Math-Net.Ru

Общероссийский математический портал

П. В. Дударин, Н. Г. Ярушкина, Формирование признаков из иерархического классификатора для кластеризации коротких текстовых фрагментов, *Нечеткие системы и мягкие вычисления*, 2017, том 12, выпуск 2, 87–96

DOI: <https://doi.org/10.26456/fssc26>

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением

<http://www.mathnet.ru/rus/agreement>

Параметры загрузки:

IP: 35.174.16.151

26 апреля 2023 г., 14:01:02

ФОРМИРОВАНИЕ ПРИЗНАКОВ ИЗ ИЕРАРХИЧЕСКОГО КЛАССИФИКАТОРА ДЛЯ КЛАСТЕРИЗАЦИИ КОРОТКИХ ТЕКСТОВЫХ ФРАГМЕНТОВ

Дударин П.В., Ярушкина Н.Г.

Ульяновский государственный технический университет, г. Ульяновск

Поступила 09.10.2017, после переработки 15.12.2017.

В данной работе предлагается подход к кластеризации коротких текстовых фрагментов, основанный на особом способе представления объектов в пространстве признаков на базе иерархического классификатора. Также в работе предлагается набор критериев оценки качества кластеризации, позволяющий всесторонне оценить получаемое решение. Проведенные результаты экспериментов показывают способность данного подхода существенно улучшить качество кластеризации.

Ключевые слова: кластеризация, нечеткий классификатор, обработка естественного языка, критерии оценки кластеризации.

Нечеткие системы и мягкие вычисления. 2017. Т. 12, № 2. С. 87–96.
<https://doi.org/10.26456/fssc26>

Введение

Современные методы машинного обучения все чаще становятся комплексными, включая в себя набор переплетенных атомарных методов. Например, кластеризация или классификация с использованием роевых [12], муравьиных [6] или генетических [9] алгоритмов оптимизации, нейронные сети глубокого обучения с отдельными решаемыми задачами на каждом слое.

Аналогично, задача кластеризации ключевых показателей эффективности (КПЭ) Системы Стратегического Планирования Российской Федерации, решаемая авторами [1], декомпозировалась и превратилась в конвейер, состоящий из отдельных задач. В предыдущих работах был получен четкий иерархический классификатор тематик, имеющихся в исследуемом наборе данных [8], а также была построена нечеткая кластеризация КПЭ на базе данного классификатора. Данная кластеризация позволила ответить на вопрос о том, какие показатели относятся к той или иной тематике. Например, если в классификаторе имеется узел, соответствующий тематике «спорт», то на основе нечеткой кластеризации мы можем получить список показателей, имеющих отношение к спорту с определенным весом. Например, узнать, что показатель «результаты игр хоккейного клуба в сезоне» относится к узлу с ключевыми словами «лыжи, хоккей, футбол, баскетбол, волейбол, спорт» со степенью 0.35. При этом остается открытой задача по получению групп КПЭ, относящихся одновременно к схожему набору тематик, т.е. к схожим наборам узлов классификатора. Например, если в классификаторе

имеются узлы, связанные со словами «учитель» и «зарплата», то логично предположить, что имеется группа КПЭ, которую можно было бы назвать «зарплата учителей». Решению данной задачи и посвящена настоящая работа.

Для решения поставленной задачи предлагается рассматривать полученный ранее иерархический классификатор как пространство признаков, т.е. каждому узлу иерархии будет соответствовать отдельная размерность в представлении КПЭ. Соотношение между КПЭ и признаком можно строить на базе отношения R , представляющего нечеткую классификацию КПЭ. При этом легко заметить, что такие признаки не равноценны, например, включение корневого элемента иерархии в пространство признаков полностью лишено смысла с точки зрения информации о кластерах. Поэтому попытка проведения кластеризации без использования особой нормировки и/или весовых коэффициентов в метрике расстояний не даст хороших результатов кластеризации.

Дальнейшее изложение построено следующим образом: в главе 1 приведена формальная постановка задачи и обсуждаются критерии оценки качества решения, в главе 2 представлены варианты решения задачи и анализируются их результаты, в заключении подведены итоги проделанной работы.

1. Постановка задачи и определение критериев оценки качества решения

Формально постановка задачи выглядит следующим образом:

- имеется иерархический классификатор $HG = \langle HV, HE \rangle$, где HV — вершины классификатора, и HE — ребра;
- имеется набор объектов SL (список КПЭ);
- для SL определена нечеткая классификация отношением принадлежности $R : (s \in SL, v \in HV_{\text{лист.}}) \rightarrow [0, 1]$, для родительских узлов отношение определяется как сумма степеней принадлежности соответствующих листовых узлов $R : (s \in SL, v \in HV_{\text{род.}}) = \sum_{n \in HV_{\text{лист.}} \cap HV_{\text{доч.}}(v)} R(s, n)$;
- требуется построить четкую кластеризацию SL на основании пространства признаков задаваемого HG :

$$Cluster(SL) = \left\{ \bigcup_{i \in [1, C]} c_i \mid \forall s \in SL \exists! i \in [1, C], s \in c_i \right\}.$$

Для оценки качества решений, полученных в ходе вычислительных экспериментов, требуется определить критерии этой оценки. В общем случае, задача оценки качества кластеризации не тривиальна, так как каждая задача и каждый алгоритм имеет свои особенности [2]. В некоторых задачах отдается предпочтение определению количества кластеров, в других важнее получить максимально чистые группы. Алгоритмы кластеризации отличаются по форме формируемых кластеров, их плотности и организации в пространстве. Но даже в тех редких случаях, когда существует эталонная или близкая к эталонной кластеризация, решение о выборе критериев не становится очевидным.

Чаще всего в литературе встречаются критерии основанные на сопоставлении результатов для пар объектов. Для вычисления этих критериев предварительно рассчитываются количества истинно-положительных (TP), истинно-отрицательных (TN), ложно-положительных (FP) и ложно-отрицательных (FN) результатов. После этого можно определить критерии такие как Rand Index (RI), Precision, Recall и F-score [7]. Данный набор критериев адекватно отражает качество кластеризации в случае, если количество кластеров совпадает с количеством классов (полученных от экспертов). Если число кластеров будет больше числа классов, то значение критерия recall будет не объективно низким и это отразится на RI и F-score, поэтому в данной работе эти критерии не рассчитываются. Характер критерия precision (попарной точности) ближе к характеру ниже определяемых критериев точности и чистоты, нежели чем к характеру одноименного критерия точности.

Другую группу возможных критериев образуют критерии, основанные на сопоставлении множеств. Наиболее часто используемым критерием такого типа является *чистота* [3]:

$$Purity = \frac{1}{N} \sum_k (\omega_k \cap c_j),$$

где $N = ||SL||$, ω_k — класс k , c_j — кластер j . Данный критерий допустим в ситуации равновеликих классов и кластеров, но не будет работать в случае, если среди классов имеется один доминирующий. В этом случае критерий становится не чувствителен к чистоте минорных кластеров. К сожалению, в рассматриваемом примере имеет место именно такая ситуация, поэтому, учитывая этот недостаток, предлагается использовать *среднюю (микро) чистоту*, а также определить *среднюю (микро) точность* и *среднюю (микро) аккуратность*. Для этого вначале определяется матрица сопоставления множеств:

$$CM = [cm_{i,j} = (\omega_i \cap c_j)],$$

где ω_k — класс k , c_j — кластер j . Для каждой строки можно определить *точность* ($Precision_i$) и *аккуратность* ($Accuracy_i$):

$$Precision_i = \frac{CM(i,j)}{\sum_j CM(i,j)},$$

$$Accuracy_i = \frac{CM(i,j)}{\sum_k CM(k, j_{max})}.$$

Для каждого столбца можно определить *чистоту* ($Purity_i$) [10]:

$$Purity_i = \frac{CM(j,i)}{\sum_j CM(j,i)}.$$

Итоговые оценки получаются усреднением.

Данная система из трех критериев позволяет всесторонне оценить качество кластеризации. Точность показывает степень сосредоточенности элементов одного класса в одном кластере, а аккуратность и чистота в совокупности позволяют судить о количестве лишних элементов в кластерах.

При этом следует отметить, что никакие два из трех критериев не являются достаточными.

Без учета точности результат, содержащий количество кластеров, равное количеству кластеризуемых объектов, будет считаться идеальным. Без аккуратности — решение, в котором почти все элементы помещены в один кластер, будет считаться идеальным.

Без чистоты, в случае, если матрица сопоставления множеств имеет прямоугольную форму, т.е. число возвращаемых кластеров больше числа классов, не получится оценить аккуратность дополнительных кластеров.

В силу специфики решаемой задачи, т.к. число кластеров заранее не известно и является достаточно субъективным, то целесообразно предложить решение из большего числа кластеров, из которых можно собрать более крупные при необходимости, чем получить небольшое количество кластеров, в которых потом придется просматривать все элементы, чтобы найти лишние, даже если их там будет немного. Поэтому при оценке предпочтение будет отдаваться аккуратности и чистоте.

2. Варианты решения и оценка качества на основе эталонной кластеризации

2.1 Описание эталонной кластеризации, используемой в экспериментах

Для проведения экспериментов¹ и получения количественных оценок качества была взята выборка КПЭ относящихся к тематике «дошкольное образование» размером в 3029 КПЭ. Экспертами была проведена кластеризация и выделены 24 кластера. В Таблице 1 для примера приведена часть классов с количеством КПЭ в них.

Таблица 1: Классы тематики «дошкольное образование»

Тематическая группа	Кол-во
Аттестация персонала	75
Вариативные формы обучения	157
Государственные стандарты	9
Готовность выпускников	8
Денежные выплаты и социальное обеспечение	5
Заработная плата персонала	8
Инклюзивное образование	30
Инновационная деятельность	76
Квалификация персонала	47
Научные, творческие и спортивные мероприятия	63
Очередь на поступление в образовательное учреждение	1394
Повышение квалификации персонала	52

Особенностью данной выборки является наличие одного большого кластера «Очередь на поступление в образовательное учреждение» (1394 КПЭ) и большого числа мелких кластеров, что определенно усложняет задачу кластеризации.

¹ Все исходные материалы для проведения экспериментов, программная реализация экспериментов и результаты доступны в проекте на github (<https://github.com/PavelDudarin/sentence-clustering>).

Все перечисленные ниже эксперименты также были проведены для тематик «общее образование» и «высшее образование». Результаты экспериментов были положительно оценены экспертами, что дополнительно подтверждает эффективность предлагаемого подхода.

2.2 Стандартная кластеризация

Итак, иерархический классификатор НГ рассматривается как пространство признаков, в этом случае для каждого $s \in SL$ требуется сопоставить точку в этом пространстве. Наиболее очевидный способ сделать это — воспользоваться отношением принадлежности R из имеющейся нечеткой классификации. Таким образом, определена функция:

$$F : SL \rightarrow R^n, \forall s \in SL F(s) = (s_1, s_2, \dots, s_n),$$

$$s_i = R(s, v_i), n \in [1, ||HV||], v_i \in HV.$$

Первый эксперимент проводился с использованием алгоритма k-means² [4]. Данный алгоритм вернул несколько крупных кластеров, в которые попало большинство КПЭ. Оценки качества кластеризации приведены в Таблице 2 в строке 1.

Второй эксперимент проводился с использованием алгоритма HDBScan [5]. Данный алгоритм в качестве альтернативного выбран по трем соображениям. Во-первых, это один из самых современных и широко используемых алгоритмов кластеризации. Во-вторых, метод поиска кластеров в нем опирается на плотность, а не на расстояние, что потенциально дает лучший эффект в многомерных пространствах. В-третьих, он имеет эффективную проверенную программную реализацию. Результаты приведены в Таблице 2 строка 2.

Алгоритма HDBScan относится к алгоритмам истинной кластеризации, а не партиционирования как k-means, т.к. в ходе его работы часть объектов расценивается как шум и не соотносится ни с одним кластером. Такие элементы объединяются в один кластер «шума» [11]. Толкование этого кластера зависит от задачи, некоторые задачи действительно содержат зашумленные данные, в этом случае эти элементы следует отбросить. В задаче кластеризации КПЭ в кластер «шума» чаще всего попадают показатели с длинной или нестандартной формулировкой, в этом случае они могут быть соотнесены с выявленными кластерами процедурой классификации.

Также можно заметить большое количество кластеров по сравнению с количеством классов, что негативно влияет на оценку точности алгоритма. Причина в указании параметра минимального количества элементов в кластере равного 7, что связано с желанием отдать приоритет аккуратности и чистоте по соображениям, приведенным в главе 2.

Расчет качества результатов работы алгоритма HDBScan производится без учета кластера «шума», т.к. эти элементы считаются не кластеризованными. При этом некоторые классы могут целиком попадать в кластер «шума», что крайне нежелательно, т.к. в этом случае последующая классификация кластера «шума» по результатам кластеризации даст недостоверный результат. Поэтому важно обращать внимание на соотношение точности с учетом кластера «шума» и без него.

²Для проведения численных экспериментов были использованы библиотеки sklearn.cluster и hdbscan на языке python. Кластеризация алгоритмом HDBScan проводилась с параметром min_cluster_size = 7.

Оба эксперимента показывают недостаточную аккуратность и чистоту кластеров, и это не удивительно. При таком подходе, учитывая, что в алгоритмах кластеризации чаще всего используется один из вариантов мер Минковского, все элементы иерархии вносят одинаковый вклад в расстояние между объектами. При кластеризации верхние узлы иерархии будут «стягивать» все объекты в один кластер, что отразится в высокой точности, но низкой аккуратности и чистоте кластеров в результатах экспериментов, что и наблюдается в результатах экспериментов.

2.3 Кластеризация с учетом весов узлов классификатора

Для того чтобы отрегулировать влияние узлов иерархии, ослабив мало значимые и усилив значимые, предлагается использовать следующую функцию:

$$F : SL \rightarrow R^n, \forall s \in SL F(s) = (s_1, s_2, \dots, s_n),$$

$$s_i = R(s, v_i) * \frac{1}{\omega(v_i) * \sigma * \sqrt{2\pi}} e^{-\frac{(\ln \omega(v_i) - \mu)^2}{2 * \sigma^2}},$$

$$\omega(v_i) = ||\{s \in SL | R(s, v_i) > 0\}||, n \in [1, ||HV||], v_i \in HV.$$

Функция весового множителя обладает следующими свойствами:³

- при $\omega(v_i) = 0$ функция обращается в 0,
- при $\omega(v_i) \rightarrow \infty$ функция стремится к 0.

На положительном участке оси абсцисс функция имеет один глобальный максимум.

Параметр σ взят равным 1, а параметр μ подбирается таким образом, чтобы максимум весовой функции приходился на среднее значение количества показателей, соответствующих вершинам второго и третьего с конца уровней иерархии⁴. В данном примере $\mu = 3.7$, так, чтобы максимум приходился на 15 показателей в одном узле иерархии.

Результаты экспериментов для алгоритмов k-means и HDBScan приведены в Таблице 2 в строках 3 и 4 соответственно. Результаты k-means по-прежнему неудовлетворительные, это связано с наличием одного мажорного кластера в выборке. А вот результаты HDBScan показывают отличный результат. Алгоритм вернул набор абсолютно аккуратных и чистых кластеров. Также можно заметить уменьшение размера кластера «шума» и количества возвращаемых кластеров, что является желательным эффектом.

При этом, сравнив значения в колонках 3 и 4 Таблицы 2, можно заметить, что в кластер «шума» целиком попала часть классов, а точнее два — размер которых меньше 7 элементов. Изменив параметры HDBScan, установив, например, размер минимального кластера равного 5, можно добиться обнаружения всех классов, заплатив за это большим числом результирующих кластеров. Результаты данного эксперимента представлены в Таблице 2 в строке 5.

³ В данном примере взята функция логнормального распределения вероятностей, но может быть использована любая функция, обладающая указанными свойствами.

⁴ Данное правило было выбрано эмпирически в результате проведения более чем 5000 экспериментов.

Таблица 2: Результаты экспериментов

Название	Точность, + шум, %	Точность, %	Аккумулятивность, %	Чистота, %	Среднее, %	Шум, шт.	Классификация, шт.
1.K-means	-	64.9	39.9	84.7	63.2	-	24
2.HDbScan	51.4	55.5	84.8	97.2	79.2	914	104
3.K-means (вес)	-	76.8	43.2	94.4	71.5	-	24
4.HDBScan 7 (вес)	65	69.6	99.1	99.1	89.2	663	128
5.HDBScan 5 (вес)	68.1	68.1	99.1	99.4	88.9	441	189

Заключение

В данной работе был предложен способ представления объектов в пространстве признаков на базе иерархического классификатора, позволяющий эффективно решать соответствующую задачу кластеризации, а также был предложен набор критериев оценки качества кластеризации, позволяющий всесторонне оценить получаемое решение. Будучи примененным в задаче кластеризации коротких текстовых фрагментов, данный способ позволил существенно улучшить получаемое решение (с 79.1% до 89.2%). Основными недостатками предложенного подхода являются: избыточное число кластеров и достаточно большой кластер шума. В рамках решаемой задачи указанные недостатки являются, скорее, особенностями и не препятствуют получению качественного решения, тем не менее, в дальнейшем планируется провести работу по улучшению предлагаемого подхода.

Список литературы

- [1] Дударин П.В., Ярушкина Н.Г. Алгоритм построения иерархического классификатора коротких текстовых фрагментов на основе кластеризации нечеткого графа // Радиотехника. 2017. № 6. С. 114–121.
- [2] Сибирев И.В. Индексы оценки результатов кластеризации // Нечеткие системы, мягкие вычисления и интеллектуальные технологии (НСМВИТ-2017): труды VII всероссийской научной-практической конференции (г. Санкт-Петербург, 3–7 июля, 2017 г.). Т. 1. С. 129–138.
- [3] Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints // Information Retrieval. 2009. Vol. 12, № 4. Pp. 461–486. <https://doi.org/10.1007/s10791-008-9066-8>
- [4] Ball G.H., Hall D.J. Isodata, a Novel Method of Data Analysis and Pattern Classification. Technical Report prepared for Information Sciences Branch, Office of Naval Research. Menlo Park, California: Stanford Research Institute, 1965.
- [5] Campello R.J.G.B., Moulavi D., Sander J. Density-Based Clustering Based on Hierarchical Density Estimates // Advances in Knowledge Discovery and Data Mining. PAKDD 2013. Series: Lecture Notes in Computer Science. Vol. 7819. Berlin, Heidelberg: Springer, 2013. Pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14

- [6] Chen Y., Han M., Zhu H. Ant spatial clustering based on fuzzy IF-THEN rule // Fuzzy Information and Engineering 2010. Series: Advances in Intelligent and Soft Computing. Vol. 78. Berlin, Heidelberg: Springer, 2010. https://doi.org/10.1007/978-3-642-14880-4_62
- [7] Manning C.D., Raghavan P., Schütze H. An Introduction to Information Retrieval. Cambridge, England: Cambridge University Press, 2009.
- [8] Dudarin P.V., Yarushkina N.G. An approach to fuzzy hierarchical clustering of short text fragments based on fuzzy graph clustering // Proceedings of the Second International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'17). IITI 2017. Series: Advances in Intelligent Systems and Computing. Vol. 679. Cham: Springer, 2018. Pp. 295–304. https://doi.org/10.1007/978-3-319-68321-8_30
- [9] Kapil S., Chawla M., Ansari M.D. On K-means data clustering algorithm with genetic algorithm // Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC). Wagnaghat, 2016. Pp. 202–206.
- [10] Mansoori E.G. GACH: a grid based algorithm for hierarchical clustering of high-dimensional data // Soft Computing. 2014. Vol. 18, № 5. Pp. 905–922. <https://doi.org/10.1007/s00500-013-1105-8>
- [11] McInnes L., Healy J., Astels S. hdbscan: Hierarchical density based clustering // Journal of Open Source Software. 2017. Vol. 2, № 11. <https://doi.org/10.21105/joss.00205>
- [12] Zhang J., Wang Y., Feng J. A hybrid clustering algorithm based on PSO with dynamic crossover // Soft Computing. 2014. Vol. 18, № 5. Pp. 961–979. <https://doi.org/10.1007/s00500-013-1115-6>

Образец цитирования

Дударин П.В., Ярушкина Н.Г. Формирование признаков из иерархического классификатора для кластеризации коротких текстовых фрагментов // Нечеткие системы и мягкие вычисления. 2017. Т. 12, № 2. С. 87–96. <https://doi.org/10.26456/fssc26>

Сведения об авторах

1. Дударин Павел Владимирович

аспирант кафедры «Информационные системы» Ульяновского государственного технического университета.

Россия, 432000, г. Ульяновск, ул. Радищева, д. 102, УлГТУ.

E-mail: p.dudarin@ulstu.ru

2. Ярушкина Надежда Глебовна

зав. кафедрой «Информационные системы» Ульяновского государственного технического университета.

Россия, 432000, г. Ульяновск, ул. Радищева, д. 102, УлГТУ.

E-mail: jng@ulstu.ru

FEATURES CONSTRUCTION FROM HIERARCHICAL CLASSIFIER FOR SHORT TEXT FRAGMENTS CLUSTERING

Dudarin Pavel Vladimirovich

PhD student at Information Systems department,
Ulyanovsk State Technical University

Russia, 432000, Ulyanovsk, Radisheva st., 102, UlSTU. E-mail: p.dudarin@ulstu.ru

Yarushkina Nadezhda Glebovna

Head of Information Systems department,
Ulyanovsk State Technical University

Russia, 432000, Ulyanovsk, Radisheva st., 102, UlSTU. E-mail: jng@ulstu.ru

Received 09.10.2017, revised 15.12.2017.

In this paper an approach to clustering of short text fragments by hierarchical classifier is proposed. This approach is based on a special way of object representation in the dimension of features constructed by hierarchical classifier. Also, in order to reasonable perform quality evaluation of clustering results there is a set of criteria proposed in this paper. Presented experimental results show the ability of this approach to improve significantly the quality of clustering.

Keywords: clustering, fuzzy classifier, natural language processing, evaluation of clustering.

Nechetkie Sistemy i Myagkie Vychisleniya [Fuzzy Systems and Soft Computing], 2017, vol. 12(2), pp. 87–96. <https://doi.org/10.26456/fssc26>

References

- [1] Dudarin P.V., Jarushkina N.G. Algorithm for constructing a hierarchical classifier of short text fragments based on the clustering of a fuzzy graph. *Radiotekhnika* [Radio Engineering], 2017, no. 6, pp. 114–121. (in Russian)
- [2] Sibirev I.V. Indices of the evaluation of clustering results. *Nechetkie Sistemy, Myagkie Vychisleniya i Intellektual'nye Tehnologii (NSMVIT-2017): Trudy VII Vserossiyskoj Nauchnoy-Prakticheskoy Konferencii* [Fuzzy Systems, Soft Computing and Intellectual Technology (NSMVIT-2017): Proceedings of the VII All-Russian Scientific and Practical Conference]. Saint-Petersburg, 3–7 July 2017. Vol. 1. Pp. 129–138. (in Russian)
- [3] Amigo E., Gonzalo J., Artiles J., Verdejo F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval*, 2009, vol. 12(4), pp. 461–486. <https://doi.org/10.1007/s10791-008-9066-8>
- [4] Ball G.H., Hall D.J. *Isodata, a Novel Method of Data Analysis and Pattern Classification*. Technical Report prepared for Information Sciences Branch, Office of Naval Research. Stanford Research Institute, Menlo Park, California, 1965.

- [5] Campello R.J.G.B., Moulavi D., Sander J. Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining. PAKDD 2013*. Series: Lecture Notes in Computer Science. Vol. 7819. Springer, Berlin, Heidelberg, 2013. Pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14
- [6] Chen Y., Han M., Zhu H. Ant spatial clustering based on fuzzy IF-THEN rule. *Fuzzy Information and Engineering 2010*. Series: Advances in Intelligent and Soft Computing. Vol. 78. Springer, Berlin, Heidelberg, 2010. https://doi.org/10.1007/978-3-642-14880-4_62
- [7] Manning C.D., Raghavan P., Schütze H. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2009.
- [8] Dudarin P.V., Yarushkina N.G. An approach to fuzzy hierarchical clustering of short text fragments based on fuzzy graph clustering. *Proceedings of the Second International Scientific Conference «Intelligent Information Technologies for Industry» (IITI'17). IITI 2017*. Series: Advances in Intelligent Systems and Computing. Vol. 679. Springer, Cham, 2018. Pp. 295–304. https://doi.org/10.1007/978-3-319-68321-8_30
- [9] Kapil S., Chawla M., Ansari M.D. On K-means data clustering algorithm with genetic algorithm. *Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)*. Wagnaghat, 2016. Pp. 202–206.
- [10] Mansoori E.G. GACH: a grid based algorithm for hierarchical clustering of high-dimensional data. *Soft Computing*, 2014, vol. 18(5), pp. 905–922. <https://doi.org/10.1007/s00500-013-1105-8>
- [11] McInnes L., Healy J., Astels S. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2017, vol. 2(11). <https://doi.org/10.21105/joss.00205>
- [12] Zhang J., Wang Y., Feng J. A hybrid clustering algorithm based on PSO with dynamic crossover. *Soft Computing*, 2014, vol. 18(5), pp. 961–979. <https://doi.org/10.1007/s00500-013-1115-6>

Citation

Dudarin P.V., Yarushkina N.G. Features construction from hierarchical classifier for short text fragments clustering. *Nechetkie Sistemy i Myagkie Vychisleniya* [Fuzzy Systems and Soft Computing], 2017, vol. 12(2), pp. 87–96. (in Russian) <https://doi.org/10.26456/fssc26>