1-й Международный Молодежный Форум
«ПРОФЕССИЯ и ЗДОРОВЬЕ»
г. Москва, 31 мая – 3 июня 2016 года

ПРИМЕНЕНИЕ БАЙЕСОВСКОЙ КЛАССИФИКАЦИИ ДЛЯ ПРЕЦЕДЕНТНОГО РАСПОЗНАВАНИЯ В ПРОФПАТОЛОГИИ

Иванов А.Г.
ведущий инженер-программист

Восточно-Сибирский институт медико-экологических исследований
Ангарск

Актуальность проблемы

В каждом профпатологическом центре аккумулируется массив данных, исследование которого позволяет извлекать информацию, представляющую ценность для научно-исследовательского и диагностического процессов. Такое исследование требует соблюдения технических требований к организации данных, и предполагает применение аналитических инструментов, реализующих программно-математические методы в сфере искусственного интеллекта. Ввиду отсутствия указанных инструментов, интегрированных в регулярный рабочий процесс, информации, потенциально содержащиеся в накопленных данных, зачастую остаются невостребованными.

Thomas Bayes

(1701-1761)

Подходы к решению проблемы

Использование современных вычислительных технологий для реализации идей Т.Байеса (1736), показало состоятельность их приложений в медицине (Ashby D., 2006; Kadane J. B., 2005)

В отечественной и зарубежной науке и практике медицины труда известны подходы к использованию байесовских моделей для решения следующих задач:

- Оценка и прогнозирование профессионально-обусловленных рисков здоровью (Бисалиев Р. В., 2015; Creely K. S. et al., 2005; Li J., et.al., 2008);
- Выделение маркеров воздействия профессиональных факторов (Бударина Л.А., 2010);
- Моделирование и анализ заболевания на основе прецедентного материала (Andreassen S. et al., 1987; Greenland S., 1982; Liu J. et al., 2009);
- Разделение случаев по классам прецедентов (Амиров Н.Х. и др., 1999; G. B. Hamra et.al., 2014).

В то же время, в известных работах не предложено комплексной информационной технологии, способной предоставить инструментарий для регулярного использования в сфере профпатологии, научными сотрудниками и практикующими врачами.

Цель работы

Цель работы - создание программного инструмента для прецедентного распознавания в зависимости от выбора задачи классификации при анализе массивов результатов клинического профпатологического обследования, обладающих потенциальной диагностической ценностью.

Методы

1. Извлечение и консолидация сведений

Многомерный анализ и исследование данных (OLAP/Data Mining) (Palaniappan S., Ling C., 2008), классификация по продукционным правилам (*Гаврилова Т. А., 2000; Fieschi. M., 1990*)

2.Построение моделей информативности

Алгоритмы «Расширенный древовидный наивный байесовский классификатор (РДНБК)» (Tree Augmented Naïve Bayesian Classifier, TAN)(Friedman et al., 1996); «Поиск сети доверия с наибольшей байесовской метрикой (СДНБМ)» (Greedy Thick Thinning, GTT) (Cheng, J. et al., 1997)

3. Выявление взаимосвязей показателей

Оценка «силы влияния» (Strength of Influence, SoI) показателей (Koiter J.R., 2006; Ratnapinda P., Druzdzel M. J., 2014)

4. Выбор показателей

Оценка перекрёстной энтропии «свидетельство - заключение» (Pavlenko T., von Rosen D., 2002) (Обзор других подходов: R.B. O'Hara, M. J. Sillanpää, 2009)

5. Построение классифицирующей модели

Алгоритмы «Наивный байесовский классификатор (НБК)» (Naïve Bayesian Classifier, NBC) (Abraham R et al., 2007; Kantardzic M., 2011, с.146) «Расширенный древовидный наивный байесовский классификатор (РДНБК)»

Исходные данные

Объединённые профессиональные группы

№ п/п	Наименование	Категория		Число случаев
1	Аппаратчик установки перегонки	C1	Аппаратчик	354
2	Аппаратчик установки подготовки сырья			
3	Аппаратчик установки регенерации			
4	Аппаратчик установки синтеза			
5	Аппаратчик установки хлорирования			
6	Заместитель начальника цеха	C2	Инженерно-технический и административный персонал (ИТиАР)	163
7	Мастер смены			
8	Мастер участка контрольно-измерительных приборов и автоматики			
9	Мастер-механик			
10	Мастер-энергетик			
11	Начальник отделения			
12	Начальник смены			
13	Начальник цеха			
14	Оператор дистанционного пульта управления			
15	Старший мастер			
16	Энергетик			
17	Технолог			
18	Программист			
19	Инженер-электрик			
20	Мастер			
21	Мастер по ремонту			
22	Слесарь	C3	Слесари	382
23	Слесарь участка контрольно-измерительных приборов и автоматики			
24	Слесарь-ремонтник			
25	Слесарь-ремонтник (дежурный)			
26	Слесарь-сантехник			
27	Слесарь-электрик			
28	Слесарь-электрик по ремонту			
29	Чистильщик			
30	Электромонтёр	C4	Электромонтёры	79
31	Электромонтёр по ремонту			
		Итого, объем выборки		**978**

Исходные данные

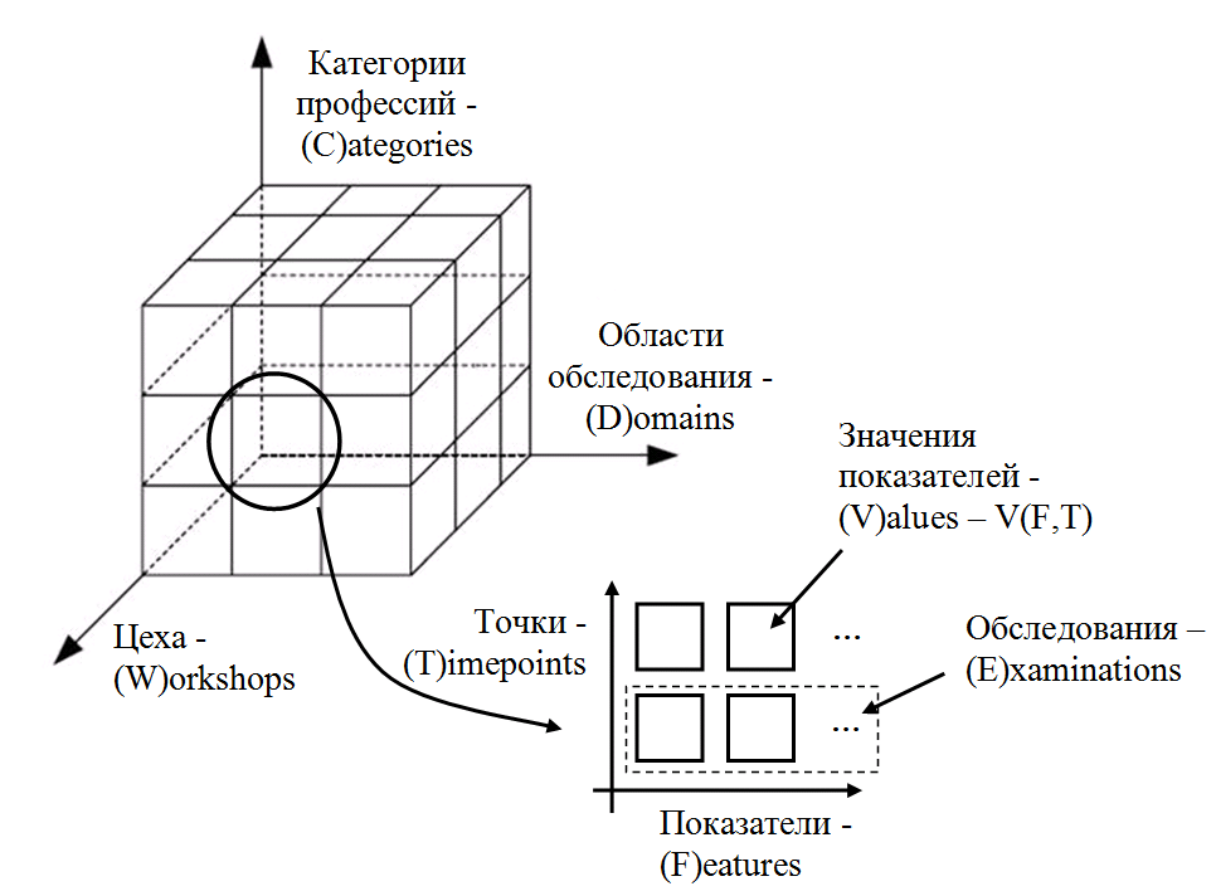
Производственные объекты в сопоставлении с видами экспозиции токсикантам

№ п/п	Предприятие	Цех		Факторы	Виды экспозиции		Количество случаев
		Шифр	Обозначение		Код	Описание	
1	АО 'Саянскхимпласт'	21	Цех по производству натра едкого, хлора и водорода методом мембранного электролиза	Пары ртути	1	Пары ртути	307
		22	Цех по производству концентрированной каустической соды				
		23	Цех по производству жидкого хлора				
		25	Цех очистки рассола и переработки продуктов электролиза ртутным методом				
		30	Производство винихлорида	Винилхлорид, дихлорэтан	2	Винилхлорид, дихлорэтан	347
		40	Цех производства поливинихлорида				
2	ООО 'Усольехимпром'	5001	Производство эпихлоргидрина и эпоксидно-диановых смол	Эпихлоргидрин, хлористый аллил	3	Эпихлоргидрин, хлористый аллил	221
		2101	Цех ртутного электролиза	Пары ртути	4	Пары ртути (постконтактный период)	103
						Итого, объем выборки	**978**

Группы показателей клинического обследования

№п/п	Наименование	Количество показателей
1	Индексы количественной оценки риска основных патологических синдромов и состояний (АСКОРС)	13
2	Биометрические и социально-личностные показатели	3
3	Обследование биохимического статуса	50
4	Обследование иммунологического статуса	12
5	Нейропсихологические методики обследования	3
6	Психологические методики обследования	17
7	Электронейромиография (ЭНМГ)	22
8	Электроэнцефалография (ЭЭГ)	7
9	Реоэнцефалография (РЭГ)	18
10	Биологический возраст (БВ)	3
11	Опросник "Краткая форма оценки здоровья" (SF-36)	8
	Итого, количество показателей	**156**

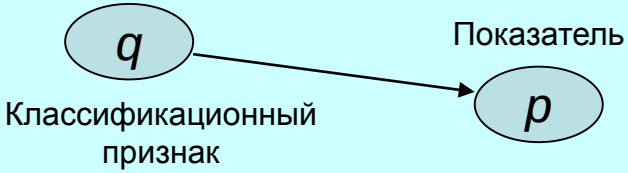
Организация исходных данных



Пространство исходных данных организовано в форме 5-мерного гиперкуба ***H(C,D,W)***, каждый элемент которого представлен плоскостью значений показателей ***V(F,P)***, разделённой временными точками ***T*** на векторы обследований ***E***. Опрос гиперкуба в соответствии с целью извлечения данных позволяет сформировать множество выборок для построения частных моделей информативности.

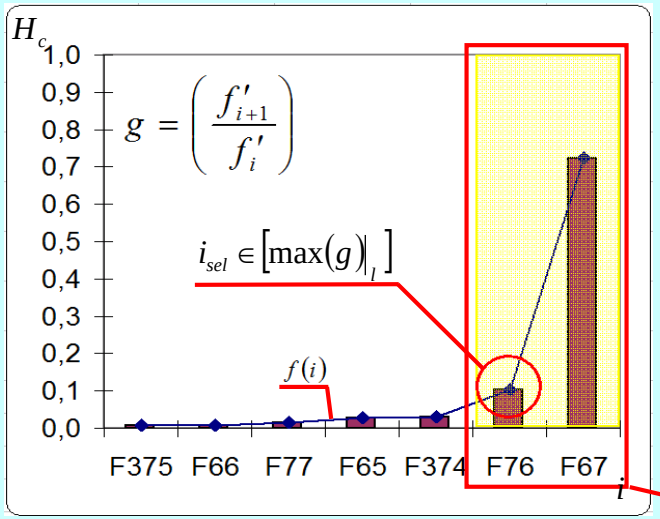
В ходе автоматизированного построения таких моделей, было сформировано 110 выборок (5 объединённых профессиональных групп, 11 направлений обследования, 2 временные точки), путём отбора и применения продукционных правил, проведена классификация 137280 значений показателей номинального и интервального типа.

Оценка информативности показателей и их отбор в классифицирующую модель



$$H_C(p,q) = -\sum p \cdot \log_2\left(\frac{p}{q}\right)$$

где информативность отдельного показателя *P* оценивается мерой перекрёстной энтропии H_c между распределением *P* априорных вероятностей состояния показателя, и распределением *Q* оценок исхода в классификационном признаке.



На примере области показателей «Психологические методики обследования»:

Наибольшей информативностью (ценностью для распознавания) в рамках модели, построенной для группы С1, обладает показатель F67 (Кратковременная память, методика Лурия '10 слов'), следует показатель F76 (Шкала астенического состояния (методика Малковой-Чертовой);

F67 и F76 подлежат отбору в классифицирующую модель.

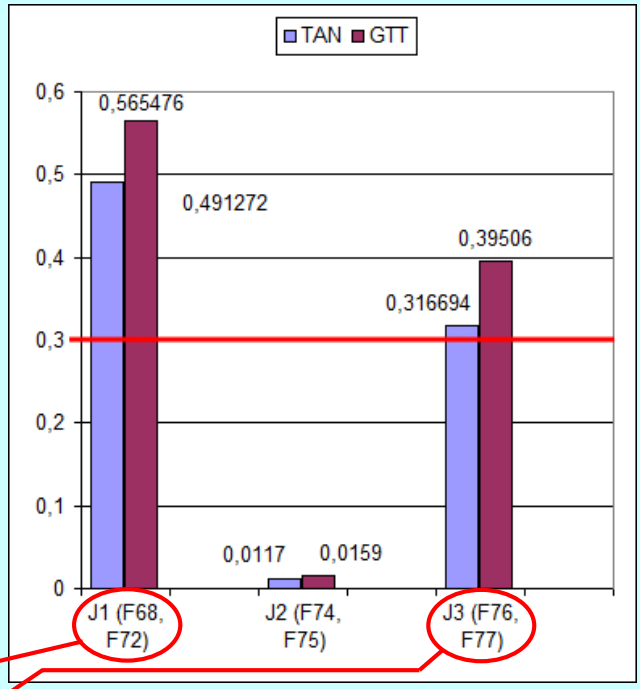
ДРНБК	С1	С2	С3	С4	С1-4
F65	0,027	0,003	0,003	0,013	0,000
F66	0,009	0,015	0,032	0,161	0,044
F67	0,724	0,200	0,229	0,220	0,266
F76	0,102	0,028	0,026	0,025	0,044
F77	0,015	0,008	0,004	0,016	0,016
F374	0,030	0,017	0,032	0,006	0,018
F375	0,008	0,048	0,022	0,082	0,002

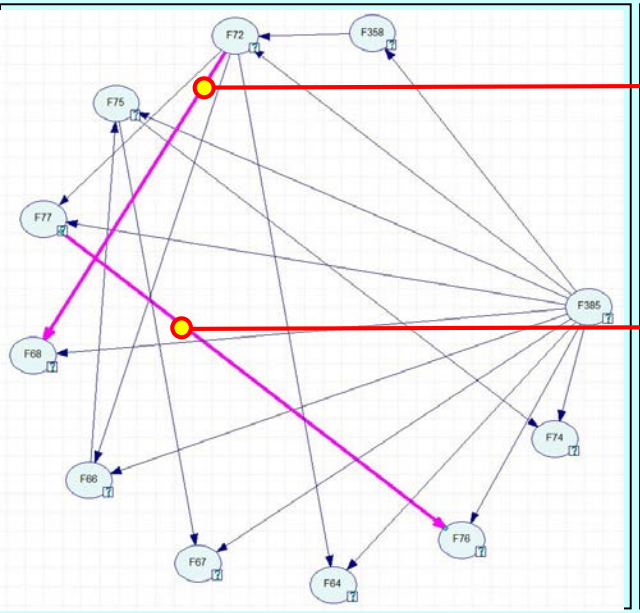
Подлежат отбору показатели **[F]**, расположенные на гистограмме, упорядоченной по возрастанию их информативности, правее точки излома i_{sel} кусочно-линейной огибающей ***f(i)*** с первым локальным максимумом ***g*** соотношения значений производной на ***n+1*** и ***n*** участках ***f(i)***. Окрестность локализации ***g*** определяется экспертной оценкой доли гистограммы ***L*** с наиболее информативными показателями (в % от общего числа показателей). В отсутствии значения ***g***, к отбору принимаются все информативные показатели правее долевой отметки ***L***. (Исследование подхода продолжается!)

Статистические взаимосвязи

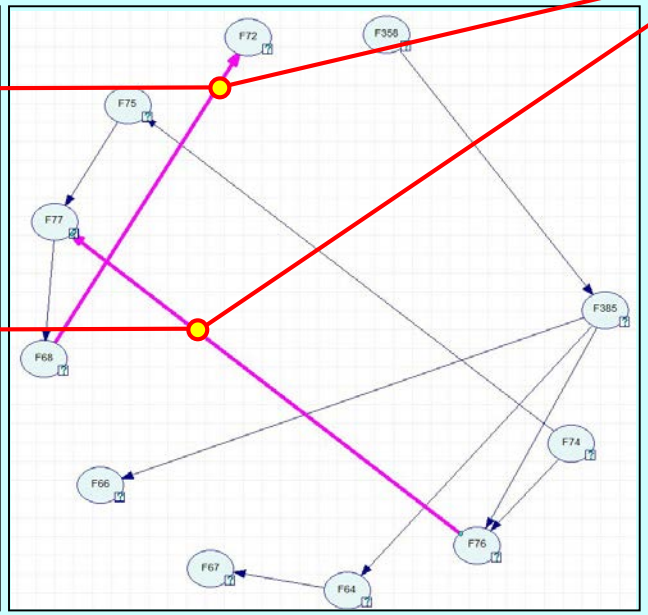
Значимость признака статистической взаимосвязи $J_i \in \{J\}$ между парой показателей ***(F1, F2)*** оценивается функцией «степень воздействия» (***SoI****) на основе меры расстояния между распределениями их априорных вероятностей. В представленном примере, условием идентификации взаимосвязи J_i является её наличие в моделях двух типов (TAN, GTT), при значении ***SoI***, большем среднего по всем выявленным взаимосвязям ***{J}.*** Взаимосвязи могут представлять исследовательскую ценность, а также помощь в принятии решения о сжатии признакового пространства путем исключения наименее информативного из комплементарных показателей. (Исследование подхода продолжается!)

*SoI – Strength of Influence (Koiter J.R., 2006)





ДРНБК (TAN)

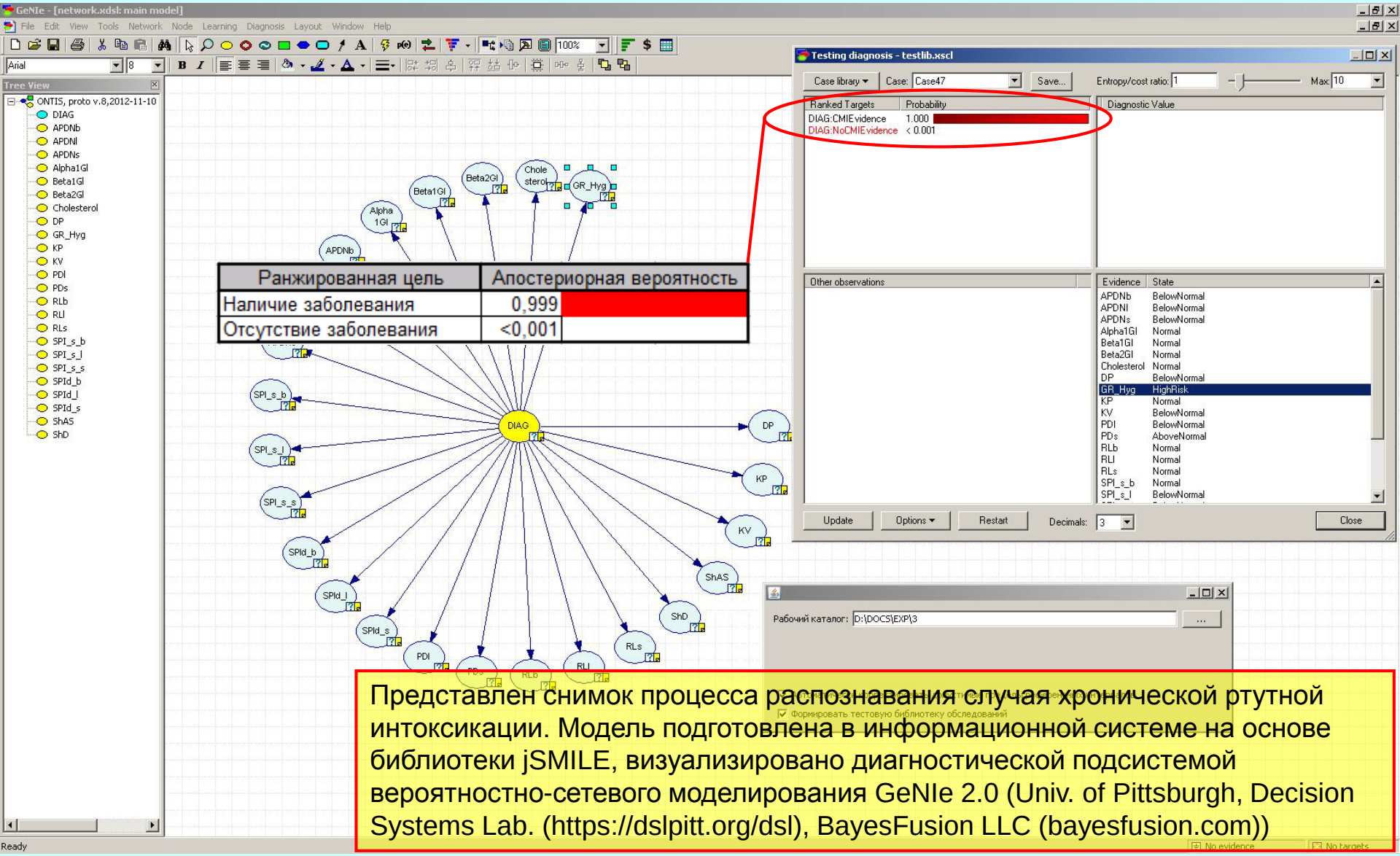


СДНБМ (GTT)

F68 - Личностная тревожность
F72 - Реактивная тревожность (методика Спилбергера – Ханина)
SoI(TAN) = 0,491272
SoI(GTT)=0,565476

F76 – Шкала астенического состояния
F77 - Шкала депрессии
SoI(TAN) =0,316694
SoI(GTT)=0,39506

Индивидуальные результаты распознавания хронической ртутной интоксикации при дифференцировании от иных нарушений нервной системы



Представлен снимок процесса распознавания случая хронической ртутной интоксикации. Модель подготовлена в информационной системе на основе библиотеки jSMILE, визуализировано диагностической подсистемой вероятностно-сетевого моделирования GeNIe 2.0 (Univ. of Pittsburgh, Decision Systems Lab. (https://dslpitt.org/dsl), BayesFusion LLC (bayesfusion.com))

Прецедентное распознавание вида экспозиции по объединённым профессиональным группам

Свойства выборки

	C1	C2	C3	C4	C1-4
	Аппаратчики	**ИТР**	**Слесари, чистильщики**	**Электрики, электромонтёры**	**Все категории**
Объем выборки (N)	354	163	382	79	978
Структура выборки (классы длительной экспозиции)					
1. Пары ртути	111	45	127	26	309
2. Винилхлорид, дихлорэтан	137	59	127	27	350
3. Эпихлоргидрин, хлористый аллил	74	55	79	20	228
4. Пары ртути (постконтактный период)	32	4	49	6	91

Результаты распознавания

	C1			C2			C3			C4			C1-4		
	Аппаратчики			**ИТР**			**Слесари, чистильщики**			**Электрики, электромонтёры**			**Все категории**		
Объем тестового раздела выборки	35			16			38			7			97		
Результат классификации	ВЕРН	ОШИБ	НЕОП	ВЕРН	ОШИБ	НЕОП	ВЕРН	ОШИБ	НЕОП	ВЕРН	ОШИБ	НЕОП	ВЕРН	ОШИБ	НЕОП
НБК	32	3	0	14	0	2	34	3	1	7	0	0	89	6	2
ДРНБК	31	3	1	15	1	0	35	3	0	7	0	0	91	5	1
НБК, %	91,43	8,57	0,00	87,50	0,00	12,50	89,47	7,89	2,63	100,00	0,00	0,00	91,75	6,19	2,06
ДРНБК, %	88,57	8,57	2,86	93,75	6,25	0,00	92,11	7,89	0,00	100,00	0,00	0,00	93,81	5,15	1,03

Полученные результаты классификации тестовых примеров по видам токсической экспозиции позволяют сделать вывод о целесообразности развития и исследования приложений байесовских методов для прецедентного распознавания в сфере профпатологии

Свидетельства о государственной регистрации программы для ЭВМ и базы данных

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2013612007

Система поддержки принятия решения в дифференциальной диагностике профессиональных заболеваний - Occupational Diseases Differential Diagnosis Decision Support System (O4D2S)

Правообладатель(ли): ***Федеральное государственное бюджетное учреждение «Восточно-Сибирский научный центр экологии человека» Сибирского отделения Российской академии медицинских наук (RU)***

Автор(ы): ***Иванов Антон Геннадьевич (RU),***
Дьякович Марина Пинхасовна (RU)

Заявка № **2012661264**
Дата поступления **17 декабря 2012 г.**
Зарегистрировано в Реестре программ для ЭВМ
11 февраля 2013 г.

Руководитель Федеральной службы по интеллектуальной собственности

Б.П. Симонов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации базы данных

№ 2013620376

База знаний «Показатели клинической диагностики профессиональных нейроинтоксикаций»

Правообладатель(ли): ***Федеральное государственное бюджетное учреждение «Восточно-Сибирский научный центр экологии человека» Сибирского отделения Российской академии медицинских наук (RU)***

Автор(ы): ***Иванов Антон Геннадьевич (RU),***
Дьякович Марина Пинхасовна (RU)

Заявка № **2013620061**
Дата поступления **10 января 2013 г.**
Зарегистрировано в Реестре баз данных
06 марта 2013 г.

Руководитель Федеральной службы по интеллектуальной собственности

Б.П. Симонов

Заключение

- Для обнаружения диагностически-значимой информации, содержащейся в массивах результатов клинического профпатологического обследования, предложен способ использования известных математико-алгоритмических подходов для приведения исходных данных к моделям прецедентного распознавания.
- Разработан программный инструмент, позволяющий сократить время исследования больших массивов данных, классифицировать представленный случай путём прецедентного распознавания, что может быть использовано как основа процедуры поддержки принятия решения в дифференциальной диагностике профессиональных заболеваний.

Автор выражает сердечную признательность научным сотрудникам и клиницистам Восточно-Сибирского института медико-экологических исследований за консультационную поддержку и информационное обеспечение при выполнении работы.

В отношении всех использованных в работе данных, было получено информированное согласие обследованных лиц на использование деперсонифицированных сведений о состоянии здоровья в научных целях.

Спасибо за внимание