УДК 004.93

VLADYSLAV KHOLIEV, postgraduate (Kharkiv National University of Radio Electronics),
OLESIA BARKOVSKA, PhD (Kharkiv National University of Radio Electronics)

Analysis of the of training and test data distribution for audio series classification

The effectiveness of machine learning algorithms for any given task largely depends on the training and test datasets. This manifests itself not only in the amount of data, but also in its content (that is, its relevance for the task at hand), as well as in its organization. Generally, the common approach is to split the dataset into training and testing sets to avoid model overfitting. In addition, to achieve better metrics for the selected criteria (accuracy, learning rate, etc.) of model performance, different ratios of training and test sets are used in the partitioning. The ***goal*** *of this paper is to analyze methods of data set partitioning for use in training neural networks and statistical models. One of the reviewed* ***methods****, specifically the cross-validation method, was applied to a dataset developed from the LibriSpeach corpus, an open English speech corpus based on the LirbiVox project of voluntarily contributed audio books. The* ***result*** *of applying the selected data partitioning method on the selected data set is demonstrated*
Keywords: datasets; pre-processing; machine learning; cross validation; librispeach; librivox.

Introduction

Despite the rapid spread of the Internet at the beginning of the 21st century and the predominantly textual nature of the information that circulated in it at the beginning of its development, a significant part of the information generated, transmitted and consumed by humanity was audiovisual in nature. This is due not only to the limitations of the Internet technology at the time, but also to the biological characteristics of humans as a species, since most of the information we receive from the environment is visual and sound information.

Over time, this trend has not only persisted, but also deepened with the development of technologies for generating, transmitting and storing information. In turn, information processing and analysis technologies have developed and continue to develop still. The degree of decision-making automation continues to grow with the use of deep learning technologies and statistical models.

In particular, as mentioned above, audio information plays one of the most widespread and important roles. Moreover, it has its advantages both in terms of data and technology. The advantages of audio information are as follows:

– independence from illumination, which allows it to serve as a spatial indicator where there is insufficient visual information, or to supplement the available visual information with additional context;

– the amount of data required to transmit the semantic load is smaller and requires cheaper equipment, which in turn means faster and more affordable analysis results.

Audio information is usually presented in the form of an analog signal and its digital encoding. Various encoding formats exist and are used, with their own advantages and disadvantages and, as a result, with their own areas of application (Table 1) [1, 2, 3, 4].

Actions performed on audio information are called audio analysis, or audio sequence analysis.

Format	Doesn't have compression		Has compression		
	WAV	AIFF	FLAC	AAC	MP3
Lossy	No	No	No	Yes	Yes
Year of development (latest release)	1991 (2007)	1988 (1991)	2001 (2022)	1997 (2019)	1993 (1998)

Table 1. Various formats for storing audio files

Audio analysis is generally referred to as the extraction of information from audio signals for further operations upon them. The widespread use of audio analysis can be explained by the wide range of its applications due to the high degree of reliance on sound and audio in a wide variety of spheres of life (online banking, virtual assistants in smartphones, PCs and other devices, user verification, automatic annotation of video conferences, tone analysis, and much more).

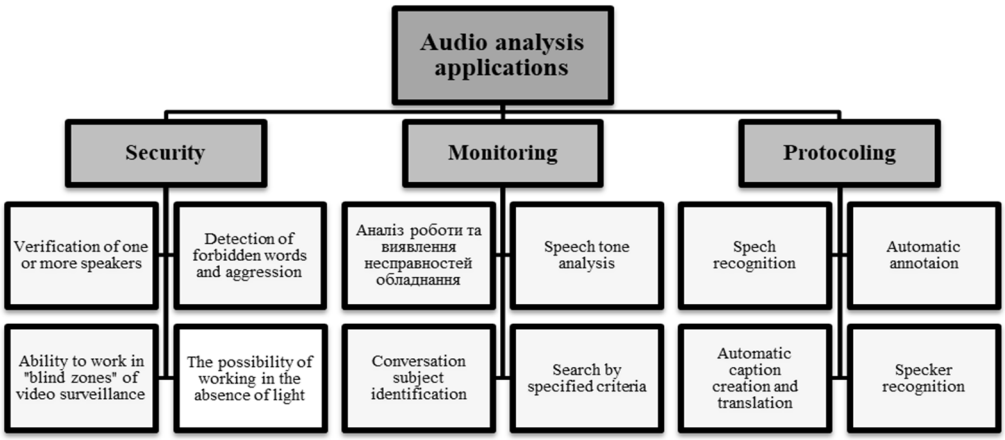

Fig. 1. Scopes of audio analysis applications**.**

Most of these tasks fall under the umbrella of the classification domain. Usually, classification consists of two stages: training a neural network model and its utilization. This means that this task requires training and test samples at the training and operation stages of the neural network, respectively.

Research task rationale

The goal of this paper is to analyze methods of data set partitioning for use in training neural networks and statistical models. To achieve this goal, the following tasks are to be accomplished:

– to review the methods, criteria and ratios of dataset partitioning to achieve the best selected performance metrics of neural network and statistical models.

– to apply one of the reviewed methods to produce a dataset that meets the requirements and criteria discussed above, based on the LibriSpeach corpus.

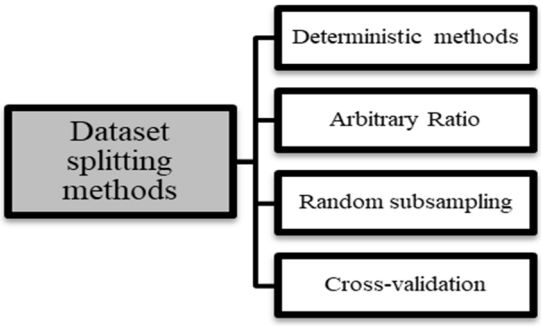

Fig. 2. Common dataset splitting methods

There exist many approaches to organizing a dataset, such as the classical random subsampling method, the cross-validation method [5], deterministic methods such as SPXY [6] or SPlit [7], as well as others.

These methods were developed to better meet general criteria, such as heterogeneity and/or balance of data on selected features, as well as to introduce new criteria that they considered key to achieving the best model performance according to selected criteria (accuracy, learning rate, etc.).

Therefore, the task of analyzing the distribution of training and test data for audio series classification is a relevant task, since the quality of classification depends on the proper distribution of sets.

The importance of dataset organization

The effectiveness of machine learning algorithms for any given task largely depends on the training and test datasets. This manifests itself not only in the amount of data, but also in its content (that is, its relevance for the task at hand), as well as in its organization. There are several stages of data preparation for samples that are applicable to most cases:

– problem formulation;
– data collection of the selected subject area;
– data normalization and formatting;
– data segmentation.

Depending on the approach, the data can be divided into training, validation, and test sets, where the validation set is used to optimize the model's hyperparameters to

achieve the best accuracy, which will be evaluated using the test set.

Another approach suggests that after separating the test set, the remaining set is divided into k parts, where during model training, k-1 of them are used as a training set, and the last one is used as a validation set, after which another part is selected to serve in its place. This is repeated k times, after which the results are averaged. This approach is known as cross-validation [5]. It is generally believed that this method shows better results with a fairly small amount of data due to the fact that the entire dataset is covered for usage as both training and validation sets, while it is less effective with medium and large datasets

The ratio of sets during the distribution is one of the parameters the exact value of which depends on the task and the nature of the data, however, in practice a certain initial value is used to start from when searching for the exact value. Previous studies have reached no consensus on which values are optimal [8]. One of the popular distribution options is 80/20 (training and validation + test sets, respectively), the sentiment of using which originates from the Pareto principle.

Alternatively, the distribution value is proposed to be obtained taking into account the characteristics of the model, as described in [8]. This study suggests using the following formula for the distribution:

$$\gamma = \frac{1}{\sqrt{p}+1}, \quad (1)$$

where γ stands for the ratio of the test set to the dataset as a whole and p stands for the number of model's parameters.

The diagram of this dependence is shown in Figure 2. As evident, the portion of the training set increases significantly with a larger number of parameters, which is logical, since a model with a larger number of parameters requires more training data to approximate all of them.

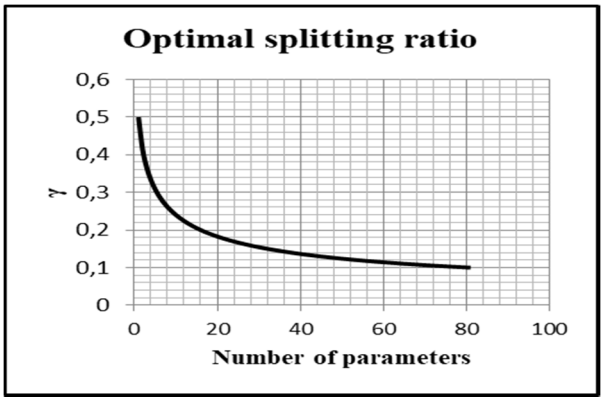

Fig. 2. Diagram of the dataset distribution based on the number of parameters

Librispeach corpus and dataset

As a dataset, an independently developed set on the basis of the LibriSpeech corpus was chosen, which, in turn, was created on the basis of the LibriVox audiobook project [9], which is in the public domain, or more specifically, its English segments.

The corpus is divided into several parts, available separately: a test set in a single archive, and a training set in three archives - 100-, 300-, and 500-hour archives. In total, the corpus contains 982 hours of recordings from 2338 speakers.

The characteristics of the corpus are presented in Table 2, which has the following columns:

– subset: name of the subset;

– hours: total duration of particular subset in hours;

– per-spk minutes: longest cumulative duration of recordings per speaker in particular subset minutes;

– female spkrs: number of female speakers in particular subset;

– male spkrs: number of male speakers in particular subset;

– total spkrs: total number of speakers in subset.

subset	hours	per-spk minutes	female spkrs	male spkrs	total spkrs
dev-clean	5.4	8	20	20	40
test-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-other	5.1	10	17	16	33
train-clean-100	100.6	25	125	126	251
train-clean-360	363.6	25	439	482	921
train-other-500	496.7	30	564	602	1166

Table 2. Data subsets in LibriSpeech

The authors of the corpus have created several annotation files: matching speakers and their characteristics with the subsets to which they are assigned, matching speakers and chapters of books whose records were taken for the corpus, and matching the books themselves with their identifiers used in other annotation files.

```
ID   |SEX| SUBSET           |MINUTES| NAME
14   | F | train-clean-360  | 25.03 | Kristin LeMoine
16   | F | train-clean-360  | 25.11 | Alys AtteWater
17   | M | train-clean-360  | 25.04 | Gord Mackenzie
20   | F | train-other-500  | 30.07 | Gesine
23   | F | train-clean-360  | 25.23 | Anita Roy Dobbs
25   | M | train-other-500  | 30.16 | John Gonzalez
26   | M | train-clean-100  | 25.08 | Denny Sayers
27   | M | train-clean-100  | 20.14 | Sean McKinley
28   | F | train-clean-360  | 25.03 | Kristin Hughes
29   | M | train-other-500  | 30.10 | Linton
31   | M | train-other-500  | 23.79 | Martin Clifton
32   | F | train-clean-100  | 24.01 | Betsie Bush
36   | M | train-other-500  | 25.85 | Chip
...
8824 | M | train-clean-360  | 25.21 | Mark Johnston
8825 | F | train-clean-360  | 23.93 | Erin Schellhase
8838 | M | train-clean-100  | 25.06 | Kevin Owens
8855 | M | train-clean-360  | 25.01 | Eric Metzler
8975 | F | train-clean-100  | 25.11 | Daisy Flaim
9022 | F | train-clean-360  | 25.17 | Claire M
9023 | F | train-clean-360  | 25.19 | P. J. Morgan
9026 | F | train-clean-360  | 21.75 | Tammy Porter
```

Listing 1. A fragment of one of the annotation files

From these parts, subsets of 100 and 300 hours were taken, totaling 464.2 hours and 1172 speakers, which is approximately half of the total volume.

Initially, the corpus was created for the task of speech recognition, as well as identification and/or classification of certain characteristics of speakers (age, gender, etc.). To create the corpus, two stages of alignment were first performed using a variety of tools and speech models [9], which were aimed at dividing the recordings into fragments and removing recordings that contained discrepancies with the text due to human error (inclusions, substitutions, deletions, and permutations). These alignment steps resulted in approximately 1200 hours of recordings up to 35 seconds long, after which the data was segmented into smaller fragments based on pauses of at least 0.3 seconds. The test data was segmented similarly to the training data, but with the additional condition of splitting fragments only at sentence boundaries to better model language usage.

Thanks to additional pre- and post-processing stages, the balance of speakers' genders was ensured, and cases of recordings with multiple speakers were eliminated (by filtering individual recordings as well as certain genres that by their nature land themselves to multi-speaker recordings).

Results and discussion

The fact that the corpus was created for the task of speech and/or speaker characteristics recognition means that it is not presented in the proper form for the method, and therefore the dataset was adapted for the current task. Since, as mentioned above, the corpus consists of voluntarily provided records by LibriVox users, not all of them were signed with identifiable names (Listing 2). Such recordings were discarded, as well as recordings with a total duration of less than 20 minutes, as this is the duration of most recordings in the dataset used (namely, the 100 and 300 hour subsets), so this cutoff threshold makes the most sense for preserving the majority of the data.

```
ID   |SEX| SUBSET           |MINUTES| NAME
249  | M | train-clean-360  | 18.69 | pww214
272  | M | train-clean-360  | 16.45 | Mr. Baby Man
288  | F | train-clean-360  | 25.13 | Bookworm
318  | F | train-clean-360  | 25.17 | Eileen aka e
1634 | M | train-clean-360  | 17.65 | daxm
2397 | M | train-clean-360  | 25.14 | texttalker
2404 | M | train-clean-360  | 25.21 | n8evv
4267 | M | train-clean-100  | 25.14 | Ric F
8396 | M | train-clean-360  | 25.16 | gloriousjob
```

Listing 2. Examples of improperly signed records that were discarded

In addition, the recordings in the corpus are stored in the FLAC (Free Lossless Audio Codec) format. This is a codec designed to compress audio without loss [1]. Although this format has advantages for storing a large number of audio files, its use would lead to significant additional costs for restoring each record from compression, so each of the records of the filtered dataset was converted to WAV (Waveform Audio File Format), which does not require additional operations to access the audio signal and is well suited for storing uncompressed audio in pulse-code modulation [2].

According to the cross-validation method, the dataset was divided into 5 parts to ensure the most equal division. The records were randomly selected for the multiclassifier task to simulate the cases of records that are not part of any of the classes (unauthorized access attempt).

After all the operations performed – filtering in several stages, converting the file format, and splitting into parts according to the selected cross-validation method -– the characteristics of the resulting dataset are as follows:

– 859 speakers (consisting of 437 males and 442 females);

– 99955 audio files (an average of 116 recordings per speaker);

– size: 21,5 Gigabytes (23 177 338 377 bytes);

– duration: 1271393.26 seconds (353 hours, 9 minutes, and 53.26 seconds);

– divided into 5 parts according to the cross-validation method.

```
fold1
| 274-121382-0000.wav
| 200-126784-0009.wav
| 335-125951-0004.wav
| 4813-248641-0000.wav
| 1958-144503-0061.wav
| ...
| 7704-106969-0010.wa

fold2
| 1313-136054-0010.wav
| 8008-271817-0039.wav
| 2764-36616-0008.wav
| 5093-39749-0016.wav
| 5126-34483-0026.wav
| ...
| 1743-142914-0034.wav

fold3
| 2427-154736-0016.wav
...
```

Listing 3. Fragment of the resulting annotation file

Conclusion

The goal of this paper was to analyze methods of data set partitioning for use in training neural networks and statistical models.. To achieve this goal, the following tasks were accomplished:

– methods and ratios of dataset partitioning to achieve the best selected performance metrics of neural network and statistical models were analyzed;

– one of the analyzed methods, namely the cross-validation method, was applied to the given dataset, which was developed on the basis of the LibriSpeach open corpus;

– described the process of developing the dataset.

Further research includes: implementing the dataset in the workflow of an intelligent user verification system, studying the feasibility of modifying the proposed and developed dataset for use in an ensemble of neural networks.

References

1. Coalson J., “FLAC – What is FLAC”, available at: https://xiph.org/flac/ (last accessed 08.12.2022).
2. “RFC 2361: WAVE and AVI codec registers”, available at: https://www.rfc-editor.org/rfc/rfc2361 (last accessed 08.12.2022).
3. Kabal P., “Audio File Format Specifications - AIFF / AIFF-C Specification”, available at: https://www.mmsp.ece.mcgill.ca/Documents/AudioFormats/AIFF/AIFF.html (last accessed 08.12.2022).
4. “MP3 and AAC Explained (archived from the original)”, available at: https://web.archive.org/web/20170213191747/https://graphics.ethz.ch/teaching/mmcom12/slides/mp3_and_aac_brandenburg.pdf (last accessed 08.12.2022).
5. Stone, M (1974). "Cross-Validatory Choice and Assessment of Statistical Predictions". Journal of the Royal Statistical Society, Series B (Methodological). 36 (2): 111–147. doi:10.1111/j.2517-6161.1974.tb00994.x.
6. R. K. H. Galvão, M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva, and T. C. B. Saldanha, A method for calibration and validation subset partitioning, Talanta 67 (2005), no. 4, 736–740.
7. V. Roshan Joseph & Akhil Vakayil (2022) SPlit: An Optimal Method for Data Splitting, Technometrics, 64:2, 166-176, DOI: 10.1080/00401706.2021.1921037
8. Joseph, V. R., Optimal ratio for data splitting, Stat. Anal. Data Min.: ASA Data Sci. J.. 15 (2022), 531–538. https://doi.org/10.1002/sam.11583.
9. V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

В. О. Холєв, О. Ю. Барковська. Аналіз розподілу навчальних та тестових даних для класифікації аудіоряду

Анотація. Ефективність виконання алгоритмами машинного навчання поставленої задачі значною мірою залежить від навчальних та тестових вибірок. Це проявляється не тільки в об'ємі даних, але й в їх змісті (тобто актуальності для поставленої задачі), а також в їх організації. Загалом набор даних прийнято розбивати на навчальну та тестову вибірки для уникнення перенавчання. Окрім того, для досягнення кращих показників (точності, швидкості навчання тощо) продуктивності моделі застосовують різний показник відношення навчальної та тестової вибірок при розбитті. **Метою** даної роботи є розглянути **методи** розбиття наборів даних для використання у навчанні нейронних мереж та статистичних моделей. Один з розглянутих методів, а саме метод перехресного затвердження, був застосований до набору даних, що був підготовлений на основі корпусу LibriSpeach – відкритого корпусу англійського мовлення, заснованого на проекті добровільно наданих аудіо книг LirbiVox. Продемонстрований **результат** застосування обраного методу розбиття даних на обраному наборі даних.

Ключові слова: дата сет, набір даних, попередня обробка, машинне навчання, крос-валідація, librispeach, librivox.

Надійшла 19.02.2023 р.

Холєв Владислав Олександрович *– аспірант кафедри "Електронно обчислювальних машин", науковий керівник – Барковська Олеся Юріївна – кандидатка технічних наук, доцентка кафедри "Електронно обчислювальних машин", Національний університет радіоелектроніки «ХНУРЕ», Харків, Україна*

Барковська Олеся Ю'ївна *– кандидатка технічних наук, доцентка кафедри "Електронно обчислювальних машин", Національний університет радіоелектроніки «ХНУРЕ», Харків, Україна*

Vladyslav Kholiev *– postgraduate of Department of Electronic Computers, doctoral supervisor – Olesia Barkovska – PhD, Associate Professor of Department of Electronic Computers, Kharkiv National University of Radio Electronics "NURE", Kharkiv, Ukraine. E-mail: vladyslav.kholiev@nure.ua, ORCID ID https://orcid.org/0000-0002-9148-1561.*

Olesia Barkovska *– PhD, Associate Professor of Department of Electronic Computers, Kharkiv National University of Radio Electronics "NURE", Kharkiv, Ukraine. E-mail: olesia.barkovska@nure.ua, ORCID ID https://orcid.org/0000-0001-7496-4353*