PROBABILITY MODEL OF CONCEPTS RECOVERY IN SMALL SAMPLE LEARNING

Alexander A. Grusho, Nick A. Grusho,
Michael I. Zabezhailo and Elena E. Timonina
Federal Research Center
"Computer Science and Control"
of the Russian Academy of Sciences
Vavilova 44-2, 119333, Moscow, Russia
Email: grusho@yandex.ru, info@itake.ru,
zabezhailo@yandex.ru, eltimon@yandex.ru

Vladislav V. Kulchenkov
Lomonosov Moscow State University
Faculty of Computational
Mathematics and Cybernetics
Leninskiye Gory 1-52, 119991, Moscow, Russia
Email: vlad.kulchenkov@gmail.com

KEYWORDS

Small sample learning, random graph model for series of small samples, concept learning.

ABSTRACT

Many information security monitoring systems and controlling of IoT systems receive information in the form of short messages, which can be considered as small samples. Concepts are considered as classes of small samples that allow you to determine the correctness of monitoring systems. The paper is devoted to the problem of recovering concepts on observations of series of small samples.

Probabilistic model of appearance of series of small samples is introduced. To define concepts, the probabilistic dependency is used within series of small samples. The case of series of length 2 of small samples is considered. This assumption allowed the construction of a random graph and provided its probability-statistical analysis. Asymptotic approximations of probability distributions in the series scheme are used to identify ranges of parameter values that better define the structure of concepts. The set of parameter values is defined, at which the structure of concepts is uniquely determined with probability which tends to 1.

INTRODUCTION

Many information security monitoring systems (Grusho et al., 2015, 2019a) and controlling of IoT systems receive information in the form of short messages, which can be considered as small samples.

Unlike classical mathematical statistics, the meaning of a "small sample" does not involve obtaining information by statistical processing of data contained in one small sample. The information required for data analysis is intended to be obtained by statistical processing of a set of small samples.

However, many problems arise with small samples analysis (Etz and Arroyo, 2015). One major problem is that the distribution on the set of small samples may not be homogeneous even under conditions of independence of obtaining small samples.

Let 's consider the example of problems with small samples. Assume that the part of small samples (k of small samples) is obtained according to the distribution P_0, and the remaining m of small samples are obtained according to the distribution P_1, $P_0 \neq P_1$. If the problem of recovering samples with distribution $P_1$is considered, and the probability of "false alarm" (a sample from distribution P_0 is by mistake taken as the sample from distribution P_1) has to be made as little as possible, then if $m/k \to 0$ and $k \to \infty$ the probability to classify incorrectly the sample from distribution P_1, as a rule, is limited from below to a constant (Grusho et al., 2019a; Axelsson, 1999).

Due to the inertia of information systems, monitoring system messages are often received in series, reflecting the close states of a computer system or a network.

The purpose of monitoring systems is to identify anomalies in the operation of monitored objects. Technologies for detecting anomalies based on the construction of models of normal behavior are known (Tukey, 1977). However, a set of incoming messages does not always have simple structure (Grusho et al., 2016). It is not always possible to apply regression methods (Tukey, 1977). For example, when the network changes its behavior, many parameters of the network functioning change. If the device shows several modes of operation, it is necessary to base their description on analysis of incoming small samples using machine learning procedures.

Recently, a lot of machine learning methods have been developed (see, for example, (Jordan and Mitchell, 2015; Bramley, 2017)). Machine learning techniques based on small samples have also been studied in detail (Shu et al., 2019). One of the main scenarios in such learning is based on Concept Learning. In (Shu et al., 2019) concepts are called classes of small samples, which allow to determine the correctness of monitoring systems. The purpose of Concept Learning is to recognize concepts by a small number of small samples based on previously observed concepts. The second purpose of this approach is to recover a set of concepts. In the future, the terminology of small sample learning theory will be used, where concepts refer to classes of samples to which membership needs to be determined for newly arrived small samples. One more problem is the concept construction (Shu et al., 2019).

The problem of recovering of concepts is recognized

Communications of the ECMS, Volume 34, Issue 1,
Proceedings, ©ECMS Mike Steglich, Christian Mueller,
Gaby Neumann, Mathias Walther (Editors)
ISBN: 978-3-937436-68-5/978-3-937436-69-2(CD) ISSN 2522-2414

as one of the main challenges in the theory of machine learning (small sample learning). As noted above, because of the inertia of information systems, messages are often received in series, reflecting the close states of the computer system or the network.

Next, we assume that the data comes from the series of small samples. Each series is often associated with a single concept. At the same time, the number of concepts is unknown, but it is finite. Each concept will be described by a set of samples. The paper (Grusho et al., 2019b) deals with the case where the series are unambiguously related to the same concept. In this paper, the mathematical model of building a series of small samples is described as a random graph. This model reflects the possibility of series from different concepts. Analysis of this model helped to prove that asymptotically with probability which tends to 1, the set of series uniquely recovers concepts and their number.

MATHEMATICAL MODEL

Let the finite alphabet A be defined. Monitoring system messages are sent to the analysis system by short formalized messages reflecting the states of different sensors. For simplicity, we consider that messages have the same length r. However not all messages are admissible, i.e. there is a set of admissible messages $X \subseteq A^r$, $|X| = n$. Each message can be considered as a small sample. The analysis system can operate in several automatically switched modes. This corresponds to the existence on set X of several the priori existing concepts that we will denote through $M_1, ..., M_k$. The number of concepts k and concepts themselves are not defined.

The task of machine learning is to recover the concepts themselves on the set of accumulated small samples. The paper assumes that small samples with a high probability are received from any one concept on series of the length l. For simplicity, let us assume that $l = 2$.

Let time T be represented by the set of natural numbers. Series of small samples of the size $l = 2$ appear sequentially and independently of each other. If a series of small samples (x_1, x_2) consists of small samples belonging to the same concept, then the probability of this event equals to $\alpha > 0$. If x_1 belongs to one concept and x_2 belongs to another concept, then the probability of such series equals to $1 - \alpha > 0$. In other words, let the control sequence of 1 and 0 be obtained by the Bernoulli scheme with the probability of appearance of 1 equaling to α and the probability of appearance of 0 equaling to $1 - \alpha$. If the element of control sequence equals to 1 at the given time, a concept from one of the sets $M_1, ..., M_k$ is equiprobably selected, and a pair of small samples (x_1, x_2) is independently and equiprobably selected from this concept. If the element of control sequence equals to 0 at the given time, two different concepts are randomly and equiprobably selected, from each of which independently and equiprobably selects a small sample x_1 and a small sample x_2.

Suppose the series sequence is infinite. Then due to independent appearance of series from Borel-Cantelli's lemma follows that each pair (x_1, x_2), $x_1, x_2 \in X$, will be met in an infinite set of times.

Let's consider the random graph $G^{(n)}$ with nodes on the set X and with edges corresponding to appearance of each pair (x_1, x_2), $x_1, x_2 \in X$, at least once. Then it follows from the above that the graph $G^{(n)}$ is complete for any n. It follows that graph $G^{(n)}$ carries no information about the structure of concepts and their number.

Let $n \to \infty$. Note that in this case the sequence $G^{(n)}$ also carries no information about the structure of concepts and their number.

Let's consider now a great number of random graphs $\{G^{(n)}\}$ which are defined on X, $|X| = n$, by means of a chain of length T of randomly chosen series of small samples from the set X. The random graph $G_T^{(n)}$ may not be a complete graph. Therefore, in graphs $G_T^{(n)}$ some regularities may be revealed related to the structure of concepts and their number. Obviously, at small T, these regularities manifest weakly, and at large T, they disappear altogether. However, there is area of T, where these regularities can be identified. Using the set of graphs $\{G_T^{(n)}\}$, the algorithm for recovering the set of concepts of $M_1, ..., M_k$ and estimating their number k will be constructed.

PROPERTIES OF GRAPHS $G_T^{(n)}$

Let's assume that the number of concepts $k < \infty$ is fixed, and it is unknown. Each small sample from X belongs to some concept M_i, the numbers of small samples in concepts $M_1, ..., M_k$ satisfy the following conditions:

$$\begin{array}{l} |M_i| = n\varepsilon_i,\ 0 < \varepsilon_i < 1, \\ \Sigma_{i=1}^{k} \varepsilon_i = 1,\ i = 1, ... k \end{array} \tag{1}$$

Let a chain of series of small samples of length T be constructed. On this data the graph $G_T^{(n)}$ is defined. This graph is the source for the concept recover algorithm.

Lemma 1. Let $n \to \infty$, $\alpha = 1 + o(1)$,

$$T = \frac{(n \ln n)^2}{\alpha}(1 + o(1)),$$

then in the random graph $G_T^{(n)}$ the subgraphs formed by nodes M_i, $i = 1, ... k$, are complete graphs with the probability which tends to 1.

Proof. Consider the classic task of allocations of particles into boxes (Kolchin et al., 1978), where boxes are pairs of the kind $(x, x\prime)$, where x, $x\prime$ belong to the same concept, and different particles are placed in these boxes according to the probability scheme built above at the moments of time when there are units in the control sequence. For each concept M_i, the number of such pairs is equal to

$$\left(\begin{array}{c} |M_i| \\ 2 \end{array} \right).$$

Let in the control sequence of the length T and random choosing of the concept M_i gets N_i of units. If N is the number of all units in the sequence of the length T, then by the definition of the defined probability measure

$N_i = \frac{N}{k}(1 + o(1))$ with probability $1 + o(1)$, where the convergence of the probability to 1 is denoted by $1+o(1)$. At the same time

$$P(N = \alpha T(1 + o(1))) = 1 + o(1).$$

From here for all $i = 1, ..., k$

$$P(N_i = \frac{\alpha T}{k}(1 + o(1))) = 1 + o(1).$$

The mathematical expectation of the number of empty boxes μ_0 (i.e. the number of edges missing in the graph $G_T^{(n)}$ on the set of nodes M_i) in equiprobable scheme of allocations for each $i = 1, ..., k$, equals to (Kolchin et al., 1978):

$$E_i(\mu_0) = \left(\begin{array}{c} |M_i| \\ 2 \end{array} \right) \left(1 - \frac{1}{\left(\begin{array}{c} |M_i| \\ 2 \end{array} \right)} \right)^{N_i}. \quad (2)$$

It follows that simultaneously for all concepts the mathematical expectation of the number of empty boxes for all concepts is equal to $o(1)$.

By Markov's inequality (Shiryayev, 1984), the probability that at least one node will not fall into the corresponding complete graph is equal to o(1).

Thus, all concepts in $G_T^{(n)}$ produce complete graphs with probability which tends to 1. Lemma 1 is proved.

In complete graphs obtained in the random graph $G_T^{(n)}$ on subsets of nodes M_i, $i = 1, ...k$, the value of considered parameter T does not take into account the following possibility.

Let node x belong to the set M_i. However, it is possible that edges resulting from zeros in the control sequence will also allow the node x to be attached to the complete graph arising on another concept. Such a situation will be called a learning error, as it introduces ambiguity in the description of the membership of the node x to one of concepts.

The following lemma shows that the probability of any error tends to zero when

$$T = \frac{1}{\alpha}(n \ln n)^2(1 + o(1))$$

and

$$1 - \alpha = O(\frac{1}{\ln n}).$$

Lemma 2. Let $n \to \infty$, $1 - \alpha = O(\frac{1}{\ln n})$,

$$T = \frac{1}{\alpha}(n \ln n)^2(1 + o(1)).$$

Then the probability of any error, i.e. that there exists a node belonging to any two concepts, tends to zero.

Proof. Let node $x \in M_i$. Then x generates an error with the set M_j , $j \neq i$, if fixed $|M_j|$ edges corresponding to some zeros of the control sequence connect x to all nodes of the set M_j. The probability of this event for fixed zeros in the control sequence is equal to

$$\left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j|}.$$

The mathematical expectation of the number of such events on the set of zeros of the control sequence is equal to

$$\left(\begin{array}{c} T - N \\ |M_j| \end{array} \right) \left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j|}. \quad (3)$$

Then the mathematical expectation of the number of errors generated on the set M_i by edges connecting M_i with M_j is equal to

$$|M_i| \cdot \left(\begin{array}{c} T - N \\ |M_j| \end{array} \right) \left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j|}. \quad (4)$$

Let's use the inequality (Riordan, 1958)

$$\left(\begin{array}{c} m \\ r \end{array} \right) \leq \frac{m^r}{r^r(m - r)^{m-r}}.$$

Having substituted the received estimates and values $|M_i|$, $|M_j|$, we receive for some $\varepsilon > 0$ that the next estimate takes place for formula (4) in conditions of Lemma 2:

$$ne^{Cn} \left(\frac{\ln n}{n} \right)^{\varepsilon n} = o(1), \quad (5)$$

where $C > 0$, $C = const$. This estimation is true for every pair (i, j), $i \neq j$, of concepts. Lemma 2 is proved.

ALGORITHM FOR CONCEPTS RECOVERY

Let's denote via $\mathcal{B}$ the algorithm for concept recovery. Let $\mathcal{A}$ be the algorithm for allocating the maximum complete subgraph in a graph. The result of $\mathcal{A}$ in $G_T^{(n)}$ we will denote via $G_T^{(n)}(1)$. Delete then graph $G_T^{(n)}(1)$ from the graph $G_T^{(n)}$ (remove nodes of the graph $G_T^{(n)}(1)$ and the associated edges). In the remaining graph, using algorithm $\mathcal{A}$, we will allocate the maximum complete subgraph $G_T^{(n)}(2)$, etc. Thus, complete subgraphs $G_T^{(n)}(1), ..., G_T^{(n)}(\hat{k})$ define sets of nodes $\hat{M}_1, ..., \hat{M}_{\hat{k}}$. These are the first $\hat{k}$ steps of the algorithm $\mathcal{B}$.

After $\mathcal{A}$ has finished its work algorithm $\mathcal{B}$ has to correct the sets $\hat{M}_1, ..., \hat{M}_{\hat{k}}$. The procedure is as follows. If the node x does not belong to any of sets $\hat{M}_1, ..., \hat{M}_{\hat{k}}$, then we calculate its connectivity with each of these sets. Let $\hat{M}_i$ be the set having maximum connectivity to the node x. Attach this node to $\hat{M}_i$. Repeat this procedure with each node that does not belong to any of sets $\hat{M}_1, ..., \hat{M}_{\hat{k}}$. If ambiguity occurs, any of the allowed sets is arbitrarily selected. If x is an isolated node, then any of sets $\hat{M}_1, ..., \hat{M}_{\hat{k}}$ is arbitrarily selected.

After all the nodes from the set X are distributed across sets $\hat{M}_1, ..., \hat{M}_{\hat{k}}$, let's find the area of values of parameter T at which $k = \hat{k}$ and with probability which tends to 1, $M_i = \hat{M}_i$, $i = 1, ..., k$.

Lemma 3. With probability which tends to 1, for each $i = 1, ..., \hat{k}$ there exists a number $j = 1, ..., k$ such that $\hat{M}_i = M_j$.

Proof. Consider the set $\hat{M}_1$ and assume that the proposition of lemma 3 for $\hat{M}_1$ is not performed. This means that the set $\hat{M}_1$ consists of several subsets of

sets M_i, $i = 1, ..., k$. Since the graph $G_T^{(n)}(1)$ is the maximum complete subgraph, and the subgraphs of the graph $G_T^{(n)}$ generated on sets M_i, $i = 1, ..., k$, are complete graphs, the number of nodes in the graph $G_T^{(n)}(1)$ is greater than or equal to the number of nodes in each set M_i, $i = 1, ..., k$. That is, there is $\delta > 0$ such that $|\hat{M}_1| \geq n\delta$. Then there exists the number $j = 1, ..., k$ such that $|\hat{M}_1 \cap M_j| \geq n\varepsilon$, $\varepsilon > 0$. In fact, if the intersection $\hat{M}_1$ with each set M_i, $i = 1, ..., k$, is equal to $o(n)$, it contradicts the condition $|\hat{M}_1| \geq n\delta$ because $k < \infty$.

In addition, there is a node $x \in \hat{M}_1 \cap M_i$ for some $i \neq j$. For a pair $(x, \hat{M}_1 \cap M_i)$, an estimate similar to (3) is fair, namely

$$\left(\begin{array}{c} T - N \\ |M_j \cap \hat{M}_1| \end{array} \right) \left(\frac{1}{|M_i| \cdot |M_j|} \right)^{|M_j \cap \hat{M}_1|}, \qquad (6)$$

where $i \neq j$. It follows that for (6) the estimation (5) is fair. It means that there are no nodes in the set $\hat{M}_1$ that do not belong to M_j with probability which tends to 1.

According to the algorithm $\mathcal{B}$, the graph $G_T^{(n)}(1)$ is removed from the graph $G_T^{(n)}$ together with the set M_j. Since the following graph is the maximum complete graph on the set of nodes $X \setminus M_j$, all reasonings fair to $\hat{M}_1$ are true for $\hat{M}_2$. Then there is a number $i = 1, ..., k$ such that $\hat{M}_2 = M_i$ with probability which tends to 1.

Obviously, at the k-th step, the maximum complete subgraph is defined on the remaining set M_s and all elements of the set X are exhausted. It follows that $k = \hat{k}$ and after the corresponding renumbering, equality $M_i = \hat{M}_i$, $i = 1, ..., k$ is fair. Lemma 3 is proved.

Theorem. *Let* $n \to \infty$, $1 - \alpha = \frac{C}{\ln n}$, *where* $C = const$, $C > 0$,

$$T = \frac{1}{\alpha}(n \ln n)^2(1 + o(1)).$$

Then

$$P(k = \hat{k},\ M_i = \hat{M}_i,\ i = 1, ..., k) \to 1.$$

Proof. The proposition of the theorem follows from lemmas 1-3 and the description of the algorithm $\mathcal{B}$ of graphs formation $G_T^{(n)}(1), ..., G_T^{(n)}(\hat{k})$.

It follows from the theorem that at large n there is an area of values of T in which concepts are correctly recovered with great probability. At the same time, at very small and very large values of T, the structure of concepts cannot be determined in principle.

CONCLUSION

1. In the paper it is proved that in the process of using machine learning to recover concepts in the original small sample data, bad results are obtained not only when the set of accumulated samples is small, but also when the set of accumulated samples becomes very large. It is shown that there is a relation between the parameters of the number of possible small samples and the accumulated data, in which the structure of the set of concepts is uniquely determined with probability which tends to 1.

   It follows that it is not true that in all tasks increasing of the set of accumulated data improves the quality of learning. Therefore, it is necessary to highlight the parameter value areas when the quality of training is the best.

2. It is convenient to use asymptotic approximations of probability distribution in the series scheme to identify parameter value areas that define the structure of concepts better.

3. The paper does not consider estimates of the complexity of algorithms for the concepts recovery. Currently, there are many algorithms for allocating complete graphs and comparing the complexities of these algorithms is not part of the task of this paper. However, it should be noted that if it is possible to construct a linear order in each of recovered concepts, the task of classifying a newly incoming small sample is solved quite quickly by known algorithms.

4. The paper does not consider the possibility of building the best estimation of the probability of the correct recovering of the concepts. As noted in paragraph 1 of the Conclusion, the problem of proving the existence of an area of parameters in which the estimation of the probability of the correct recovery of concepts tends to 1 has been solved. However, preliminary studies have shown that limitations (1) on the power of concepts can be weaken.

Acknowledgements
This work was partially supported by the Russian Foundation for Basic Research (grant No. 18-29-03081).

REFERENCES

Grusho, A., N. Grusho, E. Timonina, and S. Shorgin. 2015. "Possibilities of secure architecture creation for dynamically changing information systems". *Systems and Means of Informatics* 25, No. 3, 78–93.

Grusho, A., N. Grusho, and E. Timonina. 2019. "The bans in finite probability spaces and the problem of small samples". In *Distributed computer and communication networks*. V.M. Vishnevskiy, K.E. Samouylov, and D.V. Kozyrev (Eds.), Lecture Notes in Computer Science, vol 11965. Springer, Cham, 578–590.

Etz, K. E., J. A. Arroyo. 2015. "Small Sample Research: Considerations Beyond Statistical Power". *Prev Sci*, 1033–1036.

Axelsson, S. 1999. "The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection". In *Proc. of the 6th Conference on Computer and Communications Security*, 1–7.

Tukey, J.W. 1977. *Exploratory data analysis*. Addison Wesley. 711 p.

Grusho, A., N. Grusho, and E. Timonina. 2016. "Detection of anomalies in non-numerical data". In *Proceedings of 8th Congress (International) on Ultra Modern Telecommunications and Control Systems and Workshops*. IEEE, Picataway, N.J., 273–276.

Jordan, M. I., and T. M. Mitchell. 2015. "Machine learning: Trends, perspectives, and prospects". *Science* 349, Iss. 6245, 255–260.

Bramley, J.W. 1977. *Constructing the world: Active causal learning in cognition*. London: University College London. PhD Thesis. 361 p.

Shu, J., X. Zongben, and M. Deyu. 2019. "Small sample learning in big data era". Available at: https://arxiv.org/abs/1808.04572.

Grusho, A. A., M. I. Zabezhailo, N. A. Grusho, and E. E. Timonina. 2019. "Concepts forming on the basis of small samples". *Informatics and applications* 13, Iss. 4, 81–84.

Kolchin, V. F., B. A. Sevast'yanov, V. P. Chistyakov. 1978. *Random allocations, Scripta Series in Mathematics*. V. H. Winston and Sons, Washington, DC, xi+262 p.

Shiryayev, A. N. 1984. *Probability*. Addison Wesley. 711 p. Translated from the Russian by R. P. Boas. Graduate Texts in Mathematics, 95. Springer-Verlag, New York, xi+577 p.

Riordan, John. 1958. *An introduction to combinatorial analysis*. John Willey and Sons, New York.

AUTHOR BIOGRAPHIES

ALEXANDER A. GRUSHO,Professor (1993), Doctor of Science in physics and mathematics (1990). He is principal scientist at Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences and Professor of Moscow State University.

Research interests: probability theory and mathematical statistics, information security, discrete mathematics, computer sciences.

His email is `grusho@yandex.ru`.

NICK A. GRUSHO has graduated from the Moscow Technical University. He is Candidate of Science (PhD) in physics and mathematics. At present he works as senior scientist at Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS).

Research interests: probability theory and mathematical statistics, information security, simulation theory and practice, computer sciences.

His email is `info@itake.ru`.

VLADISLAV V. KULCHENKOV has graduated from the Moscow State University, Faculty of Computational Mathematics and Cybernetics.

Research interests: probability theory and mathematical statistics, machine learning, optimization theory, data mining, financial risk.

His e-mail address is: vlad.kulchenkov@gmail.com.

ELENA E. TIMONINA has graduated from the Moscow Institute of Electronics and Mathematics and obtained the Candidate degree (PhD) in physics and mathematics (1974). She is Doctor in Technical Science (2005), Professor (2007). Now she works as leading scientist in Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (FRC CSC RAS).

Research interests: probability theory and mathematical statistics, information security, cryptography, computer sciences.

Her email is `eltimon@yandex.ru`.

MICHAEL I. ZABEZHAILO has graduated from the Institute of Physics and Technology and gained the Candidate degree (PhD) in theoretical computer science (1983). He is Doctor of Science in physics and mathematics (2016). Now he works as Head of laboratory in Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences.

Research interests: mathematical foundations of artificial intelligence, reasoning modeling, information security, theoretical computer sciences.

His email is: `zabezhailo@yandex.ru`.