STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **8**, March 2020, pp 296–303.
Published online in International Academic Press (www.IAPress.org)

Probability Model Based on Cluster Analysis to Classify Sequences of Observations for Small Training Sets

Sergey S. Yulin*, Irina N. Palamar

Rybinsk State Aviation Technical University, Russia

Abstract The problem of recognizing patterns, when there are few training data available, is particularly relevant and arises in cases when collection of training data is expensive or essentially impossible. The work proposes a new probability model MC&CL (Markov Chain and Clusters) based on a combination of markov chain and algorithm of clustering (self-organizing map of Kohonen, k-means method), to solve a problem of classifying sequences of observations, when the amount of training dataset is low. An original experimental comparison is made between the developed model (MC&CL) and a number of the other popular models to classify sequences: HMM (Hidden Markov Model), HCRF (Hidden Conditional Random Fields), LSTM (Long Short-Term Memory), kNN+DTW (k-Nearest Neighbors algorithm + Dynamic Time Warping algorithm). A comparison is made using synthetic random sequences, generated from the hidden markov model, with noise added to training specimens. The best accuracy of classifying the suggested model is shown, as compared to those under review, when the amount of training data is low.

Keywords Sequence Classification, Markov Chain, Hidden Markov Model, Self-Organizing Map, K-Means, Probability Graphical Model.

AMS 2010 subject classifications 60J10,65C40,40B05

DOI: 10.19139/soic-2310-5070-690

1. Introduction

Sequences of observations are classified in the process of solving the problems of recognizing: speech [1, 2], hand-written text [3], gestures of hands/head [4, 5], states of technical objects [6, 7, 8] Due to intense introduction of computer-aided learning into various areas of human activities, machine learning engineers often have to deal with small-scale training sets, which structure and characteristics are almost unknown. To classify the sequences of observations, the following machine learning methods have widely been used: Hidden Markov Model (HMM), Hidden Conditional Random Fields (HCRF), Long Short-Term Memory (LSTM), k-Nearest Neighbors algorithm (kNN) with Dynamic Time Warping algorithm (DTW).

kNN method is a popular metric non-parametric algorithm of classification. It is based on computing a distance between test specimen and specimens from the training set. Several studies on applying kNN-method together with DTW algorithm were undertaken by Professor Eamonn Keogh and his colleagues. These studies have shown that kNN+DTW method demonstrates the best results to classify one-dimensional sequences of observations [9].

LSTM method is a variation of architecture of recurrent neural networks, developed to classify time sequences. There are works, where the superiority of LSTM neural network over the other machine learning algorithms is shown in the problems of recognizing hand-written text [10] and speech [11].

HMM and HCRF methods are probability models with the concept of states [12, 13, 14, 15]. The basic idea

*Correspondence to: Sergey S. Yulin (Email: julin.serg@gmail.com), Irina N. Palamar(inpalamar@rsatu.ru). Rybinsk State Aviation Technical University, 152934, Yaroslavl region, Rybinsk, st. Pushkin, 53.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

of such models is that, at each instant of time, the system (process) under review is in a state out of finite set of states, and, as time passes, a transition from one state to the other takes place based on markov assumption about conditional independence. HMM is a generative model that requires assessing parameters of distributions of probabilities of the observed data, which become available in each state [16, 17]. HCRF is a discriminative model that requires assessing parameters of separating hyperplane between the observed data of various classes in each state [18, 19, 20].

As of now, there are no publicly available experimental data about the quality of classification, demonstrated by the above models, when the amount of training data is low (up to 100 specimens). Hence, the objective of this work is to examine the behavior of the above-listed methods in solving the problem of classification, when the amount of training set is low, and to develop a new probability MC&CL (Markov Chain and CLusters) model, which remedies the shortcomings of the reviewed models. Synthetic random sequences are used as experimental data: training and test sequences of observations are generated through sampling from hidden markov model with Gaussian probability-density function. Adding noise to training sets breaks training and test sets down.

MC&CL (Markov Chain and CLusters) method implies development and modification of probability model, we have previously developed, which is based on markov chain and self-organizing map of Kohonen/Growing neural gas [21, 22]. As distinguished from the previous works, the suggested method is generalized for the condition of random algorithm for clustering, and tailored to solve the problem of classification, when the amount of training data is low.

Low amount of training data in machine learning is dangerous, because it doesnt allow to form a statistically significant representation of the research subject since, with few experiments performed, it is not possible to separate noise components from useful information about the object. This work shows that the developed MC&CL model effectively solves the problem of classification, when there are few training data, through the efficient leveling of noise components in the training set.

The following two problems may cause low accuracy of classification, when there are few training data:

1) multiple free parameters [23];

2) lack of parameters of distributing the observed data in the model [24, 25, 26].

kNN+DTW model is non-parametrical, and, it might actually be said that it has very many free parameters, if it is assumed that each specimen from training set is considered to be a parameter. Hence, kNN+DTW model is hardly suitable for classifying, when the set is small, since it corresponds to problem 1.

HCRF and LSTM models are discriminative, and they do not assess parameters of distributing the observed data. Therefore, HCRF and LSTM models are not really suitable for classifying, when the set is small, since they correspond to problem 2.

HMM model has few parameters and requires assessment of parameters of distributing the observed data. Thus, it is in line with both items and is suitable for solving the problem of classification, when there are few training data. In this paper, an attempt shall be made to reduce the number of free parameters in HMM model and to alter an algorithm of its training, having suggested thereby a new MC&CL model.

Chapter 2 contains a description of the developed MC&CL model. Chapter 3 contains information about experiments. Chapters 4 and 5 comprise Discussion and Conclusion, respectively.

2. Materials and methods

2.1. Description of probability MC&CL model

It is implied that the sequence of observations might be represented as a set of multidimensional random variables $\bar{x}_t - X = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_t, \ldots, \bar{x}_T\}$, here, a specific sequence $O = \{\bar{o}_1, \bar{o}_2, \ldots, \bar{o}_t, \ldots, \bar{o}_T\}$ is an implementation of these random variables, where $\bar{o}_t = \{o_1, o_2, \ldots, o_i, \ldots, o_n\}$ is the th element of sequence, a vector of attributes that contains n components of o_i.

The structure of HMM model is taken as the basis of the suggested model, but with the difference that the states of the model are explicit rather than hidden. The sequence of observations shall be broken down to clusters, using any known algorithm for clustering (SOM, k-means). Here, the number of cluster shall be equivalent to the number

of the explicit state of the model. Then, the structure of the model shall represent a product of distributions of probabilities of random variables of two types: observed data and numbers of clusters. A joint distribution of probabilities of random variables shall be a product of two conditional distributions of probabilities:

$$p(\bar{h}, X) = \prod_{t=1}^{T} p(h_t \mid h_{t-1})\, p(\bar{x}_t \mid h_t) \tag{1}$$

where $\bar{x}_t$ – random variable, corresponding to the t^{th} element of sequence of observations;

h_t – random variable, corresponding to the number of cluster that corresponds to the t^{th} sampling of sequence;

T – length of random sequence of observations.

Distribution of probabilities $p(\bar{x}_t \mid h_t)$ shall be specified as a multidimensional Gaussian distribution with a scalar value of dispersion. Here, the dispersion shall be considered equal for all distributions of product (1). Hence, the probability of observing element $\bar{o}_t$ of sequence in cluster with number ν shall be computed as

$$p(\bar{x}_t = \bar{o}_t \mid h_t = \nu) = \left(\frac{\beta}{2 \cdot \pi}\right)^{n/2} \cdot \exp\left\{-\frac{\beta}{2} \cdot \|\bar{c}_\nu - \bar{o}_t\|^2\right\}, \tag{2}$$

where n – size of attribute space;

β – distribution parameter (non-assessable);

$\bar{c}_\nu$ – value of the center of cluster with number;

$\bar{o}_t$ – t^{th} element of sequence of observations.

Distribution of probabilities $p(h_t \mid h_{t-1})$ shall be specified as the distribution of markov chain with regularization in the form of adding Dirichlet distribution. Then, probabilities of transitions between clusters shall be computed as:

$$p(h_t = j \mid h_{t-1} = i) = \frac{a_{i,j} + \xi_j}{\sum\limits_{\nu=1}^{N} a_{i,\nu} + \sum\limits_{\nu=1}^{N} \xi_\nu}, \tag{3}$$

where $a_{i,j}$ – number of transitions from cluster with number i to cluster with number j, $i = 0 \ldots N$, $j = 1 \ldots N$ (N – number of clusters);

ξ_j – Dirichlet distribution parameter.

Assessment of parameters of distributing probabilities in markov chain, carried out according to the method of maximum likelihood, is biased. When the amount of set increases, the bias of assessment is eliminated. Since there are few training data, an assessment shall be made pursuant to the maximum a posteriori method. Distribution of Dirichlet is a distribution a priori conjugated to the distribution of probabilities of transitions in markov chain. Adding of a priori conjugated distribution fulfills a function of regularization, not allowing the model to re-train, when there are few training data.

2.2. Training of MC&CL model

Iterative method of learning includes three stages (figure 1).

1. Select number of clusters N. Initialize centers of clusters. Select parameters ξ and β.
2. Perform clustering and assess, thereby, the values of the centers of clusters $C = \{\bar{c}_j\},\ j = 1 \ldots N$.
3. Compute probabilities of transitions $A = \{a_{i,j}\}$ between the centers of clusters:

- for the first element of the sequence ($t = 1$)

$$a_{0j} := a_{0j} + 1, \quad \text{if} \quad j = \nu_1 = \arg\min_{u=1\ldots N} \|\bar{c}_u - \bar{o}_1\|^2; \tag{4}$$

- for the other elements of the sequence ($t \neq 1$)

$$\begin{aligned} a_{0j} := a_{0j} + 1, \quad \text{if} \qquad i &= \nu_{t-1} = \arg\min_{u=1\ldots N} \|\bar{c}_u - \bar{o}_{t-1}\|^2 \\ j &= \nu_t = \arg\min_{u=1\ldots N} \|\bar{c}_u - \bar{o}_t\|^2. \end{aligned} \tag{5}$$

Probabilities of transitions between the centers of clusters shall be computed within all sequences of observations from training set, within those, on which clustering is performed.

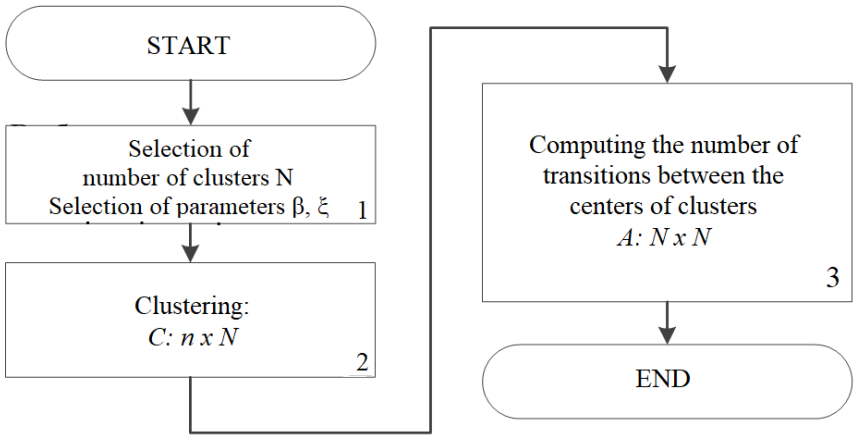

Figure 1. Flow-chart of training MC&CL model

2.3. Classification using MC&CL model

The result of training is a set of models $M_y = (C, A, \beta, \xi)$ for each class y, $y = 1 \dots Y$, where $C = \{\bar{c}_j\}$, $j = 1 \dots N$ – multitude of values of the centers of clusters; $A = \{a_{i,j}\}$ – multitude of values, equal to the number of transitions between the centers of clusters; ξ – Dirichlet distribution parameter; β – distribution parameter $p(\bar{x}_t \mid h_t)$. Parameters ξ and β shall not be assessed in the process of training, but selected empirically within a testing set.
A decision procedure, i.e. the procedure for relating data to a particular class, shall be formed through calculating likelihood of model $M_y = (C, A, \beta, \xi)$ of sequence O for all classes $y = 1 \dots Y$. Then, the model with the maximum value of likelihood shall be the model of the class, to which sequence

$$y^* = \arg\max_{y=1\dots Y} p\left(X = O \mid M_y = (C, A, \beta, \xi)\right). \tag{6}$$

belongs. Likelihood of model $M_y = (C, A, \beta, \xi)$ of sequence O is computed by summing up throughout all possible states of the model

$$\begin{aligned} p\left(X = O \mid M_y\right) = & \sum_{\forall \nu_1 \in h, \dots \forall \nu_T \in h} \lambda_{0,\nu_1} \cdot \mu_{\nu_1}(\bar{o}_1) \cdot \ldots \cdot \lambda_{0,\nu_T} \cdot \mu_{\nu_T}(\bar{o}_T), \\ & \mu_j(\bar{o}_t) = \left(\tfrac{\beta}{2 \cdot \pi}\right)^{n/2} \cdot \exp\left\{-\tfrac{\beta}{2} \cdot \|\bar{o}_t - \bar{c}_j\|^2\right\}, \\ & \lambda_{i,j} = \frac{a_{i,j} + \xi_j}{\sum\limits_{\nu=1}^{N} a_{i,\nu} + \sum\limits_{\nu=1}^{N} \xi_\nu}. \end{aligned} \tag{7}$$

where $\mu_j(\bar{o}_t)$ – probability of observing t^{th} element $\bar{o}_t$ of sequence in cluster j;
$\lambda_{i,j}$ – probability of transiting from cluster i to cluster j.

3. Experimental results

Random synthetic dataset shall be generated, using software package pmtk3 [27], to carry out a comparative assessment of classification quality. Sequences of observations for training and test datasets shall be generated from hidden markov model with random parameters of distribution. Gaussian noise (SNR = 0.1 dB) shall be added to training specimens so as to simulate differences between training and test data.
2 functions shall be used from package pmtk3: *mkRndGaussHmm(. . .)* – creation of random hidden markov model with Gaussian function of distributing probabilities of observing data in each of the hidden states of the model; *hmmSample(. . .)* – generation of the sequence of observations with the specified parameters. Gaussian noise is added to data through function *awgn()*. Parameters of generated sequences of observations are shown in Table 1. Software code on Matlab/Octave language of generating test and training sequences is presented in Listing 1.

```
1
2 -   states_number   = 6;
3 -   class_number = 5;
4 -   dimension       = 15;
5 -   timeseries_length    = 50;
6 -   nsamples  = 200;
7 -   dataTrain = cell(size(dimension,2),nsamples);
8 -   dataTest = cell(size(dimension,2),nsamples);
9
10 -  for i=1:class_number
11 -     hmmSource = mkRndGaussHmm(states_number, dimension);
12 -     [data1, ~]  = hmmSample(hmmSource, timeseries_length, nsamples);
13 -     [data2, ~]  = hmmSample(hmmSource, timeseries_length, nsamples);
14 -     dataTrain(i,:) = data1';
15 -     dataTest(i,:) = data2';
16 -  end
17
18 -  for i=1:size(dataTrain,1)
19 -      for j=1:size(dataTrain,2)
20 -          matrixT = dataTrain{i,j};
21 -          for k=1:size(matrixT,1)
22 -             matrixT(k,:) = awgn( matrixT(k,:), 0.1, 'measured');
23 -          end
24 -          dataTrain{i,j} = matrixT;
25 -      end
26 -  end
```

Listing 1. Generation of training and test sets

A comparative analysis of dependence of the quality of classification on the size of training dataset shall be made on the following models:

- kNN+DTW. Source: standard package Matlab; value k=1;
- LSTM. Source: standard package Matlab; number of sequences, trained jointly (batch size): 24;
- HMM. Source: package pmtk3; type of distributing probabilities: continuous Gaussian distribution of probabilities of observations that become available in the hidden states; initialization method: k-means; number of states: 6;
- HCRF. Source: package hcrf [28]; number of states: 6; method of regularization: L2; optimization algorithm: BFGS;
- MC&CL. Source: repository with initial codes [29], number of states (number of clusters): 6; MC&CL(SOM) – method of self-organizing Kohonen maps is used as an algorithm for clustering; (SOM); MC&CL(k-means) – k-means method is used as an algorithm for clustering).

To assess the quality of classification, an error in classification shall be used, which is computed as a portion of wrong answers: the number of mismatches between the answer of classifier and actual class mark, divided by the total number of answers.

Table 1. Parameters of generated sequences of observations

Name	Value
Number of states	6
Size of attribute space	15
Length of sequence	50
Number of sequences in a set	200
Number of classes	5

A dependence of the quality of classifying on the size of training set for the above-stated models is presented in Fig. 2 and Table 1. Training is made sequentially on 20, 30, 40, and so on, up to 110 specimens from the training set. Testing is always made on 200 specimens from testing dataset.

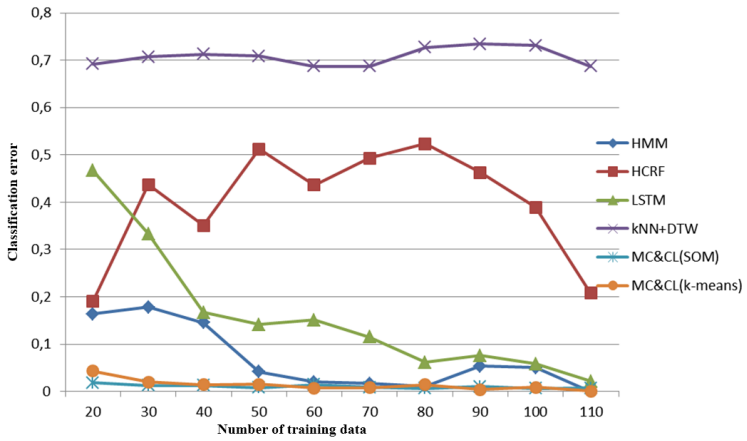

Figure 2. Dependence of quality of classification on the size of training set

Table 2. Error in classifying MC&CL (SOM) and MC&CL (k-means) models with variable amounts of training data

Size of training set	20	30	40	60	70	80	90	100	110
MC&CL (SOM)	0.018	0.013	0.013	0.008	0.009	0.018	0.010	0.006	0.007
MC&CL (k-means)	0.043	0.020	0.014	0.015	0.008	0.014	0.004	0.009	0.001

4. Discussion

After analyzing figure 2, it can be concluded that the suggested MC&CL model demonstrates the best result in classifying, when there are few training data (dozens of specimens). This is due to the fact that MC&CL model has fewer free parameters than the other models that hinders recovery of noise components from the data. From analyzing Table 2, it can be inferred that selection of clustering algorithm (SOM or k-means) has a minor effect on the final error in classification. The closest to MC&CL model result is shown by HMM model. As it was supposed, in the beginning of work, HMM model was successfully outweighed applying few training data, due to decreasing the number of free parameters in the model, particularly, the number of parameters in the probability-density function of distributing the observed data:

1. HMM model. Distribution of data, observed in each state of the model, shall be specified by multi-dimensional Gaussian function with parameters $\bar{\mu}$ – vector of mathematical expectations and Σ – covariance

matrix

$$p(\bar{x}) = \frac{1}{(2\cdot\pi)^{n/2}\cdot|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}\cdot(\bar{x}-\bar{\mu})^T\cdot\Sigma^{-1}\cdot(\bar{x}-\bar{\mu})\right), \tag{8}$$

where n – size of attribute space.
Number of parameters: $n + \frac{n\cdot(n+1)}{2}$.

2. MC&CL model. Distribution of data, observed in each state of the model, shall be specified by multi-dimensional Gaussian function with parameters $\bar{c}$ – vector of mathematical expectations (centers of clusters) and β^{-1} – scalar value of dispersion.
3.

$$p(\bar{x}) = \left(\frac{\beta}{2\cdot\pi}\right)^{n/2}\cdot\exp\left\{-\frac{\beta}{2}\cdot\|\bar{c}-\bar{x}_t\|^2\right\}. \tag{9}$$

Number of parameters: $n + 1$.

5. Conclusion

The present paper contains a description of MC&CL probability model, we have developed, generalized for the condition of random algorithm of clustering, and the original results of comparative analysis between MC&CL model and HMM, HCRF, LSTM, kNN+DTW models that use synthetic data, generated on the basis of hidden markov model with noise added to training specimens.
It is shown that the developed MC&CL model, successfully solves the problem of classification, when the amount of training data is low, through efficient leveling of noise components within the training set. Such a model may be relevant to solve the problem of classification, when it is impossible to form a large training set, since there are financial or time restrictions, and due to the fact that the occurring phenomena as such (anomalies) are rare.

REFERENCES

1. J. R. Rohlicek, W. Russell, S. Roukod, and H. Gish. *Continuous hidden markov model for speaker independent word spotting*, International Conference on Audio, Speech and Signal Processing, vol. 1, pp. 627–630, 1989
2. J. G. Wilpon, L.R. Rabiner, and C. Lee. *Automatic recognition of keywords in unconstrained speech using hidden Markov models*, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, no. 11, pp. 1870–1878, 1990.
3. B. H. Williams, M. Toussaint, A. J. Storkey. *A primitive based generative model to infer timing information in unpartitioned handwriting data*, IJCAI, vol. 2, pp. 1119–1124, 2007.
4. M. Elmezain, A. Al-Hamadi, and M. Bernd. *A Hidden Markov Model-Based Isolated and Meaningful Hand Gesture Recognition*, International Journal of Electrical & Electronics Engineering, vol. 3, iss. 4, pp. 156–163, 2009.
5. S. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrell. *Hidden Conditional Random Fields for Gesture Recognition*, In Conference on Computer Vision and Pattern Recognition (CVPR), 2006.
6. V. Chandola, A. Banerjee, and V. Kumar. *Anomaly Detection: A Survey: Technical Report*, Minneapolis, Department of Computer Science and Engineering University of Minnesota, 2007.
7. E. Khalastchi, G. A. Kaminka, M. Kalech, and R. Lin. *Online anomaly detection in unmanned vehicles*, In The 10th International Conference on Autonomous Agents and Multiagent Systems, vol. 1, pp. 115–122, 2011.
8. S. Hansen. *Fault Diagnosis and Fault Handling for Autonomous Aircraft*, Ph.D. dissertation, Technical University of Denmark, Department of Electrical Engineering, Denmark, 2012.
9. E. Keogh. *UCR Time Series Classification Archive*, URL: http://www.cs.ucr.edu/∼eamonn/time_series_data/
10. A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. *Novel Connectionist System for Improved Unconstrained Handwriting Recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence. vol.31, no. 5, 2009.
11. A. Graves, Abdel-rahman Mohamed, and G. Hinton. *Speech Recognition with Deep Recurrent Neural Networks*, Acoustics, Speech and Signal Processing (ICASSP) IEEE International Conference, pp. 6645–6649, 2013
12. D. Koller, and N. Friedman. *Probabilistic Graphical Models*, Massachusetts, MIT Press, 2009.
13. A. B. Merkov. *Recognition of patterns: Introduction to methods of statistical learning*, Moscow, Editorial URSS, 2011.
14. Z. Taushanov, and A. Berchtold, *A Direct Local Search Method and its Application to a Markovian Model*, Statistics, Optimization and Information Computing, vol. 5, no. 1, pp. 19–34, 2017.
15. Sutton, and A. McCallum. *An Introduction to Conditional Random Fields for Relational Learning*, Massachusetts, MIT Press, 2006.

16. L.R. Rabiner. *Hidden Markov models and their use in selected applications when recognizing speech: review*, TIIER, ch. 77(2), pp. 86–120, 1989.
17. R.V. Andreao, B. Dorizzi, and J. Boudy. *ECG signal analysis through hidden Markov models*, Biomedical Engineering, IEEE Transactions, vol. 53, iss. 8, pp. 1541–1549, 2006.
18. A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. *Hidden Conditional Random Fields for Phone Classification*, Interspeech, pp. 1117–1120, 2005.
19. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell. *Hidden-state Conditional Random Fields*, IEEE PAMI, 2007.
20. S. Wang, A. Quattoni, L. P. Morency, D. Demirdjian, and T. Darrell. *Hidden Conditional Random Fields for Gesture Recognition*, In Conference on Computer Vision and Pattern Recognition (CVPR), 2006.
21. I.N. Palamar, and S.S.Yulin. *Generative probabilistic graphical model based on self-organizing map*, Proceedings of SPIIRAN, Saint-Petersburg, no. 2, pp. 227–247, 2014.
22. N. Palamar, and S. S. Yulin. *Probabilistic Graphical Model Based on Growing Neural Gas for Long Time Series Classification*, Modern Applied Science, Canada (Toronto), vol.9, no 2, pp. 109–116, 2015.
23. V.P. Vapnik. *Recovery of empirical data dependencies*, Moscow, Nauka, 1979.
24. A. Ng, and M. Jordan. *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes*, In Advances in Neural Information Processing Systems 14, pp. 841–848, 2002.
25. J.-H. Xue, and D.M. Titterington. *Comment on ¡¡discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes¿¿*, Neural Processing Letters, vol. 28, iss. 3, pp. 169–187, 2008.
26. P. Liang, and M. I. Jordan. *An asymptotic analysis of generative, discriminative, and pseudo-likelihood estimators*, In Proceedings of the 25th International Conference on Machine Learning (ICML), 2008.
27. GitHub. *Probabilistic Modeling Toolkit for Matlab/Octave, [Online resource], 2010.* URL: https://github.com/probml/pmtk3.
28. SourceForge. *HCRF library (including CRF and LDCRF) [Online resource], 2011.* URL: https://sourceforge.net/projects/hcrf/.
29. S. Julin. Bitbucket. *PhD Codesource [Online resource], 2015.* URL: https://bitbucket.org/sjulin/phd_enterprisecode.