КЛАСТЕРИЗАЦИЯ ПЛАЗМИД ПАЛОЧКОВИДНЫХ ФОРМ БАКТЕРИЙ И ИХ ВИДОВ С ИСПОЛЬЗОВАНИЕМ СПЕКТРОСКОПИИ

И.В. Кухарчук
Ассистент кафедры электронных вычислительных машин БГУИР

Д.И. Самаль
Заведующий кафедрой электронных вычислительных машин БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: i.kukharchuk@bsuir.by, samal@bsuir.by

Abstract. . This paper describes the solution of the problem of classification for bacterial species and their nutrient media. The described algorithm allows classifying three species of bacteria using self-organized map, as well as the clustering of three nutrient media for E. coli bacteria using k-means. The paper presents the rationale for choosing a solution model, calculating its accuracy, and ways to increase the accuracy. The experiments show the proposed solution is capable of clustering spectrograms with an accuracy of 96%.

Возможность быстрой автоматической идентификации и/или классификации бактерий (их видов) в образце до сих пор является серьёзной проблемой в области микробиологии.

В рамках настоящей работы рассмотрены существующие подходы к проведению кластеризации содержимого штаммов бактерий на основе их спектрограммы, полученной при помощи настольного микроскопа комбинационного рассеяния.

Целью представляемого исследования являлась реализация алгоритма, самообучающегося в условиях ограниченного количества имеющихся спектрограмм; после обучения задачей алгоритма является последующая кластеризация входящих штаммов с необходимым уровнем точности, превышающим 90%. В рамках работы были решены следующие задачи:

- загрузки и подготовки данных (спектрограмм) к обработке;
- создания архитектуры совокупного алгоритма обучения и кластеризации видов;
- выбора алгоритма обучения для первичного разделения бактерий по видам;
- выбора алгоритма кластеризации для разделения бактерий по питательной среде, в которой они выращены.

Первым этапом проведённого исследования являлось формирование базы спектрограмм известных бактерий. Определение вида бактерий происходит на основе их содержимого, в частности – плазмидов, и их влияния на изменения в спектрограммах [1]. На данном этапе все спектрограммы сохраняются в виде сырых необработанных данных.

Классифицируемыми бактериями являются E. coli, shiwinella, lactobacilus. Спектрограмма с максимально различающимися интенстивностями названных бактерий представлены на рисунке 1. Однако, если рассматривать весь набор возможных спектрограмм, то различия между данными бактериями в большинстве случаев практически нивелируются. Тестовый набор каждого вида бактерий составляет 100 штаммов. Таким образом, исходная база данных составляла три набора по 100 штаммов для каждого из видов исследуемых бактерий, соответственно.

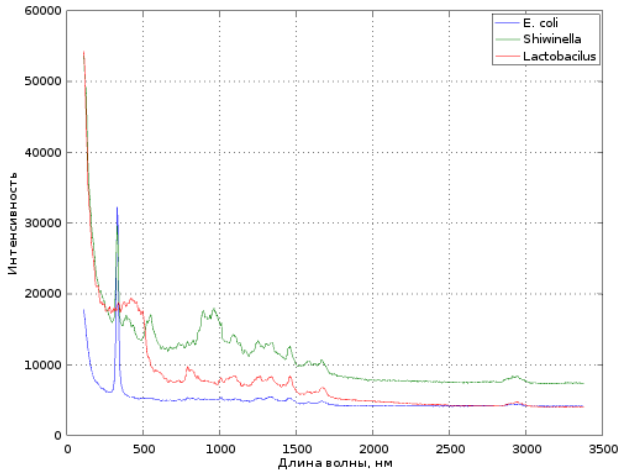

Рис. 1. Спектрограммы классифицируемых бактерий

Согласно постановке задачи, для первого классифицируемого вида бактерий – E. coli – необходимо различать бактерии по питательной среде, в которой они выращены (MMGls, LB и т.д.). Пример того, как среда влияет на изменение спектрограммы бактерии E. coli, отражён на рисунке 2.

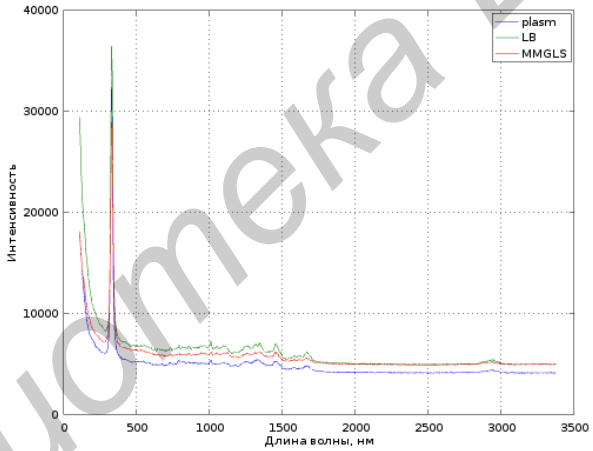

Рис. 2. Спектрограммы влияния различных питательных сред на примере бактерии E. coli

Так как имелась возможность использования тестового набора для классификации бактерий, то применялись алгоритмы с обучением, общий вид, которых, отображён на рисунке 3.

Для решения задачи классификации была выбрана нейронная сеть встречного распространения без обратных связей на базе самоорганизующейся карты Кохонена и звезды Гроссберга. Данная нейронная сеть хорошо подходит для указанной задачи ввиду следующих факторов:

– наличие свойства обобщения ввиду наличия слоя Кохонена;

– простота обучения слоя Гроссберга, накопления статистических данных;

– высокая скорость обучения [2-3].

Используемая в работе классическая схема нейронной сети отображена на рисунке 4. На данном рисунке входные данные представлены вектором X, слой Кохонена вектором K, слой Гроссберга вектором G, выходной вектор – Y. Веса для слоёв соответственно w и v.

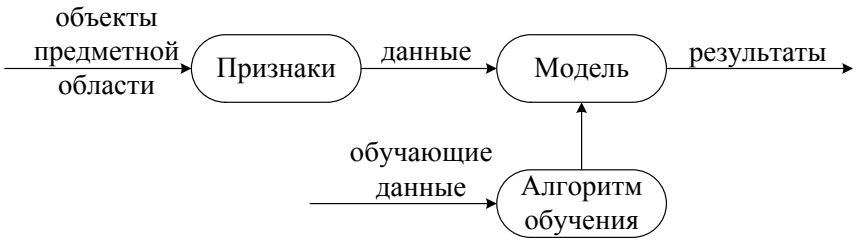

Рис. 3. Общий вид решения задачи машинного обучения

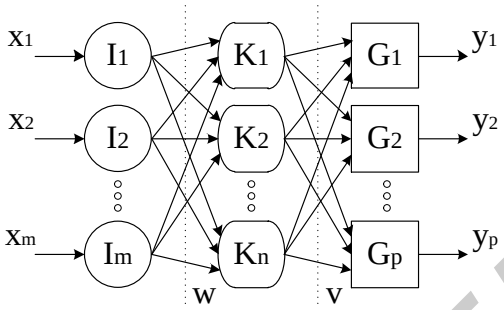

Рис. 4. Нейронная сеть с встречным распознаванием без обратных связей

После анализа сформированной базы спектрограмм в области влияния питательных сред на конечный вид графиков, была выявлена классическая закономерность: спектрограммы могут быть собраны в кластере, как это отражено на рисунке 2. Количество кластеров по постановке задачи заранее известно. Таким образом, достаточным решением для задачи кластеризации питательных сред бактерии E. coli является метод k-средних, имеющий вид [4]:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \qquad (1)$$

где k – число кластеров; S_i – полученные кластеры; $i = 1, 2, \ldots, k$ и μ_i – центры масс векторов $x_j \in S_i$.

Архитектура решения, позволяющего классифицировать бактерии, и кластеризировать питательные среды для одной из них представлена на рисунке 5.

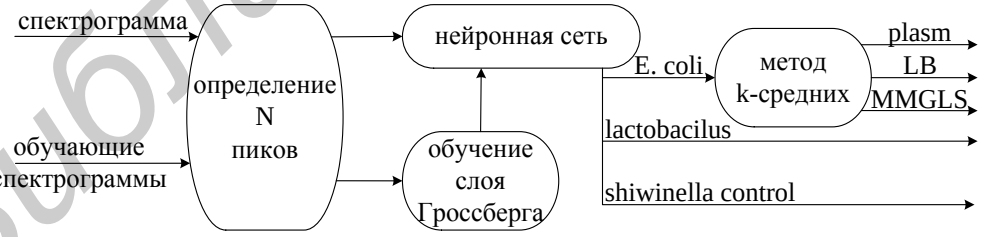

Рис. 5. Архитектура предложенного решения классификации бактерий и кластеризации их питательных сред

Входные данные построенной по образцу сети *могут* быть непрерывными, однако классификатор, работающий с таким объёмом непрерывных данных (1024 точки), продемонстрировал недостаточную точность определения бактерий (порядка 50%) на всех трёх типах. Для повышения точности было проведено редуцирование исходных данных и изменение исходных признаков, по которым проводилась классификация. В качестве признаков отбирались

пики спектрограмм и их общее количество для каждого объекта было сокращено до 100. Таким образом, операцией обработки входных данных нейронной сети стал отбор пиков. Для отбора определённого количества пиков изменялось разрешённое расстояние между ними. Данное действие позволило поднять общую точность определения бактерий до уровня 84%.

Следующими этапами модификаций нейронной сети стали:

– расширение обучающей выборки путём добавления шума к входным векторам;

– увеличение порога нейрона, который чаще остальных становится победителем;

– использование интерполяции вместо аккредитации– эмпирическая коррекция параметров α и β на этапе обучения нейронной сети.

В результате для каждого типа бактерии в классификаторе были получены изменения начального значения коэффициентов обучения следующим образом: . было создано поле в рамках $\alpha \in [0{,}41; 0{,}90]$ и $\beta \in [0{,}01; 0{,}50]$, и в результате циклов обучения получены статистические данные. Проведённые эксперименты отображены на рисунке 6. Алгоритм уменьшения коэффициентов обучения подчиняется показательному закону распределения.

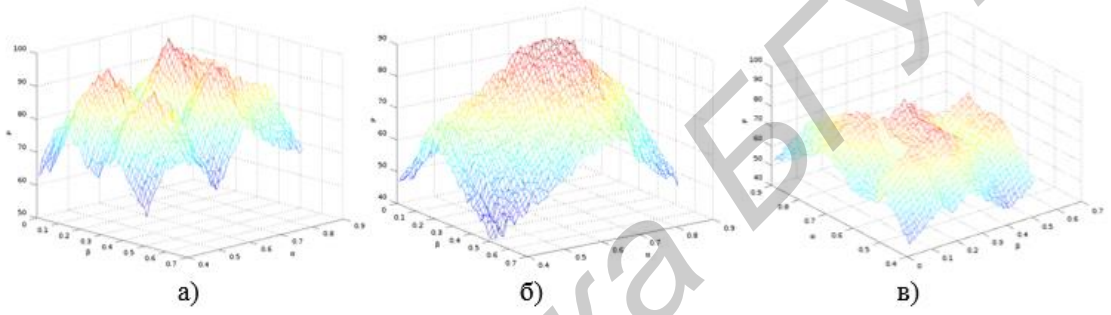

Рис. 6. Экспериментальные данные подбора коэффициентов обучения α и β нейронной сети встречного распространения без обратных связей: а) для бактерии E. coli, б) для бактерии shiwinella Control, в) для бактерии lactobacilus

На основании полученных данных выбраны следующие эмпирические коэффициенты: α = 0,85 и β = 0,15. Подбор начальных коэффициентов повысил точность модели на 2%. Совокупный эффект модификаций составил 12%.

В результате данной работы был спроектирован и реализован классификатор видов бактерий и последовательный кластеризатор питательных сред для бактерии E. coli с точностью классификации 96%.

Литература

[1]. Kong, K. Raman spectroscopy for medical diagnostics: from in-virto biofluid assays to in-vivo cancer detection / K. Kong, C. Kendall, N. Stone, I. Notingher // Advanced Drug Delivery Reviews. – 2015. – Vol.89. – Pp.121-134.

[2]. Olszewski, D. Time Series Visualization Using Asymmetric Self-Organizing Map / D. Olszewski, J. Kacprzyk, S. Zadrozny // Materials of 11th International Conference: Adaptive and Natural Computing Algorithms. – ICANNGA, 2013. – Pp.40-49.

[3]. Yu, D. Supervised Kernel Self-Organizing Map / D. Yu, J. Hu, X. Song, Y. Qi, Z. Tang // Materials of Third Sino-foreign-interchange Workshop: Intelligent Science and Intelligent Data Engineering. – IScIDE, 2012. – Pp.246-253.

[4]. Флах, П. Машинное обучение. Нака и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. – М.: ДМК Пресс, 2015. – 400с.