УДК 004.021

Погорельский А.М.
Магистрант
2 курс, факультет «Космический»
МФ МГТУ им. Н.Э. Баумана
Россия, г. Москва
Научный руководитель: Афанасьев А.В.,
кандидат технических наук

РАЗРАБОТКА АЛГОРИТМА БАЛАНСИРОВКИ НАГРУЗКИ ДЛЯ ПРОЦЕССА ОРКЕСТРИРОВАНИЯ СЕРВИСОВ ДЕТЕКТИРОВАНИЯ АНОМАЛИЙ

Аннотация: *Целью данной работы является разработка алгоритма балансировки нагрузки для процесса оркестрирования сервисов детектирования аномалий, необходимо провести исследование алгоритмов балансировки, а также возможных решений.*

Ключевые слова: *Алгоритмы балансировки, оркестрация, алгоритмы, деревья принятия решений, машинное обучение, алгоритм распределения.*

Annotation: *The purpose of this work is to develop a load balancing algorithm for the process of orchestration of anomaly detection services; it is necessary to study balancing algorithms, as well as possible solutions.*

Key words: *Balancing algorithms, orchestration, algorithms, decision trees, machine learning, distribution algorithm.*

Основные понятия

Песочница (Sanbox) - виртуальная среда, внутри которой путём автоматизированной экспертизы проверяется некий ресурс на наличие трояно-вирусных угроз;

Пул песочниц (Sandbox pool) - логически организованный набор из песочниц, собранный по некоторым критериям (зависит от конкретного заказчика и его целей);

Алгоритм базируется на принципе совместном использовании двух алгоритмов:

1. Распределения Round-Robin для пулов песочниц с приоритетами;
2. Дерево принятия решений.

Алгоритм распределения Round-Robin.

Из-за своей простоты реализации, а также довольно неплохой скорости работы он обеспечивает довольно неплохое покрытие за единицу времени и полноту выходных данных.

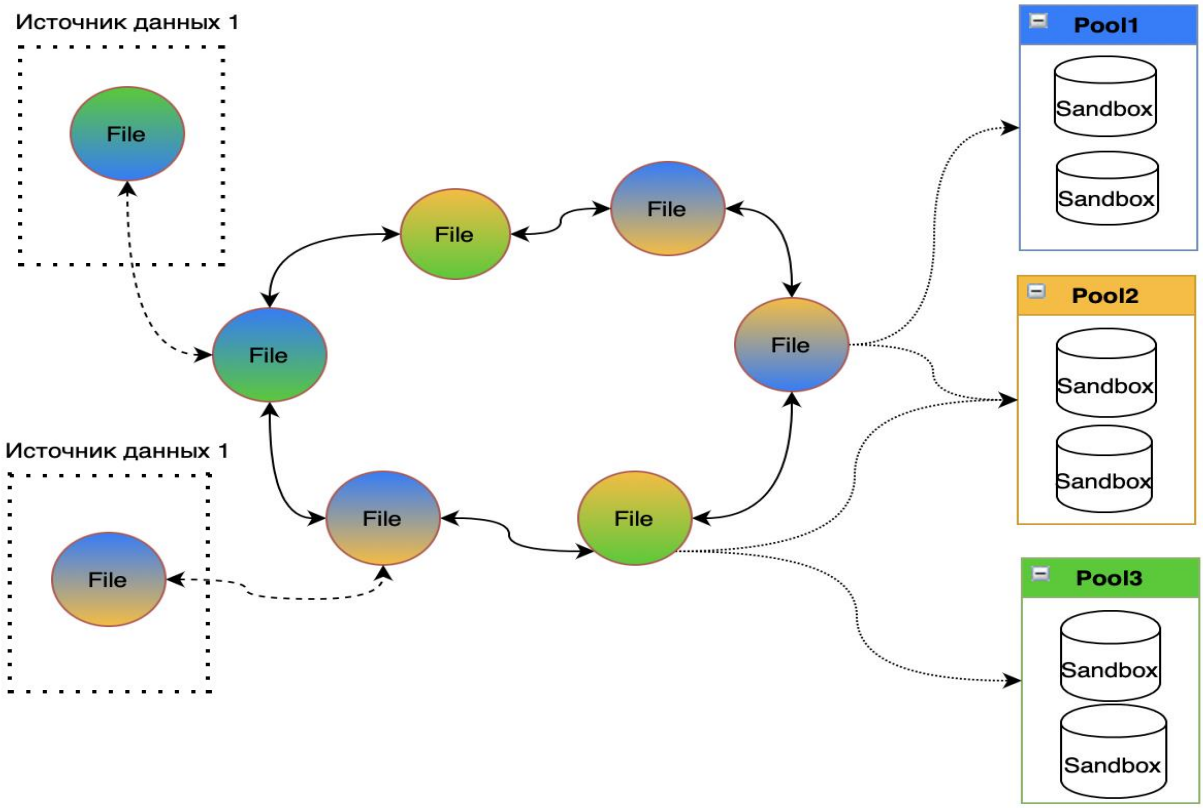

Рисунок 1. Алгоритм Round-Robin

В данном случае применяется алгоритм Round-Robin с приоритетами[1, с.298] (рисунок 1), которые определяются посредством указания в настройке входных источников данных специальных идентификаторов, которые в последующем отражают идентификаторы пулов песочниц. Файл держится

внутри кольцевого буфера до того момента, пока по каждому идентификатору он не будет загружен в нужный пул песочниц с последующим получением результирующего отчёта.

Деревья принятия решений

Деревья принятия решений являются удобным инструментом в тех случаях, когда требуется не просто классифицировать данные, но ещё и объяснить почему тот или иной объект отнесён к какому-либо классу. В случае использования для балансировки это позволяет выстроить логическую цепочку принятия решений для однозначного решения подходит ли песочница для дополнительной нагрузки или не подходит. Для реализации дерева принятия решений использовался алгоритм ID3.

Алгоритм ID3

Алгоритм ID3 (англ. Iterative Dichotomizer, итеративный дихотомайзер), предложенный Д. Куинланом, определяет очередность переменной и ее атрибутов через их информационную значимость (информационную энтропию)[2, с.204]. Для этого следует найти энтропию всех неиспользованных признаков и их атрибутов относительно тестовых экземпляров и выбрать тот, для которого энтропия минимальна (а информативность - максимальна).

Энтропия – среднее количество битов, чтобы закодировать атрибут S у множества элементов. Энтропия для бинарного свойства:

$$H(A,S) = -\frac{m}{n}\log_2\frac{m}{n} - \frac{n-m}{n}\log_2\frac{n-m}{n}$$

Если свойство S не бинарное, а может принимать s различных значений, каждое из которых реализуется в m_i случаях, то энтропия обобщается естественным образом

$$H(A,S) = -\sum_{i=1}^{S}\frac{m_i}{n}\log\frac{m_i}{n}$$

Достоинства алгоритма:

- Интерпретируемость и простота классификации.
- Гибкость: можно варьировать множество B.
- Допустимы разнотипные данные и данные с пропусками.
- Не бывает отказов от классификации.
- Наглядность

Недостатки:

- Жадный ID3 переусложняет структуру дерева, и, как следствие, сильно переобучается;
- Фрагментация выборки: чем дальше v от корня, тем меньше статистическая надёжность выбора;
- Высокая чувствительность к шуму, к составу выборки, к критерию информативности;

В нашем случае этот алгоритм не будет иметь переусложненную структуру, так как множество входных данных было отброшено как не имеющие смысла, так как их энтропия имеет максимально большое значение, для того чтобы быть оптимальной с точки зрения получения максимального прироста информативности[3, с.463].

Использованные источники:

1. Алексеев, В.Е. Графы и алгоритмы. Структуры данных. Модели вычислений / В.Е. Алексеев, В.А. Таланов. - М.: Бином. Лаборатория знаний, Интернет-университет информационных технологий, 2014 – с. 322.
2. Канцедал, С.А. Алгоритмизация и программирование / С.А. Канцедал. - М.: Форум, Инфра-М, 2014 – с. 352.
3. Нейт Сильвер. Сигнал и Шум. Почему одни прогнозы сбываются, а другие — нет, Азбука-Аттикус, КоЛибри, 2015 – с. 608.