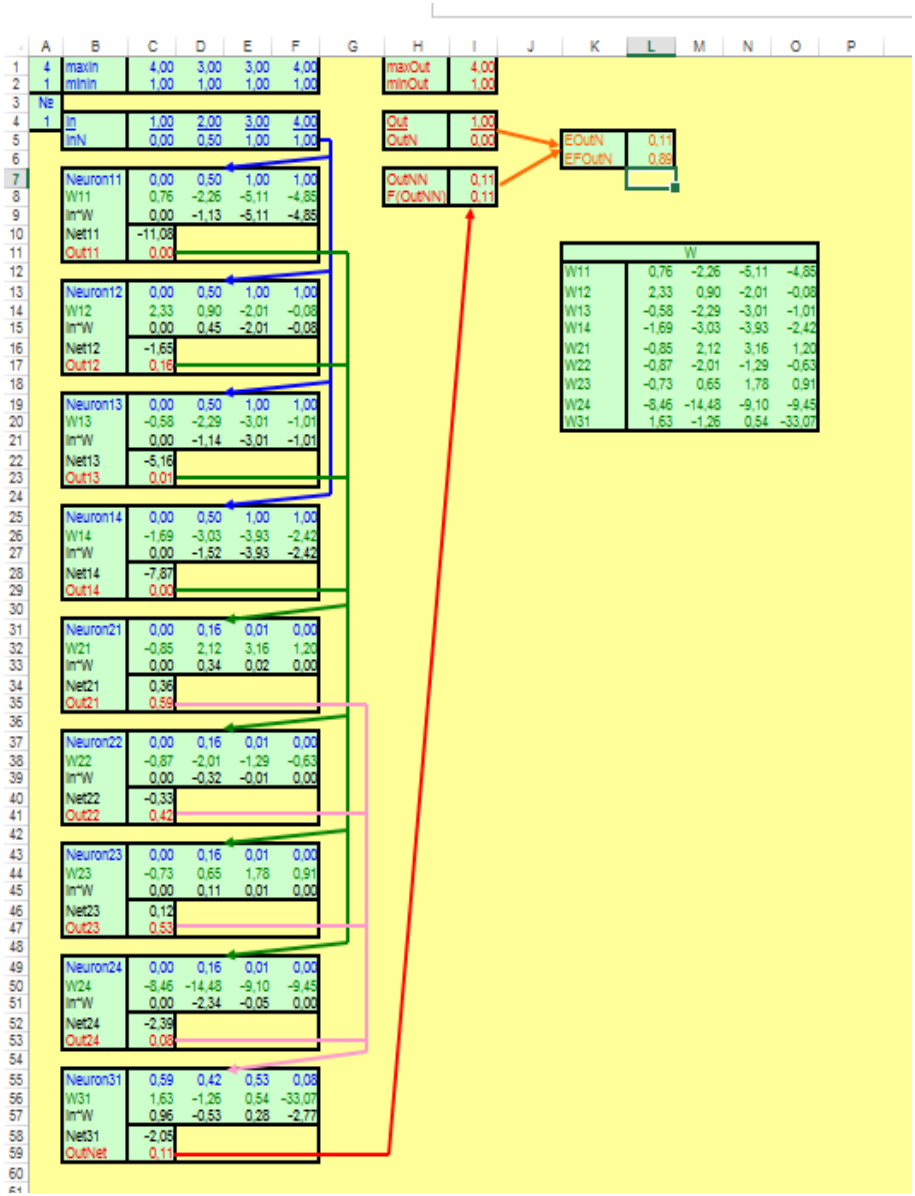


Рис 6. **Результат построения нейронной сети в EXCEL**

5. МЕТОДЫ КЛАССИФИКАЦИИ: ДЕРЕВО РЕШЕНИЙ

Методические указания: Деревья решений – один из методов автоматического анализа данных. Первые идеи создания деревьев решений восходят к работам Ховленда (Hoveland) и Ханта(Hunt) конца 50-х годов XX века. Однако, основополагающей работой, давшей импульс для развития этого направления, явилась книга Ханта (Hunt, E.B.), Мэрина (Marin J.) и Стоуна (Stone, P.J) "Experiments in Induction", увидевшая свет в 1966г. Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение. Под правилом понимается логическая конструкция, представленная в виде "если ... то ...".

Область применения деревья решений в настоящее время широка, но все задачи, решаемые этим аппаратом могут быть объединены в следующие три класса:

Описание данных: Деревья решений позволяют хранить информацию о данных в компактной форме, вместо них мы можем хранить дерево решений, которое содержит точное описание объектов.

Классификация: Деревья решений отлично справляются с задачами классификации, т.е. отнесения объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.

Регрессия: Если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых(входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

На сегодняшний день существует значительное число алгоритмов, реализующих деревья решений CART, C4.5, NewId, ITrule, CHAID, CN2 и т.д. Но наибольшее распространение и популярность получили следующие два:

CART (Classification and Regression Tree) – это алгоритм построения бинарного дерева решений – дихотомической классификационной модели. Каждый узел дерева при разбиении имеет только двух потомков. Как видно из названия алгоритма, решает задачи классификации и регрессии.

C4.5 – алгоритм построения дерева решений, количество потомков у узла не ограничено. Не умеет работать с непрерывным целевым полем, поэтому решает только задачи классификации.

Задание 1. Фермер может выращивать либо кукурузу, либо пшеницу. Вероятность того, что цены на будущий урожай этих культур повысятся, останутся на том же уровне или понизятся, равна соответственно 0,25, 0,30 и 0,45. Если цены возрастут, урожай кукурузы даст

30 000 руб. чистого дохода, а урожай пшеницы — 10 000 руб. Если цены останутся неизменными, фермер лишь покроет расходы. Но если цены станут ниже, урожай кукурузы и пшеницы приведет к потерям в 35 000 и 5 000 руб. соответственно. Постройте дерево решений. Какую культуру следует выращивать фермеру? Каково ожидаемое значение его прибыли?

Порядок выполнения задания:

1. Строим дерево решений в EXCEL (рис. 7):

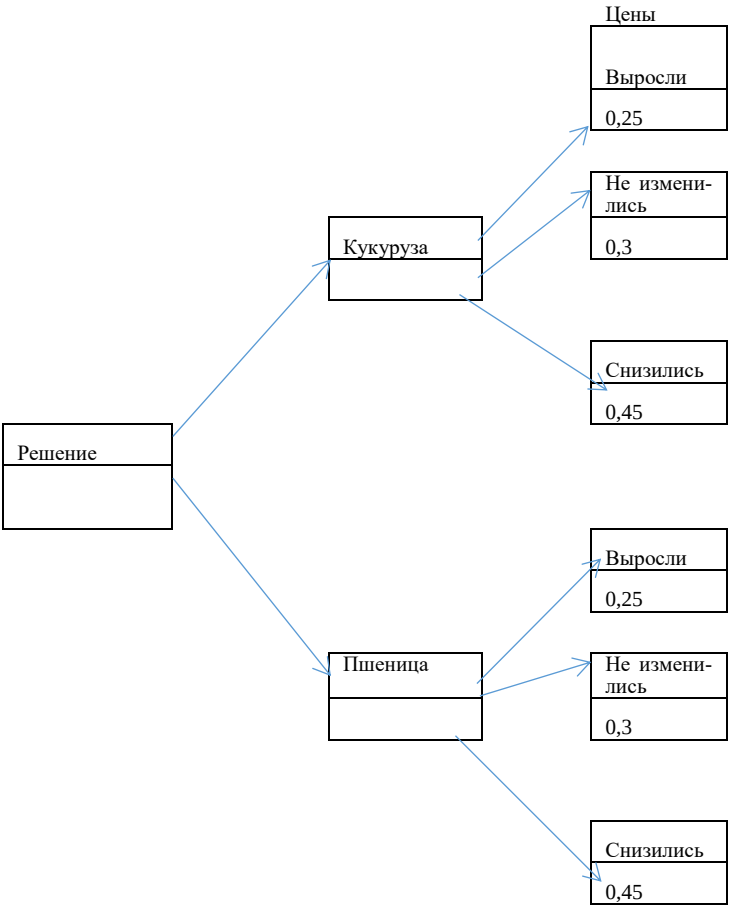


Рис 7. **Результат построения дерева решений в EXCEL**

2. Рассчитываем доход по каждой культуре (рис.8) согласно методологии Байеса (произведение вероятности на возможный доход или убытки)

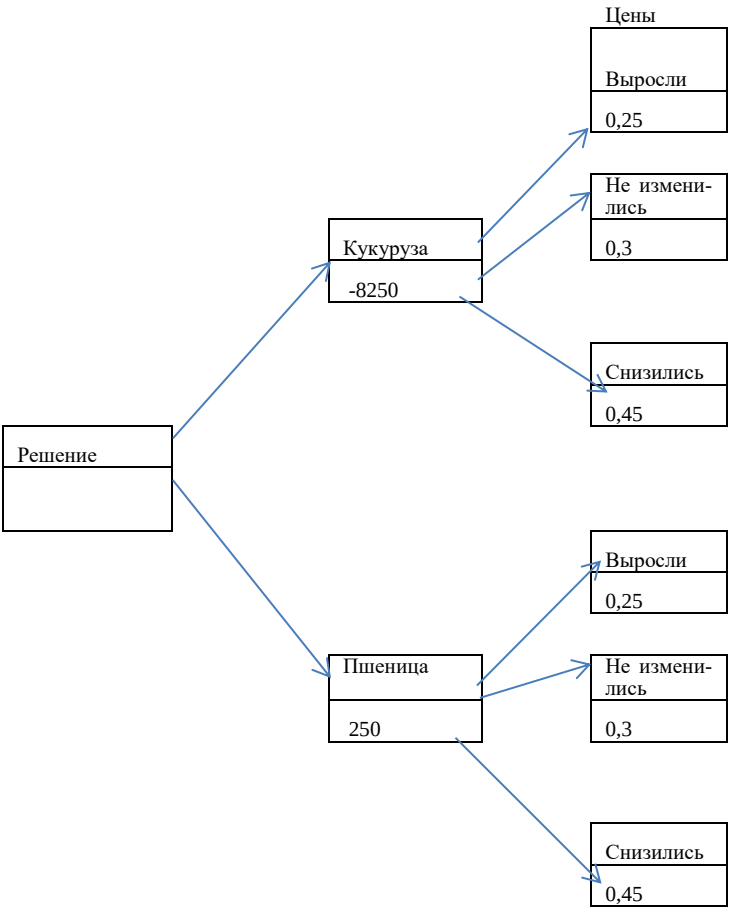


Рис 8. **Результат расчета дохода по дереву решений в EXCEL**

3. Формируем функцию принятия хозяйственного решения:

Решение	Выбираем	**Пшеница**	доход	**250,00**

Задание 2:

Предприятие рассматривает варианты капитальных вложений. Первый вариант предусматривает строительство нового цеха для увеличения объема выпуска продукции стоимостью M1 = 500 млн. руб. При этом варианте возможны большой спрос (годовой доход в размере R1 = 230 млн. руб. в течение 5 последующих лет) с вероятностью p1 = 0,7 и низкий спрос (ежегодные убытки R2 = 90 млн. руб. с вероятностью p2 = 0,3.

Второй вариант предусматривает создание нового предприятия для выпуска новой продукции Стоимостью M1 = 700 млн. руб. При этом варианте возможны большой спрос (годовой доход в размере R1 = 450 млн. руб. в течение 5 последующих лет) с вероятностью p1 = 0,6 и низкий спрос (ежегодные убытки R2 = 150 млн. руб. с вероятностью p2 = 0,4.

При третьем варианте предлагается отложить инвестиции на 1 год для сбора дополнительной информации, которая может быть позитивной или негативной с вероятностью p1 = 0,8 и p2 = 0,2 соответственно. В случае позитивной информации можно осуществить инвестиции по указанным выше расценкам, в вероятности большого и низкого спроса меняются на p1 = 0,9 и p2 = 0,1 соответственно. Доходы на последующие годы остаются на том же уровне. В случае негативной информации инвестиции осуществляться не будут.

Все расчеты выражены в текущих ценах и не должны дисконтироваться. Нарисовать дерево решений. Определить наиболее эффективную последовательность действий, основываясь на ожидаемых доходах. Какова ожидаемая стоимостная оценка наилучшего решения?

Задание 3:

Рассматривается проект покупки доли (пакета акций) в инвестиционном проекте. Пакет стоит 7 млн., и по завершению проект принесет доход 12 млн. с вероятностью 0,6 или ничего с вероятностью 0,4.

При этом через некоторое время будет опубликован прогноз аналитической фирмы относительно успеха этого проекта. Прогноз верен с вероятностью 0,7, то есть, равны 0,7 условные вероятности.

Однако, в случае положительного прогноза пакет порождает до 10,6 млн., а в случае отрицательного подешевеет до 3,4 млн. Требуется составить стратегию действий: покупать ли долю, или ждать прогноза, и совершать ли покупку при том или ином результате прогноза.

6. КЛАСТЕРНЫЙ АНАЛИЗ

Методические указания: Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Применение кластерного анализа в общем виде сводится к следующим этапам:

- отбор выборки объектов для кластеризации;
- определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных;
- вычисление значений меры сходства между объектами;
- применение метода кластерного анализа для создания групп сходных объектов (кластеров);
- представление результатов анализа.

После получения и анализа результатов возможна корректировка выбранной метрики и метода кластеризации до получения оптимального результата.

Итак, как же определять «похожесть» объектов? Для начала нужно составить вектор характеристик для каждого объекта — как правило, это набор числовых значений, например, рост-вес человека. Однако существуют также алгоритмы, работающие с качественными (т.н. категорийными) характеристиками.

После того, как мы определили вектор характеристик, можно провести нормализацию, чтобы все компоненты давали одинаковый вклад при расчете «расстояния». В процессе нормализации все значения приводятся к некоторому диапазону, например, [-1, 1] или [0, 1].

Наконец, для каждой пары объектов измеряется «расстояние» между ними — степень похожести. Существует множество метрик, вот лишь основные из них:

Евклидово расстояние. Наиболее распространенная функция расстояния. Представляет собой геометрическим расстоянием в многомерном пространстве.

Квадрат евклидова расстояния. Применяется для придания большего веса более отдаленным друг от друга объектам.

Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т.к.

они не возводятся в квадрат).

Расстояние Чебышева. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате.

Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются.

Задание: По данным таблиц 2 и 3 (тема 2) произвести кластер-анализ показателей прибыли и используемых ресурсов. В качестве метрики кластеризации использовать эвклидово расстояние. Опробовать алгоритмы кластеризации ближнего и дальнего соседа. Сравнить результаты кластеризации и классификации с использованием логистической регрессии (тема 2).

Порядок выполнения задания:

1. Используя результаты описательной статистики (тема 3) по данным таблиц 2 и 3 (ячейки в EXCEL: A1:V22), в частности, рассчитанные значения средней и стандартного отклонения (таблицы 6, 7), проведем нормализацию исходных данных (таблица 8, 9). Для этого используем формулу:

$$z_i = \frac{x_i - \bar{x}_i}{\sigma_{xi}}$$

где, z_i – нормализованные данные показателя по наблюдению i, $i=1 \div n$;

x_i – исходные данные показателя по наблюдению i;

$\bar{x}_i$ – его среднее значение;

σ_{xi} – среднеквадратическое (стандартное) отклонение x от среднего.

2. Составим таблицу расстояний между объектами по нормализованным данным (таблицы 10, 11). Для расчета расстояний между кластерами используем формулу Евклидового расстояния.

3. Проведем кластеризацию по алгоритму «ближнего» соседа. Найдем два объекта, расстояние между которыми наименьшее (НАИМЕНЬШИЙ(B2:V2;СЧЁТЕСЛИ(B2:V2;0)+1)). Затем перейдем к следующим объектам (таблицы 12, 13). Выделим однородные группы объектов по критерию наименьшей вариации данных 5-10%.

4. Так же опробуем алгоритм «дальнего» соседа. только в этом случае будем выбирать объекты с наибольшим расстоянием между ними.

5. Сравним результаты кластеризации (для показателей Y и X_1) по алгоритмам (таблица 14), а также классификации объектов по коэффициенту эффективности с помощью статистических методов.

Таблица 6. **Исходные данные к лабораторной работе № 6**

Районы	X_1 Амортизация, тыс. руб.	X_2 Оборотные активы, тыс. руб.	X_3 Долгосрочные кредиты, тыс. руб.	X_4 Энергетические мощности, тыс. л. с.	X_5 Субсидии и дотации, тыс. руб.
Белыничский	2144,06	2253,81	1896,49	104	19549
Бобруйский	3379,79	1419,01	1635,29	106	25202
Быховский	2052,07	1110,72	966,58	80	22353
Горецкий	3866,44	2100,35	2243,26	136,5	29303
Глусский	2120,14	659,23	928,73	70	12750
Дрибинский	1447,81	779,75	1219,53	63	16768
Кировский	3171,83	2233,9	2137,44	84	15187
Климовичский	1192,64	643,87	687,31	39	10684
Кличевский	2522,17	1414,85	945,69	79	15765
Костюковичский	1132,98	749,75	918,06	56	14951
Краснопольский	783,57	329,81	520,49	36	11626
Кричевский	1459,44	789,01	925,39	61	13080
Круглянский	1137,42	923,43	1623,64	70	14067
Могилевский	6509,15	3258,51	2787,91	161	46773
Мстиславльский	2827,16	1840,13	2237,55	155	29035
Осиповичский	2246,35	814,66	824,47	64	11288
Славгородский	1949,65	815,67	1093,74	87,5	25861
Хотимский	2054,84	1107,6	1330,97	61	16348
Шкловский	3926,61	4374,03	6331,85	150	27016
Чаусский	2584,35	1266,96	1354,08	100	27956
Чериковский	210,83	266,03	395,71	17,9	6038
Среднее значение	2319,97	1388,15	1571,63	84,80	19600,00
Среднеквадратическое (стандартное) отклонение	1368,99	1005,88	1260,31	39,39	9260,29

Источник: данные Национального статистического комитета Республики Беларусь. Результаты расчетов.

Таблица 7. **Исходные данные к лабораторной работе № 6 (продолжение)**

Районы	X_6 Земельные ресурсы, балло-га	X_7 Затраты труда, тыс. чел. -ч	X_8 Персонал управления, чел.	X_9 Доступ в Интернет, ед.	X_{10} Удельный вес специалистов с высшим образованием, %	X_{11} Наличие программных продуктов, ед.	$У_{прибыль}$, тыс. руб.
Белыничский	1383152	3500	252	2	0,179	5	286,03
Бобруйский	1574104	3576	274	11	0,151	5	202,99
Быховский	1410689	3108	227	4	0,146	4	146,84
Горецкий	1975612	3636	352	11	0,571	6	241,03
Глусский	994795	1966	181	9	0,122	3	59,06
Дрибинский	964286	2112	148	7	0,412	3	102,6
Кировский	1193161	3461	312	8	0,286	6	191,84
Климовичский	871917	1737	131	5	0,179	3	81,14
Кличевский	1095960	2959	249	10	0,155	5	259,14
Костюковичский	970061	1764	130	5	0,185	1	137,47
Краснопольский	780841	1055	83	1	0,076	0,001	7,63
Кричевский	923064	1824	176	4	0,314	1	113,98
Круглянский	989408	1889	126	6	0,222	4	85,87
Могилевский	2315634	6215	648	17	0,363	10	317,73
Мстиславльский	1906365	5271	354	12	0,219	11	301,32
Осиповичский	1093246	2716	215	11	0,167	4	88,04
Славгородский	1141440	2544	204	10	0,06	7	154,67
Хотимский	1056316	2539	191	2	0,027	3	103,28
Шкловский	1887543	4420	425	16	0,357	6	362,78
Чаусский	1756973	3753	266	6	0,14	3	190,59
Чериковский	289581	789	57	1	0,154	0,001	31,79
Среднее значение	1265435,62	2896,86	238,14	7,52	0,21	4,29	165,04
Среднеквадратическое (стандартное) отклонение	484055,32	1338,59	132,11	4,59	0,13	2,85	99,39

Источник: данные Национального статистического комитета Республики Беларусь. Результаты расчетов.

Таблица 8. **Нормализованные данные**

Районы	Z_1 Амортизация, тыс. руб.	Z_2 Оборотные активы, тыс. руб.	Z_3 Долгосрочные кредиты, тыс. руб.	Z_4 Энергетические мощности, тыс. л. с.	Z_5 Субсидии и дотации, тыс. руб.
Белыничский	-0,128	0,861	0,258	0,487	-0,006
Бобруйский	0,774	0,031	0,051	0,538	0,605
Быховский	-0,196	-0,276	-0,480	-0,122	0,297
Горецкий	1,130	0,708	0,533	1,312	1,048
Глусский	-0,146	-0,725	-0,510	-0,376	-0,740
Дрибинский	-0,637	-0,605	-0,279	-0,554	-0,306
Кировский	0,622	0,841	0,449	-0,020	-0,477
Климовичский	-0,823	-0,740	-0,702	-1,163	-0,963
Кличевский	0,148	0,027	-0,497	-0,147	-0,414
Костюковичский	-0,867	-0,635	-0,519	-0,731	-0,502
Краснопольский	-1,122	-1,052	-0,834	-1,239	-0,861
Кричевский	-0,629	-0,596	-0,513	-0,604	-0,704
Круглянский	-0,864	-0,462	0,041	-0,376	-0,597
Могилевский	3,060	1,859	0,965	1,934	2,934
Мстиславльский	0,370	0,449	0,528	1,782	1,019
Осиповичский	-0,054	-0,570	-0,593	-0,528	-0,898
Славгородский	-0,271	-0,569	-0,379	0,068	0,676
Хотимский	-0,194	-0,279	-0,191	-0,604	-0,351
Шкловский	1,174	2,968	3,777	1,655	0,801
Чаусский	0,193	-0,120	-0,173	0,386	0,902
Чериковский	-1,541	-1,116	-0,933	-1,698	-1,465

Источник: результаты расчетов в EXCEL.

Таблица 9. **Нормализованные данные (продолжение)**

Районы	Z_6 Земельные ресурсы, балло-га	Z_7 Затраты труда, тыс. чел. -ч	Z_8 Персонал управления, чел.	Z_9 Доступ в Интернет, ед.	Z_{10} Удельный вес специалистов с высшим образованием, %	Z_{11} Наличие программных продуктов, ед.	$Z_{прибыль}$, тыс. руб.
Белыничский	0,243	0,451	0,105	-1,204	-0,266	0,251	1,217
Бобруйский	0,638	0,507	0,271	0,757	-0,482	0,251	0,382
Быховский	0,300	0,158	-0,084	-0,768	-0,520	-0,100	-0,183
Горецкий	1,467	0,552	0,862	0,757	2,752	0,602	0,765
Глусский	-0,559	-0,695	-0,433	0,322	-0,705	-0,451	-1,066
Дрибинский	-0,622	-0,586	-0,682	-0,114	1,528	-0,451	-0,628
Кировский	-0,149	0,421	0,559	0,104	0,558	0,602	0,270
Климовичский	-0,813	-0,866	-0,811	-0,550	-0,266	-0,451	-0,844
Кличевский	-0,350	0,046	0,082	0,540	-0,451	0,251	0,947
Костюковичский	-0,610	-0,846	-0,819	-0,550	-0,220	-1,154	-0,277
Краснопольский	-1,001	-1,376	-1,174	-1,422	-1,059	-1,504	-1,584
Кричевский	-0,707	-0,801	-0,470	-0,768	0,773	-1,154	-0,514
Круглянский	-0,570	-0,753	-0,849	-0,332	0,065	-0,100	-0,797
Могилевский	2,170	2,479	3,102	2,065	1,150	2,006	1,536
Мстиславльский	1,324	1,774	0,877	0,975	0,042	2,357	1,371
Осиповичский	-0,356	-0,135	-0,175	0,757	-0,359	-0,100	-0,775
Славгородский	-0,256	-0,264	-0,258	0,540	-1,182	0,953	-0,104
Хотимский	-0,432	-0,267	-0,357	-1,204	-1,436	-0,451	-0,621
Шкловский	1,285	1,138	1,414	1,847	1,104	0,602	1,989
Чаусский	1,015	0,640	0,211	-0,332	-0,566	-0,451	0,257
Чериковский	-2,016	-1,575	-1,371	-1,422	-0,459	-1,504	-1,341

Источник: результаты расчетов.

Таблица 10. **Матрица расстояний между объектами кластеров**

Районы	Белыничский район	Бобруйский район	Быховский район	Горецкий район	Глусский район	Дрибинский район	Кировский район	Климовичский район	Кличевский район	Костюковичский район	Краснопольский район	Кричевский район
Белыничский район	0	1,230	1,402	1,337	2,284	1,914	1,209	2,175	**0,387**	1,667	2,972	1,802
Бобруйский район	1,230	0,000	1,122	0,522	1,716	1,735	**0,189**	2,014	0,844	1,769	2,731	1,664
Быховский район	1,402	1,122	0,000	2,017	2,467	2,143	1,615	2,441	**0,665**	1,968	3,288	2,032
Горецкий район	1,337	**0,522**	2,017	0,000	2,231	2,250	0,709	2,530	0,999	2,252	3,253	2,174
Глусский район	2,284	1,716	2,467	2,231	0,000	0,658	1,541	0,713	2,034	1,069	1,105	0,734
Дрибинский район	1,914	1,735	2,143	2,250	0,658	0,000	1,547	0,285	1,760	0,419	1,072	**0,115**
Кировский район	1,209	**0,189**	1,615	0,709	1,541	1,547	0,000	1,825	0,827	1,587	2,545	1,476
Климовичский район	2,175	2,014	2,441	2,530	0,713	0,285	1,825	0,000	2,037	0,568	0,798	0,384
Кличевский район	**0,387**	0,844	0,665	0,999	2,034	1,760	0,827	2,037	0,000	1,590	2,831	1,654
Костюковичский район	1,667	1,769	1,968	2,252	1,069	0,419	1,587	0,568	1,590	0,000	1,331	**0,336**
Краснопольский район	2,972	2,731	3,288	3,253	1,105	1,072	2,545	0,798	2,831	1,331	0,000	**0,547**
Кричевский район	1,802	1,664	2,032	2,174	0,734	**0,115**	1,476	0,384	1,654	0,336	0,547	0,000
Круглянский район	2,144	2,018	2,421	2,532	0,767	0,282	1,829	**0,062**	2,015	0,519	0,580	0,368
Могилевский район	3,204	2,561	4,698	2,079	4,129	4,284	2,747	4,555	2,971	4,326	4,559	4,220
Мстиславльский район	0,522	1,068	0,904	0,972	2,491	2,239	1,130	2,516	**0,479**	2,061	2,224	2,133
Осиповичский район	1,993	1,422	2,181	1,942	0,306	0,601	1,244	0,773	1,733	0,953	1,179	0,631
Славгородский район	1,329	1,152	1,514	1,648	0,970	0,639	0,968	0,924	1,131	0,621	0,869	0,544
Хотимский район	1,840	1,394	2,023	1,916	0,447	0,443	1,208	0,668	1,605	0,756	0,990	0,448
Шкловский район	1,514	1,657	2,180	1,226	3,328	3,183	1,806	3,467	1,463	3,050	3,226	3,084
Чаусский район	1,013	0,594	1,257	1,065	1,366	1,214	**0,429**	1,499	0,691	1,187	1,420	1,127
Чериковский район	2,922	2,885	3,384	3,400	1,421	1,151	2,696	0,872	2,843	1,259	1,143	1,231

Источник: результаты расчетов.

Таблица 11. **Матрица расстояний между объектами кластеров - продолжение**

Районы	Круглянский район	Могилевский район	Мстиславльский район	Осиповичский район	Славгородский район	Хотимский район	Шкловский район	Чаусский район	Чериковский район
Белыничский район	2,144	3,204	0,522	1,993	1,329	1,840	1,514	1,013	2,922
Бобруйский район	2,018	2,561	1,068	1,422	1,152	1,394	1,657	0,594	2,885
Быховский район	2,421	4,698	0,904	2,181	1,514	2,023	2,180	1,257	3,384
Горецкий район	2,532	2,079	0,972	1,942	1,648	1,916	1,226	1,065	3,400
Глусский район	0,767	4,129	2,491	**0,306**	0,970	0,447	3,328	1,366	1,421
Дрибинский район	0,282	4,284	2,239	0,601	0,639	0,443	3,183	1,214	1,151
Кировский район	1,829	2,747	1,130	1,244	0,968	1,208	1,806	0,429	2,696
Климовичский район	**0,062**	4,555	2,516	0,773	0,924	0,668	3,467	1,499	0,872
Кличевский район	2,015	2,971	0,479	1,733	1,131	1,605	1,463	0,691	2,843
Костюковичский район	0,519	4,326	2,061	0,953	0,621	0,756	3,050	1,187	1,259
Краснопольский район	0,580	4,559	2,224	1,179	0,869	0,990	3,226	1,420	1,143
Кричевский район	0,368	4,220	2,133	0,631	0,544	0,448	3,084	1,127	1,231
Круглянский район	0,000	4,565	2,494	0,810	0,912	0,693	3,451	1,492	0,868
Могилевский район	4,565	0,000	2,695	3,878	3,713	3,904	**1,940**	3,139	5,426
Мстиславльский район	2,494	2,695	0,000	2,187	1,609	2,071	1,014	1,128	3,317
Осиповичский район	0,810	3,878	2,187	0,000	0,705	**0,208**	3,024	1,061	1,591
Славгородский район	0,912	3,713	1,609	0,705	0,000	**0,523**	2,543	0,588	1,772
Хотимский район	0,693	3,904	2,071	**0,208**	0,523	0,000	2,947	0,960	1,527
Шкловский район	3,451	1,940	**1,014**	3,024	2,543	2,947	0,000	1,991	4,296
Чаусский район	1,492	3,139	1,128	1,061	0,588	0,960	1,991	0,000	2,358
Чериковский район	**0,868**	5,426	3,317	1,591	1,772	1,527	4,296	2,358	0,000

Источник: результаты расчетов.

Таблица 12. **Итерации расчета минимальных расстояний между объектами кластеров**

Районы	Итерации							
	1	2	3	4	5	6	7	8
Белыничский район	0,387	0,387	0,387	0,387	0,387	0,387	0,387	
Бобруйский район	0,189	0,189	0,189					
Быховский район	0,665	0,665	0,665	0,665	0,665	0,665	0,665	0,665
Горецкий район	0,522	0,522	0,522	0,522	0,522	0,522	0,522	0,522
Глусский район	0,306	0,306	0,306	0,306	0,306			
Дрибинский район	0,115	0,115						
Кировский район	0,189	0,189	0,189					
Климовичский район	0,062							
Кличевский район	0,387	0,387	0,387	0,387	0,387	0,387	0,387	
Костюковичский район	0,336	0,336	0,336	0,336	0,336	0,336		
Краснопольский район	0,547	0,547	0,547	0,547	0,547	0,547	0,547	0,547
Кричевский район	0,115	0,115						
Круглянский район	0,062							
Могилевский район	1,940	1,940	1,940	1,940	1,940	1,940	1,940	1,940
Мстиславльский район	0,479	0,479	0,479	0,479	0,479	0,479	0,479	0,479
Осиповичский район	0,208	0,208	0,208	0,208				
Славгородский район	0,523	0,523	0,523	0,523	0,523	0,523	0,523	0,523
Хотимский район	0,208	0,208	0,208	0,208				
Шкловский район	1,014	1,014	1,014	1,014	1,014	1,014	1,014	1,014
Чаусский район	0,429	0,429	0,429	0,429	0,429	0,429	0,429	0,429
Чериковский район	0,868	0,387	0,387	0,387	0,387	0,387	0,387	
Минимальное расстояние	**0,062**	**0,115**	**0,189**	**0,208**	**0,306**	**0,336**	**0,387**	**0,429**

Источник: результаты расчетов.

Таблица 13. **Итерации расчета минимальных расстояний между объектами кластеров - продолжение**

Районы	Итерации						
	9	10	11	12	13	14	15
Белыничский район							
Бобруйский район							
Быховский район	0,665	0,665	0,665	0,665			
Горецкий район	0,522	0,522					
Глусский район							
Дрибинский район							
Кировский район							
Климовичский район							
Кличевский район							
Костюковичский район							
Краснопольский район	0,547	0,547	0,547	0,547			
Кричевский район							
Круглянский район							
Могилевский район	1,940	1,940	1,940	1,940	1,940	1,940	1,940
Мстиславльский район	0,479						
Осиповичский район							
Славгородский район	0,523	0,523	0,523				
Хотимский район							
Шкловский район	1,014	1,014	1,014	1,014	1,014	1,014	
Чаусский район							
Чериковский район							
Минимальное расстояние	**0,479**	**0,522**	**0,523**	**0,547**	**0,868**	**1,014**	**1,940**

Источник: результаты расчетов.

Таблица 14. **Состав дендрограмм кластеров**

Состав 1-й дендрограммы		Состав 2-й дендрограммы	
Номер кластера	Объекты	Номер кластера	Объекты
1-й кластер	Круглянский+Климовичский	1-й кластер	Шкловский+Мстиславский+ +Могилевский+Белыничский+ +Кличевский
2-й кластер	Кричевский+Дрибинский	2-й кластер	+Кричевский+Дрибинский+ +Краснопольский+ +Костюковичский
3-й кластер	Кировский+Бобруйский	3-й кластер	Кировский+Бобруйский+ +Чаусский+Горецкий
4-й кластер	Хотимский+Славгородский	4-й кластер	Осиповичский+Глусский
5-й кластер	Осиповичский+Глусский	5-й кластер	Хотимский+Славгородский
6-й кластер	Кричевский+Костюковичский	6-й кластер	Круглянский+Чериковский
7-й кластер	Белыничский+Кличевский		
8-й кластер	Кировский+Чаусский		
9-й кластер	Кличевский+Мстиславский		
10-й кластер	Бобруйский+Горецкий		
11-й кластер	Кричевский+Краснопольский		
12-й кластер	Круглянский+Чериковский		
13-й кластер	Шкловский+Мстиславльский		
14-й кластер	Шкловский+Могилевский		

Источник: результаты расчетов.

7. АССОЦИАТИВНЫЕ ПРАВИЛА

Методические указания: В последнее время неуклонно растет интерес к методам "обнаружения знаний в базах данных" (knowledge discovery in databases). Объемы современных баз данных, которые весьма внушительны, вызвали устойчивый спрос на новые масштабируемые алгоритмы анализа данных. Одним из популярных методов обнаружения знаний стали алгоритмы поиска ассоциативных правил.

Ассоциативные правила позволяют находить закономерности между связанными событиями. Примером такого правила, служит утверждение, что покупатель, приобретающий "Хлеб", приобретет и "Молоко" с вероятностью 75%. Первый алгоритм поиска ассоциативных правил, называвшийся AIS был разработан в 1993 году сотрудниками исследовательского центра IBM Almaden. С этой пионерской работы возрос интерес к ассоциативным правилам; на середину 90-х годов прошлого века пришелся пик исследовательских работ в этой области, и с тех пор каждый год появлялось по несколько алгоритмов.

Практическая ситуация: Имеется некая организация агросервиса. Количество позиций (товаров (минеральные удобрения, запасные части) и услуг (ремонт сельскохозяйственной техники, известкование почв, мелиоративные работы), которыми она торгует) достигает 10000. Одной из главных задач этой организации является работа с покупателями ее товаров и услуг: сельскохозяйственных организаций, крестьянских (фермерских) хозяйств.

Опишем типичный процесс закупки товаров и услуг у агросервисной организации. Приходит покупатель с определенной суммой, допустим 10 т. р. Ему на руки выдается прайс-лист, в котором перечислены имеющиеся товары и их цена. Некоторые товары он (покупатель) точно знает, что будет брать (основываясь на предварительных собственных исследованиях, либо на личном опыте и интуиции). Далее специалист агросервисной организации вводит те позиции, которые выбрал покупатель, в расходную накладную. Заметим, что если покупатель захотел купить товар А в количестве, допустим, 10 единиц, а на складе осталось 8, то у покупателя остаются деньги, которые он планировал вложить в этот товар. В этом случае он либо вложит оставшиеся деньги в другой товар, либо не потратит их в этот раз и в этой организации, что явно ей не выгодно.

Описанная выше производственная ситуация – это явная ситуация упущенной выгоды. Задача специалиста агросервисной организации предугадать желание покупателя (сельскохозяйственной организации) и даже подсказать, если покупатель еще до конца не определился в отношении покупки "дополнительных" (не планируемых заранее), либо заменяющих отсутствующие на складе товары, если у клиента нет никаких соображений по этому поводу. Поставленную задачу в полном

объеме не в состоянии решить даже очень опытный и талантливый специалист, так как товаров очень много и некоторые закономерности совсем не очевидны. Помочь специалисту может такая система, которая будет подсказывать ему список тех товаров, которые склонен купить клиент к уже приобретенным в этот раз (в текущей транзакции).

Задание: Пользуясь алгоритмами поиска ассоциативных правил, изложенных в лекционном курсе, и методикой построения дерева (концептуального графа) темы 5, на примере распространенной «корзины» покупки материально-технических ресурсов сельскохозяйственными организациями построить граф ассоциативных правил для ряда сельскохозяйственных организаций, которые были объектами практики.

Порядок выполнения задания:

1. Производится сканирования входного набора, и все элементы каждой транзакции сортируются в порядке убывания поддержки этих элементов во всем базовом наборе.

2. Фильтрация. Производится удаление тех элементов a, для которых поддержка этого элемента меньше минимальной.

3. Построение префиксного FP-tree из оставшихся элементов.

4. Извлечение частных предметных наборов. Узлом FP-tree является структура, которая хранит значение узла, ссылки на все дочерние элементы и его значение поддержки для текущего узла.

Построение префиксного дерева происходит в несколько этапов: *Этап 1.* Построение корневого узла. *Этап 2.* Для каждого элемента каждой отсортированной транзакции из входного набора строятся узлы по следующему правилу: если для очередного элемента в текущем узле есть потомок, содержащий этот элемент, то новый узел не создается, а поддержка этого потомка увеличивается на 1, в противном случае создается новый узел-потомок с поддержкой 1. Текущим узлом при этом становится найденный или построенный узел.

Пример: для входного набора, состоящего из (a - удобрения, b - нефтепродукты, c – запасные части, d - трактора, e - комбайны), (a, c, d), (c b, a), (a, b) при минимальной поддержке T = 2 (см. a, b) построенное дерево принимает вид (рис. 9):

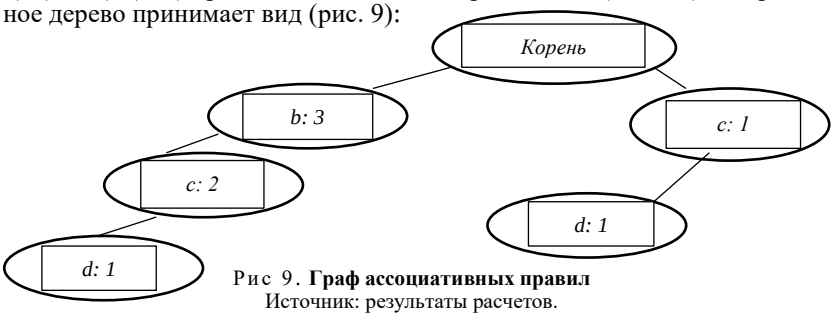


Рис 9. **Граф ассоциативных правил**
Источник: результаты расчетов.

8. ГЕНЕТИЧЕСКИЕ МОДЕЛИ

Методические указания: Данная лабораторная работа базируется на результатах расчетов темы 4, выполнение которой сопровождается созданием базы знаний нейросетевых моделей. База знаний о нейронных сетях предназначена для хранения информации об архитектурах нейронных сетях и методах их обучения, ограничениях и решаемых классах задач. Задача базы знаний по генетическим алгоритмам – накапливать информацию о том, какие методы и правила приводят к лучшему результату и применять их далее.

Управляемыми параметрами в генетических алгоритмах являются: длина хромосомы; наполнение хромосомы (локусы и аллели); параметры оператора кроссовера; параметры оператора мутации; параметры оператора инверсии; параметры выбора лучших особей; критерий остановки генерации особей и популяции; параметры генерации начальной и последующих популяций и т.д.

Введем понятие прототипа особи. Под *прототипом* будем понимать хромосому, варьируемую как по количеству генов, так и по аллели генов. При этом указывается диапазон, в котором возможны вариации значений аллелей. Так же прототип указывает, какие гены могут иметь дочерние хромосомы. Введем также понятие *дочерней и базовой хромосомы.* Базовая хромосома - это хромосома, несущая основную (базовую) информацию об объекте описания. Дочерняя - это составная часть базовой особи, ее ответвление. Дочерних хромосом может быть несколько. Дочерние хромосомы вводятся в связи с тем, что в ряде задач решаемых с помощью генетических алгоритмов, в обычной хромосоме трудно и зачастую невозможно ряд параметров переменной длины разместить в одной хромосоме. Обычно поступают следующим образом – под параметры переменной длины либо резервируют часть генов, либо такие параметры размещают в конце хромосомы. Но допустим, как поступить с варьируемым количеством слоев в нейронной сети и нейронов в каждом слое? То есть в случае функции вида f (g (x)). Поэтому в данной работе мы предлагаем делать ответвления от базовой хромосомы. Но суть и принцип работы генетического алгоритма остается прежним, за небольшими исключениями в обработке генов.

Задание: Пользуясь сформированной базой знаний нейронных сетей (тема 4) построить генетическую модель выбора наилучшей нейросети.

Порядок выполнения задания: Для решения задачи поиска нейросетевой модели с помощью разработанного генетического алгоритма будем придерживаться следующей последовательности действий. Первым шагом необходимо проанализировать входные примеры поступившей задачи, то есть определить обучающую выборку.

Здесь часть работы берет на себя постановщик задачи, который минимум должен указать, что будет выходами в случае обучения с учителем. В случае обучения без учителя определение количества выходных классов берет на себя предлагаемая топология. Далее необходимо оценить сложность задачи с точки зрения решения ее с помощью нейронных сетей, то есть оценить примерное количество скрытых слоев и нейронов в них, а также выбрать архитектуру нейронной сети и метод ее обучения (или набор архитектур и методов обучения). После предварительного оценивания сложности задачи работает конструктор нейросетей (тема 4), который и формирует один или несколько прототипов нейронных сетей. И последним шагом, самым длительным по временным затратам, является собственно сам генетический поиск адекватной нейросетевой модели.

Общий алгоритм работы топологии генетического поиска нейросетевой модели следующий:

1. На вход топологии поступает задача, с полным или частичным указанием входов, выходов, целевой функции. В общем случае поступают временные ряды входов и выходов (тема 1);

2. Временные ряды и заданные параметры анализируются блоком оценки сложности задачи и блоком анализа временных рядов;

3. На основе данных о временных рядах задачи, а также сложности задачи, блок генерации прототипов, формирует прототипы особей нейронных сетей;

4. На основе информации о сложности задачи база знаний генетических алгоритмов подает данные на стохастический генератор (см. функцию генерации: темы 1, 2) по прототипу по количеству банков особей и особей в них;

5. Стохастический генератор по прототипу по заложенному алгоритму формирует заданное количество банков особей начальной популяции 1;

6. После окончания формирования начальной популяции управление передается блоку управления популяциями, который находясь в рамках заданных ограничений используя блок получения новых особей и блок оценки особей, формирует новые особи, новые банки особей и новые популяции (см. подробнее п.2.). Новые особи, банки особей и популяции формируются либо до предельных значений ограничений, выставленных пользователем, либо по достижению целевого критерия задачи.

Если смотреть обобщено на решение задач нейронными сетями (тема 4), то можно разделить задачи на две категории. Первая – задача может быть решена только с помощью одной топологии и метода обучения, вторая – задача имеет несколько вариантов решений на различных нейронных сетях.

9. НЕЧЕТКАЯ ЛОГИКА

Методические указания: Используемая в различных видах систем модель на основе нечеткой логики представляет собой базу знаний, построенную специалистами предметной области как множество нечетких правил. Нечеткий логический вывод формируется в несколько шагов:

введение нечеткости: на этом этапе функции принадлежности применяются к фактическим значениям входных переменных;

логический вывод: вычисляется значение истинности для предпосылок каждого правила и применяется к заключениям каждого правила. Это приводит к одному нечеткому подмножеству, которое будет назначено каждой переменной вывода для каждого правила;

композиция: нечеткие подмножества, назначенные каждой переменной вывода, объединяют в одно множество для всех переменных вывода;

приведение к четкости: используется в случаях, когда необходимо преобразовать нечеткий набор выводов в четкое число.

Задание: Пользуясь навыками, приобретенными при выполнении лабораторной работы № 4, необходимо построить нейронную сеть на базе нечеткой логики. На этих принципах построено большое количество сетей, рассмотрим подробнее одну из них - сеть Ванга - Менделя. Структура такой сети представляет собой четырехслойную нейронную сеть, в которой первый слой выполняет фазификацию входных переменных, второй - агрегирование значений активации условия, третий - агрегирование М правил вывода (первый нейрон) и генерацию нормализующего сигнала (второй нейрон), тогда как состоящий из одного нейрона выходной слой осуществляет нормализацию, формируя выходной сигнал.

Порядок выполнения задания:

1. Обучение нечетких сетей, также как и классических сетей, может проводиться по алгоритму с учителем, основанному на минимизации целевой функции, задаваемой с использованием евклидовой нормы. Для обучения нечеткой нейронной сети применяют алгоритм, включающий последовательное чередование следующих шагов:

- для фиксированных значений параметров c_{ij} и d_{ij} первого слоя вычисляются значения параметров w_i третьего слоя сети (рис.10);
- при зафиксированных значениях параметров w_i третьего слоя уточняются параметры c_{ij} и d_{ij} первого слоя сети.

2. Таким образом, на первом этапе для K обучающих выборок , k=1, 2, K, получаем систему K линейных уравнений.

3. На втором этапе фиксируются значения коэффициентов полиномов третьего слоя и осуществляется уточнение (обычно многократное)

коэффициентов функции Гаусса для первого слоя сети стандартным методом градиента.

4. Поскольку в череде этапов этап уточнения параметров функции Гаусса имеет много меньшую скорость сходимости, то в ходе обучения реализацию этапа 1, как правило, сопровождает реализация нескольких этапов 2.

5. Часто требуется найти «решение» системы, которая решений (в обычном смысле) не имеет. Выходом из ситуации является нахождение таких значений неизвестных параметров, что все условия системы выполняются «в некоторой степени» (нечеткая логика).

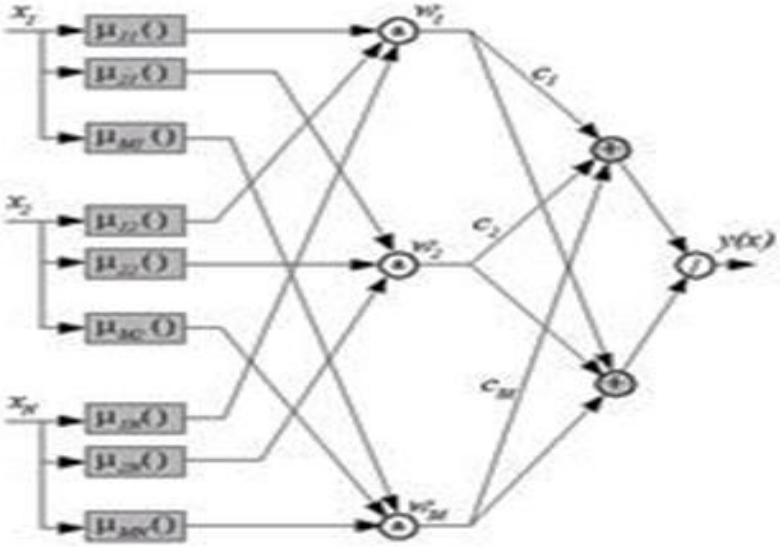


Рис 10. **Структура нечеткой нейронной сети**
Источник: результаты расчетов.

Как следует из рисунка 10, μ_{ij} - функция Гаусса с параметрами математического ожидания, которое определяет центр c_{ij} и параметрами разброса, которые определяются средним квадратическим отклонением d_{ij}. Псевдорешением (нечетким решением) системы линейных уравнений в результате нейросетевого моделирования называется решение системы с минимальной нормой среди всех столбцов, имеющих минимальную невязку (норма вектора равна квадратному корню из суммы квадратов компонент вектора, а невязкой решения системы называется норма вектора Ax-b.

10. ДОКУМЕНТАЛЬНЫЕ ИНФОРМАЦИОННО-ПОИСКОВЫЕ СИСТЕМЫ

Методические указания: Информационно-поисковая система — совокупность средств для хранения, поиска и выдачи по запросу информации. Поиск и размещение информации в информационно-поисковой системе осуществляется вручную или с помощью компьютера в соответствии с принятым информационным языком по определенным правилам (алгоритму). В число составных частей информационно-поисковой системы, кроме информационно-поискового языка, правил перевода и критерия соответствия, входят средства ее технической реализации, массив текстов (документов), в котором осуществляется информационный поиск, люди, непосредственно участвующие в поиске. Различают два вида информационно-поисковых систем — документальные и фактографические. К документальным информационно-поисковым системами относятся указатель в книге, библиотечный каталог или книгохранилище в библиотеке, к фактографическим — телефонный справочник, адресная книга, каталог изделий. Фактографическая информационно-поисковая система, в отличие от информационно-логической системы, не обеспечивает получения новой информации из имеющейся в ней, а только помогает отыскивать факты или сведения, которые были в нее введены.

Access предоставляет довольно широкий спектр возможностей для поиска и отбора информации в базе данных. К таким средствам можно отнести использование команды **Поиск**, **Фильтрация**, **Сортировка**, создание и использование запросов. Простейшим способом поиска информации в базе данных является использование директивы **Поиск**. Этот поиск может проводиться как в одном из указанных полей, так и во всех полях таблицы БД. Возможно изменение порядка просмотра записей в таблице. Обычно поиск по этой директиве начинается с активного места таблицы (активной записи, активного поля). Для просмотра всей таблицы необходимо перейти к первой записи, а затем начать поиск. Для того чтобы записи в таблице выстраивались при выводе в удобном для пользователя порядке, используется сортировка. Access может проводить сортировку по одному полю, по нескольким полям, по возрастанию или по убыванию значений ключевого признака. Для вывода только определенных записей таблицы (отбора) используется фильтрация. В Access поиск и отбор любой нужной информации можно производить с использованием запросов, имеющих большие возможности, чем рассмотренные ранее средства. Запросы используются примерно так же, как таблицы. Запрос представляет собой вопрос о данных, хранящихся в таблицах, или инструкцию на отбор записей, подлежащих изменению. С помощью Access могут быть созданы следующие типы запросов:

Запрос-выборка задает вопросы о данных, хранящихся в таблицах, и представляет полученный динамический набор в режиме формы или таблицы без изменения данных. Изменения, внесенные в динамический набор, отражаются в базовых таблицах.

Запрос-изменение изменяет или перемещает данные. К этому типу относятся: запрос на добавление записей, запрос на удаление записей, запрос на создание таблицы, запрос на обновление.

Перекрестные запросы предназначены для группирования данных и представления их в компактном виде.

Запрос с параметром позволяет определить одно или несколько условий отбора во время выполнения запроса.

Запросы SQL — запросы, которые могут быть созданы только с помощью инструкций SQL в режиме SQL: запрос — объединение, запрос к серверу и управляющий запрос.

В Access имеется возможность самостоятельно создать запрос или воспользоваться мастером по разработке запросов.

Задание: На основе электронных материалов предыдущих лабораторных работ создать в Access аналог информационно-поисковой системы базы знаний по курсу «Технологии интеллектуального анализа данных».

Порядок выполнения задания:

1. По указанию преподавателя обучающиеся загружают программу Microsoft Access. Преподаватель руководит работой обучающихся, давая команды по работе с базой данных.

2. После того как загрузилась программа, компьютер предлагает ***Создать базу данных***, ***Открыть уже ранее созданные базы данных***. Мы выбираем ***Создать новую***. Далее компьютер предлагает дать название вашей базе данных (назовем ***своей фамилией и именем***), после этого клавиша ***Создать***. В диалоговом окне предлагаются варианты построения: ***Построение таблицы в режиме конструктора***, ***Построение таблицы с помощью мастера***, ***Построение таблицы путем ввода данных*** (о каждом варианте преподаватель сообщает его суть).

3. Выбираем ***Построение таблицы с помощью конструктора***. В диалоговом окне ***Новая таблица*** выбираем подменю ***Конструктор***. Появляется окно, в котором заносятся признаки наших объектов и типы данных, то есть мы строим макет таблицы.

4. Далее сохраняем эту таблицу под названием ***Классы*** и закрываем ее нажатием на крестик в верхнем правом углу окна. В диалоговом окне базы данных появится название вашей таблицы, двойным щелчком открываем эту таблицу. Предлагается макет таблицы и здесь можно вносить значения определенных нами признаков предложенного набора объектов.

5. Повторяем операции для формирования базы знаний.

11. СИСТЕМЫ, ОСНОВАННЫЕ НА ЗНАНИЯХ

Методические указания: На начальном этапе исследований по искусственному интеллекту возникло всеобщее убеждение, что за интеллектуальным поведением человека скрываются его знания об окружающем мире. Речь идет о знаниях, которыми обладают специалисты профессионалы. Интеллектуальная система, основанная на знаниях, представляет собой такую систему, в которой с помощью логического вывода знания применяются к решению поставленных задач. Экспертные системы отличаются от традиционных вычислительных пакетов программ тем, как они организованы. С традиционных позиций программы представляет собой процедуру и данные.

Если компьютерной программе предстоит выполнить задачу эксперта, то она нуждается в большом объеме знаний, позволяющих решать сложные проблемы, подобно тому, как это делает человек. Система должна быть тщательно организована. В общем случае знания разделяются на три типа:

а) Фактические (декларативные) знания. Этот вид знаний представляет собой информацию о конкретном случае, обычно собираемую посредством диалога с пользователем, который указывает какие факты следует считать справедливыми в настоящее время. Важно то, как представлена эта информация, поскольку сама структура представления также информативна. Структуру представления следует выбирать в зависимости от содержания знаний.

б) Процедурные знания. Эти знания обычно собираются заранее путём опроса специалиста в данной предметной области и составляют ядро базы знаний. Они используются также в блоке рассуждения системы, позволяя потом выводить следствия. Процедурные знания дают возможность при необходимости генерировать и факты. Таким образом, фактические и процедурные знания тесно переплетаются. Кроме того, в ходе работы системы приходится принимать решения, какие из тех правил следует использовать.

в) Управляющие знания. Системе должен быть предусмотрен некоторый набор стратегий, чтобы можно было рассматривать альтернативные возможности в ходе работы, переход при неудаче от одной стратегии к другой.

Системы продукций представляют собой конкретный метод организации программ в виде совокупностей трех групп:

1. База знаний (БЗ).

2. Список порождающих правил (ПП).

3. Метод выбора, какое порождающее правило следует применить при данном состоянии базы знаний.

Каждое порождающее правило (продукция) имеет форму ЕСЛИ (условие), ТО (действие) или, возможно, форму ЕСЛИ (условие), ТО

(*действие 1*), В ПРОТИВНОМ СЛУЧАЕ (*действие 2*). При этом процедуры сопоставления с образцом определяет, является ли данное правило применимым вообще.

Задание: Пользуясь созданной базой знаний (тема 10) по курсу «Технологии интеллектуального анализа данных» создать экспертную систему самостоятельного обучения дисциплине в рамках стандартных вопросов и автоматического отбора информации для формирования ответов на эти вопросы.

Порядок выполнения задания:

1. Используя логические функции EXCEL сформируем ряд порождающих правил (продукций), вытекающих из стандартных контрольных вопросов, предложенных преподавателем.

Сформируем таблицу решений (см. табл. 15):

Таблица 15 **Фрагмент базы знаний учебного курса «Технологии интеллектуального анализа данных»**

Контрольные вопросы на ключевое слово	номера тем в базе данных Access								
	1	2	3	4	5	6	7	8	9
процесс интеллектуального анализа	1	2							
статистические методы			3			3; 6			
нейросети				4					4; 9
дерево решений					5				
кластеризация					5;6				
ассоциативные правила							5;7		
генетические алгоритмы								6;8	
нечеткая логика									9
интеллектуальные системы	1	2	3	4	5	6	7	8	9

Источник: результаты логического анализа структуры курса.

2. После получения порции знаний и фиксации их в виде таблицы решений, соответствующая часть понятий предметной области и их возможные значения заносятся в базу понятий экспертной системы, причем понятия, которые можно задавать как факты, обязательно размещают в ее верхней части. Некоторые из этих понятий могут не только задаваться, но и определяться при помощи соответствующих правил.

3. Базу знаний необходимо организовать в виде совокупности таблиц, каждая из которых имеет много общего с расширенной таблицей решений и является минимальной структурой информации для представления классов объектов, явлений или процессов. В базу знаний необходимо включить «рабочую память», что упрощает реализацию механизма объяснений. В каждой клетке правил, кроме ссылки на соответствующую клетку *Возможные значения* базы понятий, может находиться формула, которая включает функции. В Excel имеется возможность создавать и использовать в рабочих листах функции, которые реализуют алгоритмы любой сложности.

12. СТРУКТУРНЫЙ АНАЛИЗ СИСТЕМ

Методические указания: Структурный анализ – это анализ, который проводится для того, чтобы исследовать статические характеристики систем посредством определения в ней подсистем (элементы различного уровня), а также определение отношений и связи между ними. Объекты исследования при осуществлении структурного анализа – какие-либо варианты структур системы, которые формируются в процессе декомпозиции. Основными показателями исследуемых структур являются: совокупность выделенных связей, отношений и элементов; их характеристики, обобщающие показатели структур, которые характеризуют их воздействие на всю систему управления,

Структура системного анализа представляет собой общую процедуру, которая состоит из нескольких стадий:

1. Производится декомпозиция системы на подсистемы и элементы, которые интересуют исследователей, формируются структуры и происходит их описание.

2. Определяются характеристики (качественные и количественные) у выделенных структур, оценка структур.

3. Формируются критерии и оценки степени эффективности структур.

4. Принимается решение о совершенствовании структурных характеристик.

Задание: На примере разработанной экспертной системы обучения (тема 11) провести ее структурный анализ.

Порядок выполнения задания:

1. Составить схему структуры экспертной системы (тема 11).

2. Создать неформальное описание знаний в виде графа (карты мысли). При создании неформальной модели определяются:

- основные понятия предметной области и их характеристика;
- терминология;
- отношения между понятиями;
- структура входной и выходной информации.

3. Применить структурный подход, который в основном базируется на:

- SADT (Structured Analysis and Design Technique – методология (технология) структурного анализа и проектирования) - диаграммы, описывающие функциональные возможности системы (функциональные модели);
- DFD (Data Flow Diagrams) - диаграммы потоков данных (структурные модели);
- ERD (Entity-Relationship Diagrams) - диаграммы "сущность-связь" (информационные модели);
- блок-схемы (схемы информационных процессов) - схемы алго-

ритмов, программ, данных и систем (поведенческие модели).

4. Провести реструктуризацию имеющейся экспертной системы обучения предмету. Одним из вариантов является переход от двумерной системы (темы и информационный материал) к трехмерной. Декомпозиция, которой позволяет сформировать курс в виртуальном пространстве.

Центральная часть системы – конструктор учебных курсов – позволяет преподавателям собирать свои виртуальные миры из логических блоков, в роли которых выступают шаблоны комнат. Структура курса древовидна: корнем служат главный холл и комната выбора курсов, каждый курс распадается на темы, каждая тема представляет собой коридор, содержащий комнаты с контентом. Для построения виртуального мира системы может применяться игровой движок, в котором виртуальный мир представляется в виде сцен. В настоящее время коридоры курсов и тем генерируются в одной сцене вместе с лекционными, тестовыми и прочими комнатами, что при большом размере курса вызывает замедление работы системы. Данные, получаемые клиентской частью от сервера при генерации курса (строка особого формата), кроме структуры последнего содержат всю содержательную информацию для комнат (тексты лекций и тестов), что также может вызвать неэффективное использование ресурсов. По вышеперечисленным причинам стала актуальной задача реструктуризации организации курсов в трехмерном пространстве. При выборе курса в сцене трехмерного пространства будут генерироваться только коридоры курсов и тем с телепортаторами для перехода в содержательные комнаты. Лекционные комнаты и комнаты тестирования будут представлены в виде отдельных сцен, они также будут генерироваться по шаблону. Это позволит при генерации курса запрашивать с сервера строку, содержащую только структуру курса, а содержимое курса получать при входе в отдельные комнаты. Таким образом можно повысить производительность всей системы в целом.

ЗАКЛЮЧЕНИЕ

Обобщим некоторые ключевые методы интеллектуального анализа данных, рассматриваемые в этом практикуме. *Интеллектуальный анализ данных* (ИАД) или data mining - это процесс выявления значимых корреляций, образцов и тенденций в больших объемах данных. *Ассоциация* (или отношение) является наиболее известным методом интеллектуального анализа данных. Данный метод заключается в сопоставление двух или более элементов чаще всего одного и того же типа. Метод *классификации* используют при описании нескольких атрибутов для идентификации определенного класса, а также в качестве входных данных для других методов. Каждый класс обла-

дает определенными свойствами, которые характеризуют его объекты. Для определения классификации применяют деревья принятия решений.

Дерево решений, связанное с большинством других методов, используют в рамках критериев отбора ла так же для поддержки выбора определенных данных в рамках общей структуры. Дерево решений начинают с простого вопроса, который имеет два ответа (но возможно и больше). Каждый ответ приводит к следующему вопросу помогая классифицировать и идентифицировать данные или делать прогнозы. *Кластеризация* позволяет использовать общие атрибуты различных классификаций в целях выявления кластеров. Исследуя один или более атрибутов можно сгруппировать отдельные элементы данных, вместе получая структурированное заключение. На простом уровне при кластеризации используется один или несколько атрибутов в качестве основы для определения кластера сходных результатов. Кластеризация полезна при определении различной информации, потому что она коррелируется с другими примерами, так что можно увидеть где подобия и диапазоны согласуются между собой.

Нейросетевые технологии предоставляют сегодня широкие возможности для решения задач прогнозирования, обработки сигналов и распознавания образов. По сравнению с традиционными методами математической статистики, классификации и аппроксимации, эти технологии обеспечивают достаточно высокое качество решений при меньших затратах. Интеллектуальный анализ данных опирается на построение подходящей модели и структуры которые можно использовать для обработки выявления и создания необходимой информации. Независимо от формы и структуры источника данных информация структурируется и организуется в соответствии с форматом, который позволяет выполнять интеллектуальный анализ данных с максимально эффективной моделью.

Таким образом, наличие практикума по интеллектуальному анализу данных поможет развитию научно-исследовательских компетенций у обучающихся магистрантов. Для приобретения навыков в области анализа данных для не ИТ-специалистов в предлагаемых лабораторных работах рекомендовано сосредоточиться на демонстрации возможностей интеллектуального анализа данных с помощью существующих инструментов (в различных практиках использовались надстройки Excel, Matlab и др.). А учитывая сложность материала для непрофильных специалистов, знакомство с интеллектуальным анализом данных необходимо осуществлять сразу после изучения обязательного курса математической статистики, причем, для повышения ценностно-мотивационного компонента в лабораторных работах решено демонстрировать и выполнять анализ на наборах данных из профессиональной области: агропромышленного производства.

Вопросы и темы для самопроверки

1. Понятие Интеллектуального анализа данных (Data Mining).
2. Data Mining как часть рынка интеллектуальных технологий.
3. Набор данных и их атрибутов. Измерения. Типы наборов данных.
4. Форматы хранения данных. Метаданные.
5. Задача классификации. Процесс классификации.
6. Методы, применяемые для решения задач классификации.
7. Точность классификации: оценка уровня ошибок.
8. Оценивание классификационных методов.
9. Деревья решений.
10. Процесс конструирования дерева решений.
11. Метод опорных векторов.
12. Метод «ближайшего соседа».
13. Байесова классификация.
14. Задача прогнозирования.
15. Сравнение задач прогнозирования и классификации.
16. Прогнозирование и временные ряды.
17. Решение задачи прогнозирования.
18. Задача кластеризации.
19. Применение кластерного анализа.
20. Иерархичные методы.
21. Итеративные методы.
22. Методы поиска ассоциативных правил.
23. Методы визуализации.
24. Качество визуализации.
25. Представление пространственных характеристик.
26. Основные тенденции в визуализации.
27. Средства извлечения данных: методы и возможности.
28. Начальные этапы: анализ предметной области; постановка задачи, подготовка данных.
29. Очистка данных. Инструменты очистки данных.
30. Построение и использование модели.
31. Стандарты Data Mining.
32. Рынок инструментов Data Mining.
33. Классификация инструментов Data Mining.
34. Программное обеспечение для решения задач классификации.
35. Программное обеспечения для решения задач кластеризации и сегментации. Программное обеспечение Data Mining для поиска ассоциативных правил. Программное обеспечение для решения задач оценивания и прогнозирования.
36. Системы бизнес-интеллекта и управления знаниями.

37. Сферы применения Data Mining.
38. Применение Data Mining для бизнес-задач.
39. Data Mining для научных исследований.
40. Data Mining консалтинг.
41. Data Mining услуги. Примеры решения.
42. Техническое описание решения.
43. Технологии лингвистического анализа бизнес-информации.
44. Интеллектуальный поиск в интернете.
45. Аналитическая обработка бизнес-информации.
46. Комплексный подход к внедрению Data Mining, OLAP и хранилищ данных. Интеграция OLAP и Data Mining.
47. Хранилища данных. Преимущества хранилища данных.

ТЕСТ

1. Наибольшая степень актуальности от информационной системы требуется при решении задачи:

а) информационного поиска и выполнения заранее определённых запросов к базе данных;

б) поиска функциональных и логических закономерностей в накопленных данных;

в) оперативно-аналитического анализа данных;

г) ввода, обновления и хранения данных.

2. Основное назначение OLTP-системы (On-Line Transaction Processing):

а) автоматизация интеллектуального анализа данных;

б) долговременное хранение данных;

в) операционная (транзакционная) обработка данных;

г) поддержка реляционных хранилищ данных;

3. Основное назначение OLAP-системы (On-Line Analytical processing):

а) выполнение интеллектуального анализа данных;

б) поддержка аналитической деятельности на предприятии;

в) предварительная обработка данных перед анализом;

г) обеспечение безопасности хранения данных.

4. Основное назначение систем интеллектуального анализа (Data Mining):

а) обнаружение в сырых данных скрытых знаний;

б) проведение статистического анализа;

в) решения задач математического программирования;

г) поиск агрегированных данных;

5. При проведении интеллектуального анализа из существующих данных извлекают:

а) шаблоны и тренды;

б) функциональные зависимости;
в) свойства фактов;
г) атрибуты измерений.

6. К компонентам СППР не относится:

а) информационные хранилища данных;
б) базы данных;
в) средства и методы извлечения, обработки и загрузки данных (ETL);
г) многомерная база данных и средства анализа OLAP;
д) средства Data Mining.

7. Правильная последовательность в Business Intelligence:

а) данные-информация-знания-принятие решения
б) информация-данные-знания-принятие решения
в) данные-знания-информация-принятие решения
г) принятие решения-информация-данные-знания

8. В платформе для бизнес-анализа должны быть реализованы:

а) 10 ключевых возможностей
б) 12 ключевых возможностей
в) 15 ключевых возможностей
г) 20 ключевых возможностей

9. Ключевые возможности систем BI сгруппированы:

а) по двум основным категориям
б) по трем основным категориям
в) по четырем основным категориям
г) по пяти основным категориям

10. «BI-инфраструктура» относится к категории:

а) представление информации
б) анализ данных
в) возможность интеграции
г) является основной категорией

11. Перечислите правильную последовательность этапов Knowledge Discovery in Databases –процесса обнаружения знаний в базах данных:

а) трансформация, интерпретация результатов, выборка, очистка, построение моделей.
б) построение моделей, выборка, очистка, трансформация, интерпретация результатов.
в) построение моделей, выборка, очистка, трансформация, интерпретация результатов,
г) выборка, очистка, трансформация, построение моделей, интерпретация результатов.

12. OLAP-системы это:

а) информационные системы оперативной транзакционной обработки данных.

б) информационные системы оперативного анализа данных.
в) информационные системы автоматической обработки данных.
г) информационные системы алгоритмической обработки данных.

13. OLTP-системы это:

а) информационные системы оперативной транзакционной обработки данных.
б) информационные системы оперативного анализа данных.
в) информационные системы автоматической обработки данных.
г) информационные системы алгоритмической обработки данных.

14. С какой целью создаются хранилища данных:

а) для хранения одном месте любых данных.
б) для интеграции разрозненных данных.
в) для агрегации ранее разъединенных детализированных данных.
г) для интеграции в одном месте, согласования и, возможно, агрегации ранее разъединенных детализированных данных.

15. Что входит в состав хранилища данных:

а) организационная структура, технические средства, базы или совокупности баз данных и программное обеспечение.
б) базы или совокупности баз данных и программное обеспечение.
в) источники данных и программное обеспечение.
г) организационная структура и программное обеспечение

16. Какими свойствами должны обладать средства хранения данных:

а) интегрированные, неизменчивые, поддерживающие хронологию.
б) предметно-ориентированные, интегрированные, неизменчивые, поддерживающие хронологию.
в) предметно-ориентированные, неизменчивые, поддерживающие хронологию.
г) неизменчивые, поддерживающие хронологию.

17. Сколько уровней содержит архитектура хранилищ данных:

а) два уровня.
б) три уровня.

в) четыре уровня.
г) пять уровней.

18. Что является основными составляющими структуры хранилищ данных:

а) таблица исходной информации и таблица запросов.
б) таблица базы данных и запросы.
в) таблица фактов и таблица измерений.
г) таблица запросов и таблица данных.

19. На основе чего реализуется концептуальная многомерная модель данных:

а) на основе представления данных в виде многомерного пространства, размерность которого определяется количеством измерений.
б) на основе представления данных в виде многомерного пространства, размерность которого определяется количеством граней куба.
в) на основе представления данных в виде бесконечного пространства.
г) на основе представления данных в виде пространства, ограниченного многомерным кубом.

20. Размерность многомерного пространства данных для анализа математически определяется:

а) сложением размеров всех измерений в модели данных;
б) количеством атрибутов в реляционной таблице фактов;
в) количеством таблиц содержащих измерения;
г) перемножением размеров всех измерений в модели данных.

21. Размер или кардинальность измерения определяется:

а) количеством атрибутов и свойств в измерении;
б) количеством значений ключа в таблице измерения;
в) количеством элементов в измерении;
г) количеством записей в таблице измерений;

22. Роль унифицированной многомерной модели заключается:

а) в создании концептуальной модели хранилища данных;
б) в определении функциональной зависимости между данными;
в) в определении реляционных отношений между сущностями;
г) в создании моста между пользователем и источниками данных.

23. Схема реляционного хранилища данных носит название «снежинка», если:

а) хранилище данных содержит несколько таблиц с фактами;
б) одно из измерений хранилища данных содержится в нескольких связанных таблицах;
в) каждое измерение хранилища данных содержится в одной таблице;
г) каждое измерение хранилища данных содержится в нескольких

связанных таблицах.

24. Многомерная модель данных определяет представление данных на уровне:

а) концептуальной модели и прикладной модели;

б) концептуальной модели и физической модели;

в) физической модели и прикладной модели;

г) концептуальной, физической и прикладной моделей.

25. Сколько основных компонентов в MS SQL Server 2008:

а) два.

б) три.

в) четыре.

г) пять.

26. Какие редакторы поддерживает Management Studio:

а) редактор SQL Server запросов; редактор Analysis запросов (MDX, DMX, XMLA).

б) редактор XML; редактор обычного текста.

в) редактор SQL Server запросов; редактор Analysis запросов (MDX, DMX, XMLA); редактор XML; редактор обычного текста.

г) редактор SQL Server запросов; редактор Analysis запросов (MDX, DMX, XMLA); редактор XML.

27. Потоком данных в службах SSIS называют:

а) множество данных, характеризующих объект анализа;

б) перемещение данных от источника к приёмнику;

в) файл с множеством данных, подготовленный для анализа;

г) множество данных, перемещаемых в многомерную модель данных.

28. Архитектура служб SSIS ориентирована на операции:

а) с множествами кортежей, характеризующих объекты анализа;

б) с объектами интеллектуального анализа данных;

в) оперативного и интеллектуального анализа данных;

г) извлечения, преобразования и загрузку данных.

29. Одно из основных назначений языка XML в системах анализа данных:

а) описание методов и алгоритмов анализа данных;

б) описание процесса обмена данными между приложениями;

в) разработка пользовательских приложений в системе анализа;

г) описание

30. Службы SQL Server Management Studio предназначены для:

а) администрирования и управления многомерными объектами;

б) осуществления оперативного анализа данных;

в) осуществления интеллектуального анализа данных;

г) извлечения, преобразования и загрузки данных.

31. Процессом перегрузки данных в ETL-системах называют:

а) реализацию потока данных от единственного набора данных источника до одного или нескольких наборов данных хранилища;

б) создание копии таблицы с данными в базе данных;

в) создание резервной копии базы данных на сервере;

г) реализацию потока данных из хранилища до одного набора данных в транзакционной БД.

32. В задаче кластеризации отнесение объекта, характеризуемого множеством параметров, осуществляется:

а) к одному заранее определённому аналитиком классу;

б) к одному заранее определённому аналитиком контейнеру;

в) к одному заранее неопределённому классу;

г) к одному заранее определённому экземпляру сущности.

33. Параметры, характеризующие объекты кластерного анализа, могут принимать значения из множества:

а) комплексных чисел;

б) нечётких вещественных чисел;

в) вещественных чисел;

г) лингвистических оценок.

34. Мера близости объектов в кластерном анализе характеризуется:

а) весовыми коэффициентами для пересчёта расстояний;

б) количеством объектов, входящих в кластер;

в) расстоянием между объектами из заданного набора;

г) разностью значений между параметрами объекта.

35. В иерархических дивизимных алгоритмах кластеризации на первом шаге количество кластеров определяется:

а) количеством объектов из анализируемого набора;

б) параметрами, характеризующими алгоритмы кластеризации;

в) требованиями из поставленной задачи кластеризации;

г) требованиями лица принимающего решения.

36. В неиерархических алгоритмах процедура разбиения объектов на кластеры завершается при выполнении условия:

а) количество объектов в кластерах не меньше заданного значения;

б) расстояния между кластерами имеют минимальное значение;

в) количество сформированных кластеров равно заданному значению;

г) центры и границы сформированных кластеров не меняются.

37. Задача классификации решается в случае, если зависимая переменная принимает значения из:

а) конечного множества;

б) бесконечного множества;

в) счётного множества;

г) континуума.

38. В деревьях решений в качестве листа рассматривается:

а) внутренняя вершина дерева или узел проверки;

б) конечный узел дерева или узел решения;

в) отсечённые при построении дерева узлы решений;

г) узел дерева решений, не содержащий объектов.

39. В алгоритме CART (Classification and Regression Tree), для оценки качества разбиения объектов в процессе обучения используется:

а) статистический критерий;

б) отношение правильно классифицированных объектов к общему количеству объектов;

в) статистический, и теоретико-информационный критерий.

г) теоретико-информационный критерий.

40. Правило классификации может быть представлено в виде:

а) наборами параметров, определяющих принадлежность объекта к одному из классов заданного множества;

б) классификационного правила: если (условие), то (заключение);

в) аналитического выражения, определяющего функциональную зависимость между зависимой переменной и независимыми переменными;

г) математической функции, выражающей отношение зависимой переменной от независимых переменных;

41. Условие разделения объектов в узле дерева решений должно отвечать требованию:

а) формирования подмножеств из объектов одного класса или с минимальным количеством объектов из других классов;

б) формирования подмножеств с равным количеством объектов;

в) формирования подмножеств,

г) формирования подмножества

42. При решении задач поиска ассоциативных правил в качестве транзакции рассматривают:

а) свойства объектов входящих в набор;

б) множество обнаруженных зависимостей;

в) набор объектов, элементов или товаров;

г) количество объектов в наборе.

43. Значение поддержки набора при ассоциативном поиске определяют:

а) отношением количества транзакций, содержащих набор, к общему количеству транзакций;

б) отношением количества объектов в наборе к количеству объектов, встречающихся во всех транзакциях;

в) отношением количества объектов в наборе к количеству объектов, встречающихся во всех наборах;

г) отношением общего количества транзакций к количеству транзакций, содержащих набор.

44. Заданный набор объектов называют частым, если:

а) поддержка имеет значение близкое к единице;

б) поддержка не меньше среднего значения всех поддержек;

в) поддержка больше поддержки одноэлементных наборов;

г) поддержка больше заданного минимального значения.

45. Ассоциативные правила имеют следующий вид:

а) поддержка набора А больше поддержки набора В;

б) частота набора А меньше больше частоты набора В;

в) если (условие), то (результат);

г) набор объектов А содержит объекты набора В.

46. Полезность определенного ассоциативного правила оценивается:

а) отношением количества объектов, входящих в наборы правила, к общему количеству объектов;

б) отношением транзакций, поддерживающих правило, к общему количеству транзакций;

в) отношением общего количества объектов к количеству объектов, входящих в наборы правил.

г) отношением общего количества транзакций к количеству транзакций, поддерживающих правило.

47. Содержимое структуры интеллектуального анализа данных может быть определёно:

а) существующего представления источника данных или куба;

б) существующего источника и представления источника данных;

в) существующего источника данных и многомерного куба;

г) на основе модели интеллектуального анализа данных.

48. Мастер интеллектуального анализа предназначен для работы:

а) с моделями интеллектуального анализа данных;

б) с таблицами источника данных для интеллектуального анализа;

в) со структурами и моделями интеллектуального анализа;

г) со структурами интеллектуального анализа данных.

49. Конструктор интеллектуального анализа данных предназначен для работы с моделями анализа данных:

а) в службах SQL Server Data Mining;

б) в службах SQL Server Analysis Services;

в) в службах SQL Server Integration Services;

г) в службах SQL Server Management Studio.

50. После разбиения данных на обучающий и проверочный набор эти данные могут быть использованы:

а) одной моделью, содержащей описание наборов данных;

б) всеми моделями на основе одной созданной структуры;

в) множеством моделей на основе различных созданных структур;

г) моделями, определёнными обучающим набором данных.

Литература

1. Айзек, М. П. Графика, формулы, анализ данных в Excel. Пошаговые примеры / М..П. Айзек, М. В. Финков. - СПб.: Наука и техника, 2019. - 386 с.

2. Аксень, Э.М. Стохастическое моделирование макроэкономической динамики / Э.М. Аксень ; Белорус. гос. экон. ун-т. — Минск, 2011. — 326 с. — Деп. в БелИСА 25.10.2011 г., № Д201151 // Новости науки и технологий. — 2011. — № 2 (19). — С. 52.

3. Бенгфорт, Б. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка / Б. Бенгфорт. - СПб.: Питер, 2019. - 368 с.

4. Бергер, А. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных / А. Бергер. - СПб.: BHV, 2007. - 928 с.

5. Боровиков, В.П. Популярное введение в современный анализ данных в системе STATISTICA. Учебное пособие для вузов. +CD / В.П. Боровиков. — М.: РиС, 2015. — 288 с.

6. Буць, В. И. Методология построения социально-экономических индексов управления ресурсосбережением / В. И. Буць. – Горки: Белорусская государственная сельскохозяйственная академия, 2009. – 168 с.

7. Винстон, У. Бизнес-моделирование и анализ данных. Решение актуальных задач с помощью Microsoft Excel / У. Винстон. - СПб.: Питер, 2006. - 320 с.

8. Воскобойников, Ю.Е. Регрессионный анализ данных в пакете MATHCAD + CD / Ю.Е. Воскобойников. — СПб.: Лань, 2011. — 224 с.

9. Горяинова, Е.Р. Прикладные методы анализа статистических данных: Учебное пособие / Е.Р. Горяинова, А.Р. Панков, Е.Н. Платонов. — М.: ИД ГУ ВШЭ, 2012. — 310 с.

10. Дайитбегов, Д.М. Компьютерные технологии анализа данных в эконометрике: Монография / Д.М. Дайитбегов. — М.: Вузовский учебник, НИЦ ИНФРА-М, 2013. — 587 с.

11. Есаулов, И.Г. Регрессионный анализ данных в пакете Mathcad: Учебное пособие / И.Г. Есаулов. - СПб.: Лань П, 2016. - 224 с.

12. Ефремов, А. А. Использование оболочки данных для оценки сравнительной эффективности функционирования сельскохозяйственных организаций / А. А. Ефремов // Вестн. Могилев. гос. ун-та им. А. А. Кулешова. Сер. В. Математика. Физика. Биология. — 2016. — № 49 (1). — С. 189–191.

13. Железко, Б. А. Теория и практика построения информационно-аналитических систем поддержки принятия решений [Текст] : монография / Б. А. Железко, А. Н. Морозевич. – Минск : Армита-Маркетинг: Менедж-мент, 1999. – 143 с.

14. Змитрович, А. И. Интеллектуальные информационные системы [Текст] / А. И. Змитрович. - Минск : ТетраСистемс, 1997. – 368 с.

15. Информационные технологии и вычислительные системы: Обработка информации и анализ данных. Программная инженерия. Математическое моделирование. Прикладные аспекты информатики / Под ред. С.В. Емельянова. - М.: Ленанд, 2015. - 104 с.

16. Искусственный интеллект и принятие решений: Интеллектуальный анализ данных. Моделирование поведения. Когнитивное моделирование. Моделирование и управление / Под ред. С.В. Емельянова. - М.: Ленанд, 2012. - 108 с.

17. Кабаков, Р. R в действии. Анализ и визуализация данных в программе R / Р. Кабаков. — М.: ДМК, 2016. — 588 с.

18. Калинина, В.Н. Анализ данных. компьютерный практикум (для бакалавров) / В.Н. Калинина, В.И. Соловьев. - М.: КноРус, 2017. - 240 с.

19. Кацко, И.А. Практикум по анализу данных на компьютере / И.А. Кацко, Н.Б. Паклин. — М.: КолосС, 2009. — 278 с.

20. Козлов, А.Ю. Статистический анализ данных в MS Excel: Учебное пособие / А.Ю. Козлов, В.С. Мхитарян, В.Ф. Шишов. - М.: Инфра-М, 2018. - 80 с.

21. Крянев, А.В. Метрический анализ и обработка данных / А.В. Крянев, Г.В. Лу-

кин, Д.К. Удумян. — М.: Физматлит, 2012. — 308 с.
22. Кулаичев, А.П. Методы и средства комплексного анализа данных: Учебное пособие / А.П. Кулаичев. - М.: Форум, 2018. - 160 с.
23. Лесковец, Ю. Анализ больших наборов данных / Ю. Лесковец, А. Раджараман. — М.: ДМК, 2016. — 498 с.
24. Маккинли, У. Python и анализ данных / У. Маккинли. - М.: ДМК, 2015. - 482 с.
25. Макшанов, А.В. Технологии интеллектуального анализа данных: Учебное пособие / А.В. Макшанов, А.Е. Журавлев. - СПб.: Лань, 2018. - 212 с.
26. Малинецкий, Г.Г. Проблемы математической истории: Основания, информационные ресурсы, анализ данных / Г.Г. Малинецкий, А.В. Коротаев. - М.: КД Либроком, 2009. - 256 с.
27. Мамонтов, В.Г. Химический анализ почв и использование аналитических данных. Лабораторный практикум: Учебное пособие / В.Г. Мамонтов. - СПб.: Лань, 2019. - 328 с.
28. Марманис, Х. Алгоритмы интеллектуального Интернета. Передовые методики сбора, анализа и обработки данных / Х. Марманис, Д. Бабенко. — М.: Символ, 2011. — 480 с.
29. Марчук, Г.И. Геронтология in silico: становление новой дисциплины. Математические модели, анализ данных и вычислительные эксперименты / Г.И. Марчук. — М.: БИНОМ. Лаборатория знаний, 2009. — 535 с.
30. Мастицкий, С.Э. Статистический анализ и визуализация данных с помощью R (черно-белые графики) / С.Э. Мастицкий. — М.: ДМК, 2015. — 496 с.
31. Миркин, Б.Г. Введение в анализ данных. учебник и практикум / Б.Г. Миркин. — Люберцы: Юрайт, 2016. — 174 с.
32. Модели и методы интеллектуального анализа данных: учебно-методическое пособие / сост. О. А . Попова, 2012/ [Электронный ресурс]. Режим доступа: https://www.docme.ru/doc/1155653/882.modeli-i-metody--intellektual._nogo-analiza-dannyh---u... – Дата доступа: 06.04.2019.
33. Мусаев, А.А. Интеллектуальный анализ данных: учебное пособие/ А. А. Мусаев – СПб.: СПбГТИ (ТУ), 2018 / [Электронный ресурс]. Режим доступа: http://sa.technolog.edu.ru/repository/iad_iadl.pdf – Дата доступа: 06.04.2019.
34. Нархид, Н. Apache Kafka. Потоковая обработка и анализ данных / Н. Нархид. - СПб.: Питер, 2019. - 320 с.
35. Наследов, А.Д. IMB SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных / А.Д. Наследов. — СПб.: Питер, 2013. — 416 с.
36. Наследов, А.Д. Математические методы психологического исследования. Анализ и интерпретация данных: Учебное пособие / А.Д. Наследов. — СПб.: Речь, 2012. — 392 с.
37. Ниворожкина, Л.И. Статистические методы анализа данных: Учебник / Л.И. Ниворожкина, С.В. Арженовский, А.А. Рудяга. - М.: Риор, 2018. - 320 с.
38. О развитии цифровой экономики: Декрет № 8 от 21 декабря 2017 г./ [Электронный ресурс]. Режим доступа: http://president.gov.by/ru/ – Дата доступа: 06.04.2019.
39. Об утверждении стратегии Республики Беларусь в сфере интеллектуальной собственности на 2012–2020 годы: Постановление Совета Министров Республики Беларусь от 21 марта 2018 г. № 208 (Национальный правовой Интернет-портал Республики Беларусь, 24.03.2018, 5/44945)/ [Электронный ресурс]. Режим доступа: http://www.pravo.by/document/ – Дата доступа: 06.04.2019.
40. Орлов, А.И. Организационно-экономическое моделирование .Ч.3 Статистические методы анализов данных. / А.И. Орлов. - М.: МГТУ , 2012. - 623 с.
41. Панкратова, Е.В. Анализ данных в программе SPSS для начинающих социологов / Е.В. Панкратова, И.Н. Смирнова, Н.Н. Мартынова. - М.: Ленанд, 2018. - 200 с.
42. Петрунин, Ю.Ю. Информационные технологии анализа данных: Учебное пособие / Ю.Ю. Петрунин. - М.: КДУ , 2010. - 292 с.
43. Пилипук А. В. Механизм и модели конкурентного функционирования / А. В. Пилипук // Современная конкуренция. – 2016. – Том 10. №3(57). – С. 119-142

44. Рафалович, В. Data mining, или интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович. - М.: SmartBook, 2018. - 352 с.
45. Резник, Г.А. Методы многомерного анализа статистических данных: Учебное пособие / Г.А. Резник. - М.: Финансы и статистика, 2008. - 400 с.
46. Романко, В.К. Статистический анализ данных в психологии: Учебное пособие / В.К. Романко. - М.: БИНОМ. ЛЗ, 2013. - 312 с.
47. Сидняев, Н.И. Теория планирования эксперимента и анализ статистических данных 2-е изд., пер. и доп. учебное пособие для магистров / Н.И. Сидняев. — Люберцы: Юрайт, 2016. — 495 с.
48. Симчера, В.М. Методы многомерного анализа статистических данных / В.М. Симчера. — М.: Финансы и статистика, 2008. — 400 с.
49. Сирота, А.А. Методы и алгоритмы анализа данных и их моделирование в MATLAB / А.А. Сирота. - СПб.: BHV, 2016. - 384 с.
50. Скобцов, В. Ю. Интеллектуальный анализ данных: генетические алгоритмы: учеб. – метод. пособие / В. Ю. Скобцов, Н. В. Лапицкая, С. Н. Нестеренков. – Минск: БГУИР, 2018 / [Электронный ресурс]. Режим доступа: https://ru.b-ok.org/book/3637836/3bca77 – Дата доступа: 06.04.2019.
51. Труды ИСА РАН: Математические модели социально-экономических процессов. Динамические системы. Управление рисками и безопасностью. Оптимизация, идентификация, теория игр. Обработка и анализ изображений и сигналов. Интеллектуальный анализ данных и распознавание образов/ Под ред. С.В. Емельянова. - М.: Красанд, 2013. - 128 с.
52. Тюрин, Ю.Н. Анализ данных на компьютере / Ю.Н. Тюрин, А.А. акаров. - М.: МЦНМО, 2016. - 368 с.
53. Форман, Дж. Много цифр: Анализ больших данных при помощи Excel / Дж. Форман. - М.: Альпина Паблишер, 2019. - 461 с.
54. Чашкин, Ю.Р. Математическая статистика. Анализ и обработка данных: Учебное пособие / Ю.Р. Чашкин; Под ред. С.Н. Смоленский. — Рн/Д: Феникс, 2010. — 236 с.
55. Чесноков, С.В. Детерминационный анализ социально-экономических данных / С.В. Чесноков. — М.: Книжный дом Либроком, 2013. — 168 с.
56. Яцков, Н. Н. Интеллектуальный анализ данных: пособие / Н. Н. Яцков. – Минск: БГУ, 2015 / [Электронный ресурс]. Режим доступа: http://elib.bsu.by/handle/123456789/114127?mode=full – Дата доступа: 06.04.2019.

СОДЕРЖАНИЕ