различия в принципе построения морфологической информации к словам. В алгоритме И.А. Мельчука используется словарь основ и флексий, а в разрабатываемом синтаксическом анализаторе используется словарь с восстановленной парадигмой слов.

Разработан и программно реализован журнал операций для тестирования работоспособности блоков автоматического синтаксического анализа, так как осуществление тестирования программных блоков по отдельности очень сложно. Выбраны методы программирования алгоритмов и способы хранения данных, задействованных в работе программы.

СПИСОК ЛИТЕРАТУРЫ

1. Большаков И.А., Гельбух А.Ф. Модель «Смысл – Текст» тридцать лет спустя // Диалог 99: Труды Междунар. семинара. – М., 1999 – Т. 1. – С. 15–24.
2. Мельчук И.А. Автоматический синтаксический анализ. – Новосибирск: Изд-во АН СССР, 1965. – 358 с.
3. Андреев А.М., Березкин Д.В., Брик А.В., Кантонистов Ю.А. Вероятностный синтаксический анализатор для информационно-поисковой системы [Электронный ресурс]. – режим доступа: http://www.inteltec.ru/publish/articles/textan/1kx5_9.shtml. – 16.03.2006.
4. Волкова И.А., Мальковский М.Г., Одинцев Н.В. Адаптивный Синтаксический анализатор // Диалог 2003: Труды Междунар. семинара. – М., 2003. – Т. 1. – С. 401–406.

Поступила 09.04.2009 г.

УДК 681.3.068+681.5

ВЫЯВЛЕНИЕ СКРЫТЫХ ЗАКОНОМЕРНОСТЕЙ В СЛОЖНЫХ СИСТЕМАХ

О.Г. Берестнева, Я.С. Пеккер

Томский политехнический университет
Сибирский государственный медицинский университет
E-mail: ogb@tpu.ru; pekker@ssmu.ru

Рассмотрены основные подходы к решению задачи выявления скрытых закономерностей в сложных медико-биологических системах. Показаны особенности решения такой задачи в случае количественных и качественных экспериментальных данных. Представлена технология выявления скрытых закономерностей на основе методов Data Mining.

Ключевые слова:
Компьютерный анализ данных, скрытые закономерности, биосистемы, деревья решений.

Key words:
Computer analysis of data, hidden patterns, biological systems, decision trees.

Введение

Центральной проблемой медико-биологических исследований, независимо от их специфики, является оценка состояния биологической системы. При этом под термином «биологическая система» мы понимаем организм в целом или его часть любой степени сложности на клеточном, органном уровне или уровне функциональной системы [1]. Эти же особенности относятся к анализу как биотехнических систем, так и любых сложных систем, имеющих большое количество каналов взаимодействия с внешней средой и характеризующихся существенной стохастической составляющей функционирования.

На современном уровне исследования биологических систем имеет смысл говорить именно о системном подходе, который получил развитие в трудах И.Р. Пригожина, П.К. Анохина, К.В. Судакова, У. Эшби и других выдающихся ученых.

Начиная со второй половины XX в., биология и медицина стремительно отходят от вербального описания и все больше тяготеют к формализации процессов, происходящих в биосистемах, использованию математических моделей и технических средств, а позднее и компьютерных технологий [2, 3]. Однако, это сопряжено с колоссальными сложностями, особенно при оценке динамических характеристик поведения биосистем и процесса их адаптации.

Значительные трудности изучения количественных характеристик биологических систем предопределяются особенностями и свойствами последних, и, прежде всего [1]: структурной и функциональной сложностью; вариабельностью параметров для одного состояния; нелинейностью характеристик; невозможностью полного описания системы. Последняя особенность («ущербность» описания) существенно осложняет постро-

ение моделей поведения, оценки ее состояния и принятия решений.

Как правило, исследователя интересует: либо реакция системы на те условия, в которых она функционирует, либо при известной реакции системы необходимо выявить некие закономерности функционирования системы, определяющие ее реакцию. Как в первом, так и во втором случае требуются нетрадиционные формализованного описания систем, которые и рассмотрены в данной работе.

Моделирование процессов адаптации биосистемы

Биологические системы существуют, приспосабливаются и развиваются благодаря обмену энергией, веществом и информацией с внешней средой. Такие системы называют открытыми. Одно из наиболее перспективных направлений в исследовании таких систем основано на концепции энтропии. В отличие от классической физики, где рассматриваются детерминированные и обратимые процессы, в открытых системах, и это характерно для живых систем, может наблюдаться устойчивое неравновесие.

Используя техническую терминологию, можем отметить, что биологическая система, функционируя и взаимодействуя с внешней средой, «отрабатывает» каждое экзогенное и эндогенное воздействие. Наблюдаются процессы, связанные с флуктуацией, диссипациацией энергии, обменом веществом и информацией и, наконец, формированием новых динамических состояний, которые мы привыкли называть процессами адаптации.

Традиционные методы исследования адаптационных характеристик человека позволяют оценить динамику отдельных параметров организма и в условиях влияния большого числа неучтенных факторов, вариабельности параметров в «норме», невысокой точности неинвазивных методик дают достоверный результат лишь при значительных грубых отклонениях. Учесть характер и тенденцию изменения состояния организма, как правило, не удается, либо представляется на уровне вербальных качественных заключений.

Для получения количественных характеристик процесса адаптации нами введены энтропийные показатели состояния биосистемы [1, 4], которые позволяют оценивать не абсолютные значения физиологических (или любых других) характеристик состояния организма, а тенденцию их изменения под воздействием внешних факторов или условий.

Данный подход был использован авторами при решении различных прикладных задач: оценки состояния адаптированности нефтяников в условиях вахты [1, 5] и студентов-первокурсников в условиях «вхождения» в учебный процесс [6]; оценки состояния организма на основе анализа результатов функциональных проб [4, 7]; диагностики состояния новорожденных в раннем неонатальном периоде [4, 8]; слежения за динамикой состояния организма человека в послеоперационном периоде [9]. При этом было введено понятие «адаптационная стратегия» на основе анализа вида функции $I_{адапт}(t)$, представляющей собой значения интегрального показателя в дискретные промежутки времени. Основные типы выявленных адаптационных стратегий представлены на рис. 1.

Проведенные нами исследования [1, 4–7] показали эффективность применения энтропийных методов моделирования сложных систем для анализа особенностей адаптации биосистем. Вместе с тем применение данного подхода имеет ряд ограничений: 1) переменные состояния системы должны быть количественными и не коррелированными; 2) необходимо иметь данные о поведении системы за определенный промежуток време-

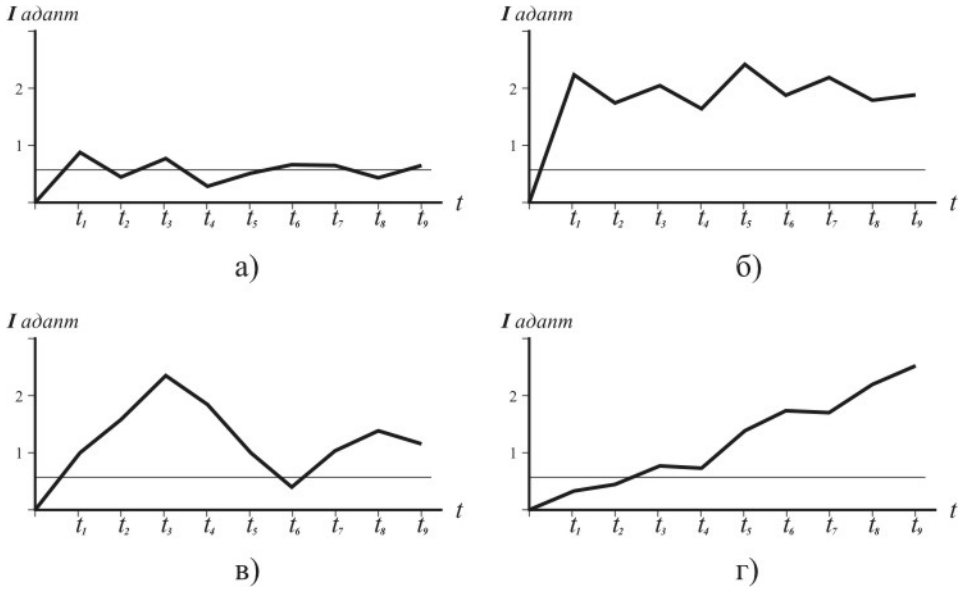

Рис. 1. *Типы адаптационных стратегий: а) адаптивный; б) компенсаторный; в) адаптивно-компенсаторный; г) дезадаптивный*

ни (динамические наблюдения). Таким образом, в тех случаях, когда для измерения переменных состояния системы используются не только количественные, но и номинальная и/или ранговая шкала, требуются другие подходы для выявления закономерностей функционирования системы.

В случае, когда выходное состояние системы характеризуется некоторой номинальной переменной (например, «исход беременности») могут быть использованы методы интеллектуального анализа данных (*Data Mining),* например, методы построения продукционных моделей.

Выявление скрытых закономерностей на основе продукционных моделей

Продукционные модели близки к логическим моделям, что позволяет организовывать на них эффективные процедуры вывода, а с другой стороны, более наглядно отражают знания, чем классические логические модели. В них отсутствуют жесткие ограничения, характерные для логических исчислений, что дает возможность изменять интерпретацию элементов продукции.

Продукционная модель, или модель, основанная на правилах, позволяет представить знания в виде предложений типа *Если* (*условие*), *то* (*действие*).

В качестве инструментария для построения продукционных моделей могут быть использованы: система *WizWhy* [9] (реализующая метод логического перебора Бонгарда [8]) и система *See5* [9] (построение деревьев решений).

Алгоритмы ограниченного перебора были предложены в середине 60-х гг. прошлого века М.М. Бонгардом для поиска логических закономерностей в данных. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей. Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Данный подход был использован авторами для выявления скрытых закономерностей функционирования сложных систем (в задачах психологии и медицины) [10, 11].

Остановимся более подробно на технологии построения **деревьев решений**. Деревья решений являются наиболее распространенным в настоящее время подходом к выявлению и визуализации логических закономерностей в данных. Чаще всего используются дихотомические деревья, когда из вершины выходит только две ветви. Каждому узлу сопоставлен некоторый признак, а ветвям – либо конкретные значения для качественных признаков, либо области значений для количественных признаков. Например, в случае решения задачи прогнозирования исхода беременности к количественным переменным относятся рост, вес, результаты лабораторных анализов, показатели динамометрии, артериального давления; к качественным переменным – анкетные данные; к ранговым – результаты некоторых психологических тестов [11].

Дерево решений позволяет построить модель линейной зависимости вида

$$Y = f(X),$$

где X – множество характеристических признаков, а Y – множество исходов.

При построении дерева решений должно соблюдаться требование непротиворечивости – на пути, ведущем из корня в лист, не должно быть взаимоисключающих значений.

Дерево решений может быть переведено в набор логических высказываний. Каждое высказывание получается при прохождении пути из корневой вершины в лист и представляет собой логическую закономерность исследуемого явления.

Качество дерева характеризуют два основных показателя: точность и сложность дерева. Точность дерева показывает, насколько хорошо разделены объекты разных классов.

В качестве показателя сложности дерева выступают такие характеристики как: число листьев дерева, число его внутренних вершин, максимальная длина пути из корня в конечную вершину и др. Показатели сложности и точности взаимосвязаны: чем сложнее дерево, тем оно, как правило, точнее.

На первом этапе на вход алгоритма поступает некоторое количество обучающих примеров (объектов), где каждый объект описывается набором характеристических признаков (в дальнейшем также разделяющие признаки) и классифицирующим признаком, который задает принадлежность к одному из диагностических классов [11].

Корню дерева соответствует самый информативный характеристический признак. Далее, в вершинах располагаются признаки в порядке уменьшения значений прироста информативности. В качестве меры информативности узла используется энтропия. Рассмотрим этот процесс подробнее.

Пусть имеется множество T объектов, разделенных по значениям классифицирующего признака на полные непересекающиеся классы $C_1,C_2,...,C_k$ (классифицирующий признак может принимать k возможных значений), тогда информация, необходимая для идентификации класса, есть

$$Info(T)=I(P),$$

где P – вероятность распределения классов $(C_1,C_2,...,C_k)$:

$$P = (p1, p2, ..., pk) = \left(\frac{|C_1|}{|T|}, \frac{|C_2|}{|T|}, ..., \frac{|C_k|}{|T|} \right),$$

а $I(P)$ – энтропия, вычисляемая по формуле:

$$I(P)=-(p1\log(p1)+p2\log(p2)+...+pk\log(pk)).$$

Основанием логарифма в указанных выше формулах служит 2.

Информация, необходимая для идентификации класса при условии, что нам известно значение

разделяющего (характеристического) признака X, считается как [10]:

$$Info(X,T)=\sum_{i=1}^{m}\left(\frac{|T_i|}{|T|}*Info(T_i)\right),$$

где T_i – одно из возможных значений разделяющего признака X; m – количество значений разделяющего признака; $Info(T_i)$ – информация для каждого значения разделяющего признака.

Тогда величина, характеризующая прирост информативности $Gain(X,T)$, может быть определена как:

$$Gain(X,T)=Info(T)-Info(X,T).$$

Прирост информативности представляет собой разницу между информацией, необходимой для идентификации класса, и информацией, необходимой для идентификации класса при условии, что нам известно значение признака X. При использовании обучающей выборки с неполным набором информации вычисление коэффициента прироста признака производится только по признакам с определенными значениями.

Понятие «прирост информации» необходимо для ранжирования характеризующих признаков при построении дерева решений. Каждый новый узел, включаемый в дерево решений, располагается так, что он приносит наивысший прирост информативности из всех разделяющих признаков, еще не включенных в путь к корню.

При последующих ветвлениях может возникнуть ситуация, когда вероятностное распределение разделяющего признака D представляет собой (1, 0). Тогда $Info(D,T)=0$ и $Gain(D,T)$ максимален. Чтобы это компенсировать вместо коэффициента $Gain$ используется следующий коэффициент:

$$GainRatio(D,T)=\frac{Gain(D,T)}{SplitInfo(D,T)},$$

где

$$SplitInfo(D,T)=I\left(\frac{|T_1|}{|T|},\frac{|T_2|}{|T|},\ldots,\frac{|T_m|}{|T|}\right),$$

а $\{T_1,T_2,\ldots,T_m\}$ – подмножества T_i, порождаемые делением множества объектов в соответствии со значениями признака D.

Если взят качественный признак, то при вычислении коэффициента прироста информации используется каждое значение. Количественный признак требует предварительных разбиений на некоторые градации или интервалы. Рассмотрим как это происходит.

Пусть признак C_i – количественный. Возможные значения признака сортируются в порядке возрастания: $A_1,A_2,\ldots,A_m$, затем для каждой величины A_j (j=1,2,...,m) записи разделяются на те, которые имеют значения до A_j включительно, и те, которые имеют значения большие A_j. Для каждого из полученных подмножеств вычисляется прирост или коэффициент прироста информации. В итоге выбирается деление на подмножества с максимальным коэффициентом прироста. Полученное пороговое значение подлежит проверке или уточнению в ходе дальнейших исследований.

Вершина относится к бесперспективным для последующего ветвления в случае, если объекты обучающей выборки для данной вершины однородны (принадлежат одному диагностическому классу), или число объектов достаточно мало (порог на число наблюдений задается в качестве входного ограничивающего параметра алгоритма).

Усечение дерева решений производится путем замещения целого поддерева узловым листом. Замещение имеет место только в том случае, если ожидаемый показатель ошибки в поддереве больше, чем в одиночном листе.

На первом шаге вычисляется количество ошибок $E_0(t)$ в поддереве с корнем в вершине t:

$$E_0(t)=\sum_{l=1}^{L}\sum_{\substack{\omega=1\\ \omega\neq Y(t)}}^{K}U_l^{\omega},$$

где L – количество вершин, на которые разделилась вершина t; K – количество диагностических классов, которые соответствуют вершине t; $\omega\neq Y(l)$ – дополнительное условие, соответствующее тому, что не рассматриваются решения (классы), которые были присвоены предыдущим вершинам; U_l^{ω} – число объектов класса ω, которые соответствуют l-ой вершине.

На втором шаге подсчитывается количество ошибок $E_1(t)$, которое будет допущено, если поддерево будет преобразовано в лист.

Затем вычисляется выигрыш $G(t)=E_0(t)-E_1(t)$.

Вершина, имеющая большое значение выигрыша, подлежит усечению. Подробно процедура отсечения рассмотрена в [10].

Каждое правило, выводимое системой, характеризуется величинами (n/m, *lift x*): n – количество объектов, соответствующих данному правилу; m – количество объектов, не принадлежащих данному диагностическому классу (ошибочное распознавание); *lift x* – уровень доверия к построенному правилу.

Для того, чтобы улучшить качество классификации, распознавания и прогнозирования, а также для получения устойчивых закономерностей (под устойчивостью авторы понимают повторение результатов) может быть использована процедура построения леса деревьев решений.

Деревья могут быть получены разными методами (или одним методом, но с различными параметрами работы), по разным выборкам. Результаты классификации и прогнозирования по каждому построенному дереву будут различаться. Для построения коллективной классификации и прогнозирования используется метод голосования, т. е. объекту приписывается тот класс, которому отдает предпочтение большинство деревьев из набора [9].

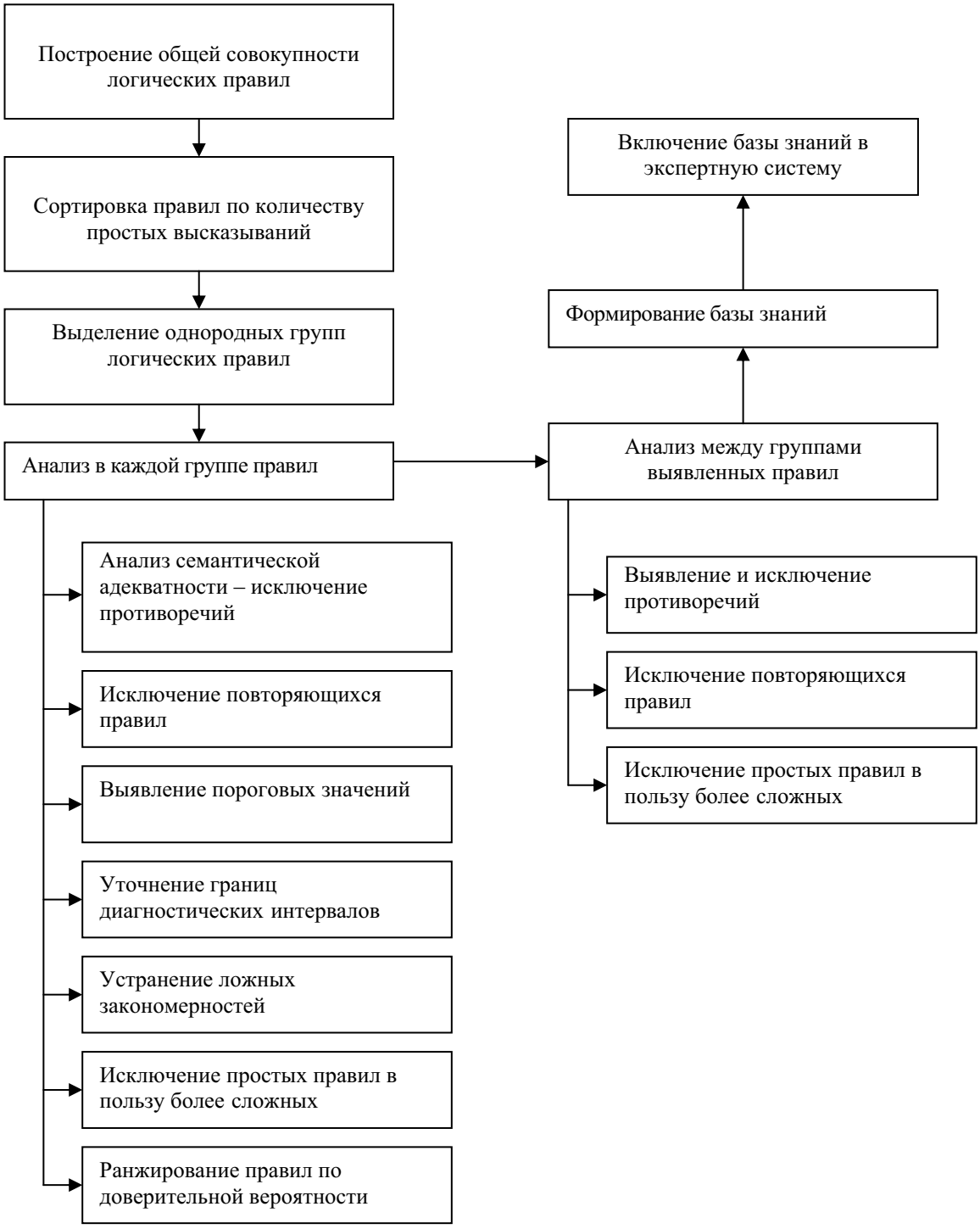

Рис. 2. *Схема выявления устойчивых закономерностей*

Поскольку результаты, полученные как в системе *WizWhy*, так и в системе *See5* представляют собой набор логических правил, для выявления устойчивых «устойчивых» закономерностей может быть использована разработанная ранее (совместно с Е.А. Муратовой [11]) технология выявления скрытых закономерностей (рис. 2).

Применение данной технологии при решении задачи исхода беременности позволило выявить некоторые раннее неизвестные социально-психологические закономерности, характерные как для благоприятного, так и для неблагоприятного исхода беременности [11–13].

Выводы

Для решения задач выявления и исследования закономерностей в зависимости от особенностей решаемой задачи и типа переменных состояния системы могут быть использованы два основных подхода: определение адаптационных стратегий поведения системы на основе энтропийных методов и технология выявления скрытых закономерностей на основе методов *Data Mining*.

Применение технологии выявления скрытых закономерностей в виде продукционных моделей продемонстрировало эффективность ее использования в области медико-биологических, когнитивных, социологических и психолого-педагогических исследований.

Работа частично поддержана грантами РФФИ (проект 08-06-00313а) и РГНФ (проект 07-06-12143в).

СПИСОК ЛИТЕРАТУРЫ

1. Ротов А.В., Медведев М.А., Пеккер Я.С., Берестнева О.Г. Адаптационные характеристики человека. – Томск: Изд-во Томского гос. ун-та, 1997. – 137 с.
2. Айдаралиев А.А., Баевский Р.М. Комплексная оценка функциональных резервов организма. – Фрунзе: Илим, 1988. – 190 с.
3. Функциональное состояние человека и методы его исследования: Сб. науч. тр. РАН / Под ред. М.В. Фролова. – М.: Наука, 1992. – 123 с.
4. Берестнева О.Г., Гергет О.М., Шаропин К.А. Моделирование адаптационных стратегий организма человека // Интеллектуальные системы (IEEE AIS'04)» и «Интеллектуальные САПР» (CAD-2004): Труды Международных научно-технических конференций. – М.: Физматлит, 2004. – Т. 2. – С. 236–240.
5. Ротов А.В., Берестнева О.Г., Пеккер Я.С. Оценка функционального состояния организма человека с применением интегральных критериев энтропийного типа // Сибирский психологический журнал. – 1996. – Вып. 2. – С. 68–69.
6. Берестнева О.Г., Карпов Г.А., Пеккер Я.С. Оценка функционального состояния беременных женщин по данным ортостатической пробы // Медико-биологические аспекты нейро-гуморальной регуляции. – 1994. – Вып. 3. – С. 4–6.
7. Берестнева О.Г., Цхай В.Ф., Пеккер Я.С. Применение интегральных критериев для оценки послеоперационного состояния при механических желтухах паразитарной природы // Медико-биологические аспекты нейро-гуморальной регуляции. – 1994. – Вып. 3. – С. 8–9.
8. Бонгард М.М. Проблема узнавания. – М.: Наука, 1967. – 320 с.
9. Дюк В., Эммануэль В. Информационные технологии в медико-биологических исследованиях. – СПб.: Питер, 2003. – 528 с.
10. Бериков В.Б. Анализ статистических данных с использованием деревьев решений. – Новосибирск: Изд-во НГТУ, 2002. – 59 с.
11. Муратова Е.А., Берестнева О.Г. Выявление скрытых закономерностей в социально-психологических исследованиях // Известия Томского политехнического университета. – 2003. – Т. 306. – № 3. – С. 97–102.
12. Берестнева О.Г., Добрянская Р.Г., Муратова Е.А. Применение методов Data Mining для формирования базы знаний экспертной системы прогнозирования исходов родов // 10-я Национ. конф. по искусственному интеллекту с международным участием (КИИ-2006): Труды конф. – М., 2006. – Т. 1. – С. 244–248.
13. Холодная М.А., Берестнева О.Г., Муратова Е.А. Структура стратегий совладания в юношеском возрасте (к проблеме валидности опросника «Юношеская копинг-шкала» // Вопросы психологии. – 2007. – № 4. – С. 143–156.

Поступила 12.10.2009 г.