DOI 10.31891/2307-5732-2018-267-6(1)-263-265
UDC 004.021::004.627

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

OPTIMIZATION OF A DATASET FOR A MACHINE LEARNING TASK BY CLUSTERING AND SELECTING CLOSEST-TO-THE-CENTROID OBJECTS

An approach to forming an optimal dataset (either of real-world objects or synthetic ones) for a machine learning task is suggested 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. Two cases of the number of objects, at which the training process is presumably close to optimal, are considered. The number may be known beforehand or unknown but included into an interval between the known integers. In this case, the optimal number of objects is determined by using the silhouette criterion. Then 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. For the known-beforehand or determined optimal number of dataset entries, 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. So, the optimal dataset consists of the closest-to-the-centroid objects found by minimizing the distance to the centroid. For relatively small datasets (required, for instance, for transfer learning tasks) of a few hundred entries, the silhouette criterion function performs much faster. If an initial number of objects is too great, it is reasonable to break them into a few groups. A subdataset will be formed from each group by using the same approach of clustering and selecting closest-to-the-centroid objects. In a wider sense, the proposed approach allows to filter surplus objects from a dataset, thus optimizing it. Generally speaking, the clustering here consumes far more resources than selection of closest-to-the-centroid objects. However, an open question is when an initial group of objects should be broken for forming optimal subdatasets. In this way, the clustering can be clustered itself to accelerate optimization of a dataset.

Keywords: machine learning, dataset, clustering, distance, cluster centroids, silhouette criterion.

В.В. РОМАНЮК
Військово-морська Академія Польщі, Польща, Гдиня

ОПТИМІЗАЦІЯ НАБОРУ ДАНИХ ДЛЯ ЗАДАЧІ МАШИННОГО НАВЧАННЯ ШЛЯХОМ КЛАСТЕРИЗАЦІЇ І ВИБОРУ НАЙБЛИЖЧИХ ДО ЦЕНТРОЇДІВ ОБ’ЄКТІВ

Пропонується підхід до формування оптимального набору даних (реальних або синтетичних об’єктів) для задачі машинного навчання, коли початкове число об’єктів значно більше, ніж потрібно. Запропонований підхід спирається на відповідно обрані алгоритм кластеризації та відстань. Розглянуто два випадки кількості об’єктів, за яких навчальний процес вважається близьким до оптимального. Ця кількість може бути відомою заздалегідь або невідомою, але включеною в інтервал між відомими цілими числами. У цьому випадку оптимальне число об’єктів визначається за допомогою силует-критерію. Тоді оптимальне число об’єктів, що підлягають включенню у відповідний набір даних, — це та оптимальна кількість кластерів, за якої досягається максимум функції силует-критерію. Для відомого заздалегідь або визначеного оптимального числа записів набору даних початкова множина об’єктів кластеризується, де кількість кластерів дорівнює кількості записів набору даних. У кожному кластері об'єкт, найближчий до кластерного центроїда, є найкращим для включення його в набір даних. Близькість тут обробляється на тій же відстані, що використовувалася раніше в функції силует-критерію та кластеризації.

Ключові слова: машинне навчання, набір даних, кластеризація, відстань, кластерні центроїди, силует-критерій.

Introduction and motivation

Along with computer systems, machine learning has influenced a lot of hardware-based information technologies, including automatics, robotics, telecommunication, credit-card control, etc. Preparation of datasets for machine learning tasks is a quite principal stage in the field. An appropriately prepared dataset is partitioned in training, validation, and testing sets, which define the quality of learning and prediction [1, 2].

The dataset volume required for successful training depends on the task itself and its complexity. For instance, by roughly the same number of classes or categories, image classification tasks are far simpler than tasks of semantic image segmentation [3]. In its turn, scene categorization is usually a more complex task than just classifying images [1, 2, 4, 5]. Building a dataset addresses collecting, selecting, processing, grouping, and labeling data. For small datasets, when no more data are available than a collected amount of entries, a few techniques of data augmentation for training are applied [6, 7].

Some machine learning tasks are based on artificial datasets (e. g., like MNIST, NORB, EEACL26 [2, 5, 6, 8]), which are used for training and testing machine learning models. A dataset wholly generated by computer is also called synthetic. In particular, EEACL26 is a synthetically generated dataset for image classification. It is an infinitely scalable set of grayscale images which can be represented in any size [8, 9]. However, the question is whether an optimal volume of a dataset can be found. If a dataset has many groups of similar objects, then the respective machine learning algorithm may be prone to overfitting, or the training process will not converge properly. Such cases happen to non-synthetic datasets as well. For example, technically different photos of an object, which were made from slight displacements of the camera, bring information equivalent to a single entry rather than a bunch of entries. Although some variations of the entry close to those different photos are generated during augmentation, the initial non-augmented dataset should contain as more original entries as possible. This is very important for proper validation and final testing, because if there is a group of a few similar entries, the group may be partitioned so that almost the same data will be in the training set and validation set. Eventually, this poor partitioning will cause poor generalization of the respective machine learning algorithm.

Hence, if a number of objects required for successful training is known, a dataset generator (either for synthetic or non-synthetic data) should produce approximately that number of objects which would be as much as dissimilar from each other. Otherwise, an estimation of such a number may be given as an interval. Then the optimal number of objects for the corresponding dataset should be determined first.

Goal of the article and tasks to be fulfilled

The goal of the article is to develop an approach to forming an optimal dataset (either of real-world objects or synthetic ones) when an initial number of objects is significantly greater than required. For achieving the article's goal, the following four tasks are to be fulfilled:

1) to circumscribe initial conditions;

2) to state breaking an initial set of objects;

3) to determine the optimal number of objects for the corresponding dataset, if this number was given initially as an interval;

4) to state finding proper objects to be included into the optimal dataset.

Initial conditions

Initially, there are two cases:

1. Integer N is a known number of objects, at which the training process is presumably close to optimal. These objects are presumed to be the most "original" for being the most dissimilar from each other.

2. Integer N is unknown but $N \in \left\{\overline{N_{\min}, N_{\max}}\right\}$ by the known integers $N_{\min}$ and $N_{\max}$. The ratio of these integer margins can be any value.

Let Q be a total number of objects $\{\mathbf{X}_q\}_{q=1}^{Q}$. Object $\mathbf{X}_q = [x_{qk}]_{1\times F}$ has F features represented as a horizontal vector. If the object is an image, its matrix (two-dimensional for grayscale images and three-dimensional for color images) can always be reversibly reshaped to the respective vector.

The initial number of objects Q is significantly greater than N (case #1) or $N_{\max}$ (case #2) but only for the case of synthetic data (or an infinitely scalable dataset) they certainly represent all classes or categories. In some cases of non-synthetic data, these Q objects will not cover the whole number of categories. This is, for instance, when a few distinct objects have a lot of similar images made from slight displacements of the camera, whereas some other distinct objects of interest are represented normally. Then the dataset optimality is understood in the sense of filtering surplus images in a few classes, although the subsequent training on such an optimal dataset is not necessary to be successful.

Breaking the initial set of objects

When number N of proper objects to be included into the optimal dataset is known, the task is to break the initial set of objects $\{\mathbf{X}_q\}_{q=1}^{Q}$ into N groups (or clusters). Subsequently, the best object in each cluster will be selected. The criterion of the selection is going to be stated later.

In formal entries, breaking set $\{\mathbf{X}_q\}_{q=1}^{Q}$ into N clusters is equivalent to mapping set $\{\mathbf{X}_q\}_{q=1}^{Q}$ into a set

$$\left\{\left\{\mathbf{X}_i^{\langle j\rangle}\right\}_{j=1}^{m_i}\right\}_{i=1}^{N} \text{ by } \{\mathbf{X}_q\}_{q=1}^{Q} \cap \left\{\left\{\mathbf{X}_i^{\langle j\rangle}\right\}_{j=1}^{m_i}\right\}_{i=1}^{N} = \{\mathbf{X}_q\}_{q=1}^{Q} \text{ and } \{\mathbf{X}_q\}_{q=1}^{Q} \cup \left\{\left\{\mathbf{X}_i^{\langle j\rangle}\right\}_{j=1}^{m_i}\right\}_{i=1}^{N} = \{\mathbf{X}_q\}_{q=1}^{Q}, \tag{1}$$

where m_i objects $\left\{\mathbf{X}_i^{\langle j\rangle}\right\}_{j=1}^{m_i}$ constitute the i-th cluster, $i = \overline{1, N}$ and $\sum_{i=1}^{N} m_i = Q$. The mapping is written as

$$\left\langle \left\{\left\{\mathbf{X}_i^{\langle j\rangle}\right\}_{j=1}^{m_i}\right\}_{i=1}^{N}, \{\mathbf{C}_i\}_{i=1}^{N} \right\rangle = a\left(\{\mathbf{X}_q\}_{q=1}^{Q}, N, \rho_{\mathbb{R}^F}\right) \tag{2}$$

by a chosen distance $\rho_{\mathbb{R}^F}$ in $\mathbb{R}^F$ and a specific algorithm of clustering denoted by $a\left(\{\mathbf{X}_q\}_{q=1}^{Q}, N, \rho_{\mathbb{R}^F}\right)$ that returns also centroid $\mathbf{C}_i$ of the i-th cluster, $i = \overline{1, N}$. In general, centroid $\mathbf{C}_i = [c_{ik}]_{1\times F}$ is not one of objects $\left\{\mathbf{X}_i^{\langle j\rangle}\right\}_{j=1}^{m_i}$ of the i-th cluster [10, 11].

Determining the optimal number of objects for the corresponding dataset

In case #2, when number N of proper objects to be included into the optimal dataset is unknown but $N \in \left\{\overline{N_{\min}, N_{\max}}\right\}$, the optimal number of objects is determined by using the silhouette criterion [12]. This criterion returns a silhouette value $v(n)$ for n clusters:

$$v(n) = s\left(n, a\left(\{\mathbf{X}_q\}_{q=1}^{Q}, n, \rho_{\mathbb{R}^F}\right)\right) \text{ by } n = \overline{N_{\min}, N_{\max}} \tag{3}$$

where the right-side term is the silhouette criterion function. Then the optimal number of clusters is

$$N \in \arg \max_{n=\overline{N_{\min}, N_{\max}}} v(n) \tag{4}$$

that is the optimal number of objects to be included into the corresponding dataset. Along with number (4), mapping

(2) inside silhouette criterion function (3) gives centroids $\{\mathbf{C}_i\}_{i=1}^{N}$ and clustered objects $\left\{\mathbf{X}_i^{\langle j\rangle} = \left[x_{ik}^{\langle j\rangle}\right]_{1\times F}\right\}_{j=1}^{m_i}$.

Finding proper objects to be included into the optimal dataset

Obviously, centroids $\{\mathbf{C}_i\}_{i=1}^{N}$ cannot be taken as entries to the dataset, unless they coincide with objects. In each cluster, the object closest to the cluster centroid is the best one for including it into the dataset. The closeness should be treated by the same distance used in mapping (2) and silhouette criterion function (3). Thus,

$$\mathbf{X}_i^* \in \arg\min_{\mathbf{X}_i^{\langle j\rangle},\, j=\overline{1,\, m_i}} \rho_{\mathbb{R}^F}\left(\mathbf{X}_i^{\langle j\rangle}, \mathbf{C}_i\right) = \arg\min_{\mathbf{X}_i^{\langle j\rangle},\, j=\overline{1,\, m_i}} \left\|\mathbf{X}_i^{\langle j\rangle} - \mathbf{C}_i\right\| \tag{5}$$

is the representative of the i -th cluster. The distance can be commonly

$$\rho_{\mathbb{R}^F}\left(\mathbf{X}_i^{\langle j\rangle}, \mathbf{C}_i\right) = \sqrt{\sum_{k=1}^{F}\left(x_{ik}^{\langle j\rangle} - c_{ik}\right)^2} \tag{6}$$

or

$$\rho_{\mathbb{R}^F}\left(\mathbf{X}_i^{\langle j\rangle}, \mathbf{C}_i\right) = \sum_{k=1}^{F}\left|x_{ik}^{\langle j\rangle} - c_{ik}\right| \tag{7}$$

or other suitable for a specific task.

Discussion and conclusion

So, the optimal dataset consists of objects $\{\mathbf{X}_i^*\}_{i=1}^{N}$ found by (5). With the chosen distance, they are defined by centroids $\{\mathbf{C}_i\}_{i=1}^{N}$ which depend also on the algorithm of clustering. Determining the optimal number by (4) is a time-consuming process, especially if $N_{\min}$ and $N_{\max}$ are of order of thousands. Nonetheless, for relatively small datasets (required, for instance, for transfer learning tasks [4]) these integers are of order of hundreds. Then, as series of experiments prove, the silhouette criterion function by (3) performs much faster. If an initial number of objects is too great, it is reasonable to break them into a few groups. A subdataset will be formed from each group by using the same approach of clustering and selecting closest-to-the-centroid objects.

The proposed approach to forming the optimal dataset relies on an appropriately selected algorithm of clustering and a distance. In a wider sense, the approach allows to filter surplus objects from a dataset, thus optimizing it. The clustering consumes far more resources than selection of closest-to-the-centroid objects. However, an open question is when an initial group of objects should be broken for forming optimal subdatasets. In this way, the clustering can be clustered itself to accelerate optimization of a dataset.

References

1. Witten I. H., Frank E., Hall M. A., Pal C. J. Data Mining (Fourth Edition). Chapter 10 — Deep learning / Morgan Kaufmann, 2017. — P. 417 — 466.

2. Romanuke V. V. Appropriateness of DropOut layers and allocation of their 0.5 rates across convolutional neural networks for CIFAR-10, EEACL26, and NORB datasets // Applied Computer Systems. — 2017. — Vol. 22. — P. 54 — 63.

3. He H.-J., Zheng C., Sun D.-W. Computer Vision Technology for Food Quality Evaluation (Second Edition). Chapter 2 — Image Segmentation Techniques / Academic Press, 2016. — P. 45 — 63.

4. Bai S., Tang H. Softly combining an ensemble of classifiers learned from a single convolutional neural network for scene categorization // Applied Soft Computing. — 2018. — Vol. 67. — P. 183 — 196.

5. Romanuke V. V. Appropriate number and allocation of ReLUs in convolutional neural networks // Research Bulletin of NTUU “Kyiv Polytechnic Institute”. — 2017. — No. 1. — P. 69 — 78.

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.

8. Romanuke V. V. Two-layer perceptron for classifying flat scaled-turned-shifted objects by additional feature distortions in training // Journal of Uncertain Systems. — 2015. — Vol. 9, No. 4. — P. 286 — 305.

9. Romanuke V. V. Optimal training parameters and hidden layer neurons number of two-layer perceptron for generalized scaled objects classification problem // Information Technology and Management Science. — 2015. — Vol. 18. — P. 42 — 48.

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

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

12. 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.

Рецензія/Peer review : 9.10.2018 р. Надрукована/Printed :29.11.2018 р.
Рецензент: