УДК 681.518:004.93.1

А.С. Довбыш, Саад Джулгам, С.О. Петров

Иерархический информационно-экстремальный алгоритм кластер-анализа результатов машинного тестирования уровня знаний учащихся

Предложены категориальная модель и алгоритм кластеризации входных данных. Формирование априорно нечеткой классифицированной обучающей матрицы осуществляется по иерархическому таксонометрическому алгоритму, а построение в процессе обучения системы четкого разбиения пространства признаков на классы распознавания – по информационно-экстремальному алгоритму.

A categorical model and an algorithm of input data clustering are suggested. The Formation of priory fuzzy classified matrix is carried out in the frame of a hierarchical taxonomatic algorithm. But the construction in the process of learning of the accurate partition of the features space is carried out in the bounds of the information-extreme algorithm.

Запропоновано категорійну модель і алгоритм кластеризації вхідних даних. Формування апріорно нечіткої класифікованої навчальної матриці здійснюється за ієрархічним таксонометричним алгоритмом, а побудова в процесі навчання системи чіткого розбиття простору ознак на класи розпізнавання – за інформаційно-екстремальним алгоритмом.

Введение. Задачу оценивания уровня знаний учащихся решает непосредственно преподаватель, относя результаты тестирования интуитивно к соответствующей оценочной шкале. Однако наметившаяся в последнее десятилетие тенденция расширения контингента учащихся всех форм обучения, особенно дистанционной, и возросшие требования к достоверности оценки уровня их знаний требуют дополнительных материальных и педагогических ресурсов. Поэтому одним из путей повышения эффективности системы управления учебным процессом, существенная составляющая которого – оценка уровня знаний студентов, есть применение машинной кластеризации входных данных, полученных в процессе тестирования [1–3]. Однако известные методы кластер-анализа, основанные на дистанционных критериях сходства, позволяют строить нечеткое разбиение пространства признаков на классы распознавания, что требует дополнительных допустимых преобразований входного математического описания компьютеризованной системы оценки (КСО) уровня знаний учащихся с целью построения безошибочных по учебной матрице решающих правил. Одним из перспективных направлений анализа и синтеза систем компьютеризации образования есть использование идей и методов информационно-экстремальной интеллектуальной технологии (ИЭИ-технология), основанной на максимизации информационной способности системы путем введения в процесс ее обучения дополнительных информационных ограничений [4–6]. В работе [6] рассмотрена задача кластеризации данных в рамках ИЭИ-технологии, хотя авторам не удалось построить безошибочные по обучающей матрице решающие правила.

Рассмотрим математическую модель и информационно-экстремальный алгоритм кластеризации входных данных – результаты машинного тестирования уровня знаний студентов.

Постановка задачи

Пусть возможные функциональные состояния процесса изучения дисциплины, состоящего из K тематических модулей, характеризуются нечетким алфавитом классов распознавания $\{X^o_{m,k} \mid m=\overline{1,M};\ \ k=\overline{1,K}\}$, где M – количество классов распознавания для одного модуля. При этом класс распознавания $X^o_{m,k}$ характеризует соответствующий уровень знаний, полученных учеником при изучении k-го модуля. Эффективность обучения КСО, основным функциональным элементом которой есть система под-

* **Ключевые слова:** кластер-анализ, информационно-экстремальный алгоритм, информационный критерий, оптимизация, тестирование, уровень знаний.

держки принятия решений (СППР) для преподавателя, будем оценивать усредненным по алфавиту классов распознавания информационным критерием функциональной эффективности (КФЭ):

$$\overline{E} = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K} E_{m,k}\ , \qquad (1)$$

где $E_{m,k}$ – информационный КФЭ обучения СППР распознавать реализации класса $X^{o}_{m,k}$.

Пусть по результатам тестового контроля осуществлена фазификация входных данных путем отражения результатов тестирования с помощью соответствующей оценочной функции на шкалу оценивания и сформирована неклассифицированная обучающая матрица, $\| y^{(j)}_{i,k} | i = \overline{1,N};\ j = \overline{1,n} \|$, где N, n – количество признаков распознавания и реализаций образа соответственно. В матрице $\| y^{(j)}_{i,k} \|$ строка есть вектором-реализацией образа $\{y^{(j)}_{m,i,k} | i = \overline{1,N}\}$, координаты которого – признаки распознавания – оценки за решение N тестов, а столбец – обучающая выборка $\{y^{(j)}_{m,i,k} | j = \overline{1,n}\}$, состоящая из полученных оценок за решение i-го теста. Дан структурированный вектор пространственно-временных параметров функционирования СППР $g =< g_1, ..., g_{\xi_1}, ..., g_{\Xi_1} >$, влияющие на функциональную эффективность обучения системы. При этом известны ограничения на соответствующие параметры функционирования $R_{\xi_1}(g_1, ..., g_{\xi_1}, ..., g_{\Xi_1}) \le 0$. Необходимо:

- для априорно неклассифицированного распределения векторов реализаций по дистанционным критериям сходства построить нечеткое разбиение пространства признаков на классы распознавания, и на этапе информационно-экстремального обучения трансформировать его путем оптимизации параметров функционирования СППР по информационному критерию (1) в четкое разбиение эквивалентности классов, т.е. построить безошибочные по обучающей матрице решающие правила;
- с целью персонализации знаний студента определить принадлежность реализации распознаваемого образа одному из классов заданного алфавита $\{X^{o}_{m,k}\}$, характеризующего уровень знаний учащегося по соответствующей шкале оценивания. Таким образом, задача информационного синтеза обучающейся СППР сводится к задаче оптимизации параметров функционирования, влияющих на функциональную эффективность СППР, путем итерационной процедуры поиска глобального максимума информационного КФЭ (1) обучения системы для эффективного управления и сопровождения учебного процесса.

Категориальная модель

Такая модель кластеризации данных приведена на рис. 1 в виде диаграммы отображения множеств, применяемых в процессе функционирования КСО уровня знаний учащихся. Здесь оператор кластер-анализа $\Phi_1: G \times T \times \Omega \times Z \to \to \widetilde{\Re}'$ где G – множество действующих на КСО факторов; T – моменты времени считывания информации; Ω – пространство признаков распознавания и Z – пространство возможных состояний КСО, строит по дистанционным критериям близости нечеткое разбиение $\widetilde{\Re}'$ классов распознавания, а операторы Φ_2 и Φ_3 соответственно формируют входную нечеткую обучающую матрицу Y и бинарную обучающую матрицу X.

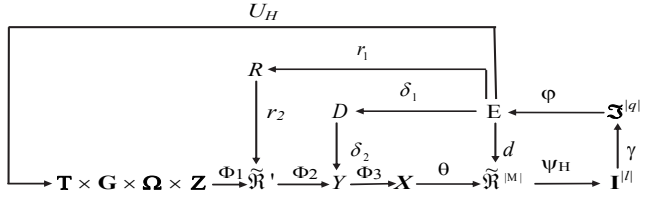

Рис. 1. Категориальная модель СППР в режиме кластер-анализа

В диаграмме (см. рис. 1) оператор θ оптимизирует в процессе обучения КСО геометрические параметры разбиения $\widetilde{\Re}^{|M|}$, а оператор $\Psi_H : \Re^{|M|} \to I^{|I|}$, где $I^{|I|}$ – множество гипотез, проверяет основную статистическую гипотезу $\gamma_1 : y^{(j)}_{m,i} \in X^{o}_{m}$ при альтернативной гипотезе $\gamma_2 : y^{(j)}_{m,i} \notin X^{o}_{m}$. Оператор $\gamma: I^{|I|} \to \mathfrak{I}^{|q|}$ по результатам оценки статистических гипотез формирует

множество $\Im^{|q|}$, где $q = l^2$ – количество точностных характеристик. Оператор $\varphi: \Im^{|q|} \to E$ формирует терм-множество E, состоящее из значений информационного КФЭ – функционала точностных характеристик. Контур оптимизации геометрических параметров разбиения $\widetilde{\Re}^{|M|}$ путем поиска максимума КФЭ обучения СППР замыкается оператором d: $E \to \widetilde{\Re}^{|M|}$. Контур оптимизации контрольных допусков содержит терм-множество D – значения системы контрольных допусков на признаки распознавания. Показанная на рис. 1 диаграмма отличается от диаграммы, приведенной в работе [4] тем, что в нее дополнительно введен контур построения разбиения $\widetilde{\Re}'$, который замыкается через терм-множество R допустимых радиусов контейнеров классов распознавания, восстанавливаемых в радиальном базисе пространства признаков распознавания.

Алгоритм кластеризации данных

Пусть априорное распределение результатов тестирования уровня знаний учащихся состоит из реализации четырех классов (класс X_5^o – «отлично», класс X_4^o – «хорошо», класс X_3^o – «удовлетворительно», класс X_2^o – «неудовлетворительно»). Анализ такого распределения позволяет сделать два допущения, упрощающих задачу кластеризации входных данных:

- мощность алфавита классов ограничена и равна $Card\{X_m^o\} = 4$;
- алфавит классов распознавания есть упорядоченным, поскольку двоичный эталонный вектор класса X_2^o – ближайший к вершине нулевого вектора-реализации (значения всех признаков находятся вне своих контрольных допусков, так как все ответы на тесты считаются ошибочными), и эталонный вектор класса X_5^o – ближайший к вершине единичного вектора-реализации (значения всех признаков находятся в своих контрольных допусках, так как все ответы на тесты считаются правильными). Кроме того, класс X_3^o – ближайший к классу X_2^o, а класс X_4^o – к классу X_5^o.

Информационно-экстремальный алгоритм обучения СППР с кластеризацией входных данных согласно категориальной модели (см. рис. 1) состоит в преобразовании неструктурированной (неклассифицированной) входной обучающей матрицы $\| y_i^{'(j)} \mid i = \overline{1,N}; j = \overline{1,n} \|$ в априорно нечеткую классифицированную многомерную матрицу $\| y_{m,i}^{(j)} \mid m = \overline{1,M}; i = \overline{1,N}; j = \overline{1,n} \|$ и отображении ее в дискретное пространство признаков распознавания, где путем допустимых целенаправленных преобразований входное математическое описание адаптируется с целью максимизации полной вероятности принятия правильных решений СППР в режиме экзамена. Рассмотрим схему иерархического агломеративного алгоритма кластер-анализа входных данных, позволяющего преобразовать неструктурированную входную обучающую матрицу в нечеткую классифицированную многомерную матрицу:

1. Формируется двоичный единичный вектор $x_5^{(1)}$ размерности N и аналогично – нулевой вектор $x_2^{(0)}$.

2. Обнуляется счетчик шагов изменения радиуса таксона: $r := 0$.

3. Инициализация счетчика шагов приращения радиуса: $r := r + 1$.

4. В вершине вектора $x_5^{(1)}$ строится таксон $T_5^{(1)}$ радиуса r.

5. Если имеет место $x^{(j)} \in T_5^{(1)}$, то выполняется п. 6, иначе – п. 3.

6. В таксоне $T_5^{(1)}$ по дистанционной мере $d[x_5^{(1)} \oplus x^{(j)}]$ определяется ближайший к единичному вектор $x_{5,\min}^{(j)}$, вершина которого принимается за центр нового таксона $T_5^{'}$, и выполняется п. 2.

Аналогично находим ближайший к нулевому вектор $x_{2,\min}^{(j)}$, вершина которого принимается за центр нового таксона $T_2^{'}$. Далее для каждого из таксонов $T_2^{'}$ и $T_5^{'}$ запускается агломеративний алгоритм поиска соответствующих цен-

тров тяжести [1]. При этом происходит инициализация счетчика шагов приращения радиусов таксонов, которая прекращается при условии $r \le d[x_{2,\min} \oplus x_{5,\min}]/2$. Выполнение этого условия позволяет построить на верхнем иерархическом уровне (рис. 2) контейнеры классов $X_2^{'}$ и $X_5^{'}$, включающие в себя все векторы-реализации из заданного распределения.

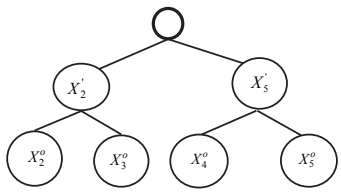

Рис. 2. Иерархическая структура классов распознавания

Для перехода на нижний уровень иерархической структуры (см. рис. 2) необходимо определить центры тяжести классов X_3^o и X_4^o. С этой целью для агломерации реализаций класса X_3^o выбран исходный вектор $x_3^{(j)}$ из условий

$$1 - \frac{2d[x_5^{'} \oplus x_3^{(j)}]}{d[x_{2,\min} \oplus x_{5,\min}]} = 0; \qquad (2)$$

$$\{d[x_2^{'} \oplus x_3^{(j)}] + [x_5^{'} \oplus x_3^{(j)}]\} \to \min_{\{x_3^{(j)}\}}. \qquad (3)$$

Условие (2) позволяет выбрать для процесса агломерации только те реализации, которые находятся на поверхности контейнера класса $X_5^{'}$, а условие (3) выбирает среди них реализацию, ближайшую к центрам классов $X_2^{'}$ и $X_5^{'}$. Аналогично начальный вектор $x_4^{(j)}$ определяется из условий

$$1 - \frac{2d[x_2^{'} \oplus x_4^{(j)}]}{d[x_{2,\min} \oplus x_{5,\min}]} = 0;$$

$$.\{d[x_2^{'} \oplus x_4^{(j)}] + [x_5^{'} \oplus x_4^{(j)}]\} \to \min_j.$$

При этом в процессе агломерации реализаций классов X_3^o и X_4^o соответственно задавались следующие ограничения на радиусы таксонов:

$$r_3 \le \frac{d[x_3 \oplus x_2^{(0)}]}{2};\ r_4 \le \frac{d[x_4 \oplus x_5^{(1)}]}{2}.$$

Кроме того, одним из ограничений есть выполнение условия, чтобы реализации, формировавшие таксон для класса X_3^o принадлежали классу $X_2^{'}$, а для класса X_4^o – классу $X_5^{'}$ (см. рис. 2). Для построения таксона класса X_2^o использовалась в качестве начальной реализация $x_{2,\min}$, а для построения таксона класса X_5^o – реализация $x_{5,\min}$, поскольку эти реализации с наибольшей вероятностью относятся к соответствующим классам. При этом на радиусы таксонов классов X_2^o и X_5^o накладываются соответственно следующие ограничения:

$$r_2 \le \frac{d[x_2 \oplus x_2^{(0)}]}{2};\ r_5 \le \frac{d[x_5 \oplus x_5^{(1)}]}{2}.$$

После построения нечеткого разбиения $\widetilde{\Re}'$ формируется входная нечеткая классифицированная обучающая матрица $\| y_{m,i}^{(j)} \|$ и запускается информационно-экстремальный алгоритм обучения СППР, позволяющий построить безошибочные (по обучающей матрице) решающие правила [4].

Таким образом, обучение СППР по информационно-экстремальному алгоритму состоит в итерационном поиске глобального максимума информационного КФЭ обучения в рабочей (допустимой) области определения его функции.

Пример реализации алгоритма кластер-анализа

Предложенный алгоритм кластер-анализа был программно реализован для оценки знаний студентов по учебной дисциплине «Интеллектуальные системы», которая читается в Сумском государственном университете студентам специальности «Информатика». Общее количество реализаций четырех классов распознавания X_2^o, X_3^o, X_4^o и X_5^o составляло 488, а количество тестов, определяющее мощность словаря признаков распознавания, равнялось $N = 141$. После формирования нечеткой классифицированной обучающей матрицы по приведенному таксонометрическому алгоритму был реализован базовый информационно-экстремальный алгоритм, оптимизирующий радиусы контейнеров классов распознавания [4]

$$d_m^* = \arg \max_{G_E \cap G_d} E_m, \qquad (4)$$

где E_m – информационный КФЭ обучения СППР распознавать реализации класса X_m^o; G_E, G_d – области допустимых значений КФЭ и радиусов контейнеров соответственно.

В качестве КФЭ обучения СППР рассматривалась модифицированная информационная мера Кульбака [4]

$$E_m^{(k)} = 0,5\log_2\left(\frac{D_{1,m}^{(k)} + D_{2,m}^{(k)}}{\alpha_m^{(k)} + \beta_m^{(k)}}\right)\times$$

$$\times\left[(D_{1,m}^{(k)} + D_{2,m}^{(k)}) - (\alpha_m^{(k)} + \beta_m^{(k)})\right], \qquad (5)$$

где $\alpha_m^{(k)}$ – ошибка первого рода, вычисляемая для k-го восстанавливаемого контейнера класса X_m^o; $\beta_m^{(k)}$ – ошибка второго рода; $D_{1,m}^{(k)}$ – первая достоверность; $D_{2,m}^{(k)}$ – вторая достоверность. При этом нормированный КФЭ представлялся в виде

$$E_m^{*(k)} = \frac{E_m^{(k)}}{E_{\max}}, \qquad (6)$$

где $E_{\max}$ – значение критерия (5) при $D_{1,m}^{(k)} = D_{2,m}^{(k)} = 1$ и $\alpha_m^{(k)} = \beta_m^{(k)} = 0$.

Реализация алгоритма (4) показала, что максимальные значения нормированных КФЭ обучения СППР не достигли своих предельных значений ($E_5^* = 0,87$; $E_3^* = 0,69$; $E_4^* = 0,75$ и $E_2^* = 0,68$). При этом радиусы восстановленных контейнеров соответственно равнялись $d_5^* = 71$, $d_4^* = 75$, $d_3^* = 87$ и $d_2^* = 51$ (здесь и далее в кодовых единицах), а среднее значение радиусов этих контейнеров равнялось $d_{\text{сред}}^* = 71$.

Таким образом, реализация базового алгоритма обучения не позволила построить безошибочные по обучающей матрице решающие правила, поскольку СКД на признаки распознавания была неоптимальной. Поэтому согласно принципу отложенных решений Ивахненко был реализован итерационный алгоритм обучения СППР с параллельной оптимизацией СКД на признаки распознавания.

На рис. 3 показан график зависимости усредненного нормированного критерия Кульбака (6) от параметра δ поля контрольных допусков, полученный в процессе обучения СППР с параллельной оптимизацией СКД. Анализ рис. 3 показывает, что оптимальное значение параметра поля контрольных допусков на признаки распознавания равно $\delta^* = \pm 11$ (в единицах 100-балльной шкалы оценивания) при максимальном предельном значении усредненного КФЭ ($\overline{E} = 1$), что свидетельствует о построении безошибочных по обучающей матрице решающих правил.

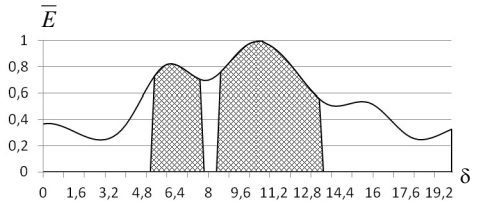

Рис. 3. График зависимости усредненного нормированного критерия Кульбака от параметра поля контрольных допусков

На рис. 4 показаны результаты восстановления контейнеров классов распознавания, полученные при оптимальном значении параметра поля контрольных допусков $\delta^* = \pm 11$.

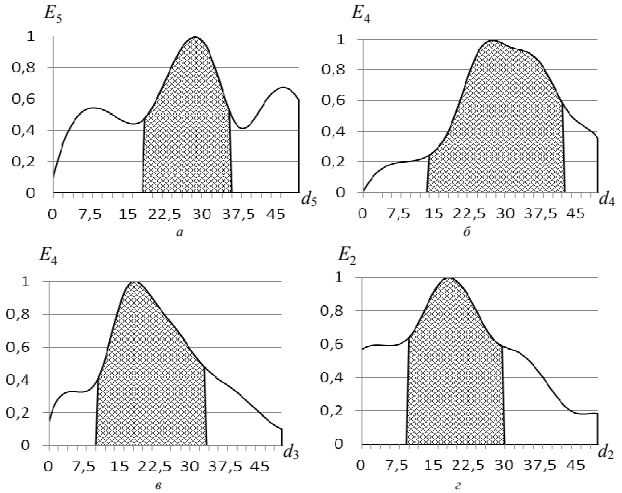

Рис. 4. Графики зависимости КФЭ от радиусов контейнеров классов распознавания: *а* – класс X_5^o; *б* – класс X_4^o; *в* – класс X_3^o; *г* – класс X_2^o

Анализ рис. 4 показывает, что оптимальные параметры контейнеров, определенные при максимальных предельных значениях КФЭ в рабочей области определения их функции (5), равны в кодовых единицах соответственно $d_5^* = 28$, $d_4^* = 26$, $d_3^* = 16$ и $d_2^* = 18$. При этом среднее

значение радиусов контейнеров равно $d^*_{\text{сред}} = 22$, т.е. существенно меньше в сравнении со значением, полученным по базовому алгоритму, что соответствует минимально-дистанционному принципу теории распознавания образов [6].

Заключение. Предложенный информационно-экстремальный метод кластер-анализа данных позволяет для структурированного алфавита классов распознавания построить по дистанционным критериям сходства априорно нечеткое разбиение пространства признаков на классы распознавания и трансформировать его в процессе обучения в четкое разбиение классов эквивалентности. При этом для заданной мощности структурированного алфавита классов распознавания в процессе обучения системы построены безошибочные по обучающей матрице решающие правила, что позволяет повысить достоверность результатов машинного оценивания уровня знаний учащихся.

1. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Ин-та математики, 1999. – 259 с.
2. *Мандель И.Д.* Кластерный анализ. – М.: Финансы и статистика, 1988. – 176 с.
3. *Jain A.K.*, *Dubes R.C.* Algorithm for Clustering Data. – Engelwood Cliffs, New Jersey Prentice Hall, 1988. – 367 p.
4. *Довбиш А.С.* Основи проектування інтелектуальних систем: Навч. посібник. – Суми: Вид-во СумДУ, 2009. – 171 с.
5. *Довбиш А.С.*, *Востоцький В.О.* Інформаційно-екстремальна система підтримки прийняття рішень у режимі кластер-аналізу // Вісн. Сум. держ. ун-ту. Сер. Техн. науки. – 2010. – № 1. – С. 73–78.
6. *Ту Дж.*, *Гонсалес З.* Принципы распознавания образов. – М.: Мир, 1978. – 401 с.

E-mail: *kras@id.sumdu.edu.ua*, *saad*710*@mail.ru*