Modern Applied Science; Vol. 9, No. 2; 2015
ISSN 1913-1844 E-ISSN 1913-1852
Published by Canadian Center of Science and Education

Probabilistic Graphical Model Based on Growing Neural Gas for Long Time Series Classification

Irina N. Palamar[1] & Sergey S. Yulin[1]

[1] Soloviev Rybinsk State Aviation Technical University, Rybinsk, Russia

Correspondence: Sergey S. Yulin, Soloviev Rybinsk State Aviation Technical University, Rybinsk, Russia.

Received: September 10, 2014 Accepted: October 16, 2014 Online Published: December 11, 2014

doi:10.5539/mas.v9n2p108 URL: http://dx.doi.org/10.5539/mas.v9n2p108

Abstract

This article proposes a generative probabilistic graphical model with hidden states (Neural Gas Graphical Model (NGGM)) based on data approximation with a grid of "neural gas" nodes aimed at solving the problem of long time series classification. Such time series include information about changes in economic, weather and health values, as well as information about changes in values of operation sensors of technical objects during a quite long period. The most difficult task of classification of such long time series using probabilistic graphical models with hidden states is the selection of the optimum number of hidden states. This work proposes a method for automatic selection of the optimum number of hidden states of the model in the course of the model learning. The model proposed in the article and the methods of its learning are based on a combination of elements used in the metric and Bayesian approaches to classification. The basic NGGM purpose is to match hidden states of a graphical model and nodes (neurons) of the approximating grid. Comparative assessment of the quality of the proposed NGGM model classification with the currently most common time series classification models has been made: the HMM (Hidden Markov Model) and the HCRF (Hidden Conditional Random Fields) applied at the data sets from the UCI repository. The quality was assessed by the macro-average F-measure criterion using the k-fold cross-validation. As a result of classification quality analysis, it was noted that the proposed NGGM model showed better classification quality on the data set being a set of multiple, labeled samples of pen tip trajectories recorded whilst writing individual characters than the HCRF and HMM models.

Keywords: probabilistic graphical model, generative model, long time series classification, growing neural gas, HMM, HCRF

1. Introduction

Probabilistic graphical models with hidden states are used for time series classification (Koller and Friedman, 2009). For solving the classification problem, the machine learning methods based on probabilistic graphical models such as HMM (Rabiner, 1989) and HCRF (Gunawardana et al., 2005) have proven to be efficient. These models imply either availability of a priori information about the number of hidden states or experimental selection of their optimal number (Soullard and Artieres, 2011). In case of the long time series representing information about a process lasting during a sufficiently long period, there is no a priori information about the required number of hidden states. Such processes include changes of economic, weather, and health indexes over time, and changes in readings of technical objects sensors (Grabocka et al., 2012). For classification of such time series, the criteria for selection of the optimum number of hidden states are most relevant, since a random increase in the number of hidden states results in a considerable increase in model parameters, which in turn results of overfitting.

The problem of selection of the optimum number of parameters of probabilistic graphical models is an important theoretical problem. The most popular criteria for selecting the number of hidden states are AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), ICL (Integrated Completed Likelihood), and PML (Penalised Marginal Likelihood). They are used for selection of the number of hidden states in HMM and are based on maximum likelihood (ML) estimation of the models (Celeux and Durand, 2008). A well-known difficulty is non-regularity in their maximum likelihood estimators, under which classical information criteria lose their theoretical justifications. Roughly speaking, Fisher information matrices around the ML estimators are singular, and thus an asymptotic second order approximation is not applicable (Fujimaki and Hayashi, 2012). The most successful solution for classification of long time series with selecting the optimal number of hidden

states is the concept of the infinite Hidden Markov Models (iHMM) (Beal et al., 2002). In order to describe the infinite number of iHMM states, the hierarchical Dirichlet process is used, and the number of states is determined based on three predefined hyper-parameters. The necessity of setting these hyper-parameters does not allow automating fully the process of selecting the optimum number of states.

Thus, the objective of this work is the development of a probabilistic graphical model for classification of long time series and automatic evaluation of the optimum number of model's hidden states during learning.

In this article, we propose a probabilistic graphical model based on approximation of training data with a grid of nodes built according to the algorithm of growing "neural gas". The algorithm of growing "neural gas" (11 Martinetz and Schulten, 1991) refers to a broad class of methods for principal manifolds building (Gorban et al., 2008). The main idea of the proposed graphical model is to set up a correspondence between hidden states and nodes (neurons) of the approximating grid. At that, observation conformity to a particular hidden state is assessed by calculating the distance between them and by the subsequent normalization in the range between 0 and 1. The advantage of the growing "neural gas" algorithm is the possibility to find the optimal number of the approximating grid nodes based on the characteristics of the data itself, and the main problem is to define the criterion for stopping the increase in the number of grid nodes. While resolving the problem of clustering, the problem is solved by using the criterion of calculating the local neuron error or by calculating the total clustering error (Holmstrom, 2002). We propose finding the optimal number of grid nodes based on the classification quality assessment by the F-measure indicator of the training and validation sampling after each new node to the grid is added.0

We will compare the proposed model to the HMM (PMTK, 2010) and HCRF (HCRFLIB, 2010) models, and evaluate the classification quality using the k-fold cross-validation method and counting the F-measure criterion averaged over all classes on data sets from the UCI repository (UCI Arabic Digit, 2010; UCI Character, 2008).

2. Materials and Methods

2.1 Neural Gas Graphical Model

The model proposed in the work and the methods for learning it are based on a combination of elements used in the metric and Bayesian approaches to classification. Let us call it the Neural Gas Graphical Model (NGGM). Some parameters of this model are the approximating grid nodes' values. Such a set is typically unique for each class, whereby it is possible to perform classification on its basis by assessing the distance between the data being classified and the grid nodes. The distance normalized in the range between 0 and 1 between the observed data and each grid node may appear as an analog to the probability of occurrence of the data observed in hidden states and be used in algorithms of probabilistic inference. The measure for calculating the distance is chosen based on the data singularities. In this work, we consider only the Euclidean distance, which is calculated as the L_2-norm of the difference vector between two points in Euclidean space. The required number of hidden states (grid nodes) in the proposed model is evaluated after each iteration of the growing "neural gas" algorithm, i.e. after adding a new node. The evaluation resides in calculating the classification quality of a trained model with the current number of nodes on the training and validation sets.

The NGGM structure is shown in Figure 1. White dots designate the observed data and black dots designate the nodes of the approximating grid (neurons).

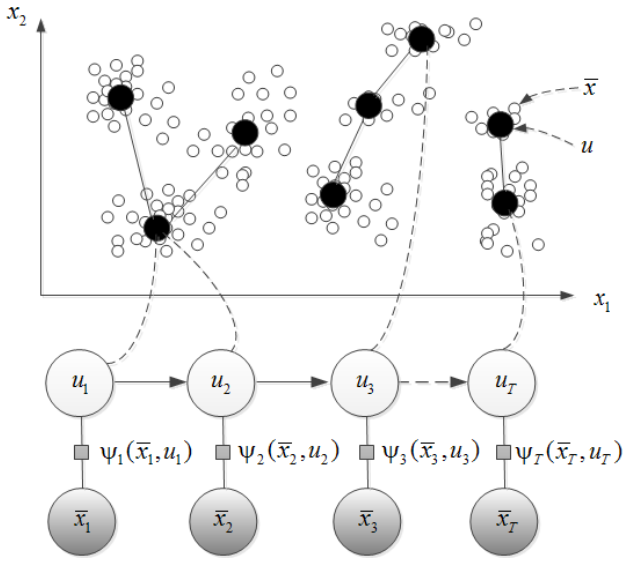

Figure 1. The NGGM structure as a probabilistic factor graph

Joint distribution of NGGM variables is represented as a product of factors described with the following expression

$$p(\bar{u},X)=\prod_{t=1}^{T}p(u_t\mid u_{t-1})\cdot\Psi(u_t,\bar{x}_t),$$

Where u_t is the number of node (neuron) of the approximating grid, taking values between 1 and N, where N is the number of nodes;

$\bar{x}_t$ is the data observed at a moment t;

t is the number of current observation;

T is length of the observed time series;

$X=\{\bar{x}_1,\bar{x}_2...,\bar{x}_T\}$ is the observed time series;

$p(u_t\mid u_{t-1})$ is the distribution of transition probabilities between states of the model (between nodes of the approximating grid);

$\Psi(u_t,\bar{x}_t)$ is the normalized factor function defining the relationship between the current node of approximating grid and current observation.

The optimum value of hidden states ("neural gas" grid nodes) is selected as

$$\begin{aligned} y^* &= \arg\max_{y=1..C}(L(X\mid M_y=\{W,A,N\})),\\ N^* &= \arg\max_{N}\frac{\sum_{X\in Z\cup J}Fmera(y,y^*)}{|Z\cup J|}, \end{aligned} \tag{1}$$

Where $Z=\{X_1^{(y)},X_2^{(y)},...,X_{m1}^{(y)}\}$ is the training set with cardinality $m1$;

$J=\{X_1^{(y)},X_2^{(y)},...,X_{m2}^{(y)}\}$ is the validation set with cardinality $m2$;

y is the true value of the class, to which the sequence X belongs, $y=1..C$ where C is the number of classes;

y^* is the classifier response about the membership of the sequence X;

$M_y=\{W,A,N\}$ is the model for the class y;

W is the set of values of approximating grid nodes;

A is the matrix of probabilities transition between grid nodes;

L is the log-likelihood function.

$Fmera$ is the function of F-measure, which is a weighted harmonic means of precision and recall of the classifier.

Estimation of distribution $p(u_t \mid u_{t-1})$ of discrete random variable by the method of maximum likelihood is limited to calculating frequency of observation of a particular event. In this case, such an event is a transition from one grid node to another, probability of which is calculated as a_{ij}

$$a_{ij} = \frac{\gamma_{ij}}{\sum_{j=1}^{N} \gamma_{ij}},$$

$$\gamma_{ij} = \begin{cases} \gamma_{ij} + 1, if\ i = u_{t-1} = \\ \quad = \underset{u}{\operatorname{argmin}} \left\| \overline{w} - \overline{x}_{t-1} \right\|_2 \\ \quad and \\ j = u_t = \underset{u}{\operatorname{argmin}} \left\| \overline{w} - \overline{x}_t \right\|_2 \\ \gamma_{ij}, in\ other\ cases \end{cases}, \tag{2}$$

Where γ_{ij} is the number of transitions from one grid node with number i to another one with number j in a single time series, $i, j = 1..N$;

u_t is the number of grid node that corresponds to the vector $\overline{x}_t$.

Let us define the value of the non-normalized factor $\tilde{\Psi}(u_t, \overline{x}_t)$ as

$$\tilde{\Psi}(u_t, \overline{x}_t) = \widetilde{B}(i,t) = \widetilde{b}_{it} = -\left\| \overline{w}_i - \overline{x}_t \right\|_2, \tag{3}$$

Where $\left\| \overline{w}_i - \overline{x}_t \right\|_2$ is the distance between the weight vector $\overline{w}_i$ of the node i and the observed vector $\overline{x}_t$ in the Euclidean space. $t = 1..T, i = 1..N$

For the values of this factor to be used for calculating the likelihood, values of the matrix $\widetilde{B}$ elements should be normalized, i.e., the matrix B satisfying the following two conditions should be formed:

1. $0 < B(i,t) \le 1$; 2. $\sum_{i=1}^{N} B(i,t) = 1$.

Normalization of values of probabilities distribution is a standard procedure in mathematical statistics that is often used to normalize the distribution histograms. It should be noted that the distance between the observation $\overline{x}_t$ and nodes of the map $i, i = 1..N$ is highly irregular in the sense that the value of distance to most nodes of the map will have a relatively large value, and will have a sufficiently small value only for one ("best matching node") or several nodes. In this regard, we use normalization in logarithmic space:

$$\begin{aligned} l_t &= \ln \sum_{i=1}^{N} \exp(\widetilde{B}(i,t)), t = \overline{1,T}, \\ B(i,t) &= \exp(\widetilde{B}(i,t) - l_t), t = \overline{1,T}, i = \overline{1,N}, \end{aligned} \tag{4}$$

where $\overline{l}$ is the vector of logarithmic normalization constants ($\overline{l} : T \times 1$).

2.2 Learning and Inference

Let us consider the steps of model training. The purpose of training is to obtain models of the observed data in each class $M_y = \{W, A, N\}$, containing the optimum number of hidden states N. For the NGGM training, an algorithm has been developed, consisting of the following steps.

Step 1. This step resides in carrying out one iteration of the growing "neural gas" algorithm for obtaining a new node (neuron) of the approximating grid. For each class *y*, a separate approximating grid of nodes is built based on the data set $Z^{(y)}$. As a result of learning, each node u_i will correspond to the weight vector $\overline{w}_i \in W$.

Step 2. This step includes estimation of conditional distribution of probabilities $p(u_t \mid u_{t-1})$ (expression 2).

Step 3. This step resides in calculation of likelihood $L(X \in Z \cup J \mid M_y = \{W, A, N\})$ of the obtained model parameters to the union set $Z \cup J$, which is the union of training and validation sets. In order to calculate the likelihood, we shall use the "forward step" recursive procedure of the "forward-backward" algorithm. Using the expression (3) for calculating distances and expression (4) for their normalization, we obtain values of the factor $\Psi(u_t, \bar{x}_t) = B(i,t) = b_{it}$. Then the log-likelihood of the model M_y of the class y of the observed time series is calculated as

$$L(X \mid M_y = \{W, A, N\}) = \ln \sum_{u_1, u_2 \ldots u_T} b_{11} \cdot a_{12} \cdot$$

$$\cdot b_{22} \cdot \ldots \cdot a_{T-1,T} \cdot b_{TT} + \sum_{t=1}^{T} l_t.$$

Step 4. This step includes classification and evaluation of classification quality by F-measure criterion averaged over all classes based on likelihood values. Classification is made based on the entire set $Z^{(y)} \cup J^{(y)}$ by calculating the expression

$$y^* = \arg \max_{y=1..C} (L(X \in Z \cup J \mid M_y = \{W, A, N\})),$$

Step 5. This step includes evaluation of changes in the F-measure criterion value (expression 1):

- if in the course of several iterations of the growing "neural gas" algorithm the value of the F-measure criterion is consistently reduced or remains unchanged, addition of new nodes (neurons) to the model should be stopped, and this model should be deemed to be an optimum one;

- if in the course of several iterations of the growing "neural gas" algorithm the value of the F-measure criterion increases, we proceed to Step 1 of the training algorithm and continue adding nodes (neurons).

3. Experimental Results

Let us perform comparative evaluation of the HMM and HCRF models, and the “Neural Gas Graphical Model” (NGGM) model proposed in this article at solving the problem of classification using on data sets from the UCI machine learning repository.

The following characteristics of the tested models have been used. Initialization of parameters of the models: HMM – with the k-means algorithm, HCRF – with random values, and NGGM does not require initialization. The learning method: HMM – by the algorithm Baum-Welch, HCRF – by the BFGS quasi-Newton optimization algorithm. Regularization method: HMM – without regularization, HCRF – L_2-regularization. The Gaussian density function was used as the function of density of data distribution in the HMM model.

Evaluation of the L_2-regularization parameter for the HCRF model, the number of hidden states for the HMM and HCRF models was carried out at a validation set which is 10% of the data set.

The “Spoken Arabic Digit Data Set” (UCI Arabic Digit, 2010) is a data set of 8,800 pronounced digits in the Arabic language that are time series of 13 cepstral coefficients. The number of classes is 10. The number of attribute space dimensions is 13. The number of instances of each class is 880. The average time series length T = 45.

The “Character Trajectories Data Set” (UCI Character, 2008) includes trajectories of pen movement obtained at writing down letters of the English alphabet on a Wacom tablet. The data consist of three parameters: the point coordinate on the axis of abscissas, axis of ordinates, and the pen pressure. The data have been smoothed and normalized. The number of classes is 20. The number of attribute space dimensions is 3. The number of instances of each class is 100. The average time series length T = 180.

The classification quality was evaluated according to the k-fold cross-validation method. Experiments were performed at the k value equal to 10. As a measure of quality evaluation, we used the balanced F-measure averaged over all classes.

The results of classification quality evaluation by the method of k-fold cross-validation are shown in Table 1.

Table 1. Results of the classification quality evaluation

	HMM	HCRF	NGGM
Spoken Arabic Digit Data Set (T = 45)			
Parameters	5 states	5 states	260 nodes
F-measure on the training set	0.9313	0.9587	0.9354
F-measure on the test set	0.8778	0.9525	0.9223
Overfitting	0.0535	0.0062	0.0131
Character Trajectories Data Set (T = 180)			
Parameters	7 states	7 states	133 nodes
F-measure on the training set	0.9568	0.9960	0.9999
F-measure on the test set	0.9329	0.9651	0.9838
Overfitting	0.0239	0.0309	0.0161

As a result of classification quality analysis assessment, the following can be noted:

- on the “Character Trajectories Data Set”, the average length of time series in which T = 180, the proposed NGGM model showed classification quality by 1.87% better than the HCRF model and by 5% better than the HMM model;

- the effect of overfitting on the proposed model has not been observed.

Other NGGM’s advantages include the possibility of using various distance functions depending on particular data. The NGGM’s disadvantages include a much larger number of hidden states compared to the HMM and HCRF models, which results in an increased classification execution time.

Based on the results of the work, we can conclude that the developed NGGM model is well suited for classification of long time series and can be used as a basis of a system of technical objects fault diagnostics based on the analysis of sensor readings during a sufficiently long period.

4. Discussion

In this article, the model for classification of long time series based on the theory of probabilistic graphical models and the growing "neural gas" algorithm have been proposed. Currently, there are many works representing hybrid models based on combining probabilistic graphical models and methods of principal manifolds building. For instance, these are the works by (Kurimo, 1997; Somervuo, 2000) that consider HMM together with Kohonen self-organizing maps, which are used to solve the problem of mathematical expectation initialization before estimation of probability distribution parameters p(x) on learning data of the Baum-Welch algorithm. Also other works are known, such as (Calinon and Billard, 2005), which use vector quantization grids as a method of reducing the number of attribute space dimensions for the HMM classifier. Beside continuous HMM with the Gaussian probability density function, there are discrete HMM with discrete distribution and approximation of distribution using vector quantization grids (Neukirchen et al., 2001; Osterndorf and Rohlicek, 1990). However, this approach proved inefficient due to the low quality of classification (worse than that of continuous HMM) and the need for a large amount of training data.

Unlike the above-mentioned works, this article proposes the generating graphic probabilistic NGGM model that does not require evaluating of the observed data occurrence probability and selecting the form of the probability distribution density function (normal, exponential, etc.). This considerably broadens applicability limits of the NGGM model, since not all real-world data correspond to the known parametric families of probability distributions, or may be described in terms of probability (frequency of occurrence). The distance normalized in the range between 0 and 1 distance between the observed data and each grid node is an analog to the probability of occurrence of the data observed in hidden states and is used in algorithms of probabilistic inference.

5. Conclusion

The work presents a generative probabilistic graphical model with hidden states based on data approximation with a grid of "neural gas" nodes for solving the problem of classification of long time series. The probabilistic graphical model structure is described, where each hidden state corresponds to a node of the approximating grid, and the probability of transition to the new node depends only on the current node. An algorithm of model learning has been proposed, which selects the optimum number of hidden states (of the approximating grid nodes) based on evaluation of classification quality through validation sets by the F-measure criterion after each iteration of the “growing neural gas" algorithm. Probabilistic inference for calculation of likelihood is made by

the "forward-backward" algorithm with normalization in logarithmic space of the distance between the data observed and each node of the grid.

Comparative evaluation of classification quality of the proposed NGGM model with the HMM and HCRF models has been made on data sets from the UCI repository. Quality was assessed by F-measure criterion using the k-fold cross-validation. As a result of classification quality evaluation analysis, it was noted that on the "Character Trajectories Data Set" data set, in which the average length of time series is T = 180, the proposed NGGM model showed classification quality by 1.87% better than the HCRF model, and by 5% better than the HMM model.

Based on the results of this work, we conclude that the developed NGGM model is well suited for classification of long time series and can be used as a basis of a system of technical objects (e.g., aviation engines) fault diagnostics based on the analysis of sensors readings (e.g. pressure and temperature sensors) during a sufficiently long period.

Disadvantages of the NGGM include a considerably larger number of hidden states compared to the HMM and HCRF models, which results in an increased classification time. Advantages of the NGGM include the possibility of efficient evaluation of the optimum number of hidden states, the high quality of classification of long time series, and the possibility to use different distance functions depending on specific data.

In the future, in order to calculate the likelihood, instead of the "forward-backward" algorithm that has high computational complexity, we expect to develop our own method based on such sampling methods as the Gibbs sampling and the Markov chain Monte Carlo, which can result in a considerable shortening of classification time without any loss of the classification quality.

References

Beal, M. J., Ghahramani, Z., & Rasmussen, C. E. (2002). The infinite hidden Markov model. Advances in Neural Information Processing Systems 14. In the Proceedings of the 2001 Conference, Vancouver, British Columbia, Canada. The MIT Press, pp. 577-585.

Calinon, S., & Billard, A. (2005). Recognition and Reproduction of Gestures using a Probabilistic Framework combining PCA, ICA and HMM. In the Proceedings of the 22nd International Conference on Machine Learning (ICML), Bonn, Germany. ACM New York, pp. 105-112. http://dx.doi.org/10.1145/1102351.1102365

Celeux, G., & Durand, J. B. (2008). Selecting Hidden Markov Model State Number with Cross-Validated Likelihood. *Computational Statistics, 23*(4), 541-564. http://dx.doi.org/10.1007/s00180-007-0097-1

Fujimaki, R., & Hayashi, K. (2012). Factorized Asymptotic Bayesian Hidden Markov Models. In the Proceedings of the 29th International Conference on Machine Learning (ICML), Edinburgh, Scotland. Omnipress, pp. 799-806.

Gorban, A., Kegl, B., Wunsch, D., & Zinovyev, A. (2008). Principal Manifolds for Data Visualization and Dimension Reduction. Springer, pp. 340. http://dx.doi.org/10.1007/978-3-540-73750-6

Grabocka, J., Bedalli, E., & Thieme, L. S. (2012). Efficient Classification of Long Time-Series. *Advances in Intelligent Systems and Computing, 207*(1), 47-57. http://dx.doi.org/10.1007/978-3-642-37169-1_5

Gunawardana, A., Mahajan, M., Acero, A., & Platt, J. C. (2005). Hidden conditional random fields for phone classification. In the Proceedings of the International Conference on Speech Communication and Technology, Lisbon, Portugal. Publisher Curran Associates, Inc., pp. 1117-1120.

HCRFLIB. (2010). HCRF library (including CRF and LDCRF). Retrieved February 14, 2014, www.sourceforge.net/projects/hcrf

Holmstrom, J. (2002). Growing Neural Gas Experiments with GNG, GNG with Utility and Supervised GNG, M. S. thesis, Uppsala University, Uppsala, Sweden.

Koller, D., & Friedman, N. (2009). Probabilistic Graphical Models: Principles and Techniques. MIT Press, pp. 1265.

Kurimo, M. (1997). Using Self-Organizing Maps and Learning Vector Quantization for Mixture Density Hidden Markov Models, Thesis of PhD of Technology, Helsinki University of Technology, Finland.

Martinetz, T., & Schulten, K. (1991). A "neural gas" network learns topologies. *Artificial neural networks, 1*(1), 397-402.

Neukirchen, C., Rottland, J., Willett, D., & Rigoll, G. (2001). A continuous density interpretation of discrete HMM systems and MMI-neural networks. *IEEE Transactions on Speech and Audio Processing, 9*(4), 367-377. http://dx.doi.org/10.1109/89.917682

Osterndorf, M., & Rohlicek, J. (1990). Joint quantizer design and parameter estimation for discrete hidden Markov models. In the Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Albuquerque, New Mexico, USA. Publisher IEEE, pp. 705-708. http://dx.doi.org/10.1109/ICASSP.1990.115865

PMTK. (2010). *Probabilistic Modeling Toolkit for Matlab/Octave.* Retrieved February 14, 2014, from www.github.com/probml/pmtk3

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257-286. http://dx.doi.org/10.1109/5.18626

Somervuo, P. (2000). Competing Hidden Markov Models on the Self-Organizing Map. In the Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN), Como, Italy, Vol. 3. Publisher IEEE, pp. 169-174. http://dx.doi.org/10.1109/IJCNN.2000.861299

Soullard, Y., & Artieres, T. (2011). Hybrid HMM and HCRF model for sequence classification. In the Proceedings of the 19th European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium. Publisher ESANN, pp. 453-458.

UCI Arabic Digit. (2010). *UCI Repository. Spoken Arabic Digit Data Set.* Retrieved February 14, 2014, from http://archive.ics.uci.edu/ml/datasets/Spoken+Arabic+Digit

UCI Character. (2008). *UCI Repository. Character Trajectories Data Set.* Retrieved February 14, 2014, from http://archive.ics.uci.edu/ml/datasets/Character+Trajectories