СИСТЕМА КЛАССИФИКАТОРОВ НА ОСНОВЕ НЕЙРОННОЙ СЕТИ ДЛЯ ИДЕНТИФИКАЦИИ АНОМАЛЬНОГО ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ КОРПОРАТИВНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Банокин П.И.
Томский политехнический университет
Pavel805@gmail.com

Введение

Внутренними утечками данных являются случаи хищения конфиденциальной информации, совершенные сотрудниками предприятия. Внутренние утечки данных – одна из главных угроз информационной безопасности предприятия, а их устранение их является причиной значительных финансовых и репутационных издержек [1].

Одним из методов идентификации утечек данных является анализ поведения пользователей программного обеспечения. Метод основан на предположении о том, что поведение недобросовестного сотрудника в момент хищения данных отличается от его обычного каждодневного поведения [2]. Так, сотрудник может использовать несвойственные для его обычного поведения комбинации программных приложений или отдельных инструментов программного приложения или могут измениться законы распределения числовых значений характеристик взаимодействий с вычислительным устройством. В настоящее время существуют реализации алгоритмов идентификации утечек данных из отдельных источников данных (в том числе из социальной сети Facebook [3], из корпоративной сети передачи данных [4]). Сотрудники предприятий ежедневно могут использовать несколько вычислительных устройств, источников данных и широкий перечень программного обеспечения, включая мобильные, веб- и многооконные программные приложения. В настоящее время алгоритмы, в полной мере учитывающие специфику работы пользователей корпоративного программного обеспечения, не реализованы в виде программного обеспечения с открытым исходным кодом.

Характеристики поведения пользователей

Единицей информации о поведении пользователя является регистрируемая запись. Регистрируемая запись является структурой данных для хранения различных характеристик действий пользователя, включая имя пользователя, название вычислительного устройства, название программного приложения, название действия пользователя, название программного инструмента. В ходе наблюдения выявлено, что обычно пользователь использует вычислительное устройство не непрерывно, а в виде коротких взаимодействий – серий действий. Перерыв между сериями действий обычно составляет не менее трех минут. Для выявления устойчивых характеристик поведения пользователя коллекция регистрируемых записей преобразуется в набор серий действий пользователя. Серия действий пользователя – структура данных, в которой перечисляются значимые действия пользователя между короткими периодами неактивности.

Анализ серий действий позволяет выявить следующие закономерности поведения пользователей:

- Наборы частых множеств признаков для категориальных характеристик поведения пользователя, таких как названия используемого программного приложения или инструмента программного приложения (отчет, справочник и т.д.), название вычислительного устройства и др. Примерами шаблонов поведения, представленных частыми множествами признаков, являются совместное следующих программных инструментов: «отчет по оплате, справочник сотрудников, справочник товаров», «справочник сотрудников, журнал учета рабочего времени».
- Распределение значений продолжительности серий действий пользователей и перерывов между ними.
- Доверительный диапазон изменения энтропии.

Для создания наборов частых множеств признаков применяется алгоритм APRIORI [5]. Для выявления факта принадлежности выборок числовых данных используется критерий Колмогорова-Смирнова. Доверительный диапазон колебания энтропии вычисляется на основе стандартного отклонения.

Классификация поведения пользователя

Для оценки качества работы классификаторов производился подсчет количества истинно положительных, ложно положительных, истинно отрицательных, ложно отрицательных результатов работы. Итоговый результат работы классификатора для каждого пользователя изображался в ROC-пространстве [6]. На рисунках 1 и 2 показаны результаты работы классификаторов на основе сравнения наборов частых множеств признаков и на основе сравнения выборок данных критерием Колмогорова-Смирнова соответственно. Результаты экспериментов показали, что не существует универсального набора закономерностей поведения для всех пользователей. Каждый пользователь обладает собственными устойчивыми характеристиками поведения.

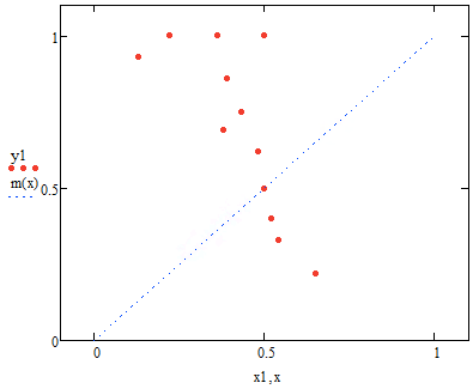

Рисунок. 1. Результаты работы классификатора на основе сравнения наборов частых множеств признаков

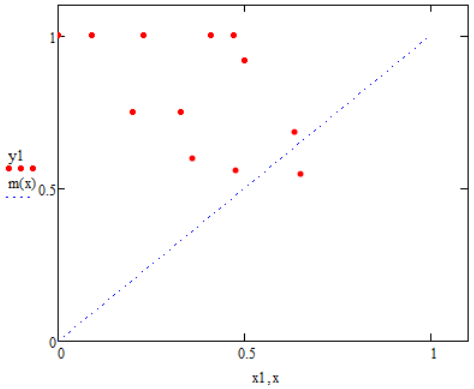

Рисунок. 2. Результаты работы классификатора на основе сравнения выборок данных критерием Колмогорова-Смирнова

Представленное на рисунке 3 объединение классификаторов в систему на основе нейронной сети прямого распространения позволит учитывать индивидуальные закономерности поведения каждого пользователя. Для учета индивидуальных характеристик поведения для каждого пользователя создается набор обучающих пар и выполняется обучение нейронной сети. В дальнейшем состояние обученной нейронной сети сохраняется для каждого пользователя. Входные нейроны K1, K2, K3, K4, K5, K6 принимают результаты работы классификаторов отдельных характеристик поведения пользователя. Значения входных нейронов находятся в диапазоне [0; 1]. Выходное значение 0 свидетельствует о том, что поведение пользователя не соответствует закономерностям каждодневного поведения. Выходное значение 1 свидетельствует об отсутствии аномального поведения пользователя.

При проведении исследования работы системы классификаторов на основе нейронной сети методом K-Fold [7] точность составила 99.62%.

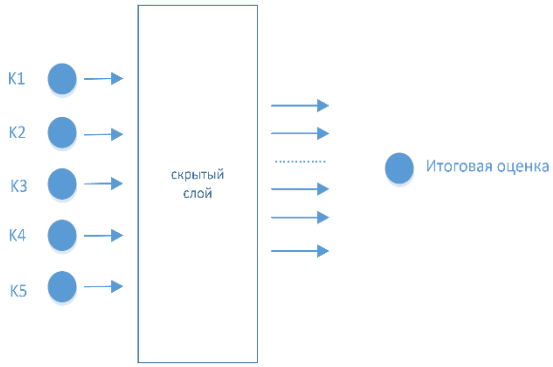

Рис. 3. Система классификаторов на основе нейронной сети

Заключение

В статье предложен способ идентификации утечек данных с системы классификаторов на основе нейронной сети прямого распространения. Экспериментальным путем подтверждено, что пользователи корпоративного программного обеспечения обладают разным набором устойчивых характеристик поведения. Предложенный способ идентификации утечек данных позволяет учитывать индивидуальные наборы устойчивых характеристик поведения пользователей.

Список использованных источников

1. Shabtai A., Elovici Y., Rokach L. A Survey of Data Leakage Detection and Prevention Solutions. - Berlin, Germany: Springer, 2012.
2. Denning D. E. An Intrusion Detection Model // IEEE transactions on software engineering. - 1987. - №SE-13, 2. - С. 222-232
3. Ghiglieri M., Stopczynski M., Waidner M. Personal DLP for Facebook // Pervasive Computing and Communications Workshops (PERCOM Workshops). - New York, USA: IEEE, 2014. - С. 629-634.
4. Linda O., Vollmer T., Manic M. Neural Network based Intrusion Detection System for critical infrastructures // International Joint Conference on Neural Networks. - Atlanta, USA: IEEE, 2009. - С. 1827 - 1834.
5. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules // VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases. - San Francisco, CA, USA: ACM, 1994. - С. 487-499.
6. Паклин Н. (2006). Логистическая регрессия и ROC-анализ – математический аппарат. URL: http://www.basegroup.ru/library/analysis/regression/logistic/.
7. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection // Proceeding IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2. - San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. - С. 1137-1143.