УДК 519.688

ИММУННЫЙ АЛГОРИТМ МУЛЬТИКЛОНАЛЬНОЙ СЕЛЕКЦИИ В РЕШЕНИИ ЗАДАЧИ ИДЕНТИФИКАЦИИ ПОЧЕРКА

Ю.И. ЕРЕМЕНКО
А.А. ШАТАЛОВ

Старооскольский технологический институт (филиал) ФГОУ ВПО Национальный исследовательский технологический университет «Московский институт стали и сплавов»

e-mail:
erem49@mail.ru
asskunst@rambler.ru

В статье обосновано применение методов искусственного интеллекта и представлен иммунный алгоритм для решения задачи идентификации автора рукописного русскоязычного текста по образцу его почерка. Алгоритм разработан на основании результатов анализа методов теории опасности, негативной селекции и идиотипических сетей Ерне.

Ключевые слова: распознавание рукописного текста, экспертиза, алгоритм клонального отбора, идентификация личности по почерку.

По оценкам консалтинговой компании Acuity Market Intelligence (*фирма-аналитик рынка биометрических технологий*), общий объем мирового биометрического рынка возрос с $4 млрд. в 2012 году до $10 млрд. и составит, по прогнозам, к 2017 году $11 млрд. [15]. Среди современных средств биометрической идентификации выделяют системы для идентификации личности по почерку, нашедшие применение в криминалистике.

Экспертиза почерка является одним из самых распространенных и в то же время наиболее сложных и спорных в плане объективности результатов. От эксперта требуется разносторонний подход, умение логически осмысливать и сопоставлять большое количество фактов, правильно применять и выбирать необходимые средства и методы исследования, оценивать их результаты, как в отдельности, так и в совокупности. При этом эксперты часто совершают ошибки, а наибольший процент выводов о невозможности решения поставленных вопросов приходится на этот вид экспертиз. Возможность ошибки в выводах в большей степени зависит от сложности решаемых задач при проведении экспертизы [2].

Текущая ситуация объясняются сложностью объекта исследования – почерка, который представляет собой высоковариативную мультикомпонентную систему и зависит от многочисленных факторов: психофизиологических свойств пишущего, особенностей обучения письму и различных эпизодических условий внутреннего и внешнего характера. Невозможно полностью исключить ошибки экспертизы, потому что их наличие определяется опытом и физическим состоянием эксперта. Другими словами, результаты почерковедческой экспертизы имеют высокий субъективный фактор.

Основной задачей исследования почерка в практической деятельности является идентификация конкретного исполнителя рукописи (подписи). Классическая методика решения данной задачи сводится к поиску индивидуальных особенностей и сравнению их между собой. Образцы почерка представляют собой ряд слов или фраз, написанных лицом, личность которого нужно установить. В компьютерной среде подобные образцы (шаблоны) сканируются и сохраняются в любом из известных графических форматов.

Применяемые в настоящее время методы для экспертизы почерка опираются на исследования, произведенные еще в прошлом веке. Исторически, для решения идентификационных задач используются методики, опирающиеся на обширный почерковой материал, который предварительно собирался на основании специальных исследований. В ходе последующей статистической обработки этих данных и вычисления идентификационных зависимостей выявлялись признаки, при помощи которых специалист-почерковед проводил экспертизу. Эти методы являются трудноформализуемыми и не предполагают использование современных вычислительных средств. Они опираются на прописи старого образца, поэтому некоторыми

исследователями отмечается снижение адекватности почерковедческих экспертиз. Для создания новых методов необходимо проводить сбор и обработку большого количества статистических данных, что представляется дорогой и трудозатратой работой и является недостатком классического подхода. Рационально при помощи методов интеллектуальной обработки данных и современных вычислительных средств обобщить ранее накопленный опыт, исключить человеческий фактор из процедуры экспертизы почерка. Использование средств автоматизации работы экспертов, не только может повысить производительность труда эксперта, но и в целом повысит объективность экспертизы почерка [4].

Несмотря на многочисленные исследования теории и практики, а так же попытки использования математических методов и компьютерной техники в криминалистике, проблема идентификации почерка по-прежнему не нашла однозначного решения [3,16]. В связи с этим в настоящее время для решения вышеупомянутой проблемы ведутся исследования с применением алгоритмов, основанных на принципах работы биологических систем [3,4,8,10,12,16]. Наилучших результатов на сегодняшний день удалось добиться при разработке методов, основанных на нейронных сетях [1, 7, 17], однако ряд исследований [8,18] указывает на превосходство алгоритмов искусственных иммунных систем (ИИС) над нейросетевыми при решении целого ряда задач идентификации. Аппарат ИИС сочетает в себе достоинства и гибкость нейронных сетей и мультиагентных систем, а в ряде работ с успехом применен с нейронными сетями. Топология иммунных сетей проста и прозрачна, что позволяет следить за процессами в сети и проводить их оптимизацию[14].

Ранее A. K. Muda и S. M. Shamsuddin успешно применили один из первых иммунносетевых алгоритмов, основанный на теории негативной селекции для идентификации исполнителя текста на английском языке. Авторство в 4-х из 5-ти опытов было установлено корректно, но для эксперимента была использована нерепрезентативная выборка, которая составила 10 человек.

Таким образом проведенный анализ показывает, что наиболее перспективным для идентификации автора рукописного текста на русском языке является применение аппарата ИИС, который может решать классификационные задачи и проводить анализ данных на основе принципов молекулярного узнавания [13].

Пусть имеется множество V, представляющее список лиц с известным почерком. Множество V разбито на N подмножеств $K_1, K_2, K_3, ..., K_n$, которые являются классами и содержат набор признаков, характерных для почерка одного лица. Набор признаков представляет множество $P = \{p_1, p_2, p_3, ... p_g\}$, где p_m – уникальный признак почерка. $K_i \subset P$ $(i=1,...,N)$.

Есть подмножество $E = \{p_g | g \in [1;G]\}$, представляющее собой образец почерка неизвестного лица. При этом выполняется условие 1:

$$m(E \cap K_i) \le N, (i = 1,...,N), K_i \cap Kj \ne \varnothing, (i \ne j); \quad (1)$$

Где, m – мощность множества.

Принимая во внимание тот факт, что при идентификации почерка для оценки результатов используются методики с нечеткой шкалой, представим в общем виде систему экспертной оценки в виде функции 2:

$$Rt(m) \in \{"не\_соответствие",...,"полное\_соответствие"\}; \quad (2)$$

Таким образом, задача идентификации почерка формулируется следующим образом:

$$F(K_i) = Rt(m(E \cap K_i)) \to "полное\_соответствие", (i = 1,...,N); \quad (3)$$

То есть, необходимо найти такую функцию (алгоритм), которая для любой E позволила бы подобрать класс или классы K, стремящиеся к максимальной степени соответствия по оценке эксперта-криминалиста. Однако такие классы могут быть не найдены, при условии, что $m(E \cap K_i) \to 0, (i = 1,...,N)$.

В настоящее время в аппарате ИИС существуют три ведущие теории: теория клонального отбора [16], теория идиотипических сетей Эрне [6] и теория опасности [11].

Результатом исследований в теории клонального отборка является алгоритм CLONALG, который можно кратко описать следующим образом:

1. Открыть базу данных образцов почерков (БД) и сформировать начальную популяцию (MP, от анг. Main Population).

2. Получить образец неизвестного почерка (антигена) и рассчитать энергию связи (аффинность) с элементами из MP (антителами).

3. Антитела с лучше аффинностью клонировать, клоны подвергнуть мутации.

4. Сформировать промежуточную популяцию из клонов (TP, от анг. Temporary Population).

5. Выбрать из TP лучшие клоны и записать их в БД. Удалить из БД худшие антитела и добавить случайно сгенерированные новые.

6. Выполнять пункты 2-5 пока не будет выполнен критерий остановки алгоритма.

На теории идиотипических сетей Ерне основан алгоритм кластеризации AINet. Особенность сети Ерне заключается в идее о том, что элементы такой сети (лимфоциты) способны распознавать друг друга, что в свою очередь наделяет сеть неким аналитическим аппаратом. Алгоритм AINet в целом похож на CLONALG, но в отличие от него обладает механизмом клонального сжатия, за счет которого реализуется идея идиотипической сети. При клональной сжатии из популяции удаляются взаимно похожие элементы.

На теории опасности (danger theory) основан алгоритм DCA, который является бинарным классификатором. Принцип действия алгоритма DCA основан на механизме работы биологической дендритной клетки. Согласно теории опасности такие клетки способны запускать иммунный ответ, собирая и анализируя сигналы, указывающие на аномальную смерть клеток в организме. В DCA модель дендритной клетки производит анализ образцов неизвестного почерка, как сигналов, с последующим их отнесением к одному из идентифицируемых классов.

На основании результатов анализа и опытного моделирования рассмотренных алгоритмов был получен алгоритм мультиклональной селекции, позволяющий проводить идентификацию почерка.

Для задач идентификации введем следующие понятия. Математическое определение антитела: *At=<Mas, inf >*, где *Mas{0-255}* – массив признаков (генов). Каждый признак представляет собой пиксель изображения, где 0 соответствует черному пикселю, 255 – белому пикселю. *inf* – значение антитела. *Mp={At}* – БД или основная популяция антител. *Ag=<Mas>* – антиген. Степень схожести или аффинность *Ag-At* может быть вычислена при использовании манхэттенской метрики:

$$D = \sum_{i=1}^{l} |at_i - ag_i|; \qquad (4)$$

где l – количество элементов массива генов, at_i – i-й ген антитела At, ag_i – i-й ген антигена Ag. Для изменения или мутации генов использована формула 5, полученная экспериментальным путем.

$$Pm(at_i) = \text{var}^* D_{ati} * Km / D; \qquad (5)$$

где, var – число, случайно принимающее значения 1 и -1 и определяющее направление мутации, Km – эмпирически установленный коэффициент (принят за 17000), D_{ati} – аффинность между at_i и i-ым геном ag_i антигена Ag.

Для расчета размеров промежуточной популяции *F(D)* предложено использовать формулу 5, также полученную эмпирически. Она позволяет получать клоны в интервале от 0 до n+1, пропорционально их аффинности. Антитела в промежуточной популяции с меньшей аффинностью производят наименьшее количество потомков, а антитела с большей аффинностью – наибольшее, даже если минимальная и максимальная аффинность в промежуточной популяции принимает значения в пределах 93 – 95% схожести.

$$F(D) = n * \frac{D*(D - D_s)}{100*(D_{\max} - D_s)} + 1; \qquad (5)$$

Где, D – аффинность между антителом At и антигеном Ag. n – целое число (принято за 6), определяющее максимальное значение клонов. Ds – средняя аффинность между всеми антителами и антигеном, D_{max} – максимальная аффинность между антителом и антигеном, полученная на текущей итерации иммунного алгоритма.

Предлагается выбирать за одну итерацию не один образец почерка, а несколько. Для этого изначальный иммунный алгоритм был дополнен внутренним циклом. На первой итерации алгоритма происходит анализ всех имеющихся образцов и из них выбираются в ТР наиболее подходящие. Все ТР для текущей выборки антител компонуются в еще одну временную популяцию ТМР, благодаря которой появилась возможность проводить анализ в совокупности.

Предложено заменять БД на ТМР, так как все интересующие антитела уже выбраны в последней. Таким образом система быстрее приходит к равновесию, и на конечных итерациях МР содержит множество антител, наиболее схожих с исследуемыми антигенами.

На рис. 1 изображена блок-схема разработанного алгоритма.

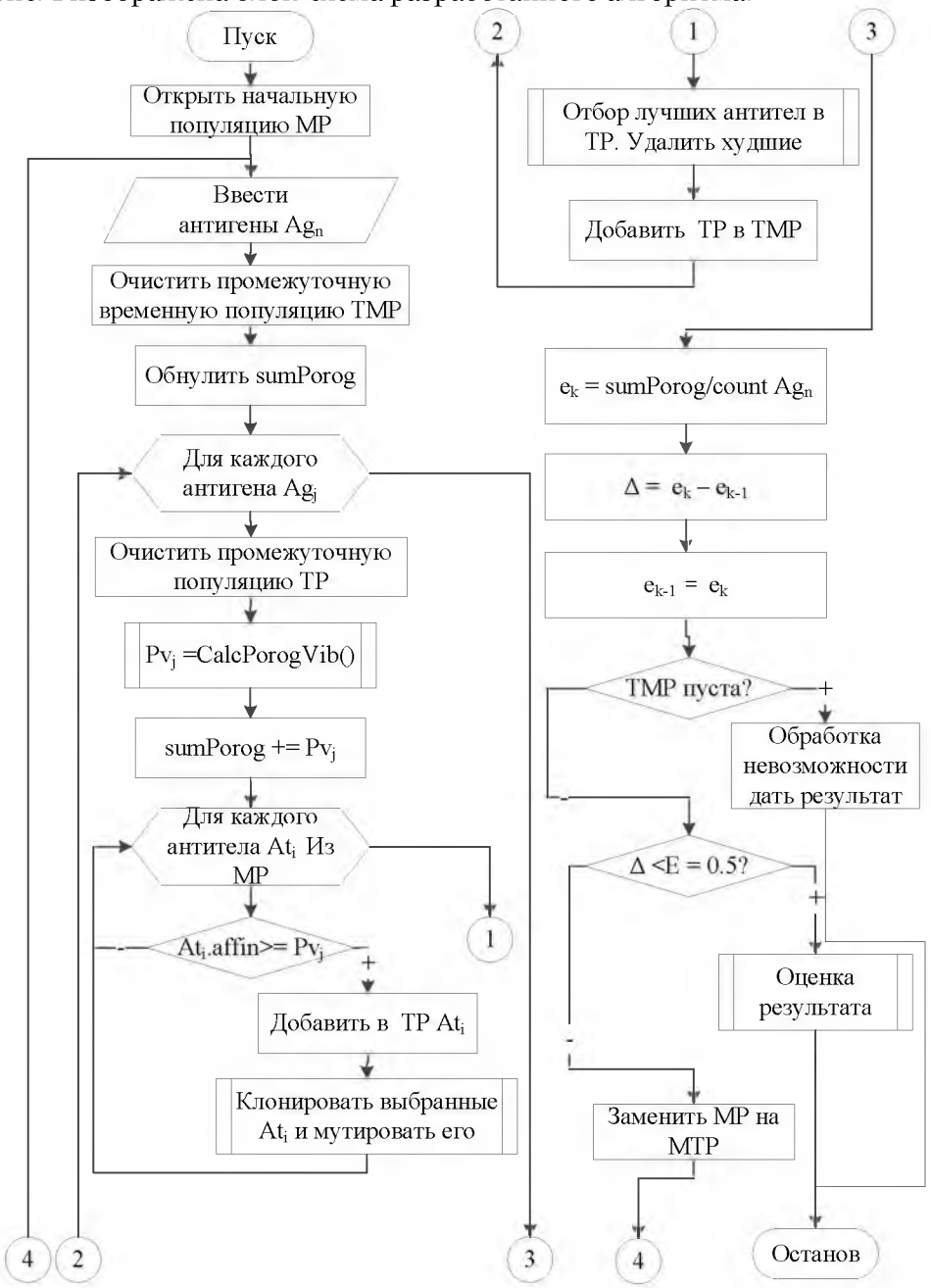

Рис. 1. Блок-схема иммунного алгоритма для решения задачи идентификации почерка

Результат идентификации предлагается оценивать следующим образом. Вес W это величина, которая обозначает суммарную аффинность образцов почерка одного лица, находящихся в финальной выборке антител. Для каждого вероятного кандидата находится

процентная доля веса *Wp* от суммарного веса всех кандидатов, фильтруя лица с *Wp<1%*. Для финальной выборки находится среднеквадратическое отклонение δ при условии, что распределение результатов подчиняется нормальному закону распределения. Искомый кандидат считается верно определенным среди выбранных, если $Wp-\delta>0$. Такое решение позволило минимизировать реакцию алгоритма идентификации на похожие варианты образцов почерка принадлежащих лицам, не являющимися авторами исследуемого текста. Рисунок 2 демонстрирует удачный результат опыта по идентификации человека по имени «Person_1_16», который и являлся неизвестным лицом.

Рис. 2. Пример правильно идентифицированной личности

Опыт проводился с использованием БД в размере 50 человек. На анализ подавалось 62 образца неизвестного почерка. На оси абсцисс отмечены вероятные кандидаты на авторство неизвестного почерка, попавшие в финальную выборку. Гистограмма показывает долю веса для этих кандидатов. График показывает разность между долей веса и среднеквадратическим отклонением.

На основе предложенного метода разработана информационная система, позволяющая пользователю решать задачу идентификации пользователя по образцу его почерка. Интерфейс программы представлен на рис. 3.

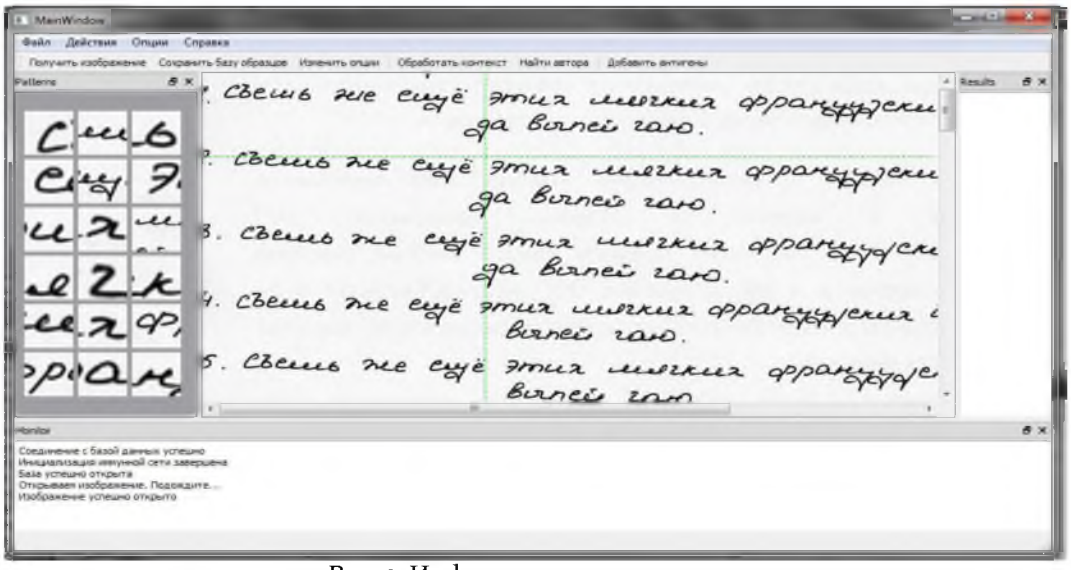

Рис. 3. Информационная система,
предназначенная для решения идентификационной задачи

Программа позволяет загружать и просматривать графические изображения. С изображений выделяется один или несколько образцов почерка, которые можно как

идентифицировать, так и сохранить как новую базу образцов. Также можно открывать и сохранять как целые базы образцов почерка, так и отдельные образцы.

Программа написала на языке Qt с использованием парадигмы объектно-ориентированного программирования (ООП). Программа использует модульную структуру и выстроена при помощи шаблона SDI(Single Document Interface). Функционал и архитектура программы позволяет работать с базами данных почерков, как расположенных локально, так и на удаленных серверах.

Полный цикл работы с программой описывается следующий образом: пользователь открывает БД, после чего загружает в центральное окно образец почерка неизвестного лица. Используя рамку выделения в центрально области, пользователь вносит в окно Patterns интересующие его образцы почерка, после чего нажмёт копку «Найти автора». Когда процедура идентификации заканчивается, программ выдает пользователю список наиболее вероятных авторов искомого почерка.

Выполнено при поддержке грантов РФФИ №12-07-00252 и № 13-08-00532 А.

Список литературы

1. Жиляков Е.Г., Лихошерстный А.Ю., Красильников В.В. Структура нейросте и для распознавания объектов аэрокосмических изображений на основе анализа распределения их энергии по частотным интервалам // Научные ведомости БелГУ. – 2012. – №7(126). С. 117-124.
2. Зеленецкий, В.С. Предупреждение экспертных ошибок. Метод. пособие. М.: ВНИИСЭ, 1990.
3. Комаров, А.С. Логические и программные средства интеллектуального анализа криминалистических данных. . дис. ... канд. тех. наук. – М.: ВИНИТИ РАН, 2010.
4. Кулик С.Д., Никонец Д.А. Автоматизация почерковедческих исследований// XIX Международная научная конференция «Информатизация и информационная безопасность правоохранительных органов» (25–26 мая 2010 г., Москва): Сборник трудов. – М.: Академия управления МВД России, 2010. С. 314–317.
5. Кулик С.Д., Никонец Д.А. Примеры использования нейросетевого алгоритма в методиках для эксперта-почерковеда // Нейрокомпьютеры: разработка и применение. – 2009. – №9. С. 61-65.
6. Литвиненко В.И., Дидык А.А., Захарченко Ю.А. Компьютерная система для решения задач классификации на основе модифицированных иммунных алгоритмов // Информационно-измерительные системы. – ААЭКС. – 2008. – Т.22. – №2.
7. Юдин Д.А., Магергут В.З. Применение метода экстремального обучения нейронной сети для классификации областей изображения // Научные ведомости БелГУ. – 2013. – №8(151). С. 95 – 103.
8. Azah Kamilah bt. draman @ muda «Authorship invarianceness for writer identincation using invariant discretization and modified immune classifier».A thesis submitted in fultilment of the requirements for the award of the degree of Doctor of Philosophy (Computer Science) .Faculty of Computer Science and Information System Universili Teknologi Malaysia, august 2009.
9. Djeddi, C. "Artificial Immune Recognition System for Arabic writer identification», Innovation in Information & Communication Technology (ISIICT), 2011 Fourth International Symposium on, 29 2011-Dec. 1 2011, Amman, Conference Publications. Page(s): 159 – 165.
10. Julie Greensmith, Amanda Whitbrook, Uwe Aickelin «Artificial Immune Systems», Handbook of Metaheuristics, 2nd edition, Springer, 2010, 27p.
11. Julie Greensmith, Uwe Aickelin, Gianni Tedesco. Information Fusion for Anomaly Detection with the Dendritic Cell Algorithm. Information Fusion 11 (1). 2010. – 21-34pp.
12. Khaled Mohammed Bin Abdl and 2Siti Zaiton Mohd Hashim, «Swarm-Based Feature Selection for Handwriting Identification», Journal of Computer Science 6 (1): 80-86, 2010.
13. L. N. De Castro, F.J. Von Zuben, 2000a.» Artificial Immune Systems: Part I I – A Survey of Application. Technical Report» –RT DCA 02/00.
14. Muda, Azah Kamilah and Shamsuddin, Siti Mariyam (2005) «A framework of artificial immune system in writer identification.» In: BIC`05, Puteri Pan Pacific, Conference or Workshop ,24 May 2007.
15. The Future of Biometrics Market Research Report [Электронный ресурс]: Market Research Report. / Millburn, USA. — URL: http://www.acuity-mi.com/FOB_Report.php (дата обращения: 17.04.2013).

16. Utpal Garain, Mangal P. Chakraborty, Dipankar Dasgupta.» Recognotion of handwritten indic script using Clonal Selection Algorithm». H. Bersini and J.Carneiro(Eds.): ICARIS 2006, LNCS 4163, pp.256-266, 2006.

17. Yu Yang «Application of Artificial Immune System in Handwritten Russian Uppercase Character Recognition», Computer Science and Service System (CSSS), 2011 International Conference on, Conference Publications, Publication Date: 27-29 June 2011 Volume-OnPage 238-24.

18. Yu Yang « Handwritten Icelandic character recognition based on artificial immune system»,Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International 20-22 Aug. 2011, Conference Publications .Volume: 2.

IMMUNE MULTICLONAL SELECTION ALGORITHM FOR HANDWRITING IDENTIFICATION PROBLEM

Y.I. EREMENKO
A.A. SHATALOV

Stary Oskol's Technological Institute (department) FGOU VPO National Technological Research University «Moscow Institute of Steel and Alloys»

e-mail:
erem49@mail.ru
asskunst@rambler.ru

The article justifies the application of artificial intelligence methods and shows the immune algorithm for solution of an author of handwritten Russian-language text identification problem. The algorithm has been developed based on analysis results of danger, negative selection and Erne idiotype network theory methods.

Key words: handwriting identification, individual's identification based on handwriting, clonal selection algorithm, handwriting analysis.