УДК 681.518:004.93.1'

ИНФОРМАЦИОННО-ЭКСТРЕМАЛЬНАЯ КЛАССИФИКАЦИЯ ДИАГНОСТИЧЕСКИХ ДАННЫХ

С. А. С. М. Джулгам, *аспирант,*
Сумский государственный университет, г. Сумы

Предложены категорийная модель и иерархический алгоритм классификации данных в рамках информационно-экстремальной интеллектуальной технологии, основанной на максимизации информационной способности системы диагностирования. При этом классификация данных осуществляется путем вложения модифицированной для бинарного пространства признаков процедуры K-means в алгоритм информационно-экстремальной классификации.

Ключевые слова: *классификация, кластер, информационный критерий, функциональная эффективность, система поддержки принятия решений, диагностирование.*

ВВЕДЕНИЕ

Повышение функциональной эффективности компьютеризированных систем диагностирования и прогнозирования течения и исхода лечения патологических процессов связано с разработкой и внедрением в практическое здравоохранение телекоммуникационной системы, на базе которой будут созданы национальный и региональные лечебно-диагностические GRID-центры. Как показывает накопленный опыт, решение этой проблемы зависит не только от экономических факторов, а в значительной степени от научно-методологических причин, связанных с созданием основ анализа и синтеза интеллектуальных систем поддержки принятия решений (СППР), которые являются основной составляющей сервисного GRID-центра и способны моделировать когнитивные процессы, присущие человеку при принятии решений. При этом важное значение приобретает задача разработки эффективных алгоритмов кластер-анализа входных данных с целью автоматизации формирования входного математического описания интеллектуальной СППР. Основным недостатком известных методов кластеризации данных [1 -3], основанных на дистанционных критериях близости, является их модельность, обусловленная игнорированием пересечения кластеров, которое имеет место в практических задачах диагностирования, и произвольными начальными условиями формирования диагностических данных. Поэтому эти методы не обеспечивают высокой достоверности и устойчивости разбиения пространства признаков на кластеры.

Одним из перспективных подходов к повышению функциональной эффективности интеллектуальных СППР, функционирующих в режиме кластер-анализа, является вложение построенных по дистанционным критериям алгоритмов кластер-анализа, позволяющих определять геометрические центры кластеров, в алгоритмы машинного обучения в рамках информационно-экстремальной интеллектуальной технологии (ИЭИ-технологии), позволяющей оптимизировать в информационном смысле геометрические параметры контейнеров кластеров, восстанавливаемых в радиальном базисе пространства признаков распознавания [4 -6]. Рассмотренные в работах [7, 8] алгоритмы кластеризации данных в рамках ИЭИ-технологии предусматривали сначала построение по дистанционным критериям априорного нечеткого разбиения кластеров с последующей их оптимизацией в рамках информационно-экстремальных алгоритмов, что существенно снижало оперативность классификации.

В статье рассматривается в рамках ИЭИ-технологии категорийная модель и алгоритм иерархической классификации данных для заданной мощности разбиения с вложенным в него алгоритмом определения геометрических центров кластеров, являющийся аналогом метода *K*-means для бинарного пространства признаков распознавания.

ПОСТАНОВКА ЗАДАЧИ

Рассмотрим формализованную постановку задачи информационного синтеза СППР, функционирующей в режиме классификации диагностических данных в рамках ИЭИ-технологии. Пусть известны мощность $Card\,\tilde{\Re} = M$ в общем случае нечёткого разбиения $\tilde{\tilde{\Re}}^{|M|}$ пространства признаков на кластеры и неклассифицированная обучающая матрица $\| y_i^{(j)} \|$, $i = \overline{1,N}$, $j = \overline{1,n}$, где N, n – количество диагностических признаков и векторов-реализаций (далее просто реализации) кластера соответственно. При этом кластер характеризует функциональное состояние патологического процесса, диагностический признак является результатом лабораторно-клинического анализа и анамнеза патологического процесса, а реализация представляет структурированное множество диагностических признаков. Кроме того, дано структурированный вектор параметров СППР $g =< x_m , d_m , \delta >$, где x_m – эталонная реализация, вершина которой определяет геометрический центр контейнера кластера X_m^o, $m = \overline{1,M}$; d_m – радиус контейнера кластера X_m^o і δ – параметр симметричного поля контрольных допусков на диагностические признаки. При этом заданы следующие ограничения: $d_m \in [0; d(x_m \oplus x_c) - 1]$, где $d(x_m \oplus x_c)$ – кодовое расстояние центра кластера X_m^o от центра ближайшего (соседнего) к нему кластера X_c, і $\delta \in [0; \delta_H / 2]$, де δ_H – нормированное (эксплуатационное) поле допусков для относительной шкалы измерения признаков, который задает область значений параметра контрольного поля допусков. Необходимо в процессе классификации диагностических данных определить геометрические центры контейнеров кластеров разбиения $\tilde{\tilde{\Re}}^{|M|}$ и оптимизировать (здесь и далее в информационном смысле) его геометрические параметры путем итерационного поиска экстремальных значений координат вектора параметров обучения, обеспечивающих максимальное значение усредненного по алфавиту кластеров информационного критерия функциональной эффективности (КФЭ) классификации диагностических данных:

$$\bar{E}^* = \frac{1}{M} \sum_{m=1}^{M} \max_{\{k\}} E_m, \qquad (1)$$

где E_m - информационный КФЭ обучения СППР распознавать реализации кластера X_m^o; $\{k\}$ – упорядоченное множество шагов обучения (восстановление в радиальном базисе пространства диагностических признаков контейнеров кластеров).

КАТЕГОРИЙНАЯ МОДЕЛЬ КЛАССИФИКАЦИИ ДАННЫХ

Входное математическое описание СППР, функционирующей в режиме классификации данных, представим в виде теоретико-множественной структуры:

$$\Delta_B =< T,\ G,\ \Omega,\ Z,\ Y,\ X;\ F_1,\ F_2 >,$$

где T – множество моментов времени снятия информации; G – пространство факторов, влияющих на СППР; Ω – пространство диагностических признаков; Z – пространство функциональных состояний СППР; Y – входная неклассифицированная обучающая матрица; X – бинарная неклассифицированная обучающая матрица; $F_1 : T \cdot \Omega \cdot Z \cdot G \to Y$ – оператор формирования обучающей матрицы Y; $F_2 : Y \to X$ – оператор формирования бинарной обучающей матрицы.

На рис. 1 показана категорийная модель СППР в виде диаграммы отображения множеств, применяемых в процессе иерархической классификации данных.

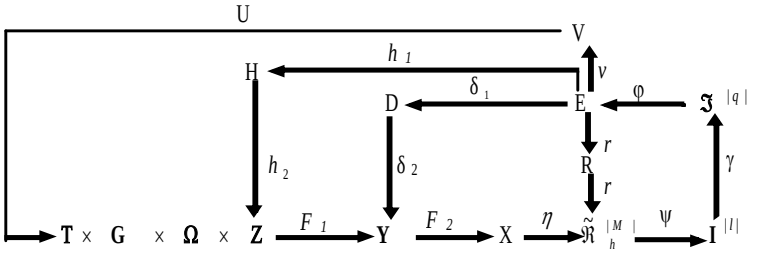

Рисунок 1 – Категорийная модель СППР, функционирующей в режиме кластер-анализа данных

В диаграмме (рис. 1) оператор $\eta : X \to \Re_h^{|M|}$ определяет геометрические центры контейнеров кластеров, образующих на h-м ярусе иерархической структуры данных разбиение $\tilde{\Re}_h^{|M|}$ пространства признаков, а оператор классификации $\Psi : \tilde{\Re}^{|M|} \to I^{|I|}$ проверяет основную статистическую гипотезу о принадлежности реализаций $\{x_{m,h}^{(j)} \mid j = \overline{1,n}\}$ классу $X_{m,h}^{o}$ и формирует множество гипотез $I^{|I|}$, где I – количество статистических гипотез. Оператор $\gamma : I^{|I|} \to \Im^{|q|}$ по результатам оценки статистических гипотез формирует множество точностных характеристик $\Im^{|q|}$, где $q = I^2$. Оператор $\varphi : \Im^{|q|} \to E$ вычисляет множество значений информационного КФЭ, который является функционалом от точностных характеристик. Контур диаграммы, содержащий терм-множество R радиусов контейнеров кластеров состоит из операторов Ψ, γ, φ, r_1 і r_2, оптимизирует геометрические параметры разбиения $\tilde{\Re}_h^{|M|}$ путем итерационного поиска глобального максимума КФЭ в рабочей (допустимой) области определения его функции. Оптимизация СКД на диагностические признаки осуществляется итерационной процедурой, в которой последовательно реализуются операторы F_2, η, ψ, γ, φ, δ_1 і δ_2. Контур оптимизации геометрических параметров контейнеров кластеров каждого яруса структуры H замыкается операторами h_1 і h_2. Оператор v при условии, что максимум КФЭ обучения СППР не достигает своего предельного значения, осуществляет переход к следующему типу решающих правил, которые формируются в радиальном базисе

пространства признаков распознавания, а оператор U: $V \to G \cdot T \cdot \Omega \cdot Z$ регламентирует процесс обучения СППР.

Таким образом, показанная на рис. 1 категорийная модель иерархической классификации данных сочетает традиционный дистанционный подход к определению центров кластеров и информационно-экстремальный подход, позволяющий осуществлять статистическую коррекцию геометрических параметров разбиения.

АЛГОРИТМ КЛАССИФИКАЦИИ

Алгоритм классификации данных согласно категорийной модели (рис. 1) в рамках ИЭИ-технологии представим как двухциклическую процедуру итерационной оптимизации параметра δ_h поля системы контрольных допусков (СКД) на диагностические признаки для алфавита кластеров h-го яруса иерархической структуры данных

$$\delta_h^* = \arg \max_{G_\delta} \{ \max_{G_E} \overline{E}_h \}, \qquad (2)$$

где $\overline{E}_h$ – КФЭ классификации данных на h-м ярусе структуры алгоритма; G_δ – допустимая область значений параметра δ_h; G_E – рабочая (допустимая) область определения функции КФЭ.

Реализация информационно-экстремального алгоритма классификации (2) осуществляется при следующих ограничениях

$$\left(\forall X_{m,h}^o \in \tilde{\Re}^{|M|} \right) \left[X_{m,h}^o \neq \varnothing, m = \overline{1, M} \right], \qquad (3)$$

$$\left(\forall X_{m,h}^o \in \tilde{\Re}^{|M|} \right) \left(\forall X_{c,h}^o \in \tilde{\Re}^{|M|} \right) \left[X_{m,h}^o \neq X_{c,h}^o \to Ker X_{m,h}^o \cap Ker X_{c,h}^o = \varnothing \right], \qquad (4)$$

$$\left(\forall X_{m,h}^o \in \tilde{\Re}^{|M|} \right) \left(\forall X_{c,h}^o \in \tilde{\Re}^{|M|} \right) [X_{m,h}^o \neq X_{c/h}^o \to (d_{m,h}^* < d(x_{m,h} \oplus x_{c,h})) \& (d_{hc}^* < d(x_{m,h} \oplus x_{c,h}))], \qquad (5)$$

$$\bigcup_{X_{m,h}^o \in \tilde{\Re}} X_{m,h}^o \subseteq \Omega, \qquad (6)$$

где $Ker X_{m,h}^o$, $Ker X_{c,h}^o$ – ядра двух ближайших (соседних) кластеров $X_{m,h}^o$ і $X_{c,h}^o$; $d_{m,h}^*$ – оптимальный радиус контейнера кластера $X_{m,h}^o$; $d(x_m \oplus x_c)$ – межцентровое кодовое расстояние кластеров $X_{m,h}^o$ і $X_{c,h}^o$; $d_{c,h}^*$ – оптимальный радиус контейнера кластера $X_{c,h}^o$.

Рассмотрим алгоритм классификации данных (2) с параллельной оптимизацией контрольных допусков на диагностические признаки, восстанавливающий в радиальном базисе пространства признаков оптимальные контейнеры кластеров. При этом формирование классифицированной обучающей матрицы $\| y_{h,m,i}^{(j)} \|$ на h-м уровне иерархической структуры будем осуществлять при условии, что обработано не менее 80 % реализаций от их общего количества. Это ограничение позволяет отфильтровывать реализации, находящиеся на периферии их распределения и имеющие поэтому малую вероятность принадлежности к восстанавливаемым кластерам. Если не попавшие в обучающую матрицу реализации будут поступать на вход СППР, функционирующей в режиме экзамена, то для их классификации

предполагается применение разработанных в рамках ИЭИ-технологии методов факторного кластер-анализа [9], позволяющих выделять новые кластеры.

Входными данными алгоритма классификации являются массив реализаций $\{Y_i^{(j)} \mid i = \overline{1,N}; j = \overline{1,n}\}$ и система нормированных допусков $\{\delta_{H,i}\}$, определяющая область значений соответствующих контрольных допусков на диагностические признаки.

Рассмотрим основные этапы реализации алгоритма разбиения пространства признаков на три кластера:

1. Обнуляется счетчик шагов изменения параметра δ поля контрольных допусков на диагностические признаки: $l := 0$.

2. Инициализируется счетчик шагов изменения параметра δ: $l := l + 1$ и вычисляются нижние $A_{HK_i}[l]$ и верхние $A_{BK_i}[l]$ контрольные допуски для всех признаков:

$$A_{HK_i}[l] = y_i - \delta \frac{\delta_{H_i}}{100}, \tag{7}$$

где y_i – i-й признак эталонного вектора-реализации неклассифицированной многомерной обучающейся матрицы $\| y_i^{(j)} \|$.

3. Формируется бинарная обучающая матрица $\| x_i^{(j)} \|$ по правилу

$$x_i^{(j)} = \begin{cases} 1, \ if \ \ A_{HK_i}[l] < y_i^{(j)} < A_{BK_i}[l]; \\ 0, \ else. \end{cases}$$

4. Обнуляется счетчик ярусов иерархической структуры алгоритма классификации: $h := 0$.

5. Инициализируется счетчик ярусов иерархической структуры алгоритма классификации: $h := 1$.

6. Разбивается исходное множество реализаций $\{\tilde{O}_i^{(j)}\}$ на два кластеры $\{X_m^o[h] \mid m = \overline{1,2}\}$:

а) находятся начальные эталонные векторы-реализации $\{x_m\}$ кластеров X_m^0 при условии, что $d(x_1 \oplus x^0) \to \min$, $d(x_2 \oplus x^1) \to \min$ и $d(x_1 \oplus x_2) \to \max$, где x^0, x^1 - нулевой и единичный векторы;

б) обнуляются значения радиусов кластеров X_m^0: $d_m[h] := 0$, $n_m := 1$, где n_m - количество реализаций, попавших в кластер $X_m^o[h]$;

в) $d_m[h] := d_m[h] + 1$;

г) если $N' = \sum_{m=1}^{M} n_m < N$, где N - общее количество реализаций, то выполняется пункт 6д, иначе – пункт 6и;

д) определяются реализации, попавшие в кластеры $X_m^o[h]$, по правилу

$$x_i \in X_1^o[h], \ if \ \ d(x_i \oplus x_1) <= d \ and \ d(x_i \oplus x_1) < d(x_i \oplus x_2),$$
$$x_i \in X_2^o[h], \ if \ \ d(x_i \oplus x_2) <= d \ and \ d(x_i \oplus x_2) <= d(x_i \oplus x_1),$$

где $x_i \mid i = \overline{1,N}$ – реализации бинарной обучающей матрицы $\| x_i^{(j)} \|$;

е) формируется множество $\{x_m\}$ эталонных реализаций кластеров $\{X_m^o[h]\}$, координаты которых определяются по правилу

$$x_{m,i} = \begin{cases} 1, \; if \; \dfrac{1}{n}\sum_{j=1}^{n} x_{m,i}^{(j)} > 0,5; \\ 0, \; else. \end{cases} \qquad (8)$$

ж) вычисляется значение информационного КФЭ (1);

з) если $d[h] < d(x_1 \oplus x_2)$, то выполняется пункт 6д, иначе – пункт 6и;

и) определяются максимальное значение КФЕ (1) и оптимальные радиусы кластеров $\{X_m^o[h]\}$ при выполнении условия $N' = \sum_{m=1}^{M} n_m < N$, где N' – количество реализаций принадлежащих разбиению $\tilde{\Re}_h$.

7. $h := h + 1$.

8. Разбивается бинарное пространство диагностических признаков на три кластера $\{X_m^o\,[h] \mid m = \overline{1,3}\}$:

а) определяется начальная реализация $\{x_3\}$ кластера X_3^o, при условии, что $d(x_1 \oplus x_3) \to \min$ и $d(x_2 \oplus x_3) \to \min$, где x_1, x_2 – эталонные реализации кластеров $\{X_m^o \mid m = \overline{1,2}\}$, восстановленных при выполнении пункта 6;

б) обнуляется значение радиуса кластера X_3^o: $d_3 := 0$;

в) если $d < d(x_1 \oplus x_3)$ и $d < d(x_2 \oplus x_3)$, то выполняется пункт 9г, иначе– пункт 9и;

г) $d_3 := d_3 + 1$;

д) определяются реализации, попавшие в кластер X_3^o по правилу

$x_i \in X_3^o, if \; d(x_i \oplus x_3) <= d \; and \; d(x_i \oplus x_3) <= d(x_i \oplus x_1) \; and \; d(x_i \oplus x_3) <= d(x_i \oplus x_2)$,

где $x_i \mid i = \overline{1,N}$ - реализации бинарной обучающей матрицы $\| x_i^{(j)} \|$;

е) формируется множество $\{x_m\}$ эталонных реализаций кластеров $\{X_m^o[h]\}$ по правилу (8);

ж) вычисляется значение информационного КФЭ (1);

з) выполняется пункт 6и;

и) определяется оптимальный радиус контейнера кластера X_3^o при выполнении условий $d < d(x_1 \oplus x_3)$ и $d < d(x_2 \oplus x_3)$.

9. Если $\delta[l] \le \delta_H / 2$, то выполняется пункт 2, иначе – пункт 10.

10. Если $\overline{E}[l] \notin G_E$, то выполняется пункт 11, иначе – пункт 2.

11. Выполняется процедура поиска глобального максимума КФЭ $\overline{E}[l]$ в рабочей области определения его функции при условии, что $N' = \sum_{m=1}^{3} n_m \ge 0,8N$.

12. $\overline{E}^{*}[l] := \underset{\{\delta\}}{extremE_m}[l]$.

13. Определяется оптимальный параметр поля контрольных допусков на диагностические признаки $\delta^{*} := \arg\overline{E}^{*}[l]$ и по формуле (7) вычисляются оптимальные контрольные допуски на признаки распознавания:

$$A^{*}_{HK_i} = y_i - \delta^{*}\frac{\delta_{H_i}}{100};\ A^{*}_{BK_i} = y_i + \delta^{*}\frac{\delta_{H_i}}{100}.$$

14. ОСТАНОВ.

В качестве критерия оптимизации параметров обучения в рамках ИЭИ-технологии может рассматриваться любая статистическая информационная мера, являющаяся функционалом от точностных характеристик. Например, для двухальтернативных решений и равновероятных гипотез можно применить модификацию энтропийного КФЭ (по Шеннону) [6]

$$E^{(k)}_m = 1 + \frac{1}{2}\left(\frac{\alpha^{(k)}_m(d)}{\alpha^{(k)}_m(d) + D^{(k)}_{2,m}(d)}\log_2\frac{\alpha^{(k)}_m(d)}{\alpha^{(k)}_m(d) + D^{(k)}_{2,m}(d)} + \right.$$

$$+\frac{\beta^{(k)}_m(d)}{D^{(k)}_{1,m}(d) + \beta^{(k)}_m(d)}\log_2\frac{\beta^{(k)}_m(d)}{D^{(k)}_{1,m}(d) + \beta^{(k)}_m(d)} + \frac{D_{1,m}(d)}{D^{(k)}_{1,m}(d) + \beta^{(k)}_m(d)}\log_2\frac{D_{1,m}(d)}{D^{(k)}_{1,m}(d) + \beta^{(k)}_m(d)} +$$

$$\left. +\frac{D^{(k)}_{2,m}(d)}{\alpha^{(k)}_m(d) + D^{(k)}_{2,m}(d)}\log_2\frac{D^{(k)}_{2,m}(d)}{\alpha^{(k)}_m(d) + D^{(k)}_{2,m}(d)}\right), \qquad (9)$$

где $\alpha^{(k)}_m(d)$ – ошибка первого рода при принятии решений на k-м шаге обучения; $\beta^{(k)}_m(d)$ – ошибка второго рода; $D^{(k)}_{1,m}(d)$ – первая достоверность; $D^{(k)}_{2,m}(d)$ – вторая достоверность; d – полная мера, определяющая радиусы гиперсферических контейнеров классов распознавания.

Таким образом, с целью обеспечения нахождения значения усредненного КФЭ в рабочей области определения его функции, в которой выполняются ограничения $D_{1,m} > 0,5$, $D_{2,m} > 0,5$ и $d^{*}_m < d(x_m \oplus x_c)$, где $d(x_m \oplus x_c)$ - межцентровое кодовое расстояние кластера X^{o}_m к его ближайшему соседу, на втором ярусе иерархической структуры одновременно с восстановлением контейнера кластера X^{o}_3 проводится коррекция контейнеров кластеров $\{X^{o}_m \mid m = \overline{1,2}\}$, восстановленных на первом ярусе.

ПРИМЕР РЕАЛИЗАЦИИ АЛГОРИТМА КЛАССИФИКАЦИИ

Рассмотрим реализацию вышеприведенного алгоритма классификации диагностических данных в самообучающейся СППР с параллельной оптимизацией контрольных допусков на диагностические признаки на примере определения стадии патологии острой кишечной инфекции (ОКИ), вызванной условно-патогенными микроорганизмами, с целью выбора врачом соответствующей схемы лечения. Входная априорно неклассифицированная многомерная обучающая матрица $\| y^{(j)}_i \|$ состояла из 90 реализаций и 19 признаков распознавания, характеризующих различные состояния патологического процесса. При этом каждые 30

реализаций матрицы $\| y_i^{(j)} \|$ характеризовали контрольную группу (практически здоровые лица) – кластер X_1^o, группу пациентов, для которых необходимо комбинированное лечение с включением в схемы коллоидного серебра (10 мг/л) – кластер X_2^o и группу пациентов, для которых необходимо одновременное назначение пробиотика и коллоидного серебра на фоне базисной терапии – кластер X_3^o. Реализации кластеров представлены в виде структурированной последовательности признаков распознавания, полученных по результатам лабораторных исследований микробиоценоза кишечника, уровня секреторного IgA, противовоспалительных цитокинов IL 1β, противовоспалительного цитокина IL 4, интегративных показателей эндогенной интоксикации.

На рис. 2 показан график зависимости усреднённого по алфавиту кластеров первого яруса иерархической структуры нормированного КФЭ (9) от параметра δ_h поля контрольных допусков на диагностические признаки. На рис. 2 темные участки графиков обозначают рабочие области, в которых значения первой и второй достоверностей превышают соответствующие ошибки первого и второго родов, а радиус контейнера кластера $X_{h,m}^o$ меньше его межцентрового расстояния с ближайшим соседом.

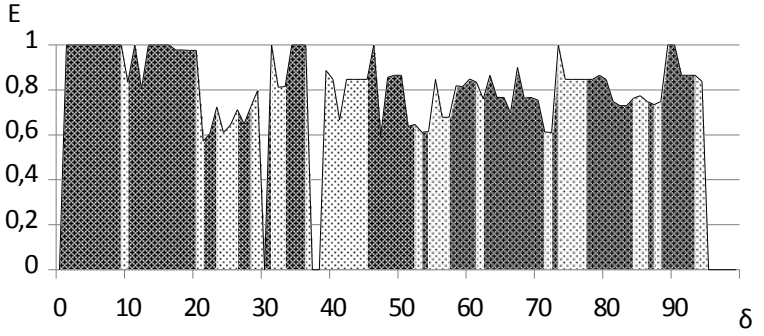

Рисунок 2– График зависимости критерия Шеннона от величины параметра поля контрольных допусков на диагностические признаки

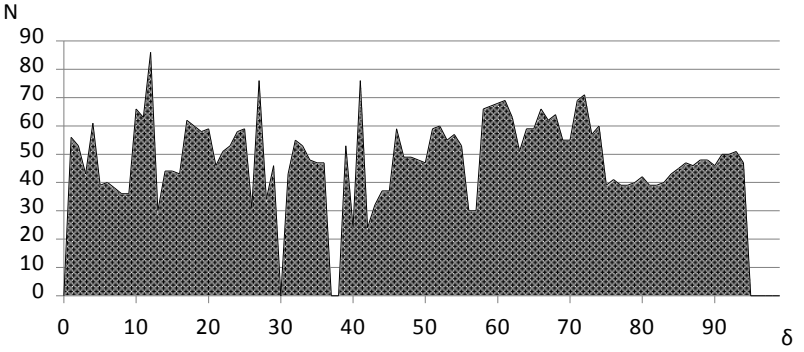

Рисунок 3 – График зависимости количества точек, попавших в кластеры первого яруса, от величины параметра поля контрольных допусков на диагностические признаки

На рис. 3 показан график зависимости количества реализаций, принадлежащих разбиению кластеров, от параметра поля допусков δ, полученный в процессе реализации алгоритма классификации на первом ярусе иерархической структуры.

На рис. 4 показана зависимость усреднённого значения КФЭ (9) от радиусов кластеров, восстанавливаемых на первом ярусе структуры алгоритма классификации, при оптимальном параметре поля допусков δ^* и количестве реализаций $N' \geq 0,8N$.

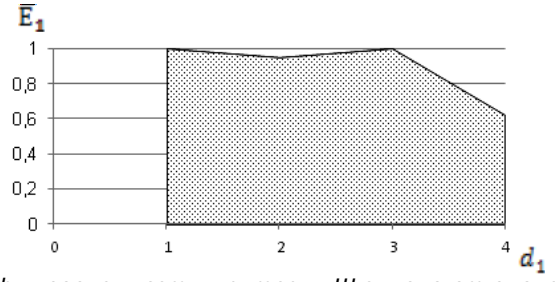

Рисунок 4 – График зависимости критерия Шеннона от значения радиусов кластеров первого яруса

Анализ рис. 4 показывает, что оптимальный радиус кластеров, обеспечивающих выполнение условий $N' = \sum_{m=1}^{M} n_m < 0,8N$ и $d < d(x_1 \oplus x_2)$, равняется $d^* = 3$ (здесь и далее в кодовых единицах), а усредненное максимальное значение КФЭ равняется $\bar{E} = 1$ при количестве точек, попавших в кластеры первого яруса, $N' = 89$. При значениях радиуса $d = 0$ и $d = 1$ значение КФЭ также равняется единице, но не выполняется первое условие.

На рис. 5 приведены зависимости среднего значения КФЭ от радиуса кластера X_3^o, восстанавливаемого на втором ярусе иерархической структуры, при оптимальном параметре поля допусков δ^* и количестве реализаций $N' \geq 0,8N$.

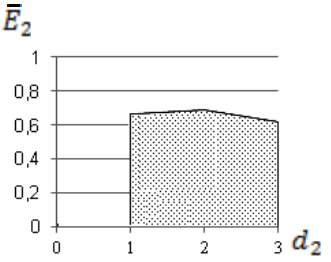

Рисунок. 5 – График зависимости критерия Шеннона от значения радиусов кластеров второго яруса

Анализ рис. 5 показывает, что оптимальный радиус кластера X_3^o, обеспечивающий выполнение условий $d < d(x_1 \oplus x_3)$ и $d < d(x_2 \oplus x_3)$ равняется $d_2^* = 2$. При этом усредненное максимальное значение КФЭ и количество точек, попавших в кластеры второго яруса, равняются соответственно $\bar{E}_2^* = 0,68$ и $N' = 87$.

С целью повышения КФЭ классификации при разбиении пространства признаков на кластеры на втором ярусе была реализована параллельно-последовательная оптимизация контрольных допусков на диагностические признаки, особенность которой состояла в том, что полученные на этапе параллельной оптимизации квазиоптимальные контрольные допуски принимались как стартовые для процедуры последовательной операции в поочередной смене контрольных допусков для каждого признака при заданных допусках для следующих признаков. Алгоритм последовательной оптимизации контрольных допусков на признаки распознавания в рамках ИЭИ-технологии представим в виде структурированной двухциклической итерационной процедуры:

$$\{\delta^*_{K,i}\} = \arg\left[\bigotimes_{s=1}^{S} \max_{G_{\delta_i}}\{\max_{G_E \cap G_d} \bar{E}^{(s)}\}\right], \quad i = \overline{1,N}, \tag{10}$$

где $\bar{E}^{(s)}$ – усредненный КФЭ обучения СППР, вычисленный на s-й итерации последовательной процедуры; G_{δ_i}, G_E, G_d – области допустимых значений контрольных допусков для i-го признака, критерия оптимизации и радиусов контейнеров соответственно; $\otimes$ – символ операции повторения.

В процедуре (10) полученные в результате параллельной оптимизации квазиоптимальные контрольные допуски на диагностические признаки принимаются как стартовые, что обеспечивает нахождение значений информационного КФЭ в рабочей области определения его функции и существенно повышает оперативность алгоритма последовательной оптимизации.

Результаты паралельно-последовательной оптимизации контейнеров кластеров на втором ярусе иерархической структуры алгоритма классификации показали, что уже на втором прогоне адгоритма (10) получено максимальное усреднённое значение КФЭ, равное $\bar{E}^* = 0,83$. При этом общее количество точек, попавших в кластеры второго яруса, равняется $N' = 88$.

Проведенное для сравнения обучение в рамках ИЭИ-технологии диагностической системы по априорно классифицированной обучающей матрице, содержащей по 30 аналогичных реализаций для каждой из трёх выше рассмотренных патологий, позволило получить максимальное значение КФЭ, равное $\bar{E}^* = 0,85$, что свидетельствует о достаточно хорошей сходимости предложенного алгоритма классификации.

ВЫВОДЫ

1. Предложенная схема двухъярусного иерархического информационно-экстремального алгоритма классификации отличается простотой реализации при разбиении пространства диагностических признаков не более чем на четыре кластера и обеспечивает приемлемую с практической точки зрения сходимость.

2. Поскольку сложность задачи классификации существенно увеличивается с ростом числа кластеров, то при мощности разбиения, превышающей четыре кластера целесообразным является использование информационно-экстремального алгоритма факторного кластер-анализа, позволяющего выделять новые кластеры с переобучением диагностической системы.

ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНА КЛАСИФІКАЦІЯ ДІАГНОСТИЧНИХ ДАНИХ

С. А. С. М. Джулгам,
Сумський державний університет, м. Суми

Запропоновано категорійну модель та ієрархічний алгоритм класифікації даних у рамках інформаційно-екстремальної інтелектуальної технології, що ґрунтується на максимізації інформаційної спроможності системи діагностування. При цьому класифікація даних здійснюється шляхом вкладення модифікованої для бінарного простору ознак процедури K-means в алгоритм інформаційно-екстремальної класифікації.

Ключові слова: *класифікація, кластер, інформаційний критерій, функціональна ефективність, система підтримки прийняття рішень, діагностування.*

INFORMATION AND EXTREME CLASSIFICATION OF DIAGNOSTIC DATA

S. A. S. M. Julgam,
Sumy State University, Sumy

A categorical model and hierarchical algorithm to classify data within the information-extreme intellectual technology, based on the maximization of the information capacity of the system diagnostics, are proposed. In this case, classification of data is implemented by embedding the modified procedure of K-means for binary features space in the information-extreme classification algorithm.

Key words: *classification, cluster, information criterion, functional efficiency, decision support system, diagnosing.*

СПИСОК ЛИТЕРАТУРЫ

1. Симанков В. С. Адаптивное управление сложными системами на основе теории распознавания образов / В. С. Симанков, Е. В. Луценко. – Краснодар : Техн. ин-т Кубан. гос. технол. ун-та, 1999. – 318 с.
2. Halkidi, Maria. On Clustering Validation Techniques /Maria Halkidi, Yannis Batistakis, Michalis Vazirgiannis // Journal of Intelligent Information Systems. - December 2001. - Volume 17 Issue 2-3. – P.107–145.
3. Усков А. А. Экспресс-диагностика ОРВИ средствами нечетко-логической экспертной системы / А. А. Усков, М. В. Шипилов, В. В. Иванов // Международный журнал Программные продукты и системы. – 2011. – № 3. – С. 62 -70.
4. Краснопоясовський А. С. Інформаційний синтез інтелектуальних систем керування: Підхід, що ґрунтується на методі функціонально-статистичних випробувань / А. С. Краснопоясовський. – Суми: Видавництво СумДУ, 2004. – 261 с.
5. Довбиш А. С. Основи проектування інтелектуальних систем : навчальний посібник / А. С. Довбиш. – Суми : Видавництво СумДУ, 2009. – 171 с.
6. Information-extreme algorithm for recognizing current distribution maps in magnetocardiography / A. S. Dovbysh, S. S. Martynenko, A. S. Kovalenko, N. N. Budnyk// Journal of Automation and Information Sciences . – 2011. – V. 43. – № 2. – P. 63 -70.
7. Оптимизация иерархической структуры алгоритма распознавания магнитокардиограмм / А. С. Довбыш, А. С. Коваленко, И. А. Чайковский, С. С. Мартыненко // Кибернетика и вычислительная техника. – 2011. – Вып. 163. – С. 65 -75.
8. Довбиш А. С. Информационно-экстремальный кластер-анализ в самообучающихся GRID-центрах / А. С. Довбыш, С. А. С. М. Джулгам // Межд. крымская конференция «СВЧ техника, телекоммуникационные технологии». материалы конференции в 2 т. – Севастополь : Вебер, 2012. - Т. 1. - С. 413 -414.
9. Краснопоясовський А. С. Факторний класифікаційний аналіз за методом функціонально-статистичних випробувань / А. С. Краснопоясовський, М. В. Козинець // Радіоелектронні та комп'ютерні системи. – 2004. – № 4. – С. 46 -50.

Поступила в редакцию 12 октября 2012 г.