4. Steinbach, M. C. Hierarhical Sparsity in Multistage Convex Stochastic Programs [Text] / M. C. Steinbach // Konrad-Zuse-Zentrum fur Informationstechnik. - Berlin, ZIB-Report, 2000.
5. Steinbach, M. C. Tree-Sparse Convex Programs [Text] / M. C. Steinbach // Konrad-Zuse-Zentrum fur Informationstechnik. - Berlin, ZIB-Report, 2001.
6. Hoeve, W.-J. Operations Research Techniques in Constraint Programming [Text] / W.-J. Hoeve // Institute for Logic, Language and Computation Universiteit van Amsterdam, 2005. - 154 p.
7. Steinbach, M. C. General Information Constraints in Stochastic Programs [Text] / M. C. Steinbach. // Berlin, ZIB, 2001.
8. Евдокимов, А. Г. Потокораспределение в инженерных сетях [Текст] / А. Г. Евдокимов, В. В.Дубровский, А. Д. Тевяшев. - М. : Стройиздат, 1979. - 199 с.
9. Евдокимов, А. Г. Оперативное управление потокораспределением в инженерных сетях [Текст] /А. Г. Евдокимов, А. Д.Тевяшев. – Харьков. : Вища школа, 1980. - 144 с.
10. Тевяшев, А. Д. Применение линеаризованных моделей установившегося потокораспределения в задачах оперативного управления [Текст] / А. Д. Тевяшев, С. И. Козыренко // Новые информационные технологии управления развитием и функционированием трубопроводных систем энергетики, 1993. – С. 20 - 33.
11. Вентцель, Е. С. Теория вероятностей [Текст] / Е. С. Вентцель. - М. : Наука, 1969. – 564 с.

Розглянуто новий підхід до обробки даних медико-біологічних досліджень з використанням методів обчислювального інтелекту. Особливістю цього підходу є нечутливість методу до співвідношення кількості об'єктів до кількості показників, що ці об'єкти характеризують і нечутливість до закону розподілу даних. Запропонований підхід дозволяє проводити обробку даних при заздалегідь відомій і невідомій кількості об'єктів

Ключові слова: обчислювальний інтелект, нейронна мережа, кластер, центроїд, ступінь належності

Рассмотрен новый подход к обработке данных медико-биологических исследований с использованием методов вычислительного интеллекта. Особенностью этого подхода является то, что метод не чувствителен к соотношению количества объектов к количеству характеризующих эти объекты показателей и нечувствителен к закону распределения данных. Предложенный подход подразумевает обработку данных при заранее известном и неизвестном количестве объектов

Ключевые слова: вычислительный интеллект, нейронная сеть, кластер, центроид, степень принадлежности

УДК 615.471:616-071

АДАПТИВНАЯ ОБРАБОТКА ДАННЫХ МЕДИКО-БИОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ МЕТОДАМИ ВЫЧИСЛИТЕЛЬНОГО ИНТЕЛЛЕКТА

И. Г. Перова
Кандидат технических наук,
старший преподаватель
Кафедра биомедицинской инженерии
Харьковский национальный университет
радиоэлектроники
пр. Ленина, 14, г. Харьков, Украина, 61166
E-mail: churyumova@mail.ru

1. Введение

Применение в медицинской диагностике интеллектуальных методов обработки данных в настоящее время получило широкое распространение. Это объясняется необходимостью оказания помощи врачу на этапе анализа количественной, а также правильно интерпретированной качественной информации о состоянии пациента. Практически любое серьезное исследование в настоящее время сопровождается сложными математическими вычислениями и анализом. Именно по этой причине предлагается метод адаптивной обработки данных медико-биологических исследований на основе методов вычислительного интеллекта, состоящий из этапа предварительной обработки, этапа компрессии данных и этапа разделения на однородные группы, так называемого этапа кластеризации данных.

2. Постановка проблемы

К настоящему времени существует большое количество методов разделения области данных на однородные группы. Эти методы могут работать, основываясь на информации, полученной из обучающей выборки либо без наличия таковой, так называемые

методы кластеризации данных. В данной статье рассматриваются только методы кластеризации данных, поскольку в такой области как обработка данных медико-биологических исследований часто существует проблема составления корректной обучающей выборки и существует ряд медицинских задач, для которых нет смысла затрачивать большое количество времени на сбор подобных данных. Именно для таких задач и предлагается использовать методы кластеризации данных. Они разнообразны по принципам разделения объектов на классы, но в основном эти методы носят «четкую» природу, то есть не позволяют кластерам пересекаться в пространстве признаков [1]. Неотъемлемой частью методов кластеризации является стандартизация, нормировка и компрессия исходных данных и методы, которые для этого используются, не должны искажать исходные данные и не должны устранять их возможную нелинейность. Поэтому целью статьи является создание нового подхода к обработке данных медико-биологических исследований, который как раз учитывает все вышеперечисленные особенности данных.

3. Литературный обзор

Принцип работы методов кластеризации заключается в разделении объектов на группы, основываясь на степени «похожести» одного объекта на другой [1]. Благодаря этому врач еще до постановки диагноза получит информацию об имеющихся «сгустках» в данных, которые образуют плотные группы (кластеры).

Типичный пример эвристического алгоритма для параллельной обработки данных – алгоритм k-эталонов – приводит автор [2]. Основной идеей является тот факт, что совокупность объектов, находящихся на одинаковом расстоянии от каждого из k-эталонов, образуют компактную группу.

В качестве примера эвристического алгоритма для последовательной обработки данных автор [3] предлагает использовать простой алгоритм, в основе которого лежит предположение, что представители одного класса не могут быть удалены друг от друга более чем на заданную пороговую величину. Исходная информация представлена в форме матрицы «Объект-Свойство». Параметрами алгоритма является функция близости между объектами и пороговое значение этой функции. Алгоритмы этой группы являются достаточно примитивными и мало используются. Для выработки первых представлений о структуре данных в [2] предлагает использовать набор быстродействующих алгоритмов, использующих понятие центра тяжести применительно к параллельной обработке данных. К алгоритмам такого типа относится известный алгоритм Форель [4] и алгоритм k-средних (k-means algorithm) [5], на основе которого строится целое семейство алгоритмов. Единственным управляющим параметром этого алгоритма является r – радиус шаров, которыми покрывается выборка X.

Рассмотренные выше алгоритмы кластеризации имеют один общий существенный недостаток. Они обеспечивают так называемую «четкую» кластеризацию, то есть области не могут пересекаться и накладываться. Но по той причине, что медицинские выборки данных имеют особенность «пересекаться» в пространстве признаков за счет необходимости персонального подхода к каждому человеку в отдельности, такая кластеризация на границе разделения кластеров будет неточной и «грубой». Выходом из создавшегося положения может служить применение искусственных нейронных сетей (ИНС) для кластеризации данных, поскольку ИНС позволяют определить любую нелинейность данных. Но хоть ИНС и способны на сложное разделение объектов на группы, все равно они при этом являются четкими системами, соответственно кластеризация, проведенная с помощью ИНС, будет носить четкий характер.

Основным недостатком рассмотренных выше методов является ограничение на форму кластеров, в основном это гипершары. Поскольку данные медико-биологических исследований имеют достаточно сложную структуру, то использование для их кластеризации гипершаров не обеспечивает необходимую точность формирования однородных кластеров. То есть необходимо искать новые подходы, которые будут обеспечивать адаптацию формы кластеров под имеющиеся у исследователя данные.

4. Адаптивная обработка данных медико-биологических исследований

В случае применения методов кластеризации для обработки данных медико-биологических исследований схема формирования диагноза имеет вид, представленный на рис. 1.

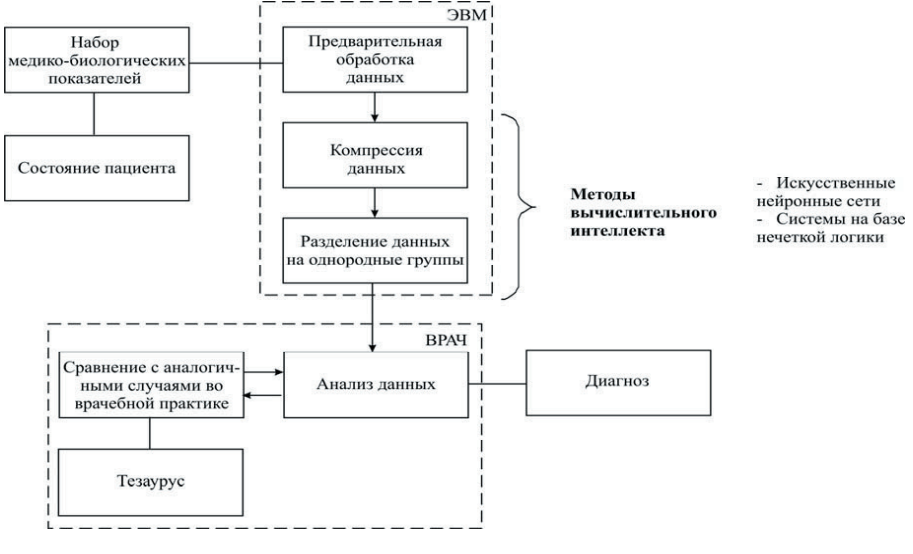

Рис. 1. Обобщенная схема формирования диагноза с использованием методов вычислительного интеллекта

Схема (рис. 1) предусматривает три этапа при обработке данных медико-биологических исследований: сбор данных, их обработка и принятие решения. Рассмотрим подробнее процесс обработки, который состоит из предварительной обработке, компрессии данных и разделения на однородные группы (т. е. кластеризации).

В качестве исходной информации используется набор медико-биологических показателей, характеризующих состояние пациента. Эти показатели образуют таблицу «Объект-Свойство», в которой строки соответствуют множеству объектов (пациенты), а столбцы – множеству свойств этих объектов (медико-биологические показатели).

Этап предварительной обработки данных включает в себя их формализацию данных, то есть сопоставление каждому значению признака соответствующего кодового числа. Это предлагается осуществить на основе медианы с целью придания процедуре робастных свойств (защита от аномальных наблюдений и возможных артефактов) с помощью рекуррентного соотношения:

$$me_i(k)=me_i(k-1)+\eta_m(k)\,sign\left(x_i(k)-me_i(k-1)\right),\ i=1,2,...,\ n, \quad (1)$$

где $\eta_m(k)$ – параметр шага поиска, выбираемый в стационарном случае в соответствии с условиями Дворецкого [6, 8].

Таким образом, мы получаем таблицу «Объект-Свойство», в которой максимальному значению признака соответствует 1, минимальному -1, а промежуточные значения находятся внутри гиперкуба [-1,1].

В качестве примера подобного кодирования можно рассмотреть таблицу «Объект-Свойство» до кодирования (табл. 1) и после него (табл. 2).

Таблица 1

Данные про пациентов до кодирования

№	Пол	Рост	Вес	Сист	Диаст	Пульс
1	м	156	50	90	60	72
2	м	168	60	105	60	60
3	м	170	53	110	65	78
4	м	170	63	120	70	64
5	м	172	56	110	50	60
6	м	171	57	120	70	60
7	ж	168	52	110	70	65
8	м	171	64	160	100	52

Таблица 2

Данные про пациентов после кодирования относительно гиперкуба [-1;1]

№	Пол	Рост	Вес	Сист	Диаст	Пульс
1	-1	-0,64	-0,68	-0,57	-0,41	-0,40
2	-1	-0,38	-0,47	-0,36	-0,41	-0,57
3	-1	-0,33	-0,62	-0,29	-0,29	-0,31
4	-1	-0,33	-0,41	-0,14	-0,18	-0,51
5	-1	-0,29	-0,56	-0,29	-0,65	-0,57
6	-1	-0,31	-0,54	-0,14	-0,18	-0,57
7	1	-0,38	-0,64	-0,29	-0,18	-0,50
8	-1	-0,31	-0,39	0,43	0,53	-0,69

Второй частью этапа обработки данных является этап компрессии данных или сокращение размерности входного пространства признаков, которое производится с помощью автоассоциативной нейронной сети типа Bottle Neck [9], схема обучения которой представлена на рис. 2.

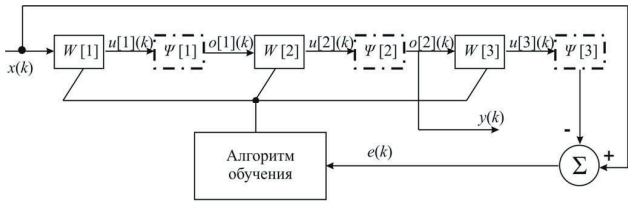

Рис. 2. Обучение автоассоциативного трехслойного персептрона типа Bottle Neck

В процессе обучения минимизируется целевая функция вида:

$$E^k=\sum_k E(k)=\frac{1}{2}\sum_k\sum_{j=1}^{n}\left(x_j(k)-o_j^{[3]}(k)\right)^2=\frac{1}{2}\sum_k\sum_{j=1}^{n}e_j^2(k). \quad (2)$$

На нулевой слой поступает n-мерный вектор входных сигналов $x(k)$ $(n_0=n)$, первый скрытый слой содержит $n_1=n$ нейронов, второй скрытый слой – $n_2<n$ нейронов и выходной слой – $n_3=n$ нейронов. Целью обучения является восстановление на выходе сети сигнала $o^{[3]}(k)$, наилучшим образом аппроксимирующего входной сигнал $x(k)$. Сжатый сигнал $y(k)=o^{[2]}(k)$ снимается с выхода второго слоя и при этом достигается оптимальное решение задачи нелинейного факторного анализа [10].

В результате на выходе скрытого слоя получим нелинейно сжатый входной сигнал, выраженный в необходимом количестве признаков.

Для кластеризации предлагается использовать адаптивный алгоритм нечеткой кластеризации данных для пакетной или последовательной обработки информации [11 – 13]. Исходной информацией является выборка наблюдений, сформированная и N n-мерных векторов признаков $X=\{x(1),x(2),...,x(N)\}$, $x(k)\in X$, k=1,2,...,N. Результат работы метода представляет собой разбиение исходного массива данных на m классов с некоторым уровнем $w_j(k)$ принадлежности k-того вектора признаков j-му кластеру. Целевая функция, подлежащая минимизации имеет вид:

$$E\left(w_j(k),c_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}w_j^{\beta}(k)d^2\left(x(k),c_j\right)\to\min, \quad (4)$$

при ограничениях:

$$\sum_{j=1}^{m}w_j(k)=1,\ k=1,...,\ n,\ 0<\sum_{k=1}^{N}w_j(k)<N,\ j=1,...,\ m.$$

Здесь $w_j(k)\in[0,1]$ – уровень принадлежности вектора x(k) к j-му кластеру, c_j - центроид j-го кластера, $d^2\left(x(k),c_j\right)$ – расстояние между x(k) и c_j в принятой метрике, β – неотрицательный параметр, именуемый «фаззификатором» (в случае использования в качестве $d^2\left(x(k),c_j\right)$ евклидового расстояния, принимается равным 2).

Работа алгоритма начинается с задания начальной случайной матрицы нечеткого разбиения W_0. В соответствии с ее значениями высчитывается начальный набор центров-прототипов c_j^0, согласно формуле

$$c_j = \frac{\sum_{k=1}^{N} w_j^{\beta}(k) x(k)}{\sum_{k=1}^{N} w_j^{\beta}(k)}. \quad (5)$$

Наоснованиирассчитанныхцентров-прототипов c_j^0 далее вычисляется матрица W^1 согласно формуле

$$w_j(k) = \frac{\left(d^2\left(x(k), c_j\right)\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(d^2\left(x(k), c_l\right)\right)^{\frac{1}{1-\beta}}}. \quad (6)$$

После этого в пакетном режиме пересчитываются $c_j^1, W^2, ..., W^t, c_j^t, W^{t+1}$ и так далее до тех пор, пока разность между нынешними и последующими значениями матрицы W не станет меньше заданного порога точности. Таким образом, вся имеющаяся выборка данных обрабатывается многократно.

В результате работы алгоритма получим матрицу нечеткого разбиения, в которой пациенты будут разделены на кластеры (диагнозы). Форма кластеров может меняться от гипершара до гиперэллипсоида в зависимости от формы исходных данных, то есть от выбора расстояния между x(k) и c_j:

$$d\left(x(k), c_j\right) = \sqrt{\left(x(k) - c_j\right)^T A_j \left(x(k) - c_j\right)}, \quad (7)$$

где A_j – матрица, которая может быть определена как обратная нечеткая ковариационная матрица каждого кластера.

Если в качестве матрицы A_j возьмем единичную матрицу, то в результате получим эвклидово расстояние $d\left(x(k), c_j\right) = \sqrt{\left(x(k) - c_j\right)^T \left(x(k) - c_j\right)}$, и форма кластеров будет округлая (гипершары).

Для придания кластерам формы гиперэллипсоидов в качестве матрицы A_j можно использовать симметрическую положительно определенную матрицу, т.е. матрицу, у которой все собственные значения являются действительными и положительными и $A_j = F_j^{-1}$, где

$$F_j = \frac{\sum_{j=1}^{m} w_j^{\beta}(k)\left(x(k) - c_j\right)\left(x(k) - c_j\right)^T}{\sum_{j=1}^{m} w_j^{\beta}(k)}. \quad (8)$$

Последовательная обработка данных востребована в случае, когда при кластеризации большого массива данных появляются дополнительные данные, которые необходимо присоединить к этой же таблице «Объект-Свойство». В случае большого объема таблицы, повторная кластеризация всех данных может занять достаточно большое время. Соответственно, появляется необходимость создания процедуры, которая бы обеспечивала работу алгоритма в реальном времени с возможностью постоянного добавления новых данных в таблицу «Объект-Свойство», которая имеет такой вид:

$$w_j(k) = \frac{\left(d^2\left(x(k), c_j(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}\left(d^2\left(x(k), c_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \quad (9)$$

$$c_j(k+1) = c_j(k) - \eta(k)\nabla_{c_j} L_k\left(w_j(k), c_j(k), \lambda(k)\right) =$$

$$= c_j - \eta(k) w_j^{\dagger}(k) d\left(x(k+1), c_j(k)\right) \nabla_{c_j} d\left(x(k+1), c_j(k)\right), \quad (10)$$

где $\eta(k)$ - параметр скорости обучения, $c_j(k)$ – центры-прототипы j-го кластера, вычисленные на выборке из k наблюдений, ∇_{c_j} – (n x 1)-вектор-градиент.

В результате работы алгоритма кластеризации мы получаем разделение наших данных на однородные кластеры, которые могут иметь форму произвольно ориентированных в пространстве гиперэллипсоидов и способны пересекать в пространстве признаков (рис. 3). Также в результате работы алгоритма будет известна степень принадлежности каждого из объектов к каждому из кластеров $w_j(k)$.

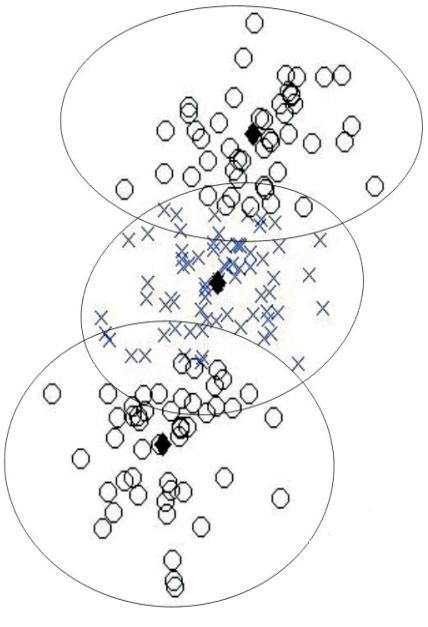

Рис. 3. Пример расположения пересекающихся кластеров

Примеры применения метода нечеткой кластеризации данных к реальным медицинским данным и данным репозитория Калифорнийского университета рассмотрены в работах [12 – 14]. Однако в данных работах в качестве метода компрессии данных использовался метод главных компонент (PCA- analysis). Поскольку отличием предлагаемого в данной статье подхода является использование в качестве метода сокращения размерности входных признаков нейронной сети, то при повторении экспериментов в новой интерпретации наблюдается повышение точности разделения объектов на классы. В частности. при использовании предлагаемого подхода к данным, которые описывают выживаемость пациентов при болезнях сердца (представлены в открытом доступе в репозитории университета Калифорнии). Каждое наблюдение описывается 10 параметрами, один из которых является атрибутом класса: класс 1 – пациент умер

48 %, класс 2 – пациент выжил 52 %. В итоге осталось 9 значимых параметров. В работе [14] точность подобной кластеризации была невысока, ошибка составила 13,36 %. При проведении полной обработки данных с помощью предлагаемой процедуры, включающей 3 этапа, было получено повышение точности разбиения данных на классы. Ошибка кластеризации составила 10,12 %, что свидетельствует о большей универсальности метода.

5. Выводы

Предложенная адаптивная обработка медико-биологических данных обеспечит врача необходимой информацией о степени близости объектов друг к другу, о форме распределения данных в пространстве признаков и о количестве однородных групп (диагнозов) в рассматриваемой выборке с учетом особенностей данных медико-биологических исследований. Следует отметить, что данный подход является универсальным лишь в случае отсутствия обучающей выборки и дополнительной информации о данных и может быть использован для получения первого представления об информации, которая имеется в медицинских данных.

Также предложенный алгоритм поможет исследователю подобрать метод для дальнейшего анализа, покажет уровень «нечеткости» данных, то есть степень пересечения кластеров, будет полезен при дальнейшей оценке информативности медико-биологических показателей.

Литература

1. Лбов, Г. С. Метод адаптивного поиска логической решающей функции [Текст] / В. М. Неделько, С. В. Неделько // Сиб. журн. индустр. матем. – 12:3 2009. – С. 66–74
2. Айвазян, С. А. Прикладная статистика и основы эконометрики. Теория вероятностей и прикладная статистика / В. С. Мхитарян — М.: Юнити, 2001.
3. Дорофеюк, А. А. Процедуры классификационного анализа в задаче формирования информативных признаков при исследовании ритмической структуры биосигнала [Текст] / А. А. Десова, В. В. Гучук, Ю. А. Дорофеюк, И. В. Покровская // Автоматика и телемеханика. - 2008. - № 6. – С. 143-152.
4. Zagoruiko, N. Principe of Natural Classification [Text] / N. Zagoruiko, I. Borisova // Int. Journal «Pattern Recognition and Image Analysis». - 2005. - Vol 15, № 1. - P. 27-29.
5. Nelles, O. Nonlinear System Identification: from classical approaches to neural networks and fuzzy models. [Text] / O. Nelles // Springer - Verlag Berlin Heidelberg New York, 2001. – 785 p.
6. Seraya, O. V. Linear regression analysis of a small sample of fuzzy input data [Text] / O. V. Seraya, D. A. Demin // Journal of Automation and Information Sciences. – 2012. – Vol. 44 (7). – P. 34 - 48.
7. Дёмин, Д. А. Нечеткая кластеризация в задаче построение моделей «состав – свойство» по данным пассивного эксперимента в условиях неопределённости / Д. А. Дёмин // Проблемы машиностроения. – 2013. – №6. – С. 15 – 23.
8. Данилова, Н. В. Применение метода нечетких с-средних для построения функций принадлежности параметров технологического процесса [Текст] / Н. В. Данилова // Сб. научн. тр. семинара «Инновационные технологии, моделирование и автоматизация в металлургии». – Санкт-Петербург, 2010. – С. 11-12.
9. Тесленко, Н. А. Нечеткая кластеризация массивов биомедицинских данных в условиях избыточности информации [Текст] / Н. А. Тесленко, И. Г. Чурюмова // Бионика интеллекта. – 2006. – №1 (64). – С. 92-95.
10. Bishop, Christopher. Pattern recognition and machine learning. Berlin: Springer. 2006. - ISBN 0-387-31073-8.
11. Чурюмова, И. Г. Система медицинской диагностики на основе нечеткой логики [Текст] / И. Г. Чурюмова // Восточно-Европейский журнал передовых технологий. – 2006. – 5/2 (23). – С. 89-91.
12. Чурюмова, И. Г. Система донозологической диагностики сердечно-сосудистых заболеваний [Текст] / И. Г. Чурюмова // Восточно-Европейский журнал передовых технологий. – 2007. – № 5/4 (29). – С. 31-33.
13. Чурюмова, И. Г. Применение методов нечеткой кластеризации для анализа медицинских данных в режиме реального времени [Текст] / И. Г. Чурюмова, Н. П. Мустецов // Электроника и связь. Тематический выпуск «Проблемы электроники». – 2007. – Ч. 2. – С. 118-121.
14. Патент України на винахід № 91767 Спосіб оцінки біологічних станів, заснований на нечіткій кластеризації даних множини вимірювальних показників [Текст]: МПК (2009) G06F 19/00 G06F 17/00 G06F 7/00 G01N 33/48/ Бодянський Є. В., Мустецов М. П., Чурюмова І. Г.; Харківський національний університет радіоелектроніки. – Заявл. від 22.12.2008; опубл. 25.08.2010. – Бюл. №16.