Azerbaijan Journal of High Performance Computing, Vol. 2, Issue 2, 2019, pp. 170-177
https://doi.org/10.32010/26166127.2019.2.2.170.177

Parallel Solution of Features Subset Selection Process for Hand-Printed Character Recognition

Elviz Ismayilov, Rahman Mammadov

Azerbaijan State Oil and Industry University, Baku, Azerbaijan, elviz.ismayilov@gmail.com, mammadli.kenta@gmail.com

*Correspondence:
Elviz Ismayilov
Azerbaijan State Oil and Industry University, Baku, Azerbaijan, elviz.ismayilov@gmail.com,

Abstract

The existence of a huge amount of features for pattern recognition problems brings to the overloading of the training and exploitation steps of the recognition; also, highly correlated features affect the accuracy of the designed systems negatively. One of the most used ways for tackling this problem is the application of genetic algorithms for the solution of the binary optimization problems that appeared during the features subset selection process. In this paper was used parallel genetic algorithms for the selection of the most informative features in Azerbaijani hand-printed character recognition system by using opportunities of the distributed cluster computing. In this way after the given number of generations most appropriate features with the high recognition rate were selected from the features database.

Keyword: feature selection; genetic algorithms; crossover methods; cluster computing; distributed systems

1. Introduction

Recent innovative theoretical and practical approaches proposed for recognition of hand-printed characters, handwritten texts have been successfully applied for the solution of recognition systems for texts in different local languages, ancient alphabets, artifacts, license plate detection (Barbuti & Caldarola, 2018; Kowsalya & Periasamy, 2019; Kumar & Ieee, 2016). There are growing appeals for proposal and applications of the novel ideas in field of recognition of texts from images made by mobile cameras, text recognition and analysis problem which is significant step for video processing, image-text matching (Bin Ahmed et al., 2019; El Bahi & Zatni, 2019; Joan & Valli, 2019; X. Y. Liu et al., 2019; Roy et al., 2019; Wang et al., 2019).

Most of the studies for the design and development of recognition systems focused on characters and text in English and other famous and most used alphabets. The solution to the problems associated with different structures mentioned above local languages and their morphological structure requires an individual approach. On that account selection of the most effective features and their application is the main indicator for ensuring accuracy of a specific character and text recognition problems (Ali & Suresha, 2019; Benchaou et al., 2018; Cilia et al., 2019).

The feature selection process personalized by the essence of the investigated problem can be considered as an appropriate solution of characteristic object

recognition problems. For instance, pattern, geometrical and graphical features used for recognition of characters can be successfully applied for classification of different surfaces from urban scenes, satellite images, maps and environments (Szendrei et al., 2011; Wang et al., 2017).

One of the challenging problems that appear in this domain is the application of different optimization methods for the selection of the most appropriate features from the domain of existing parameters. As mentioned parameter sets are usually high dimensional, the application of analytical optimization methods is out of the question. One way to prevail these problems is to apply random selection methods or approaches based on the heuristic optimization algorithms (Baniya & Gnimpieba, 2020; Wang & Feng, 2019).

Through the solution of image processing problems as recognition and classification, especially in specific object detection tasks (as the face, gesture recognition), there is an infinite number of possibilities for extraction textural, geometrical and morphological features (Shaukat et al., 2018; Zmyzgova & Ieee, 2018). By examining previous work and experimental analysis through the recognition of hand-printed characters was investigated methods for the extraction of features for better classification of the characters (Ismayilova & Ismayilov, 2018).

The essential contribution of this work is the application of genetic algorithms for the selection of the best features subset from the large-scale features database extracted for Azerbaijani hand-printed characters recognition. This gives a significant advantage, as a given problem is a pattern recognition problem and there is a very large sample space for different types of features.

2. Related work

Although feature subset selection is not separately an expert system, it is one of the most important parts of the design and development of recognition and classification systems. There is a wired choice of approaches for feature selection almost in medical expert systems (Baliarsingh et al., 2019; Rachmani et al., 2019), for prediction of business, decision – making in industrial areas (Sun et al., 2019; Zandieh & Aslani, 2019), in robotics, data analysis, etc. (Harandi et al., 2019; H. Liu et al., 2019).

Even though most popular methods applied for solution of binary optimization problem appeared in machine learning systems as feature subset selection are genetic algorithms, swarm optimization and ant colony optimization, several approaches have been proposed as result of modification, combination of these algorithms, and also based on modified methods for selection, crossover and mutation during application of evolutionary algorithms (Ghosh et al., 2019; Kouchami-Sardoo et al., 2019; Qiu, 2019).

An important question associated with the application of non-analytical optimization methods for the determination of the most informative features is using parallel algorithms for extracting the data from the large feature sets without affecting the accuracy rate of intelligent systems. There exists a considerable body of literature on the solution of features subset selection by parallel or semi-parallel algorithms and distributed computing systems (Liu & Ditzler, 2019; Tsamardinos et al., 2019;

Venkataramana et al., 2019).

3. Features extraction for hand-printed character recognition

As mentioned in the introduction section, there are different approaches for the extraction of features for pattern recognition systems. In this paper were applied geometrical features for the classification of hand-printed characters. After preprocessing and thinning, characters located in rectangles with a fixed size. The idea of the authors is to classify hand-printed characters by the number of intersection points of the characters with the lines drawn through the mentioned rectangle (fig. 1). The experimental data for the given features are huge; the length of the features vector is more than 15000.

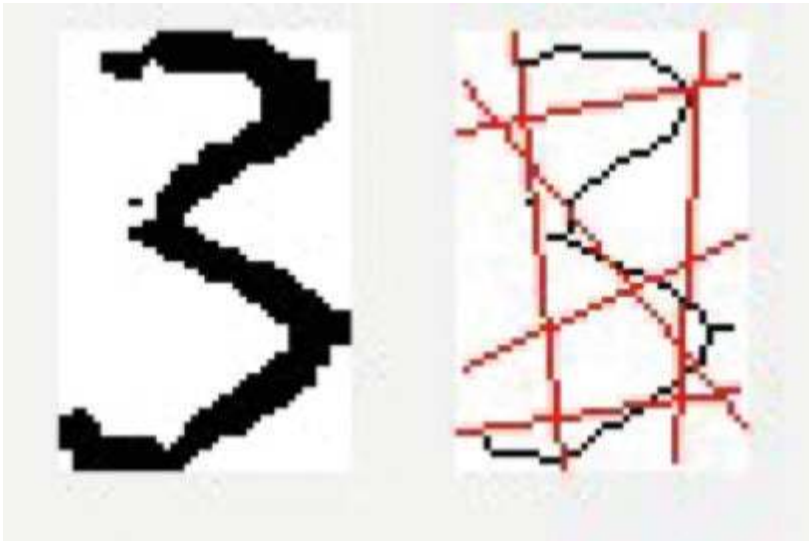

Fig. 1: Lines through the character for determination of features

For this study, we analyzed the data collected from forms filled by different people with different writing skills with Azerbaijani hand-printed characters, digits and special symbols. Classification of the training database was realized by the SVM method, which gives more accurate results for the investigated problem (Ismayilov, 2018).

For this study, we analyzed the data collected from forms filled by different people with different writing skills with Azerbaijani hand-printed characters, digits and special symbols. Classification of the training database was realized by the SVM method, which gives more accurate results for the investigated problem (Ismayilov, 2018).

4. Results of experiments

Constructed parallel algorithm based on genetic algorithms and different crossover methods were executed in ASOIU HPC center, a cluster-computing system with 20 computational nodes (https://github.com/openhpc/ohpc/wiki/System-Registry).

In tables 1, 2, 3 were described results of experiments with different indicators. Planned comparisons revealed that change of the crossover methods and different

mutation rates affects the accuracy rate of recognition and execution time of the process. This paper has investigated the application of 3 crossover methods; 1-point crossover, k-point crossover and shuffle crossover (Umbarkar & Sheth, 2015).

Superior results are seen for the combination of crossover method 2 and mutation rate<0.4 and crossover method one and mutation rate<0.5. It is interesting to note that the execution time for these experiments is very close. There were also some important differences in decreasing execution time while joining more calculation nodes to the system.

Table 1. The best features subsets and accuracy rates while evaluation process

Number of generations	Execution time (second)	Length of the features subset	Accuracy rate (%)
100	799	7682	93
250	2403	7707	94
500	4002	7650	96
1000	7963	7529	97

Table 2. Results of experiments with different number of computational nodes

Number of executive nodes	Execution time	Generation number	Length of the features subset	Accuracy rate
2	4232	500	7654	92
5	1703	500	7568	94
10	1143	1000	7787	96
15	837	1000	7731	95
20	799	1000	7598	97

Table 3. Results of experiments with different crossover methods and mutation rates

Crossover method	Mutation rate	Execution time (seconds)	Generation number	Length of the features subset	Accuracy rate (%)
Crossover 1	<0.5	800	200	7782	91
	<0.4	796	200	7674	90
	<0.3	797	200	7144	94
Crossover 2	<0.5	801	200	7689	92
	<0.4	803	200	7832	97
	<0.3	798	200	7345	93
Crossover 3	<0.5	801	200	7980	97
	<0.4	799	200	7759	90
	<0.3	787	200	7564	89

In figure is illustrated dependency between accuracy rate of the recognition system and number of generations for application of genetic algorithms. We obtain good results for 7598 features, which gives 97% of accuracy for recognition of the hand-printed characters.

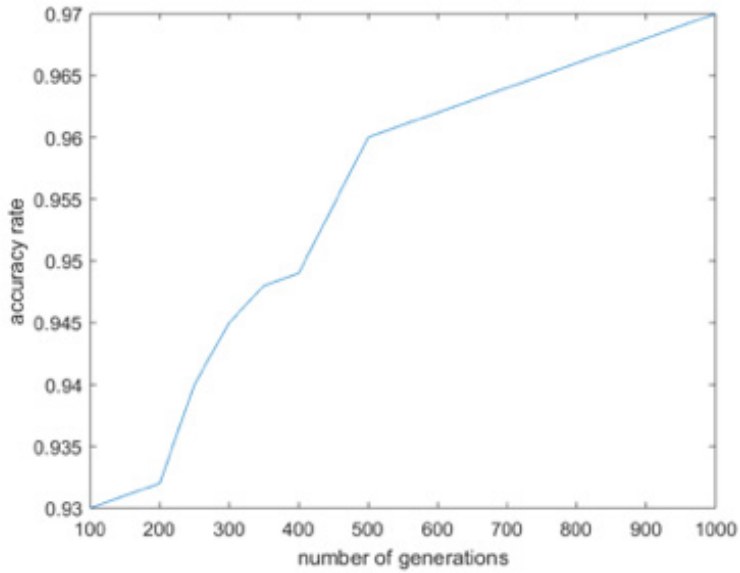

Fig. 2: Graph of dependency of the accuracy rate from generation number

5. Conclusion & Discussion

The broad implication of the present research is that given approach can be applied for the solution of any problems as data mining, image processing, signal analysis, where a huge number of features brings to the overloading of the system and high correlation between the features decrease the quality of the intelligent systems.

Future results should consider the potential effects of application of parallel genetic algorithms for optimization of the features selection step in recognition systems more carefully, for example by addition more informative and human mind based features to the system, testing different crossover, selection methods and mutation rates for determination of the best algorithm depending only the nature of the investigating problem.

References

Ali, A. A. A., & Suresha, M. (2019, Oct). A novel features and classifiers fusion technique for recognition of Arabic handwritten character script [Article]. *Sn Applied Sciences, 1*(10), 13, Article Unsp 1286. https://doi.org/10.1007/s42452-019-1294-6

Baliarsingh, S. K., Ding, W. P., Vipsita, S., & Bakshi, S. (2019, Dec). A memetic algorithm using emperor penguin and social engineering optimization for medical data classification [Article]. *Applied Soft Computing, 85, 15, Article 105773.* https://doi.

org/10.1016/j.asoc.2019.105773

Baniya, B. K., & Gnimpieba, E. Z. (2020). The Effectiveness of Distinctive Information for Cancer Cell Analysis Through Big Data [Proceedings Paper]. Advances *in Computer Vision, Vol 2, 944*, 57-68. https://doi.org/10.1007/978-3-030-17798-0_7

Barbuti, N., & Caldarola, T. (2018). An Innovative Multifunction System for Text Recognition of Digital Resources Reproducing Ancient Handwritten and Hand-Printed Artifacts. Assoc Computing Machinery. https://doi.org/10.1145/3240117.3240141

Benchaou, S., Nasri, M., & El Melhaoui, O. (2018, Jul). Feature Selection Based on Evolution Strategy for Character Recognition [Article]. *International Journal of Image and Graphics, 18*(3), 13, Article 1850014. https://doi.org/10.1142/s0219467818500146

Bin Ahmed, S., Naz, S., Razzak, M. I., & Yusof, R. (2019, Jan). Arabic Cursive Text Recognition from Natural Scene Images [Article]. *Applied Sciences-Basel, 9*(2), 27, Article 236. https://doi.org/10.3390/app9020236

Cilia, N. D., De Stefano, C., Fontanella, F., & di Freca, A. S. (2019, Apr). A ranking-based feature selection approach for handwritten character recognition [Article; Proceedings Paper]. *Pattern Recognition Letters, 121,* 77-86. https://doi.org/10.1016/j.patrec.2018.04.007

El Bahi, H., & Zatni, A. (2019, Sep). Text recognition in document images obtained by a smartphone based on deep convolutional and recurrent neural network [Article]. *Multimedia Tools and Applications, 78*(18), 26453-26481. https://doi.org/10.1007/s11042-019-07855-z

Ghosh, M., Guha, R., Singh, P. K., Bhateja, V., & Sarkar, R. (2019, Dec). A histogram based fuzzy ensemble technique for feature selection [Article]. *Evolutionary Intelligence, 12*(4), 713-724. https://doi.org/10.1007/s12065-019-00279-6

Harandi, F. A., Derhami, V., & Jamshidi, F. (2019, Dec). A new feature selection method based on task environments for controlling robots [Article]. *Applied Soft Computing, 85,* 13, Article 105812. https://doi.org/10.1016/j.asoc.2019.105812

Ismayilov, E. A. (2018). STUDY OF AZERBAIJANI HAND-PRINTED CHARACTERS RECOGNITION SYSTEM BY NEW FEATURE CLASS AND SVM METHOD. *Problems of Information Technology,* 89-94.

Ismayilova, N., & Ismayilov, E. (2018). “SOFT” FEATURES AND SVM FOR HAND-PRINTED CAHARCTERS RECOGNITION [Proceedings Paper]. *Proceedings of the 6th International Conference on Control and Optimization with Industrial Applications, Vol Ii,* 178-180.

Joan, S. P. F., & Valli, S. (2019, Mar). A Survey on Text Information Extraction from Born-Digital and Scene Text Images [Review]. *Proceedings of the National Academy of Sciences India Section a-Physical Sciences, 89*(1), 77-101. https://doi.org/10.1007/s40010-017-0478-y

Kouchami-Sardoo, I., Shirani, H., Esfandiarpour-Boroujeni, I., Alvaro-Fuentes, J., & Shekofteh, H. (2019, Nov). Optimal feature selection for prediction of wind erosion threshold friction velocity using a modified evolution algorithm [Article]. *Geoderma, 354,* 12, Article 113873. https://doi.org/10.1016/j.geoderma.2019.07.031

Kowsalya, S., & Periasamy, P. S. (2019, Sep). Recognition of Tamil handwritten character using modified neural network with aid of elephant herding optimization [Article]. *Multimedia Tools and Applications, 78*(17), 25043-25061. https://doi.org/10.1007/s11042-019-7624-2

Kumar, S., & Ieee. (2016). *A Study for Handwritten Devanagari Word Recognition.* Ieee.

Liu, H., & Ditzler, G. (2019, Aug). A semi-parallel framework for greedy information-theoretic feature selection [Article]. *Information Sciences, 492,* 13-28. https://doi.org/10.1016/j.ins.2019.03.075

Liu, H., Duan, Z., Wu, H. P., Li, Y. F., & Dong, S. Y. (2019, Dec). Wind speed forecasting models based on data decomposition, feature selection and group method of data handling network [Article]. *Measurement, 148,* 12, Article Unsp 106971. https://doi.org/10.1016/j.measurement.2019.106971

Liu, X. Y., Meng, G. F., & Pan, C. H. (2019, Jun). Scene text detection and recognition with advances in deep learning: a survey [Article]. *International Journal on Document Analysis and Recognition, 22*(2), 143-162. https://doi.org/10.1007/s10032-019-00320-5

Qiu, C. Y. (2019, Dec). A novel multi-swarm particle swarm optimization for feature selection [Article]. *Genetic Programming and Evolvable Machines, 20*(4), 503-529. https://doi.org/10.1007/s10710-019-09358-0

Rachmani, E., Hsu, C. Y., Nurjanah, N., Chang, P. W., Shidik, G. F., Noersasongko, E., Jumanto, J., Fuad, A., Ningrum, D. N. A., Kurniadi, A., & Lin, M. C. (2019, Dec). Developing an Indonesia's health literacy short-form survey questionnaire (HLS-EU-SQ10-IDN) using the feature selection and genetic algorithm [Article]. *Computer Methods and Programs in Biomedicine, 182*, 10, Article Unsp 105047. https://doi.org/10.1016/j.cmpb.2019.105047

Roy, P. P., Bhunia, A. K., Bhattacharyya, A., & Pal, U. (2019, Mar). Word searching in scene image and video frame in multi-script scenario using dynamic shape coding [Article]. *Multimedia Tools and Applications, 78*(6), 7767-7801. https://doi.org/10.1007/s11042-018-6484-5

Shaukat, A., Farhan, S., Fahiem, M. A., Tauseef, H., Tahir, F., & Usman, G. (2018, Nov). Textural and Geometrical Features Based Approach for Identification of Individuals Using Palmprint and Hand Shape Images from Multiple Multimodal Datasets [Article]. *Journal of Testing and Evaluation, 46*(6), 2281-2298. https://doi.org/10.1520/jte20160625

Sun, J., Zhou, M. J., Ai, W. G., & Li, H. (2019, Dec). Dynamic prediction of relative financial distress based on imbalanced data stream: from the view of one industry [Article]. *Risk Management-an International Journal, 21*(4), 215-242. https://doi.org/10.1057/s41283-018-0047-y

Szendrei, R., Elek, I., & Marton, M. (2011). Graph-Based Feature Recognition of Line-Like Topographic Map Symbols. In Y. Tan, Y. Shi, Y. Chai, & G. Wang (Eds.), *Advances in Swarm Intelligence, Pt Ii* (Vol. 6729, pp. 291-298). Springer-Verlag Berlin.

Tsamardinos, I., Borboudakis, G., Katsogridakis, P., Pratikakis, P., & Christophides, V. (2019, Feb). A greedy feature selection algorithm for Big Data of high dimensionality [Article]. *Machine Learning, 108*(2), 149-202. https://doi.org/10.1007/s10994-018-5748-7

Umbarkar, A. J., & Sheth, P. D. (2015). CROSSOVER OPERATORS IN GENETIC ALGORITHMS: A REVIEW. *ICTACT journal on soft computing, 6*(1).

Venkataramana, L., Jacob, S. G., Ramadoss, R., Saisuma, D., Haritha, D., & Manoja, K. (2019, Nov). Improving classification accuracy of cancer types using parallel hybrid feature selection on microarray gene expression data [Article]. *Genes & Genomics, 41*(11), 1301-1313. https://doi.org/10.1007/s13258-019-00859-x

Wang, J. L., Lu, Y. H., Liu, J. B., & Quan, L. (2017, Oct). A robust three-stage approach to large-scale urban scene recognition [Article]. *Science China-Information Sciences, 60*(10), 13, Article 103101. https://doi.org/10.1007/s11432-017-9178-8

Wang, X., Feng, X., & Xia, Z. (2019, Oct). Scene video text tracking based on hybrid deep text detection and layout constraint [Article]. *Neurocomputing, 363,* 223-235. https://doi.org/10.1016/j.neucom.2019.05.101

Wang, Y. W., & Feng, L. Z. (2019, Dec). A new hybrid feature selection based on multi-filter weights and multi-feature weights [Article]. *Applied Intelligence, 49*(12), 4033-4057. https://doi.org/10.1007/s10489-019-01470-z

Zandieh, M., & Aslani, B. (2019, Dec). A hybrid MCDM approach for order distribution in a multiple-supplier supply chain: A case study [Article]. *Journal of Industrial Information Integration, 16***,** 13, Article 100104. https://doi.org/10.1016/j.jii.2019.08.002

Zmyzgova, T. R., & Ieee. (2018). *Special Features of Structural Pattern Recognition for Digital Images of Reactions of Integral Strain Gauges Indications.* IEEE.

Submitted 08.06.2019
Accepted 22.11.2019