Марков Владимир Васильевич – e-mail: vmarkov@sfedu.ru; кафедра систем автоматизированного проектирования; доцент.

Kuliev Elmar Valerievich – Southern Federal University; e-mail: ekuliev@sfedu.ru; 44, Nekrasovskiy lane, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; associate professor.

Tsyrulnikova Elvira Sergeevna – e-mail: bolshova@sfedu.ru; the department of computer aided design; graduate student.

Kulieva Nina Vladimirovna – e-mail: holopova@sfedu.ru; the department of computer aided design; graduate student.

Markov Vladimir Vasilievich – e-mail: vmarkov@sfedu.ru; the department of computer aided design; associate professor.

УДК 004.822 DOI 10.23683/2311-3103-2019-4-89-102

В.В. Бова, Ю.А. Кравченко

БИОИНСПИРИРОВАННЫЙ ПОДХОД К РЕШЕНИЮ ЗАДАЧИ КЛАССИФИКАЦИИ ПРОФИЛЕЙ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ В ИНТЕЛЛЕКТУАЛЬНЫХ ИНТЕРНЕТ-СЕРВИСАХ*

Рассматриваются проблемы повышения эффективности организации личностно-ориентированного взаимодействия пользователя в интеллектуальных Интернет-сервисах для формирования культуры безопасного поведения в Интернет-пространстве. Для их решения предлагается метод построения профилей поведения пользователей на основании анализа их информационных потребностей, максимально соответствующих их предпочтениям, в том числе и неявным. Актуальность работы определяется растущей популярностью идеи персонализации контента и информационных услуг в Интернет-пространстве. Профиль поведения пользователя рассматривается авторами как слабоформализованный объект в пространстве признаков внутренних и внешних характеристик, описывающих его взаимодействие с Интернет-ресурсом. Предлагаемый в работе метод основан на вероятностном алгоритме EM- кластеризации исследуемых данных о характеристиках пользователя и распределенных Интернет-ресурсов для генерации структуры входных параметров классификаторов модели формирования профиля пользователя. Оптимизация структуры реализуется механизмом отбора информативных признаков профиля, основанного на идее выявления скрытых интересов и предпочтений пользователей с одной стороны, и способностью ресурса удовлетворять заинтересованных пользователей этому набору признаков – с другой. Для снижения размерности исходных данных признакового пространства в задаче классификации предлагается метаэвристический алгоритм оптимизации «кукушкин поиск», отличающийся масштабируемостью и высокой интерпретируемостью выходных данных. Оптимизация параметров классификаторов заключается в подборе параметров функции принадлежности и меток классов обобщенного профиля таким образом, чтобы численный критерий точности классификации признаков ресурсов и предпочтений пользователей сводился к максимуму на реальных данных. Для оценки эффективности предложенного алгоритма проведен вычислительный эксперимент на тестовых наборах данных из открытого репозитория UCI Machine Learning Repository. Результаты которого показали, что построенный на тестовых данных классификатор обладает более высоким уровнем интерпретируемости полученных результатов формирования профилей, сохраняя точность классификации.

Классификация; интеллектуальные Интернет-сервисы; профиль поведения пользователя; EM-алгоритм; биоинспирированные методы; метаэвристика «кукушкин поиск».

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18–29–22019.

V.V. Bova, Y.A. Kravchenko

BIOINSPIRED APPROACH FOR SOLVING THE PROBLEM OF CLASSIFICATION IN USERS BEHAVIOR PROFILES IN INTELLIGENT INTERNET SERVICES

The article deals with the problems of increasing the effectiveness of personal-oriented user interaction organization in intelligent Internet services for developing a culture of safe behavior in the Internet space. To solve them, we propose a method for constructing profiles of user behavior based on an analysis of their information needs, which most closely match their preferences, including the implicit ones. The relevance of the work is determined by the growing popularity of the idea of personalizing content and information services in the Internet space. The user's behavior profile is considered by the authors as a weakly formalized object in the feature space of internal and external characteristics describing its interaction with an Internet resource. The proposed method is based on the EM-clustering probabilistic algorithm of the studied data on user characteristics and distributed Internet resources for generating the structure of the input parameters of classifiers of the user profile generation model. The structure optimization is realized by the mechanism of selection of informative characteristics of a profile based on the idea of identifying hidden interests and preferences of users on the one hand, and the ability of a resource to satisfy interested users of this feature set - on the other. In order to reduce the dimension of the source data of the attribute space, the classification problem suggests a meta-heuristic algorithm for optimizing the "cuckoo search", which is distinguished by scalability and high interpretability of the output data. Optimization of the parameters of classifiers consists in the selection of the parameters of the membership function and the labels of classes of the generalized profile so that the numerical criterion for the accuracy of the classification of resource attributes and user preferences is reduced to a maximum on real data. To assess the effectiveness of the proposed algorithm, a computational experiment was conducted on test datasets from the open repository of the UCI Machine Learning Repository. The results of which showed that the classifier built on test data possesses a higher level of interpretability of the results of the formation of profiles, while maintaining the accuracy of classification.

Classification; intelligent Internet services; behavioral user profile; EM-algorithm; bioinspired methods; cuckoo search metaheuristics.

Введение. Развитие технологий взаимодействия человека с Интернет-пространством, возникновение в нем новых видов информационно-образовательной деятельности и форм виртуальной коммуникации свидетельствует о том, что Интернет-пространство стало новой средой социализации личности обучающегося [1–3]. Проблемы развития глобальной информационной инфраструктуры и ее влияние на личность имеют дуальную сущность. С одной стороны информационные возможности Интернет-среды, перенос активности учебной и самообразовательной деятельности из реального пространства в виртуальное создают благоприятные условия для самореализации и саморазвития личности [4]. С другой, ведут к возникновению рисков виктимного воздействия, нравственной и психологической дезориентации личности, обусловленных лавинообразном ростом распределенных Интернет-ресурсов (ИР) различной направленности, большим объемом информации, содержащейся в Интернет-пространстве, развитием контента сомнительного качества. При этом нужно отметить, что виктимное влияние сугубо индивидуально и зависит от способности индивида к самостоятельному, осознанному выбору информации, подходящей его интересам, убеждениям и ценностям [5]. Одним из подходов к повышению эффективности организации личностно-ориентированного взаимодействия пользователя и ИР является построение поведенческих профилей пользователей информационно-образовательного Интернет-пространства на основании анализа их информационных потребностей, максимально соответствующих их предпочтениям, в том числе и неявным.

С целью эффективного извлечения требующихся данных об активности пользователей применяются методы интеллектуального анализа данных. При работе с большими объемами неструктурированных данных в Интернет-сервисах

возникает необходимость осуществлять поиск скрытых закономерностей, извлечения, структуризации и классификации тематических категорий, отражающих интересы и предпочтения пользователей образовательного контента [6–9], что требует создания новых методов и моделей автоматизации этих процессов. В работе предлагается метод классификации поведенческих профилей пользователей для эффективного решения задачи оптимизации и сопровождения безопасной информационно-образовательной деятельности в Интернет-пространстве. Оптимизация параметров классификаторов профилей выполняется с помощью метаэвристического алгоритма поиска кукушки.

1. Проблематика исследования. Стремительный рост образовательных ИР, объемы потоковых и пользовательских данных, поступающих на обработку и накапливаемых в Интернет-сервисах достигли огромных размеров [1–3]. Актуальной проблемой становится создание новых методов для эффективного извлечения полезной информации из большого объема скрытых данных об интересах и предпочтениях пользователей, необходимых для решения ряда аналитических задач (рис. 1), связанных с моделированием Интернет-поведения пользователей в социально значимых целях формирования безопасного поведения в Интернет-пространстве [3, 8].

Рис. 1. Задачи моделирования Интернет-поведения пользователей

Основными проблемами, связанными с реализацией этих задач и их практическим использованием при построении поведенческого профиля пользователя на основании предпочтений, персонифицирующих образовательную деятельность в Интернет-сервисах, являются разреженность и масштабируемость исходных данных о потребностях и требованиях, предъявляемым к ИР и информационным услугам. Сложность проблемы извлечения полезных знаний о пользовательских предпочтениях и их обработки связана с большими объемами и неоднородностью накапливаемых данных, а также их быстрым изменениям (или обновлениям) во времени [6–9].

При анализе больших объемов данных и поиске в них скрытых закономерностей возникает проблема отбора информативных признаков для разработки модели классификации [7]. С одной стороны, необходимо, чтобы выполнение поиска происходило за приемлемое время, а с другой – чтобы существенные данные не были потеряны. Отбор большого количества признаков приводит к увеличению вычислительной сложности модели, и как следствие к необходимости применения значительных вычислительных ресурсов и затрат времени. Отказ от признаков,

кажущихся несущественными, или проявляющиеся на уровне шума, может привести к потере значимой информации. Таким образом, определение связи шумовых признаков с целевой переменной в решаемой задаче, является также существенной проблемой [10, 11].

Необходимость снижения размерности исходных данных, удаление зашумленных и избыточных признаков для интерпретации этих данных определяет актуальность исследования и невозможна без привлечения методов интеллектуального анализа данных, машинного обучения и биоинспирированных методов поиска оптимальных решений [10–14].

2. Постановка задачи. В общем виде постановка задачи исследования может быть сформулирована следующим образом. Пусть *Res* и *User* – множества ресурсов и пользователей соответственно. Предположим, что у каждого пользователя $u \in User$ имеется T – множество предпочтений (интересов), называемыми далее темами $t \in T$, представленными в $r \in Res$.

Для сравнения профилей пользователей и ресурсов вводится понятие обобщенного тематического профиля ресурса и пользователя, объединяющего в себе и признаки ресурсов, и предпочтения пользователей [8]. Вектор числовых значений параметров которого представлен как: $\theta_j = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}, \beta_{j1}, \beta_{j2}, \ldots, \beta_{jm})$, где $\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}$ – вес i-признака для $r_i \in Res$, $\beta_{j1}, \beta_{j2}, \ldots, \beta_{jm}$ – вес j-признака для $u_j \in User$.

Пусть X – множество объектов с заданной на нем вероятностной мерой P для $p(u,r)$ – вероятность выбора ресурса r пользователем u. Задача генерации структуры классификатора заключается в том, чтобы, зная выборку объектов $X_i = \{x_1, \ldots, x_z\} \subset X$ и число k оценить вектор параметров $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$. Для этого предположим, что плотность распределения информативных признаков на X имеет вид k распределений (классификаторов): $p(x) = \sum_{j=1,k} w_j p_j(x_i), \sum_{j=1}^{k} w_j = 1$, где $p_j(x_i) = max\ f(x_i, \theta_j)$ – функция правдоподобия j-го классификатора, w_j – её априорная вероятность, определяемая как $p(u,r)$.

Оптимизация параметров заключается в подборе параметров функции принадлежности и меток выходов классификаторов таким образом, чтобы критерий точности классификации сводился к *max* на обучающих данных:

$$E(\theta) = \frac{\sum_{i=1}^{z} \begin{cases} 1, \text{если } p(x_i) = f(x_i, \Theta) \\ 0, \text{иначе} \end{cases}}{z} \rightarrow max, \quad (1)$$

где $f(x_i, \Theta)$ – функция, описывающая классификатор, Θ – вектор параметров поведенческого профиля, z – число обучающих данных.

3. Метод построения профиля поведения пользователей. Разработка метода основана на последовательном решении следующих задач (рис. 2):

1) отбор информативных признаков профиля, основанный на идее выявления скрытых интересов и предпочтений пользователей;

2) генерация структуры классификатора (векторного описания структурных характеристик) профилей пользователя и ИР для формирования профилей групп пользователей со схожими тематическими предпочтениями на основе вероятностного EM-алгоритма;

3) оптимизация параметров классификатора с помощью метаэвристики «кукушкин поиск», направленная на снижение размерности пространства признаков из неявных данных большого объема о предпочтениях пользователя.

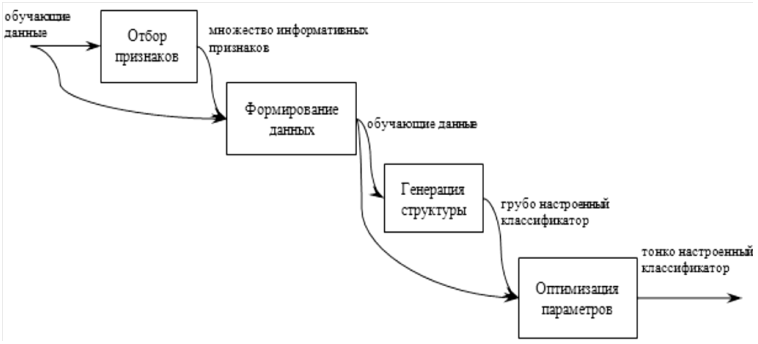

Рис. 2. Процесс построения классификатора профиля

Первоначальным этапом решения задачи data mining является процесс конструирования признаков (feature engineering) [11]. В рассматриваемом случае, в качестве объектов выступают пользователи Интернет-сервисов, а в качестве признаков – тематические категории посещенных ими Интернет-ресурсов. Такое признаковое описание позволяет формировать поведенческие (тематические) профили пользователей – векторное представление его интересов и тематических предпочтений. Тематический профиль пользователя рассматривается авторами как слабоформализованный объект в пространстве признаков внутренних и внешних характеристик поведения, описывающих его взаимодействие с ИР [12–15]. Классификация таких объектов представляет собой упорядочение и разработку схемы соотношения параметров объекта, данных о нем и его результирующего поведения в различных условиях.

После окончания этапа конструирования признаков, как правило, производят отбор подмножества наиболее информативных и достоверных признаков для построения модели [15]. Это позволяет снизить объемы обрабатываемой информации, избежать переобучения и в целом улучшить качество модели. В нашем случае будем группировать ресурсы по тематическим категориям, в предположении о том, что пользователи, интересующиеся одной тематикой, могут получать информацию из различных ИР. Признаки, встречающиеся не более чем у одного пользователя, будем считать неинформативными.

Начальная информация о признаковом пространстве используется для отбора признаков на этапе предобработки (предварительной фильтрации) для последующей генерации структуры классификатора [11, 14]. Далее классификатор тематического профиля будет уточняться по мере того, какие действия по отношению к ИР будут выполняться со стороны пользователей (действия по выбору ресурсов).

На этапе отбора признаков на основе обучающих данных формируется множество информативных признаков, представленных векторами значений вероятностей наблюдаемых оценок предпочтения выбора ИР пользователем. Целью отбора признаков является: избежать переобучения модели; уменьшить объем анализируемых данных; улучшить эффективность классификации; устранить нерелевантные и шумовые признаки; повысить интерпретируемость полученных результатов [8, 11].

На основе методов коллаборативной фильтрации [16] данные приводится к матричному представлению обучающей выборки в матрице факторизации предпочтений $D = (u_i, r_i)_{i=1}^{l}$, где l – размерность пространства признаков предпочтений.

Допустим, пользователь u имеет интерес $t \in T$ с вероятностью $p(t|u)$. Тогда профиль u определим как вектор значений вероятностей $p(t|u)$, $t = 1,\ldots,|T|$, причем $\sum_{t \in T} p(t|u) = 1$. В свою очередь, каждый ресурс r соответствует некоторому множеству тем $t \in T$ и удовлетворяет им с вероятностью $p(t|r)$, таким образом, про-

филь r – вектор вероятностей $p(t|u)$, $t = 1,\ldots,|T|$, а $\sum_{t\in T} p(t|r) = 1$. А w_j – вероятность прогноза, что выбранным будет ИР r, если выполняется $p(t|r)$ и $p(t|u)$. Тогда $p(u|r)$ – вероятность выбора ресурса r пользователем u можно представить как:

$$p(u,r) = \sum_{t\in T} p(t|r) \times p(t|u). \quad (2)$$

Для восстановления признакового пространства профилей $p(t|u)$ и $p(t|r)$ в случае разреженных данных и исключения избыточных признаков, не влияющих на результаты классификации в обучающей выборке D, воспользуемся Expectation-Maximization (EM) алгоритмом разделения смеси распределений [18]. Область применения данного алгоритма достаточна широка, он используется не только для кластеризации данных, но и в дискриминантном анализе, а также для восстановления пропусков в данных [19]. В основе алгоритма лежит методика итерационного вычисления оценок максимального правдоподобия функции, описывающей классификатор.

На этапе генерации структуры, для того чтобы получить первоначальное представление о структуре исследуемых данных, введем вспомогательный вектор скрытых переменных Q. EM-алгоритм опирается на итерационное повторение шагов: E-шаг – вычисляет ожидаемое значение Q по текущему приближению Θ – вектора параметров профиля, а на M-шаге решается оптимизационная задача максимизации правдоподобия и находится следующее приближение вектора Θ по текущим значениям Q и Θ.

На E-шаге определим через $p(x_i \,\&\theta_j)$ совместную плотность вероятности того, что получен объект x_i и этот объект сгенерирован j-й компонентой смеси: $p(x_i \,\&\theta_j) = p(x_i)P(\theta_j|x_i) = w_j\, p_j(x_i|\theta_j)$.

Пусть вектор $q_{ij} \equiv P(\theta_j|x_i)$ определяет апостериорную вероятность генерации обучающего объекта x_i j-й компонентой смеси. Примем эти величины как скрытые переменные. Обозначим $Q = (q_{ij})_{m\times k} = (q_1,\ldots, q_j)$, где q_j – j-й столбец матрицы Q в предположении, что каждый объект может быть сгенерирован единственной компонентой. Тогда для всех i, j, зная параметры w_j и θ_j, представим q_{ij} согласно:

$$q_{i,j} = \frac{w_j\, p_j(x_i,\theta_j)}{\sum_{s=1}^{k} w_s\, p_s(x_i,\theta_s)}. \quad (3)$$

На M-шаге приходим к k независимым задачам максимизации взвешенного правдоподобия для формирования вектора параметров профиля θ_j.

$$f(x) = F(X^z, \Theta) = \sum_{i=1}^{z} \ln \sum_{j=1}^{k} q_{i,j}\, p_j\left(x_i, \theta_j\right) \to \max_{\Theta}. \quad (4)$$

Объекты обучающей выборки учитываются с весами q_{ij} и их распределение свое для каждой из k компонент.

3. Метаэвристический алгоритм «кукушкин поиск». Для оптимизации параметров классификаторов применяются две группы методов. К первой относятся классические методы оптимизации, основанные на производных, например, метод наименьших квадратов [11]. Эти методы дают точные результаты, но они имеют тенденцию сходиться к локальным оптимумам. Трудности применения классических методов оптимизации, в частности проблема локального экстремума, определяют необходимость обратиться ко второй группе методов – метаэвристических, таких как алгоритмы роевого интеллекта, стайные алгоритмы. Алгоритмы, инспирированные природными системами эффективны при решении нелинейных, многомерных, многокритериальных задач оптимизации с ограничениями [20, 21].

В работе для решения задачи построения классификатора наилучшей точности, сохраняющего при этом хорошую обобщающую способность на больших данных, предлагается использовать популяционный алгоритм «кукушкин поиск». Принимая во внимание, что реальное признаковое пространство Θ имеет большую

размерность и может быть разреженным, а самих объектов достаточно много, то оптимизация параметров алгоритма (его обучение) будет проводиться на обучающей выборке – некоторому конечному подмножеству $X_i \subseteq X$.

Алгоритм «кукушкиного поиска» (англ. Cuckoo Search, CS) был предложен в 2009 г. [22]. CSA вдохновлен поведением кукушек в процессе их вынужденного гнездового паразитизма и фактически представляет собой имитацию процесса поиска случайных блужданий, перенося стратегию поведения кукушек на пространство поиска оптимума. Кукушки откладывают яйца в коллективные гнезда вместе с другими кукушками, и могут выбрасывать яйца конкурентов, чтобы увеличить вероятность появления их собственных птенцов. CSA основан на трех правилах [22–24]: во-первых, каждая кукушка может откладывать только одно яйцо за раз и подбрасывать свое яйцо в случайно выбранное гнездо; во-вторых, лучшие гнезда с высоким качеством яиц наследуются в следующие поколения; в-третьих, число доступных гнезд птиц других видов фиксировано. Так, если хозяин гнезда обнаружит в нем яйца чужого вида с некоторой вероятностью ρ_a *(0;1)*, то он либо выбросит эти яйца, либо просто покинет свое гнездо и соорудит на другом месте новое гнездо [21]. В CSA каждое яйцо в гнезде представляет собой решение, а яйцо кукушки – новое решение. Цель заключается в использовании новых и потенциально лучших (кукушкиных) решений с высоким значением пригодности, чтобы заменить худшие решения в гнёздах.

Общий процесс имитации поиска кукушки состоит в том, чтобы инициализировать несколько гнезд птиц, вычислить значение пригодности каждого гнезда, а затем позволить птице обновлять свое место обитания, следуя маршруту полета Леви, пока не будет найдена глобальная точка потенциально лучшего решения [22]. Схема имитации поведения кукушки на основе полетов Леви показана на рис. 3.

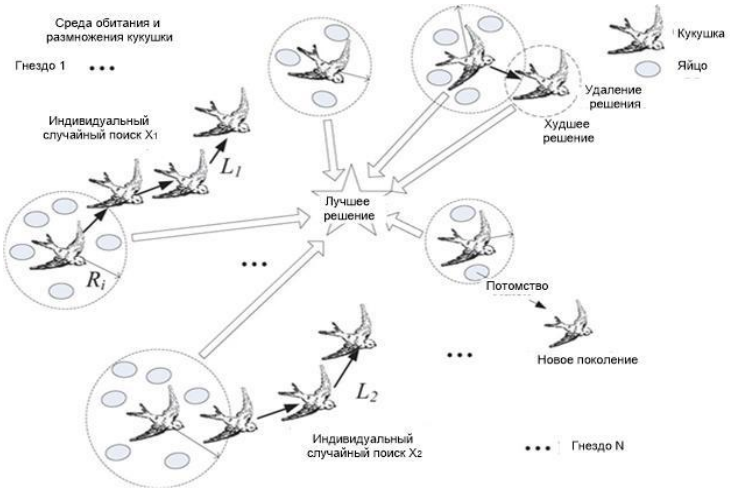

Рис. 3. Модель, имитирующая поведение кукушки

Положим, что речь идет о задаче глобальной безусловной максимизации, тогда в общем виде CSA можно описать следующим образом: 1) инициализируем популяцию S хозяйских гнезд и кукушку, т.е. определяем начальные значения компонентов векторов X_i; 2) выполняем определенное число случайных перемещений кукушки в пространстве поиска на основе полётов Леви по параметрам, описывающим масштаб области поиска решений с высоким значением пригодности [24] и вычисляем новое положение кукушки X_n; 3) случайным образом выби-

раем гнездо s_i и если $f(X_n) > f(X_i)$, то заменяем яйцо в хозяйском гнезде на яйцо кукушки, т.е. полагаем $X_i = X_n$; 4) с вероятностью ρ_a удаляем из популяции некоторое число худших случайно выбранных гнёзд (включая, возможно и гнездо s_i) далее по правилам шага 1 строим такое же число новых гнёзд; 5) если условие окончания итераций не выполнено, то переходим к шагу 2.

В каноническом CSA вероятность ρ_a и параметры полёта Леви являются фиксированными константами, а в перемещениях кукушки не учитывается информация о найденных решениях высокого качества. В целях диверсификации поиска на начальных итерациях целесообразно использовать большие значения величин ρ_a – вероятность, с которой гнездо может быть покинуто хозяином (вероятность удаления векторов решений) и ν – свободный параметр вектора вероятности полета кукушки к гнездам с высоким значением приспособленности. На завершающих итерациях для повышения точности локализации экстремума (интенсификации поисковой процедуры) применяются меньшие значения величин ρ_a и ν [23]. В работе предлагается использовать динамические значения данных параметров. Рассмотрим модифицированный CSA.

Шаг 1. Инициализация исходной популяции. Задать размер популяции S (популяция $\Theta = \{\theta^s; s = 1,2,\ldots,S\}$) и число итераций алгоритма N в качестве критерия остановки, генерация вектора начального положения кукушки θ^s, который является текущим решением и значений свободных параметров ${\rho_a}^{min}$, ${\rho_a}^{max}$, ${\nu}^{min}$, ${\nu}^{max}$.

Шаг 2. Инициализировать счетчик итераций *Counter =1* до *N*.

Шаг 3. Вычисление лучшего решения θ^{best} в популяции Θ, где *best* = $\arg\max_{0\le s\le S}(E(\theta^s))$.

Шаг 4. Проверка остановки алгоритма. Алгоритм работает итерационно, если *Counter N*, то завершить работу алгоритма найденным лучшим решением θ^{best} (*шаг 9*).

Шаг 5. Далее осуществляется итерационный процесс поиска оптимума, который состоит из генерации новых и удаления худших решений популяции. Для каждой особи популяции *s* генерируется новое решение $\theta^{s,\,new}$ путем изменения всех текущих элементов векторов решений на случайную величину полета Леви такое, что $\theta_l^{s,new} = \theta_l^{s,counter-1} + Levi_l$, где $l = 1,2,\ldots,m$ (m – размер вектора поиска θ^s); $\theta_l^{s,counter}$ – значение l-го элемента вектора популяции θ^s на итерации *Counter*; $Levi_l = \frac{\gamma_l \lambda_l}{|\nu_l|^{1/\beta}}$, γ_l – коэффициент прыжка Леви, рекомендуемое значение 0,01 [20], β = 1,5; λ_l – нормально распределенная случайная величина $\lambda_l \sim N(0; r_\lambda^2)$, $\nu_l = \nu^{max}\exp(d^n)$, $d\frac{1}{n}\ln\left(\frac{\nu^{max}}{\nu^{min}}\right)$. $r_\lambda^2 = \left\{\frac{\Gamma(1+\beta)\sin(\pi*\frac{\beta}{2})}{\Gamma\left(\frac{1+\beta}{2}\right)*2^{(\beta-1)/2}}\right\}^{1/\beta}$, где Γ(x) – гамма-функция [22].

Шаг 6. Выполнение локального поиска. После применения нового вычислительного расположения на основе случайных блужданий распределения Леви производится оценка качества нового решения. Если новый вектор решения показывает лучшую точность классификации, чем соответствующий вектор популяции, то происходит его замена на новое решение, т.е. $E(\theta^{s,\,new}) > E(\theta^{s,\,counter\,-\,1})$, то $\theta^{s,\,counter} = \theta^{s,\,new}$.

Шаг 7. После анализа решений в текущей окрестности поиска осуществляется удаление худшего решения из популяции Θ. Если $rnd \le \rho_a$, то удаляется решение такого, что $k = \arg\min_{0\le s\le S}(E(\theta^s))$ при *rnd* – случайно распределенной в интервале $[{\rho_a}^{max}, {\rho_a}^{min}]$ величины вероятности.

Шаг 8. Вместо удаленного решения новое генерируется случайным образом, переход к шагу 2.

Шаг 9. Сохранение найденного лучшего решения точности классификации.

Алгоритм возвращает вектор решения, показавший лучшее значение точности классификатора среди всех векторов популяции.

4. Экспериментальные исследования. С целью исследования производительности разработанного метода были проведены вычислительные эксперименты на наборах данных из открытого репозитория Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html [25]. Репозиторий UCI Machine Learning является крупнейшим открытым хранилищем реальных и модельных задач интеллектуального анализа данных. Для тестирования выбран набор данных User Knowledge Modeling Data Set. Предлагаемый метод сравнивался с методом машинного обучения Fuzzy C-means (Нечеткий алгоритм С-средних) с точки зрения точности классификации (на тестовых данных) и числа случаев, которые используются для обучения [7]. Задача обучить классификатор и рассчитать точность его работы на тестовой выборке. Результаты сравнительного анализа представлены в табл. 1.

Таблица 1

Сравнительный анализ точности классификации методами CSA и Fuzzy C-means

Метод	Размер обучающей выборки	Размер тестовой выборки	Точность на тестовых данных
CSA	160	40	0,941
Fuzzy C-means	160	40	0,879
CSA	120	60	0,964
Fuzzy C-means	120	60	0,867
CSA	100	80	0,935
Fuzzy C-means	100	80	0,844
CSA	90	100	0,917
Fuzzy C-means	90	100	0,833
CSA	80	120	0,914
Fuzzy C-means	80	120	0,796
CSA	60	140	0,893
Fuzzy C-means	60	140	0,770

Согласно данным табл. 1, точность классификации предложенного метода лучше результата, полученного при использовании Fuzzy C-means. При сокращении размера выборки до минимально возможных размеров (60 из 140) точность классификации CSA остается максимальной. Значит, используя этот метод, можно формировать правила на небольшом количестве тестовых данных без потери степени точности.

На основании предыдущего испытания, мы выяснили, что оптимальное число прецедентов на тестовой выборке равно 60, далее будем использовать ее для проведения серии экспериментов по определению степени влияния количества итераций на точность и время выполнения алгоритма Cuckoo Search, результаты которых отражены в табл. 2 и на рис. 4 и 5.

Таблица 2

Влияние количества итераций на точность и время выполнения алгоритмов

Количество итераций	25	50	75	100	150	200	250	500
Точность выполнения алгоритма CSA	0,75	0,86	0,91	0,95	0,96	0,95	0,98	0,98
Время выполнения алгоритма CSA (мс)	27	28	37	43	43	54	50	53
Точность выполнения алгоритма FCM	0,70	0,71	0,65	0,78	0,81	0,78	0,76	0,86
Время выполнения алгоритма FCM (мс)	36	40	44	53	59	60	64	71

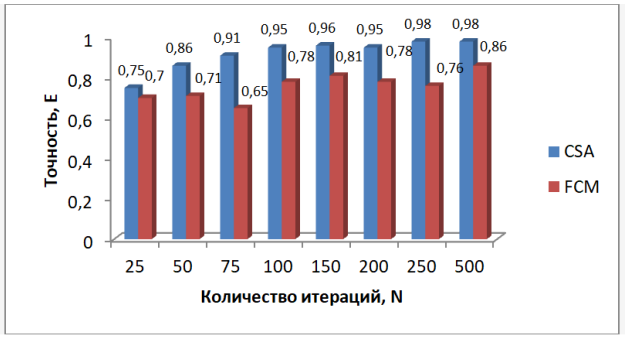

Рис. 4. Точность классификации алгоритмами Cuckoo Search и Fuzzy C-means

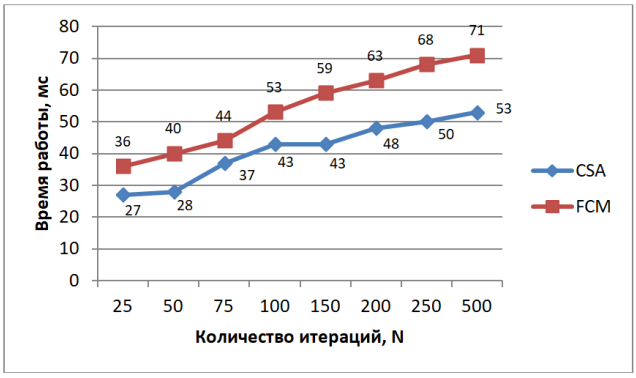

Рис. 5. График временной сложности алгоритмов Cuckoo Search и Fuzzy C-means

Согласно данным из табл. 2 предложенный алгоритм показывает сравнительно высокую точность и превосходит метод Fuzzy C-means уже начиная с 50 итерации, достигая стабильного высокого показателя. Это подтверждает гипотезу, что даже при небольшом количестве исходных данных с помощью CSA можно делать прогнозы с достаточной степенью точности. Из визуального представления полученных сравнительных результатов классификации обоими методами хорошо видно, что продолжительность работы алгоритмов растет линейно. CSA работает значительно быстрее и при увеличении количества итераций время работы увеличивается несущественно, что подтверждает его эффективность и целесообразность применения для решения задачи классификации.

Заключение. В результате проведенного исследования разработан метод классификации поведенческих (в контексте задачи – тематических) профилей пользователей, основанный на метаэвристическом подходе к оптимизации размерности пространства признаков (категорийных данных интересов и потребностей пользователей), поддающихся содержательной интерпретации для задачи персонализации образовательного контента. Для повышения точности и скорости решения задачи классификации профилей поведения пользователей разработан биоинспирированный алгоритм оптимизации параметров классификаторов профиля поведения пользователей на основе метаэвристики «кукушкин поиск», отличающийся применением суперпозиции нескольких критериев оптимальности решений с учетом зашумленности признакового пространства, что позволит повысить качество и снизить время обработки данных по сравнению с известными методами. К достоинствам разработанного алгоритма CSA следует отнести простоту реализации, возможность идентифицировать объекты, являющиеся шумовыми выбросами, хорошую обобщающую способность. Полученные результаты в серии экспериментов на открытых данных показывают, что при значительном сокращении тестовых данных предлагаемый метод сохраняет высокую степень точности классификации, что позволяет сделать заключение о перспективности его применения для решения задачи классификации в различных областях, связанных с обработкой больших объемов данных. Предложенный авторами метод позволит увеличить релевантность отбора объектов изучения в соответствии с индивидуальными особенностями, интересами и предпочтениями пользователей в Интернет-сервисах.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Tingting Z., Chen L.Y., Liang-Hsien T.* Understanding user motivation for evaluating online content: a self-determination theory perspective // Behaviour and Information Technology. – 2015. – No. 34. – P. 479-491.
2. *Deliang W., Lingling X., Chuan C.H.* Understanding the continuance use of social network sites: a computer self-efficacy perspective // Behaviour and Information Technology. – 2015. – No. 34. – P. 204-216.
3. *Алфимцев А.Н., Девятков В.В., Сакулин С.А*. Персонализация в гипертекстовых сетях на основе распознавания действий пользователей и нечеткого агрегирования // Вестник МГТУ им. Баумана. Сер. «Приборостроение». – 2012. – № 3.
4. *Писаренко В.И.* Возможности использования педагогических знаний в междисциплинарных исследованиях // Современная наука: актуальные проблемы теории и практики. Серия «Гуманитарные науки». – 2018. – № 12-3. – С. 48-52.
5. *Войскунский А., Евдокименко А., Федунина Н.* Сетевая и реальная идентичность: сравнительное исследование // Психология. Журнал Высшей школы экономики. – 2013. – Т. 10, № 2. – С. 98-121.
6. *Bova V.V., Kureichik V.V., Leshchanov D.V.* The model of semantic similarity estimation for the problems of big data search and structuring // Application of Information and Communication Technologies - AICT 2017. – P. 27-32.
7. *Газиев Г.З., Курдюкова Г.Н., Курдюков В.В.* Кластеризация Big Data для их анализа и обработки // Сб. научных статей конференции «Направления и механизмы развития науки нового времени: от теории до внедрения результатов». – 2017. – С. 150-162.
8. *Паутов К.Г., Попов Ф.А.* Метод кластеризации тематических профилей пользователей и его применение для анализа интернет-трафика // Фундаментальные исследования. – 2015. – № 74. – С. 765-769.
9. *Кравченко Ю.А.* Оценка когнитивной активности пользователя в системах поддержки принятия решений // Известия ЮФУ. Технические науки. – 2009. – № 4 (93). – С. 113-117.
10. *Курейчик В.М., Картиев С.Б.* Алгоритм классификации, основанный на принципах случайного леса, для решения задачи прогнозирования // Программные продукты и системы. – 2016. – № 2. – С. 11-15.

11. *Сарин К.С., Ворожцов С.А., Арипилов С.Н.* Построение ансамблей нечетких классификаторов на основе метаэвристики "кукушкин поиск" и горной кластеризации // Электронные средства и системы управления. – 2017. – № 1-2. – С. 26-29.
12. *Картиев С.Б., Курейчик В.М.* Разработка и исследование алгоритма решения задачи кластеризации для осуществления вопросно-ответного поиска в информационно-аналитической системе прогнозирования // Известия ЮФУ. Технические науки. – 2016. – № 7 (180). – С. 18-28.
13. *Марков В.В., Кравченко Ю.А., Кузьмина М.А.* Развитие методов семантической фильтрации на основе решения задачи кластеризации биоинспирированными алгоритмами // Известия ЮФУ. Технические науки. – 2018. – № 4 (198). – С. 175-185.
14. *Ходашинский И.А., Анфилофьев А.Е., Бардамова М.Б., Ковалев В.С., Мех М.А., Сонич О.К.* Метаэвристические методы оптимизации параметров нечетких классификаторов // Информационные и математические технологии в науке и управлении. – 2016. – № 1. – С. 73-81.
15. *Jalalirad A., Tjalkens T.* Using feature-based models with complexity penalization for selecting features // Journal of Signal Processing Systems. – 2018. – Vol. 90, Issue 2. – P. 201-210.
16. *Guo G., Zhang J., Thalmann D.* Merging trust in collaborative filtering to alleviate data sparsity and cold start // Knowledge-Based Systems. – 2014. – No. 57. – P. 57-68.
17. *Авдеенко Т.В., Макарова Е.С.* Метод определения релевантности прецедентов на основе нечетких лингвистических правил // Научный вестник НГТУ. – 2016. – Т. 62, № 1. – С. 17-34.
18. *Бова В.В., Щеглов С.Н., Лещанов Д.В.* Модифицированный алгоритм EM-кластеризации для задач интегрированной обработки больших данных // Известия ЮФУ. Технические науки. – 2018. – № 4 (198). – С. 154-166.
19. *Харченко А.М.* Адаптивный расчет функции для динамического EM-алгоритма // Математика. – 2015. – С. 134.
20. *Курейчик В.М., Каланчук С.А*. Обзор и состояние проблемы роевых методов оптимизации // Информатика, вычислительная техника и инженерное образование. – 2016. – № 1 (25). – С. 1-13.
21. *Карпенко А.П.* Современные алгоритмы поисковой оптимизации. Алгоритмы, вдохновленные природой: учеб. пособие. – М.: МГТУ им. Н.Э. Баумана, 2014. – 446 с.
22. *Yang X.S., Deb S.* Multiobjective cuckoo search for design optimization // Comput. Oper. Res. – 2013. – No. 40 (6). – P. 1616-1624.
23. *Chifu V.R., Pop C.B., Salomie I., Niculici A.N.* Optimizing the semantic web service composition process using cuckoo search // Intelligent Distributed Computing. – 2012. – No. 5. – P. 93-102.
24. *Coelho L.S., Guerra F.A., Batistela N.J., Leite J.V.* Multiobjective cuckoo search algorithm based on duffings oscillator applied to jiles-atherton vector hysteresis parameters estimation // IEEE Trans. Magn. – 2013. – No. 49 (5). – P. 1745.
25. *Репозиторий* машинного обучения. – URL: http://www.ics.uci.edu/~mlearn/MLRepository.html (дата обращения: 24.06.2019).

REFERENCES

1. *Tingting Z., Chen L.Y., Liang-Hsien T.* Understanding user motivation for evaluating online content: a self-determination theory perspective, *Behaviour and Information Technology*, 2015, No. 34, pp. 479-491.
2. *Deliang W., Lingling X., Chuan C.H.* Understanding the continuance use of social network sites: a computer self-efficacy perspective, *Behaviour and Information Technology*, 2015, No. 34, pp. 204-216.
3. *Alfimtsev A.N., Devyatkov V.V., Sakulin S.A.* Personalizatsiya v gipertekstovykh setyakh na osnove raspoznavaniya deystviy pol'zovateley i nechetkogo agregirovaniya [Personalization in hypertext networks based on recognition of user actions and fuzzy aggregation], *Vestnik MGTU im. Baumana. Ser. «Priborostroenie»* [Herald of the Bauman Moscow State Technical University. Series Instrument Engineering], 2012, No. 3.
4. *Pisarenko V.I.* Vozmozhnosti ispol'zovaniya pedagogicheskikh znaniy v mezhdistsiplinarnykh issledovaniyakh [Possibilities of using pedagogical knowledge in interdisciplinary research], *Sovremennaya nauka: aktual'nye problemy teorii i praktiki. Seriya "Gumanitarnye nauki"* [Modern Science: actual problems of theory and practice", a series of "Humanities"], 2018, No. 12-3, pp. 48-52.

5. *Voyskunskiy A., Evdokimenko A., Fedunina N.* Setevaya i real'naya identichnost': sravnitel'noe issledovanie [Network and real identity: a comparative study], *Psikhologiya. Zhurnal Vysshey shkoly ekonomiki* [Psychology. Journal of the Higher School of Economics**],** 2013, Vol. 10, No. 2, pp. 98-121.
6. *Bova V.V., Kureichik V.V., Leshchanov D.V.* The model of semantic similarity estimation for the problems of big data search and structuring, *Application of Information and Communication Technologies - AICT 2017*, pp. 27-32.
7. *Gaziev G.Z., Kurdyukova G.N., Kurdyukov V.V.* Klasterizatsiya Big Data dlya ikh analiza i obrabotki [Clustering of Big Data for their analysis and processing], *Sb. nauchnykh statey konferentsii «Napravleniya i mekhanizmy razvitiya nauki novogo vremeni: ot teorii do vnedreniya rezul'tatov»* [Collection of scientific articles of the conference "Directions and mechanisms of development of modern science: from theory to implementation of results"], 2017, pp. 150-162.
8. *Pautov K.G., Popov F.A.* Metod klasterizatsii tematicheskikh profiley pol'zovateley i ego primenenie dlya analiza internet-trafika [Clustering method of thematic user profiles and its application for Internet traffic analysis], *Fundamental'nye issledovaniya* [Fundamental study], 2015, No. 74, pp. 765-769.
9. *Kravchenko Yu.A.* Otsenka kognitivnoy aktivnosti pol'zovatelya v sistemakh podderzhki prinyatiya resheniy [Assessment of cognitive activity of the user in decision support systems], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2009, No. 4 (93), pp. 113-117.
10. *Kureychik V.M., Kartiev S.B.* Algoritm klassifikatsii, osnovannyy na printsipakh sluchaynogo lesa, dlya resheniya zadachi prognozirovaniya [Classification algorithm based on the principles of random forest to solve the problem of forecasting], *Programmnye produkty i sistemy* [Software products and systems], 2016, No. 2, pp. 11-15.
11. *Sarin K.S., Vorozhtsov S.A., Arimpilov S.N.* Postroenie ansambley nechetkikh klassifikatorov na osnove metaevristiki "kukushkin poisk" i gornoy klasterizatsii [Building ensembles of fuzzy classifiers on the basis of metaheuristic "Kukushkin search" and mountain clustering], *Elektronnye sredstva i sistemy upravleniya* [Electronic means and control systems], 2017, No. 1-2, pp. 26-29.
12. *Kartiev S.B., Kureychik V.M.* Razrabotka i issledovanie algoritma resheniya zadachi klasterizatsii dlya osushchestvleniya voprosno-otvetnogo poiska v informatsionno-analiticheskoy sisteme prognozirovaniya [Development and research of an algorithm for solving the problem of clustering for the implementation of question and answer search in the information and analytical forecasting system], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2016, No. 7 (180), pp. 18-28.
13. *Markov V.V., Kravchenko Yu.A., Kuz'mina M.A.* Razvitie metodov semanticheskoy fil'tratsii na osnove resheniya zadachi klasterizatsii bioinspirirovannymi algoritmami [Development of semantic filtering methods based on the solution of the clustering problem by bio-inspired algorithms], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 4 (198), pp. 175-185.
14. *Khodashinskiy I.A., Anfilof'ev A.E., Bardamova M.B., Kovalev V.S., Mekh M.A., Sonich O.K.* Metaevristicheskie metody optimizatsii parametrov nechetkikh klassifikatorov [Metaheuristic methods for optimization of fuzzy classifier parameters], *Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii* [Information and mathematical technologies in science and management], 2016, No. 1, pp. 73-81.
15. *alalirad A., Tjalkens T.* Using feature-based models with complexity penalization for selecting features, *Journal of Signal Processing Systems*, 2018, Vol. 90, Issue 2, pp. 201-210.
16. *Guo G., Zhang J., Thalmann D.* Merging trust in collaborative filtering to alleviate data sparsity and cold start, *Knowledge-Based Systems*, 2014, No. 57, pp. 57-68.
17. *Avdeenko T.V., Makarova E.S.* Metod opredeleniya relevantnosti pretsedentov na osnove nechetkikh lingvisticheskikh pravil [Method of determining the relevance of precedents based on fuzzy linguistic rules], *Nauchnyy vestnik NGTU* [Scientific Bulletin of NSTU], 2016, Vol. 62, No. 1, pp. 17-34.
18. *Bova V.V., Shcheglov S.N., Leshchanov D.V.* Modifitsirovannyy algoritm EM-klasterizatsii dlya zadach integrirovannoy obrabotki bol'shikh dannykh [The modified algorithm, EM-clustering for task integrated big data processing], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 4 (198), pp. 154-166.

19. *Kharchenko A.M.* Adaptivnyy raschet funktsii dlya dinamicheskogo EM-algoritma [The adaptive function is calculated for the dynamic em algorithm], *Matematika* [Mathematics], 2015, pp. 134.
20. *Kureychik V.M., Kalanchuk S.A.* Obzor i sostoyanie problemy roevykh metodov optimizatsii [Review and state of the problem of swarm optimization methods], *Informatika, vychislitel'naya tekhnika i inzhenernoe obrazovanie* [Informatics, computer science and engineering education], 2016, No. 1 (25), pp. 1-13.
21. *Karpenko A.P.* Sovremennye algoritmy poiskovoy optimizatsii. Algoritmy, vdokhnovlennye prirodoy: ucheb. posobie [Modern search engine optimization algorithms. Nature-inspired algorithms: textbook]. Moscow: MGTU im. N.E. Baumana, 2014, 446 p.
22. *Yang X.S., Deb S.* Multiobjective cuckoo search for design optimization, *Comput. Oper. Res.*, 2013, No. 40 (6), pp. 1616-1624.
23. *Chifu V.R., Pop C.B., Salomie I., Niculici A.N.* Optimizing the semantic web service composition process using cuckoo search, *Intelligent Distributed Computing*, 2012, No. 5, pp. 93-102.
24. *Coelho L.S., Guerra F.A., Batistela N.J., Leite J.V.* Multiobjective cuckoo search algorithm based on duffings oscillator applied to jiles-atherton vector hysteresis parameters estimation, *IEEE Trans. Magn*., 2013, No. 49 (5), pp. 1745.
25. Repozitoriy mashinnogo obucheniya. Available at: http://www.ics.uci.edu/~mlearn/MLRepository.html (accessed 24 June 2019).

Статью рекомендовал к опубликованию д.т.н., профессор Ю.А. Гатчин.

Бова Виктория Викторовна – Южный федеральный университет; e-mail: vvbova@yandex.ru; 347928, г. Таганрог, Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; доцент.

Кравченко Юрий Алексеевич – e-mail: yakravchenko@sfedu.ru; кафедра систем автоматизированного проектирования; доцент.

Bova Victoria Victorovna – Southern Federal University; e-mail: vvbova@yandex.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; associate professor.

Kravchenko Yury Alekseevich – e-mail: yakravchenko@sfedu.ru; the department of computer aided design; associate professor.

УДК 004.6:004.8 DOI 10.23683/2311-3103-2019-4-102-114

Д.В. Балабанов, А.В. Ковтун, Ю.А. Кравченко

БУСТИНГ БИОИНСПИРИРОВАННЫХ АЛГОРИТМОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ИНТЕГРАЦИИ ДАННЫХ*

В настоящее время интеграция данных является актуальной проблемой. Интеграция данных может быть представлена на различных уровнях. Современные методы решения задач интеграции не могут решать семантическую проблему, к тому же они слишком сложны. На основе исследований методов, используемых на данный момент, можно сделать вывод, что самыми часто используемыми методами, для подходов к решению задачи неоднородности на семантическом уровне, используются такие эвристики, которые применяют результирующую онтологию. В большинстве случаев, данные информационных систем представлены как объекты информации, которые в свою очередь формируют некую предметную область или ее часть, в тоже время к каждой части (области) относится ее собственная онтология. Исходя из этого, при решении задачи семантической неоднородности данных нужно привести определения предметных областей и взаимодействия их объектов. Таким образом можно построить взаимодействие информационных систем,

* Работа выполнена при поддержке РФФИ (проекты: № 19–07–00099, № 18–07–00055).