Mathematical Problems of Computer Science 33, 102–110, 2010.

Error Probability Exponents of Multiple Hypotheses Testing Illustrations

Naira M. Grigoryan

State Engineering University of Armenia
nar.gri@gmail.com

Abstract

The paper presents an application of multiple hypothesis testing for two Markov sources with virtual example in terms of text categorization problem. Some numerical experiments concerning Markov sources are considered. Our goal is to present numerical illustrations of interdependencies of error probability exponents as a supplementary to our previous theoretical paper [9].

1. Introduction

Ahlswede and Haroutunian [2] formulated and solved a group of problems concerning logarithmicaly asymptotically optimal (LAO) hypotheses identification and hypotheses testing (HT) for many objects, which is based on result of [4] for testing of many hypotheses for one object. In [3]–[5] the problem of multiple hypotheses testing for one Markov chain was considered. In [8] three hypotheses testing problem for two independent objects with independent observations was studied. Some numerical illustrations of the HT problem for two independent Markov sources considered in [9].

We give a short survey of multiple HT applications areas for Markov sources.

In paper [12] Shannon founded the field of information theory which revolutionized the telecommunications industry. In that paper, Shannon also proposed using a Markov chain to create a statistical model of the sequences of letters in a piece of English text. Markov chains are now widely used in speech recognition, handwriting recognition, information retrieval, data compression, spam filtering and text classification. Now there are numerous text documents available in electronic form. Such documents represent a massive amount of information that is easily accessible. Seeking value in this huge collection requires organization; much of the work of organizing documents can be automated through text classification. The accuracy and our understanding of such systems greatly influences their usefulness.

Now we make a short survey of some basic concepts and facts needed for further expansion of the subject. The Markov source (MS) is a version of discrete memoryless source (DMS) with a Markovian dependence of consecutive messages. A formal representation of MS is as follows. Let $\mathcal{X}$ be the source alphabet and

$$G \triangleq \{G(x|u), \quad x, u \in \mathcal{X}\}, \tag{1}$$

be a matrix of Markovian conditional probabilities.

MS produces a sequence of random variables $\{X_n\}_{n=0}^{\infty}$. Let transition probability of stationary Markov chain be

$$G(X_n = x|X_{n-1} = u) = G(x|u), \qquad x, u \in \mathcal{X}, \ , \quad n = 1, 2, \ldots$$

The source studied in this paper is defined by Markov chain with not necessarily unique stationary distributions $Q = \{Q(u),\ u \in \mathcal{X}\}$ corresponding to transition probability distributions G (we will note this circumstance by the following way $G \to Q$)

$$\sum_{u \in \mathcal{X}} Q(u)G(x|u) = Q(x), \quad x \in \mathcal{X}. \tag{2}$$

We denote the joint distribution of two consecutive messages as follows

$$Q \circ G \triangleq \{Q(u)G(x|u),\ u, x \in \mathcal{X}\}.$$

The conditional probability of the vector $\mathbf{x} = (x_0, x_1, \ldots, x_N) \in \mathcal{X}^{N+1}$ of the Markov chain with transition probability G and stationary distribution Q, is defined as the following product

$$Q \circ G^N(\mathbf{x}) \triangleq Q(x_0) \prod_{n=1}^{N} G(x_n|x_{n-1}). \tag{3}$$

The conditional probability of a subset $\mathcal{A}_N \subset \mathcal{X}^{N+1}$ is the sum

$$Q \circ G^N(\mathcal{A}_N) \triangleq \sum_{\mathbf{x} \in \mathcal{A}_N} Q \circ G^N(\mathbf{x}).$$

We are given M hypotheses about distributions of each of two independent sources (1),

$$H_{m_k} : G_{m_k}, \quad m_k = \overline{1, M}, \ k = 1, 2 \tag{4}$$

for each source only one of them to be true.

The statistician based on $N+1$ observations of the sources, namely on $\mathbf{x}^1 = \{x_0^1, \ldots, x_N^1\}$ and $\mathbf{x}^2 = \{x_0^2, \ldots, x_N^2\}$ should make decision which of M hypotheses is true. The test Φ^N for this model can be composed by the pair of tests φ_1^N and φ_2^N for the corresponding objects: $\Phi^N = (\varphi_1^N, \varphi_2^N)$. A test φ_k^N is a partition of $\mathcal{X}^{N+1}$ into M disjoint subsets $\mathcal{A}_N^{m_k}$. If $\mathbf{x}^k \in \mathcal{A}_N^{m_k}$ then the test adopts the hypothesis H_{m_k}. Naturally, in decision making in favor of one of M alternatives he/she may commit different kinds of errors, which are denoted by $\alpha_{l_1,l_2|m_1,m_2}(\Phi^N)$, this is the probability of the erroneous acceptance by the test Φ^N of the pair of hypotheses (H_{l_1}, H_{l_2}) provided that the pair (H_{m_1}, H_{m_2}) is true,

$$\alpha_{l_1,l_2|m_1,m_2}(\Phi^N) \triangleq Q_{m_1} \circ G_{m_1}^N(\mathcal{A}_N^{l_1}) \quad Q_{m_2} \circ G_{m_2}^N(\mathcal{A}_N^{l_2}),$$
$$(l_1, m_1) \neq (l_2, m_2), \quad m_k, l_k = \overline{1, M}, \ k = 1, 2. \tag{5}$$

The probability to reject a true pair of hypotheses (H_{l_1}, H_{l_2}) is the following

$$\alpha_{m_1,m_2|m_1,m_2}(\Phi^N) = \sum_{(l_1,l_2) \neq (m_1,m_2)} \alpha_{l_1,l_2|m_1,m_2}(\Phi^N). \tag{6}$$

We study error probability exponents of the sequence of tests Φ, which are called "reliabilities":

$$E_{l_1,l_2|m_1,m_2}(\Phi)\triangleq\limsup_{N\to\infty}-\frac{1}{N}\log\alpha_{l_1,l_2|m_1,m_2}(\Phi^N),\quad m_k,l_k=\overline{1,M},\ k=\overline{1,2}. \tag{7}$$

From (6) and (7) it is easy to see that

$$E_{m_1,m_2|m_1,m_2}(\Phi)=\min_{(l_1,l_2)\neq(m_1,m_2)}E_{l_1,l_2|m_1,m_2}(\Phi).$$

The matrix $E(\Phi)=\{E_{l_1,l_2|m_1,m_2}(\Phi)\}$ is called the reliability matrix of the sequence Φ of tests.

As in [1] we call the test sequence Φ^* logarithmically asymptotically optimal (LAO) for this model if for given values of the elements $E_{m,m|M,m}$, $E_{m,m|m,M}$, $m=\overline{1,M-1}$ it provides maximal values for all other elements of the matrix $E(\Phi^*)$.

The rest of the paper is organized as follows. In Section 2 the theorem on multiple LAO HT problem for two Markov sources is formulated. The numerical experiments results are presented in Section 3.

2. Multiple LAO HT for Two Markov Sources

In this section we present the main result of multiple HT with reliability requirement for pair of Markov sources, which was considered in [9].

For knowing correctly in which set of tests the elements of the reliability matrix $E_{m,m|M,m}$, $E_{m,m|m,M}$, $m=\overline{1,M-1}$ of the tests for two objects can be positive we divide the set of all tests $\Phi=(\varphi_1,\varphi_2)$ into following classes:

$\mathcal{A}\triangleq\{\Phi=(\varphi_1,\varphi_2):E_{m|m}(\varphi_k)>0,\ m=\overline{1,M-1},\ k=1,2\}$,

$\mathcal{B}\triangleq\{\Phi=(\varphi_1,\varphi_2)$: one, or two m' from $[1,2]$ exist such that $E_{m'|m'}(\varphi_k)=0$ for one value of k, but $E_{m'|m'}(\varphi_j)>0$, $k\neq j$, and for other $m<M, E_{m|m}(\varphi_k)>0$, $k,j=\overline{1,2}\}$,

$\mathcal{C}\triangleq\{\Phi=(\varphi_1,\varphi_2)$: one or two m' from $[1,M-1]$ exist such that $E_{m'|m'}(\varphi_k)=0$, and for other $m<M$, $E_{m|m}(\varphi_k)>0$, $k=\overline{1,2}\}$.

In other words we divide set of tests into classes taking in consideration zero values of elements of reliability matrix of one MS, because if there are a zero element in reliability matrix of a MS, then the corresponding element of the reliability matrix of both MS equals to zero too.

Let us define the following family of sets for given positive elements $E_{m,m|M,m}$, $E_{m,m|m,M}$, $m=\overline{1,M-1}$ to determine LAO test Φ^*:

$$R_m^1\triangleq\{Q\circ G:D(Q\circ G\|Q\circ G_m)\le E_{M,m|m,m},\ \exists Q_m:D(Q\|Q_m)<\infty\},\ m=\overline{1,M-1},$$

$$R_m^2\triangleq\{Q\circ G:D(Q\circ G\|Q\circ G_m)\le E_{m,M|m,m},\ \exists Q_m:D(Q\|Q_m)<\infty\},\ m=\overline{1,M-1},$$

$$R_M^1\triangleq\{Q\circ G:D(Q\circ G\|Q\circ G_m)>E_{M,m|m,m},\ m=\overline{1,M-1},$$

$$R_M^2\triangleq\{Q\circ G:D(Q\circ G\|Q\circ G_m)>E_{m,M|m,m},\ m=\overline{1,M-1}.$$

The optimal values of the reliabilities of the LAO test sequence will be the following:

$$E^*_{M,m|m,m}(E_{M,m|m,m})\triangleq E_{M,m|m,m},\qquad E^*_{m,M|m,m}(E_{m,M|m,m})\triangleq E_{m,M|m,m}, \tag{8}$$

$$E^*_{l_1,l_2|m_1,m_2}(\Phi^*) \triangleq \inf_{Q \circ G \in R^*_{l_k}} D(Q \circ G || Q \circ G_{m_k}), \qquad l_k \neq m_k, \qquad l_{3-k} = m_{3-k}, \quad k = 1,2, \tag{9}$$

$$E^*_{l_1,l_2|m_1,m_2}(\Phi^*) \triangleq E^*_{m_1,l_2|m_1,m_2}(\Phi^*) + E^*_{l_1,m_2|m_1,m_2}(\Phi^*), \qquad m_k \neq l_k, \quad k = 1,2, \tag{10}$$

$$E^*_{m_1,m_2|m_1,m_2}(\Phi^*) \triangleq \min_{(m_1,m_2) \neq (l_1,l_2)} E^*_{l_1,l_2|m_1,m_2}(\Phi^*). \tag{11}$$

Now we will formulate the result on multiple LAO HT for two Markov sources.

Theorem 1. *Let all distributions G_m, $m = \overline{1,M}$, be different and absolutely continuous relative to each other: $0 < D(Q_m \circ G_m || Q_l \circ G_l) < \infty$, $l \neq m$. If positive elements $E_{m,m|M,m}$, $E_{m,m|m,M}$, $m = \overline{1,M-1}$ are given and the following inequalities hold*

$$E_{M,1|1,1} < \min_{l=\overline{2,M}} [\min_{Q_l} D(Q_l \circ G_l || Q^1_l \circ G_1)], \tag{12}$$

$$E_{1,M|1,1} < \min_{l=\overline{2,M}} [\min_{Q_l} D(Q_l \circ G_l || Q_l \circ G_1)], \tag{13}$$

$$E_{m,M|m,m} < \min[\min_{l=\overline{1,m-1}} E^*_{m,l|m,m}, \min_{l=\overline{m+1,M}} \inf_{Q_l} D(Q_l \circ G_l || Q_l \circ G_m)], \ m = \overline{2,M-1}, \tag{14}$$

$$E_{M,m|m,m} < \min \min_{l=\overline{1,m-1}} E^*_{l,m|m,m}, \min_{l=\overline{l+1,M}} [\inf_{Q_l} D(Q_l \circ G_l || Q_l \circ G_m)], \ m = \overline{2,M-1}, \tag{15}$$

then

a) there exists a LAO test sequence $\Phi^* \in \mathcal{A}$, the reliability matrix of which $E(\Phi^*)$ is defined as in to (8)-(11) and all elements of it are positive,

b) when even one of the inequalities (12)-(15), written for multiple hypotheses is violated, then there exists at least one element of the matrix $E(\Phi^*)$ equal to 0,

c) for given positive numbers $E_{l,l|M,l}$, $E_{l,l|l,M}$, $l = \overline{1,M-1}$, the reliability matrix $E(\Phi)$ of the tests Φ from the class B or C necessarily contains elements equal to zero.

The proof of this theorem is presented in [9].

Here are presented how the elements of reliability matrix corresponding to two independent object can be expressed by the elements of reliability matrix corresponding to each of two objects.

Lemma 1. *If for given $E_{m|m}(\varphi_k)$, $m = \overline{1,M-1}$, $k = 1,2$, elements of the reliability matrix satisfy to the following conditions:*

$$E_{1|1} < \min[\min_{Q_m} D(Q_m \circ G_m || Q_m \circ G_1), \ m = \overline{2,M}]$$

$$E_{l|l} < \min[\min_{m=\overline{1,l-1}} E^*_{m|l}(E^*_{m|m}), \min_{m=\overline{l+1,M}} D(Q_m \circ G_m || Q_m \circ G_l)], \ l = \overline{2,M-1},$$

then for $\Phi = (\varphi_1, \varphi_2)$ test there exist the following equalities:

$$E_{l_1,l_2|m_1,m_2}(\Phi) = E_{l_1|m_1}(\varphi_1) + E_{l_2|m_2}(\varphi_2), \quad m_1 \neq l_1, \quad m_2 \neq l_2$$

If $m_k \neq l_k$, $m_{3-k} = l_{3-k}$, $k = 1,2$, then

$$E_{l_1,l_2|m_1,m_2}(\Phi) = E_{l_k|m_k}(\varphi_k).$$

3. An Example Showing the Interrelation of Error Probabilities

In this section we consider a numerical illustration of multiple HT problem. As an application of multiple HT for Markov chains we can consider text classification problem.

Over the recent years, text classification has become one of the key techniques for organizing online information [13], [14]. It can be used to organize document databases, filter spam from people's email. A well-known approach to the automatic learning of linguistic categories is based on Elman's Simple Recurrent Network (SRN) [15].

The goal of text classification is the automatic assignment of documents to a fixed number of categories. In general, each document can be in multiple, exactly one, or no category at all. Using machine learning, the objective is to learn classifiers from examples which assign categories automatically. This is a supervised learning problem.

A desirable property of a feature is for its distribution to be highly dependent on the class. Words that occur independently of the class give no information for classification. A natural approach to developing a metric for filtering features is to determine whether each word has a class-independent distribution and to eliminate the word if it has such a distribution. In statistics, the procedure of determining whether data is generated from a particular distribution is known as hypothesis testing.

Let us create the following model of text categorization. Each document can be exactly in one category with class labels assigned to the documents. We propose to model English text as a Markov process where the probability of observing any text word is dependent on the previous word.

Suppose that a document is comprised of an ordered sequence of word events. Next we make a Markovity assumption: we assume that the probability of each word in the document is dependent of the previews word, but it is independent of its position in the document. In other words if we have vocabulary $X = \{x_1, \ldots, x_L\}$ each category of the document is described by the conditional probabilities matrix $G = \{G(x|u), u, x \in \mathcal{X}\}$. Now our goal is to assign each document to the appropriate category, based on the designed rules. So we have M hypothesis and based on sequence of words the classifier has to decide if a particular feature vector is likely to be drawn from a given category or not and try to minimize misclassifications (error probabilities).

For a good perception of the HT and text categorization theories it would be pertinent to discuss an example with the binary set $\mathcal{X} = \{0, 1\}$.

In the example we suppose that there are given two Markov sources with alphabet $\mathcal{X} = \{0, 1\}$.

Suppose an outcome of language research that enables a representation of different (3) languages genres reflected in the following transition matrices as hypothesis to test for each of two texts:

Example 1

$$H_1 : G_1 = \begin{pmatrix} 0.3 & 0.7 \\ 0.1 & 0.9 \end{pmatrix}, \quad H_2 : G_2 = \begin{pmatrix} 0.49 & 0.51 \\ 0.92 & 0.08 \end{pmatrix}, \quad H_3 : G_3 = \begin{pmatrix} 0.9 & 0.1 \\ 0.45 & 0.55 \end{pmatrix}.$$

Example 2

$$H_1 : G_1 = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}, \quad H_2 : G_2 = \begin{pmatrix} 0.49 & 0.51 \\ 0.92 & 0.08 \end{pmatrix}, \quad H_3 : G_3 = \begin{pmatrix} 0.9 & 0.1 \\ 0.45 & 0.55 \end{pmatrix}.$$

With the same success this kind of model (namely, conversation of research subjects to hypotheses which identify those subjects) can be considered for other areas of classification.

e.g., speech recognition, handwriting recognition, information retrieval, data compression, spam filtering, etc.. Normally in this kind of categorization problems the performance of algorithms is discussed in complexities point of view. In term of this example we would like to introduce a framework of problems where the quality of categorization of objects is considered via error exponents analysis.

For above mentioned hypotheses, applying Theorem 1. and definitions (8)-(10) we got values for all elements of reliability matrix, given fixed elements $E_{3,1|1,1}$, $E_{3,2|2,2}$, $E_{1,3|1,1}$, $E_{2,3|2,2}$. For a numerical experiments we generate a sequence of those reliability matrices in the following way. At first we initialize a matrix with fixed components equal to 0.01. By increasing of those values by step $\delta = 0.1$, so as to keep the (12) - (15) conditions valid, we got sequence of reliability matrixes. Based on that sequence we draw the surface of $(E_{1,1|2,2}, E_{1,2|1,2}, E_{2,1|2,1})$ in Fig. 1. Applying Lemma 1 for each object we get the planes $(E_{1|2}(E_{1|1}), E_{1|1})$ and $(E_{1|3}(E_{1|1}), E_{1|1})$ with the graphs in Fig. 2 and Fig. 3 (Fig. 4 and Fig. 5 in case of Example 2). respectively.

The surface in Fig. 1 illustrates the interdependence of reliabilities $E_{1,1|2,2}$, $E_{1,2|1,2}$ and $E_{2,1|2,1}$.

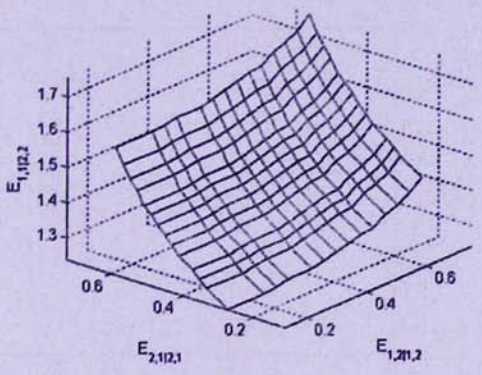

Fig. 1. Surface of $E_{1,1|2,2}$, $E_{1,2|1,2}$ and $E_{2,1|2,1}$ for Example 1.

Note that in Fig. 2 starting from the value of $E_{1|1} \approx 0.35$ the value of reliability $E_{1|2}(E_{1|1})$ decreases faster.

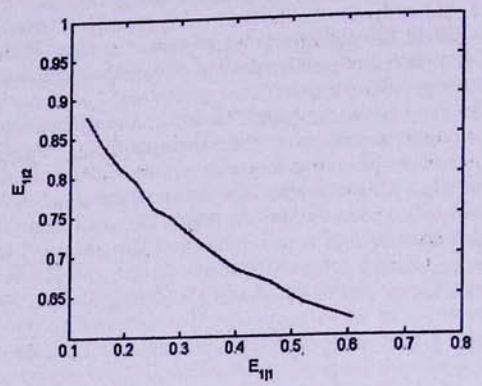

Fig. 2. Curve of $E_{1|2}(E_{1|1})$ for the first object of Example 1.

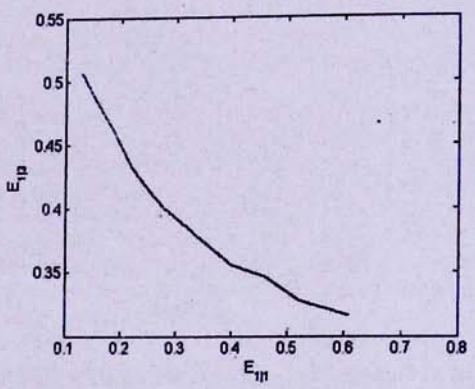

Fig. 3. Curve of $E_{1|3}(E_{1|1})$ for the second object of Example 1.

In Fig. 3 the value of reliability $E_{1|3}(E_{1|1})$ decreases faster starting from the value of $E_{1|1} \approx 0.25$. The last two figures show that, when one of the inequalities (12)-(15) is violated, then the element $E_{1|3}$ of reliability matrix tends to zero.
Now we will present the graphs concerning two second example:

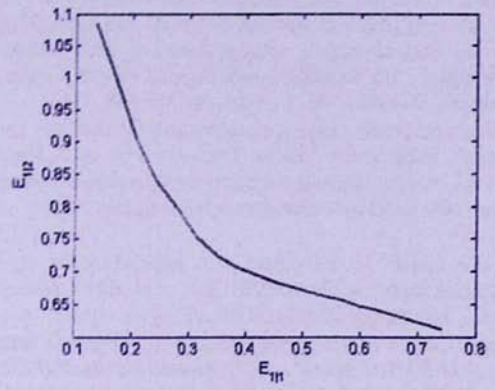

Fig. 4. Curve of $E_{1|2}(E_{1|1})$ for the first object of Example 2 .

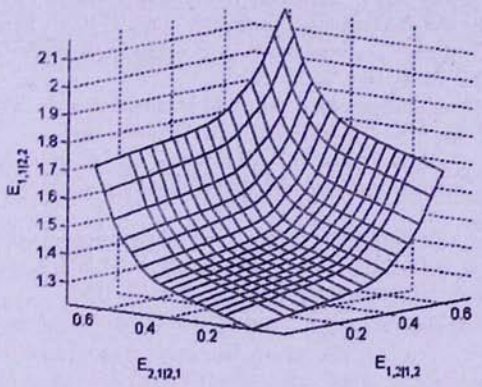

Fig. 5. Surface of $E_{1,1|2,2}$, $E_{1,2|1,2}$ and $E_{2,1|2,1}$ for the Example 2.
All calculation are made in MathLab environment.

References

[1] L. Birgé, "Vitess maximals de dćroissance des erreurs et tests optimaux associes", *Z. Wahrsch. Verw. Gebiete*, vol. 55, pp. 261–173, 1981.

[2] R. F. Ahlswede and E. A. Haroutunian, "On logarithmically asymptotically optimal testing of hypotheses and identification", *Lecture Notes in Computer Science*, vol. 4123," GeneralTheory of Information Transfer and Combinatorics", Springer, pp. 462 – 478, 2006.

[3] S. Natarajan, "Large deviations, hypotheses testing, and source coding for finite Markov chains", *IEEE Trans. Inform. Theory*, vol 31, no. 3, pp. 360-365, 1985.

[4] E. A. Haroutunian, "On asymptotically optimal criteria for Markov chains", (in Russian), *First World Congress of Bernoulli Society*, section 2, vol. 2, no. 3, pp. 153-156, 1989.

[5] E. A. Haroutunian, "Asymptotically optimal testing of many statistical hypotheses concerning Markov chain", (in Russian), *5-th Intern. Vilnius Conferance on Probability Theory and Mathem. Statistics*, vol. 1 (A-L), pp. 202-203, 1989.

[6] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics", IEEE Trans. Inform. Theory, vol 35, no 2, March, 401-408, 1989.

[7] E. A. Haroutunian, "Logarithmically asymptotically optimal testing of multiple statistical hypotheses", *Problems of Control and Information Theory*, vol. 19, no. 5-6, pp. 413–421, 1990.

[8] E. A. Haroutunian and P. M. Hakobyan, "On logarithmically asymptotically optimal hypotheses testing of three distributions for pair of objects", *Mathematical Problems of Computer Science*, vol. 24, pp. 76 - 81, 2005.

[9] E. A. Haroutunian and N. M. Grigoryan, "Reliability approach for testing of many distributions for pair of Markov chains", *Mathematical Problems of Computer Science*, vol. 28, pp. 109–116, 2007.

[10] E. A. Haroutunian, M. E. Haroutunian, and A. N. Harutyunyan. "Reliability Criteria in Information Theory and in Statistical Hypotheses Testing", *Foundation and Trends in Comunications and Information Theory*, vol. 4, no. 2–3, pp. 97–263. 2008.

[11] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*, Academic Press, New York, 1981.

[12] C. E. Shannon, "A Mathematical theory of communication", *Bell System Technical Journal* , vol. 27, pp. 379–423, 623–656, July, October, 1948

[13] P. F. Brown, V. J. Della Pietra, P. V. de Souza, J. C. Lai, R. L. Mercer, "Class-based n-gram models of natural language", *Computational Linguistics*, vol. 18(4), pp. 467–479, 1992.

[14] L. D. Baker and A. K. McCallum, "Distributional clustering of words for text classification", *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 96–103, 1998.

[15] J. L. Elman. "Finding structure in time", *Cognitive Science*, vol. 14, 179–211, 1990.

Բազմակի վարկածների ստուգման սխալների հավանականությունների ցուցիչների լուսաբանումը օրինակներով

Ն.Գրիգորյան

Ամփոփում

Ուսումնասիրվել է բազմակի վարկածների ստուգման խնդիրը երկու մարկովյան աղբյուրներից բաղկացած հաղորդակցության համակարգի դեպքում և ցույց են տրոշ կիրառություններ: Մասնակի օրինակի դեպքում կատարվել են թվային հաշվարկներ և ներկայացվել հուսալիության մատրիցի տարրերի գրաֆիկական պատկերներ: