САМООБУЧЕНИЕ СИСТЕМЫ ОБНАРУЖЕНИЯ ВРЕДОНОСНЫХ ПРОГРАММ КОМПЛЕКСНОЙ ЗАЩИТЫ ЭЛЕКТРОННОЙ ПОЧТЫ

Т.Ю. Голиков
Научный руководитель – Тонкович И.Н.
канд.хим.наук, доцент
Белорусский государственный университет информатики и радиоэлектроники

Для защиты информации от вредоносного программного обеспечения разработаны специальные программы для их обнаружения называемые антивирусами. При работе антивируса над поиском вирусов и вредоносных программ возникает проблема определения, является ли программа безопасной для работы операционной системы.

Для решения данной задачи предлагается использовать модели (дерево решений и случайный лес), основанные на машинном обучении. Машинное обучение – это набор алгоритмов для анализа данных, их изучения и последующего определения данных. С помощью машинного обучения, программа может анализировать данные, делать прогнозы или выбирать подходящий из предложенных вариантов.

При работе антивируса с изолированной средой программа получает отчёт о работе исследуемой программы, в котором содержится информация о её действиях. Используя обучение с учителем, можно получить ответ на вопрос, является ли данная программа вредоносной.

На основе регрессионной модели дерева решений составляют ряд вопросов, которые позволят определить угрозу от программы. Модель дерева принятия решений – это ряд вопросов, ответы на которые являются ответвлениями к узлам условий перехода к другому вопросу, в конце которых стоят значения целевой функции. Другими словами, проводя опрос, система получает информацию, позволяющую определить класс программы. Например, первым вопросом может быть вопрос об издателе программы. В зависимости от ответа последует следующий вопрос, например, о действиях программы. Если издатель не является производителем операционной системы, и программа пытается изменить содержание системных файлов, то можно сделать вывод, что она является вредоносной.

Также можно использовать модель случайный лес. Данная модель использует множество деревьев, каждое из которых работает со своей выборкой. При формировании выборки рассматриваются следующие метаданные: положение секций файла, размеры секций, информация от производителя или отправителя, вызываемые функции, используемые библиотеки или иная информация из заголовка исполняемого файла, строковые данные, извлекаемые из файла. В процессе регрессии все ответы от деревьев усредняются. Итоговый результат принимается по голосу большинства.

Однако стоит заметить, что обе модели имеют как преимущества, так и недостатки. Модель дерева решений проста в понимании, что даёт возможность переобучения, вероятность успеха зависит от количества узлов, описанных в программе. Модель случайный лес имеет высокую точность, однако она не может дать гарантий выполнения поставленной задачи.