СКРЫТАЯ МАРКОВСКАЯ МОДЕЛЬ ДЛЯ ПРИНЯТИЯ РЕШЕНИЙ О КЛАССИФИКАЦИИ ВРЕМЕННОГО РЯДА СОБЫТИЙ В КОМПЬЮТЕРНОЙ СЕТИ

Бубнов Я.В.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Иванов Н.Н. – канд. физ.-мат. наук., доцент

В статье предлагается метод решения проблемы бинарной классификации событий, наблюдаемых в компьютерной сети, и представленных в виде временного ряда с помощью скрытой марковской модели. Метод позволяет принять решение об отнесении серии результатов индивидуальной классификации каждого события от зашумленных детекторов к одному из двух классов.

Обнаружение вредоносной активности узлов корпоративной компьютерной сети является одной из важных прикладных задач. Ее решение позволяет предотвратить целевые кибератаки, направленные как на отказ системы в целом, так и на кражу конфиденциальной информации. На практике данная задача решается с помощью детекторов [1, 2], анализирующих содержимое передаваемых через сеть пакетов, или детекторов занимающиеся сбором системной информации непосредственно с узлов сети. Собранная информация сохраняется в виде временных рядов в таких системах событийного мониторинга, как: Prometheus, Zabbix, Nagios [3]. Оперирование подобными системами предполагает установку допустимых диапазонов значений, при которых обеспечивается нормальное функционирования системы. При выходе анализируемых значений за пределы установленных границ, операторы получают уведомление о необходимости ручного вмешательства.

Как правило, допустимые границы выбираются опытным путем [4], к тому же, в некоторых случаях действия оператора строго регламентированы и поддаются автоматизации. Таким образом, ставится задача определить в какой момент времени требуется вмешательство операторов, для ее решения предлагаем использовать скрытую марковскую модель.

Рассмотрим частный случай детекторов, вычисляющий вероятность отнесения передаваемого сетевого пакета к вредоносному трафику, другими словами – бинарный классификатор. Тогда наблюдаемым событием $Y_t \in \mathbf{I}$ является результат работы классификатора для монотонно возрастающего времени $t \in \mathbf{Z}$. Определим марковскую сеть, представленную на рисунке 1, где каждому наблюдаемому событию поставлено в соответствие два возможных состояния, или класса.

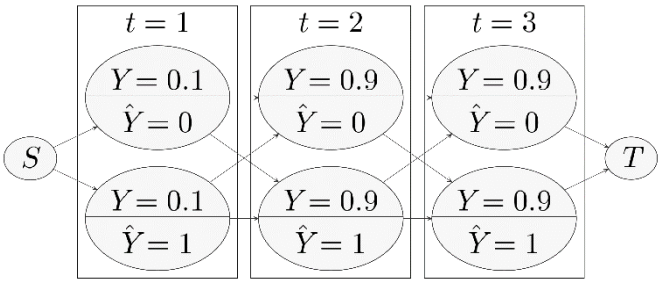

Рисунок 1 – Схема цепи Маркова серии наблюдаемых событий в компьютерной сети, где Y_t – вероятность отнесения наблюдения к одному из классов $\hat{\mathbf{Y}}$

Определим вероятность нахождения в конкретном состоянии i с помощью нормального распределения:

$$p_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{w}{\sigma}\right)^2}, \quad (1)$$

$$w = Y_i - \hat{Y}, \quad (2)$$

где среднеквадратическое отклонение составляет $\sigma = \sqrt{1/5}$. Другими словами, вероятнее всего скрытое состояние достоверно наблюдается, когда выход детектора находится в пределах интервала дисперсии эталонного класса.

Условимся, что наблюдаемые события образуют ординарный поток однородных событий, а значит могут быть описаны процессом Пуассона. Таким образом, определим вероятность перехода из состояния i в состояние j с помощью экспоненциального распределения:

$$p_{i,j} = e^{-z}, \tag{3}$$

$$z = \frac{t_i - t_j}{e^{1-Y_i \oplus Y_j}}, \tag{4}$$

где $Y_i \oplus Y_j$ представляет собой полином Жегалкина – то есть, наиболее вероятны переходы из одинаковых классов, а переходы между различными классами дополнительно штрафуются.

Пусть задан некоторый интервал упорядоченных времен $t \in \mathbf{T}$, состоящий из N элементов, тогда задача определения результирующего класса события по наблюдениям в границах $\mathbf{T}$ заключается в нахождении наиболее вероятного пути из состояния t_0 в одно из состояний t_N.

Для поиска наиболее вероятного пути в цепи Маркова, представленной на рисунке 1, воспользуемся алгоритмом Витерби и определим следующие рекуррентные соотношения:

$$V_{1,n} = p_1, \tag{5}$$

$$V_{t,n} = \max_i^N \left(p_t \cdot p_{i,n} \cdot V_{t-1,i} \right). \tag{6}$$

Тогда скрытое конечное состояние определяется из уравнения:

$$x_N = \arg\max_i^N \left(V_{t_N,i} \right) \tag{7}$$

На самом деле нас в первую очередь интересует не реальные состояния цепи Маркова, а вероятность, что серия наблюдаемых событий относится к конкретному классу. Решение данной задачи напрямую: путем подсчета количества наблюдаемых классов в восстановленной последовательности, приводит к результату, который не учитывает временную локальность: события, произошедшие недавно в прошлом, влияют на результат больше, чем события из далекого прошлого.

Для решения описанной проблемы введем два виртуальных состояния в цепь Маркова, где два конечных состояния – равновероятные. Новая цепь Маркова представлена на рисунке 2.

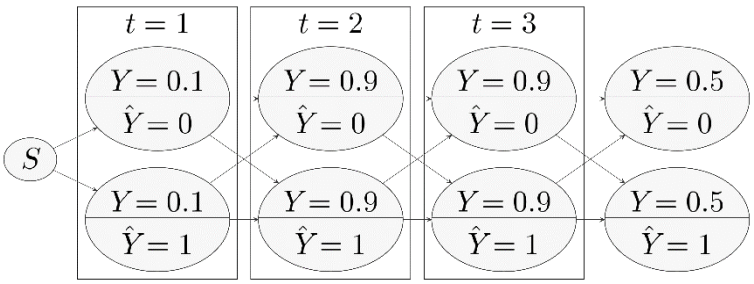

Рисунок 2 – Схема цепи Маркова серии наблюдаемых событий с равновероятными конечными состояниями

Вычислим вероятности нахождения в каждом из конечных состояниях с помощью алгоритма Витерби:

$$P_N^{\hat{Y}=0} = V_{t_N,N_0}, \tag{8}$$

$$P_N^{\hat{Y}=1} = V_{t_N,N_1}. \tag{9}$$

Тогда результирующий класс рассматриваемого временного ряда событий вычисляется по формуле:

$$Y = \arg\max(P_N^{\hat{Y}=0}, P_N^{\hat{Y}=1}) \tag{10}$$

Таким образом, в работе представлен бинарный классификатор временных рядов с обучением без учителя. Данный подход на практике позволяет принять решение о блокировке узла, при повторяющихся во времени событий с вредоносной активностью. Описанный метод при необходимости легко обобщается на многоклассовые задачи классификации.

Список использованных источников:

1. Qi, C. A bigram based real time DNS tunnel detection approach / C. Qi, X. Chen, C. Xu, J. Shi, P. Liu // Procedia Computer Science, Elsevier B.V. – 2013. – Vol. 17, P. 852-860.

2. Born, K. Detecting DNS Tunneling Using Character Frequency Analysis / K. Born, D. Gustafson // Proceedings of the 9th Annual Security Conference, Las Vegas 7-8 Apr 2010. – Las Vegas, 2010. - P. 2-3.

3. Skvortsov, P. Monitoring in the Clouds: Comparison of ECO2Clouds and EXCESS Monitoring Approaches / P. Skvortsov, D. Hoppe, A. Tenschert, M. Geinger // Proceedings of the 2nd International Workshop on Dynamic Resource Allocation and Management in Embedded, High Performance and Cloud Computing DREAMCloud, Prague 19 Jan 2016. – Prague, 2016. – P. 1-8.

4. Rong, K. ASAP: Prioritizing Attention via Time Series Smoothing / K. Rong, P. Bailis // Proceedings of Very Large Data Bases Endowment, Munich 28 Aug-1 Sep 2017. – Munich, 2017. – P. 1358-1369.