ИНФОРМАТИКА

2008 **июль-сентябрь** **№ 3**

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

УДК 004:519.22

И.Э. Том, О.В. Красько, Н.А. Новоселова

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ АНАЛИЗА МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ В ДЕТСКОЙ ЛЕЙКОЗОЛОГИИ

Рассматриваются две информационные технологии, реализованные в виде соответствующих программных комплексов и предназначенные для анализа многомерных медико-биологических данных: первая – система анализа данных, реализующая комплекс разнообразных статистических и нейросетевых методов на основных стадиях обработки данных; вторая – система классификации групп риска пациентов на основе гибридной нечеткой классификационной модели, обеспечивающей построение нескольких подмножеств Парето-оптимальных компактных и хорошо интерпретируемых классифицирующих правил.

Введение

Одной из важнейших медицинских задач, на которую ориентировано большинство разрабатываемых в мире методов и технологий анализа данных, является диагностика заболеваний. Диагностика в медицине представляет собой традиционную задачу классификации, методы решения которой интенсивно разрабатываются во многих областях исследований, таких как распознавание образов, искусственный интеллект и др. Для медицинских задач классификации объектом данных является пациент, предсказывающими или диагностическими являются признаки, которые присутствуют в данных пациента (эпидемиологические сведения, симптомы и клиническая картина заболевания, результаты лабораторных тестов и др.) и которые требуется выделить из всей совокупности разнородных данных пациента, а класс представляет собой диагноз (заболевание, клиническое состояние пациента, группу риска и т. п.). Целью классификации является определение метки класса для некоторого объекта данных, описываемого значениями нескольких диагностических признаков (переменных). Построение классификационной модели заключается в нахождении отображения $F: X_1 \times X_2 \times \ldots \times X_n \to \Omega$, оптимального относительно некоторого критерия *J*, где $\Omega = \{C_1, C_2, \ldots, C_M\}$ – множество возможных меток классов, X_1, X_2, …, X_n – множество значений диагностических признаков, а каждый пациент *p* описывается *n*-мерным вектором признаков $x_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$ или так называемым индивидуальным профилем факторов, где $x_{pi} \in X_i$.

Ниже рассматриваются две программные системы, разработанные в Объединенном институте проблем информатики (ОИПИ) НАН Беларуси и предназначенные для анализа медицинских данных: программный комплекс анализа прогностических факторов (ПК АПФ) – система анализа данных, реализующая комплекс разнообразных статистических и нейросетевых методов на основных стадиях обработки данных – от редуцирования размерности и предобработки до классификации и прогнозирования; ПК «Гибрид» – система классификации пациентов на основе гибридной нечеткой классификационной модели, более специализированная и предназначенная для построения нескольких подмножеств Парето-оптимальных компактных и хорошо интерпретируемых медиками классифицирующих правил, получаемых на основе верифицированного обучающего множества X_1, X_2, … , X_n прецедентных клинических, лабораторных и эпидемиологических данных пациентов.

Особо следует отметить, что разработка указанных систем была бы невозможна без специалистов Республиканского научно-практического центра детской онкологии и гематологии (РНПЦДОГ), возглавляемого профессором О.В. Алейниковой, в частности профессора М.П. Потапнева (в настоящее время директора РНПЦ гематологии и трансфузиологии) и доцента Т.А. Угловой, которые принимали самое деятельное участие в постановке задач, анализе,

проверке работоспособности и оценке эффективности применения как отдельных методов анализа, так и программных комплексов в целом.

1. Информационно-аналитическая система анализа прогностических факторов риска и прогнозирования исхода индукционной терапии лейкозов у детей. ПК АПФ

Создание информационной технологии анализа многомерных медицинских онкогематологических данных на основе совместного использования комплекса статистических, интеллектуальных нейросетевых методов являлось основной целью авторов и их коллег. Разработанная информационная технология была применена к разведочному анализу прецедентных медицинских данных в детской онкогематологии. Практические цели анализа: повышение обоснованности идентификации группы риска больного по его индивидуальному профилю прогностических факторов риска (ПФР) и оценка прогноза результативности индукционной терапии детей с острыми лейкозами на самых ранних этапах лечения.

Задача создания информационной технологии решалась в рамках Государственной программы прикладных исследований «Создание новых оптико-электронных систем и информационных технологий» по заданию «Разработать информационную технологию разведочного анализа прецедентных медицинских данных на основе нейросетевых и статистических методов для применения в детской онкогематологии». Предметом исследования являлись статистические и нейросетевые методы анализа разнородных многомерных медицинских данных, которые интегрировались в единую информационную технологию анализа.

В результате исследований был создан макет ПК АПФ, который представляет собой специализированный медико-ориентированный комплекс программных средств статистического и нейросетевого анализа многомерных медицинских данных, являющийся составной частью информационно-аналитической системы сбора лабораторных, клинических и персональных данных о больных острыми лейкозами, хранения и выдачи их по запросам для последующего анализа. Эта система была разработана специалистами РНПЦДОГ и ОИПИ НАН Беларуси в рамках совместного выполнения международного проекта МНТЦ В-522, посвященного созданию компьютерной системы анализа прогностических факторов риска для выбора адекватной терапии острых лейкозов у детей (рис. 1).

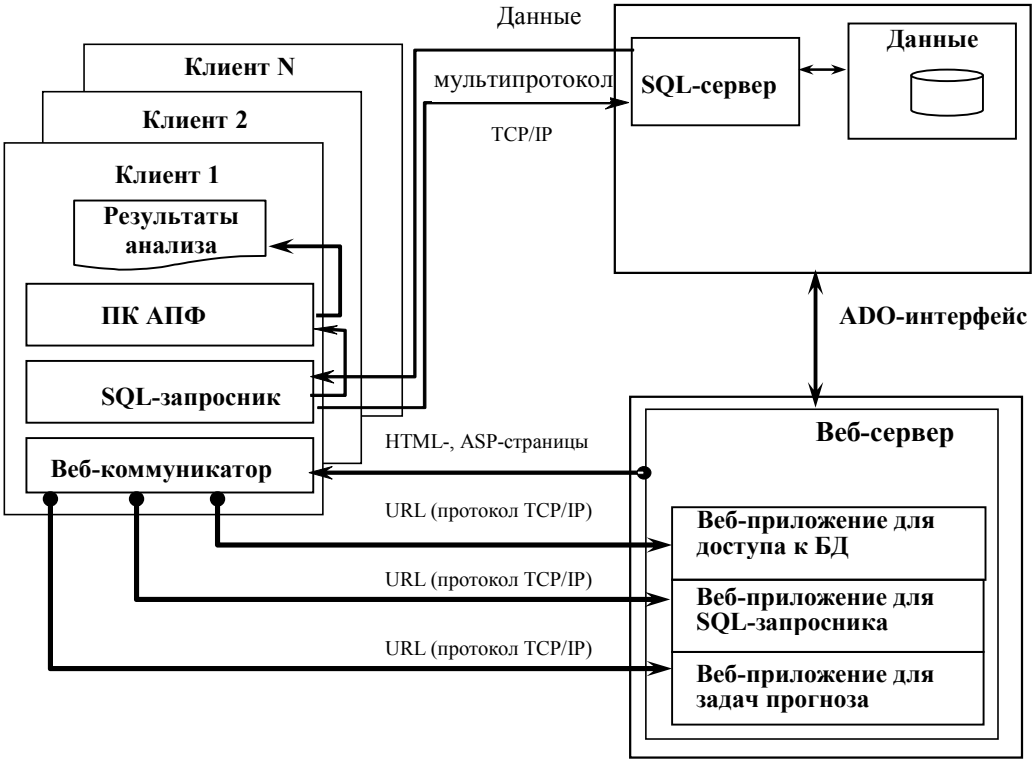

Рис. 1. Структура компьютерной системы анализа прогностических факторов риска

Комплексное использование в ПК АПФ статистических и интеллектуальных методов анализа данных объясняется необходимостью взаимной компенсации ограничений, присущих этим двум методологиям. Так, нейросетевые методы обладают интеллектуальным свойством обучения и высокой классификационной эффективностью, менее чувствительны к форме анализируемых данных, но при этом практически исключают возможность объяснения, почему получен тот или иной результат при классификации или прогнозировании. И наоборот, результаты, достигаемые статистическими методами (в случае возможности построения точной математической модели классификации), всегда хорошо интерпретируются и поэтому незаменимы при решении многих важных задач, например при отборе значимых переменных и т. п. Вместе с тем недостатком статистических методов является необходимость предварительных знаний о свойствах данных. При решении реальных медицинских задач классификации в большинстве случаев приходится работать с разнородными многомерными данными, имеющими большое число пропусков, в условиях отсутствия информации о законах распределения данных или когда имеющиеся данные не подчиняются нормальному закону распределения. Поэтому процесс построения точной математической модели решения задачи классификации является зачастую практически нереализуемым. Кроме того, имеющаяся у медиков информация о заболевании обычно представлена в форме неполных эмпирических знаний и набора входных/выходных данных о поведении исследуемых биологических объектов. Поэтому для решения таких классификационных задач, по убеждению авторов, необходимо использовать интеллектуальные методы, которые способны обрабатывать такую не полностью определенную информацию, и дополнять их по возможности результатами соответствующих, хорошо себя зарекомендовавших статистических методов. Консилиум из статистических и интеллектуальных нейросетевых методов на каждом из этапов анализа данных, которые реализованы в ПК АПФ, обеспечивает более высокий уровень доверия медиков (конечных пользователей) к результатам решения задачи классификации больных по группам риска или прогнозирования исхода терапии, чем при использовании только статистических или только нейросетевых методов.

В процессе выполнения задания была разработана и реализована в виде макета программного обеспечения (ПК АПФ) медико-ориентированная четырехэтапная информационная технология (рис. 2) анализа онкогематологических данных, характеризующихся разнородностью шкал измерений, большим числом (свыше 150) клинико-лабораторных и эпидемиологических переменных, небольшим объемом обучающей и тестовой выборок. Для каждого из следующих этапов информационной технологии анализа реализован комплекс статистических и нейросетевых методов, призванных выполнять роль консилиума:

Этап 1. *Понижение размерности клинико-лабораторных данных* и их предобработка с помощью методов однофакторного анализа (t- и χ^2-критериев), главных компонент в классической и нейросетевой постановке, множественного анализа соответствий, комбинированной оценки энтропии и χ^2-критерия, дискретизации и кодирования непрерывных переменных.

Этап 2. *Отбор комбинаций наиболее информативных прогностических факторов* путем использования комплекса пошаговых статистических методов отбора: пошагового алгоритма отбора с добавлением в задачах линейной регрессии, алгоритма полного перебора в задачах квадратичной регрессии, дискриминантного анализа на основе расстояния Махаланобиса (пошагового с добавлением и полным перебором).

Этап 3. *Верификация отобранных комбинаций* путем классификации наблюдений тестового множества с использованием линейного дискриминантного анализа, нейросети типа «многослойный персептрон» с различными алгоритмами обучения и различной нормализацией, нейросети типа «самоорганизующаяся сеть Кохонена» для задачи классификации.

Этап 4. *Идентификация группы риска больного по его индивидуальному профилю* прогностических факторов риска и прогнозирование результативности индукционной терапии с помощью алгоритмов множественной регрессии с экспоненциальной и логистической функциями, нейросети типа «многослойный персептрон» для задач регрессии, классификационных CART-деревьев (деревьев решений) с автоматическим и интерактивным вариантами построения.

Кроме программной реализации указанных выше методов, часть из которых применяется в других пакетах обработки данных, разработан и программно реализован ряд новых методов. В частности, для этапов предобработки и отбора комбинаций ПФР реализован комбинированный метод оценки энтропии и χ^2-критерия, для этапа верификации – алгоритм обучения классифицирующей нейронной сети с использованием метода Гаусса – Зейделя, для этапа прогнозирования – интерактивный вариант классификационного CART-дерева. Следует отметить, что при создании ПК АПФ изначально была заложена идеология последовательного наращивания его функциональных возможностей за счет включения в состав новых, более совершенных методов и алгоритмов анализа (см., например, ПК «Гибрид», описываемый ниже), что позволяет надеяться на получение уникального по своим функциональным возможностям программного продукта, способного конкурировать или даже превосходить лучшие зарубежные системы анализа непрерывных многомерных медико-биологических данных. Разработанные методы, алгоритмы и их программное обеспечение представлены более чем в 20 научных публикациях как в нашей стране, так и за рубежом (см., например, [1–3]).

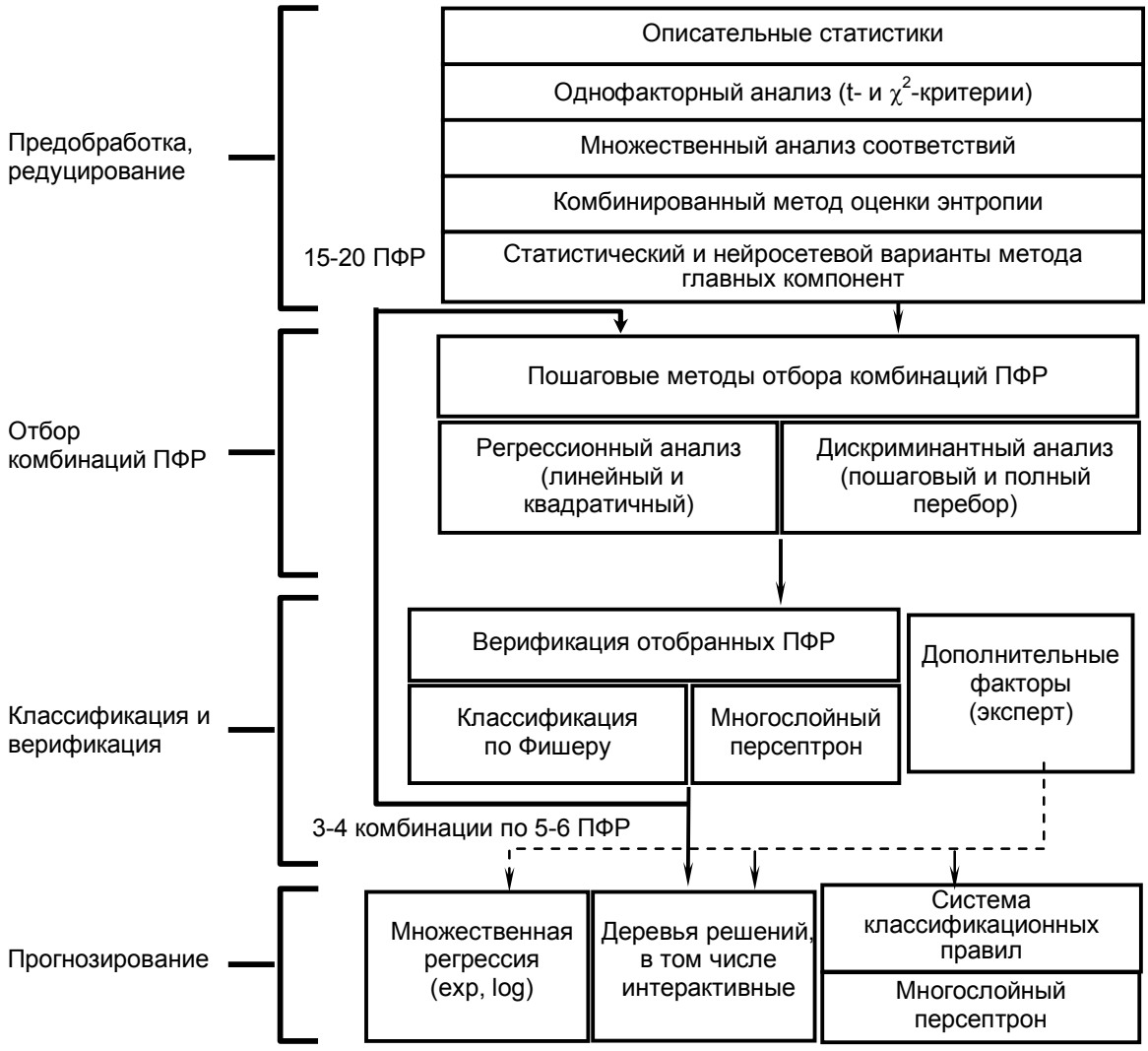

Рис. 2. Этапы информационной технологии анализа данных

Макет ПК АПФ апробирован в РНПЦДОГ и принят для опытного применения при решении задач стратификации групп риска детей с острыми лейкозами и прогнозирования исхода индукционной терапии. Информационная технология прошла проверку в детской клинике Высшей медицинской школы г. Ганновера (Германия). В процессе апробации системы в РНПЦДОГ на реальных данных удалось получить новые, нигде ранее не опубликованные в ми-

ровой медицинской литературе, но верифицированные в РНПЦДОГ медицинские результаты (новые прогностические факторы риска на этапе индукционной терапии, а также комбинированные факторы риска для прогноза раннего ответа организма пациентов на терапию и выхода в ремиссию [4]), представляющие значительный интерес для медиков-гематологов (клиницистов) при лечении острых лейкобластных и миелобластных лейкозов у детей.

2. Информационная технология анализа медико-биологических данных на базе гибридной нечеткой модели классификации. ПК «Гибрид»

Дальнейшие работы по анализу медико-биологических данных авторы развивали в направлении исследований современных интеллектуальных методов (нечеткой логики, генетических алгоритмов, нейросетей и нейросетевых алгоритмов обучения) анализа медицинских данных для повышения достоверности классификации и прогнозирования исхода терапии в детской лейкозологии.

Примером исследований в области развития интеллектуальных средств анализа медико-биологических данных является НИОКР «Разработать информационную технологию анализа медицинских данных на базе гибридных нейросетевых моделей для задач классификации и прогнозирования в детской лейкозологии», выполнявшаяся в ОИПИ НАН Беларуси в рамках Государственной научно-технической программы «Передовые информационные и телекоммуникационные технологии». Цель работы – разработка и внедрение современных интеллектуальных методов анализа медицинских данных для повышения достоверности классификации и прогнозирования исхода терапии в детской лейкозологии.

Разработанная информационная технология в виде ПК «Гибрид» предназначена для анализа лабораторных, клинических, эпидемиологических и других медицинских данных пациентов и генерирования на основе такого анализа комплекса эффективных классифицирующих правил, которые обладают компактностью и хорошо интерпретируются медиками-клиницистами.

Первоочередным местом применения разработанной технологии являются онкогематологические учреждения Беларуси, которые при выборе протоколов лечения сталкиваются с необходимостью решения задач классификации пациентов по группам риска и прогнозирования исхода терапии. Возможно ее применение и в других областях медицины для решения задач классификации и прогнозирования, где требуется получить не только точные, но и интерпретируемые решения. Так, в настоящее время проводятся работы по применению разработанной информационной технологии анализа данных для дифференциальной диагностики подтипов транзиторных ишемических атак, а также в ортопедии и травматологии. В конце 2006 г. подписан договор о передаче прав на использование ПК «Гибрид» в РНПЦ гематологии и трансфузиологии Министерства здравоохранения Беларуси.

ПК «Гибрид» реализован на базе гибридной нечеткой модели (ГНМ) классификации (рис. 3). Комплекс разработанных методов и алгоритмов построения ГНМ позволяет использовать прецедентную информацию, накопленную в базах данных, обучающий набор верифицированных медицинских данных и знания медиков-экспертов для формирования начальной структуры ГНМ, оптимизации структуры и настройки параметров модели. В конечном итоге построенная ГНМ классификации генерирует набор классификационных лингвистических правил, характеризующихся не только высокой классификационной эффективностью (до 96,79 % на тестовом наборе WBC из международного архива данных по машинному обучению [5]), но и по своей структуре являющихся компактными и легко интерпретируемыми. Процесс построения ГНМ-классификации базируется [6–8]:

– на кластерном алгоритме инициализации ГНМ, позволяющем в автоматическом режиме выделить определенное число кластеров в данных и построить на основе полученных кластеров начальный набор нечетких правил путем проектирования многомерных кластеров на пространство значений отдельных переменных с последующей аппроксимацией полученного набора точек с использованием различных видов функций принадлежности: треугольной, трапециевидной, гауссовой;

– структурно-ориентированном методе инициализации начальной структуры ГНМ-классификации, позволяющем использовать имеющийся набор данных для построения соот-

ветствующего набора начальных правил с применением специально разработанной процедуры наращивания количества правил по мере поступления очередного объекта данных;

– специально разработанном эвристическом методе адаптивного обучения параметров ГНМ-классификации, рассчитывающем значения модификаций параметров отдельных функций принадлежности с учетом ограничений, позволяющих сохранять интерпретируемость результирующего набора правил после обучения;

– методе оптимизации структуры ГНМ на основе многокритериального генетического алгоритма, обеспечивающем поиск наборов нечетких правил, оптимальных согласно нескольким критериям. Отличительной особенностью данного метода является возможность получения нескольких недоминируемых наборов нечетких правил, которые оптимальны относительно трех целевых функций: точности классификации, количества нечетких правил в наборе и общей длины нечетких правил. Конечному пользователю предоставляется возможность среди полученных наборов правил выбрать наиболее подходящий с точки зрения соотношения точности классификации и интерпретируемости.

Рис. 3. Обобщенная схема этапов построения ГНМ

Для сравнения предложенного подхода к извлечению набора классифицирующих правил в виде ГНМ классификации с наиболее эффективными зарубежными методами (нечетким генетическим подходом, методом деревьев решений C4.5, методом декомпозиции для задачи классификации, генетическим алгоритмом для отбора оптимального по двум критериям набора правил [9–12]) был использован тестовый набор данных по раку молочной железы (WBC). Результаты сравнения приведены в таблице.

Таблица

Точность классификации набора данных WBC с использованием различных методов

Исследуемые методы	Точность классификации на тестовом множестве, %	Количество правил
C4.5 [9]	94,90	> 4
Нечеткий генетический подход [11]	96,88	4 (средняя длина правила 3,44)
Метод декомпозиции [10]	95,30	> 4
Генетический отбор [12]	95,59	4 (средняя длина правила 2,06)
ГНМ (ПК «Гибрид»)	**96,79**	**4** (средняя длина правила **2,28**)

Из таблицы видно, что использование ГНМ классификации для набора данных WBC позволяет в трех случаях из четырех получить более точные результаты классификации и во всех четырех случаях – более компактную структуру конечного набора нечетких правил классификации. Тестирование ПК «Гибрид» на данных международных архивов и на реальных медицинских данных, собранных в РНПЦДОГ, показало высокую эффективность предложенных методов анализа.

Заключение

Применение ПК АПФ в медицинской практике позволило выявить новые комбинации прогностических факторов риска при лечении детских лейкозов, что способствовало повышению обоснованности принимаемых решений в процессе индукционной терапии, улучшению качества жизни пациентов после терапии и более рациональному расходованию средств на лечение. Применение ПК «Гибрид» при проведении испытаний предложенных алгоритмов на наборах медицинских данных из архива по машинному обучению показало идентичную или даже более высокую точность классификации по сравнению с другими интеллектуальными методами генерирования набора классифицирующих правил при одновременном получении более простой и компактной структуры нечеткого классификатора. Классификационные правила, построенные для реального набора данных, были проанализированы специалистами в области лейкозологии, которые констатировали их высокую прогностическую значимость и согласованность с медико-биологическими представлениями.

В ходе дальнейших исследований предполагается осуществить разработки:

– алгоритмов модификации гибридной нечеткой модели классификации, учитывающих возможность дообучения структуры и настройки параметров модели по мере поступления новых данных;

– многоуровневого генетического алгоритма построения структуры и обучения гибридной нечеткой модели классификации, позволяющего объединить два заключительных этапа общей схемы построения классификатора с получением нескольких недоминируемых наборов нечетких правил классификации;

– специализированной программной системы поддержки принятия решения при постановке диагноза, стратификации групп риска и решении других классификационных задач.

Список литературы

1. Технология анализа медицинских данных статистическими и нейросетевыми методами / И.Э. Том [и др.] // Искусственный интеллект. – 2004. – № 2. – С. 398–402.

2. Новоселова, Н.А. Предварительный отбор информативных признаков для улучшения точности предсказания с помощью нейронной сети / Н.А. Новоселова // Искусственный интеллект. – 2004. – № 2. – С. 150–154.

3. Development of the information analytical system for childhood oncohematology / I.E. Tom [et al.] // Annual Proceedings of Medical Science. – 2005. – Vol. 50, suppl. 2. – P. 43–45.

4. Development of the computer-based system for prognostic risk factors analysis to select adequate therapy of childhood acute leukemias (ISTC Project B-522) // UIIP NASB [Electronic resource]. – 2004. – Mode of access: http://itk1.bas-net.by/b522/pr_result.htm. – Date of access: 01.08.2008.

5. UC Irvine Machine Learning Repository // University of California, Irvine, CA [Electronic resource]. – 2007. – Mode of access: http://itk1.bas-net.by/b522/pr_result.htm. – Date of access: 01.08.2008.

6. Новоселова, Н.А. Нечеткое нейросетевое моделирование для получения интерпретируемого набора классифицирующих правил / Н.А. Новоселова, И.Э. Том, О.В. Красько // Искусственный интеллект. – 2006. – № 2. – С. 211–214.

7. Новоселова, Н.А. Построение нечеткой нейросетевой модели для решения задач классификации / Н.А. Новоселова // Информатика. – 2006. – № 3. – С. 5–14.

8. Новоселова, Н.А. Построение нечеткой модели классификации с использованием многокритериального генетического алгоритма / Н.А. Новоселова // Искусственный интеллект. – 2006. – № 3. – С. 613–622.

9. Elomaa, T. General and efficient multisplitting of numerical attributes / T. Elomaa, J. Rousu // Machine Learning. – September, 1999. – Vol. 36, № 3. – P. 201-244.

10. Guan, S.U. Class decomposition for GA-based classifier agents – A Pittapproach / S. U. Guan, F. Zhu // IEEE Trans. on Systems, Man and Cybernetics, Part B: Cybernetics. – February 2004. – Vol. 34, №1. – P. 381–392.

11. Ishibuchi, H. Multi-objective evolutionary design of fuzzy rule-based systems / H. Ishibuchi, T. Yamamoto // Proc. of International Conference on Systems, Man and Cybernetics (IEEE SMC 2004), The Hague, The Netherlands, October 10-13, 2004. – The Hague, 2004 – Vol. 3. – P. 2362– 2367.

12. Pena-Reyes, C.A. A fuzzy-genetic approach to breast cancer diagnosis / C.A. Pena-Reyes, M. Sipper // Artificial Intelligence in Medicine. – 1999. – Vol. 17. – P. 131–155.

Поступила 21.02.08

Объединенный институт проблем
информатики НАН Беларуси,
Минск, Сурганова, 6
e-mail: tom@newman.bas-net.by

I.E. Tom, O.V. Krasko, N.A. Novoselova

INFORMATION TECHNOLOGIES FOR MEDICOBIOLOGICAL DATA ANALYSIS OF CHILDHOOD LEUKEMIA

Two information technologies for multivariate medico biological data analysis are considered. The first technology is a data analysis system, realizing the complex of different statistical and neural network methods for the main stages of data processing. The second one is a system for classification of patients risk groups on the basis of hybrid fuzzy classification model, providing the construction of several Pareto-optimal subsets of compact and interpretable classification rules.