Testing the Algorithm of Area Optimization by Binary Classification with Use of Three State 2D Cellular Automata in Layers

1st Miroslaw Szaban
Institute of Computer Science,
Siedlce University of Natural Sciences and Humanities,
Poland
mszaban@uph.edu.pl

2nd Anna Wawrzynczak
Institute of Computer Science,
Siedlce University of Natural Sciences and Humanities
and National Centre for Nuclear Research,
Poland
awawrzynczak@uph.edu.pl

Abstract—The paper is dedicated to a new algorithm of optimization in the sense of the area. Proposed method joins a few issues. First one is utilizing data from the set of sensors monitoring the area put into optimization. The second one is using the classification method based on two-dimensional three-state cellular automata, working on the data reported by the sensors. This method classifies all points of the area based on the data received from the sensors and designates optimal subarea. The third issue is applying the categorization layers to the data received from sensors. Such, approach gives a possibility to specify the areas in the different levels and, in consequence, after analysis, optimal subarea or subarea including the optimal point can be designated. This method can be used in different optimization tasks, starting from simple one as optimization of n-dimensional function, through specifying the contaminated area utilizing data from mobile sensors and finally estimating the contamination source-term. In this paper are presented results of testing for the proposed algorithm on a few selected functions from the set of dedicated for this purpose.

Index Terms—area optimization, cellular automata, classification, sensors

I. Introduction

In a classification problem, we wish to determine to which class new observations belong, based on the training set of data containing observations whose class is known. The binary classification deals with only two classes, whereas in a multiclass classification observations belong to one of the several classes. The well-known classifiers are neural networks, support vector machines, k-NN algorithm, decision trees, and others. The idea of using cellular automata (CA) in the classification problem was described by Maji et al. [2], Povalej et al. [3] and by Fawcett [1]. Fawcett designed the heuristic rule based on the von Neumann neighborhood (so-called voting rule); moreover, tested its performance on different sets of data. Results of Fawcett's study indicated his method, based on CA, as better than the other compared methods, like as (a) J48, a decision tree induction algorithm, (b) k-NN, a nearest-neighbor learning algorithm, (c) SMO, implementation of support vector machines. Recently, in the papers [4] were proposed and analyzed the Fawcett's method modifications into a probabilistic form of such method. These

modifications were examined on the different sets of data, and obtained results show in general its higher effectiveness for classification (lower number of incorrect classifications), also in general better accuracy (the shortest scattering range).

The reconstruction of the source of an airborne contaminant may be obtained by using forward approaches, in which source characteristics are inferred from concentration or deposition measurements at different locations and time intervals by establishing source-concentration relationships. In e.g.[10] authors presented the reconstruction of the airborne contaminant source utilizing the Bayesian approach in conjunction with Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC). A comprehensive literature review of past works on solutions of the inverse problem for atmospheric contaminant releases can be found in (e.g.[9]). This class of problems is a potential area of application for the newly presented algorithm of area optimization and binary classification with use of three state 2D cellular automata.

This paper is organized as follows. Section 2 describes two-dimensional CAs, binary classification problem and binary classification methods based on 3-state CA. In Section 3 is presents the construction of the algorithm of area optimization. The stages of proposed approach examining and experimental results are presented in Section 4. The last Section concludes the paper, and is a study of application possibilities of the newly proposed algorithm and plans of future work.

II. Two-Dimensional Cellular Automata and Binary Classification Problem

CAs and they potential to efficiently perform complex computations are described by S. Wolfram in [8]. In this paper is considered two-dimensional CA. CA is a rectangular grid of $N \times M$ cells, each of which can take on k possible states. After determining initial states of all cells (i.e. the initial configuration of a CA), each cell changes its state according to a rule - transition function TF which depends on states of cells in a neighborhood around it. In this paper is considered finite CA (finite length of CA) with the periodic boundary conditions (bordered cells are neighbour cells each other).

Two types of the neighborhood are commonly used: the von Neumann neighborhood (the four cells orthogonally surrounding the central cell) which can be described as $a_{i,j}^{(t+1)} = TF[a_{i,j-1}^{(t)}, a_{i-1,j}^{(t)}, a_{i,j}^{(t)}, a_{i+1,j}^{(t)}, a_{i,j+1}^{(t)}]$, where $a_{i,j}^{(t)}$ denotes the state of a cell at position i, j in the two-dimensional cellular grid, at time step t. Also, the Moore neighborhood (the eight cells around the central cell) which can be described as $a_{i,j}^{(t+1)} = TF[a_{i-1,j-1}^{(t)}, a_{i,j-1}^{(t)}, a_{i+1,j-1}^{(t)}, a_{i-1,j}^{(t)}, a_{i,j}^{(t)}, a_{i+1,j}^{(t)}, a_{i-1,j+1}^{(t)}, a_{i,j+1}^{(t)}, a_{i+1,j+1}^{(t)}]$.

The square state of the data space in classification problem should be i.e. $[0, 1] \times [0, 1]$. Suppose that $N \times M$ data-points $p_{(i,j)} = (x_i, y_j)$, where $i=1, 2, ..., N$ and $j=1, 2, ..., M$ are given as a training set from two classes: class 1 and class 2. When each of $p_{(i,j)}$ data-points is known as one of two classes then we have the classification. On the other hand, when even one of the data-points is not one of two known classes we have the classification problem. Moreover, to answer the question, to which of class (1 or 2) unclassified data points belong to, the classification method should be applied. In CA the data space of such problem should be mapped from $[0, 1] \times [0, 1]$ into the grid of $N \times M$ cells (in this paper $N \times N$ for the simplicity). Each cell can take one of 3 states, classified the state 1 (class 1) and state 2 (class 2) and also unclassified state (class 0). Classifier - the rule of CA will analyze the unclassified cells and changes its states into one of two known.

The classification methods based on two-dimensional three-state cellular automata was applied for classifying whole points of the area on the base data received from the sensors. The goal was expected designation of the optimal subarea. For this purpose, three kinds of the classifiers were studied. The first classification method was proposed in [1] (the rule of CA known as $n4\_V1\_stable$). The second and third one were modifications of Fawcett's rule into two patterns: partially and fully probabilistic (see, [4]). A proposed modification should strengthen an original and more accurately classify binary data, especially for large CA grid.

III. ALGORITHM OF THE AREA OPTIMIZATION METHOD

The proposed method uses the data reported by the set of sensors monitoring the area put into optimization. This data are the input for the classification method based on two-dimensional three-state cellular automata, which classifies all points of the area to designate optimal subarea. The layers are categorized based on the level of values received from sensors. The steps of the algorithm of the optimization method are presented below.

The algorithm of area optimization by layers and binary classification with use of three state 2D cellular automata:

1) Downloading input parameters:
   - CA size,
   - Threshold - the minimum value for which the recorded indication is acceptable (sensor is in positive state - class 1), the lover values recorded by the sensor are considered as 0 (sensor is in negative state - class 2),
   - Step of Threshold - the value by which the Threshold is increased during processing, it designates the levels of the layers,
   - Number of sensors,
   - Method of data classification (including type of neighborhood);
2) Preparation of cellular automaton,
   - Mapping optimizing area into (discrete) CA grid,
   - The random distribution of sensors in CA grid - CA cells with included sensors are the classified cells (class 1 or class 2), other CA cells are unclassified (class 0),
3) Searching for solution:
   - Preparation of layers for cellular automaton work (with use of Threshold and Step of Threshold): values of sensors in layer $i \in [Threshold + (i-1) * Step; Threshold + i * Step)$, where $i = 1, 2, ...,$
   - For each layer, specify classes of CA cells with sensors: where the sensor value is $> Threshold + (i-1) * Step$, where $i = 1, 2, ...$ Then the sensor is in a positive state (class 1). Otherwise, the sensor is in negative state - class 2,
   - For each layer perform classification - during processing of a cellular automaton are specifying classes 1 or 2 for CA cells being unclassified (class 0):
4) Elaboration of received results:
   - Designation of optimal area - development of common parts from classified as class 1 areas on each of analyzed layers,

IV. EXPERIMENTAL RESULTS

The above presented and described algorithm for area optimization by layers, which apply in its work binary classification on three states two-dimensional CA should be examined in the sense of its efficiency for optimization. In this kind of test are usually used dedicated sets of different n-variables functions, as Test Functions for Unconstrained Global Optimization [7]. This set contains the testing functions with one or multi-global optima. For our problem were used functions having one optimum. Because functions have one global minimum, and proposed algorithm searching for maximum, as it was mentioned earlier, the functions $f(x)$ were inverted into $f^*(x) = f_{max}(x) - f(x)$, $f_{max}(x)$ is the maximal value of function in analyzed search domain. In the tests were used three selected and inverted functions: Booth and Matyas Function for which the results are described in this paper, and Sphere Function for which results are presented in [5], [6].

In conducted experiments, the applied CA was two-dimensional with size 500×500. The higher size of CA the more accurate results we retrieve. For proper analysis, Threshold in the algorithm was specified as: 1000 for Booth function and 10 for Matyas. The tests were conducted for a varying number of sensors ($\{5, 10, ..., 45, 50\}$) and steps of threshold ($\{2, 4, 6, 8, 10\}$).
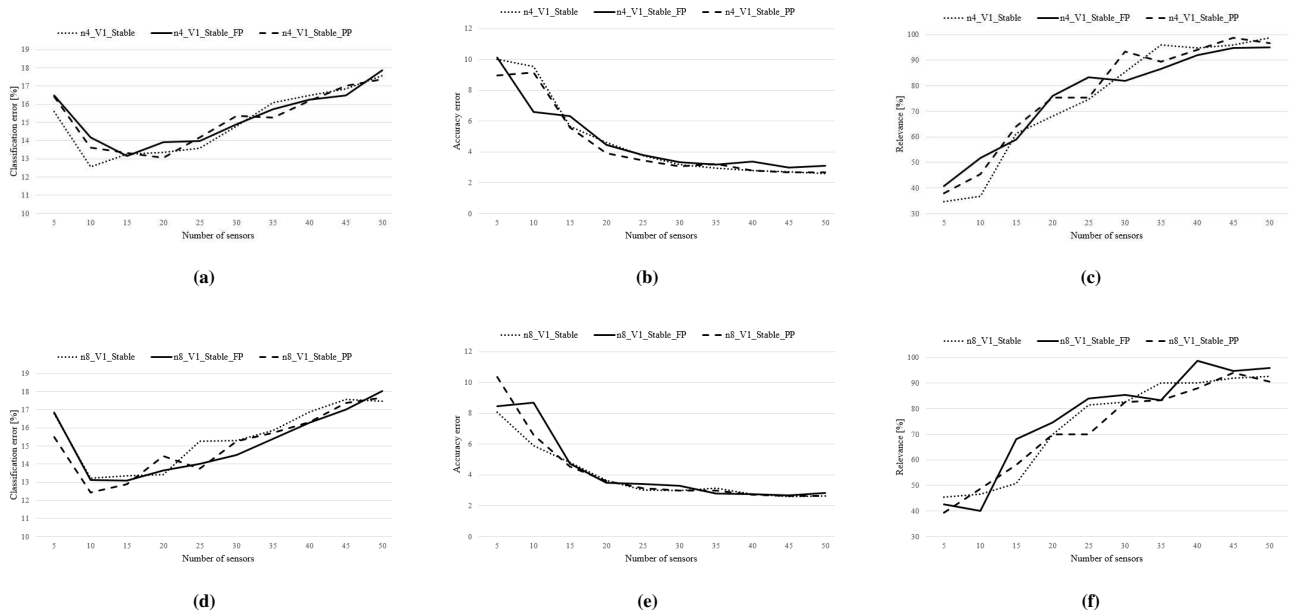
Fig. 1. Comparison of average values of **(a)** Classification error, **(b)** Accuracy error (real distance) and **(c)** Relevance (in [%]) resulted for varying number of sensors with use of each classification method ($n4\_V1\_stable$, $n4\_V1\_stable\_PP$ and $n4\_V1\_stable\_FP$) for von Neumann neighborhood on Booth function. Also, comparison of average values of **(d)** Classification error, **(e)** Accuracy error (real distance) and **(f)** Relevance (in [%]) obtained for varying number of sensors with use of each classification method ($n8\_V1\_stable$, $n8\_V1\_stable\_PP$ and $n8\_V1\_stable\_FP$) for Moore neighborhood on Booth function.

Each of tests series contains 500 single runs of the algorithm with the different random sensors spatial setup. From the set of conducted experiments were calculated the average values of classification error, accuracy error, and relevance for varying number of sensors.

A. Testing With the Booth Function

Figure 1 presents these results for Booth function with use of each classification method with von Neumann neighborhood. We can see that with growing number of sensors, the quality of results is generally getting better, except the classification error. The classification error presented in Figure 1(a) is not higher than 18% for the bordered numbers of sensors and has a parabolic trend. The lowest classification error has value about 13% for a number of sensors fluctuated from 10 to 25. This result is independent of the classification method. The error of accuracy (see, Figure 1(b)) is not higher than ~ 10 for only five sensors case and is generally going down near to 3 for 50 sensors setup. From the level of 25 sensors, accuracy error is lower than 4. The Relevance for 20 sensors is not lower than 70% and is going up. Since the setup of 40 sensors relevance being higher than 90%, where for the partially probabilistic method of classification ($n4\_V1\_stable\_PP$) and $n4\_V1\_stable$ method is near to 100% (see, Figure 1(c)).

Next step of experiments for Booth function is presented in Figure 1, where each of classification method was used with Moore neighborhood. In this case, we can see similar results and trends in particular for errors of classification and accuracy. Interesting is the fact that the fully probabilistic method of classification ($n8\_V1\_stable\_FP$) characterizes better re-

sults than for von Neumann neighborhood. Furthermore, in the case of relevance, the scores for $n8\_V1\_stable\_FP$ are slightly better than for other analyzed methods of classification. Relevance for this method and for 40 sensors is near to 100% (see, Figure 1(f)).

B. Testing With the Matyas Function

In this subsection are presented results of testing for Matyas function. Figure 2 shows these results for each of classification method with von Neumann neighborhood. As we can expect, with growing number of sensors, the quality of results is generally getting better, except the classification error. The classification error presented in Figure 2(a) is not higher than 18% only up to 15 sensor setup. For higher number of sensors classification error going up to $\sim 25\%$ for 50 sensors setup. This result is independent of the classification method. The error of accuracy (see, Figure 2(b)) is not higher than ~ 12 for only five sensors case and is generally going down near to 1 for 50 sensors setup (it is generally better than for Booth function). From the level of 20 sensors, accuracy error is lower than 4. The Relevance for 25 sensors is not lower than 70% and is going up to $\sim 90\%$, independently on the classification method (see, Figure 2(c)).

The experiments for Matyas function, where each of classification method was used with Moore neighborhood is presented in Figure 2. In this case, we can see similar results and trends to obtained for Matyas function and von Neumann neighborhood for each of assessment criteria. Moreover, classification error for the fully probabilistic method of classification ($n8\_V1\_stable\_FP$) seems to be better than other methods of classification (see, Figure 2(d)).
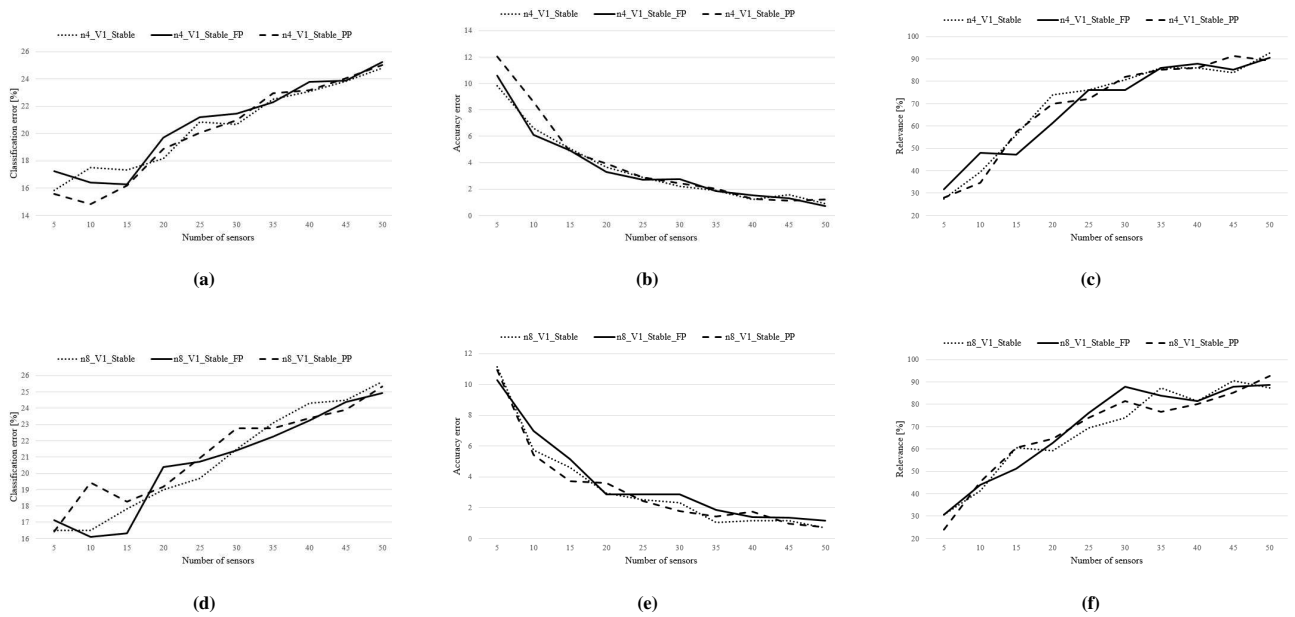
Fig. 2. Comparison of average values of **(a)** Classification error, **(b)** Accuracy error (real distance) and **(c)** Relevance (in [%]) resulted for varying number of sensors with use of each classification method ($n4\_V1\_stable$, $n4\_V1\_stable\_PP$ and $n4\_V1\_stable\_FP$) for von Neumann neighborhood on Matyas function. Also, comparison of average values of **(d)** Classification error, **(e)** Accuracy error (real distance) and **(f)** Relevance (in [%]) obtained for varying number of sensors with use of each classification method ($n8\_V1\_stable$, $n8\_V1\_stable\_PP$ and $n8\_V1\_stable\_FP$) for Moore neighborhood on Matyas function.

CONCLUSIONS

In this paper is proposed an algorithm of area optimization by layers and binary classification with use of three state 2D cellular automata. The proposed algorithm as an input utilizes the data reported by the set of sensors monitoring the area put into optimization. This data are subject to the classification based on two-dimensional three-state cellular automata, which classifies all points of the area to designate optimal subarea. The layers are categorized based on the level of values received from sensors. Such, approach gives a possibility to specify the areas in the different levels and after analysis could be selected optimal subarea or subarea included the optimal point.

The algorithm was verified and tested with use of two functions: Booth and Matyas included in the set of the functions applied for testing optimization/optimizing algorithms. The methods for interpretation of obtained results were introduced in conjunction with algorithms assessment criteria, like classification error, accuracy error, and relevance. The values of the algorithm characteristics corresponding to the algorithm run with each of three classification methods for different setups, i.e., the varying number of sensors (input data) were presented.

Conducted studies show that quite good results characterize proposed algorithm. Reasonably high relevance value, i.e. $\sim$ 90%, and higher was reached. Furthermore, the accuracy error are characterized by the low value. Performed tests show that this method could be used in different optimization problems starting from simple ones, as optimization of n-dimensional functions, and in more complicated tasks as designating the contaminated area based on the restricted number of mobile sensors data or estimating the source of airborne toxin.

Presented experiments prove that proposed algorithm is able to reduce the scanned area to the little size (even optimal). Furthermore, studies of the relevance of results obtained by the proposed algorithm indicate that it can be used as an optimization tool, which going to indicate the area including the optimum.

REFERENCES

[1] T. Fawcett, "Data mining with cellular automata." *ACM SIGKDD Explorations Newsletter*, 10(1), pp. 32–39, 2008.
[2] P. Maji, B. Sikdar, and P. Chaudhuri, "Cellular automata evolution for pattern classification." *LNCS 3305* , pp. 660–669. Springer Verlag 2004.
[3] P. Povalej, M. Lenic, and P. Kokol, "Improving ensembles with classificational cellular automata." *LNCS 3305*, pp. 242–249. Springer 2004.
[4] M. Szaban: "Probabilistic 2D Cellular Automata Rules for Binary Classification." M. Ganzha, L. Maciaszek, M. Paprzycki (eds.): Annals of Computer Science and Information Systems, Volume 8 (FedCSIS 2016), 2016, pp. 161-164, DOI: 10.15439/978-83-60810-90-3
[5] M. Szaban, A. Wawrzynczak: "The Algorithm of Area Optimization by Layers and Binary Classification with Use of Three State 2D Cellular Automata." Proceedings of the International Conference on Control, Artificial Intelligence, Robotics and Optimization, 2018 (in print)
[6] L. Swider: "Optimization of two variables functions with use of data classification algorithms based on cellular automaton". Master thesis, (2016), (in Polish).
[7] Test Functions for Unconstrained Global Optimization: $http://www-optima.amp.i.kyoto-u.ac.jp/member/student/hedar/Hedar\_files/TestGO\_files/Page364.htm$
[8] S. Wolfram: *A New Kind of Science.* Wolfram Media, 2002.
[9] Hutchinson, M., Oh, H. & Chen, W.H.: "A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors." Information Fusion 36, (2017), pp. 130-148.
[10] Wawrzynczak A., Kopka P., Borysiewicz M.: "Sequential Monte Carlo in Bayesian assessment of contaminant source localization based on the distributed sensors measurements.", LNCS 8385, (2014), pp. 407-417.