Probabilistic model parameter optimization for the problem solving algorithm

Mikhaylov A.A., Bazuyeva S.A.*

Platov South-Russian State Polytechnic University (NPI), 132, St. Prosvescheniya, Rostov region, Novocherkassk, 346428, Russian Federation

Abstract

The variety of problem solving algorithms models over set of the alternative solutions determines the application of the principle of equivocation logical reduction by narrowing of the solutions set. The choice of an optimum decision comes to the logical conversion of alternative solutions set to the feasible solutions set and to the effective solutions set. The alternative solution set is transformed to the feasible solution on the constraints set. The article explains the application of Hidden Markov Model (HMM) for the choice of optimum problem solving algorithm concerning the observable consistency. In this case we use the maximum likelihood criterion with the constraints in the form of normalizing conditions and semantic measure of the information expediency of A.A. Harkevich for the optimization of unknown parameters of the problem solving algorithm. The “committee” constructions are used for the “integration” of some algorithms for collective decision. We receive the optimal parameters for the algorithm of the collective decision using estimation of the posterior probabilities of algorithm appliance.

Keywords: Hidden Markov Model, alternative solution set, feasible solution set, effective solution set, optimal solution.

1. Introduction

In the real-world problems concerning decision-making the great information uncertainty is remained during the choice of optimal solution algorithm. This uncertainty is governed by the great initial variety of the problem solving algorithms models on alternative solutions set \( Y \). It is difficult to make choice of the problem solving algorithm from the given set. In this connection we use the principle of the logical reduction of uncertainty. It comes to the narrowing of the problem solving set.
As is known there are three stages of this procedure. During the first stage the initial set of the alternative solutions \( Y \) is narrowed down to the feasible solution set \( Y^g \). During the second stage the feasible solution set is narrowed down to the effective solution set \( Y^0 \). During the third stage we realize the choice of the optimal solution. That is the choice of the optimal solution comes to the sequence of the set transformations \( Y^g \rightarrow Y^0 \rightarrow Y^* \).

The alternative solution set is transformed down to the feasible solution set (we take into account constraints set). Performing constraints is the pre-requisite for the solution algorithm choice. That’s why the singular finish making decision \( Y^* \) is in the good-enough solution set. Narrowing of the solution set down to the feasible solution set begins at the stage of the initial set forming. Narrowing of the feasible solution set down to the effective solution set raises the definiteness of optimal solution choice on the effective solution set. All initial information is completely used for the extraction of the effective solutions from the feasible solution set. That’s why the optimal solution choice is possible when we receive new information, method of its introduction.

The optimal solution choice concerning observable consistency \( O \) comes to the characterization using Hidden Markov Model \( \Theta^* \). They maximize probability of the observable sequence appearance \( O = o_1 o_2 \ldots o_T \), \([1-5]\)

\[
\Theta^* = \max_{\Theta} (P(O|\Theta)) .
\] (1)

The aim of this article is the characterization \( \Theta \) of the problem solving algorithms using HMM. Parameters must satisfy the optimality criterion of maximum likelihood.

2. Constraint qualification for criterion

The important element of the algorithm model is the constraint of the performance criterion of the problem solving. These problems are determined by the information features used in the algorithm. We can be limited by the minimum constraint group during the algorithm parameter optimization. We allocate two constraints required for algorithm synthesis.

1. The constraints in the form of the normalizing conditions which are applied on the density of distribution of the random variables. Thus for probability \( P_k \) of the model belonging \( k \) is non-negative and the sum of the probabilities equal to one

\[
\sum_{k=1}^{I} P_k = 1, \quad P_k \geq 0 \quad \forall k=1, l .
\]

2. The second constraints type for the criterion of the problem solving algorithm optimization is determined by the feature of the algorithm function which comes to the purposeful logical transformation of the A.A. Harkevich information semantic measure \([6]\). We appreciate amount of information taking into account the change of the action expediency degree, the goal attainment of the algorithm function which is controlled by the information influence in the form of the some message. Using this constraint we have alternative solution set. Feasible solution set is formed from these alternative solutions in accordance with the aim of the algorithm function. Given constraint decreases the limitations of Bayesian approach because it is determined by the posterior information on the goal attainment of the algorithm function. Numerical measure of the conditionality of the information value introduced by Bongard and Harkevich \([7, 8]\) is determined by the goal attainment probability during the information pull and equal

\[
I = \log_2 \left( \frac{P_1}{P_0} \right) ,
\]

where \( P_0 \)– the goal attainment probability before the information pull; \( P_1 \)– the goal attainment probability after the information pull and its using. This expression is transformed into (2) when the control action influences the algorithm:

\[
I_j^* = \log_2 \left( \frac{P_j^*}{P_j} \right) ,
\] (2)

where \( I_j^* \) – numerical measure of the controlling factor force in the form of the machine instruction \( v_i \) to the algorithm transition from \( i \) state to \( j \) state.
Given constraint in the condition of the synthesis of the problem solving algorithm determines the condition of the problem solving goal attainment in the form of: $P_{V \rightarrow Y}^V = 1$, i.e.

$$
\sum_{i=0}^{V} \log(P_{i,i+1}^V / P_i) = \sum_{i=0}^{V} I^V_i = \log(1 / P_0),
$$

where $P_{i,i+1}^V$—probability of the transitions from $i$ state to $j=i+1$ state. It’s given by the value of the machine instruction $v_i$; $V$—sequence of transitions forming Viterbi method which receives the most probable sequence of the states [9] and gives the problem solving algorithm.

Other constraints can be used besides given constraints. Thus the constraints of the variance of the distribution law are required for the algorithm of signal extraction from the signal-noise mixture. The constraints of the signal swing are required for the algorithm of the signal determination.

3. General approach to the models parameter estimation

Conceptual model of the deterministic algorithm is represented as the sequence of the machine instructions $v_i$ determined on the machine instructions set $V=\{v_i\}, i=1, n_{mk}$, where $n_{mk}$—number of the machine instructions which are in the algorithm class fetch. They are represented as the system of the instructions of used processor class.

The stochastic algorithm is represented as the doubly stochastic problem which consists of the set of the known discrete observed variables $O=\{a_1, ..., a_N\}$, which describe the appearance of the machine instructions $o_n \in R^d$ and hidden variables $Q=\{q_1, ..., q_N\}$. Hidden variables $Q$ determine the changes $N$ of the model states (state variables). Values of the observed vector $o_i$ taken in the instant $i$ depend on the $i$ state

$$
P(o_i | a_i, a_{i-1}, ..., a_1) = P(o_i | a_i),
$$

i.e. it doesn’t depend on the time and hidden state $q_i$, in previous instant $q_{i-1}$, that is the transition function

$$
P(q_i | q_{i-1}, ..., q_1, o_1) = P(q_i | q_{i-1}),
$$

we don’t know how many states and how much connection there are between them.

Given axioms determine the algorithm in the form of the hidden Markov process. It is represented as bicomponent random process with hidden component and observed component of the observation symbol appearance (handles of the assembler language machine instructions).

Hidden Markov Model HMM [2, 4] has the form

$$
\Theta = (N, M, A, B, \Pi),
$$

where $N$—number of the model states. Model states set is represented as $S=\{s_i\}, i=1, N$, model state-of-the-art in the instant $t$—as $q_i$ from the sequence $Q$, which is the implementation of the hidden process; $M$—number of the possible symbols in the observed sequence.

These symbols are appeared from the model and form the alphabet $V=\{v_k\}, k=1, M$; $A=\{a_j\}$—transition probability matrix, where $a_j=P(d(q_i=s_j) | q_{i-1}=s_i), 1 \leq j, i \leq N$—probability of the model transition from the state $q_{i-1}=s_i$ in instant $t$ into the state $q_i=s_j$ next moment $t$, $q_i$—the state in the instant $t=2, T, T$—length of sequence; $B=\{b_j(k)\}$—probability distribution of the symbols appearance in the state $j$, $(1 \leq j \leq N)$, where $b_j(k)=P(b(s_i=v_k | q_j=s_j))$ — reference probability distribution that in the instant $t$ system in $j$ state (state $s_j$) gives $k$ symbol ($v_k$) into observed sequence $O$, $k=1, M$—number of the different observation symbols ok. They can be given by the model (dimension of the discrete alphabet) $V=\{v_k\}; \Pi=\{\pi_i\}$—probability distribution of the initial state $\pi_i=P(q_i=s_i), 1 \leq i \leq N$, where $q_i$—the state
in the instant \( t=1 \) from the sequence \( Q \), which is the implementation of the hidden process i.e. the probability that \( s_i \) – initial state of the model.

HMM realizes cybernetic model “black box” which generates observed sequence after performing of the given steps number \( O=o_1o_2...o_T \).

This observed sequence is formed by the symbols of discrete alphabet \( V \), which consists of the machine instructions handles \( v_i=a \), where \( a \in R \) – observation fixed in the instant \( t=1 \), \( T \) – number of the symbols in the observed sequence.

4. Parameter optimization of hidden Markov model problem formalization

*Problem statement:* observed sequence in the form of the machine instruction chains is the initial data. HMM is determined for every machine instruction chain. Three problems connected with HMM are underlined [2].

The first problem comes to the estimated probability \( P(O|\Theta) \) that given observed sequence \( O=o_1, o_2, ..., o_T \) was created only for model \( \Theta=(A, B, \Pi) \)

\[
P(O|\Theta) = \sum_{Q} P(O|Q, \Theta).
\]

In the second problem for the sequence of observations \( O=o_1, o_2, ..., o_T \) and HMM we must choose the states sequence \( Q=q_1, q_2, ..., q_T \), which determines the sequence of observations \( P[Q|O, \Theta] \) and has maximum probability \( P(Q|O, \Theta) \). In the third problem we must select model parameters \( \Theta=(A, B, \Pi) \) maximizing \( P(O|\Theta) \).

*Required:* for three problems solving connected with the determination of structure and parameters \( \Theta \) of the problem solving algorithms (1) we must choose correct criterion of the maximum likelihood optimization.

Task solution: we introduce hidden variables \( Q \) of HMM model for problem solving given by incomplete function of likelihood (1) with the observed variables \( O \) and parameters \( \Theta \). We estimate the vector of parameters \( \Theta \) of complete function of likelihood \( \log p(O, Q|\Theta) \), for which

\[
\log p(O|\Theta) = \log \left( \sum_{Z} p(O, Q|\Theta) \right).
\]

We determine the joint distribution of HMM variables. From the conditions of independence (3, 4) in the determination of HMM we realize that the generation probability by the sequence \( Q=q_1, q_2, ..., q_T \) of the hidden states the observation sequence \( O=a_1, a_2, ..., a_T \) is calculated as:

\[
P(O|Q, \Theta) = P(a_1, a_2, ..., a_T|q_1, q_2, ..., q_T, \Theta) = \prod_{t=1}^{T} p_B(a_t|q_t).
\]

The probability of sequence appearance \( Q=q_1, q_2, ..., q_T \) equal

\[
P(Q|\Theta) = P(q_1, q_2, ..., q_T|\Theta) = \pi_{q_1} \prod_{t=1}^{T-1} p_A(q_t|q_{t+1}),
\]

i.e. probability of observation of some symbol depends on the model state in the given instant.

Because of the appearance of the concrete sequence of the states and the appearance of the sequence of observation for HMM is autonomous, probability of the sequence observations \( O \) which was formed from the sequence \( Q \) in the form of the joint distribution is given by the formula

\[
P(O, Q|\Theta) = P(O|Q, \Theta),
\]

and

\[
P(O|\Theta) = p_n(q_1) \prod_{n=1}^{N} p_B(a_n|q_n) \prod_{n=2}^{N} p_A(q_n|q_{n-1}),
\]
where \( O = \{ o_1, o_2, ..., o_N \} \) – observed states of the model with the hidden variables; \( Q = \{ q_1, q_2, ..., q_N \} \) – hidden states of the model which describe its inner state; \( \Theta \) – model parameters with hidden variables.

Apart from the probability of transition between the hidden states \( q_n \) of the model algorithm which describes magnitude of connection between the states we’ll introduce Viterbi variable in the form of \( k \)-measure binary random vector \( z_{nj} \), \( j \in \{ 0, 1 \} \) its component equal to one. This vector determines the hidden states choice and their sequence. Given sequence in the form of the Viterbi approach \( V \) gives the conversion direction of the dataflow in the algorithm

\[
P(O, Q|\Theta) = \prod_{j=1}^{K} \pi_j^{z_j} \left( \prod_{n=2}^{N} \prod_{i=1}^{K} \prod_{j=1}^{K} P(q_n|q_{n-1})^{1-z_{nj}} \prod_{n=1}^{K} P(\alpha_n|q_n)^{z_{nk}} \right).
\]

For the parameters estimation \( \Theta \) we use the method of maximum likelihood

\[
\Theta_{ML} = \arg \max p(O, Q|\Theta) = \arg \max \log p(O, Q|\Theta),
\]

for the logarithmic likelihood function

\[
\log p(O, Q|\Theta) = \left( \sum_{j=1}^{K} \log \pi_j \right) + \left( \sum_{n=2}^{N} \sum_{k=1}^{K} z_{nj} \log P(q_n|q_{n-1}) \right) + \left( \sum_{n=1}^{N} z_{nk} \log P(\alpha_n|q_n) \right).
\]

Parameters which enter \( \Theta \) cannot have arbitrary values. Hence the optimization takes place in the conditions of the constraints

\[
\sum_{j=1}^{K} \pi_j = 1, \quad \sum_{j=1}^{K} P_j = 1, \quad \forall i = 1, K,
\]

where \( N_v \)–number of the transitions forming algorithm Viterbi approach.

We use the rule of Lagrange multiplier [10] and we have

\[
L(\Theta, \lambda, \mu, \eta) = \log p(O, Q|\Theta) + \lambda \left( \sum_{j=1}^{K} \pi_j - 1 \right) + \sum_{i=1}^{K} \mu_i \left( \sum_{j=1}^{K} P_j - 1 \right) + \sum_{i=1}^{K} \mu_i \left( \sum_{j=1}^{K} \log \frac{P_j}{P_0} + \log \frac{P_0}{P_j} \right) + \sum_{k=1}^{K} \eta_k \left( \sum_{n=1}^{K} P(q_n|\alpha_n) - 1 \right) \rightarrow \text{extr}. \quad (5)
\]

5. Characterization of hidden Markov model

Lagrangian (5) allows determining HMM structure and parameters.

1. For the distribution of the probability of the initial states we determine Lagrangian derivative using the element \( \pi_j \) of the matrix of the initial states probability \( \Pi \)

\[
\frac{\partial L(\Theta, \lambda, \mu, \eta)}{\partial \pi_j} = \frac{z_{1j}}{\pi_j} + \lambda = 0, \quad \text{hence,} \quad \pi_j = -\frac{z_{1j}}{\lambda}.
\]

We sum this expression using the Viterbi variable taking in account that \( \sum_{j=1}^{K} \pi_j = -1 \), we have \( \lambda = -\sum_{j=1}^{K} z_{1j} = -1 \), i.e. the element \( \pi_j \) estimation of the matrix \( \Pi \) has form
\[ \pi_j = z_{1j} \]  \hspace{1cm} (6)

2. For the distribution of the probability of the model transition from the state \( i \) into the state \( j \) we determine Lagrange derivative using the element \( P_{ij} \) of the matrix of transitional probabilities \( A \)

\[ \frac{N}{\sum_{n=2}^{N} z_{n-1,i} z_{nj}} \mathbf{1} - \mu_i \frac{1}{P_{ij}} = -\mu_j, \text{ hence, } P_{ij} = \frac{\sum_{n=2}^{N} z_{n-1,i} z_{nj} + \mu_i}{-\mu_i}. \]

Sum this expression taking into account that \( \sum_{j=1}^{K} P_{ij} = 1 \) \( \text{ and } \sum_{j=1}^{K} z_{nj} = 1 \) and receive

\[ \sum_{j=1}^{K} P_{ij} = 1 = \frac{\mu_i + \sum_{j=1}^{K} \sum_{n=2}^{N} z_{n-1,i} z_{1j}}{-\mu_i} = \frac{\mu_i + \sum_{n=2}^{N} z_{n-1,i}}{-\mu_i}, \text{ i.e. } \mu_i = -\frac{\sum_{n=2}^{N} z_{n-1,i}}{2}. \]

As a result the estimation of the elements \( P_{ij} \) of the matrix \( A \)

\[ P_{ij} = \frac{\sum_{n=2}^{N} z_{n-1,i} z_{nj}}{\sum_{n=2}^{N} z_{n-1,i}} - 1. \]  \hspace{1cm} (7)

Sum the expression (7) on the set \( K \)

\[ \sum_{i=1}^{K} \mu_i = -\frac{\sum_{n=2}^{N} \sum_{i=1}^{K} z_{n-1,i}}{2}, \]

and receive the expression for the average value of the coefficient \( \mu \)

\[ \frac{\sum_{i=1}^{K} \mu_i}{K} = \frac{1}{N-1} = \frac{1}{2}. \]

3. For the distribution of the probabilities of the symbol appearance in \( j \) state we determine Lagrange derivative using the element \( P(o_n|q_n) \) of the matrix of the conditional probabilities \( B \)

\[ \frac{\partial L(\Theta, \lambda, \mu, \eta)}{\partial P(o_n|q_n)} = \frac{\sum_{k=1}^{K} \eta_k}{P(o_n|q_n)} + \sum_{k=1}^{K} \eta_k P(q_n) = 0 \text{ i.e. } 1 = -P(q_n|o_n) \sum_{k=1}^{K} \eta_k. \]

Sum this expression and have \( K = -\sum_{k=1}^{K} \eta_k \). Taking into account this result we receive the element \( P(o_n|q_n) \) of the matrix \( B \)

\[ P(q_n|o_n) = \frac{1}{K} \text{ or } P(o_n|q_n) = \frac{P(o_n)}{KP(q_n)}. \]  \hspace{1cm} (8)
6. Parameter optimization of the algorithm (committee) model

For the collective decision we use some algorithms mix “integration” into the “committee” instructions which use the estimation of the posterior probabilities of the belonging to the initial algorithm class.

At the “committee” model of the algorithm in the form of the mixture $l$ of algorithm the initial dependence $p(y|x)$ is expressed as the models composition $p(y|x, \Theta_k)$:

$$p(y|x) = \sum_{k=1}^{l} G(\Theta_k|x) p(y|x, \Theta_k) = \sum_{k=1}^{l} g_k p(y|x, \Theta_k),$$  \hspace{1cm} (9)

where $g_k=G(\Theta_k|x)$ – the gate of mixture in the form of the belonging to the model $\Theta_k$ with the normalizing condition

$$\sum_{k=1}^{l} g_k = 1, \ g_k \geq 0 \ \forall k.$$  \hspace{1cm} (10)

As the objects in the sampling are autonomous, the density of the joint distribution (9) is transformed into the product of the distribution densities of every object

$$p(y|x) = \prod_{i=1}^{n} p(y_i|x_i, \Theta_k) = \prod_{i=1}^{n} \sum_{k=1}^{l} g_k p(y_i|x_i, \Theta_k).$$  \hspace{1cm} (11)

For the characterization of the mixture parameters we maximize $p(y|x)$. We change the order of summation and multiplication and use the principle of maximum likelihood. We form Lagrange’s function [10] from (10), (11) in the form of:

$$L=\sum_{i=1}^{m} \ln \left[ \sum_{k=1}^{l} g_k p(y_i|x_i, \Theta_k) \right] - \gamma \left( \sum_{k=1}^{l} g_k - 1 \right).$$  \hspace{1cm} (12)

For the determination of the models gates we equate Lagrange’s function derivative (12) on $g_k$ to zero:

$$\frac{\partial L(\Theta_k, \gamma)}{\partial g_k} = \sum_{i=1}^{m} \frac{p(y_i|x_i, \Theta_k)}{\sum_{k=1}^{l} g_k p(y_i|x_i, \Theta_k)} - \gamma = 0.$$  \hspace{1cm} (13)

For next transformation of this equation we denote the probability $P(y, \Theta_k|x)$ that the object $(x, y)$ is determined by the component $\Theta_k$, $P(\Theta_k|y_i, x_i)$ – the probability that $k$ component of the model is determined by $i$-object. Every object was created by some model with the formula of the complete probability

$$\sum_{k=1}^{l} P(\Theta_k|y_i, x_i) = 1, \ \forall i.$$  \hspace{1cm} (14)

For the object $(x, y)$ the probability of its determination by the model $\Theta_k$ with the formula of conditional probability equal:

$$P(y, \Theta_k|x) \rightarrow p(y|x, \Theta_k) = g_k p(y|x, \Theta_k).$$  \hspace{1cm} (15)

We substitute the equality (15) in the Bayes’ formula for $P(\Theta_k|y_i, x_i)$ and receive

$$P(\Theta_k|y_i, x_i) = \frac{g_k p(y|x_i, \Theta_k)}{\sum_{k=1}^{l} g_k p(y|x_i, \Theta_k)}.$$  \hspace{1cm} (16)
We multiply both parts of the equality by $g_k$ and sum on $k=1...l$. Taking into account the equality (16) we receive
\[ m = \sum_{k=1}^{l} \sum_{i=1}^{m} g_k p(y_i | x_i, \Theta_k) = \gamma \sum_{k=1}^{l} g_k = \gamma. \] (17)

Using received result we have from 13
\[ g_k = \frac{1}{m} \sum_{i=1}^{m} g_k p(y_i | x_i, \Theta_k) = \frac{1}{m} \sum_{i=1}^{m} P(\Theta_k | y_i, x_i). \] (18)

The equality 18 allows determining the gate $g_k$ of the committee model.

2. For the characterizing of the committee model components we calculate Lagrange’s function derivative using the parameters $k$ of model $\Theta_k$:
\[ \frac{\partial L(\Theta_k, \gamma)}{\partial \Theta_k} = \sum_{i=1}^{m} g_k p(y_i | x_i, \Theta_k) \frac{\partial \ln p(y_i | x_i, \Theta_k)}{\partial \Theta_k} = \frac{\partial}{\partial \Theta_k} \sum_{i=1}^{m} P(\Theta_k | y_i, x_i) \ln p(y_i | x_i, \Theta_k) = 0. \] (19)

Received equality determines necessary conditions of the maximum of the committee model likelihood function. These conditions match the conditions of the maximum of the likelihood function of the committee model components.

7. Problem solving algorithms of the algorithm parameter optimization

The choice of the problem solving algorithm of the algorithm parameter optimization is performed for the given select $(O, Q)$ in the form of one or several sequences with known values of the hidden components. Matrices of the probability distributions $A, B, \Theta$ of the model are filled by the equiprobable values (stage of the initialization).

I. For the probability of the generation $P(O | \Theta)$ of the sequence of observations $O=o_1, o_2,..., o_T$ for the model $\Theta=(A, B, \Pi)$ in the first task [2] we use the algorithm of forward-backward procedure [2,3,11-13].

Forward procedure. We calculate logically the intervening forward variable $\alpha_t(i)$ as
\[ \alpha_t(i)=P(o_1, o_2...o_t, q=s_t|\Theta), \]
i.e. the probability that for given model $\Theta$ till the instant $t$ we observed the sequence $o_1, o_2...o_t$. In this instant it is situated in the state $s_t$. Required probability $P(O|\Theta)$ is represented as
\[ P(O|\Theta) = \sum_{i=1}^{n} a_T (i). \]

We calculate the value $\alpha_t(i)$ by the method of induction using next algorithm:
1. At the stage of initialization we calculate the probability of the state $s_j$ and the first observation $o_1$ overlap $\alpha_t(i)=\pi_j(b_t(o_1)), 1\leq t \leq N$.
2. At the stage of induction we find the method which shows how the system in the instant comes into the state $s_j$ from $N$ possible states of previous instant. As $\alpha_t(i) = \text{joint probability of observation display } o_1o_2...o_t$ and system location in the state $s_i$ in the instant $t$
\[ \alpha_{t+1}(j)=\sum_{i=1}^{N} \alpha_t (i) a_{ij} b_j (o_{t+1}), 1 \leq t \leq T-1, 1 \leq j \leq N. \]
3. Finish at the step $T$:
\[ P(O|\Theta) = \sum_{i=1}^{N} a_T (i). \]
**Backward procedure.** We introduce backward variable $\beta_t(i)$ – conditional probability that system will be situated in the state $i$ by the instant $t$. The sequence of its observations $o_{t+1}, o_{t+2}, \ldots, o_T, o_T$

\[
\beta_t(i) = P(o_{t+1}\ldots o_T | q_t = s_i, \Theta).
\]

1. For all $i$ from 1 to $N$ we take $\beta_T(i) = 1$, then using induction.
2. For all $i$ in the reverse direction from $T-1$ to 1 and for all $i$ from 1 to $N$

\[
\beta_t(i) = \sum_{j=1}^{N} a_{ij} \beta_{t+1}(j).
\]

3. At the completion phase we determine $P(O|\Theta) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$.

II. For the solving of the second class problems we denote Viterbi algorithm [14-17] which uses dynamic programming for finding the calculation of the best chain of the states (fig.1) with the maximum probability $P[O|O, \Theta]$.

We introduce the auxiliary variables in the form of the maximum probability to reach the state $s_i$ at the stage $t$ among all methods with observed variables

\[
\delta_t(i) = \max_{q_1,...,q_t} P(q_1,q_2,...,q_t = x_1,o_1,o_2,...,o_t|\Theta) .
\]

1. At the stage of the initialization we determine $\delta_1(i) = \pi_i b_i(o_1)$ and $\Psi_1(i)$ – the most probable states sequence, reliable for the appearance of the first observed symbols. It finishes in the state $i$.
2. At the stage of induction the most probable states sequence $q_1,...,q_T$ is given by the recurrent relations

\[
\delta_t(j) = \max_{1 \leq i \leq N} \delta_{t-1}(i) a_{ij} b_j(o_t), \Psi_t(j) = \arg \max_{1 \leq i \leq N} [\delta_{t-1}(i) a_{ij}] .
\]

3. At the final stage $T$ we calculate

\[
P^* = \max_{1 \leq i \leq N} \delta_T(i), \quad q_T^* = \arg \max_{1 \leq i \leq N} \delta_T(i)
\]

the most probable sequence of the hidden states $q_T^*$, reliable for the appearance of the first $t$ observed symbols. It finishes in the state $n$

\[
q_T^* = \Psi_{T+1}(q_{T+1}^*) .
\]

III. In the third task we use the iteration Baum-Welch algorithm. It’s the modification of the Estimation Maximization algorithm (EM). It allows determining the parameters $A, B, \Pi$ which maximize the likelihood function
of the given model $\Theta=(A, B, \Pi)$ [2]. It performs this action for the appropriate sequence of the observed values $O$ iteratively with given accuracy threshold $\varepsilon$ [2].

At the E-step we calculate expectation model parameters values $a_{ij}$, $b_{i}(n)$, $\pi_i$ in the condition of the given data. We introduce the third auxiliary variable as the probability that with the given sequence of observations in instant $t$ – in the state $s_i$, in instant $t+1$ – in the state $s_j$

$$\xi_{t}(i,j)=P(\eta_{t}=s_i, \eta_{t+1}=s_j|O, \Theta),$$

which have the form through the first and the second auxiliary variables

$$\xi_{t}(i,j)=\frac{a_{i}(i)a_{j}b_{j}(\eta_{t+1})\beta_{t+1}(j)}{P(O|\Theta)} = \frac{a_{i}(i)a_{j}b_{j}(\eta_{t+1})\beta_{t+1}(j)}{\sum_{j} a_{i}(i)a_{j}b_{j}(\eta_{t+1})\beta_{t+1}(j)},$$

and the forth variable

$$\zeta_{t}(i)=P(\eta_{t}=s_i|O, \Theta)=\frac{a_{i}(i)\beta_{t}(i)}{\sum_{i} a_{i}(i)\beta_{t}(i)} = \frac{\sum_{j=1}^{N} \xi_{t}(i,j)}{\sum_{i} a_{i}(i)\beta_{t}(i)},$$

which is represented as containment probability in the instant $t$ in the state $s_i$ with the given sequence of observations $O$ and the model $\Theta$.

At the M-step (maximization) we have next probability approaching:

- expected rate of the $i$ state in the instant $t$ $\bar{\pi}_i=\xi_{t}(i)$;
- expected rate according to the expression (7) as the doubled ratio of the number of the transitions from the $i$ state to the $j$ state to the number of the output appearance in the $i$ hidden state without $1$

$$\bar{a}_{ij} = \frac{2 \sum_{i=1}^{T-1} \xi_{t}(i,j)}{\sum_{i=1}^{T-1} \xi_{t}(i)} - 1;$$

- expected rate according to the expression (8) as the ratio of the couples number $(q_n, o_n)$, $m_q$—number of the hidden states, $m_o$—number of the observed states (symbol $v_n$) to number of the appearances $n$ hidden states $q_n$

$$\bar{b}_{i}(n)=P(o_{n}|q_{n})=\frac{P(q_{n}|o_{n})P(o_{n})}{P(q_{n})} = \frac{P(q_{n}|o_{n})}{P(q_{n})} = \frac{1}{K} \frac{m_{o}}{m_{q}} = \frac{1}{K} \frac{m_{o}}{m_{q}},$$

when $P(q_{n}|o_{n}) = \frac{1}{K}$.

In the iterative repetition of two steps EM-algorithm beginning from $\Theta=(A, B, \Pi)$, we determine $\tilde{\Theta}=(\tilde{A}, \tilde{B}, \tilde{\Pi})$, then we calculate the parameters again and so on.

Given algorithm was offered by M.I. Schlesinger [18]. It was covered all over again as the EM-algorithm (expectation – maximization) [19]. Application of the auxiliary hidden variables $Q$ into EM-algorithm provides algorithm convergence [20, 21] and its conditioning with HMM. It simplifies the calculation of the likelihood maximum for the determination of the values of the parameters vector $\Theta=(A, B, \Pi)$. 
8. Summary

1. Great information ambiguity typical during the choice of optimal problem solving algorithm and caused by the great initial variety of algorithms models on the alternative solution set determines necessity of the logical transformation of initial set of the alternative solution to the feasible solutions and to the set of the effective solutions which is narrowed down to the set of the optimal solution.

2. We take into account many constraints during the transformation of the alternative solution set to the feasible solution set. These constraints must conform to the informal specificities of the problem solving. We must take into consideration the specificities of the operation formalization of the synthesizable algorithm during the algorithm synthesis. Constraint satisfaction is the necessary condition for the choice of the solution algorithm that’s why the final solution is situated in the set of the feasible solutions.

3. The likelihood function in the form of the Lagrange equation is formed for the parameter optimization. The constraints in the least compound are formed by way of the normalizing conditions for each model variable and in the form of the Harkevich information.

4. The common approach to the estimation of the HMM algorithm optimal parameters determines the probability distribution of the initial states by the expression (6), model transition probability from the state \( i \) to the state \( j \) (7), probabilities of the symbol appearance in the \( j \) state (8).

5. We use Mixture of Experts in the form of the “committee” constructions during making collective decision. They use the gates estimation (a posteriori probability of the algorithm belonging to the class).

6. During the optimization of the algorithm committee model:
   - the optimal values of the gates parameters are determined by the expression (18);
   - the optimal parameters of the algorithm committee model are given by the necessary conditions of the likelihood function maximum of the algorithm committee model. They match the conditions of the likelihood function maximum of the committee model components.

References