Критерий качества кластеризации на основе отбора признаков размеченной выборки с приложением в области разработки интерфейсов мозг-компьютер

А. Мазурин[1] А. Бернадотт[2]

В прикладных задачах машинного обучения часто встречается проблема неоднородности выборки. Например, это приводит к трудностям в решении задачи распознавания паттернов электрической активности мозга при разработке нейроинтерфейса у людей разных социальных характеристик.

В работе мы предложили новый метод оценки работы алгоритма кластеризации, имеющий низкие вычислительные затраты и основанный на способности алгоритма распознать скрытые закономерности, то есть выделять группы, схожие по внешнему признаку. Мы показали области практического применения алгоритма, в частности в задачах классификации данных электрической активности мозга при произнесении 8 слов у людей с разными социальными характеристиками.

Ключевые слова: кластеризация, критерий качества кластеризации, нейроинтерфейс.

1. Введение

В прикладных задачах машинного обучения распространена проблема неоднородности выборки. Одним из решений данной проблемы может

[1] *Мазурин Александр Дмитриевич* — студент каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, разработчик в управлении экспериментальных систем машинного обучения департамента SberDevices, Сбер, e-mail: mazurin1567@gmail.com.

Mazurin Alexander — student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems; developer at Experimental ML Systems Subdivision, SberDevices, PJSC Sberbank

[2] *Бернадотт Александра* — исполнительный директор и лидер команды в управлении экспериментальных систем машинного обучения департамента SberDevices, Сбер; доцент каф. инженерной кибернетики в МИСиС; аспирантка каф. математической теории интеллектуальных систем мех.-мат. ф-та МГУ, e-mail: bernadotte.alexandra@intsys.msu.ru.

Bernadotte Alexandra — Executive Director at Experimental ML Systems Subdivision, SberDevices, PJSC Sberbank; assistant professor, Department of Information Technologies and Computer Sciences, National University of Science and Technology MISIS (NUST MISIS); graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of Mathematical Theory of Intellectual Systems.

быть разделение выборки на несколько групп, состоящих из более однородных данных. Последующее обучение с использованием нейронных сетей на каждой из групп по отдельности может повысить точность решения прикладных задач, например, задач классификации. Простое применение классических методов кластеризации не всегда выделяет явный признак, по которому новый элемент выборки может быть отнесён к той или иной группе. Для решения этой задачи можно, с одной стороны, индуцировать разбиение выборки на подгруппы, построив конечное число отображений (внешних признаков) на элементах выборки с дискретной областью значений. С другой стороны, применяя алгоритм кластеризации, можно получить разбиение на кластеры по признаку схожести самих данных и сравнить соответствие этих двух разбиений путём создания метрики сходства.

Для оценки работы алгоритмов кластеризации сейчас используется две категории критериев качества кластеризации – внутренние и внешние. Внешние критерии оценивают результат работы алгоритма по его сходству с известным (согласно мета-данным выборки) разбиением данных на классы, в то время как внутренние критерии – по геометрическим признакам, таким как компактность получившихся кластеров и их разделимость.

В нашей работе мы представляем метод оценки качества кластеризации, соответствующий задаче кластеризации данных электроэнцефалографии согласно социальным мета-данным и используемый для повышения точности последующей классификации мыслительных команд (слов, задающих направление движения и передаваемых в устройстве, соединяющем мозг и компьютер) в проекте по разработке нейроинтерфейса. Представленный метод позволяет оценить сходство полученных в результате работы алгоритма кластеризации классов с разбиением, индуцированным наличием у объектов выборки (на данных электрической активности мозга от 270 испытуемых) внешних признаков – социальных мета-данных, таких, как возраст, профессия и образование. В 2022 году планируется доступ к использованному в нашей работе набору данных электрической активности мозга под лицензией Free BSD 3.

2. Критерий качества кластеризации, основанный на выделении признака

Пусть алгоритм кластеризации завершился разбиением обучающей выборки X на множество C кластеров, имеющее мощность K. Пусть также имеется множество A всех внешних признаков – отображений, ставящих в соответствие элементу выборки x численное значение признака a из

области его значений $dom(a)$. Принципом, на котором основан предлагаемый критерий качества кластеризации, является следующая

Кластеризация выполнена качественно $\Longleftrightarrow$ для каждого признака a из множества признаков A существует кластер из разбиения, все элементы которого принимают одно значение признака a, в то время как значение признака a у всех остальных элементов других кластеров отличается. Формализуем этот критерий в математической форме:

$$M(C) = \sum_{a \in A} \frac{1}{|dom(a)|} \sum_{i \in dom(a)} \max_{c_k \in C} \frac{\frac{1}{|c_k|} \sum\limits_{x_j \in c_k} \mathbb{I}\{a(x_j) = i\}}{\frac{1}{|X|-|c_k|} \left(\sum\limits_{c_l \in C \setminus c_k} \sum\limits_{x_j \in c_l} \mathbb{I}\{a(x_j) = i\} + \varepsilon \right)},$$

где ε – произвольное малое число, необходимое для устранения обращения знаменателя в нуль.

Теорема 1. *Вычислительная затрата предложенного критерия M составляет $O(NK\tilde{A})$ операций, где $N = |X|$ – размер обучающей выборки, K – число кластеров в разбиении, $\tilde{A} = \sum\limits_{a \in A} |dom(a)|$. Если зафиксировано K, а также множество всех внешних признаков и их значений,трудоёмкость вычисления критерия M составит $O(N)$ операций.*

Ситуация, при которой K не является фиксированной величиной, может возникать в задаче подбора оптимального количества кластеров K. При этом отпимальное K может по размерам быть сопоставимо с N, и тогда критерий M имеет квадратичную вычислительную сложность.

В таблице на рис. 1 приведено сравнение по сложности наиболее используемых внешних критериев качества кластеризации и предложенного индекса M, а также объективизация сравнения адекватности различных индексов качества кластеризации в виде представления результатов оптимальной, согласно критериям, кластеризации для восьми слов русского языка, представленных данными электроэнцефалограммы (см. Практическое применение алгоритма).

3. Практическое применение алгоритма

Применение алгоритма существенно повышает точность решения задачи классификации данных электрической активности мозга нейросетевым сверточным рекурретным классификатором[4] на выборке, состоящей из индивидов с разными социальными признаками, такими как пол, возраст (до 25, от 25 до 35, старше 35), образование (гуманитарное, техническое, прикладное). Индекс M позволил выявить более тонкие закономерности

Название критерия	Математическая запись	Описание критерия	Вычислительная сложность	Результаты применения на датасете ЭЭГ (оптимальное число кластеров)
Скорректированный индекс Ренда	$\frac{\sum_{ij} C_{n_{ij}}^2 - \left(\sum_i C_{a_i}^2 \sum_j C_{b_j}^2\right)/C_n^2}{\frac{1}{2}\left(\sum_i C_{a_i}^2 + \sum_j C_{b_j}^2\right) - \left(\sum_i C_{a_i}^2 \sum_j C_{b_j}^2\right)/C_n^2}$, где $n_{ij} = \lvert X_i \cap Y_j \rvert, a_i = \sum_j n_{ij}, b_j = \sum_i n_{ij}$	внешний; выражает схожесть двух разных кластеризаций одной и той же выборки; отличается от индекса Rand введением нормировки; максимизируем	$O(N)$	слово 1: 0.1333 (10 кластеров) слово 2: 0.0813 (19 кластеров) слово 3: 0.0944 (13 кластеров) слово 4: 0.0886 (18 кластеров) слово 6: 0.0641 (20 кластеров) слово 7: 0.0734 (11 кластеров) слово 8: 0.1013 (17 кластеров)
Индекс Жаккара	$\frac{TP}{TP + FN + FP}$	внешний; в отличие от Rand и Adjusted Rand, не учитывает пары элементов находящиеся в разные классах и разных кластерах; максимизируем	$O(N)$	слово 1: 0.1551 (2 кластера) слово 2: 0.1568 (3 кластера) слово 3: 0.1638 (3 кластера) слово 4: 0.1643 (3 кластера) слово 6: 0.2323 (3 кластера) слово 7: 0.2006 (2 кластера) слово 8: 0.2649 (3 кластера)
Индекс Фоулкса-Мэллова	$\sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$	внешний; используется для определения сходства между двумя кластерами и хорошо работает на зашумленных данных; максимизируем	$O(N)$	слово 1: 0.5241 (2 кластера) слово 2: 0.6274 (2 кластера) слово 3: 0.6418 (2 кластера) слово 4: 0.6444 (2 кластера) слово 6: 0.6986 (2 кластера) слово 7: 0.6989 (2 кластера) слово 8: 0.5710 (2 кластера)
Индекс M	$\sum_{a \in A} \frac{1}{\lvert dom(a) \rvert} \sum_{i \in dom(a)} \max_{c_k \in C} R(a,i,k)$, где $R(a,i,k) = \frac{\frac{1}{\lvert c_k \rvert} \sum_{x_j \in c_k} \mathbb{I}\{a(x_j) = i\}}{\frac{1}{\lvert X \rvert - \lvert c_k \rvert}\left(\sum_{c_l \in C \setminus c_k} \sum_{x_j \in c_l} \mathbb{I}\{a(x_j) = i\} + 1\right)}$	внешний; используется для определения сходства между кластерами и заведомо неизвестным разбиением выборки на классы; требует размеченную выборку; максимизируем	$O(N)$	слово 1: 0.5378 (18 кластеров) слово 2: 0.5439 (19 кластеров) слово 3: 0.7926 (18 кластеров) слово 4: 0.7912 (16 кластеров) слово 6: 0.4731(19 кластеров) слово 7: 0.4732 (19 кластеров) слово 8: 0.4760 (20 кластеров)

Рис. 1. В таблице представлены: скорректированный индекс Ренда [1], индекс Жаккара [2], индекс Фоулкса-Мэллова [3], и предложенный в данной статье индекс M. В последней колонке представлено применение индексов для выявления оптимального числа кластеров на данных электрической активности мозга при произнесении 8 слов-команд.

в данных и разделить выборку на кластеры методом k-средних, соответствующие комбинациям мета-данных. Так, выбор оптимального согласно индексу M числа кластеров (18 кластеров) для данных электроэнцефалограммы позволяет точно выделить гомогенные группы испытуемых относительно рода занятий: представителей технического и гуманитарного образования. Уменьшение количества кластеров до оптимального, согласно другим индексам качества кластеризации, не приводило к выделению гомогенных групп испытуемых. Полученное разделение на группы может существенно повысить качество классификации слов при обучении нейросетевого классификатора на группах из гомогенных относительно социальных характеристик данных.

4. Заключение

Подводя итоги вышесказанному, мы предложили новый метод оценки работы алгоритма кластеризации, имеющий низкие вычислительные затраты и основанный не на близости/отдалённости объектов выборки, а на способности алгоритма распознать скрытые от него закономерности, то есть выделять группы, схожие по внешнему признаку. Более того, мы показали области практического применения алгоритма, в частности в задачах классификации данных электрической активности мозга при произнесении 8 слов у людей с разными социальными характеристиками.

Clustering quality criterion based on the features extraction of a tagged sample with an application in the field of brain-computer interface development

Mazurin A., Bernadotte A.

In applied machine learning, the problem of sample heterogeneity is often encountered. For example, the sample heterogeneity leads to serious difficulties in solving the problem of brain electrical activity patterns recognition when developing a brain-computer interface for people of different social characteristics.

In this work, we proposed a new criterion of clustering quality based on the features selection, which has low computing needs and is based

not on the proximity / remoteness of the sampled objects, but on the ability of an algorithm to recognize hidden patterns, that is, to select groups that are similar in features. We have shown the areas of practical application of the algorithm, in particular, in the task of brain electrical activity patterns recognition when pronouncing 8 words by people with different social characteristics.

Keywords: clustering, criterion of clustering quality, brain-computer interface.

References

[1] W. M. Rand, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association. American Statistical Association*, **66**:336 (1971), 846–850.

[2] Paul Jaccard, "The Distribution of the Flora in the Alpine Zone", *New Phytologist*, **11**:2 (1912), 37–50.

[3] E. B. Fowlkes, C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings", *Journal of the American Statistical Association*, **78**:383 (1983), 553.

[4] D. Vorontsova, I. Menshikov, A. Zubov, K. Orlov, P. Rikunov, E. Zvereva, L. Flitman, A. Lanikin, A. Sokolova, S. Markov, A. Bernadotte, "Silent EEG-Speech Recognition Using Convolutional and Recurrent Neural Network with 85% Accuracy of 9 Words Classification", *Sensors*, **21**:20 (2021), 6744.