



САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



Методы «раскопки данных» - Data Mining

Авторы

Тишков Артем Валерьевич

Эюбова Наргиз Идаят кызы

Делакова Екатерина Александровна

Семенова Елена Михайловна

2013



САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



Медицинские данные

*Результаты медико-биологических исследований –
большое количество данных различного характера*

- Результаты лабораторных исследований;
- Социально-паспортные и антропометрические данные;
- Факторы риска;
- Данные медицинских приборно-компьютерных систем.



САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



Анализ медицинских данных

- Статистические методы
- Методы, основанные на знаниях
 - «Раскопка данных» (Data Mining)
 - Экспертные системы

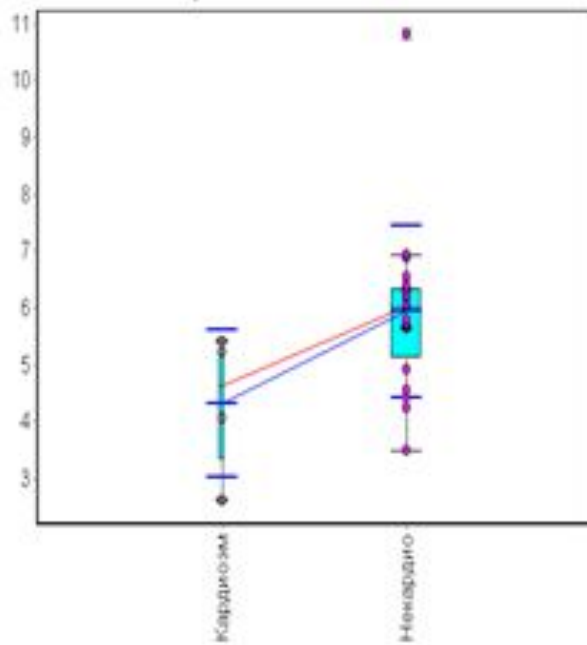
Data Mining «Раскопка данных»

- поиск (неочевидных) закономерностей в данных
- обнаружение скрытых знаний



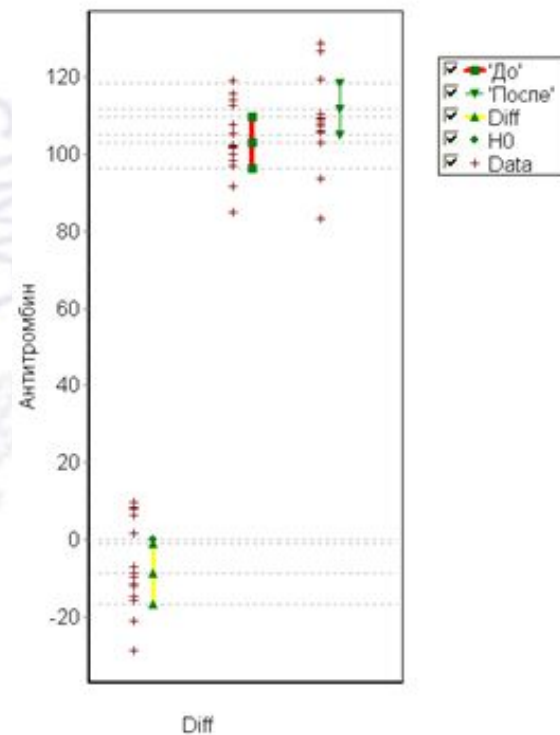
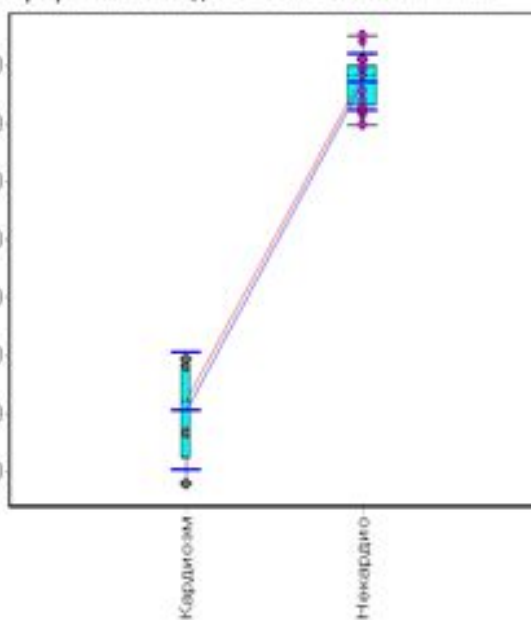
Статистические методы

Холестерин по Патогенетическим типам



5

Протромбиновый индекс 2 по Патогенетическим типам



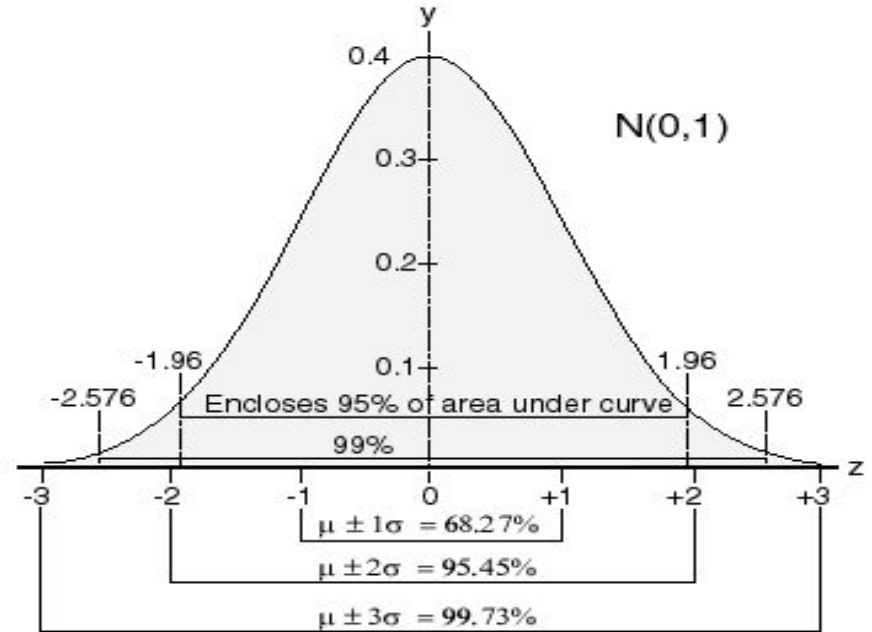
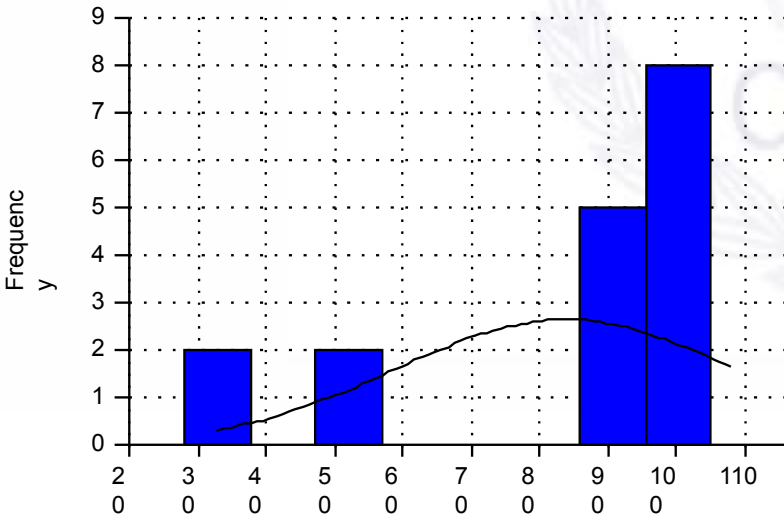
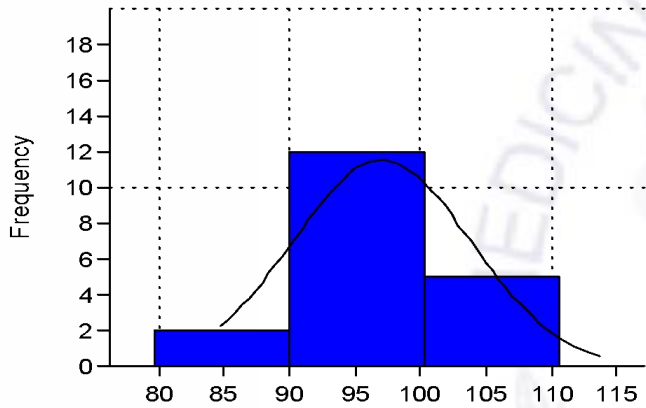


САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



Согласованность с нормальным законом распределения





Корреляционный анализ

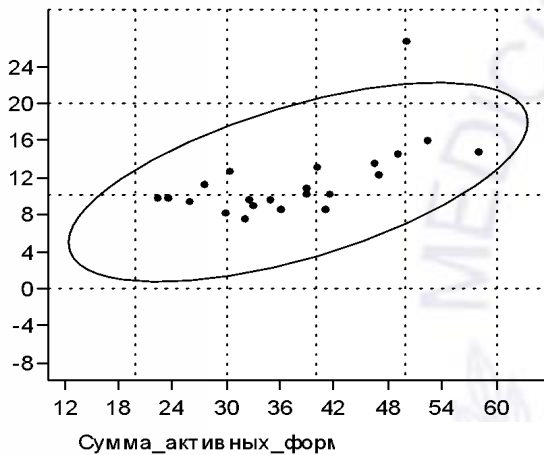
r-коэффициент
корреляции

$$r_n = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

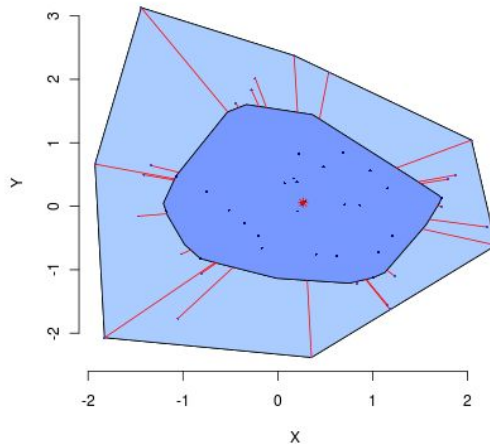
Pearson

ранговые:
 ρ Spearman
 τ Kendall

Число_тромбоцитов_вовлеченных_в_агрегаты_1



Bag Plot





САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА



Кафедра физики, математики и информатики

Гармонизированный анализ

	Протромбиновое время 2		МНО 2		Протромбиновый индекс 2	
	Среднее	N	Среднее	N	Среднее	N
Кардиоэмболический тип	15,19; 27,42 ; 39,65	4	1,44; 2,59 ; 3,73	4	24,24; 40,41 ; 57,57	4
Некардиоэмболический тип	10,52; 10,85 ; 11,18	13	1,00; 1,03 ; 1,07	13	94,03; 96,96 ; 99,89	13
Критерий Стьюдента	8,34		8,37		15,74	
p-значение двустороннее p_{2s}	5,08E-07		4,90E-07		9,80E-11	
Размер эффекта $ES = z \cdot (1 - \alpha) \times 95\%$	2,64; 4,53 ; 6,43		1,17; 1,56 ; 1,94		48,9; 56,55 ; 64,2	
Бейзов фактор BF_{JZS}	4.581506e-05		4.401148e-05		1.870535e-08	
Мощность критерия	0,99		0,83		1,00	
Размер выборки $N (\alpha = 0,01; 1 - \beta = 0,95)$	10		34		8	
Предсказательный интервал для размера эффекта	[2,87; 6,17]		[0,98; 2,12]		[45,72; 67,38]	
Предсказательная вероятность для последующих экспериментов	0,99; 0,99; 1,00		0,99; 0,99; 1,00		1,00; 1,00; 1,00	
Критерий Уэлча	4,31		4,32		10,76	
p-значение двустороннее	0,02		0,02		0,00091	
Размер эффекта по Козну	4,77		4,70		9,01	
Разница дисперсий SD	3,47		0,32		6,28	



Нестатистические методы: «раскопка данных»

эотакс И Н	эотакс И Н - 2	интерлейки н-8	MIP-1alp h a	MIP-1beta a	RANTE S	CCR1	CCR3	CCR5	CXCR1	CXCR2	resu
6,7	13	113	13	33	23	126	99	63	113	86	y
5,1	9,8	89	29	15	23	102	27	32	49	74	y
27	11,4	48	73	36	61	116	118	118	114	85	y
9	20	44	14	11	52	111	102	101	101	84	y
7	3	16	49	49	61	79	58	39	109	111	y
13	10	11	28	10	15	100	91	102	76	66	y
1,2	0	1,2	0,8	1,6	1,4	140	96	158	96	122	n
1	0	0,8	0,1	0,06	10	165	88	66	1,8	73	n
0	3,2	26	0	0,1	0	94	104	74	27	54	n
0,5	1	7	0	4	18	146	81	112	27	90	n
1,5	4	10	0,7	1	43	145	103	127	36	103	n
1	20	33	0	0	98	152	122	140	50	103	n
3	0	17	0	0	39	128	103	18	35	74	n
0,4	2	29	0,3	0,06	32	112	98	96	36	79	n
0,3	0,6	34	0,6	0,3	6	108	112	92	34	98	n
0,3	2	9	0,3	0,2	1	121	114	105	8	120	n



САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



Кластеризация (обучение без учителя)

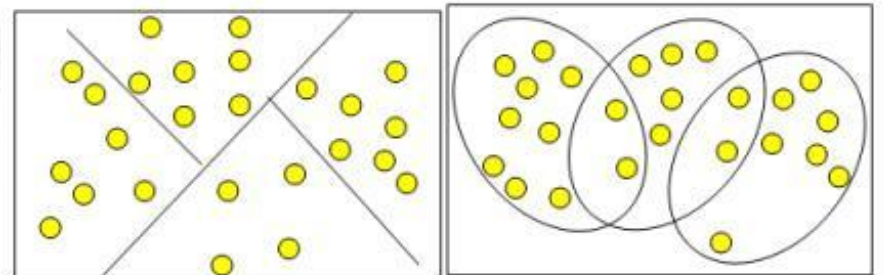
Кластеризация предназначена для разбиения совокупности объектов на однородные группы — *кластеры*.

Цель *кластеризации* — построить оптимальное разбиение объектов на группы: разбить N объектов на k кластеров.

Характеристиками *кластера* можно назвать два признака:

- внутренняя однородность;
- внешняя изолированность.

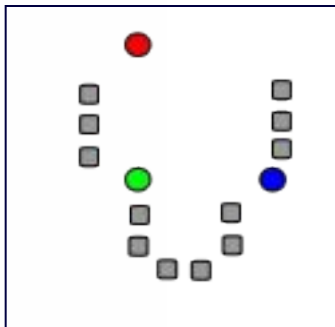
Непересекающиеся и пересекающиеся кластеры



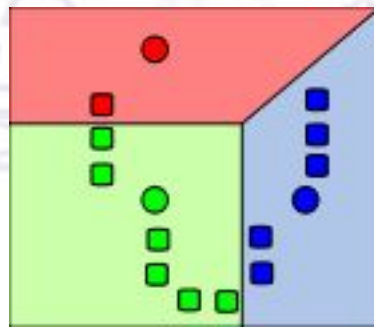


Кластеризация. K-means

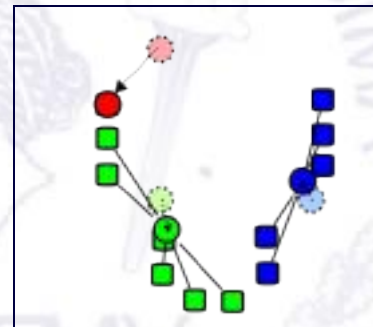
- Разделить образцы на k групп (классов) автоматически, без информации о настоящем классе образца



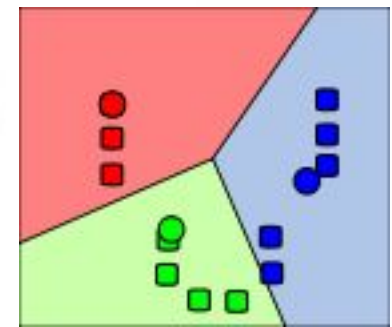
1. Выбрать начальное положение центров классов



2. Сгруппировать образцы по принципу близости к центрам



3. Вычислить новые положения центров



4. Повторить шаги 2 и 3 до схождения алгоритма



Классификация (обучение с учителем)

- Цель классификации:

отнести имеющиеся статические образцы (например, данные медосмотра) к определенному классу (например, диагнозу).

Методы:

- Классификатор Байеса
- Дерево решений
- Нейронная сеть
- Метод к ближайших соседей



САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА



Кафедра физики, математики и информатики

Классификация

- 25 пациентов, перенесших ишемический инсульт; 44 показателя
- Факторы риска
 - ишемическая болезнь сердца
 - артериальная гипертензия
 - сахарный диабет
 - курение
 - ...
- Классифицирующий признак: патогенетический тип инсульта
 - кардиоэмболический
 - некардиоэмболический (атеротромботический, лакунарный, криптогенный, гемореологический)
- Другие признаки
 - применяемые препараты
 - шкала NIHSS (**National Institutes of Health Stroke Scale**)



Наивный классификатор Байеса

- Классификатор Байеса—вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости.
- Достоинством данного классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации.

accuracy: 93.33% +/- 13.33% (micro: 92.00%)

	true Некардиозмболический	true Кардиозмболический	class precision
pred. Некардиозмболический	20	2	90.91%
pred. Кардиозмболический	0	3	100.00%
class recall	100.00%	60.00%	



Наивный классификатор Байеса



Формула Байеса для **совместной вероятности**

$$p(C|F_1, \dots, F_n) = \frac{p(C) p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}$$

$$\begin{aligned} p(C, F_1, \dots, F_n) &= p(C) p(F_1, \dots, F_n|C) \\ &= p(C) p(F_1|C) p(F_2, \dots, F_n|C, F_1) \\ &= p(C) p(F_1|C) p(F_2|C, F_1) p(F_3, \dots, F_n|C, F_1, F_2) \\ &= p(C) p(F_1|C) p(F_2|C, F_1) p(F_3|C, F_1, F_2) p(F_4, \dots, F_n|C, F_1, F_2, F_3) \end{aligned}$$

Наивное предположение: **свойства F_i и F_j условно независимы**

$$p(F_i|C, F_j) = p(F_i|C)$$

И тогда

$$p(C, F_1, \dots, F_n) = p(C) p(F_1|C) p(F_2|C) p(F_3|C) \cdots = p(C) \prod_{i=1}^n p(F_i|C)$$



Нейронные сети

При обучении сети предлагаются различные образцы образов с указанием того, к какому классу они относятся. Образец, как правило, представляется как вектор значений признаков. При этом совокупность всех признаков должна однозначно определять класс, к которому относится образец

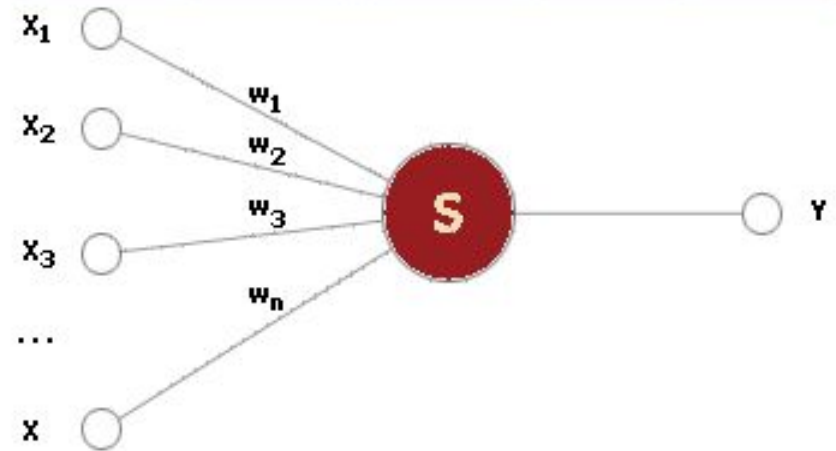


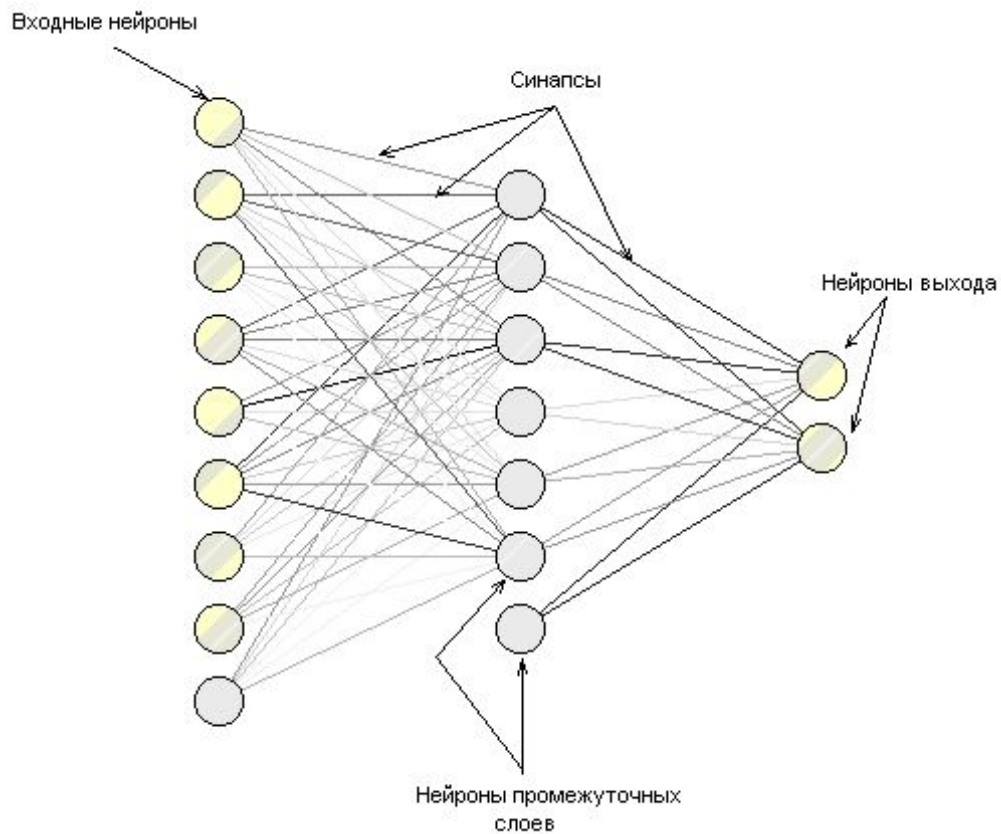
Table View Plot View

accuracy: 93.33% +/- 13.33% (micro: 92.00%)

	true Некардиоэмболический	true Кардиоэмболический	class precision
pred. Некардиоэмболический	19	1	95.00%
pred. Кардиоэмболический	1	4	80.00%
class recall	95.00%	80.00%	



Нейронные сети



Чем сильнее связь между нейронами тем более четкой линией она отображается, чем слабее — тем линия прозрачнее

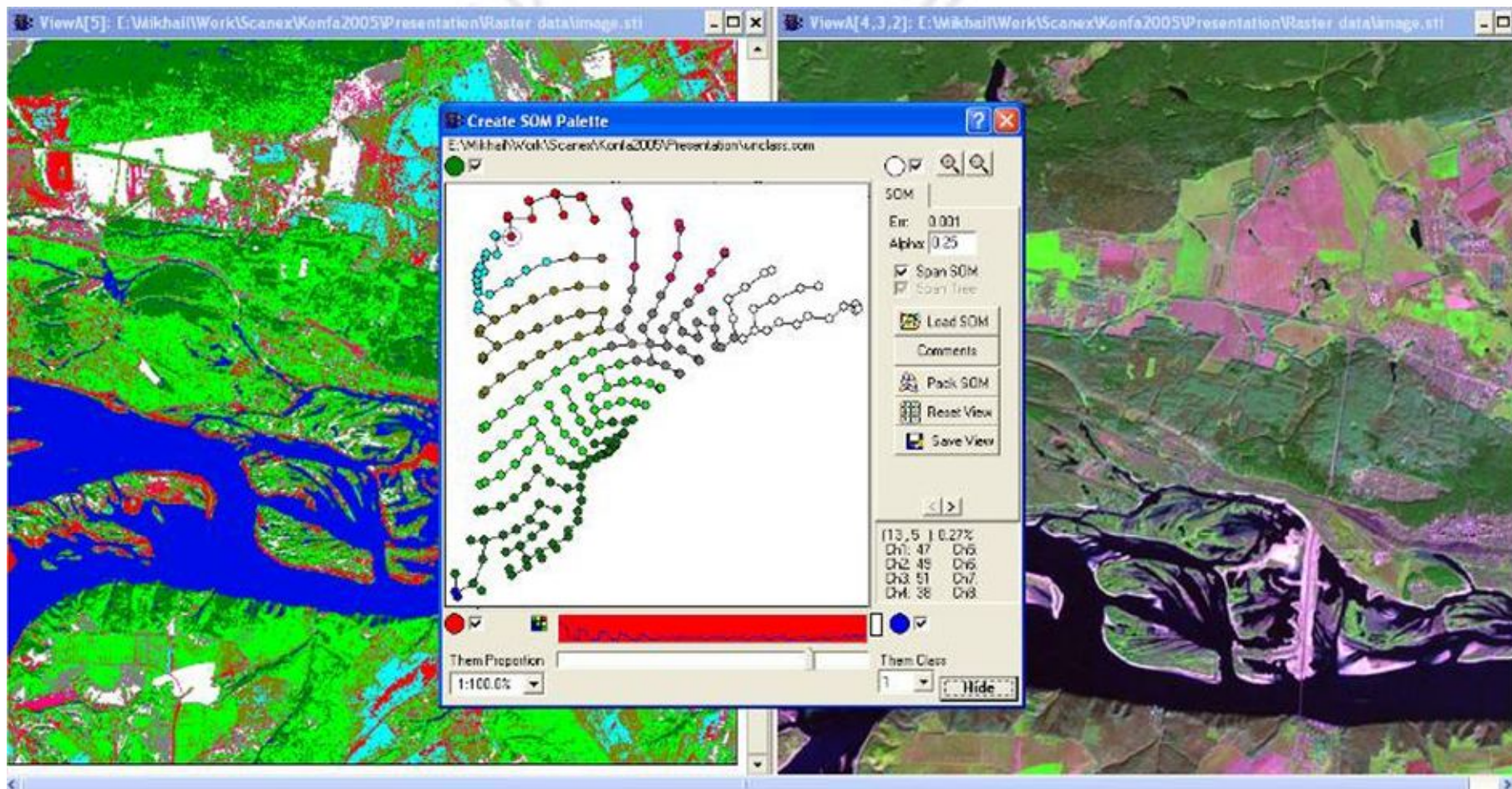


САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



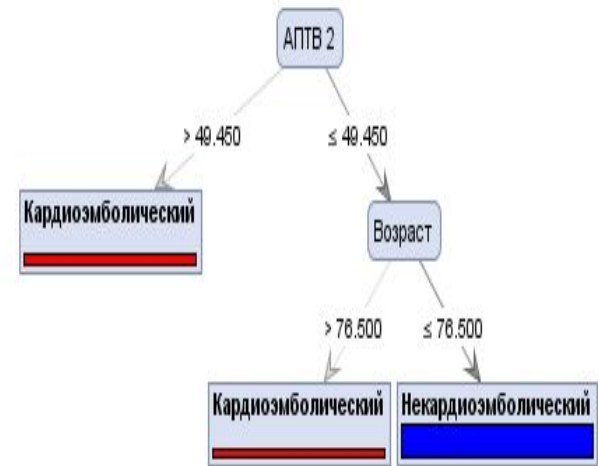
Нейронные сети, изучение космических снимков





Дерево решений

Деревья принятия решений- это дерево, на ребрах которого записаны атрибуты, от которых зависит целевая функция, в листьях записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи.



accuracy: 88.33% +/- 18.33% (micro: 88.00%)

	true Некардиоэмболический	true Кардиоэмболический	class precision
pred. Некардиоэмболический	20	3	86.96%
pred. Кардиоэмболический	0	2	100.00%
class recall	100.00%	40.00%	



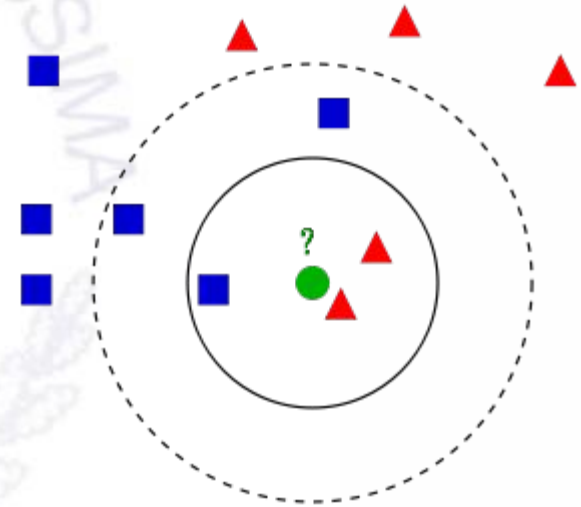
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики

Метод к ближайших соседей

Метод к ближайших соседей ([англ.](#) *k-nearest neighbor algorithm*, kNN) - метод автоматической [классификации](#) объектов. Основным принципом **метода ближайших соседей** является то, что объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента.

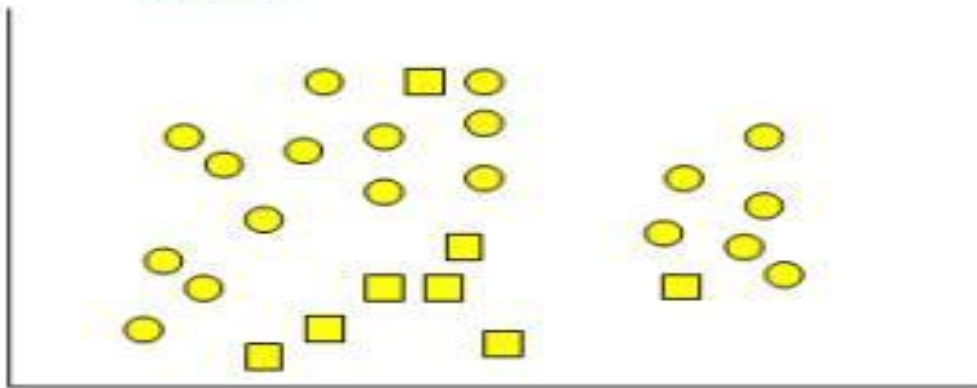
Соседи берутся исходя из множества объектов, классы которых уже известны, и, исходя из ключевого для данного метода значения k высчитывается, какой класс наиболее многочислен среди них.



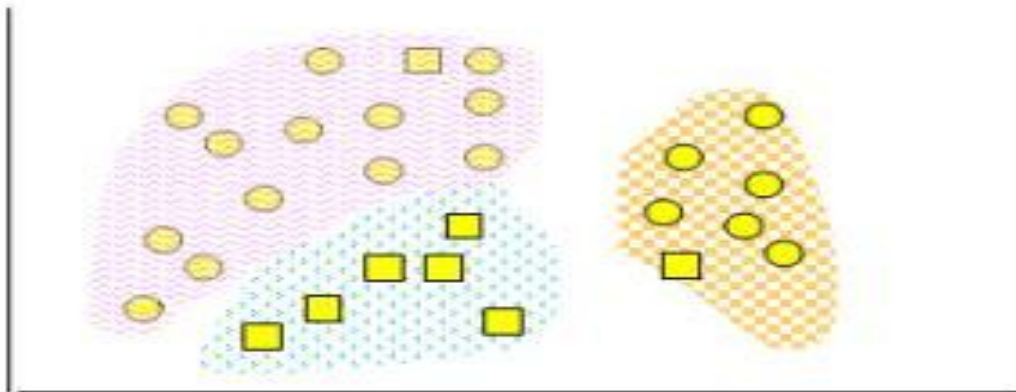
Википедия
Свободная энциклопедия



Сравнение классификации и кластеризации



*Классификация: классы
предопределены
изначально*



*Кластеризация: классы
не предопределены,
осуществляется поиск
наиболее похожих,
однородных групп*





САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА



Кафедра физики, математики и информатики

Бесплатный Data Miner: RapidMiner

The screenshot displays the RapidMiner software interface. The main workspace shows a workflow with three operators: 'Read Excel', 'Set Role', and 'Validation'. The 'Validation' operator is highlighted, and its parameters are visible on the right. The 'Problems' panel at the bottom indicates '3 potential problems'.

Process

- Read Excel
- Set Role
- Validation

Parameters

- average performar
- leave one out
- number o... 10
- sampling ... strati...
- use local random s
- parallelize training
- parallelize testing

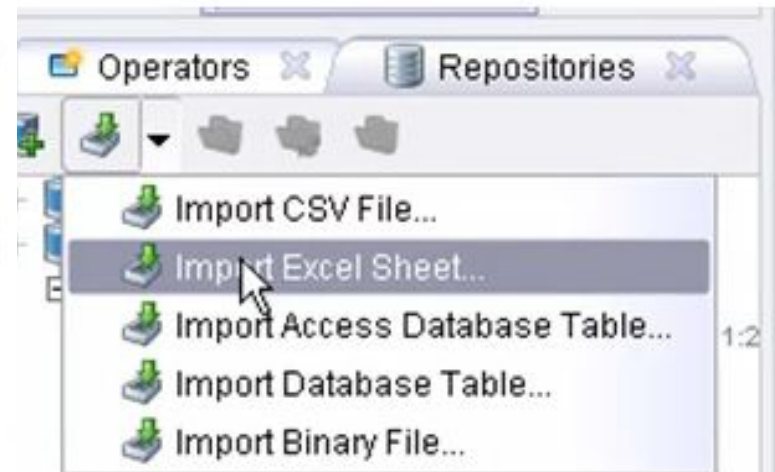
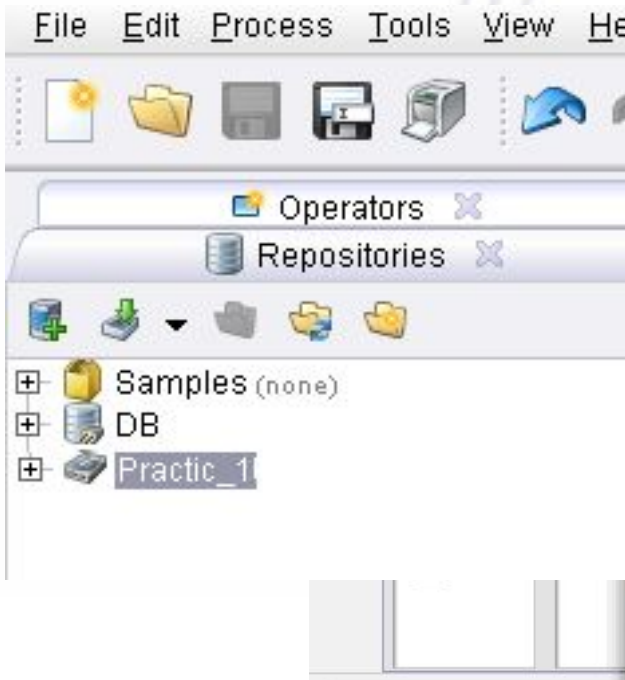
Problems

3 potential problems

Message	Fixes	Location
The attribute 'Res' is missing...	Change valu...	Set Rol...
Input example set must have	Select an att	Decisin



Репозиторий и загрузка данных



Data import wizard - Step 3 of 5

This wizard guides you to import your data.
Step 3: In RapidMiner, each attribute can be annotated. The most important annotation of an attribute is its name. In your data file, you can assign them here.

Annotation	A	B	C	D	E	F	G	H
Name	№	Диагноз	холодовая г	отн. плотнос	бактерии	Ph	Осмолярнос	Фосфор нес
-	1	МКБ	0	1010	0	5	525	1.39
-	2	МКБ	0	1018	2	5	618	1.02
-	3	МКБ	1	1022	0	5	828	0.81
-	4	МКБ	1	1018	2	5	906	0.83



Определение видов и типов данных

Data import wizard - Step 4 of 5

This wizard guides you to import your data.

Step 4: RapidMiner uses strongly typed attributes. In this step, you can define the data types of your attributes, defining what they can be used for by the individual operators. These roles can be also used entirely.

Preview uses only first 100 rows.

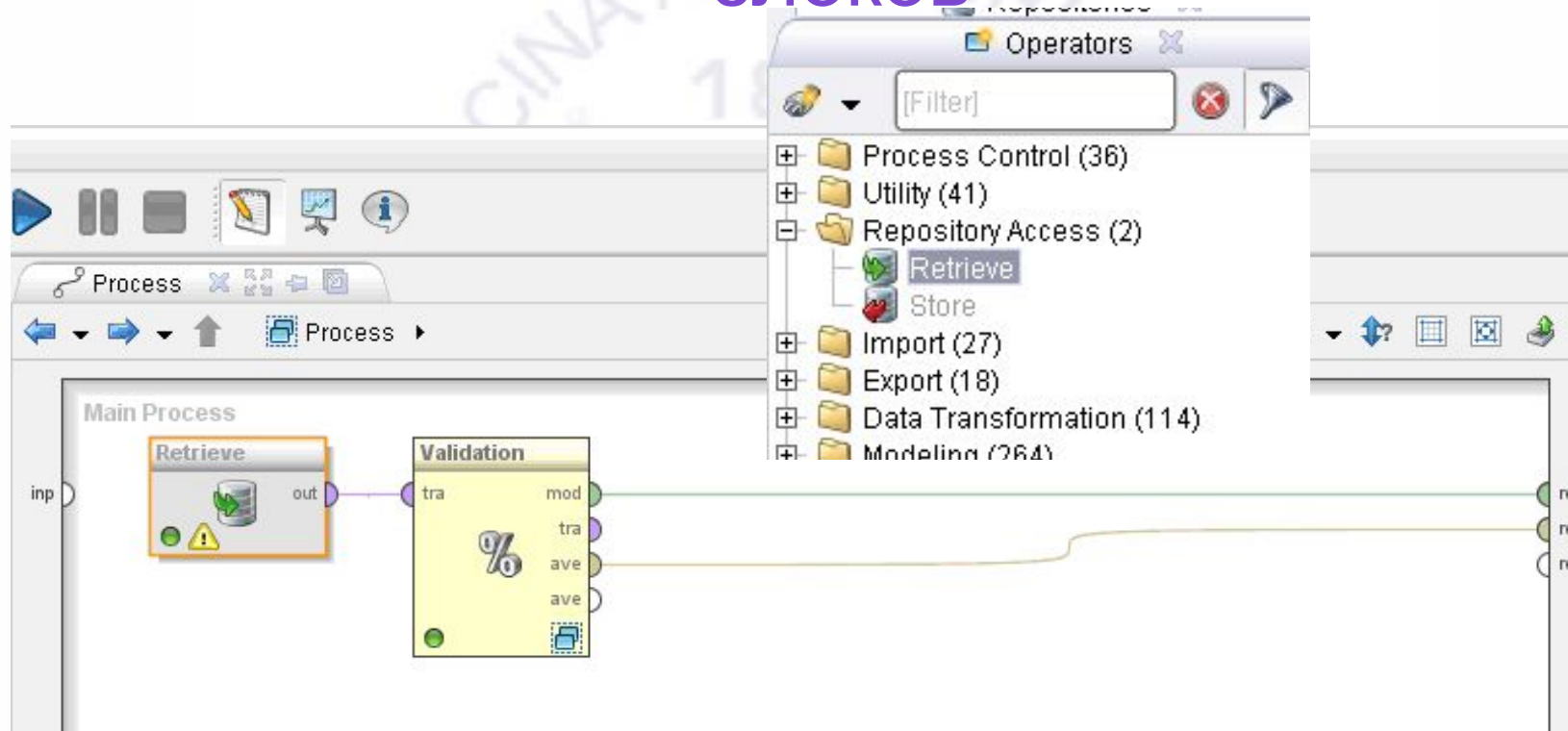
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
№	Диагноз	холодовая г	отн. плотно	бактерии	Ph	чи(500-900)	Фосфор не
integer	polyno...	nominal	integer	nominal	real	integer	real
id	label	attribute	attribute	attribute	attribute	attribute	attribute
1	МКБ	0	1010	0	5	525	1.390
2	МКБ	0	1018	2	5	618	1.020
3	МКБ	1	1022	0	5	828	0.810
4	МКБ	1	1018	2	5	906	0.830
5	МКБ	0	1010	1	5	472	1.170



САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

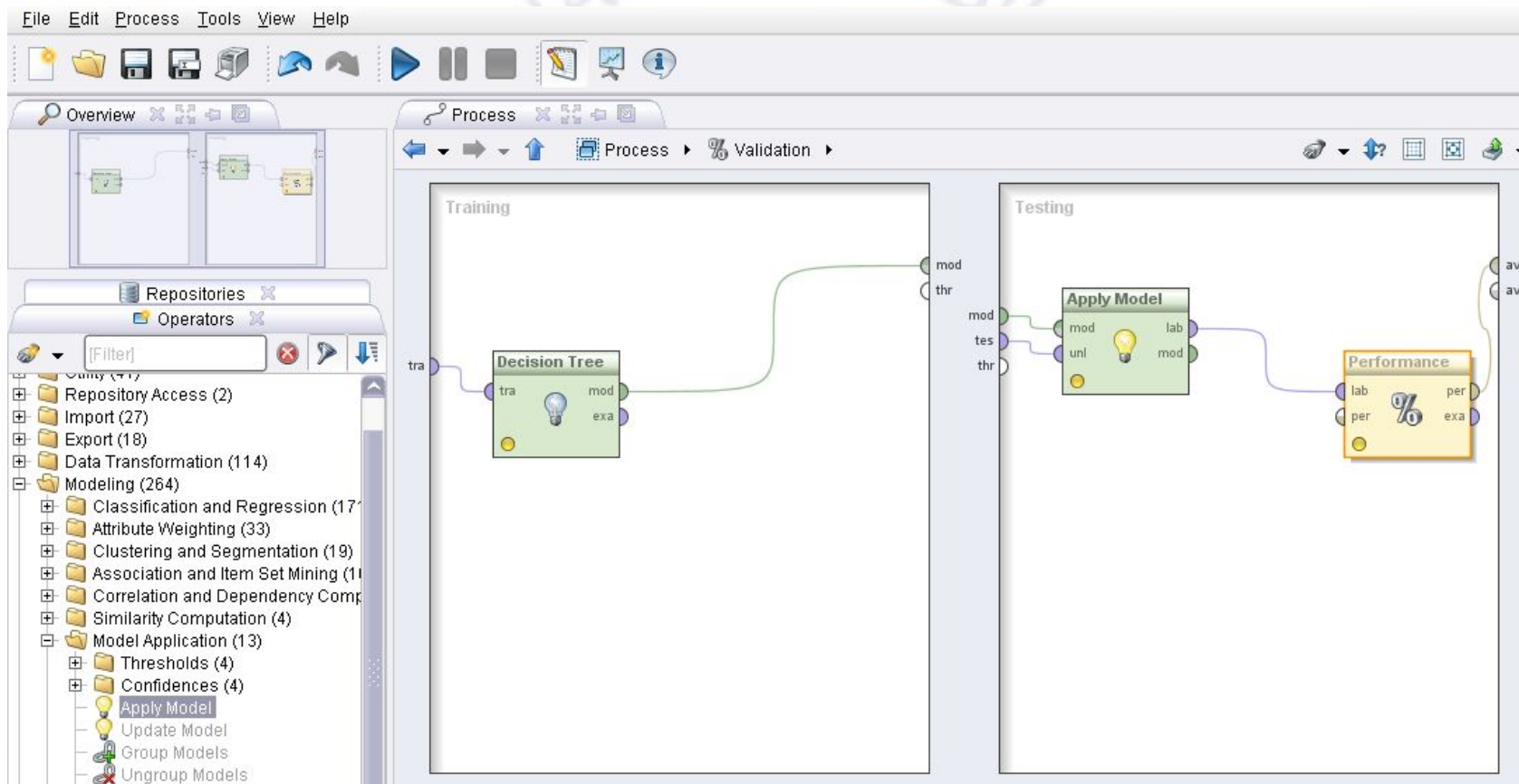
Кафедра физики, математики и информатики

Создание алгоритма анализа данных из блоков





Построение и X-валидация





Результат – точность классификации

Table / Plot View Text View Annotations

Criterion Selector

- accuracy
- kappa

Multiclass Classification Performance Annotations

Table View Plot View

accuracy: 61.67% +/- 23.63% (mikro: 62.07%)

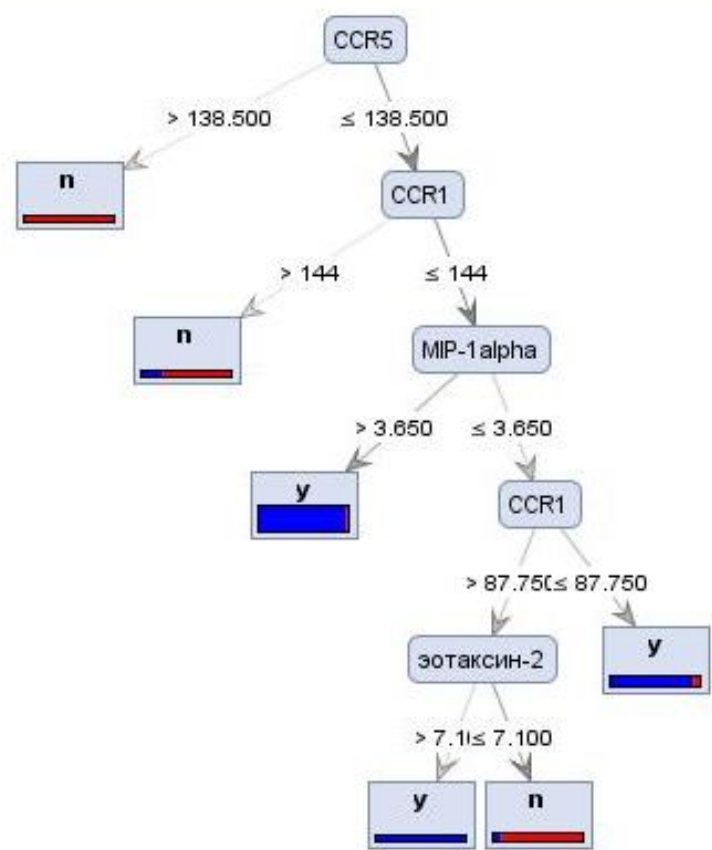
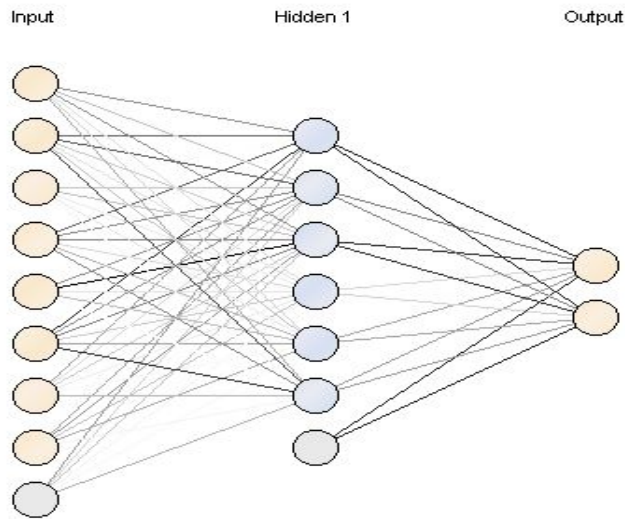
	true МКБ	true варикоцеле	true пиелонефрит	class precision
pred. МКБ	15	6	1	68.18%
pred. варикоцеле	3	3	1	42.86%
pred. пиелонефрит	0	0	0	0.00%
class recall	83.33%	33.33%	0.00%	



Результат запуска: построенный классификатор

	A	B	C	D	E	F	G	H	I	J	K	L
1	accuracy: 82.64% +/- 8.25% (mikro: 82.76%)											
2												
3		true y	true n	class precision								
4	pred. y	61	6	91.04%								
5	pred. n	9	11	55.00%								
6	class recall	87.14%	64.71%									
7												
8												
9												
10	Input	Hidden 1			Output							

11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												



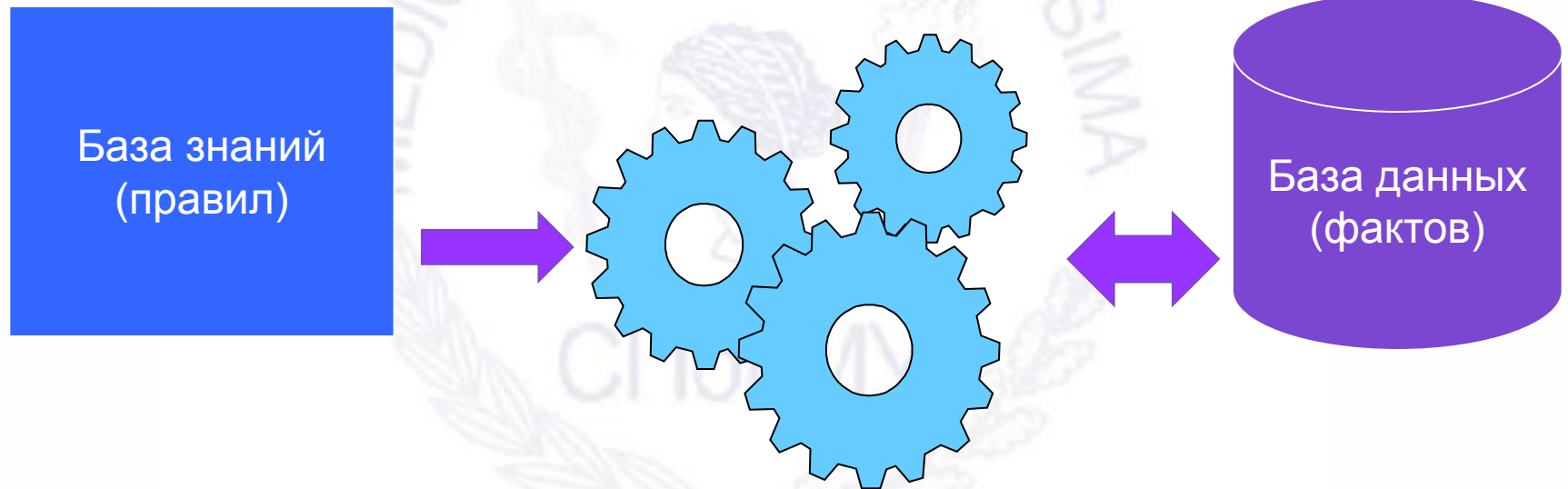


САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА

Кафедра физики, математики и информатики



Экспертные системы



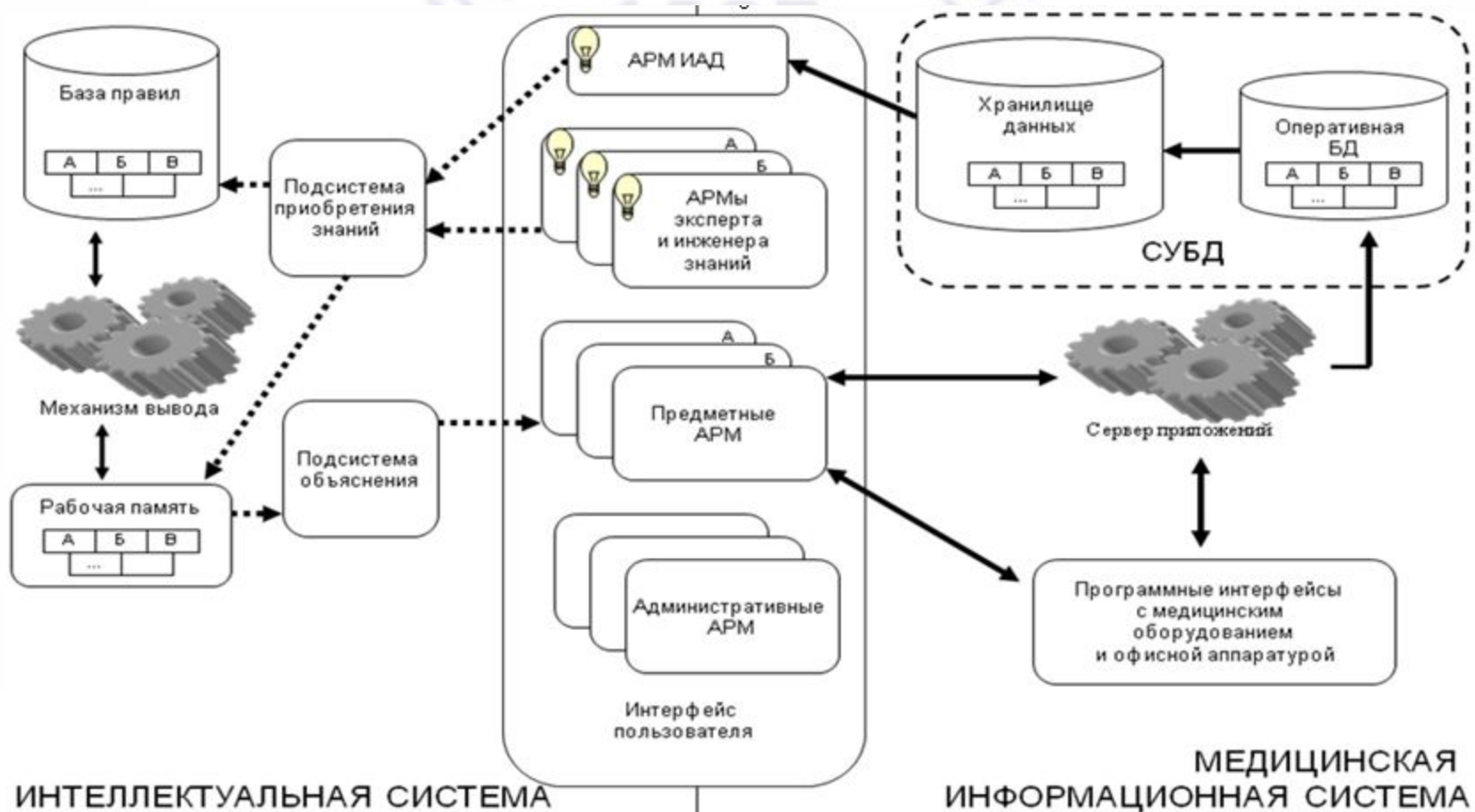


САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
ИМЕНИ АКАДЕМИКА И. П. ПАВЛОВА



Кафедра физики, математики и информатики

Интеллектуальная медицинская информационная система



Диагностика мочекаменной болезни

Показатель
степени МКБ

Pictures

- picture 1
- picture 2
- picture 3
- picture 4

Open

Select Images

- Image
- Zona
- Kristal

Small

Medium

Big

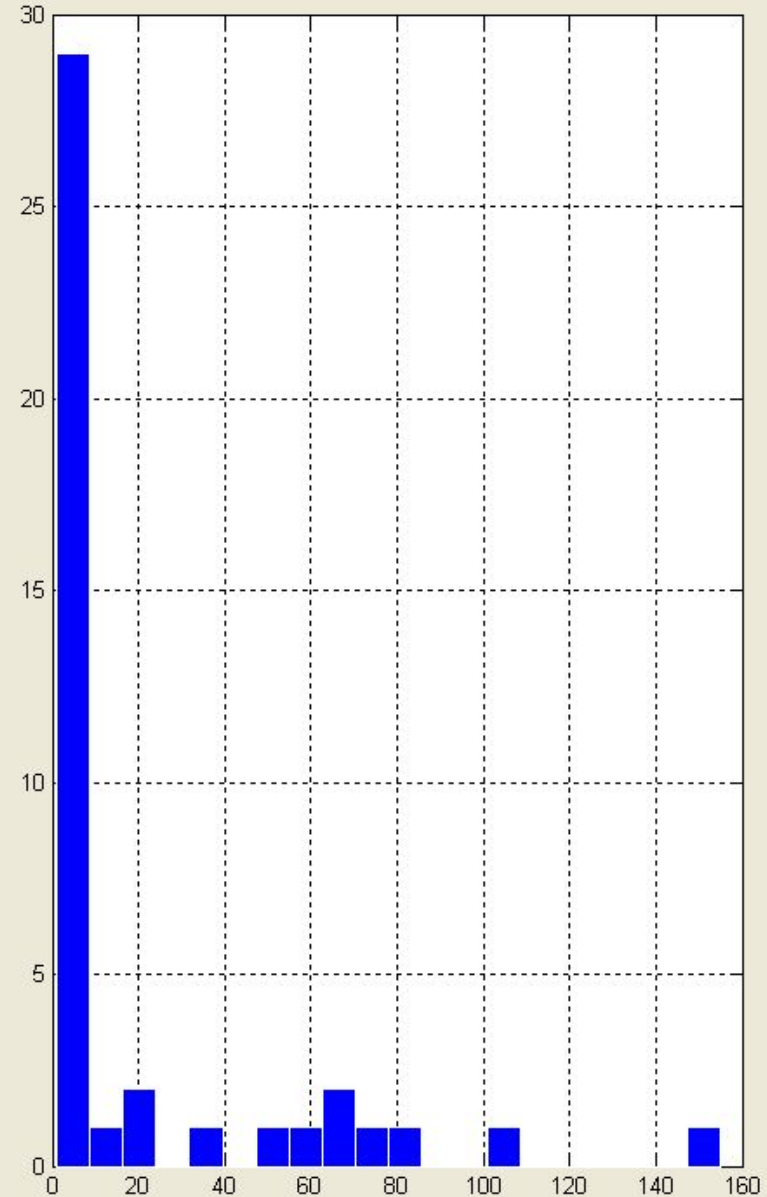
Число кристаллов
в пограничной зоне



Изображение,
полученное
с микровизора

Degrees

Select threshold



Диагностика мочекаменной болезни

