Знания-Онтологии-Теории (ЗОНТ-07)

Реализация универсальной системы извлечения знаний «Discovery» и ее применение в задачах медицинской диагностики

Демин А. В.[1], Витяев Е. Е.[2], Полоз Т. Л.[3]

[1]*Институт систем информатики СО РАН, пр. Лаврентьева, д. 6, г. Новосибирск, 630090, Россия.*
[2] *Институт математики СО РАН, пр. ак. Коптюга, д. 4, г. Новосибирск, 630090, Россия.*
[3] *НУЗ Дорожная клиническая больница, Владимировский спуск, д. 2а, г. Новосибирск, 630003, Россия.*

vityaev@math.nsc.ru, alexandredemin@yandex.ru

Аннотация. *Для решения задач извлечения знаний из данных и получения непротиворечивых прогнозов была разработана универсальная система извлечения знаний «Discovery». Данная система дает возможность произвольно задавать класс обнаруживаемых гипотез и находить в данных все закономерности заданного класса. При помощи разработанной системы было проведено два исследования возможности диагностики фолликулярной опухоли щитовидной железы, основанные на использовании данных цитологического анализа и данных УЗИ. В обоих случаях система смогла извлечь из данных набор диагностических правил, позволяющих с высокой степенью точности диагностировать фолликулярную аденому и рак. Тестирование методом скользящего контроля показало, что точность прогнозов системы достигает 96%. Проведенные сравнения с нейронными сетями показали преимущество системы «Discovery» в качестве прогнозов.*

Ключевые слова

Обнаружение закономерностей, предсказание, знание,
извлечение знаний, медицинская диагностика

1 Введение

В настоящий момент разработано достаточно большое количество различных KDD&DM (Knowledge Discovery in Data Bases and Data Mining) методов и реализующих их программных систем. Данное направление продолжает бурно развиваться и совершенствоваться. Однако ни один из используемых в данный момент KDD&DM методов не способен извлечь из информации знания в полном объеме.

Анализ методов KDD&DM показывает [1-3], что для любого метода можно выделить его онтологию, включающую типы данных, с которыми работает метод и язык оперирования и интерпретации данных, также можно выделить класс гипотез, которые проверяет метод. Это накладывает на KDD&DM методы ряд ограничений:

1. информация, содержащаяся в данных, определяется множеством отношений и операций, интерпретируемых в онтологии предметной области. Существующие методы KDD&DM могут работать только с конкретными типами данных и использовать только конкретные виды отношений и операций. Тем самым, они, во-первых, не могут использовать всю информацию, содержащуюся в данных, во-вторых, могут получать результаты, не интерпретируемые в онтологии предметной области;

2. методы обнаруживают в данных только вполне определенные типы закономерностей.

Нами был разработан реляционный подход (Relational Data Mining) к методам извлечения знаний и реализующая его программная система «Discovery» [1-4], снимающие практически все ограничения с методов KDD&DM за счет использования языка первого порядка, который практически неограниченно расширяет множество типов используемых данных, а также позволяет описывать любые виды гипотез. Проведенные практические сравнения системы «Discovery» с такими широко распространенными методами как нейронные сети, решающие деревья, ассоциативные правила, статистические методы, FOIL, показывают, что система «Discovery» работает лучше и точнее других методов [1-2,5-6].

Существующие методы не в состоянии поддерживать режим исследования данных, когда обнаруживаемая закономерность заранее неизвестна. Каждый KDD&DM-метод обнаруживает свой специфический класс гипотез. Система «Discovery» способна поддерживать режим исследования данных. Кроме того, система «Discovery» может обнаружить и проверить на данных произвольный класс гипотез, который захочет проверить эксперт.

Система «Discovery» обнаруживает гипотезы, которые сформулированы в заданных экспертом (например, финансистом) терминах – множестве интерпретируемых отношений и операций, определенных на данных. Интерпретируемость получаемых закономерностей очень важна, например, для задач медицинской диагностики. К примеру, если речь идет о диагностики заболевания и у нас есть два диагноза, полученные нейронными сетями и системой «Discovery», то доверие будет к тому прогнозу, который понятен и интерпретируем. Невозможно принимать ответственные решения, не понимая, как они получены. Нейронные сети воспринимаются как черный ящик и поэтому их прогнозу доверять трудно. Прогнозы, получаемые на основании интерпретируемых правил понятны, и по ним можно принимать решения.

Другой важной задачей, которую решает система «Discovery», является задача максимально полного извлечения знаний из данных. Полнота извлечения знаний системой «Discovery» обеспечивается двумя путями:

1. использованием теории измерений, позволяющей извлечь практически всю информацию из данных и представить ее множеством интерпретируемых отношений и операций, определённых на многосортной эмпирической системе;
2. обнаружением практически любого класса гипотез в терминах выявленных отношений и операций на этой эмпирической системе.

Все эти задачи показывают актуальность создания «универсальной» версии системы «Discovery». Однако ввиду чисто практической сложности такой задачи, ранее разрабатывались только ограниченные версии системы для решения конкретных задач. На данный момент разработана достаточно «универсальная» версия системы [7], основным отличием которой является то, что она позволяет пользователю самому описывать виды гипотез, при помощи которых будет осуществляться извлечение знаний.

2 Метод обнаружения закономерностей

Определим фрагмент языка первого порядка $L(\Sigma)$, содержащий в качестве внелогических символов только множество Σ символов отношений и операций, интерпретируемых в онтологии предметной области и извлекающие нужную нам информацию из данных. Сигнатурой $\Im(\Sigma)$ языка $L(\Sigma)$ будем называть набор $\Im(\Sigma) = < P_1,...,P_K, f_1,...,f_M >$, где $P_1,...,P_K$ – символы отношений, а $f_1,...,f_M$ – символы операций.

Пусть $U(\Sigma)$ – множество всех атомарных формул вида $P(f_1(x_1,...,x_m),...,f_n(x_1,...,x_n))$, $f_i(x_1,...,x_m) = f_j(x_1,...,x_n)$, где x_i - переменные. Известно, что совокупность универсальных формул логически эквивалентна совокупности правил вида

$$\forall x_1,...,x_k (A_1 \& ... \& A_m \to A_0), \quad (1)$$

где A_i – литера вида $A_i = P^\varepsilon$, $P \in U(\Sigma)$, $\varepsilon \in \{0,1\}$ – обозначает наличие отрицания, т.е. если $\varepsilon = 0$, то $P^\varepsilon = P$, если $\varepsilon = 1$, то $P^\varepsilon = \neg P$.

Работа системы «Discovery» основана на семантическом вероятностном выводе [1], который позволяет находить все статистически значимые закономерности вида (1), с максимальной вероятностью предсказывающие литеру A_0.

Семантический вероятностный вывод основан на определении *вероятностной закономерности* [1], которое звучит следующим образом. Правило $A_1 \& ... \& A_m \to A_0$ является *вероятностной закономерностью*, если для любого правила $A_{i1} \& ... \& A_{ik} \to A_0$, такого что $\{A_{i1},...,A_{ik}\} \subset \{A_1,...,A_m\}$, условная вероятность $p(A_0 \mid A_{i1} \& ... \& A_{ik}) < p(A_0 \mid A_1 \& ... \& A_m)$.

Для дальнейшего описания введем понятие *уточнения* правила. Правило $A_1 \& ... \& A_m \& A_{m+1} \to A_0$ является *уточнением* правила $A_1 \& ... \& A_m \to A_0$, если оно получено добавлением в посылку правила $A_1 \& ... \& A_m \to A_0$ произвольной литеры A_{m+1}.

Рассмотрим алгоритм семантического вероятностного вывода.

Суть алгоритма заключается в последовательном уточнении правил с проверкой условия принадлежности к вероятностной закономерности. Очевидно, что перебор всех возможных правил – это вычислительно сложная задача, которая требует много времени. Поэтому для практической реализации семантического вероятностного вывода используется два последовательных перебора: *базовый перебор* и *дополнительный перебор*.

Алгоритм семантического вероятностного вывода.

- На первом шаге генерируется множество вероятностных закономерностей REG_1 путем полного перебора всех правил $A_1 \& ... \& A_m \to A_0$ для $1 \le m \le d$ с проверкой выполнения условия быть вероятностной закономерностью. Т.е. $REG_1 = \{R_i\}$, где $i \in I_1$, $R_i = A_1 \& ... \& A_m \to A_0$, $1 \le m \le d$, R_i – вероятностная закономерность. Данный шаг называется *базовым перебором*, а величина d – *глубиной базового перебора*.
- На k-м ($k > 1$) шаге генерируется множество вероятностных закономерностей REG_k путем уточнения всех закономерностей, найденных на предыдущем шаге, с проверкой выполнения условия быть вероятностной закономерностью. Т.е. $REG_k = \{R_i\}$, $i \in I_k$, $R_i = A_1 \& ... \& A_m \to A_0$, $m = d + (k-1)$, R_i – вероятностная закономерность, $R_i \in Spec(REG_{k-1})$, где $Spec(REG_{k-1})$ – множество всех уточнений правил из REG_{k-1}.
- Алгоритм останавливается, когда невозможно далее уточнить ни одно правило. Результирующее множество всех закономерностей REG будет равно объединению всех REG_k : $REG = \bigcup_k REG_k$.

Глубина базового перебора, т.е. максимальная длина правил, которые будут рассматриваться во время базового перебора, задается исследователем. В практических задачах глубина базового перебора обычно берется равной двум или трем.

Условная вероятность правил при выполнении алгоритма оценивается при помощи обучающего множества. Для того чтобы избежать генерации статистически незначимых правил, на практике обычно вводятся различные дополнительные критерии, оценивающие статистическую значимость. Правила, не удовлетворяющие этим критериям, отсеиваются, даже если они имеет высокую точность на обучающем множестве.

3 Особенности реализации системы «Discovery»

Система работает с исходными данными, представленными в виде набора таблиц $\{D\}$, строки которых соответствуют объектам, а колонки – признакам объектов. Т.е. $D = \{D(1),...,D(N)\}$, где $D(i)$ – i-я строка таблицы (объект с номером i), $D(i) = \{D_1(i),...,D_m(i)\}$, $D_j(i)$ – значение j-го признака объекта $D(i)$.

Для обеспечения универсализации способа задания видов гипотез, в программе реализована иерархия элементов конструирования гипотез, исполняющих роли соответствующих математических объектов. В качестве таких элементов выступают следующие объекты:

Переменные могут либо принимать значения фиксированных констант $x_j(i) = const$, где $const$ – произвольное действительное число, либо содержать обращение к исходным массивам данных $x_j(i) = D_k(i+b)$, где k – номер признака, значения которого принимает переменная, b – смещение относительно текущей строки i.

Терм $\theta(i) = f(x_1(i),...,x_k(i))$, где $f(x_1(i),...,x_k(i))$ – функция, задающая преобразование на множестве исходных данных, $x_j(i)$ – переменная, i – номер строки таблицы данных, к которой применяется преобразование. Терм используется для задания интерпретации предиката на исходных данных.

Предикат – определяет отношение на множестве данных. Предикат содержит термы, связанные этим отношением. Общий вид предиката: $P(i) = P(\theta_1(i),...,\theta_n(i))$, где $\theta_j(i)$ – терм. Чтобы предикат мог быть применен к множеству исходных данных, он должен быть проинтерпретирован на этом множестве данных. Для этого каждому терму должна быть присвоена какая-нибудь функция, задающая преобразование на множестве исходных данных, т.е. $\theta_j(i) = f_j(i) = f_j(x_1(i),...,x_k(i))$. Таким образом, проинтерпретированный предикат имеет вид: $P(i) = P(f_1(i),...,f_n(i))$.

Правило – служит для представления гипотез. Правило состоит из посылки и заключения. Посылка правила представляет собой конъюнкцию предикатов, заключение – некоторый целевой предикат. Общий вид правила: $R = P_1(i) \& ... \& P_n(i) \to P_0(i)$, где $P_1,\ldots,P_n$ – предикаты посылки; P_0 – целевой предикат. Каждое правило R характеризуется условной вероятностью $p(R)$, с которой оно предсказывает истинность заключения при условии истинности посылки.

Таким образом, чтобы определить вид гипотез, которые будут использованы для анализа исходных данных, пользователю достаточно указать виды предикатов, которые будут участвовать в формировании правил, виды функций, которые могут быть присвоены термам, и множества значений, которые могут принимать переменные.

В данной реализации системы «Discovery» пользователь, описывая гипотезы, может использовать только встроенные в систему виды предикатов и интерпретаций. Однако уже эта возможность позволяет произвольно задавать широкий класс гипотез. А разработанный и реализованный в данной системе способ конструирования гипотез в виде иерархии элементов конструирования гипотез, является достаточно универсальным и вполне согласуется с теоретической базой реляционного подхода, что открывает широкие возможности в направлении дальнейшей универсализации данной программной системы.

Рассмотрим способ представления исходной задачи в системе «Discovery».

Предполагается, что задача, решаемая при помощи системы «Discovery», может быть сведена к задаче предсказания нескольких логических утверждений, выраженных набором целевых предикатов. Таким образом, предполагается, что исходную задачу можно разбить на несколько отдельных задач, каждая из которых предсказывает один целевой предикат. Для предсказания каждого целевого предиката в системе определен объект *предиктор*. Предиктор объединяет в себе все правила, предсказывающие соответствующий целевой предикат. Поиск правил, которые составляют предиктор, происходит при помощи семантического вероятностного вывода.

Основной задачей предиктора является формирование итогового прогноза значения целевого предиката на основе прогнозов отдельных правил, входящих в состав данного предиктора. Под *прогнозом предиктора* PR для объекта с номером i будем понимать величину $pr_{PR}(i) \in [0,1]$, где pr_{PR} – отображение, определяющее способ формирования итогового прогноза. Отображение pr_{PR} ставит в соответствие множеству прогнозов отдельных правил значение из интервала $[0,1]$, т.е. $pr_{PR} : \{pr_R : R \in PR\} \to [0,1]$, где pr_R – прогноз правила R: $pr_R = p(R)$, если правило R применимо к объекту с номером i, $pr_R = 0$ в противном случае.

Наиболее естественным способ определения pr_{PR} является его задание равным прогнозу правила, имеющего максимальную условную вероятность, т.е. $pr_{PR}(i) = \max_R \{pr_R(i) : R \in PR\}$. Однако возможны и другие способы задания pr_{PR}, и система «Discovery» поддерживает несколько различных способов формирования прогноза предиктора.

Для создания прогнозирующей системы необходимо определить *решающее правило* $\boldsymbol{DecRule}$, которое должно на основе множества прогнозов отдельных предикторов выдавать определенный сигнал системы, т.е. $\boldsymbol{DecRule}:\{\boldsymbol{pr}_{\boldsymbol{PR}}\}\to\{\boldsymbol{S}_j\}$, где $\{\boldsymbol{S}_j\}$ – множество сигналов системы.

В системе «Discovery» реализован следующий механизм определения решающего правила. Пусть имеется некоторый набор предикторов $\{\boldsymbol{PR}_j\}$, $\boldsymbol{j}=\boldsymbol{1},...,\boldsymbol{n}$. Каждому предиктору $\boldsymbol{PR}_j$ ставиться в соответствие (задается пользователем) некоторый сигнал системы $\boldsymbol{S}_j$. Обозначим через $\boldsymbol{pr}_{\boldsymbol{PR}}{}^{\boldsymbol{j}}(\boldsymbol{i})$ прогноз $\boldsymbol{j}$-го предиктора для объекта с номером $\boldsymbol{i}$. Сигнал системы $\boldsymbol{Sys}(\boldsymbol{i})$ для $\boldsymbol{i}$-го объекта определяется следующим образом. Для каждого предиктора рассчитывается показатель согласованности его прогноза по формуле $\boldsymbol{Ctr}_j(\boldsymbol{i})=\boldsymbol{pr}_{\boldsymbol{PR}}{}^{\boldsymbol{j}}(\boldsymbol{i})-\max_{\boldsymbol{k}\neq\boldsymbol{j}}\{\boldsymbol{pr}_{\boldsymbol{PR}}{}^{\boldsymbol{k}}(\boldsymbol{i})\}$, т.е. как разность между величиной прогноза данного предиктора и максимальной величиной прогнозов остальных предикторов. В качестве сигнала системы для $\boldsymbol{i}$-го объекта выбирается сигнал, соответствующий предиктору, показатель согласованности которого строго больше заданного пользователем порога $\boldsymbol{\delta}>\boldsymbol{0}$, т.е. $\boldsymbol{Sys}(\boldsymbol{i})=\boldsymbol{S}_{\boldsymbol{k}}$, где $\boldsymbol{k}=\arg\max_{\boldsymbol{j}=\boldsymbol{1},..,\boldsymbol{n}}\{\boldsymbol{Ctr}_j(\boldsymbol{i}):\boldsymbol{Ctr}_j(\boldsymbol{i})>\boldsymbol{\delta}\}$. В случае, если не существует прогноза, показатель согласованности которого выше указанного порога, то считается, что прогноз для этого объекта отсутствует. Таким образом, регулируя порог $\boldsymbol{\delta}$, пользователь получает возможность контролировать допустимую степень противоречивости прогнозов.

4 Применение в задачах медицинской диагностики

Мы использовали разработанную нами систему для решения задачи дифференциальной диагностики фолликулярного рака и фолликулярной аденомы щитовидной железы.

Дооперационная диагностика заболеваний щитовидной железы является актуальной задачей, поскольку является определяющим фактором в дальнейшей тактике ведения больного, выбора консервативного или хирургического метода лечения [8,9]. Повышение точности дооперационной диагностики позволяет снизить количество неоправданных хирургических вмешательств и дает возможность планировать рациональное лечение.

В настоящее время основным диагностическим методом заболеваний щитовидной железы является цитологическое исследование материала, полученного при аспирационной тонкоигольной пункции. В тоже время, сейчас все более важным становиться роль УЗИ в определении вида поражения, поскольку, как правило, именно при УЗИ обследовании впервые обнаруживается узел в щитовидной железе.

Наибольшие трудности при дооперационной диагностики заболеваний щитовидной железы вызывает диагностика так называемых фолликулярных опухолей при попытке дифференцировать фолликулярную аденому и фолликулярный рак. Как показывает опыт, при дооперационной цитологической диагностики фолликулярной опухоли совпадение цитологических и окончательных гистологических диагнозов не превышает 56% [9]. Что же касается УЗИ обследования, то на данный момент в медицине вообще не существует методов, позволяющих диагностировать фолликулярную опухоль по данным УЗИ обследования [8].

При помощи системы «Discovery» мы провели два исследования возможности диагностики фолликулярной опухоли, основанные на использовании данных цитологического анализа и данных УЗИ обследования.

4.1 Диагностика по цитологическим признакам

Исходными данными для анализа послужили результаты цитологического исследования мазков-отпечатков с ткани щитовидной железы 197 больных с уже известными гистологическими диагнозами (86 раков и 111 аденом). Все цитологические препараты были проанализированы по 30 признакам. Отмечали наличие или отсутствие коллоида, его консистенцию, преобладающие клеточные структуры, наличие значительного количества разрозненных клеток. Фолликулярные структуры оценивались по степени размерного полиморфизма, наличию шаровидных (трехмерных) и атипических структур. Цитоплазму клеток характеризовали по наличию или отсутствию четкого контура и локализации вакуолей.

Отмечали форму и полиморфизм ядер, неровность их контура, наслоение ядер друг на друга, структуру и равномерность распределения хроматина, наличие ядрышек.

Для наиболее объективной оценки прогностических возможностей системы, мы использовали метод скользящего контроля. Скользящий контроль осуществлялся следующим образом. Из общей выборки из 197 примеров циклически исключался один пример, система обучалась на оставшихся 196 примерах, а затем тестировалась на исключенном примере. После чего исключенный пример возвращался назад в общую выборку, и из нее исключался следующий пример, и т.д. Данный процесс продолжался до тех пор, пока система не пройдет через всю выборку данных.

В Таблице 1 представлены результаты тестирования системы при использовании правил, имеющих условную вероятность выше заданного порога.

Таблица 1. Результаты тестирования на цитологических данных.

Порог условной вероятности правил	Количество правильных прогнозов	Процент отказов от принятия решения
90 %	93 %	8 %
100 %	96 %	13 %

Как видно из таблицы, повышение порога условной вероятности правил ведет к увеличению точности прогноза, однако при этом также увеличивается число отказов от принятия решения. При этом необходимо отметить, что отказ от принятия решения не является ошибкой системы, а говорит лишь о недостатке информации для выдачи диагноза с заданной степенью точности.

Приведем несколько примеров найденных правил:

1. **ЕСЛИ** Фолликулы: атипичные **И** Наслоение ядер **И** Структура хроматина: неравномерная **И** Структура хроматина: мелкозернистая **ТО** Рак с вероятностью 100%.
2. **ЕСЛИ НЕ** Коллоид: обильный **И** Виды фолликулов: выражен полиморфизм **И** Наслоение ядер **И** Структура хроматина: неравномерная **ТО** Рак с вероятностью 100%.
3. **ЕСЛИ НЕ** Преобладающие структуры: разрозненные клетки **И НЕ** Фолликулы: атипичные **И** Полиморфизм ядер слабо выражен **ТО** Аденома с вероятностью 100%.

Мы провели сравнение точности прогнозов системы «Discovery» с прогнозами, полученной при помощи нейронной сети. Ранее мы проводили исследования возможности диагностики фолликулярной опухоли при помощи нейронных сетей и получили достаточно высокую степень точности прогнозов [10,11]. Тестирование нейронной сети описанным выше методом скользящего контроля показало точность, равную 91%, что соответствует проведенным ранее исследованиям.

Таким образом, проведенное сравнение показало, что система «Discovery» имеет более высокую точность прогнозов, чем нейронные сети. Кроме того, необходимо отметить, что система «Discovery» позволила получить интерпретируемые диагностические правила. Эти правила делают автоматизированный диагностический процесс принятия решения прозрачным и понятным для врача.

4.2 Диагностика по УЗИ признакам

Для анализа нами использовались данные об УЗИ обследовании 170 больных с уже известными гистологическими диагнозами (70 раков и 100 аденом). Результаты УЗИ обследования каждого больного были проанализированы по 9 признакам, которые включают общую эхогенность, однородность структуры, неровность и четкость контура, наличие кальцинатов, «гало», а также локализацию и данные о размерах узла.

В Таблице 2 представлены результаты тестирования системы методом скользящего контроля. Как видно из таблицы, обнаруженные системой «Discovery» диагностические правила позволяют с достаточно высокой степенью точности (до 96 %) диагностировать

фолликулярный рак и фолликулярную аденому, основываясь только на данных УЗИ обследования.

Таблица 2. Результаты тестирования на УЗИ данных.

Порог условной вероятности правил	Количество правильных прогнозов	Процент отказов от принятия решения
80 %	83 %	4 %
90 %	87 %	19 %
100 %	96 %	33 %

Примеры обнаруженных правил:

1. **ЕСЛИ** Структура неоднородная **И** Нет кальцинатов **И** Эхогенность: гипо **И** Размер узла > 16 **ТО** Рак с вероятность 100%.
2. **ЕСЛИ** Структура неоднородная **И НЕ** «Гало» **И** Размер узла < 23 ТО Рак с вероятностью 93%.
3. **ЕСЛИ** Нет кальцинатов **И** Контур: четкий **И НЕ** Эхогенность: гипо **И** Размер узла < 20 **ТО** Аденома с вероятностью 100%.

Мы также провели сравнение прогнозов системы «Discovery» с прогнозами нейронной сети. Нейронная сеть при скользящем контроле на данных УЗИ показала точность 86%. Таким образом, на данных УЗИ системы «Discovery» также превзошла нейронные сети в точности прогнозов.

5 Заключение

Проведенные нами исследования подтверждают вывод, сделанный в предыдущих работах [5, 12-14], о возможности успешного применения системы «Discovery» для решения сложных задач медицинской диагностики. Как и в предыдущих работах, полученные нами диагностические правила позволяют с достаточно высокой степенью точности (до 96 %) диагностировать фолликулярный рак и фолликулярную аденому и превосходят по точности другие методы, такие как нейронные сети, решающие деревья и статистические методы [5, 12-14]. Кроме того, необходимо отметить, что при помощи системы «Discovery» нами были получены интерпретируемые правила, которые дают не только вероятностный прогноз, но и его объяснение.

Литература

[1] Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов. – Новосибирск: НГУ, 2006. – 293 с.

[2] Kovalerchuk B., Vityaev E. Data Mining in Finance: Advances in Relational and Hybrid methods. – Kluwer Academic Publishers, 2000. – p.308.

[3] Vityaev E., Kovalerchuk B. Empirical Theories Discovery based on the Measurement Theory // Mind and Machine, Vol.14, N 4. – 2004. – pp.551-573.

[4] Витяев Е.Е., Москвитин А.А. Введение в теорию открытий. Программная система Discovery // Логические методы в информатике. – Новосибирск, 1993. – Вып. 148: Вычислительные системы. – С.117-163.

[5] Kovalerchuk B., Vityaev E., Ruiz J.F. Consistent and Complete Data and "Expert" Mining in Medicine // Medical Data Mining and Knowledge Discovery. – Springer, 2001. – pp. 238-280.

[6] Vityaev E., Kovalerchuk B. Data Mining For Financial Applications // Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers / Ed. by Maimon O., Rokach L. – Springer, 2005. – pp.1203-1224.

[7] Демин А.В., Витяев Е.Е. Реализация универсальной версии системы «DISCOVERY» // Тез. докл. конференции-конкурса «Технологии Microsoft в теории и практике программирования», Новосибирск, 24–26 февраля 2007. – С. 106–108.

[8] Богин Ю. Н., Бондаренко В. О., Шапиро Н. А., Орлов В. М. Комплексная экспресс-диагностика заболеваний щитовидной железы // Метод. рекомендации. – Москва, 1992. – 175 с.

[9] Пупышева Т. Л. Морфометрия клеток фолликулярных пролифератов щитовидной железы в тонкоигольных аспиратах // Новости клинической цитологии России. – 2002. – Т.6. – № 1-2. – С.24-26.

[10] Shapiro N.A., Poloz T.L., Shkurupij V.A., Tarkov M.S., Poloz V.V., Demin A.V. Application of Artificial Neural Network for Classification of Thyroid Follicular Tumors // Anal. Quant. Cytol. Histol. – 2007. – V. 29, – P. 122-119.

[11] Пупышева Т.Л., Демин А.В. Применение искусственных нейронных сетей в цитологической диагностике фолликулярных пролифератов щитовидной железы // Системный анализ и управление в биомедицинских системах – 2003. – Т.2. – №1. – С. 38-40.

[12] B.Kovalerchuk, E.Vityaev, James F.Ruiz Design of consistent system for radiologists to support breast cancer diagnosis. Joint Conference of Information Sciences, v.2 Computational intelligence, Neural network and Semiotics, Elsevier Publishing Company, 1997, pp. 118-121.

[13] B. Kovalerchuk, E. Vityaev, J. Ruiz. Consistent knowledge discovery in medical diagnosis. IEEE Engineering in Medicine and Biology Magazine (special issue “Data Mining and Knowledge Discovery”), July/August 2000, pp.26-37.

[14] Kovalerchuk, B., Triantaphyllou, E, Ruiz, J., Torvik, V., Vityaev, E. The Reliability Issue of Computer-Aided Breast Cancer Diagnosis, Computers and Biomedical Research, 2000 Aug., 33(4): 296-313.