# Метод индуктивного формирования баз медицинских диагностических знаний

© С. В. Смагин
Институт автоматики и процессов управления ДВО РАН,
Дальневосточный федеральный университет,
Владивосток
sergey.v.smagin@gmail.com

**Аннотация**

В работе представлено введение в проблематику индуктивного формирования баз знаний, приведены традиционные постановки основных задач обучения в этой области, выделены существующие актуальные проблемы, в том числе связанные с интерпретируемостью получаемых результатов. Для решения обозначенных проблем в предметной области медицинской диагностики предложен метод индуктивного формирования хорошо интерпретируемых баз медицинских диагностических знаний, который включает в себя: новые постановки основных задач обучения (классификации и кластеризации) для моделей зависимости с параметрами, алгоритм обучения (решающий эти задачи в их новых постановках) для хорошо интерпретируемой и полезной на практике математической модели зависимости с параметрами (онтологии медицинской диагностики, приближенной к реальной), а также комплекс программ InForMedKB, в котором реализован этот алгоритм. Указанный комплекс программ позволяет создавать обучающие выборки (состоящие из историй болезни различных разделов медицины) и на их основе индуктивно формировать базы медицинских диагностических знаний, обладающие высоким уровнем качества и при этом представленные в форме, принятой в медицинской литературе, а также генерировать объяснение этих баз знаний (на основе информации из обучающих выборок). Формальное представление баз знаний позволяет использовать их в интеллектуальных системах медицинской диагностики. Работа выполнена при частичной финансовой поддержке РФФИ, проекты №14-07-00270 и № 15-07-03193. Автор выражает глубокую признательность своему учителю д.ф.-м.н., профессору А.С. Клещеву.

**Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных», Обнинск, 13-16 октября 2015 г.**

## 1 Введение

Индуктивное формирование баз знаний на основе эмпирических данных является основным способом получения новых эмпирических знаний в науке и практике. Он заключается в получении общего знания о некоторой совокупности объектов на основании анализа единообразного описания конечного множества отдельных представителей этой совокупности – обучающей выборки (данных) [7]. Моделирование такого способа познания лежит в основе целого ряда направлений исследований, получивших в англоязычной литературе названия: Data Mining (интеллектуальный анализ данных), Machine Learning (машинное обучение), Knowledge Discovery in Databases (обнаружение знаний в базах данных), Pattern Recognition (распознавание образов), Knowledge Extraction (извлечение знаний), Data Archaeology (археология данных) и т.д., каждое из которых характеризуется собственным подходом к проблеме индуктивного формирования баз знаний, собственными постановками задач и многообразием методов их решения. По данной тематике опубликовано большое число работ, в которых сформулированы общие постановки основных задач индуктивного формирования баз знаний (задач классификации и кластеризации), изучены разнообразные модели зависимости между классами (кластерами) и объектами, а также разработано большое число алгоритмов обучения (классификации и кластеризации), решающих поставленные задачи на этих моделях [1, 3, 7, 8].

## 2 Традиционные постановки задач классификации и кластеризации

В традиционной постановке задачи классификации дано конечное множество объектов (образов, ситуаций, прецедентов), называемое обучающей выборкой, по каждому из которых собраны (измерены) некоторые данные. Данные об объекте называют также его описанием, при этом наиболее распространенным способом описания объектов является признаковое описание. Также дано конечное множество возможных классов (ответов, откликов, реакций). Предполагается, что

существует некоторая зависимость между классами и объектами, но она неизвестна. Предполагается, что для каждого объекта обучающей выборки задан его (единственный) правильный класс. При этом описание объектов обучающей выборки в той или иной степени отражает эту зависимость [1, 7]. На основе этой информации требуется для некоторого класса моделей зависимости (класса моделей некоторой предметной области (ПО)), к которому относится эта неизвестная зависимость (между классами и объектами), разработать алгоритм классификации, который по обучающей выборке строит такое решающее правило, вероятность правильной классификации которого любых новых объектов является как можно более высокой [4, 8]. Качество решающего правила оценивается на основе контрольной выборки, которая отличается от обучающей только способом ее использования.

Традиционные постановки задач классификации и кластеризации отличаются тем, что в последней разбиение множества объектов на классы неизвестно, и поэтому для объектов обучающей выборки правильные классы не могут быть заданы. Задача кластеризации сводится к разбиению обучающей выборки на непересекающиеся подмножества (называемые кластерами) – так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались; при этом строятся описания кластеров, позволяющие относить к ним новые объекты [3, 8].

### 2.1. Задача классификации

Пусть $x$ – вектор переменных (признаков), каждая из которых имеет свою область возможных значений, а $X$ – многомерное (признаковое) пространство, координатами которого являются переменные вектора $x$. Обозначим $y = \{y_1,...,y_n\}$ – множество классов. Модель зависимости $m(x, y)$ есть система соотношений между вектором переменных $x$ и значениями переменной $y$ (классами).

Модель зависимости $m$ индуцирует отношение $K \subset X \times y$ между объектами пространства $X$ и классами $y$, к которым они относятся. Обозначим $T$ (множество обучающих выборок) множество непустых конечных подмножеств $K$, $C$ (множество контрольных выборок) – множество непустых конечных подмножеств объектов пространства $X$.

Предполагается, что по модели зависимости $m$ может быть определено множество решающих правил $\Re_m$, каждый элемент которого есть отображение $R_m : X \to y$, которое сопоставляет каждой точке (объекту) пространства $X$ ее класс $y_i$. Алгоритмом классификации для модели $m$ называют отображение $L_m : T \to \Re_m$, которое по обучающей выборке $t \in T$ строит решающее правило $R_m : X \to y$ из множества $\Re_m$.

Для решающего правила $R_m$ обозначим $P(R_m)$ – вероятность правильной классификации объектов из $X$, оценка которой может быть получена с помощью контрольной выборки $c$.

Традиционно в работах по машинному обучению рассматривается следующая постановка задачи классификации. Для модели $m$ построить такой алгоритм классификации $L_m$, что для заданной обучающей выборки $t$ вероятность $P(L_m(t))$ имеет возможно большее значение [11].

Поскольку задача поиска наибольшего значения для всех возможных алгоритмов классификации и обучающих выборок не имеет шансов быть решенной, как правило, рассматривается более конкретная постановка задачи классификации, например, задача чемпионата среди алгоритмов классификации. Пусть имеются алгоритмы классификации $L_{m_1},...,L_{m_z}$, для моделей $m_1,...,m_z$ соответственно, и обучающая выборка $t$. Для модели $m$ требуется построить такой алгоритм классификации $L_m$, что для заданной обучающей выборки $t$ имеет место $P(L_m(t)) > P(L_{m_1}(t)),...,$ $P(L_m(t)) > P(L_{m_z}(t))$. В соответствии с этой постановкой задачи проводятся чемпионаты среди алгоритмов классификации ($L_m$ в данном случае считается победителем среди участников $L_{m_1},...,L_{m_z}$). Нередко предлагаемый алгоритм классификации $L_m$ сравнивается таким способом с другими известными алгоритмами классификации $L_{m_1},...,L_{m_z}$. Приведенную постановку задачи классификации можно назвать «слабой», т.к. в ней рассматривается всего одна обучающая выборка $t$. Если же в постановке задачи рассматривается множество обучающих выборок $t_1,...,t_v$, то такую постановку можно назвать «сильной». В последнем случае алгоритм классификации $L_m$ считается победителем среди участников $L_{m_1},...,L_{m_z}$, если его вероятность правильной классификации превосходит вероятности других алгоритмов на всех выборках $t_1,...,t_v$. Сильная постановка задачи поддерживается системой тестирования алгоритмов классификации «Полигон алгоритмов» [14].

В обеих приведенных постановках можно указать следующий недостаток. Оценка качества алгоритма классификации $P(R_m)$ зависит от конкретной обучающей выборки $t$ (в случае слабой постановки) или от набора обучающих выборок

$t_1,...,t_v$ (в случае сильной постановки).

### 2.2. Задача кластеризации

Пусть $Y$ – бесконечное множество номеров (имен, меток) кластеров предметной области (кластеров ПО). Предполагается, что модель зависимости $m$ такова, что с ней согласуется некоторое отношение эквивалентности на пространстве объектов $X$ (классы эквивалентности которого называются кластерами ПО и обозначаются номерами из множества $Y$) и по ней может быть построена функция расстояния (метрика) $\rho(x,x')$ между объектами пространства $X$, обладающая следующим свойством: для любых номеров кластеров ПО $y_i$, $y_j$ и любых объектов $x_a, x_b \in y_i$, $x_c \in y_j$ имеет место

(*) $\rho(x_a,x_b) < \rho(x_a,x_c)$ и

$$\rho(x_a,x_b) < \rho(x_b,x_c).$$

Назовем алгоритмом кластеризации такое отображение $M_m$, которое по обучающей выборке $t \in T$ строит некоторое отношение эквивалентности $E(t)$ на выборке $t$ (т.е. разбивает выборку $t$ на непересекающиеся подмножества, называемые индуктивными кластерами, при этом каждому объекту $x_i \in t$ приписывается некоторый номер индуктивного кластера $y_l$).

Традиционно задача кластеризации состоит в построении такого алгоритма кластеризации $M_m$, что для любой обучающей выборки $t$, любых номеров индуктивных кластеров $y_i$, $y_j$ и любых объектов $x_a, x_b \in y_i$, $x_c \in y_j$ выполнены неравенства (*).

Задача кластеризации отличается от задачи классификации тем, что номера кластеров ПО $y_l$ для объектов обучающей выборки $t$ изначально не заданы, и неизвестно само множество $Y$. Помимо этого решение задачи кластеризации в такой постановке принципиально неоднозначно по следующим причинам [1]: не существует однозначно наилучшего критерия качества кластеризации, число кластеров ПО неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием, а результат кластеризации существенно зависит от метрики $\rho$, выбор которой, как правило, субъективен и определяется конкретным экспертом.

## 3 Постановки задач классификации и кластеризации для моделей зависимости с параметрами

В работе [11] предложены новые постановки основных задач индуктивного формирования баз знаний (классификации и кластеризации) для моделей зависимости, имеющих вид необогащенных систем логических соотношений с параметрами из работы [10].

Предложенные постановки лишены недостатков традиционных постановок и представлены как частный случай задачи оценки значений параметров модели зависимости (при этом критерием качества обучения является близость получаемых оценок значений параметров к значениям, характеризующим ПО, а не традиционно используемая вероятность правильной классификации решающим правилом (получаемым в результате обучения) нового объекта, и не качество используемой метрики).

### 3.1. Задача классификации

Пусть модель зависимости $m(x,y,q)$ есть система соотношений между вектором переменных $x$, значениями переменной $y$ (классами) и вектором параметров $q$ с областью возможных значений $Q$.

Будем считать, что модель зависимости $m$ такова, что по ней может быть построено зависящее от значений параметров $q$ решающее правило $R_{mq} : X \to y$, которое сопоставляет каждой точке (объекту) пространства $X$ ее класс $y_i$.

Для решающего правила $R_{mq}$ обозначим $P(R_{mq})$ – вероятность правильной классификации объектов из $X$, оценка которой может быть получена с помощью контрольной выборки $c$.

В этом случае алгоритмом классификации для модели $m$ назовем отображение $L_m : T \to Q$, которое по обучающей выборке $t \in T$ вычисляет значения параметров $q$.

Будем считать, что ПО, в которой решается задача классификации, характеризуется значениями параметров $q^*$. Обозначим $P_{L_m}(\mu)$ – вероятность того, что $L_m(t) = q^*$ (для дискретных значений рассматривается вероятность совпадения значений параметров, а для числовых вещественных значений – вероятность того, что они отличаются не более чем на заданное число) для обучающих выборок $t$ мощности $\mu$.

В работе [11] рассматривается следующая новая постановка задачи классификации: для модели $m$

построить такой алгоритм классификации $L_m$, что $P_{L_m}(\mu)$ стремится к 1 при $\mu$, стремящемся к бесконечности. Очевидно, что функция $P_{L_m}(\mu)$ не зависит от конкретных обучающих выборок, т.е. такая постановка задачи классификации не обладает недостатком, отмеченным в разделе 2.1.

### 3.2. Задача кластеризации

Будем считать, что модель зависимости $m(x, y, q)$ такова, что она индуцирует на пространстве $X$ отношение эквивалентности, с которым согласуется $m$. Кластерами ПО $b = \{b_1, ..., b_u\}$ будем называть классы эквивалентности этого отношения. В этом случае число параметров $q$ зависит от числа кластеров ПО (т.к. каждый кластер ПО имеет свой набор параметров). Предполагается, что ПО характеризуется значениями параметров $q^*$.

Алгоритмом кластеризации для модели $m$ назовем такое отображение $M_m : T \to Q$, которое по обучающей выборке $t \in T$ строит отношение эквивалентности $E(t)$ на выборке $t$ (т.е. разбивает выборку $t$ на непересекающиеся подмножества – индуктивные кластеры $b' = \{b'_1, ..., b'_s)\}$, а затем вычисляет для классов эквивалентности отношения $E(t)$ (для индуктивных кластеров) значения параметров $q$.

Обозначим $PC_{M_m}(\mu)$ – условную вероятность того, что $M_m(t) = q^*$ (для дискретных значений рассматривается вероятность совпадения значений параметров, а для числовых вещественных значений – вероятность того, что они отличаются не более чем на заданное число) для обучающих выборок $t$ мощности $\mu$, при условии, что алгоритм кластеризации $M_m(t)$ устанавливает (некоторым определенным способом) взаимно-однозначное соответствие между кластерами ПО $b$ и индуктивными кластерами $b'$.

Обозначим $PC_{bb'}(\mu)$ – вероятность того, что алгоритм кластеризации $M_m(t)$ устанавливает такое взаимно-однозначное соответствие между кластерами ПО $b$ и индуктивными кластерами $b'$, при котором $PC_{M_m}(\mu)$ стремится к 1 при $\mu$, стремящемся к бесконечности.

В работе [11] рассматривается следующая новая постановка задачи кластеризации: для модели $m$ построить такой алгоритм кластеризации $M_m(t)$, что $PC_{bb'}(\mu)$ стремится к 1 при $\mu$, стремящемся к бесконечности. Очевидно, что такая постановка задачи кластеризации не обладает недостатками, отмеченными в разделе 2.2.

## 4 Проблема интерпретируемости индуктивно сформированных баз знаний

Как отметил Д. Мики в работе [2], автоматически (индуктивно) сформированные базы знаний (описания классов или кластеров) могут быть использованы в интеллектуальных системах только в том случае, если они понятны экспертам соответствующих ПО. В этом случае эксперты смогут не только сами пользоваться такими базами знаний в своей профессиональной деятельности, но и будут доверять интеллектуальным (экспертным) системам, использующим модели этих знаний, а также смогут проверять выводы таких систем, сформированные подсистемами объяснений [9].

Проведенный анализ показывает, что степень интерпретируемости баз знаний, которые формируют существующие алгоритмы обучения (решающие задачи классификации и кластеризации в их традиционных постановках) для полезных на практике ПО, не позволяет экспертам таких ПО в полной мере использовать полученные базы знаний в своей практической деятельности. Например, медицинский работник несет юридическую ответственность за принимаемые им решения (такие как диагностика, назначение лечения, проведение операций и т.д.) в отношении пациента. Поэтому степень доверия к медицинской интеллектуальной системе (в частности, к наиболее важной ее части – базе знаний), является для него принципиальным вопросом. В то же время среди бурно развивающихся в последнее время отечественных медицинских информационных систем (см. обзор в работе [6]) задача обеспечения интерпретируемости используемых в них баз знаний не ставится. А в передовых системах, таких как Watson [15], при всех их достоинствах, база знаний либо не понятна, либо скрыта и не может быть проверена пользователями, что вынуждает их принимать решения систем на свой страх и риск.

В [10] рассмотрен класс детерминированных моделей зависимости, имеющих вид необогащенных систем логических соотношений с параметрами. Если система логических соотношений является реальной онтологией ПО, полученной как результат формализации представлений экспертов об этой ПО (т.е. взятой из практики), то модель зависимости является хорошо интерпретируемой. В такой модели описанием классов или кластеров является набор значений параметров, названный базой знаний. Такое описание является хорошо интерпретируемым по построению, что является важным достоинством подобных моделей зависимости. При этом неявно предполагается, что существуют объективные значения параметров (база знаний), характеризующие ПО. Описание объектов для таких моделей зависимости имеет внутреннюю логическую структуру, одна часть которой известна

(она включает в себя значения наблюдаемых неизвестных и неинтересных параметров), другая – нет (она включает в себя значения ненаблюдаемых неизвестных и интересных параметров) [9]. Цель алгоритма обучения для таких моделей зависимости состоит в том, чтобы найти значения интересных параметров, которые вместе со значениями неинтересных параметров образуют базу знаний. Из этого следует, что качество алгоритма обучения зависит от выбранной модели зависимости.

При этом, как отмечалось выше, среди предлагаемых в литературе постановок задач классификации и кластеризации [1, 3, 4, 7, 8], не рассматриваются их специфические постановки для моделей зависимости с параметрами, которые требуют от алгоритмов обучения индуктивного формирования баз знаний, обладающих определенным уровнем качества. Поэтому актуальной проблемой является разработка алгоритмов обучения для полезных на практике, хорошо интерпретируемых и адекватных математических моделей зависимости с параметрами, являющихся реальными онтологиями ПО, и формирующих такие описания классов или кластеров (наборы значений параметров этих моделей зависимости, т.е. базы знаний), которые эксперты ПО оценивают как достаточные для решения практических задач в этих ПО.

## 5 Алгоритм индуктивного формирования баз медицинских диагностических знаний, решающий задачи классификации и кластеризации в их новых постановках

В работе [11] представлен алгоритм обучения (алгоритм индуктивного формирования баз медицинских диагностических знаний, решающий задачи классификации и кластеризации в их новых постановках) для полезной на практике, хорошо интерпретируемой и адекватной математической модели зависимости, являющейся онтологией медицинской диагностики, приближенной к реальной. Эта модель является важным частным случаем онтологии, опубликованной в работе [12].

В онтологии медицинской диагностики, приближенной к реальной, рассматривается один вид причинно-следственных отношений – клинические проявления заболеваний (классов $y$) и учитываются многократные наблюдения пациента, результаты которых зависят от времени наблюдения. Каждое заболевание обладает «клинической картиной» – набором признаков, значения которых зависят от заболевания и изменяются его клиническими проявлениями. Каждое описание клинического проявления (ОКП) заболевания по признаку может иметь несколько вариантов динамики (ВД) этого признака при этом заболевании.

Обучающие выборки $T$ представляют собой конечные совокупности описаний историй болезни (ИБ). Для каждого объекта $x_i$ обучающей выборки $t$ известно, к какому классу он принадлежит, т.е. в каждой ИБ указан диагноз, который совпадает с названием заболевания (с названием одного из классов $y_j$). В ИБ результаты наблюдений каждого признака, входящего в клиническую картину этого заболевания, согласуются с одним из ВД этого признака при этом заболевании. Объекты обучающей выборки представлены в признаковом пространстве $X$. Значением каждого признака является функция времени с конечной областью определения (конечным множеством моментов наблюдения) и конечной областью значений (подмножеством возможных значений признака).

В соответствии с [11] параметры $q$ модели ПО $m$ могут быть разделены на неинтересные и интересные. В рассматриваемой онтологии неинтересными параметрами являются: «признаки», «заболевания», «возможные значения» и «клиническая картина», а интересными параметрами являются: «нормальные значения», «варианты», «число периодов динамики», «значения для периода» («область значений»), «верхняя граница» и «нижняя граница».

Подставляя известные значения неинтересных параметров в модель ПО, получаем, что система логических соотношений с параметрами распадается на группы логических соотношений, соответствующие парам (заболевание × признак из клинической картины этого заболевания). Любые две такие группы логических соотношений не имеют общих интересных параметров. В результате такой декомпозиции исходная задача классификации в признаковом пространстве сводится к множеству задач для каждой группы соотношений. Из вида этих соотношений следует, что каждая такая задача является задачей кластеризации, а объект обучающей выборки в ней представляется функцией времени, значениями которой являются значения признака. Совокупность ОКП всех заболеваний по всем признакам (т.е. вектор значений параметров $q$ модели ПО $m$) образует базу медицинских диагностических знаний.

Решением задачи классификации для онтологии медицинской диагностики, приближенной к реальной, является совокупность решений задач кластеризации для логических соотношений с параметрами, соответствующих всем возможным парам (заболевание × признак из клинической картины этого заболевания).

Алгоритм решения задачи кластеризации для произвольной пары (заболевание × признак из клинической картины этого заболевания) сводится к последовательному решению следующих трех оптимизационных задач: обобщение всех ИБ обучающей выборки; формирование набора индуктивных кластеров; вычисление оценок значений параметров всех ВД. Данный алгоритм обучения в первую очередь ориентирован на

обработку ИБ с диагнозами, которые соответствуют заболеваниям, обладающим изменением значений признаков во времени, т.е. динамикой. Описание клинических проявлений таких заболеваний является наиболее трудоемким и сложным с точки зрения экспертов ПО, а потому наиболее востребован ими. Условием применимости алгоритма является хорошая обследованность ИБ, входящих в обучающую выборку. Если обучающая выборка содержит также и плохо обследованные ИБ, качество индуктивно сформированной базы медицинских диагностических знаний (качество описаний клинических проявлений заболеваний) зависит исключительно от доли хорошо обследованных ИБ в обучающей выборке.

## 6 Комплекс программ InForMedKB для индуктивного формирования баз медицинских диагностических знаний

Алгоритм обучения из работы [11] реализован в комплексе программ InForMedKB v1.1 (INductive FORmation of MEDical Knowledge Bases), который предназначен для решения следующих задач:

- создание онтологии (математической модели зависимости, заданной системой логических соотношений с параметрами) медицинской диагностики в структурной форме (с возможностью представления этой онтологии в виде, удобном для использования ее в качестве медицинского стандарта);
- создание обучающей выборки, состоящей из (сформированных на основе онтологии медицинской диагностики) ИБ по набору заболеваний некоторого раздела медицины (для каждого заболевания формируется своя обучающая выборка – как часть целого);
- формирование подвыборки обучающей выборки (в зависимости от ряда критериев отбора ИБ, которые могут быть: числовыми, логическими, перечислимыми);
- индуктивное формирование (на основе алгоритма обучения из [11]) базы медицинских диагностических знаний (которая представляет собой совокупность описаний клинических проявлений заболеваний, каждое из которых сформировано на основе соответствующей этому заболеванию обучающей выборки) в структурной форме, а также формирование объяснения этой базы знаний (также на основе информации из обучающих выборок);
- преобразование сформированной базы знаний в форму, принятую в медицинской литературе, понятную практикующему врачу.

Комплекс программ InForMedKB включает в себя следующие подсистемы (Рис.1):

- подсистему ввода онтологии медицинской диагностики и ИБ обучающей выборки;
- подсистему преобразования обучающей выборки во внутренний формат алгоритма обучения;
- подсистему индуктивного формирования баз медицинских диагностических знаний, реализующую алгоритм обучения;
- подсистему представления баз медицинских диагностических знаний в форме, принятой в медицинской литературе.


Рис.1. Архитектура комплекса программ InForMedKB

Подсистема ввода онтологии медицинской диагностики и ИБ обучающей выборки представляет собой структурный редактор информации и используется в комплексе программ в качестве сторонней компоненты (описание подсистемы дано в работе [5]). Онтология медицинской диагностики, приближенная к реальной, является формализацией экспертных знаний о ПО медицинской диагностики и содержит следующие разделы: дата поступления пациента, анатомо-физиологические особенности, диагноз, история настоящего заболевания, жалобы, данные объективного обследования, лабораторные и инструментальные методы исследования, сопутствующие заболевания. Каждый раздел (кроме даты поступления и диагноза) содержит структурированный набор признаков, каждый из которых может иметь числовой (в том числе быть целочисленным или вещественным интервалом), логический или перечислимый тип данных. В случае последних двух типов для признака также указывается набор его возможных значений.

Структура и внутренний формат онтологии и ИБ совпадают. Их различие состоит в том, что онтология содержит типы, интервалы и нумерованные наборы возможных значений признаков, а в ИБ указываются их номера или конкретные значения (в случае числовых значений – конкретное число или интервал, в случае логических значений – номер одного из них, в случае перечислимых значений – номера

возможных значений из онтологии). Набор ИБ, введенный при помощи данной подсистемы, образует обучающую выборку (которая представляет собой совокупность обучающих выборок для каждого заболевания), каждая ИБ которой представлена в формате XML.

Подсистема преобразования обучающей выборки во внутренний формат алгоритма обучения из [11], используя онтологию медицинской диагностики, приближенную к реальной, преобразовывает обучающую выборку во внутренний формат этого алгоритма. Ввиду того, что в данной онтологии признаки считаются независимыми (и в любой момент наблюдения любой признак имеет единственное значение), обрабатываются они также независимо, поэтому каждая обучающая выборка (содержащая ИБ по конкретному заболеванию) реорганизовывается в набор описаний тех признаков, которые наблюдались в ее ИБ. Описание каждого признака содержит его внутренний номер, количество ИБ (в которых он хотя бы единожды наблюдался), общее количество моментов наблюдения признака (суммарное количество по всем ИБ, где он наблюдался), а также таблицу со столбцами: номер ИБ, момент наблюдения признака (количество часов, прошедшее с момента начала заболевания), значение признака в момент наблюдения. После преобразования обучающей выборки на ее основе может быть сформирована подвыборка, удовлетворяющая (или нет) ряду критериев отбора ИБ, которые могут быть: числовыми (например, диапазон возрастов пациентов), логическими (например, пол пациентов) или перечислимыми (например, анатомо-физиологические особенности пациентов). На основе таких подвыборок возможно формирование специализированных баз медицинских диагностических знаний, учитывающих особенности конкретных групп пациентов.

Подсистема индуктивного формирования баз медицинских диагностических знаний, реализующая алгоритм обучения из [11], на основе обучающей выборки (совокупности обучающих выборок для каждого заболевания) во внутреннем формате алгоритма обучения, индуктивно формирует описания клинических проявлений соответствующих заболеваний. Их совокупность образует базу диагностических медицинских знаний в структурной форме, которая имеет следующий формат. Для каждого заболевания имеется набор признаков, входящих в его клиническую картину. Это означает, что каждый из таких признаков хотя бы в одном из своих периодов динамики имеет хотя бы одно значение, отличное от нормального. Описания заболеваний, входящих в базу медицинских диагностических знаний, состоят из описаний клинических проявлений признаков, входящих в клинические картины этих заболеваний. В таком описании для каждого признака указывается количество ИБ, в которых он наблюдался, число его вариантов динамики, а также описание этих вариантов. ВД соответствует типу реакций организма на данное заболевание по данному признаку. Описание ВД содержит информацию о числе периодов динамики в нем, о значениях признака в этих периодах и о границах длительности этих периодов.

При этом для каждого ВД формируется его объяснение – указывается количество ИБ, которые его поддерживают (в которых этот ВД проявился), а также номера ИБ, которые входят в определяющее подмножество данного ВД (исключение этих ИБ из обучающей выборки не позволит сформировать данный ВД). Совокупность объяснений всех сформированных ВД образует объяснение заболевания. Совокупность объяснений всех заболеваний образует объяснение базы медицинских диагностических знаний. Также описание может включать дополнительную информацию об ИБ обучающей выборки (если эта информация содержится в ней): количество часов, прошедших с момента начала заболевания до поступления пациента в клинику, была ли проведена операция (и, если да, то через сколько часов после поступления), количество дней, проведенных в клинике, и т.п.

Подсистема представления баз медицинских диагностических знаний в форме, принятой в медицинской литературе, осуществляет преобразование базы знаний, сформированной предыдущей подсистемой, на основе грамматики преобразования баз знаний из внутреннего формата алгоритма обучения в форму, принятую в медицинской литературе (при этом формальное представление базы медицинских диагностических знаний позволяет конвертировать его в любой формат, необходимый для использования в интеллектуальных системах медицинской диагностики). Предлагаемая грамматика преобразования (созданная на основе анализа описаний заболеваний, принятых в медицинской литературе) может быть изменена непосредственно в комплексе программ или же может быть задана внешним файлом, что позволяет гибко настраивать вид результата для различных групп пользователей. Преобразованная база медицинских диагностических знаний, а также используемая при ее формировании онтология медицинской диагностики, сохраняются в формате MS Word.

## 7 Применение комплекса программ и экспертная оценка результатов его работы

При помощи разработанного комплекса программ InForMedKB, на основе обучающей выборки реальных данных (содержащей 74 хорошо обследованные ИБ заболевания «Острый аппендицит») индуктивно сформирована база медицинских диагностических знаний (далее представлен ее краткий фрагмент), получившая в работе [13] экспертную оценку. Базой медицинских диагностических знаний является совокупность описаний клинических проявлений заболеваний, которые являлись диагнозами в ИБ, входящих в

обучающую выборку. Каждое заболевание обладает набором таких признаков, значения которых зависят от заболевания и изменяются его клиническими проявлениями. Каждое клиническое проявление может иметь несколько вариантов динамики, соответствующих типу реакций организма на это заболевание по признаку. Каждый вариант динамики – это последовательность периодов динамики признака, количество которых задается значением параметра «число периодов динамики» (ЧПД). Каждый период динамики характеризуется возможными значениями признака в нем, а также верхней и нижней границами его длительности.

Признак «Боль в животе (Присутствие)» наблюдался в 74 ИБ, имеет 2 варианта динамики: 1 вариант поддерживают 73 ИБ, ЧПД – 1: *имеется*. 2 вариант поддерживает 1 ИБ, ЧПД – 3: *имеется*, затем через 5 часов *отсутствует*, затем через 6 часов *имеется*.

Экспертное заключение по признаку «Боль в животе (Присутствие)»: оба варианта соответствуют знаниям, имеющимся в научной и учебной медицинской литературе.

Признак «Боль в животе (Характер)» наблюдался в 63 ИБ, имеет 7 вариантов динамики: 1 вариант поддерживают 30 ИБ, ЧПД – 2: *острая*, затем через 5-55 часов *ноющая*. 2 вариант поддерживает 21 ИБ, ЧПД – 1: *ноющая*. 3 вариант поддерживают 7 ИБ, ЧПД – 1: *острая*. 4 вариант поддерживает 1 ИБ, ЧПД – 2: *давящая*, затем через 3 часа *ноющая*. 5 вариант поддерживает 1 ИБ, ЧПД – 3: *острая*, затем через 6 часов *ноющая*, затем через 11 часов *острая*. 6 вариант поддерживают 2 ИБ, ЧПД – 2: *ноющая*, затем через 4 часа *острая*. 7 вариант поддерживает 1 ИБ, ЧПД – 3: *ноющая*, затем через 3 часа *острая*, затем через 4 часа *ноющая*.

Экспертное заключение по признаку «Боль в животе (Характер)»: варианты 1-4 соответствуют знаниям, имеющимся в медицинской литературе; варианты 5 и 6 являются редко встречающимися в литературе; вариант 7 является нетипичным для острого аппендицита, что требует дополнительных исследований. Можно предположить, что в ИБ, поддерживающих вариант 6, были пропущены наблюдения в одном из периодов динамики (если в первом, то варианты 6 и 5 являются одним вариантом, а если в третьем, то одним вариантом являются варианты 6 и 7).

Признак «Клинический анализ крови (Лейкоциты)» наблюдался в 67 ИБ, имеет 14 вариантов динамики: 1 вариант поддерживают 16 ИБ, ЧПД – 1: *выраженный лейкоцитоз*. 2 вариант поддерживают 16 ИБ, ЧПД – 1: *умеренный лейкоцитоз*. 3 вариант поддерживают 5 ИБ, ЧПД – 2: *выраженный лейкоцитоз*, затем через 7-15 часов *умеренный лейкоцитоз*. 4 вариант поддерживают 2 ИБ, ЧПД – 2: *умеренный лейкоцитоз*, затем через 22 часа *выраженный лейкоцитоз*. 5 вариант поддерживает 1 ИБ, ЧПД – 2: *умеренный лейкоцитоз*, затем через 7 часов *выраженный лейкоцитоз*. 6 вариант поддерживает 1 ИБ, ЧПД – 2: *умеренный лейкоцитоз*, затем через 9 часов *выраженный лейкоцитоз*. 7 вариант поддерживает 1 ИБ, ЧПД – 2: *норма*, затем через 5 часов *умеренный лейкоцитоз*. 8 вариант поддерживает 1 ИБ, ЧПД – 2: *норма*, затем через 19 часов *умеренный лейкоцитоз*. 9 вариант поддерживает 1 ИБ, ЧПД – 2: *норма*, затем через 34 часа *умеренный лейкоцитоз*. 10 вариант поддерживает 1 ИБ, ЧПД – 2: *норма*, затем через 6 часов *выраженный лейкоцитоз*. 11 вариант поддерживают 19 ИБ, ЧПД – 1: *норма*. 12 вариант поддерживает 1 ИБ, ЧПД – 2: *умеренный лейкоцитоз*, затем через 27 часов *норма*. 13 вариант поддерживает 1 ИБ, ЧПД – 2: *умеренный лейкоцитоз*, затем через 48 часов *норма*. 14 вариант поддерживает 1 ИБ, ЧПД – 1: *выраженная лейкопения*.

Экспертное заключение по признаку «Клинический анализ крови (Лейкоциты)»: варианты 1-10 соответствует знаниям, имеющимся в медицинской литературе; варианты 11-13 не характерны для острого воспалительного заболевания, но имеют место; вариант 14 является нетипичным для острого аппендицита и говорит о плохой сопротивляемости организма бурно развивающемуся болезненному процессу (данный вариант сформирован по ИБ с диагнозом «флегмонозный аппендицит»). Можно предположить, что варианты 4-6 являются одним вариантом динамики, аналогично, варианты 7-9, аналогично, варианты 12 и 13. Объединение соответствующих вариантов возможно при увеличении объема обучающей выборки.

По общему мнению эксперта работы [13], полученная база медицинских диагностических знаний представлена в форме, понятной практикующему врачу – так, как зачастую она бывает представлена в медицинской литературе. При этом сформированное описание заболевания «Острый аппендицит» соответствует знаниям, имеющимся в научной и учебной медицинской литературе, а в ряде случаев дополняет их описанием динамики клинических проявлений.

## 8 Заключение

Предложенный в работе метод индуктивного формирования баз медицинских диагностических знаний позволяет повышать уровень организации и качества деятельности в области медицинской диагностики, за счет внедрения качественных и хорошо интерпретируемых баз медицинских диагностических знаний, сформированных на основе обучающих выборок большого объема (а также подтвержденных и объясненных реальными историями болезни). Формальное представление таких баз знаний, сформированных комплексом программ InForMedKB, позволяет использовать их в системах медицинской диагностики.

## Литература

[1] MachineLearning.ru Профессиональный информационно-аналитический ресурс, посвященный машинному обучению,

распознаванию образов и интеллектуальному анализу данных http://machinelearning.ru/

[2] Michie D. Expert Systems // The Computer Journal, 1980, Vol. 23, No. 4, pp. 369-376.

[3] Вагин В.Н., Головина Е.Ю. Достоверный и правдоподобный вывод в интеллектуальных системах / Под ред. В.Н. Вагина, Д.А. Поспелова. М.: ФИЗМАТЛИТ, 2004. 704 с.

[4] Витяев Е.Е. Извлечение знаний из данных. Компьютерное познание. Модели когнитивных процессов: Моногр. / Новосибирск: Новосиб. гос. ун-т, 2006. 293 с.

[5] Грибова В.В., Тарасов А.В., Черняховская М.Ю. Система интеллектуальной поддержки обследования больных, управляемая онтологией // Программные продукты и системы. Тверь: НТП Фактор, 2007. №2. С. 49-51.

[6] Гусев А.В. Медицинские информационные системы в России: текущее состояние, актуальные проблемы и тенденции развития // Информационные технологии в медицине, 2011-2012., М.: «Радиотехника», 2012.

[7] Донской В.И. Алгоритмические модели обучения классификации: обоснование, сравнение, выбор. Симферополь: ДИАЙПИ, 2014. 228 с.

[8] Загоруйко Н.Г. Когнитивный анализ данных. Новосибирск: Академическое издательство «ГЕО», 2012. 203 с.

[9] Клещев А.С. Задачи индуктивного формирования знаний в терминах непримитивных онтологий предметных областей // Научно-техническая информация. Серия 2. М.: ВИНИТИ РАН, 2003. № 8. С. 8-18.

[10] Клещев А.С., Артемьева И.Л. Необогащенные системы логических соотношений. В 2 Ч. // Научно-техническая информация. Серия 2. М.: ВИНИТИ РАН, 2000: № 7. С. 18-28, № 8. С. 8-18.

[11] Клещев А.С., Смагин С.В. Задачи индуктивного формирования знаний для онтологии медицинской диагностики // Научно-техническая информация. Серия 2. М.: ВИНИТИ РАН, 2012. №1. С. 9-21.

[12] Клещев А.С., Черняховская М.Ю., Москаленко Ф.М. Модель онтологии предметной области «Медицинская диагностика». В 2 Ч. // Научно-техническая информация. Серия 2. М.: ВИНИТИ РАН, 2005-2006. №12, №2.

[13] Коктышева Г.А., Петряева М.В., Смагин С.В. Экспертный анализ индуктивно сформированной базы знаний заболевания «острый аппендицит» // XXXVIII Дальневосточная математическая школа-семинар имени академика Е.В. Золотова, 1-5 сентября 2014 г., Владивосток: сб. материалов [электронный ресурс]. Владивосток: ИАПУ ДВО РАН, 2014. С. 397-403.

[14] «Полигон алгоритмов» – распределенная система тестирования алгоритмов классификации на данных реальных прикладных задач [http://poligon.machinelearning.ru/]

[15] Роб Хай. Эпоха когнитивных систем: Принцип построения и работы IBM Watson http://www.redbooks.ibm.com/redpapers/pdfs/redp4955-ru.pdf

## Method for Inductive Formation of Vedical Diagnostic Knowledge Bases

Sergey V. Smagin

The paper provides an introduction into the area of inductive formation of knowledge bases and highlights the current topical problems, including the interpretability of the results. A method is suggested for inductive formation of easily interpretable medical diagnostic knowledge bases. It includes the new definitions of classification and clustering problems for dependence model with parameters, the learning algorithm (solving mentioned problems in their new definitions) is developed for the practically useful and easily interpretable mathematical dependence model with parameters, which is a near real-life ontology of medical diagnostics (defined by a system of logical relationships with parameters). Also included is the software package InForMedKB (INductive FORmation of MEDical Knowledge Bases), which implements this mentioned learning algorithm.