UDC 0048:681.3 DOI: 10.26661/2413-6549-2019-2-08

DESIGN SOUND CLASSIFICATION IOT SYSTEM WITH GENETIC ALGORITHMS

A. G. Kryvokhata, O. V. Kudin, V. I. Gorbenko

Zaporizhzhia National University
avk256@gmail.com

Key words:
Acoustic Data Classification, Convolutional Neural Network, Ensemble Learning, IoT.

This paper discusses the models and methods of machine learning used in IoT (Internet of Things) systems. Particularly some issues of methods development for sound data classifying of various nature, such as speech, music, environmental sounds etc. A sound classification subsystem could be implemented in the various Smart City, Smart Farming systems etc. Development of IoT software includes following challenges: the lack of computational resources, the relevant small RAM memory size etc. Basically, an automated sound classification system could be decomposed into four parts: the audio representation, the features extraction, the machine learning algorithm and the accuracy estimation. This paper deals with a machine learning algorithms. We use a convolutional neural network for sound classification and the Snapshot method for ensemble learning. A genetic algorithm is a typical strategy for both neural network and ensembles structural optimization. Various methods of the solution representation and crossover functions have been studied in order to find optimal configuration of genetic operators. The objective of the paper is to develop the optimal classifier for embedded sound classification system. The solution representation – genotype – is the set of neural network hyper-parameters includes the number and the type of neural network layers, the type of activation functions, the initial values of weights, the learning rate, etc. Both the Snapshot ensemble method and combination of different neural networks are used for ensemble learning. The key idea of this paper is the optimization with genetic algorithms both neural networks and the ensemble construction method. We compare different genetic operators in order to obtain optimal solution for IoT system.

ВИКОРИСТАННЯ ГЕНЕТИЧНИХ АЛГОРИТМІВ ПРИ РОЗРОБЦІ IOT СИСТЕМ КЛАСИФІКАЦІЇ ЗВУКУ

А. Г. Кривохата, О. В. Кудін, В. І. Горбенко

Запорізький національний університет
avk256@gmail.com

Ключові слова:
Класифікація звуку, згорткова нейронна мережа, ансамблеве навчання, інтернет речей.

У роботі розглядаються моделі та методи машинного навчання, що застосовуються в системах IoT (Internet of Things), зокрема для вирішення проблеми класифікації акустичних даних різного походження, таких як мова, музика, звуки природи тощо. Підсистеми класифікації звуку можуть бути впроваджені у різних системах інтернету речей, наприклад, Smart City, Smart Farming тощо. Розробка програмного забезпечення IoT включає такі виклики: обмеженість обчислювальних ресурсів, відносно невеликий об’єм оперативної пам’яті тощо. Отже, існує потреба у розробці оптимального класифікатора з найвищою здатністю до узагальнення. Здебільшого автоматизована система класифікації звуку може бути розбита на чотири частини: звукове подання, функції вилучення ознак, алгоритм машинного навчання та оцінка точності. Основна увага в цій роботі приділяється алгоритмам машинного навчання. Використовуються згорткові нейронні мережі для класифікації звуку та порівнюються декілька підходів до ансамблевого навчання. Генетичні алгоритми є типовою стратегією як для структурної оптимізації нейронних мереж, так і для їх ансамблів. Для побудови оптимальної конфігурації генетичних операторів розглянуто різні методи представлення

розв'язку та кросовер-функції. Метою роботи є розробка оптимального класифікатора звуку для вбудованих систем. Представлення розв'язку – генотип – це сукупність гіперпараметрів нейронної мережі, що включає кількість та тип шарів нейронної мережі, тип функцій активації, початкові значення ваг, швидкість навчання тощо. Метод ансамблю Snapshot і комбінація різних нейронних мереж використовуються для ансамблевого навчання. Ключова ідея даної роботи – оптимізація засобами генетичних алгоритмів як нейронних мереж, так і методу побудови ансамблю. У роботі порівнюються різні генетичні оператори кросовера з метою отримання оптимальної конфігурації IoT системи.

1. Introduction

The Internet of Things (IoT) and Machine Learning (ML) are a promising technologies for automation in the different domains e.g. Smart Cities, Smart Farming etc. Such systems could be used for noise emission determining, criminal activity detecting, animal classification, livestock tracking and so on. There are various methodologies of IoT and ML implementation to the practice use cases [1-3]. In this paper we focus on a machine hearing system. The system allows to classify natural sounds in the Smart Farming software application. Manual processing of the sound data is complicated; therefore, our goal is to *develop automatic machine hearing systems, particularly, sound classification software*. Such systems have controversial requirements such as the real time classification, the high accuracy and compatibility with single-board computers, e.g. Raspberry Pi, Odroid etc. Therefore, the software for IoT and ML applications should be sufficiently efficient and not demanding on large computing resources.

2. The aim and objectives of the study

Basically, automated audio detecting and classification systems could be roughly decomposed into four parts: audio representation, features extraction, machine learning algorithm, and accuracy estimation. The audio representation stage implies that a raw signal is subject to segmentation into shorter signal chunks by some windowing process. Typically, at this stage, the original acoustic signal is converted into the frames of a fixed length. The aim of the feature extraction stage is to receive a compact representation of the acoustic characteristics of a signal. This stage exploits special coefficients such as the Zero-crossing rate, the Spectrum shape, the Short-Time Fourier Transform and Mel-frequency cepstral coefficients. Audio classification traditionally involves such machine learning methods like K-means, support vector machine (SVM), decision trees etc [4, 5]. During the last two decades the deep learning based methods have gained popularity for audio tagging also. The methods based on convolutional neural networks or recurrent neural networks should be referenced in this context. Deep neural networks could be used on both raw acoustic signal and features extracted from it. The accuracy estimation stage deploys quality assessment methods [5].

The aim of this paper is to develop a proof-of-concept system for the classification of acoustic data on the basis of convolutional neural networks and to optimize its hyper-parameters using the genetic algorithm. The system classifies acoustic data such as "vehicle noise", "human speech", "siren", "dog bark", "engine sound" and could be used in Smart Farming applications as a part of sound event detection system.

In order to reach the mentioned aim, the following objectives were formulated for the study:

– to review state-of-the-art systems in the Smart Farming;

– to develop classification system using convolutional neural networks;

– to develop implementation of snapshot ensemble method for accuracy improving;

– to outline a direction for further development of machine hearing systems in Smart Farming.

3. Motivation and state of the art

Sound event detection and classification systems become a significant part of modern Smart Farming applications. Such software could be used to distinguish between species of insects or to detect some vermin animals.

For example, there are papers dedicated to identification bee's state by producing sounds in the beehive [6-10] or sound analysis in the livestock farming [11]. SVM and Convolutional neural networks are used in [6-10] for sound classification. There are the multi-sensor platform includes sensors for temperature, humidity, weight, CO2 and the microphone. The system is deployed with hardware including the Raspberry Pi, the sound card and sensor shields [10].

Generally, the number of articles in the Smart Farming field has been increasing in the last decade [12, 13]. *This is due* to the development of machine learning methods and technology of single-board computers such as Raspberry Pi, Odroid etc. For example, the newest revision of Raspberry Pi 4 includes 4 GB RAM and CPU 1,5 GHz. This hardware allows neural networks deploying on a single-board computer and real-time calculations performing.

4. Literature review

Among the large number of review papers on the subject of the machine hearing systems development, there are several that are the most comprehensive. Thus, review articles [14-16] provide a description of the components of an automatic sound classification system, which contains pre-processing modules, feature extraction, training algorithm and calculation module.

In [14], approaches to signal feature extraction are discussed in details. There are methods based on the physical properties of the signals and the characteristics of the human perception of sounds. Feature extraction methods represent the acoustic signal in the time, frequency, cepstral and wavelet domains.

Reviews [17, 18] provide an analysis of general approaches and publications for the automatic classification of music by genre. The majority of the most informative tags that could be used as classes in training classifiers are discussed. The most commonly used sources of labeled acoustic data that could be included in training systems are discussed. Typically, these are open music databases on the Internet that are recorded by users, from social network, and data that is generated specifically for the purpose of performing machine hearing tasks. In [18], the issue of assessing the effectiveness of genre music classification systems is discussed.

In recent years, more and more works have been devoted to the use of neural networks both in the feature extraction and the classification [16, 19-22]. Convolutional neural networks allow obtain both pre-processed and raw acoustic data sets. The effectiveness of this approach is explained by the layered architecture of convolutional neural networks. There are several types of layers: convolution layers that distinguish a certain type of feature, pooling layers that reduce dimension, and several fully dense layers in which classification is performed [23]. The disadvantages of this approach include the complexity of setting up neural networks with complex architecture and the demand for computing resources.

Ensemble training involves combining several models, such as classifiers, into a common model, followed by alignment of the results of all models by some algorithm. Studies show that the efficiency of the ensemble is usually higher than the efficiency of individual models [24]. For this approach the lack of correlation between the models of the ensemble is a mandatory requirement.

The combination of different types of classifiers using (e.g. decision trees, SVMs, Bayes classifiers, etc.) and a common prediction forming by simple voting, a mean calculating or special machine learning algorithm is one of the approaches to building ensembles. An alternative approach is to train the same types of classifiers on different subsets of training data, with further averaging of the outcome of the forecast. This approach named Bagging (bootstrap aggregating). Another class of ensemble teaching methods is Boosting. The consistent using of such class methods as AdaBoost and gradient boosting for train the classifiers reinforces the result [24].

Recently, the use of ensembles of deep neural networks has led to the development in the practical application of machine learning [17, 18]. But despite its high accuracy, ensemble training of neural networks is not as widely used as ensembles of more classical machine learning methods. This is due to the high demand for time and space resources.

Most of the works devoted to the use of an ensemble of neural networks are aimed to study the methods for generating a common result from the results of trained classifiers.

In [17] it has been proposed to train one network instead of training M neural networks. During applying the stochastic gradient descent method the basic idea is to store the values of the weight matrix in case the hitting of M local points. Thereafter, a corresponding neural network is generated for each of the M weight matrices. Thus, the learning time of the ensemble is almost the same to the learning time of a single neural network.

5. Methodology

The workflow of developing an automated sound classification has shown in the Fig. 1.

In practice, the direct sound analysis in the time domain (amplitude-time dependence) is almost not used because it is not efficient enough and requires additional time and space resources. For the most rational presentation of an acoustic signal, classical methods of digital signal processing are used. These methods include transformations that decompose a signal by orthogonal basis functions: the Fourier transform, Hartley, Mellin, wavelets etc., as well as various signal attributes that are calculated on the basis of these transformations, for example, mel-cepstral coefficients, centroids, signal energy etc. [5]. The mel, bar etc. are the frequently used units that to relate for the psychophysical features of human perception of frequency and volume. For example, mel is a unit of subjective sound frequency to perceive by humans.

The data preprocessing stage includes calculating of mel frequency cepstral coefficient (MFCC) for the giving sound files. This approach allows to unify and simplify the sound files presentation in the memory. Further we feed MFCC arrays to the convolutional neural net. It is important at this stage to configure the network optimally for the most compact storage. This is due to the need to use platforms such as Raspberry Pi to deploying neural network.

To test the proposed approaches, we use data from www.kaggle.com, namely the Urban Sound Classification dataset. The dataset contains 3449 wav audio files for training and system testing. The training sample contains sound files related to 9 categories and these are sounds like traffic noise, car sirens and human speech etc. The minimum number of files in one category is 94, the maximum is 300. The duration of audio files is mostly 4 seconds.

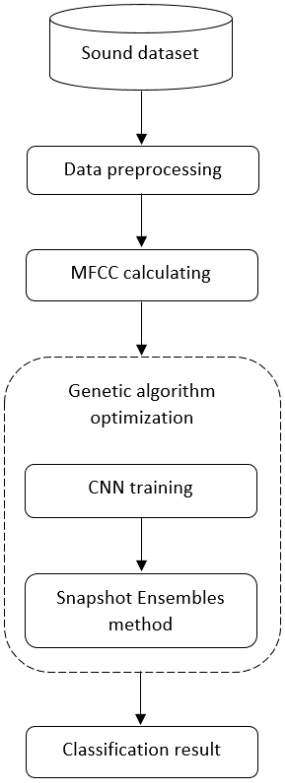

Figure 1. Sound classification system workflow

Thus, after training, the system could be deploy on the single-board computer and could be use autonomously as part of Smart Farming systems.

6. Numerical Result

The Keras library is used for a neural network software implementation in Python. The Librosa library is employed For data preprocessing. The source code for sound data preparation is shown in the Fig. 2.

```
def PrepareData(self, df, config):
    X = np.empty(shape=(df.shape[0], config.dim[0], config.dim[1], 1))
    input_length = config.audio_length
    for i, fname in enumerate(df):
      print(fname)
      file_path = self.TRAIN_DIR + '/' + str(int(fname)) + ".wav"
      data, _ = librosa.core.load(file_path, sr=config.sampling_rate,
res_type="kaiser_fast")
      data = librosa.feature.mfcc(data, sr=config.sampling_rate,
n_mfcc=config.n_mfcc)
      data = np.apply_along_axis(np.mean,1,data)
      data = data.reshape(config.dim[0],config.dim[1])
      data = np.expand_dims(data, axis=-1)
      X[i,] = data
    return X
```

Figure 2. MFCC calculation method

As mentioned above, an important step is to do optimal configuration the hyper-parameters of the neural network. A generic algorithm is a

general method here. In this work we are going to determine genetic operators as follows. We use different neural network and Snapshot Ensemble hyper-parameters such as number of dense layers, number of convolutional layers, number of units in each layer, type of pooling layer, type of activation function, number of networks in the Snapshot Ensemble.

We use binary chromosome encoding for representation of possible solution. Crossover operators that we use are the following: single point, two-point, uniform.

The results of computational experiments with different crossover operators are given below in the table 1. We suppose that the best snapshot is used after Snapshot Ensemble method. We also compare different approaches to snapshots averaging. The results are shown in the table 2.

Table 1. Accuracy with different crossover operators

Crossover operator	Accuracy
single point crossover	0,954
two-point crossover	0,942
uniform crossover	0,877

Table 2. Snapshot averaging method

Method	Accuracy
best snapshot	0,954
Average	0,940

Thus, we can see that best configuration of the system is obtained using single point crossover and best snapshot averaging method.

7. Conclusion and discussion

A convolutional neural network ensemble is built to solve the problem of acoustic data classification. The accuracy of the model on the test data set is 95%. Optimal network structure includes two packets of convolution-activation-sub-sampling layers.

The system can be used in Smart Farming or Smart City applications to filter out unnecessary sounds or as part of the sound detection system.

The basic distinction of this paper from papers [25] is that we are studying influence of different parameters of genetic algorithms like type of crossover operator.

The prospect of further research is related to the extension of the considered approach to sound data of a more complex structure.

References

1. Bagri, M. & Aggarwal, N. (2019). Machine Learning for Internet of Things. International Journal Of Engineering And Computer Science, Vol. 8, Issue 7, pp. 24680–24782. doi: 10.18535/ijecs/v8i07.4346
2. More, S. & Singla, J. (2019). Machine Learning Techniques with IoT in Agriculture. International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, Issue 3, pp. 742–747. doi: https://doi.org/10.30534/ijatcse/2019/63832019
3. Piccialli, F., Cuomo, S., di Cola, V.S. & Casolla, G. (2019). A machine learning approach for IoT cultural data. Journal of Ambient Intelligence and Humanized Computing. doi: https://doi.org/10.1007/s12652-019-01452-6
4. Kryvokhata, A. G., Kudin, O. V. & Lisnyak, A. O. (2018). A Survey of Machine Learning Methods for Acoustic Data Classification. Visnyk of Kherson National Technical University, Vol 3, Issue 66, pp. 327–331 (in Ukrainian).
5. Camastra, F. & Vinciarelli, A. (2015). Machime learning for Audio. Image and Video analysis. London: Springer-Verlag.
6. Cecchi, S., Terenzi, A., Orcioni, S., Riolo, P., Ruschioni, S. & Isidoro N. (2018). A Preliminary Study of Sounds Emitted by Honey Bees in a Beehive. 144th AES convention. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=19498
7. Nolasco, I. Terenzi, A., Cecchi, S., Orcioni, S., Bear, H. L. & Benetos E. (2018). Audio-based identification of beehive states. Retrieved from https://arxiv.org/abs/1811.06330.
8. Nolasco, I. & Benetos E. (2018). To bee or not to bee: investigating machine learning approaches for beehive sound recognition. Retrieved from https://arxiv.org/abs/1811.06016.
9. Cejrowski, T., Szymaski, J., Mora, H. & Gil D. (2018) Detection of the Bee Queen Presence using Sound Analysis. In Intelligent Information and Database Systems. ACIIDS. Lecture Notes in Computer Science, Vol. 10752. Springer.

10. Cecchi, S., Terenzi, A., Orcioni, S., Spinsante, S., Primiani, V. M., Moglie, F., Ruschioni, S., Mattei, C., Riolo, P. & Isidoro, N. (2019). Multi-sensor platform for real time measurements of honey bee hive parameters. IOP Conf. Series: Earth and Environmental Science, Vol. 275. doi: https://doi.org/10.1088/1755-1315/275/1/012016.
11. Bishop, J. C., Falzon, G., Trotter, M., Kwan, P. & Meek, P. D. (2017). Sound analysis and detection, and the potential for precision livestock farming - a sheep vocalization case study, 1st Asian-Australasian Conference on Precision Pastures and Livestock Farming. doi: https://doi.org/10.5281/zenodo.897209.
12. Wolfert, S. Ge, L., Verdouw C. & Bogaardt, M.-J. (2017). Big Data in Smart Farming – A review. Agricultural Systems, Vol. 153, pp. 69–80. doi: https://doi.org/10.1016/j.agsy.2017.01.023.
13. Hallett, S. H. (2017). Smart cities need smart farms. Environmental Scientist, Vol. 26, Issue 1, pp. 10–17. Retrieved from https://www.the-ies.org/resources/feeding-nine-billion
14. Alías, F., Socoró, J.C. & Sevillano, X. (2016). A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Applied Sciences, Vol. 6(5):143. doi: https://doi.org/10.3390/app6050143
15. Bertin-Mahieux, T., Eck, D. & Mandel, M. (2011). Automatic tagging of audio: the state-of-the-art. Machine audition: principles, algorithms and systems. IGI Global, pp. 334–352. doi: https://doi.org/10.4018/978-1-61520-919-4.ch014.
16. Salamon, J., Jacoby, C. & Bello, J. P. (2017). A dataset and taxonomy for urban sound research. Proceedings of the 22nd ACM international conference on Multimedia, pp. 1041–1044. doi: https://doi.org/10.1145/2647868.2655045
17. Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E. & Weinberger, K. Q. (2017). Snapshot Ensembles: Train 1, Get M for Free. Retrieved from http:// arxiv.org/abs/1704.00109
18. Stastný, J., Skorpil, V. & Fejfar, J. (2013). Audio Data Classification by Means of New Algorithms. 36th International conference on Telecommunications and Signal Processing 2013, Rome, Italy, pp. 507–511. doi: https://doi.org/10.1109/TSP.2013.6613984.
19. Xu, Y., Kong, Q., Huang, Q., Wang, W. & Plumbley, M. D. (2017). Convolutional gated recurrent neural network incorporating spatial features for audio tagging. The 2017 International Joint Conference on Neural Networks (IJCNN 2017), Anchorage, Alaska. doi: https://doi.org/10.1109/IJCNN.2017.7966291.
20. Rizzi, A., Buccino, M., Panella, M. & Uncini, A. (2006). Optimal short-time features for music/speech classification of compressed audio data. International Conference on Intelligent Agents. Sydney, NSW, Australia. doi: https://doi.org/10.1109/CIMCA.2006.160
21. Sturm, B. L. (2014). A Survey of Evaluation in Music Genre Recognition. Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation. AMR 2012. Lecture Notes in Computer Science, Vol. 8382, pp. 29–66. doi: https://doi.org/10.1007/978-3-319-12093-5_2.
22. Xu, Y., Huang, Q., Wang, W., Foster, P., Sigtia, S., Jackson, P. J. B. & Plumbley, M. D. (2017). Unsupervised Feature Learning Based on Deep Models for Environmental Audio Tagging. IEEE/ACM transactions on audio, speech and language processing, Vol. 25, Issue 6, pp. 1230–1241. doi: https://doi.org/10.1109/TASLP.2017.2690563
23. Zaccone, G. & Karim, Md. R. (2018). Deep learning with TensorFlow.Packt Publishing.
24. Geron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. Sebastopol: O`Reilly.
25. Gonzalez, J. A., Hurtado, L.-F. & Pla, F. (2019). ELiRF-UPV at SemEval-2019 Task 3: Snapshot Ensemble of Hierarchical Convolutional Neural Networks for Contextual Emotion Detection. Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). pp. 195–199. doi: https://doi.org/10.18653/v1/S19-2031.