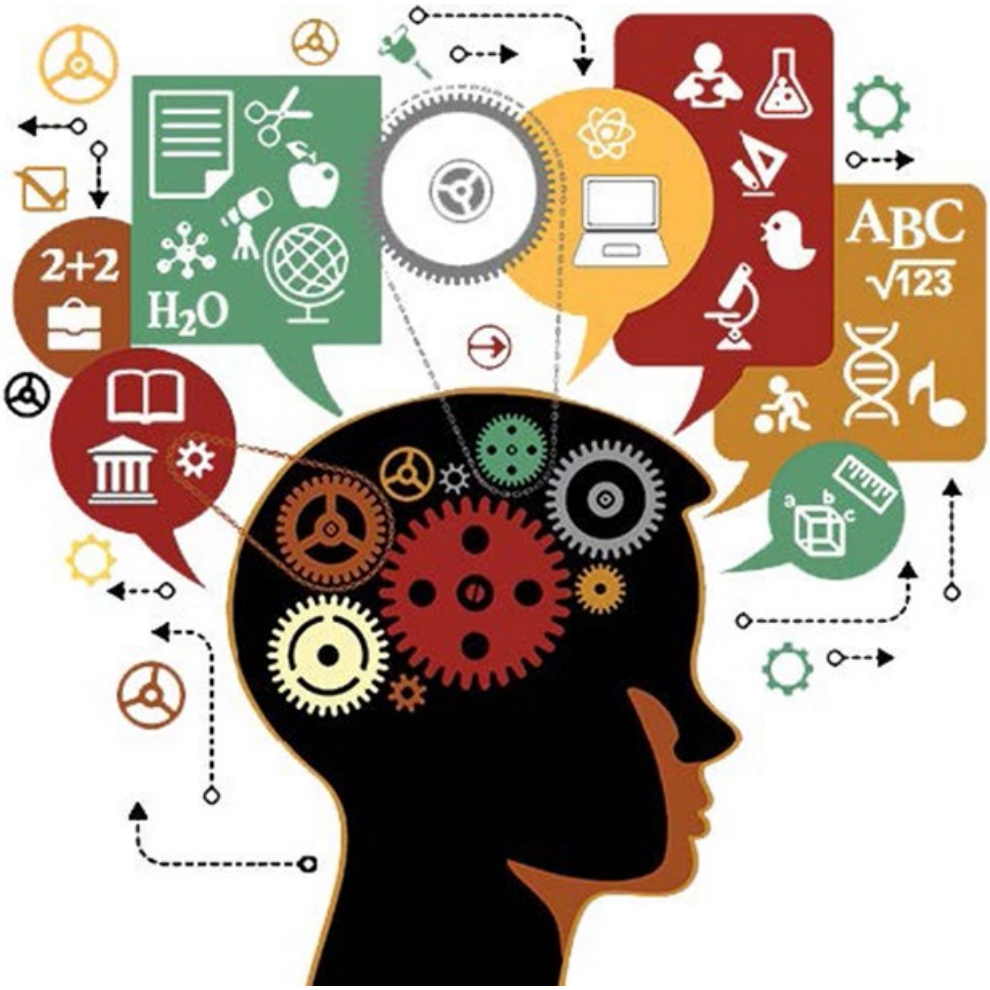
УНИВЕРСИТЕТ ИТМО

Н.Ф. Гусарова, Н.В. Добренко

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ



Санкт-Петербург
2019

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

Н.Ф. Гусарова, Н.В. Добренко

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

Учебно-методическое пособие

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлениям подготовки (специальностям)
45.03.04 - Интеллектуальные системы в гуманитарной сфере,
09.03.04. Информационные системы и технологии
в качестве учебного пособия для реализации основных профессиональных образовательных программ высшего образования бакалавриата

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург

2019

Гусарова Н.Ф., Добренко Н.В. **Интеллектуальные системы и технологии.** – СПб: Университет ИТМО, 2019. – 55 с.

Рецензент: Шалыто А.А., д.т.н., профессор

Пособие охватывает курс практических работ, необходимого для усвоения программы курса «Интеллектуальные системы и технологии». Представленные практические задания рассматриваю различные, в том числе междисциплинарные, прикладные области применения интеллектуальных систем и технологий, базирующихся на применении методов искусственного интеллекта.

Пособие предназначено для бакалавров по направлению подготовки 45.03.04 - Интеллектуальные системы в гуманитарной сфере, 09.03.04. Информационные системы и технологии и содержит материалы практических работ по дисциплине «Интеллектуальные системы и технологии».

УНИВЕРСИТЕТ ИТМО

Университет ИТМО – ведущий вуз России в области информационных и фотонных технологий, один из немногих российских вузов, получивших в 2009 году статус национального исследовательского университета. С 2013 года Университет ИТМО – участник программы повышения конкурентоспособности российских университетов среди ведущих мировых научно-образовательных центров, известной как проект «5 в 100». Цель Университета ИТМО – становление исследовательского университета мирового уровня, предпринимательского по типу, ориентированного на интернационализацию всех направлений деятельности.

Оглавление

Введение

Интеллектуальные системы и технологии все шире применяются как для решения задач первичного отбора информации и накопления отобранной информации, так и для поиска документов по заданной тематике и аналитической обработки накопленной информации. В связи с этим необходимость повышения эффективности систем поиска и анализа информации (оперативность и обоснованность получаемых решений) обуславливает быструю динамику их развития на научно-методологическом, технологическом и аппаратно-программном уровнях [1]. Развитие современных технологий уже вышло далеко за рамки разработок решений в уже существующих областях, на сегодняшний день для создания интеллектуальной системы привлекаются специалисты из прикладной области, лингвисты, нейрофизиологи, психологи, экономисты, информатики, программисты и т.д. Принципы и методы ИТ активно применяются («транслируются») во все новые междисциплинарные сферы деятельности человека.

Характерными признаками, согласно [2], интеллектуальных систем можно выделить:

1. Развитые коммуникативные способности — способ взаимодействия (интерфейса) конечного пользователя с системой, в частности возможность формулирования произвольного запроса в диалоге с ИС на языке, максимально приближенном к естественному.

2. Умение решать сложные, плохо формализуемые задачи, задачи, которые требуют построения оригинального алгоритма решения в зависимости от конкретной ситуации, для которой могут быть характерны неопределенность и динамичность исходных данных и знаний.

3. Способность к самообучению — автоматическое извлечение знаний для решения задач из накопленного опыта конкретных ситуаций.

4. Адаптивность — способность к развитию системы в соответствии с объективными изменениями модели проблемной области.

В данном пособии представлен цикл практических работ, базирующихся на применении методов искусственного интеллекта. Описание каждой практической работы содержит основные теоретические сведения, задание и порядок выполнения, состав отчета по работе. В каждой работе приводятся примеры постановки и решения практических задач с применение интеллектуальных систем и технологий. Практические работы могут выполняться как в подгруппах, так и индивидуально. Текущий контроль проводится в виде защиты отчета в форме доклада обучающегося по выполненной работе и ответов на вопросы преподавателя. Выполнение практических работ направлено на формирование компетенций, теоретического и практического освоения материала дисциплины. Промежуточный контроль проводится в устной форме.

В конце пособия приводится список рекомендуемых источников литературы.

В результате освоения материалов пособия обучающийся приобретает следующие профессиональные компетенции, а также умения и навыки:

- знание основных понятий и концепции теории построения интеллектуальных систем;
- умение ставить и решать типовые задачи, возникающие при поиске информации для организации работы с интеллектуальными системами;
- владение навыками работы в прикладных программных продуктах, использующих средства баз данных и лингвистического обеспечения;
- знание основных алгоритмов, используемых при построении интеллектуальных систем;
-умение разрабатывать алгоритмы интеллектуального и лингвистического анализа данных;
-владение навыками работы в прикладных программных продуктах, поддерживающих интеллектуальную обработку информации;
- знание основных подходов к построению архитектуры интеллектуальных систем;
- умение ставить и решать типовые задачи, возникающие при работе с интеллектуальными системами;
- владение навыками разработки архитектуры интеллектуальных систем.

Распределение трудозатрат студентов в аудитории и в процессе СРС представлено в соответствии с программой изучаемой дисциплины. В рамках самостоятельной работы студентам рекомендуется отвести на изучение теоретических материалов по плану лекций каждого раздела ориентировочно по 10 часов, на подготовку и выполнение практических работ и подготовку к промежуточному контролю также ориентировочно по 10 часов, и по 15 часов на самостоятельное изучение дополнительных источников информации.

ПРАКТИЧЕСКАЯ РАБОТА №1. ФАКТОРНЫЙ АНАЛИЗ

ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

Факторный анализ — совокупность методов многомерного статистического анализа, применяемых для изучения взаимосвязей между значениями переменных. Цели факторного анализа:

- сокращение числа переменных;
- определение взаимосвязей между переменными, их классификация.

Корреляционная матрица – это квадратная таблица, в которой на пересечении строк и столбцов располагаются коэффициенты корреляции между соответствующими параметрами.

Методики факторного анализа:

1. Анализ главных компонент.

После того, как мы получили несколько групп переменных, объединенных одним фактором, мы можем повторить процедуру регрессии данных. Таким образом, факторы последовательно выделяются один за другим. Так как, каждый последующий фактор, определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, то факторы оказываются независимыми друг от друга. Другими словами, некоррелированными или ортогональными.

Существует два критерия для вычисления количества главных компонентов:

- критерий Кайзера;
- критерий «каменистой осыпи».

2. Анализ главных факторов

В анализе главных компонент предполагается, что должна быть использована вся изменчивость переменных, тогда как в анализе главных факторов вы используете только изменчивость переменной, общую и для других переменных. В большинстве случаев эти два метода приводят к весьма близким результатам. Однако анализ главных компонент часто более предпочтителен как метод сокращения данных, в то время как анализ главных факторов лучше применять с целью определения структуры данных.

3. Факторные нагрузки.

Факторными нагрузками называются корреляции между переменными и несколькими факторами (или «новыми» переменными, которые выделены по умолчанию).

4. Вращение факторной структуры.

Можно изобразить факторные нагрузки в виде диаграммы рассеяния. На такой диаграмме каждая переменная представлена точкой, а её координаты равны факторным нагрузкам. Можно повернуть оси в любом направлении без изменения относительного положения точек; однако действительные координаты точек, то есть факторные нагрузки, должны, без сомнения, меняться. Если поворачивать оси относительно начала координат, то можно достичь ясного представления о нагрузках, определяющих группы переменных.

Типичными методами вращения являются стратегии варимакс, квартимакс, и эквимакс.

5. Косоугольные факторы.

6. Иерархический факторный анализ.

ЦЕЛЬ РАБОТЫ

Основной задачей практической работы является выделение наиболее показательных системных счётчиков, которые косвенно могут давать нам информацию об остальных параметрах системы.

ЗАДАНИЕ НА ПРАКТИЧЕСКУЮ РАБОТУ И ПОРЯДОК ВЫПОЛНЕНИЯ

Для получения исходных данных необходимо:

1. Подобрать подходящие для анализа данные из репозитория, предоставляемого преподавателем.

2. Создать таблицу данных в MS Excel.

Обработка исходных данных:

Необходимо заменить не числовые данные. Для этого вводится любая шкала соответствия, на выбор студента, с обязательным внесением в отчет легенды по шкале.

3. Следует удалить из анализа переменные, дисперсия которых равна нулю.

4. Постройте корреляционную матрицу и сделайте соответствующие выводы.

5. Откройте программу STATISTICA. Создайте новую таблицу данных (*Файл – >Создать*). Откройте модуль факторного анализа (*Анализ –> Многомерный разведочный анализ –> Факторный анализ*). (Переменные все, максимальное число факторов = 2).

6. Постройте график каменистой осыпи и сделайте соответствующие выводы о конечном количестве факторов.

7. Постройте График нагрузок. На получившемся графике наглядно можно увидеть взаимосвязь исследуемых параметров: сильно зависящие друг от друга параметры сконцентрированы в одной группе. Исследуйте каждую группу таких параметров и выведите главный – наиболее показательный. Удалите остальные переменные группы из дальнейшего

рассмотрения. При удалении переменных стоит ориентироваться на их пояснение в журнале производительности: в нем описывается значение и важность каждой. Повторяйте данный пункт до тех пор, пока не останется необходимого количества переменных.

8. Если на графике из-за наложения сложно разобрать названия переменных, то стоит обратиться к таблице факторных нагрузок из того же окна факторного анализа. В ней в числах выражена зависимость переменных от факторов. Группируя переменные по наиболее схожим показателям, мы можем понять, какие из них образуют группу на графике.

СОДЕРЖАНИЕ ОТЧЕТА

1. Описание принципа выбора счетчиков для наблюдения. Краткие описания каждого из них.
2. Описание условий, при которых снимались показания (запущенные действия и программы).
3. Исходная таблица данных в Excel.
4. Оценка дисперсии исследуемых параметров. Описание первого этапа сокращения данных.
5. Корреляционная матрица. Смысловое объяснение выявленных по матрице сильных зависимостей.
6. Результаты факторного анализа

6.1 График каменистой осыпи.

6.2 Описание каждого этапа сокращения данных: графики, таблицы собственных значений факторов, удаляемые переменные.

7. Выводы.

ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Цель работы: выделение наиболее показательных системных счётчиков, которые косвенно могут давать нам информацию об остальных параметрах системы.

Выполнение:

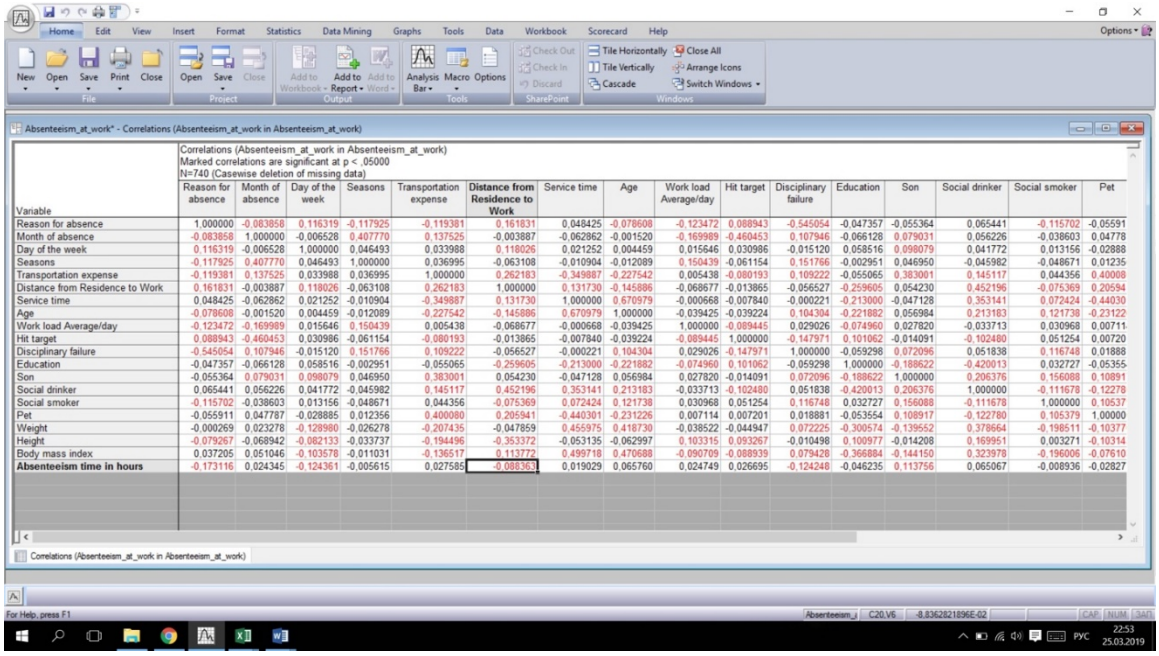
Для выполнения практической работы №1 мы используем программное обеспечение Statsoft® STATISTICA, MS Excel. В качестве датасета будет использован датасет Absenteeism at work, взятый с сайта https://archive.ics.uci.edu/.

Датасет содержит следующие признаки:

1. Individual identification (ID).
2. Reason for absence (ICD).
3. Month of absence.
4. Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6)).

5. Seasons.
6. Transportation expense.
7. Distance from Residence to Work (kilometers).
8. Service time.
9. Age.
10. Work load Average/day.
11. Hit target.
12. Disciplinary failure (yes=1; no=0).
13. Education (high school (1), graduate (2), postgraduate (3), master and doctor (4)).
14. Son (number of children).
15. Social drinker (yes=1; no=0).
16. Social smoker (yes=1; no=0).
17. Pet (number of pet).
18. Weight.
19. Height.
20. Body mass index.
21. Absenteeism time in hours (target).

Все признаки датасета являются числовыми, поэтому не требует преобразований. Рассчитав дисперсию в MS Excel мы не получили нулевых значений, поэтому будем рассматривать все признаки.



Variable	Reason for absence	Month of absence	Day of the week	Seasons	Transportation expense	Distance from Residence to Work	Service time	Age	Work load Average/day	Hit target	Disciplinary failure	Education	Son	Social drinker	Social smoker	Pet
Reason for absence	1,000000	-0,083858	0,116319	-0,117925	-0,119381	0,161831	0,048425	-0,078608	-0,123472	0,088943	-0,545054	-0,047357	-0,055364	0,065441	-0,115702	-0,05591
Month of absence	-0,083858	1,000000	-0,006528	0,407770	0,137525	-0,003887	-0,062862	-0,001520	-0,169989	-0,460453	0,107946	-0,066128	0,079031	0,056226	-0,038603	0,04778
Day of the week	0,116319	-0,006528	1,000000	0,046493	0,033988	0,118026	0,021252	0,004459	0,015646	0,030986	-0,015120	0,058516	0,098079	0,041772	0,013156	-0,02888
Seasons	-0,117925	0,407770	0,046493	1,000000	0,036995	-0,063108	-0,010904	-0,012089	0,150439	-0,061154	0,151766	-0,002951	0,046950	-0,045982	-0,048671	0,01235
Transportation expense	-0,119381	0,137525	0,033988	0,036995	1,000000	0,262183	-0,349887	-0,227542	0,005438	-0,080193	0,109222	-0,055065	0,383001	0,145117	0,044356	0,40008
Distance from Residence to Work	0,161831	-0,003887	0,118026	-0,063108	0,262183	1,000000	0,131730	-0,145886	-0,068677	-0,013865	-0,056527	-0,259605	0,054230	0,452196	-0,075369	0,20594
Service time	0,048425	-0,062862	0,021252	-0,010904	-0,349887	0,131730	1,000000	0,670979	-0,000668	-0,007840	-0,000221	-0,213000	-0,047128	0,353141	0,072424	-0,44030
Age	-0,078608	-0,001520	0,004459	-0,012089	-0,227542	-0,145886	0,670979	1,000000	-0,039425	-0,039224	0,104304	-0,221882	0,056984	0,213183	0,121738	-0,23122
Work load Average/day	-0,123472	-0,169989	0,015646	0,150439	0,005438	-0,068677	-0,000668	-0,039425	1,000000	-0,089445	0,029026	-0,074960	0,027820	-0,033713	0,030968	0,00711
Hit target	0,088943	-0,460453	0,030986	-0,061154	-0,080193	-0,013865	-0,007840	-0,039224	-0,089445	1,000000	-0,147971	0,101062	-0,014091	-0,102480	0,051254	0,00720
Disciplinary failure	-0,545054	0,107946	-0,015120	0,151766	0,109222	-0,056527	-0,000221	0,104304	0,029026	-0,147971	1,000000	-0,059298	0,072096	0,051838	0,116748	0,01888
Education	-0,047357	-0,066128	0,058516	-0,002951	-0,055065	-0,259605	-0,213000	-0,221882	-0,074960	0,101062	-0,059298	1,000000	-0,188622	-0,420013	0,032727	-0,05355
Son	-0,055364	0,079031	0,098079	0,046950	0,383001	0,054230	-0,047128	0,056984	0,027820	-0,014091	0,072096	-0,188622	1,000000	0,206376	0,156088	0,10891
Social drinker	0,065441	0,056226	0,041772	-0,045982	0,145117	0,452196	0,353141	0,213183	-0,033713	-0,102480	0,051838	-0,420013	0,206376	1,000000	-0,111678	-0,12278
Social smoker	-0,115702	-0,038603	0,013156	-0,048671	0,044356	-0,075369	0,072424	0,121738	0,030968	0,051254	0,116748	0,032727	0,156088	-0,111678	1,000000	0,10537
Pet	-0,055911	0,047787	-0,028885	0,012356	0,400080	0,205941	-0,440301	-0,231226	0,007114	0,007201	0,018881	-0,053554	0,108917	-0,122780	0,105379	1,00000
Weight	-0,000269	0,023278	-0,128980	-0,026278	-0,207435	-0,047859	0,455975	0,418730	-0,038522	-0,044947	0,072225	-0,300574	-0,139552	0,378664	-0,198511	-0,10377
Height	-0,079267	-0,068942	-0,082133	-0,033737	-0,194496	-0,353372	-0,053135	-0,062997	0,103315	0,093267	-0,010498	0,100977	-0,014208	0,169951	0,003271	-0,10314
Body mass index	0,037205	0,051046	-0,103578	-0,011031	-0,136517	0,113772	0,499718	0,470688	-0,090709	-0,088939	0,079428	-0,366884	-0,144150	0,323978	-0,196006	-0,07610
Absenteeism time in hours	-0,173116	0,024345	-0,124361	-0,005615	0,027585	-0,088363	0,019029	0,065760	0,024749	0,026695	-0,124248	-0,046235	0,113756	0,065067	-0,008936	-0,02827

Рис. 1.1 Скриншот исходных данных

	Reason	Month	Day of t	Seasons	Transpo	Distanc	Service	Age	Work lo	Hit targ	Disciplir	Educati	Son	Social d	Social s	Pet	Weight	Height	Body m	Absente	m time in hours
Reason fo	1,0000																				
Month of	-0,0839	1,0000																			
Day of the	0,1163	-0,0065	1,0000																		
Seasons	-0,1179	0,4078	0,0465	1,0000																	
Transporta	-0,1194	0,1375	0,0340	0,0370	1,0000																
Distance fr	0,1618	-0,0039	0,1180	-0,0631	0,2622	1,0000															
Service tir	0,0484	-0,0629	0,0213	-0,0109	-0,3499	0,1317	1,0000														
Age	-0,0786	-0,0015	0,0045	-0,0121	-0,2275	-0,1459	0,6710	1,0000													
Work load	-0,1235	-0,1700	0,0156	0,1504	0,0054	-0,0687	-0,0007	-0,0394	1,0000												
Hit target	0,0889	-0,4605	0,0310	-0,0612	-0,0802	-0,0139	-0,0078	-0,0392	-0,0894	1,0000											
Disciplinar	-0,5451	0,1079	-0,0151	0,1518	0,1092	-0,0565	-0,0002	0,1043	0,0290	-0,1480	1,0000										
Education	-0,0474	-0,0661	0,0585	-0,0030	-0,0551	-0,2596	-0,2130	-0,2219	-0,0750	0,1011	-0,0593	1,0000									
Son	-0,0554	0,0790	0,0981	0,0470	0,3830	0,0542	-0,0471	0,0570	0,0278	-0,0141	0,0721	-0,1886	1,0000								
Social drin	0,0654	0,0562	0,0418	-0,0460	0,1451	0,4522	0,3531	0,2132	-0,0337	-0,1025	0,0518	-0,4200	0,2064	1,0000							
Social smo	-0,1157	-0,0386	0,0132	-0,0487	0,0444	-0,0754	0,0724	0,1217	0,0310	0,0513	0,1167	0,0327	0,1561	-0,1117	1,0000						
Pet	-0,0559	0,0478	-0,0289	0,0124	0,4001	0,2059	-0,4403	-0,2312	0,0071	0,0072	0,0189	-0,0536	0,1089	-0,1228	0,1054	1,0000					
Weight	-0,0003	0,0233	-0,1290	-0,0263	-0,2074	-0,0479	0,4560	0,4187	-0,0385	-0,0449	0,0722	-0,3006	-0,1396	0,3787	-0,1985	-0,1038	1,0000				
Height	-0,0793	-0,0689	-0,0821	-0,0337	-0,1945	-0,3534	-0,0531	-0,0630	0,1033	0,0933	-0,0105	0,1010	-0,0142	0,1700	0,0033	-0,1031	0,3068	1,0000			
Body mass	0,0372	0,0510	-0,1036	-0,0110	-0,1365	0,1138	0,4997	0,4707	-0,0907	-0,0889	0,0794	-0,3869	-0,1442	0,3240	-0,1960	-0,0761	0,9041	-0,1210	1,0000		
Absenteeis	-0,1731	0,0243	-0,1244	-0,0056	0,0276	-0,0884	0,0190	0,0658	0,0247	0,0267	-0,1242	-0,0462	0,1138	0,0651	-0,0089	-0,0283	0,0158	0,1444	-0,0497	1,0000	

Рис. 1.2. Корреляционная матрица

Построим корреляционную матрицу (рис.1.2.) и выделим наиболее значимые коэффициенты корреляции, проанализируем их значения. По матрице корреляции можно увидеть, что наиболее коррелирующие параметры:

– Service time и Age (Стаж и Возраст).
– Вес и Индекс Массы тела

Далее согласно условию построим График каменистой осыпи (рис. 1.3).

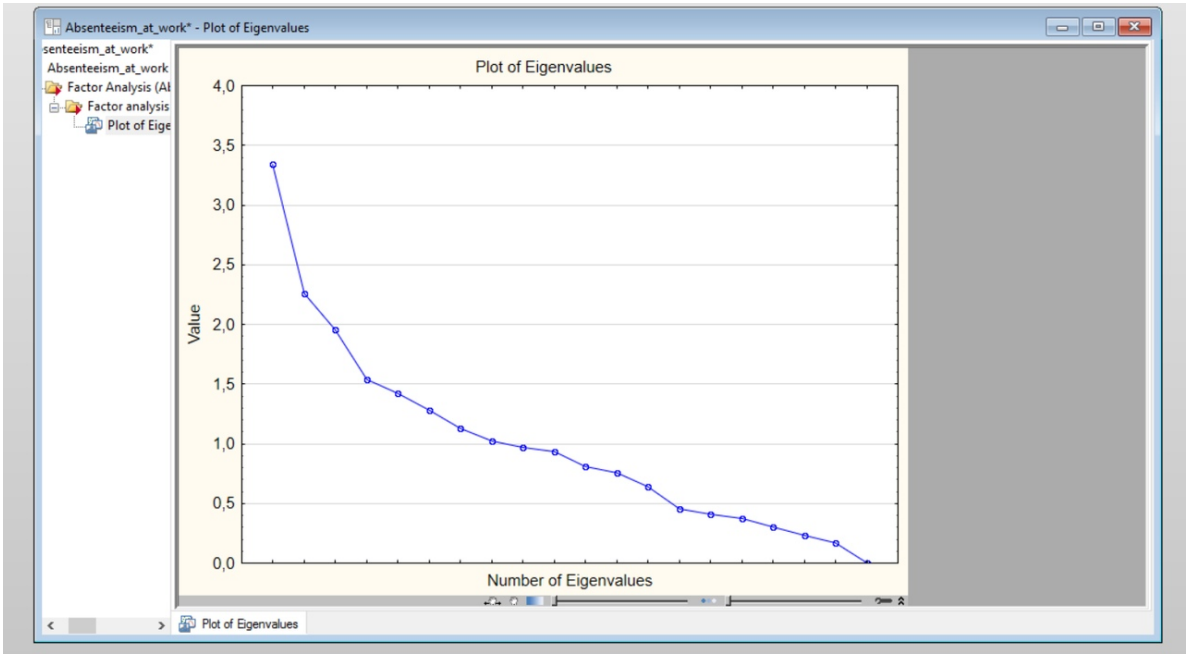


Рис. 1.3. График каменистой осыпи

Из полученного графика можно увидеть, что «перегиб» происходит на значении 4. Таким образом, можно выделить 4 фактора. Значит можно выделить 4 группы признаков.

Построим график нагрузок (рис. 1.4):

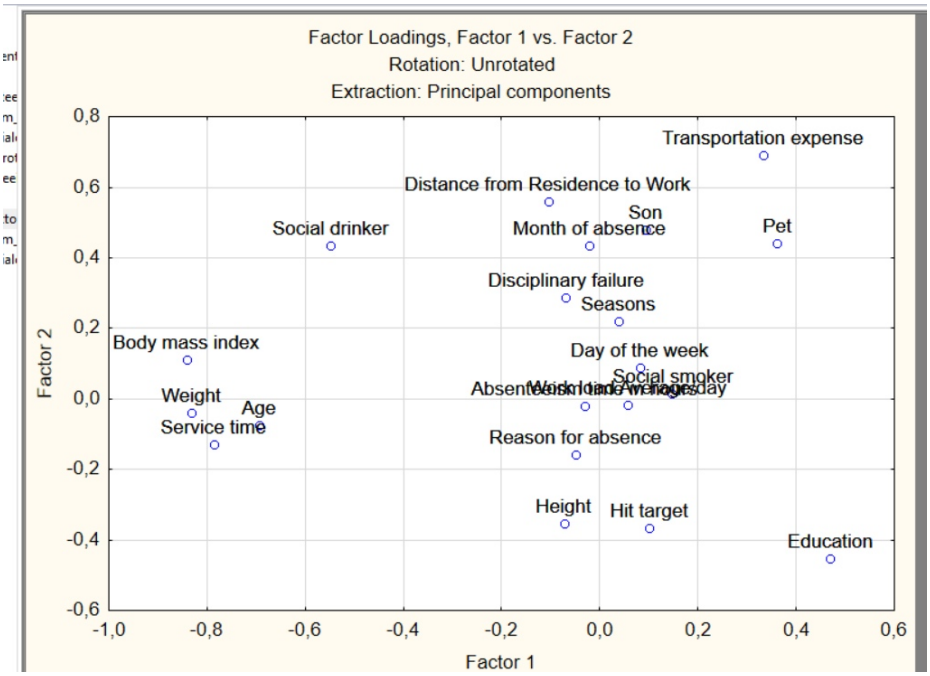


Рис. 1.4. График нагрузок

На графике отображены все параметры датасета. Мы видим, что они сгруппированы в разных областях. Из-за большого количества признаков сложно разобрать наименования параметров, а также непонятно, к какой области отнести тот или иной параметр, поэтому отобразим график нагрузок в формате 3D (рис. 1.5) и попробуем разметить группы признаков.

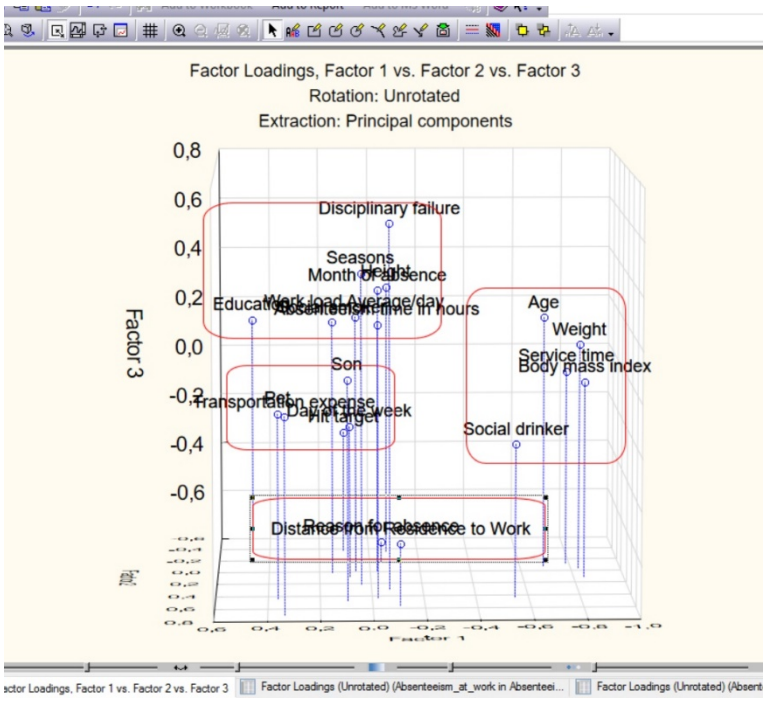


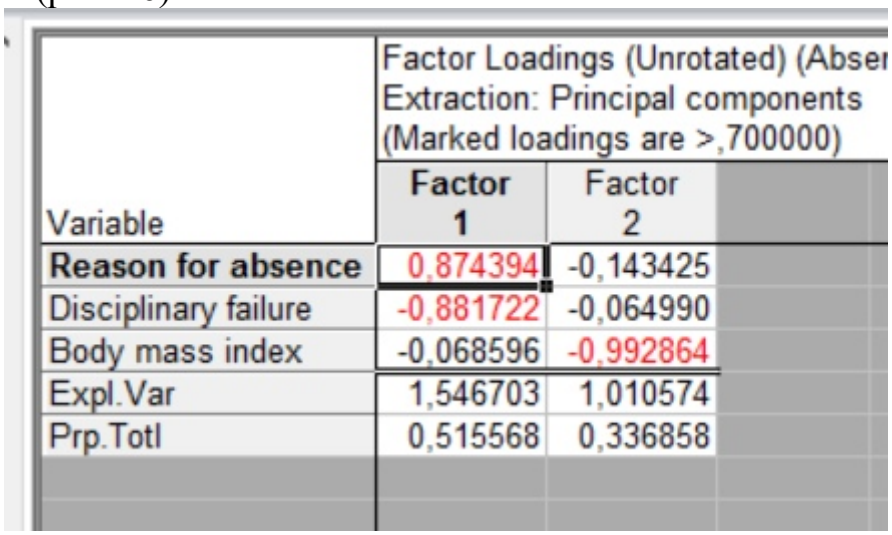
Рис. 1.5. График нагрузок в формате 3D

Таким образом, мы получили следующие группы признаков:

- Reasons for absence, day of the week, distance from residence to work;
- Month of absence, Seasons, Work load Average/day, Disciplinary failure, Son, Social smoker, Height, Absenteeism time in hours;
- Service time, Age, Social drinker, weight, body mass index;
- Transportation expense, Education, Pet.

Они формируются на основе того, что эти признаки взаимосвязаны, например, третья группа содержит индекс массы тела и вес - индекс массы тела вычисляется на основе веса, стаж работы зависит от возраста и к этим параметрам еще добавляется употребление алкоголя.

Попробуем уменьшить количество факторов, исключая те факторы, которые меньше всего влияют на целевой признак. Таким образом, получаем следующий результат, целевой признак очень сильно зависит от веса работников (рис.1.6).



Variable	Factor Loadings (Unrotated) (Abser Extraction: Principal components (Marked loadings are >,700000) Factor 1	Factor 2
Reason for absence	0,874394	-0,143425
Disciplinary failure	-0,881722	-0,064990
Body mass index	-0,068596	-0,992864
Expl.Var	1,546703	1,010574
Prp.Totl	0,515568	0,336858

Рис. 1.6. Скриншот таблицы нагрузок

Вывод:

Таким образом, проведя факторный анализ, мы выявили системные показатели, которые дают нам представление о системе признаков с помощью корреляционных матриц.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Для решения каких задач применяется факторный анализ?
2. Что показывает корреляционная матрица?
3. Что такое вращение факторной структуры?
4. Какие методы вращения вы знаете?

ПРАКТИЧЕСКАЯ РАБОТА №2. ОТБОР ПРИЗНАКОВ

ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

Практически в любой задаче моделирования возникает вопрос: какую модель зависимости применить, т.е. какие признаки использовать, а какие нет? Проблема отбора признаков (features selection) возникает из-за того, что на этапах постановки задачи и формирования данных ещё не ясно, какие признаки действительно важны, а какие не несут полезной информации или дублируют друг друга. Стремление учесть как можно больше потенциально полезной информации приводит к появлению избыточных (шумовых) признаков. По мере увеличения числа используемых признаков (сложности модели) средняя ошибка на обучающей выборке, как правило, монотонно убывает. Однако средняя ошибка на независимых контрольных данных сначала уменьшается, а затем возрастает. Это явление называют переобучением. В чрезмерно сложных моделях избыточные степени свободы ¾расходуются не столько на восстановление искомой зависимости, сколько на аппроксимацию ошибок измерений и погрешностей самой модели. Отбор признаков позволяет находить модель оптимальной сложности, при которой переобучение минимально.

Для проверки качества отбора признаков используется критерий – функционал Qint(μ,Xℓ), характеризующий качество метода μ по обучающей выборке Xℓ, например, ошибка обучения (training error). Чем меньше значение критерия Q(μ), тем выше качество метода μ. Критерий должен быть внешним, т.е. проверять качество метода μ по тем данным, которые не использовались в процессе обучения. Наиболее известные типы внешних критериев:

- Критерий средней ошибки на контрольных данных.
- Критерий скользящего контроля: берут несколько различных разбиений исходной выборки на обучение и контроль, и среднюю ошибку на контроле усредняют по разбиениям. Обратите внимание на то, что во всех критериях, использующих случайные разбиения, обучающие и контрольные подвыборки должны обладать теми же статистическими характеристиками, что и полная выборка.
- Критерии непротиворечивости: если модель алгоритмов A и метод обучения μ подобраны правильно, то настройка параметров модели по различным представительным подвыборкам должна приводить к одинаковым или почти одинаковым алгоритмам.
- Критерии регуляризации – наложить ограничения на вектор параметров алгоритма, либо ввести штраф за выход вектора параметров из некоторой допустимой области (например, чтобы норма вектора параметров || a || в алгоритме a = μ(Xℓ) не становилась слишком большой).

Как правило, используется совокупность критериев. Практическая рекомендация – отобрать некоторое количество лучших методов по

критерию скользящего контроля; а из них выбрать тот, для которого критерий регуляризации (либо критерий непротиворечивости) принимает наименьшее значение.

Задача отбора информативных признаков состоит в следующем. Будем считать, что объекты описываются набором признаков $F = \{f1,..., fn\}$. Вектор $f1(x), \ldots, fn(x) \in D1 \times \ldots \times Dn$, где Dj –множество допустимых значений признака fj, называется признаковым описанием объекта x. Пусть $G \subseteq F$ произвольное подмножество признаков. Будем обозначать через μG метод обучения, который строит алгоритмы, используя только признаки из подмножества G . Будем предполагать, что метод μG выбирает алгоритм из модели алгоритмов A(G), использующей только признаки из G . Число используемых признаков |G | будем называть сложностью модели A(G).

Для отбора информативных признаков используются различные методы:

- Полный перебор.

- Последовательное добавление признаков – простая стратегия жадного наискорейшего спуска: алгоритм добавляет к набору G по одному признаку, каждый раз выбирая тот признак, который приводит к наибольшему уменьшению внешнего критерия. Возможно также последовательное удаление признаков из полного набора, а также комбинация – алгоритм последовательного добавления–удаления.

- Построение и обход дерева возможных наборов признаков. Вершины дерева соответствуют наборам признаков. Корневая вершина соответствует пустому набору. Каждый дочерний набор образуется путём присоединения некоторого признака к родительскому набору. Чтобы избежать появления в дереве одинаковых наборов, отличающихся только порядком признаков, к дочерним наборам присоединяются только те признаки, номера которых превышают максимальный номер признака в родительском наборе. Как известно, существуют две стратегии полного обхода дерева: поиск в глубину (depth-first search, DFS) и поиск в ширину (breadth-first search, BFS). Обе позволяют вводить различные эвристики для сокращения перебора.

- Генетический алгоритм. Первое поколение наборов генерируется случайным образом. К этим наборам применяются операции скрещивания и мутации для порождения большого числа новых наборов. Затем производится .селекция: во второе поколение отбираются только B наборов, лучших по заданному внешнему критерию Q. Ко второму поколению также применяются операции скрещивания, мутации и селекции, и порождается третье поколение. Эволюционный процесс переходит от поколения к поколению до тех пор, пока не наступит стагнация, т.е. качество лучшего набора в поколении перестанет улучшаться.

- Случайный поиск с адаптацией. Если упростить генетический алгоритм, отказавшись от скрещивания, то получится алгоритм случайного поиска (stochastic search).

- Кластеризация признаков. Методы кластеризации в общем случае позволяют разбить выборку объектов на кластеры, состоящие из схожих объектов, и выделить в каждой группе по одному наиболее типичному представителю. То же самое можно проделать и с признаками, если определить функцию расстояния между признаками, например, через коэффициент корреляции или метрику Хемминга.

- Методы математического программирования. Используются для отбора признаков, главным образом, в линейных моделях регрессии и классификации. По существу, здесь также реализуется перебор признаков, но перебор внутри стандартных процедур математического программирования при поиске активных ограничений.

Более подробно со спецификой применения методов можно ознакомиться [3], в разделе «Критерии выбора моделей и методы отбора признаков».

ЦЕЛЬ РАБОТЫ

Реализовать алгоритм по отбору признаков на выбранном наборе данных.

ЗАДАНИЕ НА ПРАКТИЧЕСКУЮ РАБОТУ И ПОРЯДОК ВЫПОЛНЕНИЯ

1. Выбрать предметную область и набор данных, согласовать выбор с преподавателем. Можно использовать репозиторий ресурса http://archive.ics.uci.edu/ml/.
2. Выбрать алгоритм отбора признаков, согласовать выбор с преподавателем.
3. Реализовать алгоритм по отбору признаков на выбранном наборе данных. Выбор способа реализации алгоритма предоставляется студенту.
4. Проверить качество реализованного отбора признаков с помощью одного из критериев.

ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Для выполнения лабораторной работы был выбран датасет, содержащий показатели исследуемых пациентов, у которых был диагностирован рак молочной железы. Датасет взят из архива UCI: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra#.

Описание отобранного датасета:

Существует 10 предикторов, все количественные, и бинарная зависимая переменная, указывающая на наличие или отсутствие рака молочной железы. Предикторами являются антропометрические данные и параметры, которые могут быть собраны при анализе крови.

Атрибуты:
1. Age (years) – возраст.
2. BMI (kg/m2) – индекс массы тела.
3. Glucose (mg/dL) – уровень сахара в крови.
4. Insulin (μU/mL) – содержание инсулина в организме.
5. HOMA – индекс инсулинорезистентности.
6. Leptin (ng/mL) – содержание лептина в организме.
7. Adiponectin (μg/mL) – гормон адипонектин.
8. Resistin (ng/mL) – гормон резистин.
9. MCP-1(pg/dL) – содержание компонента MCP-1.

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	Classification
0	48	23.500000	70	2.707	0.467409	8.8071	9.702400	7.99585	417.114	1
1	83	20.690495	92	3.115	0.706897	8.8438	5.429285	4.06405	468.786	1
2	82	23.124670	91	4.498	1.009651	17.9393	22.432040	9.27715	554.697	1
3	68	21.367521	77	3.226	0.612725	9.8827	7.169560	12.76600	928.220	1
4	86	21.111111	92	3.549	0.805386	6.6994	4.819240	10.57635	773.920	1

Рис. 2.1. Пример данных из датасета

Импортируем необходимые модули и библиотеки:

In [1]:
```
import numpy as np
from sklearn.metrics import mean_squared_error
import pandas as pd
import seaborn as sns
from sklearn.cross_validation import train_test_split
from sklearn.linear_model import LinearRegression, Lasso, LassoCV, Ridge, RidgeCV
from sklearn.tree import DecisionTreeClassifier, export_graphviz

%matplotlib inline
data = pd.read_csv('dataR2.csv')
data.head()
```

Для работы с данными осуществим необходимые преобразования:

In [2]:
```
categorical_columns = [c for c in data.columns if data[c].dtype.name == 'object']
numerical_columns   = [c for c in data.columns if data[c].dtype.name != 'object']
print(categorical_columns)
print(numerical_columns)
```

```
[]
['Age', 'BMI', 'Glucose', 'Insulin', 'HOMA', 'Leptin', 'Adiponectin', 'Resistin', 'MCP.1', 'Classification']
```

In [3]:
```
data_numerical = data[numerical_columns]
data_numerical = (data_numerical - data_numerical.mean()) / data_numerical.std()
data_numerical.describe()
```

Out[3]:

	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP.1	C
count	1.160000e+02	1.160000e+02	1.160000e+02	1.160000e+02	1.160000e+02	1.160000e+02	1.160000e+02	1.160000e+02	1.160000e+02	1.
mean	1.397350e-16	-2.704135e-15	2.660707e-16	5.665966e-16	7.465293e-17	1.110223e-16	-2.268300e-16	1.081510e-16	-8.384098e-16	9.
std	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.
min	-2.066791e+00	-1.835032e+00	-1.677817e+00	-7.529064e-01	-6.116289e-01	-1.162682e+00	-1.245715e+00	-9.294081e-01	-1.413085e+00	-1
25%	-7.634769e-01	-9.180839e-01	-5.346511e-01	-5.614786e-01	-4.879188e-01	-7.455135e-01	-6.877622e-01	-6.330746e-01	-7.651317e-01	-1
50%	-8.078837e-02	1.599667e-02	-2.571837e-01	-4.060072e-01	-3.607999e-01	-3.307086e-01	-2.671475e-01	-3.146104e-01	-1.830650e-01	8.
75%	8.501505e-01	7.289308e-01	1.867643e-01	1.169240e-01	4.470017e-02	5.610726e-01	2.389324e-01	2.444781e-01	4.782652e-01	8.
max	1.967277e+00	2.190508e+00	4.581849e+00	4.812180e+00	6.138135e+00	3.318769e+00	4.070983e+00	5.437492e+00	3.364413e+00	8.

Отбор признаков будем осуществлять на основе их важности. Первым методом выберем линейную регрессию. Используем функцию LinearRegression. Для оценки точности алгоритма линейной регрессии будем использовать RMSE.

```
In [8]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 11)

In [9]: lin = LinearRegression(normalize=True)
        lin.fit(X_train, y_train)

Out[9]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=True)

In [10]: y_lin_pred = lin.predict(X_test)

In [11]: def rmse(y, p):
             return np.sqrt(mean_squared_error(y, p))
         def beatiful_coef(coefs, feature_names=data.columns):
             return pd.DataFrame(coefs, index=feature_names,
                                 columns=['coef']).sort_values('coef',
                                                               ascending=False)

In [12]: rmse(y_test, y_lin_pred)

Out[12]: 0.9861422652300279
```

Полученный коэффициент точности равен 0,98614.
В результате проделанной работе получили следующий результат:

```
In [15]: beatiful_coef(ridge.coef_,feature_names=X_train.columns)
```

Out[15]:

	coef
Glucose	0.136434
Resistin	0.083716
Insulin	0.063370
HOMA	0.033702
MCP.1	0.030060
Adiponectin	0.004415
Age	-0.025034
Leptin	-0.036648
BMI	-0.086257

Самыми показательными параметрами получились содержание сахара в крови и гормон резистин.В качестве второго метода для отбора признаков выберем метод случайного леса.

Ансамблевые алгоритмы на основе деревьев решений, такие как случайный лес (random forest), позволяют оценить важность признаков.

```
In [24]: from sklearn.ensemble import RandomForestRegressor
         forest = RandomForestRegressor(n_estimators=100, random_state=17)
         forest.fit(X_train, y_train)
         forest_test_pred = forest.predict(X_test)

         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\ensemble\weight_boostin
         th_tests is an internal NumPy module and should not be imported. It will b
           from numpy.core.umath_tests import inner1d

In [25]: rmse(y_test, forest_test_pred)

Out[25]: 0.6212527085531415

In [26]: beatiful_coef(forest.feature_importances_,feature_names=X_train.columns)
```

Out[26]:

	coef
Resistin	0.187068
Glucose	0.163129
Age	0.162173
BMI	0.127742
Adiponectin	0.096024
Leptin	0.087352
HOMA	0.073114
MCP.1	0.060333
Insulin	0.043066

В результате выявили два важных признака: резистин и содержание сахара в крови. Получили результат RMSE равный 0.6213, что намного меньше в сравнении с оценкой алгоритма линейной регрессии. Значит, метод случайного леса является более точным для отбора признаков.

Выводы

Сравнив результаты двух методов отбора признаков (линейная регрессия и случайный лес), мы получили схожие результаты. Оба метода указывают на важность признаков Glucose и Resistin, но точность метода Random Forest больше, чем метода Линейной регрессии. В данном случае, учитывая специфику выборки, имеет смысл брать выявленные признаки в рассмотрение. Подводя итог, было отобрано 2 признака из девяти рассматриваемых. Сокращение числа признаков увеличит скорость обработки данных и качество результатов, так как малоинформативные признаки будут отброшены.

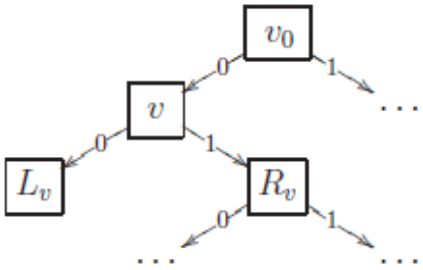
КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое нормализация данных?
2. Перечислите основных преимущества отбора признаков?
3. В чем основная идея отбора признаков с помощью генетического алгоритма?
4. В чём отличия внутренних и внешних критериев?

ПРАКТИЧЕСКАЯ РАБОТА №3. ДЕРЕВЬЯ РЕШЕНИЙ

ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

Бинарное решающее дерево – это алгоритм классификации, задающийся бинарным деревом, в котором каждой внутренней вершине $v \in V$ приписан предикат $\beta_v : X \to \{0, 1\}$, каждой терминальной вершине $v \in V$ приписано имя класса $c_v \in Y$. При классификации объекта $x \in X$ он проходит по дереву путь от корня до некоторого листа, в соответствии с Алгоритмом 1.



Алгоритм 1. Классификация объекта $x \in X$ бинарным решающим деревом

```
1: v := v0;
2: пока вершина v внутренняя
3: если βv(x) = 1 то
4: v := Rv; (переход вправо)
5: иначе
6: v := Lv; (переход влево)
7: вернуть cv.
```

Объект x доходит до вершины v тогда и только тогда, когда выполняется конъюнкция $K_v(x)$, составленная из всех предикатов, приписанных внутренним вершинам дерева на пути от корня v_0 до вершины v. Естественное требование максимизации информативности конъюнкций $K_v(x)$ означает, что каждая из них должна выделять как можно больше обучающих объектов, допуская при этом как можно меньше ошибок. Задача построения дерева минимальной сложности, правильно классифицирующего заданную выборку, в общем случае является NP-полной задачей. На практике применяют различные эвристики.

В данной практической работе рассматривается алгоритм 2 построения решающего дерева ID3 (Induction of Decision Tree). Идея алгоритма заключается в последовательном дроблении выборки на две части до тех пор, пока в каждой части не окажутся объекты только одного класса. Алгоритм записывается в виде рекурсивной процедуры LearnID3, которая строит дерево по заданной подвыборке U. Для построения полного дерева она применяется ко всей выборке и возвращает указатель на корень построенного дерева:

v0 := LearnID3 (Xℓ).

Алгоритм 2. Рекурсивный алгоритм синтеза бинарного решающего дерева ID3

```
Вход: U – обучающая выборка; B – множество элементарных предикатов;
Выход: возвращает корневую вершину дерева, построенного по выборке U;
1:  ПРОЦЕДУРА LearnID3 (U);
2:  если все объекты из U лежат в одном классе c ∈ Y то
3:  создать новый лист v;
4:  cv := c;
5:  вернуть (v);
6:  найти предикат с максимальной информативностью:
      β := arg max I(β, U); β∈B
7:  разбить выборку на две части U = U0UU1 по предикату β:
    U0 := {x ∈ U : β(x) = 0};
    U1 := {x ∈ U : β(x) = 1};
8:  если U0 = ∅ или U1 = ∅ то
9:  создать новый лист v;
10: cv := класс, в котором находится большинство объектов из U;
11: иначе
12: создать новую внутреннюю вершину v;
13: βv := β;
14: Lv := LearnID3 (U0); (построить левое поддерево)
15: Rv := LearnID3 (U1); (построить правое поддерево)
16: вернуть (v);
```

Проблемы, возникающие при реализации алгоритма:

П1. Как определить информативность предиката? – 2 варианта.

Точный критерий Фишера основан на вычислении вероятности реализации пары событий: информативность предиката φ(x), разделяющего события правильного обнаружения p_c нужных объектов и правильного необнаружения N_c ненужных объектов, относительно класса c ∈ Y по выборке X^ℓ

$$I_c(\varphi, X^\ell) = -\ln \frac{C_{P_c}^{p_c(\varphi)} C_{N_c}^{n_c(\varphi)}}{C_{P_c+N_c}^{p_c(\varphi)+n_c(\varphi)}}.$$

в случае произвольного числа классов Y = {1, . . . ,M}

$$I(\varphi, X^l) = -\ln \frac{C_{P1}^{p1} \ldots C_{PM}^{pM}}{C_l^p} \qquad (*)$$

где P_c - число объектов класса c в выборке X^ℓ, из них p_c объектов выделяются предикатом φ, p = p1+…+ pM.

Энтропийный = информационный критерий основан на сравнении энтропии (математического ожидания количества информации) о выборке до и после применения предиката φ. Если в выборке есть P объектов класса c и N остальных, то ее энтропия равна

$$H(P,N) = -\frac{P}{P+N}\log_2\frac{P}{P+N} - \frac{N}{P+N}\log_2\frac{N}{P+N}.$$

Если предикат φ выделил p объектов из P, принадлежащих классу c, и n объектов из N, не принадлежащих классу c, то энтропия выборки после получения этой информации стала равной

$$H_\varphi(P,N,p,n) = \frac{p+n}{P+N}H(p,n) + \frac{P+N-p-n}{P+N}H(P-p,N-n).$$

Тогда энтропийный = информационный критерий для двух классов вводится как

$$IGain(\varphi, X^\ell) = H(P,N) - H(P,p,N,n),$$

а для большого числа классов – как

$$I(\varphi, X^\ell) = \sum_{c\in Y} h\left(\frac{P_c}{\ell}\right) - \frac{p}{\ell}\sum_{c\in Y} h\left(\frac{p_c}{p}\right) - \frac{\ell-p}{\ell}\sum_{c\in Y} h\left(\frac{P_c-p_c}{\ell-p}\right),$$

где введена функция $h(z) \equiv -z\log_2 z$.

(**)

П2. На шаге 6 Алгоритма 2 выбирается предикат β из заданного семейства B, задающий максимально информативное ветвление дерева – разбиение выборки на две части $U = U_0 \cup U_1$.

На практике применяются различные критерии ветвления.

1. Критерий, ориентированный на отделение заданного класса $c \in Y$.
$I(\beta, U) = \max I_c(\beta, U)$, $c \in Y$.
2. Критерии, ориентированные на отделение сразу нескольких классов – (*) и (**).
3. D-критерий _ число пар объектов из разных классов, на которых предикат β принимает разные значения. В случае двух классов он имеет вид
$I(\beta, U) = p(\beta)(N - n(\beta)) + n(\beta)(P - p(\beta))$.

Другие варианты критериев можно найти в [2].

П3. Как выбрать условия деления выборки в каждом узле?

На практике в качестве элементарных предикатов чаще всего берут простые пороговые условия вида β(x) = [fj(x) >< dj]. Конъюнкции, составленные из таких термов, хорошо интерпретируются и допускают запись на естественном языке. Однако никто не запрещает использовать в вершинах дерева любые разделяющие правила: шары, гиперплоскости, и, вообще говоря, произвольные бинарные классификаторы. Когда мощность |B| не велика, эта задача решается полным перебором, дальше нужны эвристики.

П4. Как удалить поддеревья, имеющие недостаточную статистическую надёжность? = проблема редукции решающих деревьев.

Наиболее распространенные эвристики:

Предредукция (pre-pruning) или критерий раннего останова досрочно прекращает дальнейшее ветвление в вершине дерева, если информативность $I(\beta, U)$ для всех предикатов $\beta \in B$ ниже заданного порогового значения I_0. Для этого на шаге 8 условие ($U0 = \varnothing$ или $U1 = \varnothing$) заменяется условием $I(\beta, U) \le I_0$. Порог I_0 является управляющим параметром метода.

Постредукция (post-pruning) просматривает все внутренние вершины дерева и заменяет отдельные вершины либо одной из дочерних вершин (при этом вторая дочерняя удаляется), либо терминальной вершиной. Процесс замен продолжается до тех пор, пока в дереве остаются вершины, удовлетворяющие критерию замены. Критерием замены является сокращение числа ошибок на контрольной выборке, отобранной заранее, и не участвовавшей в обучении дерева. Стандартная рекомендация - оставлять в контроле около 30% объектов.

Для реализации постредукции контрольная выборка Xk пропускается через построенное дерево. При этом в каждой внутренней вершине v запоминается подмножество $S_v \subseteq X_k$ попавших в неё контрольных объектов. Если $Sv = \varnothing$, то вершина v считается ненадёжной и заменяется терминальной по мажоритарному правилу: в качестве c_v берётся тот класс, объектов которого больше всего в обучающей подвыборке U, пришедшей в вершину v. Затем для каждой внутренней вершины v вычисляется число ошибок, полученных при классификации выборки Sv следующими способами:

1) $r(v)$ - классификация поддеревом, растущим из вершины v;
2) $r_L(v)$ - классификация поддеревом левой дочерней вершины L_v;
3) $r_R(v)$ - классификация поддеревом правой дочерней вершины R_v;
4) $r_c(v)$ - отнесение всех объектов выборки S_v к классу $c \in Y$.

Эти величины сравниваются, и, в зависимости от того, какая из них оказалась минимальной, принимается, соответственно, одно из четырёх решений:

1) сохранить поддерево вершины v;
2) заменить поддерево вершины v поддеревом левой дочерней вершины Lv;
3) заменить поддерево вершины v поддеревом правой дочерней вершины Rv;
4) заменить поддерево v терминальной вершиной класса $c = \arg\min_{c \in Y} r_c(v)$.

Предпросмотр (look ahead) заключается в том, чтобы на шаге 6, вместо вычисления информативности для каждого $\beta \in B$ построить поддерево небольшой глубины h. Во внутреннюю вершину v помещается тот предикат β, при котором поддерево допускает наименьшее число ошибок. Этот алгоритм работает заметно дольше, но строит более надёжные и простые

деревья. Более подробно со спецификой применения методов можно ознакомиться в [8].

ЦЕЛЬ РАБОТЫ

Реализовать и сравнить два разных алгоритма деревьев решений.

ЗАДАНИЕ НА ПРАКТИЧЕСКУЮ РАБОТУ И ПОРЯДОК ВЫПОЛНЕНИЯ

1. Выбрать статистический ряд, подходящий для построения дерева решений, выделить на нем обучающую и контрольную выборку, согласовать с преподавателем.
2. Реализовать алгоритм 2. В ходе реализации алгоритма выбрать способы решения проблем П1-П4, причем хотя бы одну из проблем решить двумя способами.

ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Цель работы: на выбранном наборе данных реализовать алгоритм случайного дерево и случайного леса, классифицировать важность влияния признаков на целевой признак. Сделать соответствующие выводы.

Практическая работа выполнена на наборе данных Parkinson Dataset with replicated acoustic features Data Set.
Ссылка:
http://archive.ics.uci.edu/ml/datasets/Parkinson+Dataset+with+replicated+acoustic+features+

Аннотация: набор данных содержит акустические характеристики, извлеченные из 3-х голосовых записей, на которые записана непрерывная речь каждого из 80 объектов исследования (40 из которых имеют болезнь Паркинсона).

Датасет содержит следующие признаки:

1. **ID**: уникальный идентификатор объекта исследования.
2. **Recording**: порядковый номер голосовой записи.
3. **Status**: 0=здоров; 1= болезнь Паркинсона
4. **Gender**: 0=муж; 1=жен
5. **Pitch local perturbation measures** (фазовое дрожание цифрового сигнала данных):

– **Relative jitter** (Jitter_rel): средняя абсолютная разница между последовательными периодами, деленная на средний период

– **Absolute jitter** (Jitter_abs): изменение основной частоты между циклами (средняя абсолютная разница между последовательными периодами).

– **Relative average perturbation** (Jitter_RAP): относительное среднее возмущение (средняя абсолютная разница между периодом и средним значением для него и его двух соседей, деленная на средний период).

– **Pitch perturbation quotient** (Jitter_PPQ): коэффициент возмущения периода (средняя абсолютная разница между периодом и средним значением, деленная на средний период).

6. **Amplitude perturbation measures**:

– **local shimmer** (Shim_loc): средняя абсолютная разница между амплитудами двух последовательных периодов, деленных на среднюю амплитуду. 3,81% - предел обнаружения паталогий.

– **shimmer in dB** (Shim_dB): средняя абсолютная разница от ln разности между двумя последовательными периодами. Предел для обнаружения патологий составляет 0,350 дБ.

– **3-point amplitude perturbation quotient** (Shim_APQ3): это трехточечный коэффициент дрожания амплитуды, средняя абсолютная разница между амплитудой периода и средней амплитудой его соседей, деленная на среднюю амплитуду.

– **5-point amplitude perturbation quotient** (Shim_APQ5): это коэффициент дрожания амплитуды из пяти точек, средняя абсолютная разница между амплитудой периода и средней амплитуды его и его четырех ближайших соседей, деленная на среднюю амплитуду.

– **11-point amplitude perturbation quotient** (Shim_APQ11): это коэффициент дрожания амплитуды, равный 11 точкам, средняя абсолютная разница между амплитудой периода и средней амплитуды его и его десяти ближайших соседей, деленная на среднюю амплитуду. Параметр APQ и дает 3,070% в качестве порога для патологии.

7. **Harmonic-to-noise ratio measures**: (отношение гармоник к шуму)
   – in the frequency band 0-500 Hz (HNR05).
   – in 0-1500 Hz (HNR15).
   – in 0-2500 Hz (HNR25).
   – in 0-3500 Hz (HNR35).
   – in 0-3800 Hz (HNR38).
8. **Mel frequency cepstral coefficient-based spectral measures** (еденица высоты тона) of order 0 to 12 (MFCC0 – MFCC12) and their **derivatives** (производные) (Delta0 – Delta12).
9. **Recurrence period density entropy** (RPDE): энтропия плотности периода повторения.
10. **Detrended fluctuation analysis** (DFA): метод Пенга или анализ отклонения колебаний.

11. **Pitch period entropy** (PPE): энтропия основного периода.
12. **Glottal-to-noise excitation ratio** (GNE): соотношение возбуждения гортань-шум.

Импортируем библиотеки, необходимые для выполнения практической работы:

```
In [1]: import numpy as np
        import pandas as pd
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.model_selection import KFold
        from sklearn.model_selection import cross_val_score
```

Укажем путь к датасету и импортируем его:

```
In [2]: df = pd.read_csv('C:/Users/Po/Downloads/ReplicatedAcousticFeatures-Parkinson
```

Модифицируем датасет:

```
In [3]: del df['ID']
        del df['Recording']
        df.head()
```

Out[3]:

	Status	Gender	Jitter_rel	Jitter_abs	Jitter_RAP	Jitter_PPQ	Shim_loc	Shim_dB	Shim_AF
0	0	1	0.25546	0.000015	0.001467	0.001673	0.030256	0.26313	0.017
1	0	1	0.36964	0.000022	0.001932	0.002245	0.023146	0.20217	0.013
2	0	1	0.23514	0.000013	0.001353	0.001546	0.019338	0.16710	0.011
3	0	0	0.29320	0.000017	0.001105	0.001444	0.024716	0.20892	0.014
4	0	0	0.23075	0.000015	0.001073	0.001404	0.013119	0.11607	0.006

5 rows × 46 columns

1. Реализуем алгоритм «Случайное дерево».

```
In [6]: clf_tree = DecisionTreeClassifier(random_state=17)

In [7]: df1 = df['Status'] # целевой столбец
        del df['Status']

In [8]: clf_tree.fit(df, df1)

Out[8]: DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                    max_features=None, max_leaf_nodes=None,
                    min_impurity_decrease=0.0, min_impurity_split=None,
                    min_samples_leaf=1, min_samples_split=2,
                    min_weight_fraction_leaf=0.0, presort=False, random_state=17,
                    splitter='best')
```

```
In [9]: feature_importances = pd.DataFrame(clf_tree.feature_importances_,
                                           index = df.columns,
                                           columns=['importance']).sort_values('importance',ascending=False)
```

```
In [10]: feature_importances.head(10)
```

Out[10]:

	importance
Delta0	0.375076
Shim_loc	0.102298
MFCC4	0.076957
HNR38	0.055652
PPE	0.049365
MFCC3	0.034794
Delta5	0.027778
Delta8	0.026825
RPDE	0.026813
Delta11	0.026804

```
In [11]: kf = KFold(shuffle = True, random_state = 17, n_splits = 4)
```

```
In [12]: score = cross_val_score(estimator = clf_tree, X = df, y = df1, cv=kf, scoring ="accuracy")
```

```
In [13]: print(score)
         score.mean()
```

```
[0.7        0.7        0.73333333 0.65       ]
```

```
Out[13]: 0.6958333333333333
```

2. Реализуем алгоритм «Случайный лес».

```
In [15]: from sklearn.ensemble import RandomForestClassifier
```

```
In [16]: clf = RandomForestClassifier(random_state=17, n_estimators = 100, n_jobs=-1)
```

```
In [18]: clf.fit(df, df1)
```

```
Out[18]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                     max_depth=None, max_features='auto', max_leaf_nodes=None,
                     min_impurity_decrease=0.0, min_impurity_split=None,
                     min_samples_leaf=1, min_samples_split=2,
                     min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=-1,
                     oob_score=False, random_state=17, verbose=0, warm_start=False)
```

```
In [19]: feature_importances = pd.DataFrame(clf.feature_importances_,
                                           index = df.columns,
                                           columns=['importance']).sort_values('importance',ascending=False)
```

```
In [20]: feature_importances.head(10)
```

Out[20]:

	importance
HNR35	0.049012
MFCC4	0.046259
HNR38	0.045557
Delta11	0.044276
MFCC10	0.041315
Delta0	0.038235
Delta3	0.037973
Delta7	0.035280
MFCC11	0.034364
Delta1	0.033248

```
In [17]: kf = KFold(shuffle = True, random_state = 17, n_splits = 4)
         score = cross_val_score(estimator = clf, X = df, y = df1, cv=kf, scoring ="accuracy")
         print(score)
         print(score.mean())
```

```
[0.83333333 0.83333333 0.81666667 0.71666667]
0.8
```

Выводы

При применении первого алгоритма мы выявили признаки, которые влияют на систему, но не дают достаточной точности при вычислении. В исследовании с применением второго алгоритма непосредственное значение важности признаков упало, но точность стала достаточной для заключения о качестве работе алгоритма. Некоторые признаками, проявились при исследовании влияния дважды. Из значимых признаков были выделены следующие:

– отношение гармоник к шуму на высоких частотах 0-3500 Гц и 0-3800 Гц (HNR38 и HNR35);

Гармоника – это элементарная составляющая сложного гармонического колебания (сигнала). Отношение "сигнал/шум" – это отношение среднеквадратического значения величины входного сигнала к среднеквадратическому значению величины шума, выраженное в децибелах, данное значение позволяет определить долю шума в измеряемом сигнале по отношению к полезному сигналу.

– высота тона и производная этого значения (MFCC11 и Delta11).

Данный параметр измеряется в психофизических единицах высоты звука *Мел*. Он основан на статистической обработке большого числа данных о субъективном восприятии высоты звуковых тонов. С помощью формулы перевода можно преобразовать значение частоты звука (Гц) в значение высоты (мел). График данной зависимости имеет экспоненциальную форму, поэтому производная данной характеристики также выделяется значимым

параметром, влияющим на систему, ведь при экспонента мало изменяющаяся функция.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Как деревья решений, так и глубокие нейронные сети являются нелинейными классификаторами, т. е. они разбивают пространство посредством сложной границы решений. Почему в таком случае модель дерева решений настолько интуитивно понятнее глубокой нейронной сети?
2. Чем отличаются алгоритмы деревья решения от других алгоритмов классификации?
3. Опишите достоинства и недостатки решающих деревьев?
4. Что такое решающее дерево?

ПРАКТИЧЕСКАЯ РАБОТА №4. НЕЙРОСЕТЕВОЕ РАСПОЗНАВАНИЕ ПЕЧАТНЫХ СИМВОЛОВ

ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

1. Подготовка эталонных образов

Набор эталонных образов задается преподавателем. Примером такого набора является последовательность из десяти цифр от 0 до 9. В этом примере число образов M=10. В случае, когда каждый класс образов характеризуется лишь своим эталоном, имеем число классов, также равное M.

Каждый образ формируется в виде графического файла в битовом формате. Тип файла (расширение) определяется используемыми в среде MATLAB типами графических файлов. Рекомендуется использовать расширение ***tif***.

Для создания графических файлов образов удобно использовать среду **"*Adobe Photoshop*"**. В этом случае при создании каждого файла необходимо проделать следующую последовательность операций:

1) создать новый файл, задав его параметры:

- имя : *XXXX*;
- ширина: N1 пикселей;
- высота: N2 пикселей;
- цветовой режим: битовый формат.

При этом значения N1, N2=8…20 задаются преподавателем.

2) используя инструменты типа «Кисть», «Ластик» и др. создать требуемый образ символа.

3) с помощью команды «Сохранить как» сохранить созданный образ в виде файла типа ***tif*** в заданной преподавателем папке.

На рис. 4.1 приведены примеры графических символов цифр при N1=10, N2=12 пикс.

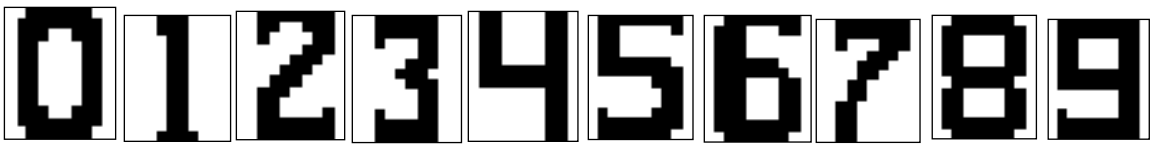


Рис. 4.1. Примеры графических символов цифр

2. Создание и обучение НС в среде MATLAB

На данном этапе выполнение работы в среде MATLAB производится с помощью программы ***sr_newff*** , которая реализует следующие функции:

- формирование числовых массивов эталонных образов, используемых в качестве обучающих;

- подготовка данных, необходимых для создания нейронной сети (НС);

- создание НС, задание параметров обучения НС и обучение НС.

Эталонный образ каждого символа представлен в виде вектора-столбца [N,1], число элементов N которого равно числу признаков (иначе говоря, N – размерность пространства признаков). Такой вектор-столбец формируется из двумерного массива-изображения [N1,N2], который, в свою очередь, формируется при считывании графического файла образа с помощью команд:

imread (FILENAME) - процедура чтения графического файла;

X = reshape (A,[N,1]) - процедура преобразования двумерного массива A[N1,N2] в одномерный вектор-столбец X[N,1], где N=N1*N2.

Процедура умножения массива на 1 приводит к смене типа элементов массива с *logical* (для элементов битового формата) на *double*.

Для удовлетворительной работы НС недостаточно формирования лишь одного обучающего образа для каждого класса (типа символа) образов. Это связано с тем, что распознаваемые образы (на этапе работы НС в режиме распознавания) всегда отличаются от обучающих по ряду причин:

- различие шрифтов и стилей печатных символов;

- погрешности сканирования и неточности совмещения символа и окна сканирования;

- низкое качество печати, дефекты бумаги и т.д.

В силу указанных причин для надежного распознавания образов НС следует обучать на достаточно представительном множестве образов, входящих в один и тот же класс.

В программе ***sr_newff*** формирование дополнительных обучающих образов производится путем незначительного искажения эталонных образов, считываемых из графических файлов. Искажение образа-эталона каждого класса реализуется путем добавления к нему равномерного (по площади изображения) шума типа «Соль и перец», представляющего собой случайное искажение отдельных пикселей изображения. Степень искажения характеризуется числом p=[0;1], определяющим долю искаженных пикселей.

Такой подход при формировании образов позволяет, во-первых, быстро получать большое число обучающих образов, и, во-вторых, регулировать (путем изменения значения p) степень разброса множества образов в пределах одного класса.

Подготовка данных, необходимых для создания НС, включает в себя:

1) формирование двумерного массива обучающих образов XR[N,K], каждый столбец которого представляет собой набор N признаков одного образа, а число столбцов K равно числу обучающих образов;

2) формирование двумерного массива желаемых откликов YR[NY,K], где NY – число выходов НС (т.е., число нейронов выходного слоя); K – число

обучающих образов. Отклик YR[:,k] (в общем случае – вектор-столбец) соответствует k-му обучающему образу – вектору XR[:,k];
3) формирование двумерного массива R[N,2], определяющего минимальное R(n,1) и максимальное R(n,2) значение n-го признака, n=1,…,N.

Создание НС. В общем случае НС ***net*** создается с помощью команды:

net = nnnnn (P1,P2,…PL), где

- nnnnn – тип НС;
- P1,…,PL – параметры НС.

В настоящей работе используется НС типа многослойного персептрона newff, которая задается командой:

net = newff (R, [A1 A2 … AL], {F1 F2 … FL}, BTF, PF), где

R - массив минимальных и максимальных значений входных нейронов (признаков);
Ai - число нейронов i-го слоя, начиная с первого скрытого слоя, i=1,…,L;
Fi - функция активации нейронов i-го слоя, по умолчанию ‘tansig’;
BTF - функция обучения сети, по умолчанию ‘trainlm’;
PF - критерий остановки, по умолчанию ‘mse’ (минимум ско).

Дополнительные параметры, используемые при создании сети:
net.performFcn=’msereg’ - обучение НС производится с помощью метода регуляризации;
net.performParam.ratio=0.1 - значение параметра регуляризации;
net.trainParam.show=5 - число эпох, через которое производится вывод параметров процедуры обучения;
net.trainParam.epochs=500 - максимальное число эпох при обучении НС;
net.trainParam.goal=0.02 - значение целевой функции, по достижении которого процесс обучения НС прекращается.

Процесс обучения НС запускается командой:

net = train (net, XR, YR) .

Для решения задач распознавания печатных символов рекомендуется использовать трехслойную НС (один скрытый слой) с числом нейронов:

N=120 - во входном слое;
A1=20 - в скрытом (промежуточном) слое;
A2=1 - в выходном слое.

При использовании большего числа нейронов процедура обучения НС может занять слишком много времени. Для рекомендованных значений параметров НС (в том числе и дополнительных) и общем числе обучающих образов (для всех заданных классов) K=100…200 время обучения НС составляет 5…20 мин.

3. Распознавание печатных символов с помощью обученной НС

Работа НС, т.е. формирование отклика Y при входном воздействии в виде вектора-столбца X[N,1] производится командой:

Y = sim (net, X).

В случае, когда желаемый отклик принимает целочисленные значения, рекомендуется использовать округление до ближайшего целого, т.е.

Y = round (sim (net, X)).

Тестирование работы НС при распознавании печатных символов с различной степенью искажения производится с помощью программы ***sr_work*** , исходными данными для которой являются:

SX.tif - имя графического файла образа-эталона;

N - число пикселей изображения образа;

NT - число тестируемых образов, полученных путем искажения эталона;

P - доля искаженных пикселей [0; 1].

На рисунках 4.2-4.6 представлены некоторые примеры распознавания символов, изображенных на рис. 4.1, с помощью обученной НС. Обучение проводилось при числе обучающих образов M=10 для каждого вида символа и параметре искажения символов p=0,1.

Рис. 4.2. Неверные распознавания символа **«0»**, искаженного 20% шума «Соль и Перец», результат распознавания (слева направо): **«2», «3», «5»**

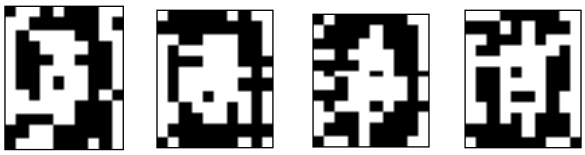


Рис. 4.3. Правильные распознавания символа **«0»**, искаженного 20% шума «Соль и Перец», результат распознавания (слева направо): **«0», «0», «0», «0»**

Рис. 4.4. Неверные распознавания символа **«4»**, искаженного 20% шума «Соль и Перец», результат распознавания (слева направо): **«3», «5», «6»**

Рис. 4.5. Правильные распознавания символа **«4»**, искаженного 20% шума «Соль и Перец», результат распознавания (слева направо):
«4», «4», «4», «4», «4»

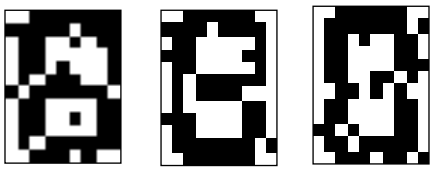


Рис. 4.6. Результаты распознавания символа **«8»**, искаженного 10% шума «Соль и Перец», результат распознавания (слева направо):
«8», «8», «6»

Результаты распознавания символов, представленные на рисунках 4.2-4.6, демонстрируют хорошее распознавание с помощью НС даже при сильном искажении (параметр p>0,1). Для объективной оценки качества работы НС необходимо вычисление вероятностных характеристик распознавания. При правильном выборе параметров обучения сети и использовании не менее 100 обучающих образов можно получить вероятность правильного распознавания символов порядка 0,6…0,9 (в зависимости от вида распознаваемого символа) при параметре искажения p=0,1…0,2.

Качество работы НС характеризуется вероятностями правильной классификации образа i-го класса, i=1,…,M. Оценка вероятностей производится по формуле:

$$\hat{P}_{np}(i) = \frac{N_{np}}{N_0} \ ,$$

где N_{np} - число правильных распознаваний образа i-го класса; N_0 - общее число распознаваний образов i-го класса. Число определяется экспериментально при запуске программы ***sr_work*** при значениях =10…100.

ЦЕЛЬ РАБОТЫ

Исследование возможностей распознавания печатных символов с помощью нейронных сетей, а также построение нейронных сетей в среде MATLAB.

ЗАДАНИЕ НА ПРАКТИЧЕСКУЮ РАБОТУ И ПОРЯДОК ВЫПОЛНЕНИЯ

1. Подготовить графические файлы эталонных образов для символов, заданных преподавателем, или самостоятельная подготовка эталонных (обучающих) образов печатных символов в виде набора графических файлов.

2. В среде MATLAB создать и обучить нейронную сеть (НС), предназначенную для распознавания печатных символов.

3. Исследовать зависимость качества работы НС от:

- степени искажения символов (параметр p);
- числа нейронов в скрытом слое.

ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Цель работы: обучить нейронную сеть распознавать рукописные цифры.

Перед нами стоит задача классификации: есть десять эталонных образцов, цифры от 0 до 9.

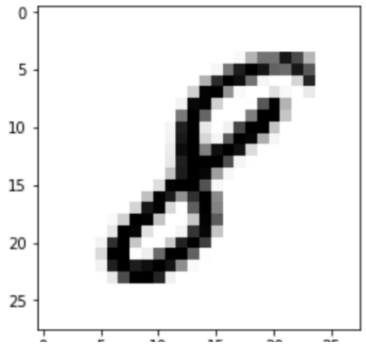
1. Импорт библиотек для работы и загрузка датасета MNIST. MNIST - объёмная датасет образцов рукописного написания цифр. Он включает в себя 60,000 изображений для тренировки and 10,000 тестовых изображений.

In [2]:
```
import ssl

ssl._create_default_https_context = ssl._create_unverified_context

import tensorflow as tf
(x_train, y_train), (x_test, y_test) = tf.keras.datasets.mnist.load_data()
```

In [24]:
```
from skimage.io import imread, imshow, imsave
from skimage import img_as_float
```

2. Части x_train и x_test содержат изображения в цветовом пространстве RGB в оттенках серого (уровни яркости от 0 до 255). Выведем пример рукописной цифры из датасета MNIST.

```
In [3]: import matplotlib.pyplot as plt
        %matplotlib inline
        image_index = 3434 # Индекс числа из MNIST
        print(y_train[image_index]) # Вывод числа
        plt.imshow(x_train[image_index], cmap='Greys')

        8

Out[3]: <matplotlib.image.AxesImage at 0x1308e2860>
```



3. Тренировочная выборка: 60000 изображений 28*28 пикселей. Тестовая выборка: 10000 изображений 28*28 пикселей.

```
In [4]: x_train.shape

Out[4]: (60000, 28, 28)

In [5]: x_test.shape

Out[5]: (10000, 28, 28)
```

4. Нормализация изображений. Мы должны нормализовать наши данные, как это всегда требуется в моделях нейронных сетей. Мы можем достичь этого, разделив коды RGB на 255 (это максимальное значение для цветового пространства RGB). Преобразуем данные в float. Теперь уровень яркости определяется значением от 0 до 1.

```
In [6]: # Преобразование массива
        x_train = x_train.reshape(x_train.shape[0], 28, 28, 1)
        x_test = x_test.reshape(x_test.shape[0], 28, 28, 1)
        input_shape = (28, 28, 1)
        # Преобразование float
        x_train = x_train.astype('float32')
        x_test = x_test.astype('float32')
        # Нормализация под цветовое пространство RGB
        x_train /= 255
        x_test /= 255
        print('x_train shape:', x_train.shape)
        print('Number of images in x_train', x_train.shape[0])
        print('Number of images in x_test', x_test.shape[0])

        x_train shape: (60000, 28, 28, 1)
        Number of images in x_train 60000
        Number of images in x_test 10000
```

5. Построение нейронной сети. Мы будем конфигурировать нашу сеть с помощью Kernas API с использованием библиотеки TensorFlow, которая включает инструменты для работы с сетью.

Входные значения: 784 (28*28) значений от 0 до 255 (уровней яркости изображения) На входной слой – 800 нейронов. На выходной слой 10 нейронов и вероятность что на изображении эталонная цифра.

Используется последовательная модель: модель, в которой слои нейронной сети идут друг за другом.

```
In [7]: # Импорт модулей из Keras включ. модели и слои
        from keras.models import Sequential
        from keras.layers import Dense, Conv2D, Dropout, Flatten, MaxPooling2D

        model = Sequential()
        model.add(Conv2D(28, kernel_size=(3,3), input_shape=input_shape))
        model.add(MaxPooling2D(pool_size=(2, 2)))
        model.add(Flatten()) #
        model.add(Dense(128, activation=tf.nn.relu))
        model.add(Dropout(0.2))
        model.add(Dense(10,activation=tf.nn.softmax))

        Using TensorFlow backend.
```

6. Компиляция. Метод обучения – Adam adaptive moment estimation. Он сочетает в себе и идею накопления движения и идею более слабого обновления весов для типичных признаков.

```
In [8]: model.compile(optimizer='adam',
                      loss='sparse_categorical_crossentropy',
                      metrics=['accuracy'])
        model.fit(x=x_train,y=y_train, epochs=10)
```

7. Запуск (10 эпох) и определение точности.

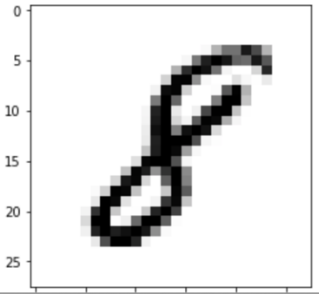
```
Epoch 1/10
60000/60000 [==============================] - 21s 343us/step - loss: 0.2120 - acc: 0.9347
Epoch 2/10
60000/60000 [==============================] - 20s 333us/step - loss: 0.0858 - acc: 0.9734
Epoch 3/10
60000/60000 [==============================] - 20s 333us/step - loss: 0.0582 - acc: 0.9821
Epoch 4/10
60000/60000 [==============================] - 20s 338us/step - loss: 0.0440 - acc: 0.9860
Epoch 5/10
60000/60000 [==============================] - 20s 332us/step - loss: 0.0357 - acc: 0.9880
Epoch 6/10
60000/60000 [==============================] - 21s 345us/step - loss: 0.0290 - acc: 0.9901
Epoch 7/10
60000/60000 [==============================] - 20s 337us/step - loss: 0.0241 - acc: 0.9921
Epoch 8/10
60000/60000 [==============================] - 20s 337us/step - loss: 0.0224 - acc: 0.9922
Epoch 9/10
60000/60000 [==============================] - 20s 341us/step - loss: 0.0213 - acc: 0.9930
Epoch 10/10
60000/60000 [==============================] - 20s 340us/step - loss: 0.0185 - acc: 0.9937
```

Точность модели после сравнения составила 98%.

```
In [9]: model.evaluate(x_test, y_test)

        10000/10000 [==============================] - 1s 133us/step

Out[9]: [0.055517800836151585, 0.9868]
```

8. Проверим индивидуальное распознавание цифр.

```
In [3]: import matplotlib.pyplot as plt
        %matplotlib inline
        image_index = 3434 # Индекс числа из MNIST
        print(y_train[image_index]) # Вывод числа
        plt.imshow(x_train[image_index], cmap='Greys')

        8

Out[3]: <matplotlib.image.AxesImage at 0x1308e2860>
```



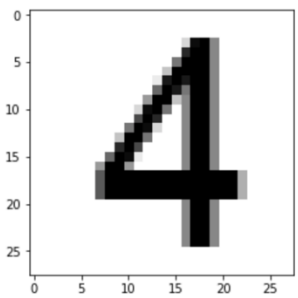
9. Проверим сеть, как она будет работать с искаженными цифрами. Для этого с помощью Adobe Photoshop создаем изображение размером 32*32 и рисуем цифру. Для более оптимизированной работе тестовое изображение сохраняется в формате .tif.

Нормализация изображения для теста.

```
In [76]: test = imread('/Users/thecontrey/Documents/6.tif')
         imshow(test)
         test = test[:,:,1]
         print(test)
         test.shape
```

Без искажений.

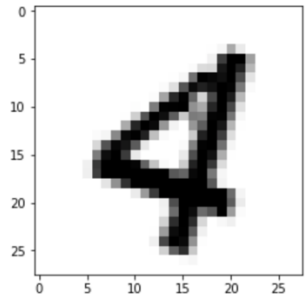
```
In [52]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())

         4
```



Фильтр «diffusion» 10% + поворот вправо на 20*.

```
In [56]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

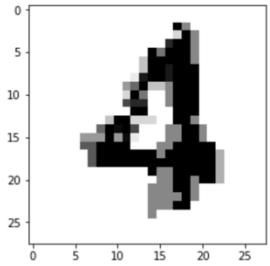
```
4
```



Фильтр «diffusion» 20%.

```
In [54]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

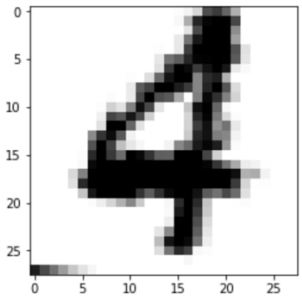
```
2
```



Фильтр «diffusion» 15%, масштабирование, лишний элемент, поворот.

```
In [60]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

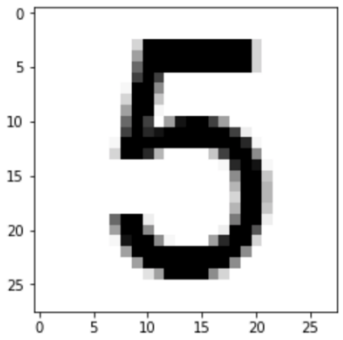
```
6
```



Попробуем другую цифру. Без искажений.

```
In [66]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

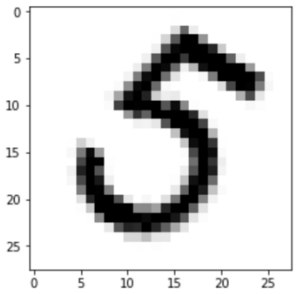
```
5
```



Поворот на 30%.

```
In [69]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

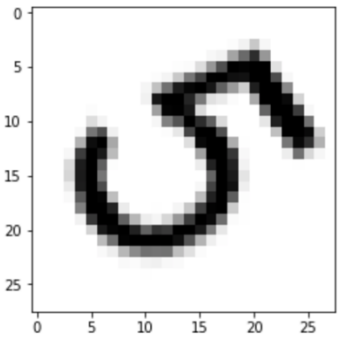
```
5
```



Поворот на 40%.

```
In [71]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

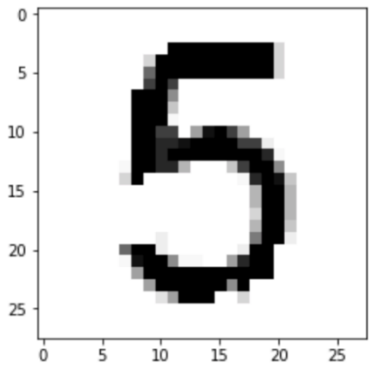
```
0
```



Фильтр «diffusion» 10%.

```
In [73]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

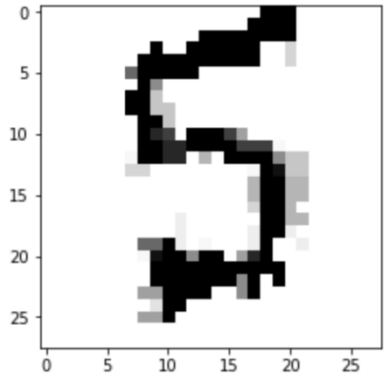
```
5
```



Фильтр «diffusion» 30%.

```
In [75]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

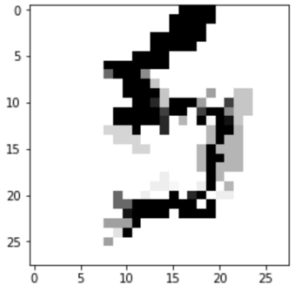
```
5
```



Фильтр «diffusion» 50%.

```
In [77]: image = test
         plt.imshow(image.reshape(28, 28),cmap='Greys')
         pred = model.predict(image.reshape(1, 28, 28, 1))
         print(pred.argmax())
```

```
5
```



Вывод

НС довольно восприимчива к изменениям очертаний символов, особенно на схожих цифрах (например 3, 8 и 9). Однако показывает хорошо на цифрах с более уникальными очертаниями (например 5, 1). Одинаково хорошие результаты получаются при повороте цифр, вплоть до 40 градусов в любую сторону цифры продолжают распознаваться.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Назовите основные достоинства и недостатки нейронных сетей?
2. Какие практические задачи решаются с применением нейронных сетей?
3. Назовите негативные последствия переобучения нейронной сети?
4. Дайте характеристику основных этапов построения нейронной сети?

ПРАКТИЧЕСКАЯ РАБОТА №5. АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ

ОСНОВНЫЕ ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ

Современные информационные системы позволяют накапливать большие объемы, в том числе и медицинских данных. В них содержаться знания, которые можно анализировать, выявлять и извлекать закономерности. При этом размерность хранимых данных, определяемая числом различных признаков, описывающих, например, состояние здоровья пациента, весьма велика и порой достигает нескольких десятков и сотен показателей. Поэтому задача снижения размерности признакового пространства и выделения наиболее информативных признаков является весьма актуальной.

Краулинг (crawling) – процесс сканирования огромного количества веб-страниц специализированными программами краулерами (спайдерами, пауками, сканерами), которые симулируют действия пользователя поисковой системы. Другими словами, краулеры загружают различные веб-страницы, читают содержание, анализируют их на наличие требуемой информации, которая копируется в соответствующее хранилище в случае обнаружения и предоставляется другим частям поисковой системы.

Например, у человека болит голова, он хочет узнать, в чем возможные причины. Он вводит поисковый запрос, краулер загружает страницы, которые были прочитаны и проанализированы на наличие нужной информации.

Краулеры также переходят по всем ссылкам, указанным на странице (на сайтах могут быть еще ссылки, ведущие на страницы с болями в спине), и процесс сканирования, анализа повторяется до тех пор, пока не будет проанализировано требуемое число веб-страниц или пока не будет достигнута определенная цель. Веб-краулер в основном используется для создания копий всех посещаемых страниц в Интернет-пространстве (посещаемые страницы обычно отмечены другим цветом. Если человек не раз ищет причины головной боли в интернете, и он был не удовлетворён результатами поиска в прошлый раз, то на уже посещённые сайты он не зайдет), которые в последующем обрабатываются и индексируются поисковой системой, в результате чего пользователь имеет возможность осуществлять быстрый поиск информации (например, с помощью истории браузера. Если наоборот была важная информация на сайте о головных болях, например, название лекарства, то в случае забытия этого названия, можно с помощью истории снова быстро зайти на страницу и вспомнить). Кроме того, краулеры могут использоваться для автоматизации работы по обслуживанию на веб-сайте: проверка ссылок, HTML-кода и т. д. Краулеры также применяются для сбора определенных типов информации из веб-страниц. Например, сбор электронных почтовых адресов для рассылки спама.

Одной из наиболее перспективных и быстроразвивающихся платформ по анализу медицинских и биомедицинских текстов является система cTAKES [9]. Это модульная система с открытым исходным кодом, которая разрабатывается сообществом исследователей из разных институтов совместно с частной клиникой Mayo. Основное предназначение cTAKES заключается в создании основы для систем извлечения информации из клинических текстов. Система реализована на платформе UIMA с применением фреймворка OpenNLP. cTAKES предоставляет набор программных модулей, которые используют наборы правил и модели машинного обучения, настроенные на анализ клинических текстов. Система позволяет проводить: 1) базовую обработку медицинских текстов, включая морфологический анализ и различные виды синтаксического анализа; 2) извлечение именованных сущностей, определение их типа и характера упоминания их в тексте; 3) сопоставление терминов с концептами метатезауруса UMLS; 4) поверхностный семантический анализ (установление семантических ролей); 5) разрешение кореференции; 6) извлечение семантических связей между сущностями. В системе также присутствуют готовые модули для решения прикладных задач: извлечения названий лекарственных препаратов и определения статуса пациента курящий/некурящий. Платформа интегрирует в себе большое число современных разработок в области анализа медицинских и биомедицинских текстов и продолжает развиваться. Заметим, что cTAKES ориентирована на обработку только английских текстов и на сегодняшний день не поддерживает другие языки.

ЦЕЛЬ РАБОТЫ

Провести теоретическое и практическое исследование анализа медицинских данных из социальных сетей и тематических форумов.

ЗАДАНИЕ НА ПРАКТИЧЕСКУЮ РАБОТУ И ПОРЯДОК ВЫПОЛНЕНИЯ

1. Провести анализ, какой спектр задач можно решить путем краулинга медицинских данных из соц. сетей и из тематических форумов?
2. Привести подробное описание применяемых механизмов, обеспечивающих гарантированное качество обработки и интерпретации медицинских данных? Описать по каким параметрам контролируется качество?
3. Найти научную статью, посвященную исследованиям в данной научной области, провести ее подробный анализ и представить отчет, содержащий подобное исследование медицинских текстов из социальных сетей или тематических форумов.

ПРИМЕР ВЫПОЛНЕНИЯ РАБОТЫ

Цель работы: Провести теоретическое и практическое исследование анализа медицинских данных из социальных сетей и тематических форумов.

Задание на данную практическую работу должно содержать два раздела, первый раздел – это теоретический анализ современных подходов к решению задач краулинга медицинских данных, второй раздел – это подробный анализ научной статьи из данной предметной области и подобное исследование медицинских текстов из социальных сетей или тематических форумов.

Первая (теоретическая) часть.

Типы краулеров [10]

1. Сфокусированный краулер (Focused Web Crawler) – это краулер, задача которого заключается в том, чтобы загрузить связанные с друг другом страницы по конкретной теме. Еще такой вид краулеров называют тематическим (Topic Crawler). Принцип его работы основывается на том, что каждый раз переходя к новому документу, данный краулер проверяет насколько он релевантен обозначенной теме, переход осуществляется только на документы, соответствующие теме. Его достоинства состоят в том, что он экономичен и не требует значительных вычислительных ресурсов.

2. Инкрементный краулер (Incremental Crawler) – традиционный вид краулера, который периодически обновляет собранные в своем хранилище документы. Помимо замены старых версий документов на новые, он может обновлять ранг документов, сортируя их на менее важные и более важные.

3. Распределенный краулер (Distributed Crawler) – это тип краулера базирующегося на распределенных вычислениях. Как правило он включает несколько вычислительных узлов, один из которых назначается мастером, а другие дочерними узлами. Это тип краулеров использует алгоритм Page Rank для улучшения релевантности поиска. Достоинство этого вида краулеров в их надежности и отказоустойчивости.

4. Параллельный краулер (Parallel Crawler) – такой краулер состоит из нескольких краулер-процессов, каждый из которых работает над своим выбранным множеством данных.

5. Кросплатформенный краулер (Cross-platform Crawler) – такой краулер должен одинаково устанавливаться и настраиваться на машинах с различными операционными системами описание наиболее популярных краулеров с открытым исходным кодом, описанных в приведенных выше источниках.

6. Nutch (http://nutch.apache.org), тип: инкрементный, параллельный, распределенный, кросплатформенный) — это инструмент поиска в Вебе и краулер, написанный на java, поддерживающий граф связей узлов, различные парсеры, фильтры и нормализаторы URL. Он интегрирован с Apache Lucene, Apache Hadoop 1.X, 2.X и в своей ветке 2.X позволяет использовать

различные хранилища данных, такие как Cassandra, Hbase и др. Данный краулер является масштабируемым (до 100 узлов в кластере и легко настраивается и расширяется, в полной мере является “вежливым”.

7. Scrapy (http://scrapy.org), тип: сфокусированный, параллельный, кросплатформенный) — это расширяемый и гибкий краулер, написанный на Python, который легко устанавливается, прозрачный, поддерживает выгрузку данных в JSON, XML, CSV. Отлично подходит для сфокусированного краулинга (под словом «краулинг» мы понимаем процесс сбора данных в Вебе, состоящий из: навигации на веб-страницы, анализа их ссылок и содержимого). “Вежливый” и устойчивый, расширяемый инструмент.

Он активно развивается сообществом и сейчас. Но данный краулер куда менее производительный нежели Apache Nutch или же Heritrix.

8. Open Search Server (http://www.open-search-server.com, тип: инкрементный, параллельный, кроплатформенный) – это краулер и поисковый движок, ядро которого написано на java, являющийся “вежливым” и поддерживающий современные подходы к полнотекстовому поиску, в частности поддерживает геолокацию. Кроме того, он обеспечивает автоматическую классификацию текстов и поддержку синонимов, поддерживает 18 языков, краулит базы данных, веб и файловые хранилища. Он использует такие технологии как: Apache Lucene, Zkoss, Apache Tomcat и т.д. Это надежный и производительный инструмент.

9. Norconex HTTP Collector

(http://www.norconex.com/product/collector-http, тип: инкрементный, параллельный, кросплатформенный) - позволяет выполнять задачи крулинга и сохранять собранные данные в любое настроенное хранилище в том числе и в поисковый индекс, написан на java, является “вежливым”. Он весьма производительный, поддерживает все востребованные функции: фильтрацию, нормализацию, а также распознавание языков, гибкость извлечения данных и т.д. Краулер активно развивается и поддерживается коммерческим проектом.

10. Bixo (http://openbixo.org, тип: инкрементный, параллельный, распределенный, кросплатформенный) — это инструмент анализа и извлечения данных из веба, написанный на java. Он работает на основе Cascading, позволяет переносить данные в хранилища данных (Data Warehouse, DWH). Расширяемый, настраиваемый, устойчивый и “вежливый”. Bixo ориентирован именно на анализ данных, он масштабируем до 1000 вычислительных узлов и подходит для анализа данных больших объемов.

Разработчики этого инструмента построили его после разработки одного из проектов для вертикального поиска, основанном на Nutch, поэтому они постарались учесть недостатки Nutch для этого типа задач. Данный инструмент рекомендуют использовать для таких задач как: найти и проанализировать комментарии продукта где-либо, отслеживать популярность в социальной сети, проанализировать данные о стоимости продукта и т.д.

11. Crawler4j (https://github.com/yasserg/crawler4j), тип: параллельный, кросплатформенный) — это краулер, написанный на java, с простым API, с его помощью можно легко организовать многопоточный краулинг. Данный инструмент можно легко встроить в проект, но при этом не поддерживает индексирования, является “вежливым”, но может порождать излишнюю нагрузку на исследуемый хост.

12. mnoGoSearch, тип: инкрементный, параллельный платформа - Unix) - это индексатор, написанный на C, управляемый из командной строки, который может быть запущен из под Apache Web Server или любого другого HTTP-сервера поддерживающего стандарт интерфейса связи внешней программы с веб-сервером (Common Gateway Interface, CGI). Этот инструмент может быть проинтегрирован с различными базами, в зависимости от нужд пользователя. Данный проект продолжает развиваться, но никакой информации о расширяемости он не предоставляет.

13. Arachnode.net (http://arachnode.net, тип: инкрементный, параллельный, кросплатформенный) — это краулер, написанный на C# и использующий платформу .NET и SQL Server 2008. Данный краулер является “вежливым” и поддерживает загрузку, индексацию и сохранение интернет-контента, включая адреса e-mail, файлы и веб-страницы. Активно продолжает развиваться, ценится сообществом, но отсутствует в некоторых рейтингах.

14. GNU Wget (http://www.gnu.org/software/wget, тип: сфокусированный, платформа Linux) — это продолжающий развиваться инструмент, написанный на C, который позволяет получать файлы из наиболее широко использующихся интернет-протоколов (HTTP, HTTPS, FTP). Данный инструмент полезен, но подходит в основном для выгрузки конкретных данных или сайта, но не масштабного краулинга сегментов Веба. Приведем еще некоторые краулеры, которые до сих пор в некоторых рейтингах присутствуют и даже упоминаются среди передовых, но которые по факту давно уже не развиваются.

15. Heritrix (http://crawler.archive.org/) — это веб-краулер с открытым исходным кодом, расширяемый, предназначенный для краулинга большого объема данных, «вежливый» по отношению к robots.txt, он собирает данные так, чтобы не нагружать вебсайт чрезмерной нагрузкой. Но несмотря на то, что в западных публикациях данный краулер считается одним из лучших, сейчас он недоступен на территории РФ, т.к. был заблокирован Роско надзором.

16. http://Dig (http://www.htdig.org/) — это веб-краулер и поисковая система, до 2004г. он был одним из самых популярных краулеров и именно с ним сравнивался в свое время Nutch. Но после 2004г. он больше не выпускал релизов и растерял свою аудиторию.

17. Grub (http://grub.org/, https://sourceforge.net/projects/grub/) — это распределенный веб-краулер, позволяющий строить индекс Веба. Grub client выполняет задачи краулинга в период простоя клиентской машины и

отправляет собранные данные на сервер. Проблема этого краулера, согласно обзору Apache Nutch, состоит в том, что распределенный таким образом краулинг не оправдает себя, т.к. он будет генерировать еще больше нагрузки своим трафиком при обработке запросов. После 2013г. Релизов этого проекта больше не было.

18. ASPseek (https://github.com/pld-linux/aspseek) — это интернет-поисковик, состоящий из индексирующего робота, поискового фонового потока и CGI интерфейса. Он может проиндексировать несколько миллионов URLs и поддерживает поиск по словам, фразам и логический поиск. Поиск может быть уточнен по времени или ранжирован с помощью pagerank.

Перечень задач, решаемых на основе краулинга медицинских данных:

Поисковый робот (он же crawler, краулер, паук, бот) — программа для сбора контента в интернете. GoogleBot – самый известный из представленных. Для медицинских исследований краулеры полезны тем, что ускоряют и облегчают процесс исследования, спасая от мануального поиска и скачивания большого объема информации. Но у краулеров есть свое слабое место- многие сайты, включая facebook, не допускают роботов к информации.

Например, исследование «Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups», использовало краулинг трех форумов, для выявления среди пользователей больных эпилепсией.

Выделение медицинских данных является одной из задач выделения знаний из текста в целом. В современных системах используется двухфазная технология аналитической обработки. В первой фазе (ETL) производится автоматизированный анализ отдельных документов, структуризация их контента и формирование хранилищ исходной и аналитической информации. Во второй фазе (Text Mining, Data Mining etc.) — извлечение в оперативном режиме знаний из хранилища или из полученной по запросу подборки документов.

Первичная аналитическая обработка в фазе ETL:

- выполнение индексирования;
- построение семантической сети;
- построение рубрик;
- создание аннотации и ключевых тем;
- терминологические векторы документов;
- хранилище аналитических данных;
- база данных фактографической информации, объединенной в досье.

В ходе аналитической обработки происходит выделение текста фактографической информации об объекте, причем с учетом всех ссылок. Для этого сначала выделяются все предложения с упоминаниями об объекте, в

которых могут встречаться названия объекта, ссылки на него, а также обобщающие определения.

Data mining (рус. интеллектуальный анализ данных) — собирательное название, используемое для обозначения совокупности методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний. Основу методов data mining составляют всевозможные методы классификации, моделирования и прогнозирования, основанные на применении деревьев решений, искусственных нейронных сетей, генетических алгоритмов, эволюционного программирования, ассоциативной памяти, нечёткой логики. К методам data mining нередко относят статистические методы (дескриптивный анализ, корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов, анализ выживаемости, анализ связей).

Интеллектуальный анализ текстов (ИАТ, англ. text mining) — направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении эффективных в практическом плане методов машинного обучения и обработки естественного языка. Название «интеллектуальный анализ текстов» перекликается с понятием «интеллектуальный анализ данных» (ИАД, англ. data mining), что выражает схожесть их целей, подходов к переработке информации и сфер применения; разница проявляется лишь в конечных методах, а также в том, что ИАД имеет дело с хранилищами и базами данных, а не электронными библиотеками и корпусами текстов.

Механизмы применяющиеся, для обеспечения гарантированного качества обработки и интерпретации:

Задача FEATURES SELECTION может быть обобщена путем определения преобразования $Z = F(X)$, позволяющего из X формировать новое пространство признаков Z, $|Z| < |X|$. В такой постановке задача называется задачей FEATURES EXTRACTION (извлечение или конструирование признаков). При решении FEATURES EXTRACTION формируются новые признаки на основе уже имеющихся в X. Самое простое преобразование $Z=F(X)$ – это линейное преобразование.

Основными способами решения задачи FEATURES EXTRACTION являются методы факторного анализа и метод экстремальной группировки признаков. Факторный анализ позволяет выделить обобщенные признаки (факторы), каждый из которых представляет сразу несколько исходных признаков.

Метод накопленных частот Суть МНЧ заключается в следующем. Пусть имеются два набора значений признака $x \in X$, принадлежащие двум матрицам «объект – признак» обучающих выборок A1 и A2. Далее не будем делать различия между выборками и соответствующими им матрицами «объект – признак». По двум наборам значений признака x строятся

эмпирические распределения и подсчитываются накопленные частоты как суммы частот от начального до текущего интервала распределения. Мерой информативности признака x служит модуль максимальной разности накопленных частот.

Метод Шеннона В методе Шеннона в качестве меры информативности признака x рассматривается средневзвешенное количество информации, которое свойственно анализируемому признаку.

Метод Кульбака В данном методе в качестве меры информативности признака x рассматривается величина, называемая дивергенцией Кульбака и отражающая расхождение между выборками A1 и A2.

Вторая (практическая) часть

В качестве научной статьи в данной практической работе использован источник:
https://academic.oup.com/bioinformatics/article/30/1/104/236067.

Краткое описание данной статьи: при использовании современных поисковиков выдача по определенному поисковому запросу составляет тысячи, а то и миллионы страниц. Авторы же реализовали систему, способную при определенном медицинском запросе выдавать меньший по количеству, но более качественный набор веб-страниц. Система включает в себя модель, оценивающую релевантность страницы к заданному пользователем запросу и краулер, собирающий страницы для оценки.

Таким образом, согласованная с преподавателем конечная работа: реализовать модель оценки веб-страницы.

Модель $\Phi(wp, \Omega)$, получающая на вход wp – веб-страницу, Ω – тему, и выдающая в ответ значение принадлежности страницы к теме в диапазоне от 0 до 1.

Выбранная тема – рак печени.

Ход выполнения:

1 этап: сбор веб-страниц форумов

Сперва были собраны исходные данные в виде 20 веб-страниц, являющиеся форумами. Они были найдены по простым поисковым запросам в Яндекс и Google, по типу «рак печени форум», «обсуждение рака печени» и т.д.

Список страниц с личными оценками:

```
[
('https://www.oncoforum.ru/forum/showthread.php?t=107634', 2),
('http://www.rakpobedim.ru/forum/index.php?/topic/1592-рак-печени/', 8),
('http://www.woman.ru/kids/healthy/thread/4714336/', 7),
('https://khabmama.ru/forum/viewtopic.php?t=135141', 1),
('https://www.e1.ru/talk/forum/read.php?f=36&t=483540', 3),
```

('https://forum.guns.ru/forummessage/80/2111885.html', 6),
('https://samaraonko.ru/forums/viewtopic.php?id=16880', 5),
('https://forum.ngs.ru/board/health/flat/1873376851/?fpart=1&per-page=50', 7),
('https://forums.rusmedserv.com/showthread.php?t=200647', 6),
('http://golodanie.su/forum/showthread.php?t=11761', 6),
('https://forums.playground.ru/talk/rak_pecheni_x-128505/', 3),
'https://www.nn.ru/community/biz/medicine/proshu_pomoshchi_i_sovetov_u_papy_rak_metastazy.html', 5),
('http://www.onconet.ru/showthread.php?p=9318', 3),
('http://crimea-med.net/index.php?/topic/3759-rak-pecheni/', 4),
('https://eva.ru/kids/messages-3240292.htm', 5),
('http://forumjizni.ru/archive/index.php/t-9567.html', 5),
('https://www.u-mama.ru/forum/family/health/788648/', 5),
('http://rak.flyboard.ru/topic2210.html', 2),
('https://forum.ykt.ru/viewmsg.jsp?id=24732435', 7),
]

Каждая страница была субъективна оценена в работе по шкале от 0 до 10. В дальнейшем это значение делилось на 10. Оценивалась степень информативности ответов людей на тему, полнота ответа.

На втором этапе был произведен сбор информации с веб-страницы. Разработка велась на языке программирования Python. Как предлагалось в статье, с веб-страницы необходимо собрать три набора данных:

T1 – основной текст страницы.

T2 – название заголовков в html-тэгах h1, h2, h3, h4, h5, h6.

T3 – название тэгов, содержащих в себе ссылку на другие страницы.

Каждый набор был собран при помощи библиотеки BeautifulSoup4.

```
def getTextes(url):
    # загружаем страницу
    r = requests.get(url)
    html_text = r.text
    soup = BeautifulSoup(html_text)

    # удаляем лишний текст на странице (js-скрипты и стили)
    for script in soup(["script", "style"]):
        script.extract()

    ### достаем текст файла
    text = soup.get_text()
    lines = (line.strip() for line in text.splitlines())
    chunks = (phrase.strip() for line in lines for phrase in line.split("  "))
    textT1 = '\n'.join(chunk for chunk in chunks if chunk)

    ### достаем все заголовки для T2
    h_list = ['h1', 'h2', 'h3', 'h4', 'h5', 'h6']
    textT2 = []

    for hn in h_list:
        hn_list = soup.find_all(hn)
        for h in hn_list:
            textT2.append(h.text)

    ### Достаем все текста с url ну и пусть url тоже будет храниться
    textT3 = [i.text for i in soup.find_all('a', href=True)]
    textT3href = [i['href'] for i in soup.find_all('a', href=True)]

    return (textT1, textT2, textT3)
```

Третий этап включал в себя выделение ключевых слов. Авторами предлагается метод RAKE для выделения из текста ключевых слов и фраз. Для этого использовалась библиотека multi-rake позволяющая работать с множеством языков, в том числе и русским. При дальнейшем отборе ключевых слов авторы предлагают убрать все слова, получившие оценку RAKE ниже определенной.

Посмотрев вручную, ниже оценки 1.1 собирается много различных слов, большая часть из которых бесполезны для темы медицины.

```
def getKeywords(texts):
    # Получаем список ключевых слов из каждого текста
    keywordsT1 = rake.apply(texts[0])
    keywordsT2 = rake.apply(''.join(texts[1]))
    keywordsT3 = rake.apply(''.join(texts[2]))

    # Производим отбор самых важных из них
    T = 1.1
    keywordsT1 = [x[0] for x in keywordsT1 if x[1] >= T]
    keywordsT2 = [x[0] for x in keywordsT2 if x[1] >= T]
    keywordsT3 = [x[0] for x in keywordsT3 if x[1] >= T]
```

Здесь показано, что при значении ниже 1.1 (список дальше идет еще на 100 слов) меньше подходящих слов, чем выше это значение.

```
('врач', 1.5),
('уберечь', 1.5),
('прошло', 1.5),
('извините', 1.5),
('лежит', 1.5),
('ситуации', 1.5),
('слабость', 1.5),
('давление', 1.5),
('serenasandra', 1.5),
('nembutal', 1.5),
('juanpablohernandez', 1.5),
('n3rgy', 1.5),
('лечению', 1.5),
('oncoforum', 1.5),
('использована', 1.5),
('онкофорум', 1.5),
('болей', 1.3333333333333333),
('свяжитесь', 1.3333333333333333),
('com', 1.3333333333333333),
('печени', 1.1428571428571428),
('2017 сообщений', 1.0833333333333333),
('2016 сообщений', 1.0833333333333333),
('2018 сообщений', 1.0833333333333333),
('ракеназад', 1.0),
('близкими', 1.0),
('флакон', 1.0),
('5 шт', 1.0),
('juli25', 1.0),
('2 поблагодарил', 1.0),
('2 поблагодарили 0', 1.0),
('столкнулся', 1.0),
```

Следующим этапом отбора слов авторы предлагают каждому ключевому слову выдать значение по следующей формуле.

$$\psi(kw_{i,j}^{k}, \Omega, \tau, \mathbf{wp}) = \frac{p_{11}(kw_{i,j}^{k}, \mathbf{wp})p_{00}(kw_{i,j}^{k}, \mathbf{wp})}{p_{01}(kw_{i,j}^{k}, \mathbf{wp})p_{10}(kw_{i,j}^{k}, \mathbf{wp})},$$

Числа p имеют два индекса принимающие значения 1 или 0. Число p является количеством страниц включающих (или не включающих) в себя конкретное ключевое слово среди страниц имеющих оценку больше (или меньше) знания τ (тау).

Опишем алгоритм:

Берется отдельно каждое ключевое слово в каждой веб-странице и проверяется его нахождение в каждой другой имеющейся странице. Если же первый индекс равен 1, то проверяется наличие ключевого слова в наборе страницы, при 0 проверяется его отсутствие. Второй индекс при 1 включает только те страницы, чья оценка выше тау. При 0 ниже тау.

$$p_{11}(kw_{i,j}^k, \mathbf{wp}) = |\{wp_x \in \mathbf{wp} | kw_{i,j}^k \in wp_x, \hat{\Phi}(wp_x, \Omega) \geq \tau\}|;$$

$$p_{01}(kw_{i,j}^k, \mathbf{wp}) = |\{wp_x \in \mathbf{wp} | kw_{i,j}^k \notin wp_x, \hat{\Phi}(wp_x, \Omega) \geq \tau\}|$$

$$p_{00}(kw_{i,j}^k, \mathbf{wp}) = |\{wp_x \in \mathbf{wp} | kw_{i,j}^k \notin wp_x, \hat{\Phi}(wp_x, \Omega) < \tau\}|$$

$$p_{10}(kw_{i,j}^k, \mathbf{wp}) = |\{wp_x \in \mathbf{wp} | kw_{i,j}^k \in wp_x, \hat{\Phi}(wp_x, \Omega) < \tau\}|$$

На четвертом этапе проводится обучение модели. Модель в статье основана на полиноминальной регрессии. Для определения подходящей степени полинома отыщем значения квадрата ошибки при построении полиномов различной степени. На рис. 5.1. видно, что при переходе от полинома 14 степени к полиному 15 степени происходит огромный скачок порядков. Поэтому выберем 14 степень полинома. Модель обучена.

p	E
2	0,4272
3	0,3743
4	0,3646
5	0,3645
6	0,3606
7	0,2907
8	0,2293
9	0,2197
10	0,2162
11	0,2103
12	0,1573
13	0,1296
14	0,066
15	3E-07
16	9E-09

Рис. 5.1. График полиноминальной регрессии

На пятом этапе проводилась проверка контрольной выборки. Полная выборка из 20 веб-страниц была разделена на обучающую и контрольную. Обучающая выборка составляла 80% процентов всей выборки. Эксперимент проводился несколько раз, каждый раз изменяя набор каждой выборки, не повторяющий предыдущий. Новые выборки определяются так: сдвигаются на 1 пункт списка.

Каждый раз записывался результат оценки качества проверки. Сравнивались известное значение оценки страницы с тем, что выдала модель. Процент точности высчитывался следующим образом:

(модуль расстояния между оценками)/(максимальное расстояние до истинного значения оценки).

Для каждого эксперимента выводился средний процент ошибки.

Таблица 5.1. Процент ошибки

№	1	2	3	4	5	6	7	8	9	10
процент	28,71	45,00	12,09	25,53	28,33	51,32	32,11	27,07	51,04	39,06

Как видно из таблицы, ошибка может достигать 12%, но может и превышать 50%. Чаще всего она варьируется около 25-30%.

Таким образом, получили обучающуюся модель, которая по определенному запросу и начальной ранжированной выборке, дает оценку другим страницам по данной теме.

Недостатки данного эксперимента:

- Малая выборка данных. Для хороших экспериментов необходима выборка в 50 веб-страниц, а то и 100 и более.
- Субъективная оценка страниц «эксперта», не имеющего достаточных знаний в отрасли. Для повышения качества, необходимы квалифицированные «эксперты», и не один, а несколько. Учитывать оценку каждого. А также увеличить диапазон оценки «от 0 до 100».

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Что такое краулинг?
2. Для чего необходимо сокращать признаковое пространство?
3. Какие задачи можно решать при анализе медицинских текстов?
4. Какие метрики используются для оценки качества интерпретации данных?

Литература

1. Советов Б.Я. Интеллектуальные системы и технологии: учебник для студ. Учреждений высш. проф. Образования /Б.Я. Советов, В.В. Цехановский, В.Д. Чертовской. — М.: Издательский центр «Академия», 2013. — 320 с.

2. Остроух А.В. Интеллектуальные системы / А.В. Остроух. – Красноярск: Научно-инновационный центр, 2015. – 110 с.

3. Воронцов К.В. Курс лекций Машинное обучение Электронный ресурс www.machinelearning.ru/wiki/index.php?title =Машинное_обучение [дата обращения - 02.10.2019].

4. http://technic.itizdat.ru/Uploads/vidonskoy/FIL13651001400N977746001/.pdf (она же http://intellectualarchive.com/getfile.php?file=KbEp3gZPNPX&orig_file=Donskoy%20VI%20Splitting %20criteria.pdf).

5. Шайдуров А. Нейросетевой анализ медицинских данных; LAP Lambert Academic Publishing - M., 2012. - 140 с.

6. Масленникова, О. Е. Основы искусственного интеллекта : учеб. пособие / И. В. Гаврилова, О. Е. Масленникова . – 2-е изд., стер. – М. : ФЛИНТА, 2013 .— 284 с. : ил. — ISBN 978-5-9765-1602-1 – 284 с.

7. Психология применения интеллектуальных систем в гуманитарной сфере : учебное пособие / Г. О. Артемова, Н. Ф. Гусарова ; М-во образования и науки РФ, СПбНИУ ИТМО, Каф. ИТГС. — СПб. : НИУ ИТМО, 2012. — 140 с.

8. Воронцов К.В. Курс Лекции по логическим алгоритмам классификации 21 декабря 2007 г. Электронный ресурс http://www.ccas.ru/voron/download/LogicAlgs.pdf#2 [дата обращения - 02.10.2019].

9. Баранов А.А, Намазова-Баранова Л.С. и др. Методы и средства комплексного интеллектуального анализа медицинских данных, Труды ИСА РАН. Том 65. 2/2015 – С. 81-93.

10. Печников А.А., Сотенко Е.М. Программы-краулеры для сбора данных о представительских сайтах заданной предметной области – аналитический обзор // Современные наукоемкие технологии. – 2017. – № 2. – С. 58-62.

Гусарова Наталия Федоровна

Добренко Наталья Викторовна

ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

Учебно-методическое пособие

В авторской редакции

Редакционно-издательский отдел Университета ИТМО

Зав. РИО Н.Ф. Гусарова

Подписано к печати

Заказ № 3311

Тираж 100

Отпечатано на ризографе

Редакционно-издательский отдел

Университета ИТМО

197101, Санкт-Петербург, Кронверкский пр., 49