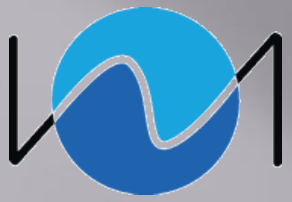


10-Я МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ «ИНТЕЛЛЕКТУАЛИЗАЦИЯ ОБРАБОТКИ ИНФОРМАЦИИ»

Идентификация потенциально опасных клиентских запросов на основе многоклассового распознавания образов

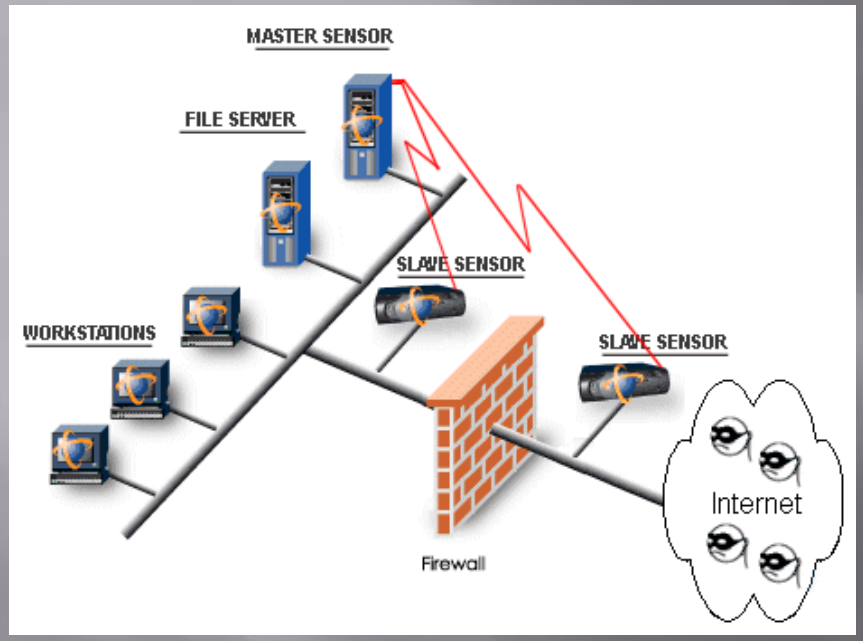
Сычугов Алексей Алексеевич xru2003@list.ru
Моттль Вадим Вячеславович
Середин Олег Сергеевич
Маленичев Антон Александрович
Тула, ФГБОУ ВО «Тульский государственный университет»,
Кафедра «Информационная безопасность»

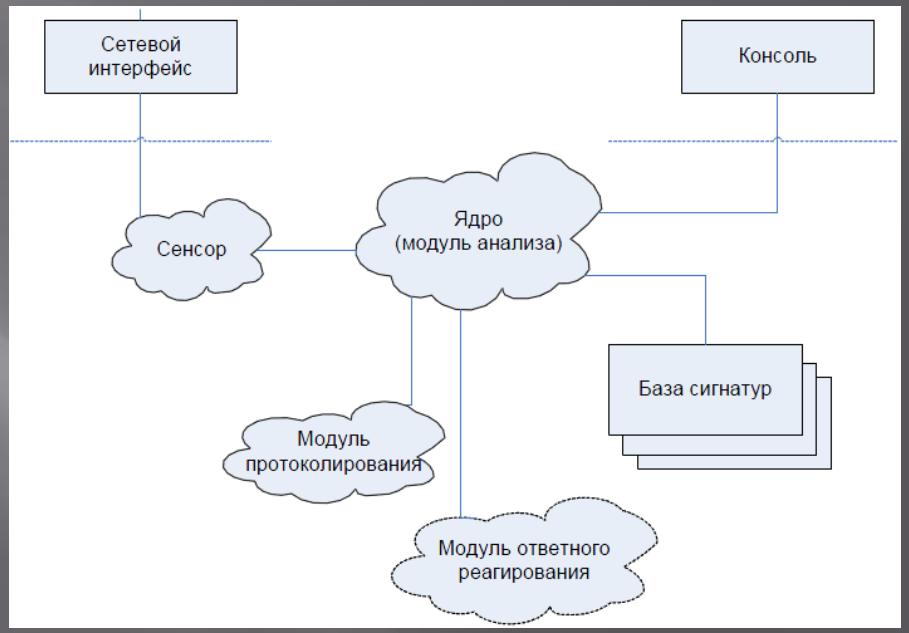
о. Крит, Греция

Вводные замечания

- Важность сетевых технологий неоспорима
- Велики временные и материальные затраты на создание гарантированно защищенных систем
- Современные средства ИБ разнообразны и надежны
- Системы обнаружения вторжений/системы предотвращения вторжений (IDS/IPS) – одно из основных средств обеспечения информационной безопасности

Системы обнаружения вторжений и системы предотвращения вторжений





- Обнаружение злоупотреблений и аномалий
- Различные методы построения, основной из которых – сигнатурный.

Системы обнаружения вторжений и системы предотвращения вторжений

- **Проблема** – необходимость постоянной корректировки сигнатур и уязвимость к модифицированным атакам при использовании сигнатурного метода в IDS/IPS
- **Необходимость решения этой проблемы** обуславливается постоянно растущим объемом трафика, который обрабатывается узлами сети, а также возрастающей активностью и изощренностью сетевых вторжений.
- **Цель** – ускорение настройки IDS/IPS и повышения её эффективности в процессе непрерывного анализа сетевого трафика и обнаружения новых видов вторжений.
- **Задача** – автоматизация процесса корректировки и дополнения списка сигнатур сетевых атак

Математическая модель исследуемого процесса

$\omega \in \Omega$ – множество всех возможных входных заявок

всякая заявка характеризуется индексом ее принадлежности к одному из нескольких классов $y(\omega): \Omega \to Y = \{0,1,...,m\}$

$Y_0 = Y_{безопасные}$ - класс безопасных заявок

$Y_1 \cup ... \cup Y_m = Y_{опасные}$ - совокупность классов опасных заявок

Задача

При поступлении очередной заявки $\omega \in \Omega$

определить ее класс, т.е. выработать оценку $\hat{y}(\omega): \Omega \to Y = \{0,1,...,m\}$

С учетом возможных ошибок в определении класса (неуверенность), выработать вектор апостериорных вероятностей

$$\pi^k(\omega) \geq 0 \qquad \sum_{k=0}^{m} \pi^k(\omega) = 1$$

Математическая модель исследуемого процесса

Всякая заявка представлена конечным числом ее непосредственно вычисляемых характеристик (вектором числовых признаков заявки):

$$\mathbf{x}(\omega) = \left(x_i(\omega), ..., x_n'(\omega) \right) \in R^n$$

Тогда задача формирования правила «размытой» классификации заявки сводится к поиску подходящей функции

$$\hat{\pi}(\mathbf{x}): R^n \to R^m$$

Имеется конечная обучающая совокупность $\Omega^* = \{\omega_1, ..., \omega_N\}$, в которой объективно определена связь между вектором признаков заявки и ее принадлежностью к одному из классов

$$\left\{ (\mathbf{x}_j, y_j),\ j = 1, ..., N \right\} \qquad \mathbf{x}_j = \mathbf{x}(\omega_j),\ y_j = y(\omega_j),\ , \omega_j \in \Omega^*,\ y_j \in \{0, 1, ..., m\}$$

Введем также отдельное обозначение для подмножества объектов класса *k* в обучающей совокупности

$$\Omega_k^* = \left\{ \omega_j \mid y(\omega_j) = k \right\} \subset \Omega^*$$

Таким образом, задача построения классификатора заявок сводится к задаче обучения многоклассовому распознаванию образов.

Многоклассовое распознавание образов

Сложность этой вычислительной задачи в условиях прикладной задачи распознавания опасных заявок определяется следующими обстоятельствами

1) Задача многоклассового распознавания образов изучена на порядок меньше, чем задача двухклассового распознавания.

$k = 0,1,...,m$ - целые числа, обозначающие разные классы объектов (заявок)

Для каждой пары классов $(k,l) \in Y \times Y \quad k < l$ необходимо построить отдельный классификатор, способный каждому предъявленному вектору признаков $\mathbf{x}(\omega) \in R^n$ поставить в соответствие пару вероятностей $p_{kl}(\mathbf{x}) + p_{lk}(\mathbf{x}) = 1$

$$p_{kl}\left(\mathbf{x}(\omega)\right) = P\left(y(\omega) = k \mid y(\omega) = k \text{ либо } y(\omega) = l\right) =$$
$$1 - p_{lk}\left(\mathbf{x}(\omega)\right) = 1 - P\left(y(\omega) = l \mid y(\omega) = k \text{ либо } y(\omega) = l\right).$$

$(1/2)m(m+1)$ - общее количество пар классов пар классов,

Будет использоваться вероятностная версию метода обучения Support Vector Machine SVM[1)] , определяющую для каждой пары классов вероятностную решающую функцию $p^k(\mathbf{x}) = 1 - p^l(\mathbf{x}) \quad \mathbf{x} \in R^n$

[1)] Татарчук А.И. Байесовские методы опорных векторов для обучения распознаванию образов с управляемой селективностью отбора признаков. Дисс. к.ф.-м.н. ВЦ РАН, 2014.

Вероятностная версия SVM

Первая составляющая заключается в решении обычной задачи SVM, т.е. в определении по двухклассовой обучающей совокупности направляющего вектора и сдвига дихотомической дискриминантной гиперплоскости $\hat{\mathbf{a}}_{kl} \in R^n \quad \hat{b}_{kl} \in R$ позволяющей принимать качественные решения в пользу одного из классов

$$\hat{\mathbf{a}}_{kl}^T \mathbf{x} + \hat{b}_{kl} \begin{cases} > 0 \Rightarrow \hat{y}(\mathbf{x}) = k, \\ < 0 \Rightarrow \hat{y}(\mathbf{x}) = l. \end{cases}$$

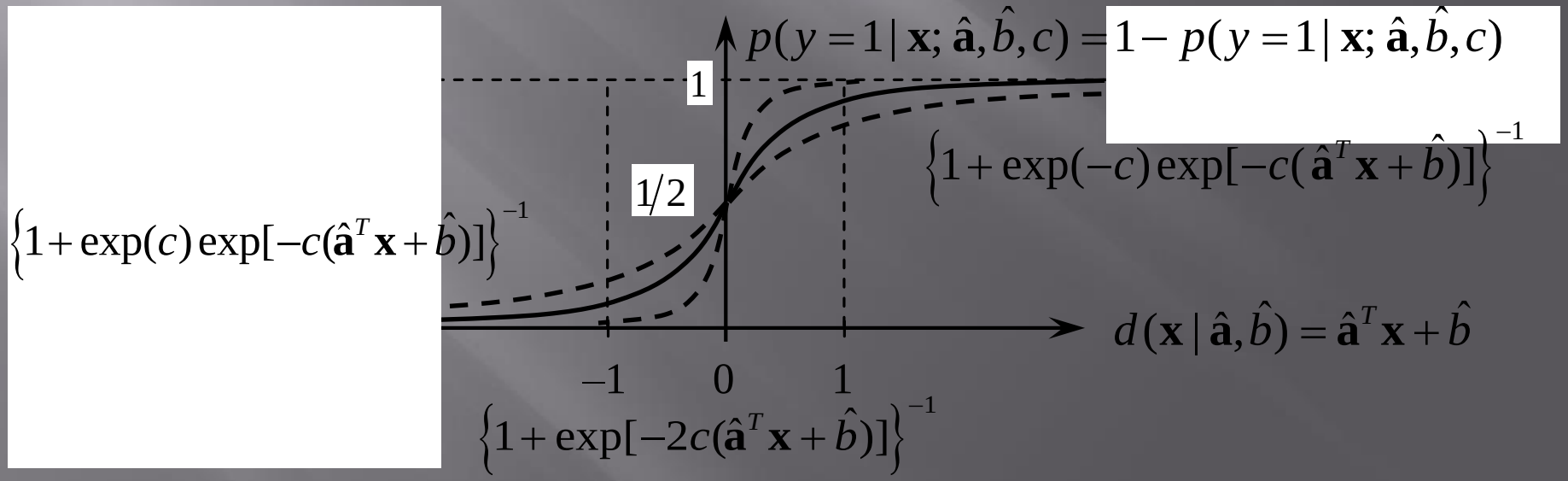
Параметры $(\hat{\mathbf{a}}_{kl}, b_{kl})$ находятся как решение задачи SVM для подмножества объектов обучающей совокупности, отнесенных «учителем» к соответствующим классам

$$\mathbf{a}_{kl}^T \mathbf{a}_{kl} + c \sum_{\substack{j:\, \omega_j \in \Omega_k^* \cup \Omega_l^* \\ y_{j,kl}(\mathbf{a}_{kl}^T \mathbf{x}_j + b_{kl}) < 1}} \left(1 - y_{j,kl}(\mathbf{a}_{kl}^T \mathbf{x}_j + b_{kl})\right) \to \min(\mathbf{a}_{kl}, b_{kl}) \qquad y_{j,kl} = \begin{cases} 1, \omega_j \in \Omega_k^*, \\ -1, \omega_j \in \Omega_l^*, \end{cases}$$

Вероятностная версия SVM

Вторая составляющая заключается в формировании пары дихотомических апостериорных вероятностей классов для любого нового объекта, представленного вектором признаков $\mathbf{x} \in R^n$

$$p_k(\mathbf{x}) = \begin{cases} \left\{1+\exp(c)\exp[-c(\hat{\mathbf{a}}_{kl}^T\mathbf{x}+\hat{b}_{kl})]\right\}^{-1}, \hat{\mathbf{a}}_{kl}^T\mathbf{x}+\hat{b}_{kl} < -1, \\ \left\{1+\exp[-2c(\hat{\mathbf{a}}_{kl}^T\mathbf{x}+\hat{b}_{kl})]\right\}^{-1}, -1 \le \hat{\mathbf{a}}_{kl}^T\mathbf{x}+\hat{b}_{kl} \le 1, \\ \left\{1+\exp(-c)\exp[-c(\hat{\mathbf{a}}_{kl}^T\mathbf{x}+\hat{b}_{kl})]\right\}^{-1}, \hat{\mathbf{a}}_{kl}^T\mathbf{x}+\hat{b}_{kl} > 1, \end{cases} \qquad p_l(\mathbf{x}) = 1 - p_k(\mathbf{x})$$



Объединение результатов обучения

Правило объединения результатов обучения отдельных классификаторов описаны ранее [2)]

Если предположить, что для каждого $\mathbf{x} \in R^n$ существует апостериорное распределение на множестве классов объекта с таким вектором признаков

$$\pi_k(\mathbf{x}),\ k = 0,1,\ldots,m \quad \sum_{k=0}^{m} \pi_k(\mathbf{x}) = 1$$

то полностью определены и дихотомические апостериорные вероятности

$$p_{kl}(\mathbf{x}) = \frac{\pi_k(\mathbf{x})}{\pi_k(\mathbf{x}) + \pi_l(\mathbf{x})} \qquad p_{lk}(\mathbf{x}) = \frac{\pi_l(\mathbf{x})}{\pi_k(\mathbf{x}) + \pi_l(\mathbf{x})}$$

Можно доказать, что обратное соотношение выражается формулами

$$\hat{\pi}_k(\mathbf{x}) = \left(1 + \sum_{l=0}^{k-1} \frac{p_{lk}(\mathbf{x})}{1 - p_{lk}(\mathbf{x})} + \sum_{l=k}^{m} \frac{1 - p_{kl}(\mathbf{x})}{p_{kl}(\mathbf{x})}\right)^{-1} \qquad k = 0,1,\ldots,m$$

Не зависимо от совместности или несовместности вероятностей, необходимо их нормировать к единичной сумме

$$\pi_k(\mathbf{x}) = \frac{\hat{\pi}_k(\mathbf{x})}{\sum_{l=0}^{m} \hat{\pi}_l(\mathbf{x})} \qquad k = 0,1,\ldots,m$$

---

2)Двоенко, С.Д., Моттль В.В., Середин О.С. Процедура распознавания образов для случая многих классов, опирающаяся на совокупность функций степени достоверности для пар классов. Известия ТулГУ, серия «Вычислительная техника, автоматика и управление», 1999, т. 2, вып. 2, с. 28-35.

Многоклассовое распознавание образов

Второе обстоятельство заключается в том, что число объектов в обучающей совокупности может оказаться очень большим (несколько миллионов)

В этом случае традиционный способ численного решения задачи SVM через двойственную задачу становится неприемлемым, поскольку двойственная задача является задачей квадратичного программирования относительно множителей Лагранжа при ограничениях, соответствующих объектам обучающей совокупности, а ее вычислительная сложность является полиномиальной относительно числа переменных.

Для преодоления этого препятствия нами разработан алгоритм прямого решения задачи SVM путем итерационного градиентного спуска непосредственно в пространстве $(\mathbf{a}_{kl}, b_{kl})$, имеющий линейную вычислительную сложность относительно числа обучающих объектов.

Следует отметить, что сходимость этого метода теоретически пока не доказана, однако, выполнен эксперимент на модельных данных, который позволил убедиться в адекватности этого метода путем сравнения результатов с классическим SVM.

Экспериментальное исследование

Для эксперимента были взяты данные конкурса KDD CUP (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html), которые представляют собой клиентские запросы, характеризуемые $n = 41$ признаком.

Каждый запрос имеет метку класса, общее количество классов $Y = 23$, из которых один класс – безопасные запросы, остальные 22 – опасные, которые, в свою очередь распределены по 4 группам:

группа 1: DOS – атаки;

группа 2: R2L – неавторизованный доступ от удаленных пользователей (например, подбор пароля);

группа 3: U2R – неавторизованный доступ от локальных суперпользователей (например, атака «переполнение буфера»);

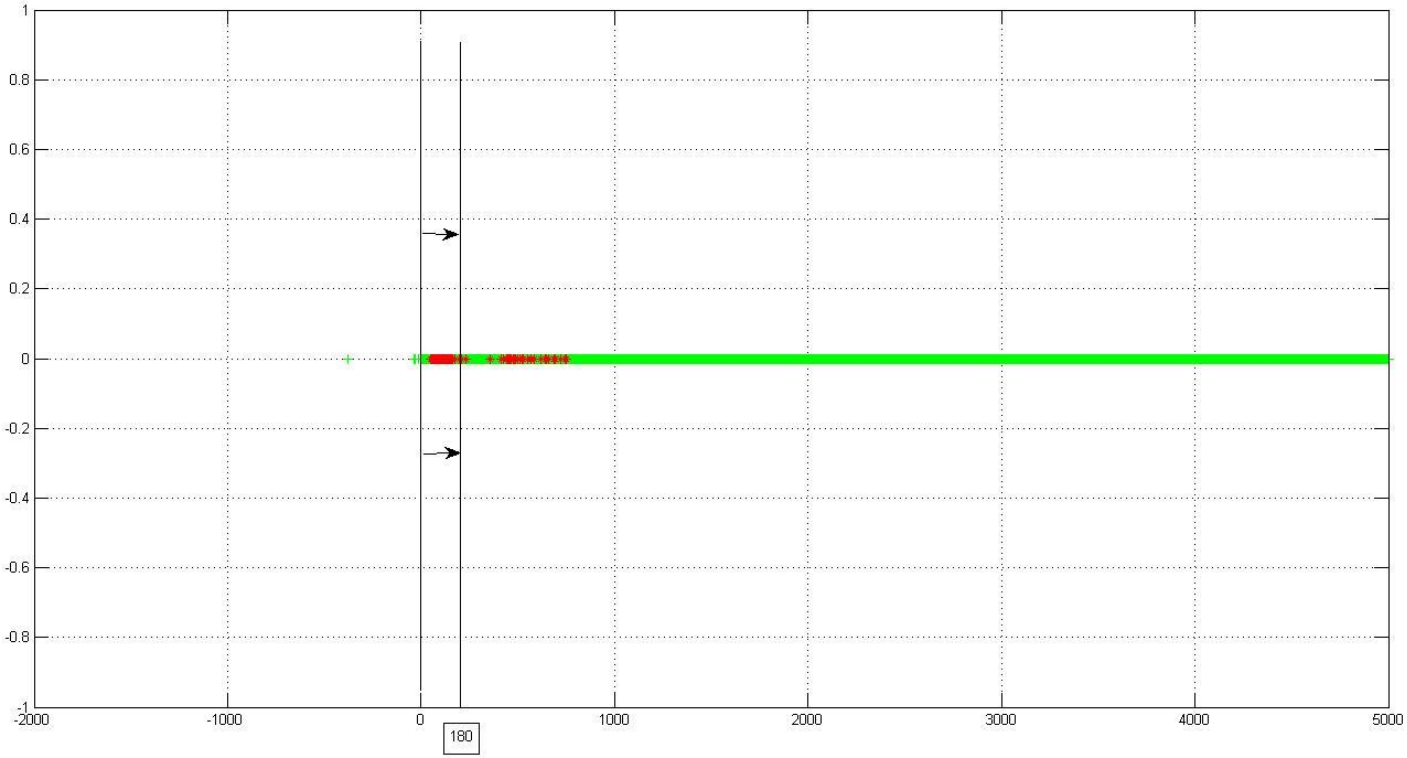
Группа 4: probing – пассивные воздействия (например, сканирование портов)).

Общее количество запросов около 5 млн., из которых около $N = 2$ млн. взяты в качестве обучающей совокупности, которая формировалась методом кросс-валидации, обеспечивающим её репрезентативность.

Экспериментальное исследование

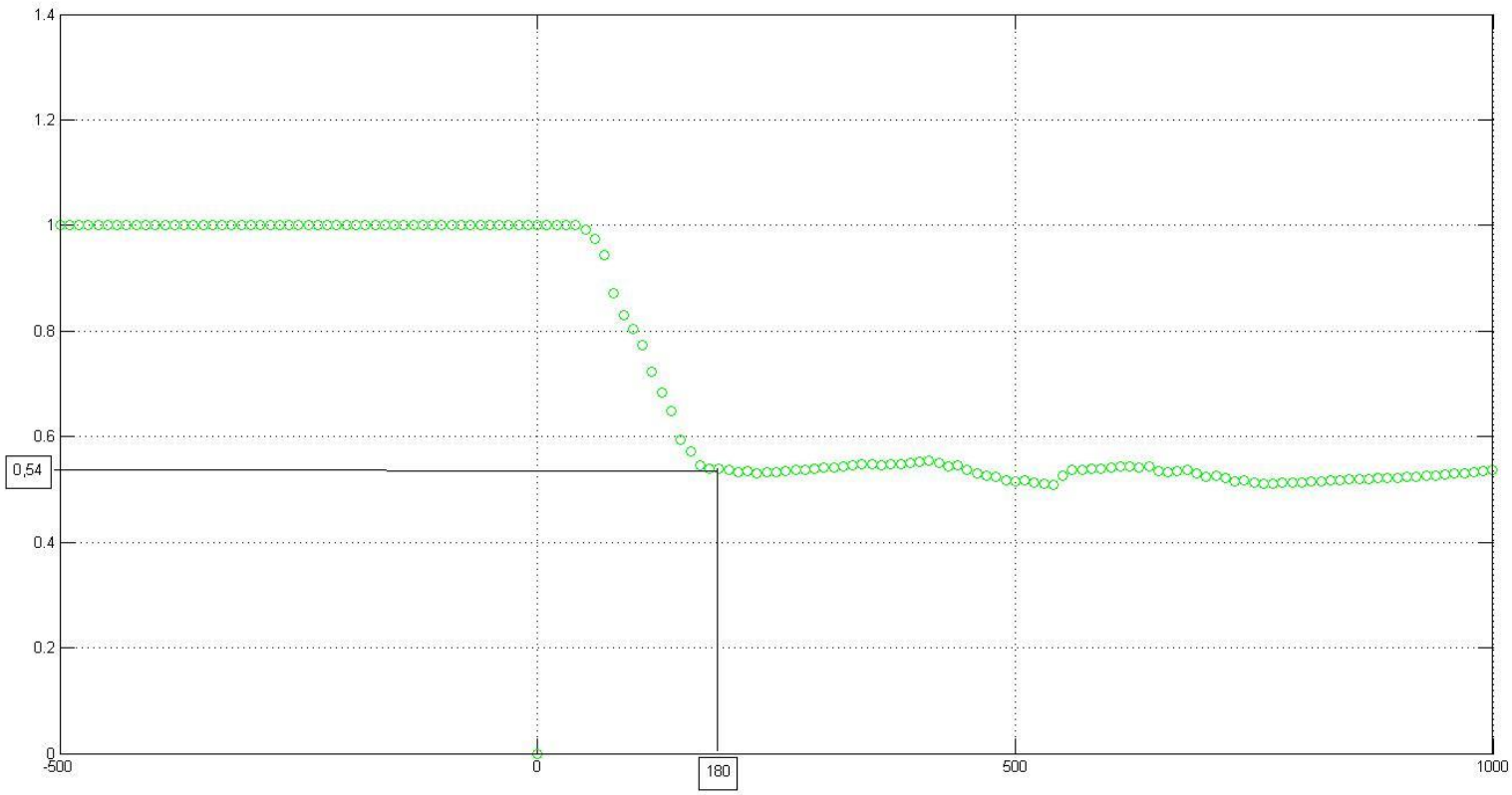
Группа	Кол-во объектов в группе	% объектов в группе	Класс	Кол-во объектов в классе	% объектов в классе
0	972781	19,859	1	972781	19,85903
1	41102	0,839	10	10413	0,21258
			11	12481	0,25480
			16	15892	0,32443
			18	2316	0,04728
2	3883390	79,278	5	1072017	21,88491
			6	2807886	57,32215
			8	264	0,00539
			9	979	0,01999
			12	21	0,00043
			14	2203	0,04497
			20	20	0,00041
3	59	0,001	2	30	0,00061
			3	9	0,00018
			4	3	0,00006
			19	7	0,00014
			23	10	0,00020
4	1099	0,022	7	53	0,00108
			13	8	0,00016
			15	12	0,00024
			17	4	0,00008
			21	1020	0,02082
			22	2	0,00004

Экспериментальное исследование



Вероятность распознавания объектов 0 группы – 99,5 -> 98%
Вероятность распознавания объектов 4 группы – 0 -> 56 %

Экспериментальное исследование



Сумма ошибок 1-го и 2-го рода.

СПАСИБО ЗА ВНИМАНИЕ

Сычугов Алексей Алексеевич
К.т.н., доцент, зав. каф. «Информационная безопасность»
ФГБОУ ВО «Тульский государственный университет»
xru2003@list.ru
Контактный телефон: 8-960-594-88-53

10-я Международная конференция «Интеллектуализация обработки информации»
4-11 октября 2014 г., о. Крит, Греция