Evaluation of parametric acyclic Markov models for dependent objects*

***Dvoenko S. D.*[1], *Sang D. V.*[2]**
dsd@tsu.tula.ru, dvietsang@gmail.com
[1]Tula State University, 300600, Tula, Lenin Ave., 92; [2]Hanoi University of Science and Technology, Hanoi, Dai Co Viet St., 1, Vietnam

In modern theory of pattern recognition objects are often classified with regard to interrelations (data coherence, spatial and temporal cohesion, etc.) between them. Markov random fields (MRFs) are most popular to model such objects. The interrelations between neighboring objects are represented by an adjacency graph. In general, the inference in MRFs is *NP*-hard when the adjacency graph contains cycles. The main idea of this work is to replace the graph with cycles by a linear combination of a finite or countable set of acyclic (treelike) parametric Markov models, for which the problem of recognizing MRFs can be efficiently solved. We propose a simplified cross-validation procedure to statistically evaluate the quality of solutions and to adjust the parameters of the linear combination, in which the Markov ones are treated as hyper-parameters.

Keywords: *Markov random fields, Markov chain, graphical models, image processing, pattern recognition.*

Оценка параметрических ациклических марковских моделей для зависимых объектов*

***Двоенко С. Д.*[1], *Шанг Д. В.*[2]**
[1]Тульский государственный университет, 92, пр. Ленина, г. Тула; [2]Ханойский научный и технический университет, 1, ул. Дай Ко Вьет, г. Ханой, Вьетнам

В современной теории распознавания образов объекты часто классифицированы с учетом взаимосвязей между ними. Марковские случайные поля являются наиболее популярными моделями таких объектов. Взаимосвязи между соседними объектами представлены графом соседства. Как правило, для графов общего вида с циклами задача распознавания марковских случайных полей обладает свойствами задачи класса *NP*. В данной работе предлагается заменить граф с циклами линейной комбинацией конечного или счетного множества ациклических (древовидных) параметрических марковских моделей, для которых проблема распознавания марковских случайных полей может быт эффективно решена. Предлагается упрощенная процедура скользящего контроля для статистической оценки качества решения и настройки параметров линейной комбинации, где марковские параметры рассматриваются как структурные.

Ключевые слова: *марковские случайные поля, марковская цепь, графические модели, обработка изображений, распознавание образов.*

Introduction

In classical pattern recognition problem objects are independently classified. However, in the modern theory of pattern recognition and machine learning the set of objects is usually

*Supported by the RFBR grants 13-07-00529, 14-07-00964.

treated as an array of interrelated data. The interrelations between objects are represented by an undirected adjacency graph $G = (T, E)$, where T is the set of objects $t \in T$ and E is the set of edges $(s, t) \in E$ connecting two neighboring objects $s, t \in T$. The adjacency graph of the linearly ordered array is a chain.

Hidden Markov models (HMMs) have proved to be very efficient for processing data in the form of a chain. But for arbitrary adjacency graphs with cycles, e.g., 4-connected grid of image pixels, finding the maximum of a posteriori probability (MAP) of a MRF is a *NP*-hard problem [1]. The standard way for solution is to specify a posteriori distribution of the MRF by clique potentials, and solve the problem in terms of Gibbs energy [1, 2]. Hereby, the MAP estimation corresponds to minimizing of Gibbs energy over all cliques of the graph G.

In this work we use acyclic Markov models developed in [3, 4, 5], for which we directly estimate the MAP solutions of MRFs. Instead of the graph with cycles, we use a finite or countable set of acyclic graphs to combine corresponding acyclic Markov models proposed in [5].

According to [3, 4, 5], a transition matrix Q, which is a parameter of so-called one-sided Markov model, was specified by a unique diagonal element q only. Later, algorithms for adjusting the diagonal element and weights of the finite set of acyclic models were developed in [6].

On the other hand, Markov parameters can be treated as hyper-parameters. However, adjusting them based on the full cross-validation scheme takes a very high time complexity.

In this work we propose also new algorithms for adjusting the diagonal element for a countable set of acyclic adjacency graphs and propose a simplified cross-validation scheme to adjust Markov parameters as hyper-parameters with much lower time complexity while maintaining a high quality of solutions.

Data array recognition problem

Let denote by T the array of objects $t \in T$. The interrelations between objects are defined by an undirected graph G without cycles.

Using the idea of HMM, the array T is treated as a two-component field (X, Y), where the observed part $Y = (\mathbf{y}_t, t \in T)$ consists of the feature vectors, which take values from a certain set $\mathbf{y}_t \in \Phi$. The hidden part $X = (x_t, t \in T)$ consists of classes that need to be labeled from a finite set $\Omega = \{1, 2, \ldots, m\}$, where m is the number of classes. Let the hidden part X be restored using the MAP estimation:

$$\begin{aligned} \hat{X}(Y) &= (\hat{x}_t, t \in T), \text{where} \\ \hat{x}_t(Y) &= \arg\max_{x_t \in \Omega} p_t(x_t|Y). \end{aligned} \tag{1}$$

Let the feature vectors $\mathbf{y}_t, t \in T$ be conditionally independent with respect to (w.r.t.) the hidden part:

$$\psi_t(\mathbf{y}_t|X) = \psi_t(\mathbf{y}_t|x_t). \tag{2}$$

Let the hidden part X be a MRF w.r.t. the acyclic graph G. It was shown in [3], that a priori and a posteriori fields of hidden classes X are one-sided MRFs with respect to the same graph G.

In pattern recognition problem a posteriori distributions $p_t(x_t|\mathbf{y}_t)$ are usually evaluated by using some supervised learning method. Then the MAP solution of the hidden part X can be estimated by using the so-called basic algorithm in [4, 5, 6] with two passes along the acyclic graph G. The forward pass from the terminal vertexes to the root of G estimates filtering a

posteriori marginal distributions of hidden classes. The backward pass from the root to the terminal vertexes estimates interpolating a posteriori marginal distributions of hidden classes.

In [3, 4, 5] the particular model of a MRF with some simplifications was introduced. At first, the model assumes that the one-sided MRF X is a homogeneous ergodic reversible finite Markov chain. The next simplification is that the transition matrix $Q(m \times m)$ of the hidden field X is a symmetric and doubly stochastic matrix (i.e., a matrix with unit sums of elements over each row and each column), which consists of identical diagonal and identical non-diagonal elements as follows:

$$Q = \begin{bmatrix} q & \frac{1-q}{m-1} & \dots & \frac{1-q}{m-1} \\ \frac{1-q}{m-1} & q & \dots & \frac{1-q}{m-1} \\ \dots & & & \\ \frac{1-q}{m-1} & \frac{1-q}{m-1} & \dots & q \end{bmatrix} \tag{3}$$

Such a matrix Q can be specified by a single diagonal element q, which is the unique Markov parameter that needs to be adjusted.

Parametric adjusting of a set of acyclic graphs

Parametric adjusting of a finite set of acyclic graphs. Three algorithms for adjusting diagonal elements and weights of the linear combination of a given set of acyclic graphs were proposed in [6]. The general idea of them is to include the searching for the diagonal element in the process of adjusting graph weights based on the Gauss-Seidel scheme used in [5]. Variations of the diagonal element q in the range $1/m \leqslant q < 1$ and graph weights w in the range $0 \leqslant w \leqslant 1$ are considered like the coordinate-wise descent. In the first algorithm A1 all Markov models for graphs $G_k, k = 1, \dots, K$ are defined by a single diagonal element q. In two other algorithms A2 and A3 each graph $G_k, k = 1, \dots, K$ corresponds to a separate acyclic Markov model with its own diagonal element $q_k, k = 1, \dots, K$.

These algorithms demonstrate [6] a high level of recognition quality. It is the same, for instance, like the algorithm TRWS [7], which is considered as the most efficient one nowadays.

Parametric adjusting of a countable (full) set of acyclic graphs. Two recognition algorithms named A4 (basic) and A5 (sequential basic) were previously developed [5] for this case. Because we don't know real acyclic graphs from the whole infinite countable set (or from its some unknown limited subset in a particular case), we don't know their weights in a linear combination. More over, we don't need to use graph weights in A4 and A5 algorithms.

The general idea of these algorithms consists in expanding vicinities of all objects w.r.t. the arbitrary adjacency graph G (with cycles, e.g., raster lattice) in the data array up to the moment when all objects will be in all maximal vicinities (Fig. 1).

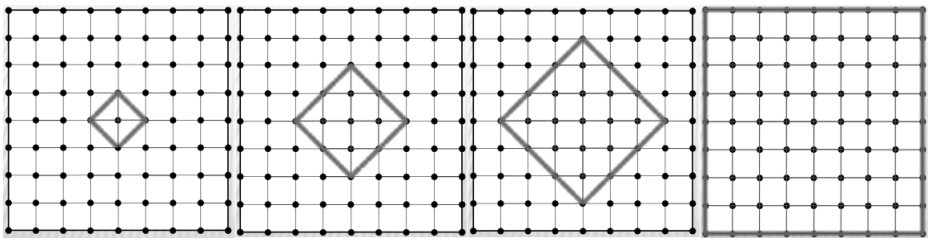

Fig. 1. The expanded vicinity of the central point of a lattice 9×9 after steps 1, 2, 3, 8

Let each object $t \in T$ be the root of some its personal and unknown for us acyclic graph G from the countable set. The step-by-step vicinity expansions w.r.t. every graphs G can evaluate

the step-by-step filtering MAP solutions for each object just by single pass along its current vicinity sub graph of each G, and can do it simultaneously with others. The idea of simultaneous step-by-step expansions of all vicinities is the idea of the A4 algorithm (basic), which realizes, per se, the parallel scheme of the set processing. The idea of sequential step-by-step expansions of all vicinities is the idea of the A5 algorithm (sequential basic).

The final vicinities give the final MAP solutions for all $t \in T$ as interpolating a posteriori marginal distributions of hidden classes.

As a result, the linear combination becomes a real integration of some unknown subset of acyclic adjacency graphs from the countable set represented by the given arbitrary graph G.

Both A4 and A5 algorithms have a single value of the diagonal element q, which is the unique Markov parameter that needs to be adjusted.

It was shown in [5], A4 and A5 algorithms without parametric adjusting demonstrate a high level of recognition quality too, which is statistically the same, for instance, with the TRWS algorithm [7]. The algorithm for parametric adjusting of A4 and A5 consists in just variations of the diagonal element q in the range $1/m \leqslant q < 1$ without weights. Hence, we suppose only the single integrated Markov model for the arbitrary graph G.

Here we focus on the A4 algorithm only for the countable (full) set of acyclic graphs. In fact, if the scanning direction of the set of objects isn't defined (sequential basic), we can use really only the parallel scheme of the set processing. This scheme is very promising in view of the fast processing of large sets.

Hyper-parameter adjusting problem

In general case, parameters can be divided into natural ones and hyper-parameters. Natural parameters reflect the essential properties of a model. But in many cases, it is required to determine some extra-parameters to control the configuration of the parametric model or a priori distribution of natural parameters. These extra parameters are called as hyper-parameters.

Sometimes the difference between natural and hyper-parameters is quite unclear. The interpretation of parameters as natural or hyper-parameters can be determined by investigator's point of view. In particular, such a problem arises in adjusting parameters for the combination of acyclic adjacency graphs. Markov parameters and graph weights can be viewed as either natural or hyper-parameters. It is a complicated problem to determine the correct hyper-parameters, because they have to restrict decision rules in order to avoid overfitting. Thus they can't be directly used in the error minimization process based on the training set to estimate the natural parameters. In model selection hyper-parameters are usually selected by the cross-validation. In practice the k-folds cross-validation is often used, where data are split into roughly equal-sized parts.

Some overfitting can occur in the cross-validation-based hyper-parameter optimization. Therefore, it may be necessary to use another test set that doesn't participate in the validation process. The statistical efficiency of the adjusted model is finally evaluated based on the test set. The standard way of hyper-parameter adjusting is the grid search, which is an exhaustive searching through a specified value set. This requires considerable time for multiple training runs that are applicable only to fast learning algorithms. As a rule, no more than 2-3 hyper-parameters can be validated together.

In order to reduce the computational complexity we consider graph weights as natural parameters, and Markov parameters as hyper-parameters for the finite set. At last, we propose a hybrid approach [6] consists in adjusting Markov parameters as natural in the same procedure with adjusting graph weights based on the Gauss-Seidel scheme. But the quality of the decision

rule is evaluated by a simplified cross-validation scheme that is similar to hyper-parameter adjusting scheme.

Simplified cross-validation scheme

In this work we apply our algorithms to the problem of image segmentation. Particularly, we use our algorithms to segment 100 simulated raster textured images (Fig. 2) with the size of 201×201. Each image contains three classes of textures which are realizations of normally distributed two-dimensional random variables with slightly different means in the feature space of RG-color components. The error rate of independent recognition for such images is not less than 30%.

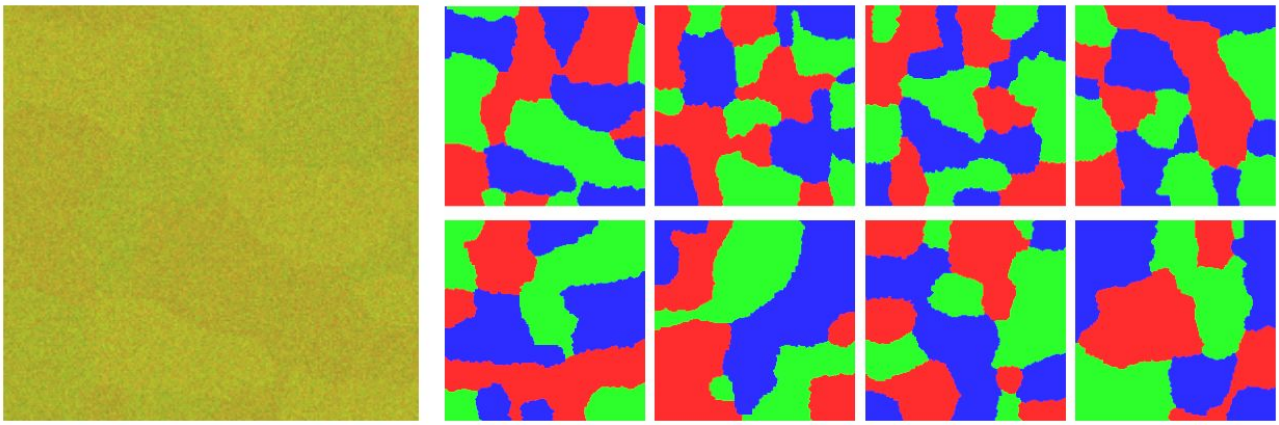

Fig. 2. The first simulated image and the exact segmentations of first eight simulated images

Generally, the set of images is divided into training and test sets, where the test set usually contains 25% of the total volume of the initial set. Such a partition of the initial set is often repeated 100 times. By repeating partitions some images may never participate in test, and some others may never participate in training.

In the simplified scheme of cross-validation we alternately leave one image as a test and the remaining images are used for cross-validation. It guarantees that each image is used for test once, and all images are involved in training with the same number of times.

To reduce the computational complexity we apply the non-classical cross-validation with only one image for training. It is known that a relatively small amount of samples in the training set may lead to overfitting. Hence, at first glance, training on one image seems not to be sufficient. However, an image of size 201×201 consists of about 40 thousands of pixels. This means, in fact, we train classifiers on about 40 thousands of objects. This large number of objects ensure us to train good classifiers with high generalization ability.

Another reason to use the non-classical cross-validation scheme is that adjusting Markov parameters for one simulated image and for most of them have almost similar results. Indeed, for each image let us consider the error line obtained by combining acyclic adjacency graphs with equal weights and varying value of diagonal element. The similar error line is built on the set of remaining images.

Experiments show that the values of diagonal element corresponding to minimums of errors in both cases are almost identical (Fig. 3).

Furthermore, performing validation on most of images allows us to judge about the generalization ability of the obtained decision rule just right on the validation set. Therefore, among

all decision rules we can choose the one that potentially has the best generalization ability with the lowest validation error, i.e., the average number of errors on the validation set.

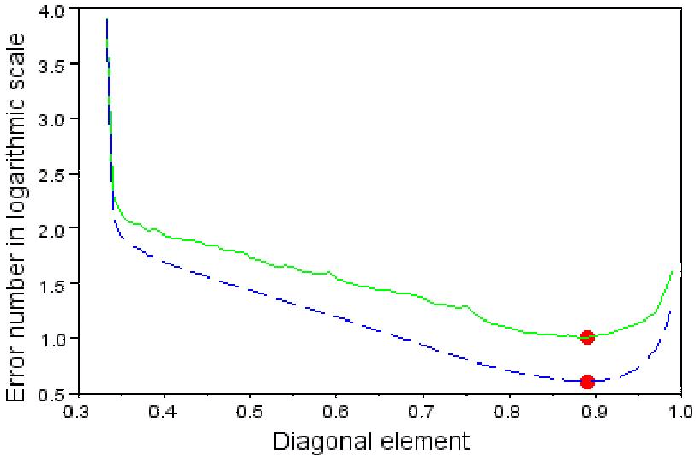

Fig. 3. Lines of recognition errors: solid for single image, dashed for a set of images

We can create the simplified scheme of cross-validation for estimating the general error as follows:

Algorithm 1 The simplified scheme of cross-validation

Input: A set of N images.

Output: General error.

1: **Exclude a testing image :**
Alternately exclude the j^{th} image $1 \leqslant j \leqslant N$ from the given image set and treat it as a test.

2: **Perform non-classical cross-validation on the remaining images:**

3: - The other $N-1$ images form the training set.

4: - Alternately select the k^{th} image from the training set ($1 \leqslant k \leqslant N-1$) and train parameters θ_k on the k^{th} image.

5: - The remaining $N-2$ images are used for validation. Using parameters θ_k, estimate the validation error $ValErr_k$ as the average value of recogntion errors on the validation set.

6: - Define the set of parameters corresponding to the lowest validation error:
$$\hat{\theta}_{opt} = \underset{\theta_k, k=1,\ldots,N-1}{\arg\min} \; ValErr_k.$$

7: **Recognize the test image:**
Estimate recognition error $TestErr_j$ on the test image using parameters $\hat{\theta}_{opt}$.

8: **Repeat:**
Return to step 1 until all images are tested once.

9: **Calculate generate error:**
Estimate the average value of test errors. This value is considered as the recognition error on the general set, i.e., general error, which is calculated as follows:
$$GeneralError = \frac{1}{N} \sum_{j=1}^{N} TestErr_j.$$

10: **Output:**
Return $GeneralError$ as General error.

The simplified scheme of cross-validation is illustrated in Fig. 4. In such a scheme the cross-validation error is defined by averaging all validation errors obtained by different sets of parameters. However, the test error is evaluated using the best set of parameters among them. Therefore, the general error should be less than the cross-validation error.

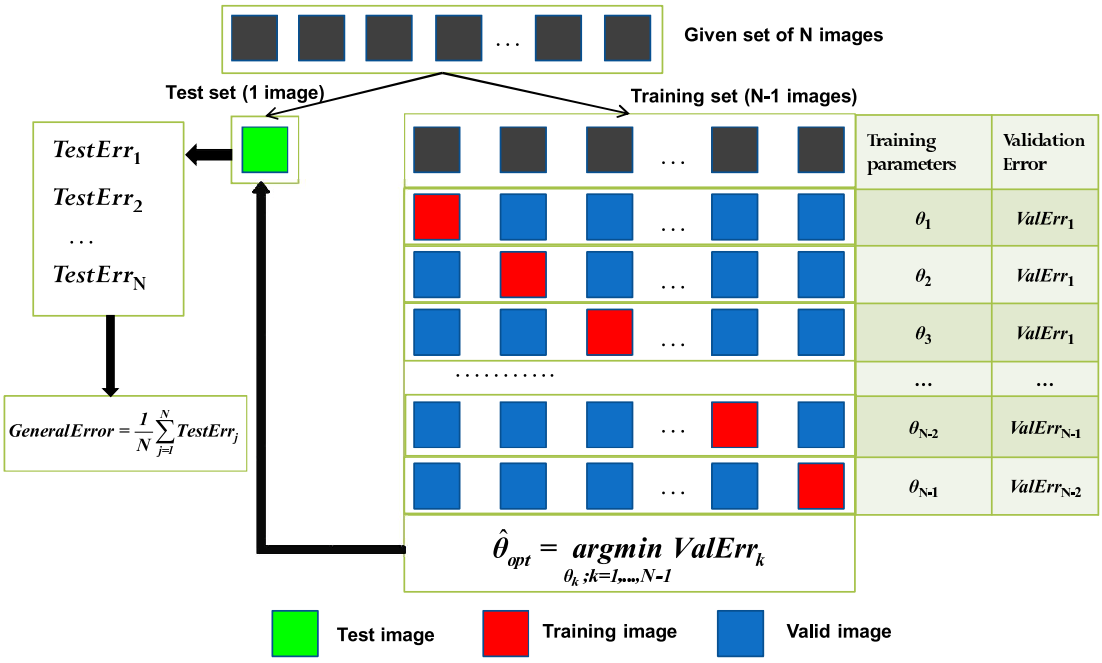

Fig. 4. The simplified scheme of cross-validation

To get the optimal parameters only, we can build a procedure without tests. The cross-validation is performed throughout a given image set and we just need to select parameters corresponding to the least validation error.

Experiments

Average cross-validation and general errors are shown in Fig. 5. Overall time for the simplified cross-validation scheme of each algorithm is shown in Fig. 6. In comparing to others, the Weight fitting algorithm based on the Gauss-Seidel method performs the simplified cross-validation scheme very fast, but it has the highest CV and general errors.

The algorithms A1, A2, A3 for simultaneously adjusting the diagonal element and graph weights are much more slower than the Weight fitting algorithm in performing the CV scheme. However, the algorithms A1, A2, A3 considerably improve the quality of solutions with lower CV and general errors.

The algorithm A1 with unique diagonal element performs the CV scheme with a little faster than the algorithms A2 and A3 with multiple diagonal elements, while having higher CV and general errors.

The algorithm A4 performs the CV scheme very fast with almost the same speed as the Weight fitting algorithm while maintaining a high quality of recognition. Especially, the algorithm A4 obtains the lowest CV error and the lowest difference between the CV and general errors. It means the algorithm A4 is the best one in this paper in dealing with the overfitting problem.

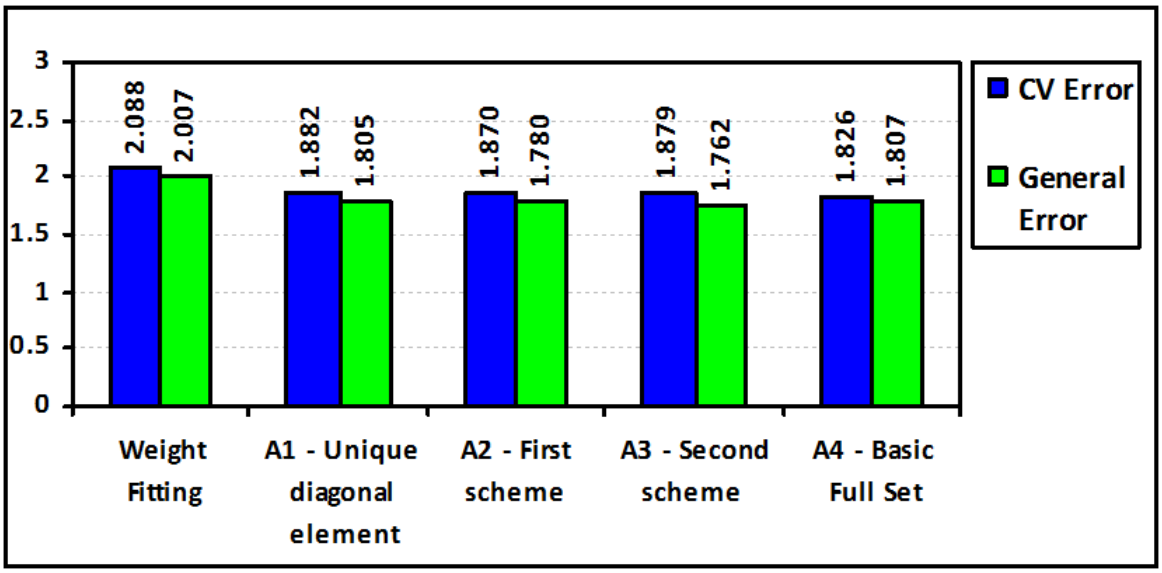

Fig. 5. Recognition errors

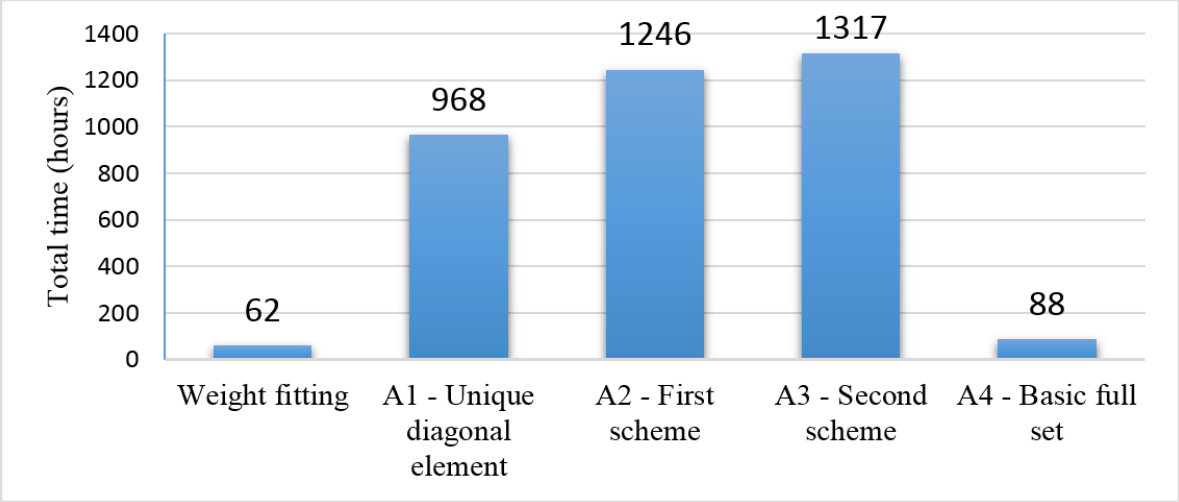

Fig. 6. Overall time for the simplified cross-validation scheme

Conclusion

In this paper we propose the simplified cross-validation scheme with reduced time complexity in adjusting hyper-parameters. This procedure evaluates the statistical properties of decision rules and determines the optimal combination of acyclic adjacency graphs with Markov parameters treated as hyper-parameters.

In view of the large volume of the real data it is very promising to use the parallel scheme of the set processing. This scheme is based on the idea of the countable (full) set of acyclic graphs to represent the real arbitrary ones (lattices for raster images). The experiments show the high statistical quality of processing results.

References

[1] *Wainwright M. J., Jordan M. I.* Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning. 2008. V. 1. P. 1–305.*

[2] *Szeliski R., Zabih R., Scharstein D., Veksler O., Kolmogorov V., Agarwala A., Tappen M, Rother C.* A comparative study of energy minimization methods for markov random fields with smoothness-based priors. *IEEE Trans. PAMI. 2008. V. 6, No. 6. P. 1068–1080.*

[3] *Dvoenko S. D., Kopylov A. V., Mottl V. V.* The Problem of Pattern Recognition in Arrays of Interconnected Objects. Statement of the Recognition Problem and Basic Assumptions. *Automat. Remote Control. 2004. V. 65, No. 1. P. 127–141.*

[4] *Dvoenko S. D., Kopylov A. V., Mottl V. V.* A Problem of Pattern Recognition in Arrays of Interconnected Objects. Recognition Algorithm. *Automat. Remote Control. 2005. V. 66, No. 12. P. 2019–2032.*

[5] *Dvoenko S. D.* Recognition of dependent objects based on acyclic Markov models. *Pattern Recognition and Image Analysis. 2012. V. 22, No. 1. P. 28–38.*

[6] *Dvoenko S. D., Sang D. V.* Recognition of raster textured images based on parametric acyclic Markov models. *In Proc. 22-th Graphicon (Moscow, 2012). M.: Maks Press, 2012. P. 139–143.*

[7] *Kolmogorov V.* Convergent Tree-Reweighted Message Passing for Energy Minimization. *IEEE Trans. PAMI. 2006. V. 28, No. 10. P. 1568–1583.*