УДК 004.056.5

Обнаружение сетевых атак по анализу статистики проходящего трафика

М.С. Рыжов [1,2] **, А.П. Овсянников** [3]

[1] Московский физико-технический институт (государственный университет)

[2] Институт проблем управления им. В.А. Трапезникова РАН

[3] Межведомственный суперкомпьютерный центр РАН

Была поставлена задача реализовать программу, всесторонне исследующую проходящий сетевой траффик, и создать штатную систему обнаружения активных атак и несанкционированного доступа к сетевым ресурсам на основе анализа статистики удалённого доступа методами машинного обучения. Традиционно выделяются 3 типа задач машинного обучения, и для того, чтобы провести качественный анализ свойств, были реализованы самые популярные представители каждого типа:

- обучение с учителем - *логистическая регрессия*, *градиентный бустинг*
- обучение без учителя - *K-Means*, *агломеративная кластеризация*
- частичное обучение - *LabelPropagation, LabelSpreading*

Характеристикой качества построенного анализатора принята доля верно определённых вредоносных сетевых данных среди их полного числа на известной атаке подбора пароля.

Логистическая регрессия

Метод строит линейный алгоритм классификации $a: X \to Y$ вида

$$a(x, w) = sign(\sum_{i=1}^{n} w_i f_i(x) - w_0) = sign < x, w >$$

где w_j — вес j-го признака, w_0 - порог принятия решения, $w = (w_1, ..., w_n)$ - вектор весов, $f_i(x)$ - i-ый признак объекта, $\langle x,w \rangle$ - скалярное произведение признакового описания объекта на вектор весов . Вектор весов выбирается из условия минимизации полинома ошибки на обучающей выборке, с известным набором результатов $\{y_i\}$ [1]:

$$Q(w) = \sum_{i=1}^{n} \ln(1 + \exp(-y_i < x_i, w >))$$

Полученные результаты, показали, что метод был недостаточно точен (~50%).

Градиентный бустинг

Алгоритм является последовательным объединением базовых алгоритмов, линейных классификаторов в данном случае. Пусть к некоторому моменту обучены N − 1 алгоритмов $b_1(x), ..., b_{N-1}(x)$,тогда композиция имеет вид $A_{N-1}(x) = \sum_{i=1}^{N-1} b_i(x)$. Затем к текущей композиции добавляется еще один алгоритм $b_N(x)$. Этот алгоритм обучается так, чтобы как можно сильнее уменьшить ошибку композиции на обучающей выборке:

$$\sum_{i=1}^{l} L(y_i, A_{N-1}(x_i) + b_N(x_i)) \to \min_b$$

где l – размер обучающей выборки, x – элементы выборки, $L(y, x)$ – функция ошибки [1].

Метод показал результаты не лучше, чем логистическая регрессия при различных размерах обучающей выборки (~50%).

K-Means

Для выборки $(x_1,...,x_n)$ метод строит кластеры $S=(S_1,...,S_k), k \le n$. Если у кластеров центры $m=(m_1,...,m_k)$, метод оптимизирует среднее внутриклассовое расстояние, меняя состав кластеров на каждой своей итерации [1, 2]:

$$\mu = \min_S \frac{1}{n}\sum_{i=1}^{n} \| m - x_i \|$$

Метод K-Means показал следующие результаты, при разных количествах итераций алгоритма: при количестве итераций меньше, чем 70, то алгоритм выдает 70% точность предсказания, при количестве итераций ~70 – точность может достигать 90\%, при количестве итераций от 70 до 100 – точности не будет хуже, чем 57%, при 100 итерациях – 87%, при прочих – 80%. Это говорит о том, что данные имеют чётко выраженные центры кластеризации.

Агломеративная кластеризация

Инициализируются множества кластеров из одной точки $C0 = \{\{x1\},\ldots,\{xl\}\}$. На каждой итерации $\forall t = 1,\ldots,l$ ищутся два ближайших кластера, которые потом объединяются в новый слитый кластер $W = U \cup V, C_t = C_{t-1} \cup \{W\} \setminus \{U,V\}$ [1,3].

Результаты работы алгоритма приведены в таблице 1.

LabelPropagation, LabelSpreading

Алгоритмы LabelPropagation и LabelSpreading – графовые алгоритмы, для которых не важны зависимости по данным и расположение центров кластеров. Определяющими факторами для них являются классы соседних объектов выборки в пространстве признаков – объект будет принадлежать классу, которому принадлежит наибольшее число его соседей [1,3]. Исследования показали, что увеличивая число соседей, можно достичь точности ~96\%, что делает их идеальными методами для анализаторов.

Таблица 1. Точность метода агломеративной кластеризации при различных типах расстояний и норм.

Норма/Расстояние	L1	L2	Манхэттенская	Косинусное сходство	Евклидова
Расстояние Уорда	-	-	-	-	0,85
Между всеми объектами класса	0,5005	0,5005	0,5005	0,805	-
Среднее расстояние	0,5005	0,5005	0,5005	0,691	-

Литература

1. *Рыжов М.С., Овсянников А.П.* Обнаружение вторжений по статистике проходящего трафика, 2016.
2. *E.M. Mirkes* K-means and K-medoids applet. - University of Leicester, 2011.
3. *Usha Nandini Raghavan, Reka Albert, Soundar Kumara* - Near linear time algorithm to detect community structures in large-scale networks, 19 Sep 2007