DOI 10.31891/2307-5732-2018-267-6(2)-213-215
UDC 004.021::004.627

V.V. ROMANUKE
Polish Naval Academy, Poland, Gdynia

OPTIMAL PARTITIONING OF AN INITIAL DATASET INTO SUBDATASETS TO BE CLUSTERED FOR GETTING RID OFF THE DATASET SUPERFLUITIES FOR A MACHINE LEARNING TASK

As preprocessing huge datasets may consume far more resources than solving a machine learning task, an approach to optimal partitioning an initial dataset into subdatasets is suggested. Every subdataset is subsequently clustered in order to filter surplus objects from it. This is fulfilled based on the previously suggested approach to optimizing a dataset by clustering it and selecting closest-to-the-centroid objects, which constitute thus a refined dataset. Firstly, it is described how subdatasets are obtained. An initial number of objects is 2 to the power of some integer. This integer is presumed to be greater than 6 because a dataset of less than 100 objects is counted small, and so partitioning it further is hardly reasonable. Both the number of entries in a subdataset and the number of subdatasets of the same type also are 2 to the power of some integers. When, secondly, the dataset hugeness is explained and formalized, a factor showing how the subdataset may be maximally "squeezed" is introduced. This factor can be equal to from 2 up to the power integer for the initial number of objects decreased by 1. If this factor is great, it shows superfluities in the initial dataset. Besides, it defines the number of different types of subdatasets. For lesser values of the factor, the dataset superfluity is less, but the number of subdatasets grows. The minimal value of the factor gives the least hugeness of the dataset, where it can be 4 times "squeezed" at the most. Finally, the initial dataset is optimally partitioned by a squeezing factor and a certain subdataset type, at which the time taken by clustering is minimized. A data augmentation technique can be used to achieve such a 2-to-the-power representation. This is for the approach could be efficiently parallelized.

Keywords: machine learning, dataset, clustering, dataset partitioning, subdatasets.

В. В. РОМАНЮК
Військово-морська академія Польщі, Гдиня

ОПТИМАЛЬНЕ РОЗБИТТЯ ПОЧАТКОВОГО НАБОРУ НА ПІДРОЗДІЛИ ДАНИХ ДЛЯ ЇХ ПОДАЛЬШОЇ КЛАСТЕРИЗАЦІЇ З МЕТОЮ ЗНЯТТЯ НАДЛИШКІВ У НАБОРІ ДЛЯ ЗАДАЧІ МАШИННОГО НАВЧАННЯ

Оскільки передпроцесінг великих наборів даних може потребувати набагато більше ресурсів, ніж вирішення задачі машинного навчання, пропонується підхід до оптимального розбиття вихідного набору даних на підрозділи. Кожен підрозділ даних потім кластеризується, щоб відфільтрувати надлишкові об'єкти з нього. Це виконується на основі запропонованого раніше підходу до оптимізації набору даних шляхом кластеризації та вибору найближчих до центроїдів об'єктів, які, таким чином, утворюють удосконалений набір даних. Початкове число об'єктів дорівнює 2 у деякій цілій степені. І кількість записів у підрозділі, і кількість підрозділів того ж типу також дорівнюють 2 у деяких цілих степенях. Вводиться фактор, який показує, як можна максимально "стиснути" підрозділ. Для менших значень цього фактора надлишковість набору даних менша, але кількість підрозділів зростає. Його мінімальне значення означає найменшу надлишковість набору даних, де він може бути "стиснутий" в 4 рази щонайбільше. Початковий набір даних оптимально розбивається за такого фактора стискування та певного типу підрозділів, за яких час кластеризації мінімізується. Для здійснення вказаного подання у формі "2 у степені" можна використати техніку збільшення даних.

Ключові слова: машинне навчання, набір даних, кластеризація, розбиття набору даних, піднабори даних.

Introduction and motivation

The dataset is a crucial part in solving any machine learning task. Recently, it was shown in [1] that a dataset may be optimized by clustering. Article [1] explains that, in a wider sense, a dataset is optimized by filtering surplus objects from it. Namely, an approach to forming an optimal dataset (either of real-world objects or synthetic ones) for a machine learning task was suggested in [1] for when an initial number of objects is significantly greater than required. The proposed approach relies on an appropriately selected algorithm of clustering and a distance [2, 3]. It considers two cases of the number of objects, at which the training process is presumably close to optimal. In the case #1, the number is unknown but included into an interval between the known integers. Then, the optimal number of objects is determined by using the silhouette criterion [4]. Here, the optimal number of objects to be included into the corresponding dataset is the optimal number of clusters at which the maximum of the silhouette criterion function is achieved. When the optimal number of dataset entries is known, i. e. determined by using the silhouette criterion or known-beforehand (the case #2), the initial set of objects is clustered, where the number of clusters is equal to that number of dataset entries. In each cluster, the object closest to the cluster centroid is the best one for including it into the dataset. The closeness is treated by the same distance used previously in the silhouette criterion function and clustering [3]. The closest-to-the-centroid objects are found by minimizing the distance to the centroid. So, the optimal dataset consists of such objects.

Article [1] also suggested that if an initial number of objects is too great, it would be reasonable to break them into a few groups. This is crucial for accelerating the process of clustering. Thus, an optimal subdataset will be formed from each group by using the same approach of clustering and selecting closest-to-the-centroid objects. However, when should an initial group of objects be broken for forming optimal subdatasets? It is obvious that a criterion for partitioning an initial dataset into subdatasets to be optimized is the amount of resources spent for the process of clustering. In particular, it is the time which is taken by clustering. Besides, if the optimal number of objects for optimal subdatasets is unknown, then using the silhouette criterion is too much time-consuming.

So, the question is how should huge initial datasets be partitioned before clustering? Is it possible to know a close-to-optimal number of subdatasets, each of which will be clustered separately? Before answering these questions, the hugeness of a dataset must be formalized. Representation of classes and a partitioning must be linked.

Goal of the article and tasks to be fulfilled

The goal of the article is to develop an approach to optimal partitioning an initial dataset into subdatasets to be clustered. This will provide a faster and more efficient preparation of an optimal dataset for a machine learning task. For achieving the article's goal, the following three tasks are to be fulfilled: 1) to describe how subdatasets are obtained; 2) to formalize the dataset hugeness; 3) to describe how an optimal partitioning is searched.

Subdatasets and superfluities in the initial dataset

Let Q be an initial number of objects. These objects constitute the initial dataset. For convenience of partitioning the dataset into subdatasets of an equal volume, let $Q = 2^W$ by $W \in \mathbb{N} \setminus \overline{\{1, 6\}}$. The reason for such W is that a dataset of less than 100 objects is counted small, and so partitioning it further is hardly reasonable. A matter of object representation in each class will be discussed below.

In the u-th subdataset type, let integer N_u be a number of objects, at which the training process is presumably close to optimal. If there are initially, say, $Q = 1024$ objects, then such a dataset can be partitioned into the following versions of subdatasets: 1) two subdatasets by 512 objects; 2) four subdatasets by 256 objects; 3) eight subdatasets by 128 objects; ...; 7) 128 subdatasets by eight objects. Surely, the case when no partition is reasonable is additionally included, where, potentially, $N_1 \in \overline{\{256, 512\}}$ by holding at a kind of principle of double-integer uncertainty [5]. For the case of two subdatasets, $N_2 \in \overline{\{128, 256\}}$ and $N_3 \in \overline{\{64, 128\}}$ for the case of four subdatasets, and so on. And it is reasonable that $N_8 \in \overline{\{2, 4\}}$ because of only eight objects in each of 128 subdatasets. Thus, every subdataset is at least 2 times "squeezed" (i. e., optimized or filtered). At the most, it may be 4 times "squeezed".

However, this has been just a single example of how the dataset is partitioned (eight cases including the case with no partition by $u = 1$). Here is another example (for the same initial set of 1024 objects), where every subdataset is from 4 to 8 times "squeezed": 1) two subdatasets by 512 objects and $N_2 \in \overline{\{64, 128\}}$; 2) four subdatasets by 256 objects and $N_3 \in \overline{\{32, 64\}}$; 3) eight subdatasets by 128 objects and $N_4 \in \overline{\{16, 32\}}$; ...; 6) 64 subdatasets by 16 objects and $N_7 \in \overline{\{2, 4\}}$. The case when no partition is reasonable comes at $N_1 \in \overline{\{128, 256\}}$.

Continuing on the same logics, there is a case of an ultimate squeezing, where the initial dataset is not partitioned at all but only from 2 to 4 best objects are filtered out. In other words, here the dataset is from 512 down to 256 times "squeezed". Formally, there is the single subdataset type represented singly by the initial dataset, and $N_1 \in \overline{\{2, 4\}}$ for such a case.

Henceforward, a factor

$$z \in \overline{\{2,\ \log_2 Q - 1\}} \quad \text{or} \quad z \in \overline{\{2,\ W - 1\}} \tag{1}$$

shows how the subdataset may be maximally "squeezed": the number of squeezing times is equal to 2^z. Integer 2^z is the maximally possible squeezing coefficient. If factor (1) is great, it shows superfluities in the initial dataset. Besides, it defines the number of different types of subdatasets, which is equal to $W - z$. The total number of subdatasets of the u-th subdataset type is 2^{u-1} by $u = \overline{1,\ W - z}$. Every subdataset of u-th type has $Q_u = Q/2^{u-1}$ entries. For the three cases, exemplified above, factor z is equal respectively to 2, 3, and 9. Moreover, the number of objects, at which the training process is presumably close to optimal for every subdataset of the u-th type, is

$$N_u \in \overline{\left\{N_{\min}^{\langle u \rangle},\ N_{\max}^{\langle u \rangle}\right\}} \quad \text{by} \quad N_{\min}^{\langle u \rangle} = 2^{W-z-u+1} \quad \text{and} \quad N_{\max}^{\langle u \rangle} = 2N_{\min}^{\langle u \rangle} \quad \text{for} \quad u = \overline{1,\ W - z} \tag{2}$$

by factor (1). Obviously, setting $z = W - 1$ may serve as an evidence of hugeness of the dataset. For lesser z, the dataset superfluity is less, but the number of subdatasets grows. Factor $z = 2$ gives us the least hugeness of the dataset, where it can be 4 times "squeezed" at the most. Superfluity and hugeness imply here much the same.

Searching for an optimal partitioning

Obviously, rationality of the partitioning depends on the number F of object features. Therefore, the time t which is taken by clustering can be considered as a function $t(z, u, F)$. Then the initial dataset is optimally partitioned by the parameters

$$[u^*(F) \quad z^*(F)] \in \arg \min_{z \in Z \subset \overline{\{2,\ W-1\}}} \left\{ \min_{u = \overline{1,\ W-z}} t(z, u, F) \right\} \tag{3}$$

where $u^*(F)$ is the optimal subdataset type, and $z^*(F)$ is the optimal factor for squeezing the subdatasets by the given number of object features, and $Z \subset \overline{\{2,\ W - 1\}}$ is an admissible subset of the squeezing factors suitable for the given task. Occasionally, if the number of object features can be varied, then, for instance, it is

$$F \in \{\alpha F_{\min}\}_{\alpha=1}^{\alpha_{\max}} \quad \text{by} \quad \alpha_{\max} \in \mathbb{N} \setminus \{1\} \tag{4}$$

where $F_{\min}$ is a minimal number of features (see Figure 1 along with Figure 2 showing influence of this number).

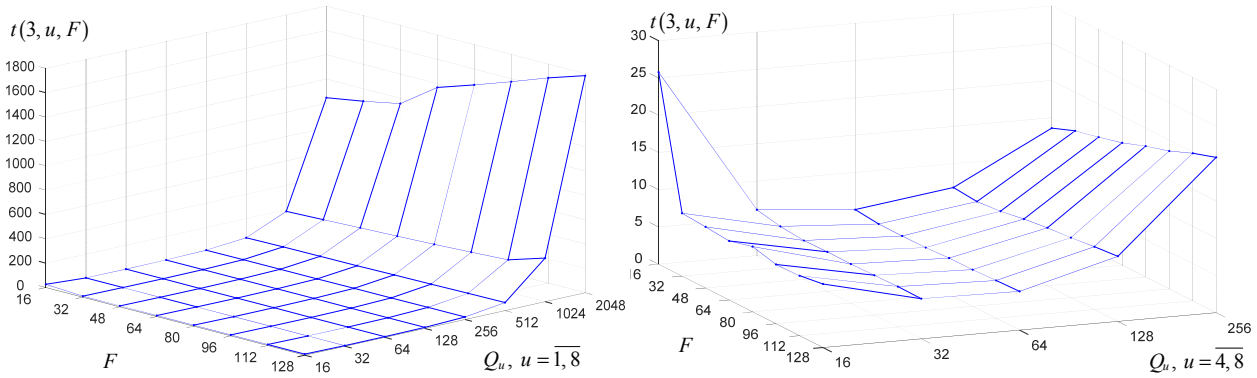

Figure 1. An example of the time spent for the clustering by varying the number of object features according to (4), where $z = 3$

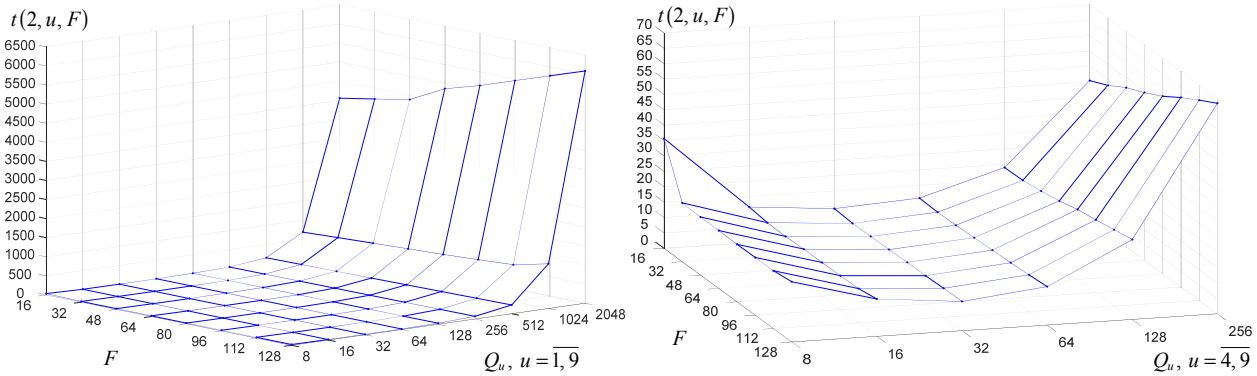

Figure 2. The added example of the time spent for the clustering by varying the number of object features according to (4), where $z = 2$

The example visualized above stands for not much superfluities in the initial dataset. It solves problem (3) for $Z = \{2, 3\}$ and any number of features from 16 to 128: 32 datasets by 64 entries are clustered into 8 to 16 clusters faster (see Figure 1 to the right) than 64 datasets by 32 entries are clustered into the same interval of clusters (see Figure 2 to the right). Influence of the number of object features on the time spent for the clustering is minor.

Discussion and conclusion

If the dataset represents D classes, with d_i objects in the i-th class, then they should satisfy a statement

$$\min_{i=\overline{1,\,D},\ k=\overline{1,\,D}} \{\max\{d_i/d_k,\, d_k/d_i\}\} \quad \text{by} \quad \sum_{i=1}^{D} d_i = Q = 2^W. \tag{5}$$

If initially $Q \neq 2^W$, this can be achieved by using a data augmentation technique [6, 7] allowing also to achieve approximately equal representation of classes. Similar equal representation of each class should be achieved for every subdataset, where the principle in (5) can be used. In the naive way, the classes must be "allocated" randomly.

The proposed 2^W-approach is intended to be efficiently parallelized. It relies on statement (5) holds, whereupon problem (3) is solved numerically. After subdatasets are determined, the approach of clustering and selecting closest-to-the-centroid objects [1] will filter surplus objects from every subdataset. Then, the optimized subdatasets are gathered into the refined dataset, which is considered to be optimal itself for a machine learning task.

References

1. Romanuke V. V. Optimization of a dataset for a machine learning task by clustering and selecting closest-to-the-centroid objects // Herald of Khmelnytskyi national university. Technical sciences. — 2018. — No. 6, Vol. 1. — P. 263 — 265.

2. Nylen E. L., Wallisch P. Neural Data Science. Chapter 9 — Classification and Clustering / Academic Press, 2017. — P. 249 — 276.

3. Larsson C. 5G Networks. Chapter 6 — Clustering / Academic Press, 2018. — P. 123 — 141.

4. Campello R. J. G. B., Hruschka E. R. A fuzzy extension of the silhouette width criterion for cluster analysis // Fuzzy Sets and Systems. — 2006. — Vol. 157, Iss. 21. — P. 2858 — 2875.

5. Romanuke V. V. Interval uncertainty reduction via division-by-2 dichotomization based on expert estimations for short-termed observations // Journal of Uncertain Systems. — 2018. — Vol. 1, No. 12. — P. 3 — 21.

6. Romanuke V. V. Training data expansion and boosting of convolutional neural networks for reducing the MNIST dataset error rate // Research Bulletin of NTUU "Kyiv Polytechnic Institute". — 2016. — No. 6. — P. 29 — 34.

7. Lv J.-J., Shao X.-H., Huang J.-S., Zhou X.-D., Zhou X. Data augmentation for face recognition // Neurocomputing. — 2017. — Vol. 230. — P. 184 — 196.

Рецензія/Peer review : 19.11.2018 р. Надрукована/Printed :19.12.2018 р.
Стаття прорецензована редакційною колегією