МЕТОДИКА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ О СОБЫТИЯХ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ, СОДЕРЖАЩИХСЯ В НЕСТРУКТУРИРОВАННЫХ ПОЛЬЗОВАТЕЛЬСКИХ СООБЩЕНИЯХ

THE DATA MINING TECHNIQUE ABOUT INFORMATION SECURITY, CONTAINED IN UNSTRUCTURED USER MESSAGES

A. Gaynov
I. Zavodtsev

Summary. In work the technique of data mining on event security incident contained in unstructured user message, which through the introduction of additional procedures of filtering allows to increase the accuracy of classification of unstructured user message, that is, to determine the presence of data on the event security incident relating to one of the specified classes of information security incidents.

Keywords: an event security incident; an information security incident; unstructured user message; a SIEM-system.

Гайнов Артур Евгеньевич;
Соискатель, Кубанский институт информационной защиты
ArturGaynov@mail.ru
Заводцев Илья Валентинович;
К.т.н., доцент, Кубанский институт информационной защиты
nirls@mail.ru

Аннотация. В работе предложена методика интеллектуального анализа данных о событиях информационной безопасности (СИБ), содержащихся в неструктурированных пользовательских сообщениях (НПС), которая за счет введения дополнительной процедуры фильтрации позволяет повысить точность классификации НСП, то есть определять наличие в них данных о СИБ, относящихся к одному из заданных классов инцидентов информационной безопасности (ИИБ).

Ключевые слова: событие информационной безопасности; инцидент информационной безопасности; неструктурированное пользовательское сообщение; SIEM-система.

В условиях проведения масштабных компьютерных атак неотъемлемой частью непрерывного функционирования информационных систем любых предприятий и организаций является оперативное и точное обнаружение всех ИИБ [4, 9].

На сегодняшний день потребность сбора и обработки в информационных системах большого объема данных о СИБ обуславливает необходимость широкого использования SIEM-систем. Имея несомненные преимущества перед ручной обработкой данных, такие системы имеют ряд ограничений: они способны проводить анализ данных о СИБ, полученных исключительно из журналов регистрации базового и прикладного программного обеспечения (ПО), средств защиты информации [5–8, 10], а сведения, поступающие непосредственно от пользователей в них не учитываются. Это делает общую картину не полной, снижая точность анализа, а, следовательно, и полноту ее восприятия аналитиком (специалистом) информационной безопасности 1 линии.

При этом следует отметить, что не взирая отсутствие у пользователей необходимых знаний в области информационной безопасности и изложение ими проблем под воздействием эмоций [1–3], сами сведения могут содержать ценные данные о СИБ, а, следовательно, должны быть обработаны с заданными требованиями к точности, полноте и оперативности [4, 9].

Анализ широкого круга публикаций показал, что отсутствие проработанного методического аппарата не позволяет решать эту проблему в современных SIEM-решениях в полной мере. Поэтому в работе предлагается подход к обработке пользовательских данных, поступающих в виде неструктурированных сообщений на естественном русском языке, содержащих неточности и ошибки.

Цель статьи — разработка методики обработки НПС с использованием процедуры классификации на основе глубоких искусственных нейронных сетей (ИНС).

Пусть $\boldsymbol{ise}^{\boldsymbol{user}} \in ISE^{user}$ — единичное НПС, представляющее собой последовательность слов $\boldsymbol{ise}^{\boldsymbol{user}} = [x_1^{user}, \dots, x_i^{user}]$, где j — длина сообщения ise^{user}. Определим $K^{ISE} = \{k_1^{ISE}, k_2^{ISE}\}$ — совокупность классов НПС, где k_1^{ISE}– это класс, к которому относятся НПС, содержащие данные о СИБ, а k_2^{ISE}– это класс, к которому относятся НПС, не содержащие данные о СИБ, и $K^{ISI} = \{k_1^{ISI}, \dots, k_q^{ISI}\}$ — совокупность классов ИИБ, где $q = \overline{1, Q}$ — количество классов ИИБ

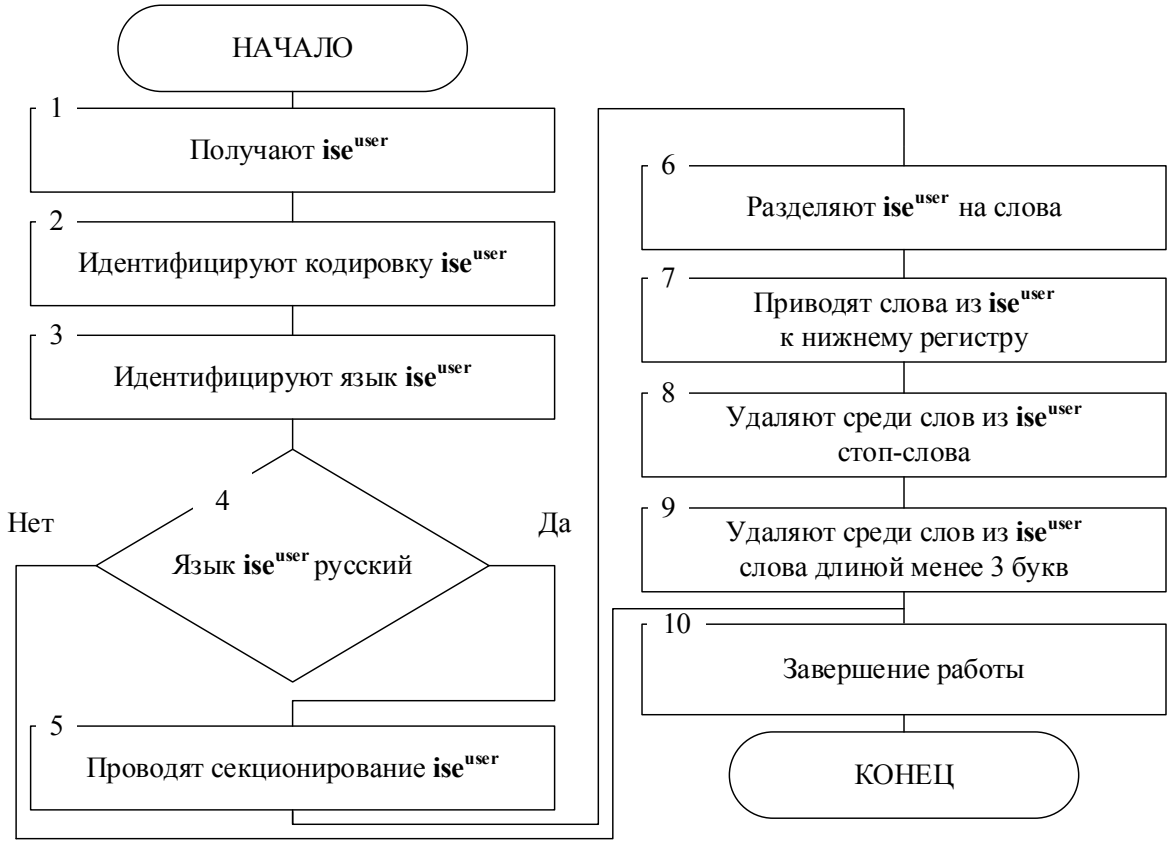

Рис. 1. Алгоритм предварительной обработки НПС

Тогда исходными данными методики выступают ise^{user}, K^{ISE} и K^{ISI}.

В качестве целевой функции определим функцию потерь (f_{los}) — меру расхождения между истинными и полученными значениями о принадлежности НПС к одному из классов ИИБ: $f_{los} \to 0$.

При этом функция потерь на этапе фильтрации — среднеквадратичная функция потерь (f_{los}^{fil}), а на этапе классификации — перекрестная энтропия (f_{los}^{clas}):

$$f_{los}^{fil} = {}^1/_2 \sum_{i=1}^{n} (d_i - y_i)^2, \quad (1)$$

где d_i — истинное значение о принадлежности НПС к K^{ISE}, y_i — фактическое значение о принадлежности НПС к K^{ISE}.

$$f_{los}^{clas} = -\sum_{i=1}^{n} d_i \cdot \log y_i, \quad (2)$$

где d_i — истинное значение о принадлежности НПС к K^{ISI}, y_i — фактическое значение о принадлежности НПС к K^{ISI}.

Рассмотрим алгоритм классификации НПС (рис. 1).

Предварительная обработка НПС: идентификация кодировки ($f_1^{p\_t}$) и языка сообщения ($f_2^{p\_t}$), секционирование сообщения ($f_3^{p\_t}$), разделение сообщения на слова ($f_4^{p\_t}$), приведение слов к нижнему регистру ($f_5^{p\_t}$), удаление среди них стоп-слов ($f_6^{p\_t}$) и слов длиной менее 3 букв ($f_7^{p\_t}$):

$$ise^{user\_p\_t} = f^{p\_t}(ise^{user}) =$$

$$= f_7^{p\_t} \circ f_6^{p\_t} \circ f_5^{p\_t} \circ f_4^{p\_t} \circ f_3^{p\_t} \circ f_2^{p\_t} \circ f_1^{p\_t}(ise^{user}). \quad (3)$$

1. Перевод слов из предварительно обработанного НПС в вещественное пространство признаков с использованием языковой модели Continuous Bag-of-Words:

$$\boldsymbol{ise}^{user\_vec} = f^{vec}(\boldsymbol{ise}^{user\_p\_t}). \quad (4)$$

2. Проведение фильтрации НПС (рис. 2) через определение содержания в нем данных о СИБ или их отсутствия:

$$f^{fil} : ISE^{user\_vec} \to K^{ISE}. \quad (5)$$

Для чего выполняется:

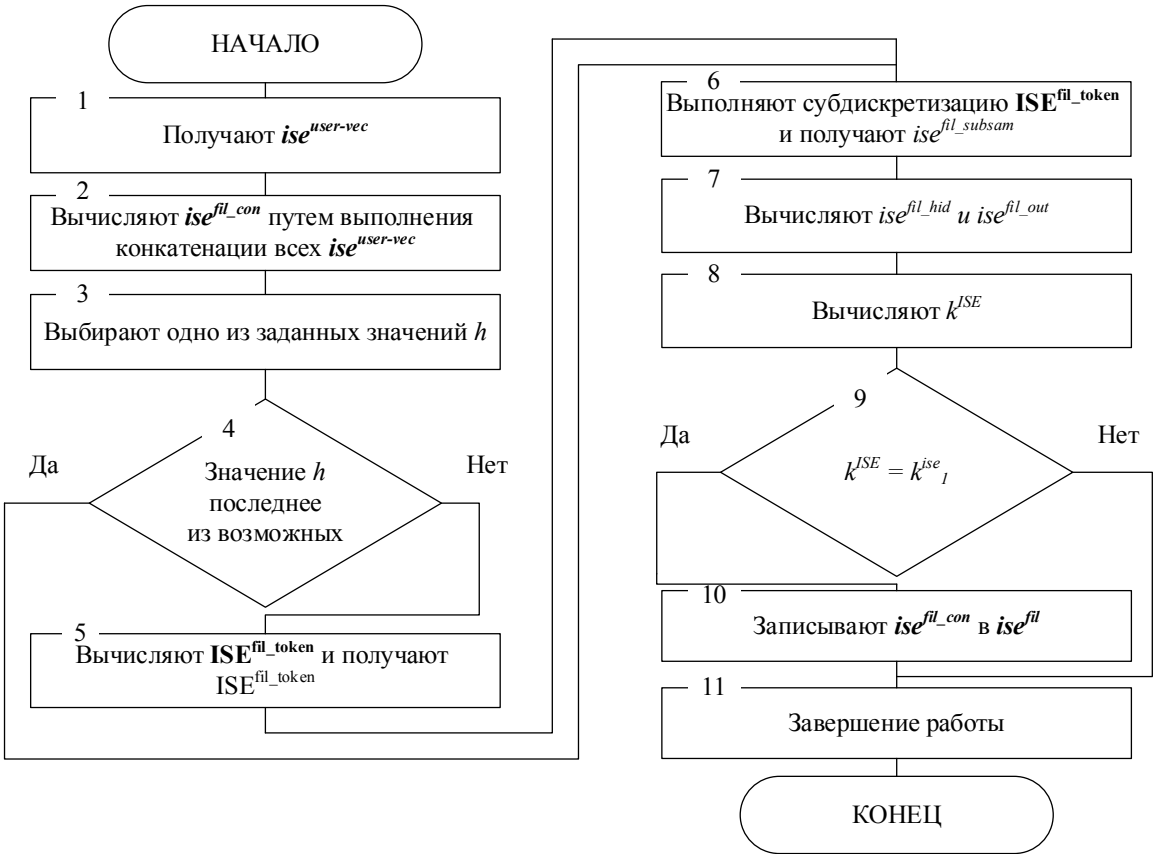

Рис. 2. Алгоритм фильтрации НПС

конкатенация входной последовательности из j векторных представлений слов:

$$\boldsymbol{ise}_{1:j}^{\boldsymbol{fil\_con}} = \\ = \boldsymbol{ise}_{\mathbf{1}}^{\boldsymbol{user\_vec}} \oplus \boldsymbol{ise}_{\mathbf{2}}^{\boldsymbol{user\_vec}} \oplus \ldots \oplus \boldsymbol{ise}_{\boldsymbol{j}}^{\boldsymbol{user\_vec}}, \quad (6)$$

где $\oplus$ — оператор конкатенации.

получение множества карт признаков $ISE^{fil\_token} = \left[\boldsymbol{ISE}_{\mathbf{1}}^{\boldsymbol{fil\_token}}, \ldots, \boldsymbol{ISE}_{\boldsymbol{n-h+1}}^{\boldsymbol{fil\_token}}\right]$.

Для этого путем осуществления операции свертки с применением фильтра $\boldsymbol{FIL}^{\boldsymbol{fil}} \in FIL^{fil}$, $\boldsymbol{FIL}^{\boldsymbol{fil}} \in \mathbb{R}^{hk}$, где h — параметр, определяющий размер окна, k — параметр, равный размерности $\boldsymbol{ise}_{\boldsymbol{j}}^{\boldsymbol{user\_vec}}$, вычисляется карта признаков $\boldsymbol{ISE}_{\boldsymbol{i}}^{\boldsymbol{fil\_token}}$:

$$\boldsymbol{ISE}_{\boldsymbol{i}}^{\boldsymbol{fil\_token}} = f^{fil\_token}\left(\boldsymbol{FIL}^{\boldsymbol{fil}} \cdot \boldsymbol{ise}_{\boldsymbol{i:i+h-1}}^{\boldsymbol{fil\_con}} + b\right), \quad (7)$$

где f^{fil_token} — функция активации ReLU, $b \in \mathbb{R}$ — параметр смещения, который, как и h, задается специалистами службы информационной безопасности предприятия.

Применяя (7) ко всем возможным выборкам слов из (6) и изменяя параметр h, вычисляется ISE^{fil_token}.

выполнение субдискретизация ISE^{fil_token} с применением операции max-over-time polling:

$$ise_i^{fil\_subsam} = f^{fil\_subsam}\left(\boldsymbol{ISE}_{\boldsymbol{i}}^{\boldsymbol{fil\_token}}\right), \quad (8)$$

где $ise_i^{fil\_subsam}$ — скаляр, полученный путем взятия максимального элемента из части $\boldsymbol{ISE}_{\boldsymbol{i}}^{\boldsymbol{fil\_token}}$, f^{fil_subsam} — операция max-over-time polling.

За счет повторения (8) для всех $\boldsymbol{ISE}_{\boldsymbol{i}}^{\boldsymbol{fil\_token}}$ формируются карты подвыборочного слоя $ISE^{fil\_subsam} = \left[\boldsymbol{ISE}_{\mathbf{1}}^{\boldsymbol{fil\_subsam}}, \ldots, \boldsymbol{ISE}_{\boldsymbol{n-h+1}}^{\boldsymbol{fil\_subsam}}\right]$.

подача выходных значений подвыборочного слоя на входы многослойного персептрона (нейроны каждой $\boldsymbol{ISE}_{\boldsymbol{i}}^{\boldsymbol{fil\_subsam}}$ связаны с одним нейроном многослойного персептрона), после чего вычисляются выходные значения скрытого и выходного слоев многослойного персептрона и определяется содержатся или нет в НПС данные о СИБ:

$$ise_i^{fil\_hid} = f^{fil\_hid}\left(\sum_j ise_j^{fil\_subsam} \cdot \boldsymbol{W}_{\boldsymbol{j,i}}^{\boldsymbol{fil\_subsam}} +\right.$$

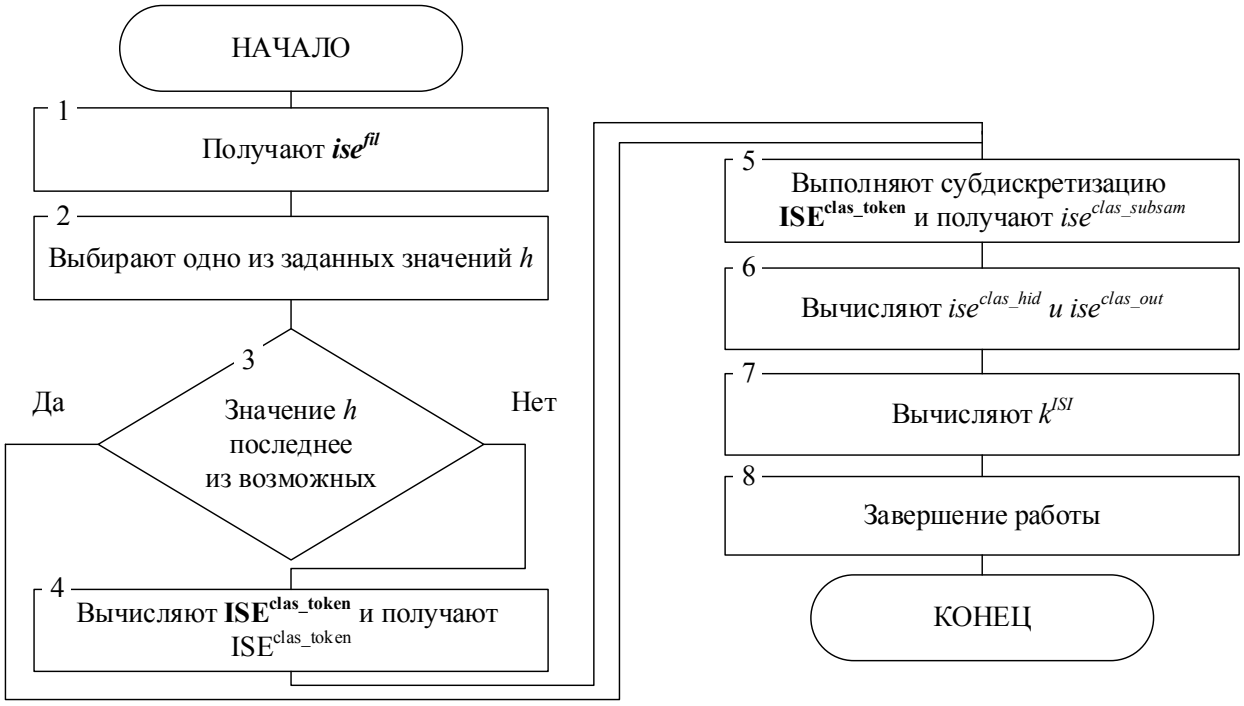

Рис. 3. Алгоритм классификации НПС

$$+ b^{fil\_subsam}\Big), \quad (9)$$

где $ise_i^{fil\_hid}$ — выходное значение нейрона скрытого слоя многослойного персептрона, f^{fil_hid} — функция активации ReLU, $\boldsymbol{W}_{j,i}^{\boldsymbol{fil_{subsam}}}$ — матрица весов скрытого слоя многослойного персептрона, b^{fil_subsam} — параметр смещения скрытого слоя многослойного персептрона.

$$ise^{fil\_out} =$$

$$= f^{fil\_out}\left(\sum_i ise_i^{fil\_hid} \cdot \boldsymbol{W}^{\boldsymbol{fil\_out}} + b^{fil\_out}\right), \quad (10)$$

где ise^{fil_out} — выходное значение нейрона выходного слоя многослойного персептрона, f^{fil_out} — функция активации гиперболический тангенс, $\boldsymbol{W}^{\boldsymbol{fil\_out}}$ — матрица весов выходного слоя многослойного персептрона, b^{fil_out} — параметр смещения выходного слоя многослойного персептрона.

$$k^{ISE} = \begin{cases} k_1^{ISE}, ise^{fil\_out} = 1 \\ k_2^{ISE}, ise^{fil\_out} = -1 \end{cases} \quad (11)$$

При $k^{ISE} = k_1^{ISE}$ $\boldsymbol{ise^{fil}} = \boldsymbol{ise_{1:j}^{fil\_con}}$, $\boldsymbol{ise^{fil}} \in ISE^{fil}$.

Проведение классификации НПС (рис. 3), то есть отнесение к одному из заранее известных классов ИИБ: $f^{clas} : ISE^{fil} \rightarrow K^{ISI}$. Для чего выполняются следующие этапы:

получение множества карт признаков $ISE^{clas\_token} = \left[\boldsymbol{ISE}_1^{\boldsymbol{clas\_token}}, \ldots, \boldsymbol{ISE}_{n-h+1}^{\boldsymbol{clas\_token}}\right]$.

Для этого путем осуществления операции свертки с применением фильтра $\boldsymbol{FIL}^{\boldsymbol{clas}} \in FIL^{clas}$, $\boldsymbol{FIL}^{\boldsymbol{clas}} \in \mathbb{R}^{hk}$, где h — параметр, определяющий размер окна, k — параметр, равный размерности $\boldsymbol{ise}_j^{\boldsymbol{user\_vec}}$, вычисляется карту признаков $\boldsymbol{ISE}_i^{\boldsymbol{clas\_token}}$:

$$\boldsymbol{ISE}_i^{\boldsymbol{clas\_token}} =$$

$$= f^{clas\_token}\left(\boldsymbol{FIL}^{\boldsymbol{clas}} \cdot \boldsymbol{ise}_{i:i+h-1}^{\boldsymbol{fil}} + b\right), \quad (12)$$

где f^{clas_token} — функция активации ReLU, $b \in \mathbb{R}$ — параметр смещения, который, как и h, задается специалистами службы информационной безопасности предприятия.

Применяя (12) ко всем возможным выборкам слов из $\boldsymbol{ise^{fil}}$ и изменяя параметр h, получить ISE^{clas_token}.

выполнение субдискретизации ISE^{clas_token} с применением операции max-over-time polling:

$$ise_i^{clas\_subsam} = f^{clas\_subsam}\left(\boldsymbol{ISE}_i^{\boldsymbol{clas\_token}}\right), \quad (13)$$

где $ise_i^{clas\_subsam}$ — скаляр, полученный путем взятия максимального элемента из части $\boldsymbol{ISE}_i^{\boldsymbol{clas\_token}}$, f^{clas_subsam} — операция max-over-time polling.

Таблица 1. модели машинного и глубокого обучения, созданные и обученные в ходе эксперимента

Модель машинного обучения	Модель перевода текста на естественном языке в вещественное пространство признаков				
	Bag of Words	Bag of Words & TF IDF	Bag of Ngrams & TF IDF		Continuous Bag-of-Words
			n = 2	n = 3	
Логистическая регрессия	№ 1	№ 2	№ 3	№ 4	№ 5
k ближайших соседей	№ 6	№ 7	№ 8	№ 9	№ 10
Опорные векторы	№ 11	№ 12	№ 13	№ 14	№ 15
RNN	-	-	-	-	№ 16
LSTM	-	-	-	-	№ 17
CNN	-	-	-	-	№ 18
Предложенная методика	-	-	-	-	№ 19

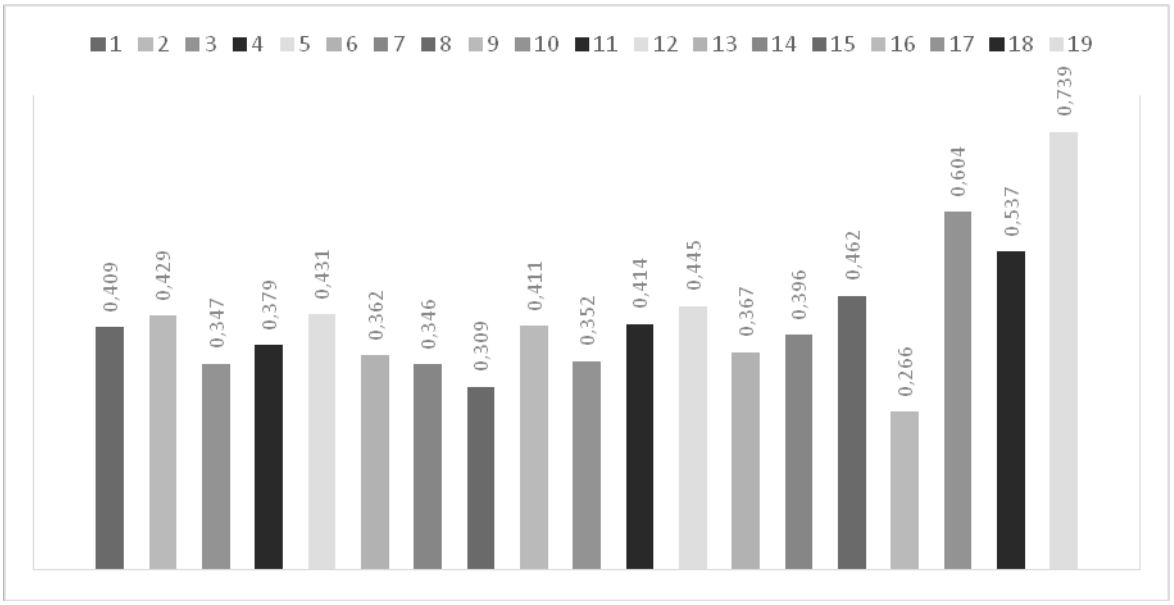

Рис. 4. Гистограмма значений точности

Повторяя *(13)* для всех $\boldsymbol{ISE_i^{clas\_token}}$, сформируются карты подвыборочного слоя $ISE^{clas\_subsam} = \left[\boldsymbol{ISE_1^{clas\_subsam}}, \ldots, \boldsymbol{ISE_{n-h+1}^{clas\_subsam}}\right]$.

подача выходных значений подвыборочного слоя на входы многослойного персептрона (нейроны каждой $\boldsymbol{ISE_i^{clas\_subsam}}$ связаны с одним нейроном многослойного персептрона), после чего вычисляются выходные значения скрытого и выходного слоев многослойного персептрона и определить к какому классу ИИБ относится НПС, содержащие данные о СИБ:

$$ise_i^{clas\_hid} = \\ = f^{clas\_hid}\left(\sum_j ise_j^{clas\_subsam} \cdot \boldsymbol{W_{j,i}^{clas\_subsam}} + \right. \\ \left. + b^{clas\_subsam}\right), \quad (14)$$

где $ise_i^{clas\_hid}$ — выходное значение нейрона скрытого слоя многослойного персептрона, f^{clas_hid} — функция активации ReLU, $\boldsymbol{W_{j,i}^{clas\_subsam}}$ — матрица весов скрытого слоя многослойного персептрона, b^{clas_subsam} — параметр смещения скрытого слоя многослойного персептрона.

$$ise^{clas\_out} = \\ = f^{clas\_out}\left(\sum_i ise_i^{clas\_hid} \cdot \boldsymbol{W^{clas\_out}} + b^{clas\_out}\right), \quad (15)$$

где ise^{clas_out} — выходное значение нейрона выходного слоя многослойного персептрона, f^{clas_out} — функция активации softmax, $\boldsymbol{W^{clas\_out}}$ — матрица весов выходного слоя многослойного персептрона, b^{clas_out} — параметр смещения выходного слоя многослойного персептрона.

Таким образом, НПС, содержащее данные о СИБ, относится к тому классу ИИБ, для которого выходное значение нейрона из выходного слоя многослойного персептрона будет максимальным:

$k^{ISI} = k_i^{ISI}$ при

$$ise_i^{clas\_out} = max\left(ise_1^{clas\_out}, \dots, ise_q^{clas\_out}\right). \quad (16)$$

Для оценки эффективности предложенного научно-методического аппарата проведен эксперимент, в ходе которого обучено 19 моделей машинного обучения (таблица 1).

С использованием сформированной выборки НСП (10000 сообщений) было проведено тестирование. Полученные результаты (рис. 4) показали, что точность определения НПС, содержащих данные о СИБ, при применении предлагаемого научно-методического аппарата составляет 0,739 ± 0,002, что на 0,135 лучше относительно моделей № 1–18.

При этом значения показателей оперативности и полноты у предлагаемого научно-методического аппарата не уступают значениям, полученным при использовании остальных моделей.

Таким образом, предлагаемая методика обработки НПС с использование процедуры классификации на основе глубоких ИНС с более высокой эффективностью может быть использована для обеспечения непрерывности функционирования информационных систем различных предприятий и организаций.

ЛИТЕРАТУРА

1. Антипов М. А. Виртуальные коммуникации как атрибут постсовременности / М. А. Антипов // Сборник конференции НИЦ Социосфера. 2013. № 55. С. 30–33.
2. Антипов М. А. Клиповое мышление как атрибут техногенного общества / М. А. Антипов // XXI век: итоги прошлого и проблемы настоящего плюс. 2015. № 6 (28). Т. 2. С. 20–28.
3. Гольдштейн Б. С. Call-центры и компьютерная телефония / Б. С. Гольдштейн, В. А. Фрейкман. — СПб.: БХВ-Петербург, 2014. — 368 с.: ил.
4. ГОСТ Р ИСО/МЭК ТО 18044–2007 «Информационная технология. Методы и средства обеспечения безопасности. Свод норм и правил менеджмента информационной безопасности».
5. Коноваленко С. А. Анализ систем мониторинга вычислительных сетей / С. А. Коноваленко, И. Д. Королев // Молодой ученый. 2016. № 23 (127). С. 66–73.
6. Коноваленко С. А. Базовые функциональные возможности существующих систем мониторинга вычислительных сетей / С. А. Коноваленко, И. Д. Королев, Д. А. Новоселов // Приволжский научный вестник. 2016. № 12–1 (64). С. 65–70.
7. Котенко И. В. Технологии управления информацией и событиями безопасности для защиты компьютерных сетей / И. В. Котенко, И. Б. Саенко, О. В. Полубелова, А. А. Чечулин // Проблемы информационной безопасности. Компьютерные системы. 2012. № 2. С. 57–68.
8. Котенко И. В. Применение технологии управления информацией
9. и событиями безопасности для защиты информации в критически важных инфраструктурах / И. В. Котенко, И. Б. Саенко, О. В. Полубелова, А. А. Чечулин // Труды СПИИРАН. 2012. Вып. 1 (20). С. 27–56.
10. Милославская Н. Г. Управление инцидентами информационной безопасности и непрерывностью бизнеса / Н. Г. Милославская, М. Ю. Сенаторов, А. И. Толстой. — М.: Горячая линия-Телеком, 2014. — 170 с.: ил.
11. Ниязаров Т. Какой часовой механизм точнее, или сравнение SIEM-решений / Т. Ниязаров // Jet Info. 2015. № 8 (265). С. 13–25.

Журнал «Современная наука: актуальные проблемы теории и практики»