Д. В. Бакро

Прогнозирование состояния технического оборудования методами Data Mining.

Аннотация. В статье рассматриваются пакеты программ для анализа сырых данных. Решается задача выбора подходящего инструмента для прогнозирования состояния технического оборудования.

Ключевые слова и фразы: программные пакеты, сбой датчика, программирование.

Введение

Каждое оборудование со временем выходит из строя. Прогнозировать состояние технического оборудования можно по показаниям датчиков. Рассмотрим данные со спутника. В процессе слежения за космическим аппаратом (КА)[1] наземными измерительными комплексами создается файл показаний датчиков положения и вспомогательных расчетных значений, включая время, дальность, скорость передвижения, курсовые и тангажные углы, интегральные значения, ускорения, результаты интерполяции и др. Как показывает практика, работа датчиков сопровождается частыми сбоями, например, в виде кратковременных изменений показаний, которые противоречат некоторым физическим характеристикам или возможностям аппарата. Возникновение подобных сбоев необходимо обнаружить для того, чтобы иметь точную информацию о том, что же в действительности происходит с КА.

1. Постановка задачи

Пусть имеется комплект измерений от n датчиков с временем t между отсчетами. Требуется определить места сбоев и выделить датчики, показания которых неверны. Для обнаружения подобных сбоев необходимо разработать соответствующее алгоритмическое и математическое обеспечение, либо воспользоваться готовыми инструментами. В настоящей работе попробуем сделать второе. Для выбора подходящего программного обеспечения рассмотрим доступные пакеты

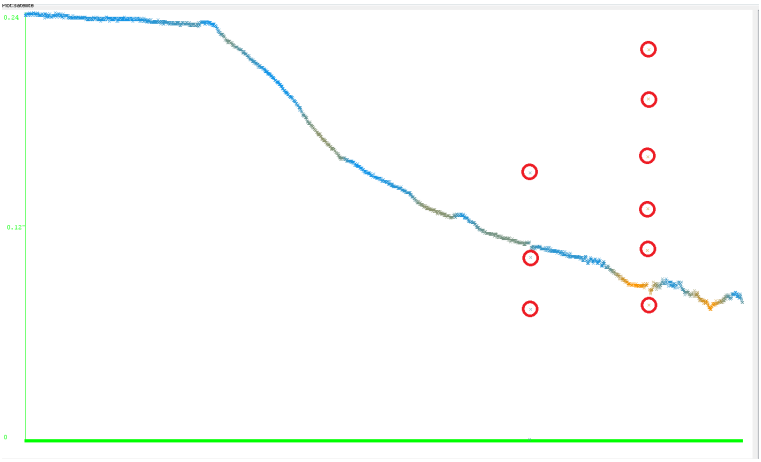

Рис. 1. Пример показаний датчика D

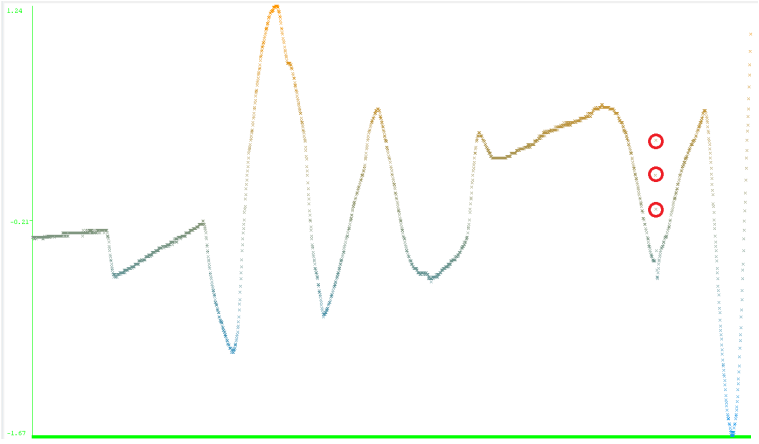

Рис. 2. Пример показаний с датчика UTP

программ для анализа сырых данных, возможно, какие-то инструменты будет уместно применить для достижения поставленной цели. Стоит учитывать, что некоторые датчики взаимосвязаны между собой по отдельным признакам.

На рисунках 1 и 2 изображен график показаний датчиков D и UTP. Красным выделены предполагаемые точки сбоев.

1.1. Интеллектуальная обработка данных (Data Mining[2])

Data Mining (интеллектуальная обработка данных) — название, используемое для обозначения совокупности методов обнаружения в данных знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Задача "Data Mining":

- Поиск закономерностей (скрытых знаний) в данных, необходимых для принятия решений в различных сферах человеческой жизни.

Методы Data Mining:

- Классификация - система группировки объектов исследования или наблюдения по общим признакам;
- Моделирование – исследование объектов исследования или наблюдения на их моделях;
- Прогнозирование - это выдвижение некоторого суждения относительно неизвестных событий.

Постановка задачи в Data Mining:

- имеется база данных;
- предполагается, что в БД есть некоторые **скрытые знания**;
- разработать методы обнаружения этих знаний.

Скрытые знания – это знания, которые должны быть **новыми** (а не подтверждающие какие-то ранее полученные сведения), **нетривиальными** (которые нельзя просто так увидеть при визуальном анализе), **практически полезными** (ценными для исследователя или потребителя) и **доступными для интерпретации** (могут быть представлены в наглядной для пользователя форме и легко объяснить в терминах предметной области).

Необходимо разработать методы обнаружения этих знаний.

Методы «Data Mining» имеет смысл применять только для достаточно больших баз данных. Знания, добываемые методами «Data Mining», принято представлять в виде ассоциативных правил, деревьев решений, кластеров и математических функций. Алгоритмы поиска таких закономерностей входят в область Искусственного интеллекта, Математической статистики, Математического программирования, Визуализации, OLAP.

Могут скопиться достаточно большие объемы телеметрии от КА, такие, что в рамках обычных средств и программных пакетов обрабатываются неэффективно, то есть долго. Например, когда мы

обрабатываем данные за большой временной период или даже собранные сразу с нескольких космических аппаратов. Mining Data выполняет поиск и анализ скрытых знаний, что требуется для прогнозирования состояния технического оборудования. Когда объемы обрабатываемых данных становятся весьма и весьма существенными, то начинают говорить о Больших Данных (Big Data). В первую очередь о Big Data начинаю говорить, когда начинают работать с неструктурированной или слабоструктурированной информацией.

2. Обзор пакетов для Data Mining

2.1. Deductor[3]

Deductor – программное обеспечение, которое содержит в себе инструменты, необходимые для осуществления процесса извлечения скрытых закономерностей из массивов данных. Deductor позволяет решить задачи анализа данных: от сбора информации из различных источников до прогнозирования и оптимизации. Назначение:

- консолидация данных из десятков разнородных источников;
- очистка, систематизация и обогащение собранной информации;
- отчетность, визуализация, OLAP-анализ [1], расчет KPI [2];
- моделирование, прогнозирование, оптимизация;
- самообучение на новых данных и адаптация моделей.

Deductor может применяться в любом бизнесе, где есть большие объемы данных.

2.2. Theano[4]

Theano — это библиотека Python и оптимизирующий компилятор, которая позволяет определить, оптимизировать и вычислять математические выражения, эффективно используя многомерные массивы. Возможности библиотеки:

---

[1]OLAP (англ. online analytical processing, аналитическая обработка в реальном времени) — технология обработки данных, заключающаяся в подготовке суммарной (агрегированной) информации на основе больших массивов данных, структурированных по многомерному принципу.

[2]KPI (Key performance indicator) - показатель в денежном или натуральном выражении, который нужен для оценки эффективности работы предприятия

- тесная интеграция с NumPy [3];
- прозрачное использование GPU;
- эффективное дифференцирование переменных;
- быстрая и стабильная оптимизация;
- динамическая генерация кода на C;
- расширенные возможности юнит-тестирования и самопроверок;

Theano используется в высокоинтенсивных вычислительных научных исследований. По сути, программирование под Theano не является программированием в полном смысле этого слова, так как пишется программа на Python, которая создает выражение для Theano. С другой стороны, это является программированием, так как мы объявляем переменные, создаем выражение которое говорит что делать с этими переменными компилируем эти выражения в функции, которые используются при вычислении. Если кратко, то вот список того, что именно может делать:

- может использовать g++ или nvcc для того, чтобы откомпилировать части части вашего выражения в инструкции GPU или CPU, которые выполняются намного быстрее чистого Python;
- дифференцирование переменных: Theano может автоматически строить выражения для вычисления градиента;
- стабильность оптимизации: Theano может распознать некоторые численно неточно вычисляемые выражения и рассчитать их используя более надежные алгоритмы.

2.3. STATISTICA Data Miner[5]

STATISTICA Data Miner содержит полный набор методов Data Mining на рынке программного обеспечения. Система STATISTICA Data Miner содержит удобные инструменты для всего процесса Data Mining – от построения запросов к БД до создания итоговых отчетов.

Особенности:

- предлагает множество возможностей и методов. Эти функции могут иметь решающее значение для максимизации ROI [4] в конкурентной среде;

[3]NumPy — это расширение языка Python, добавляющее поддержку больших многомерных массивов и матриц, вместе с большой библиотекой высокоуровневых математических функций для операций с этими массивами

[4]ROI - коэффициент доходности бизнеса

- может быть использован как новичками, которым предлагается автоматическое построение моделей с помощью Мастера Data Mining, так и экспертами, которым предоставляется самый широкий выбор методов и технологий для решения даже самых сложных задач;
- является универсальным средством Data Mining, что дает все необходимые инструменты для быстрого понимания критически важных процессов и немедленного воздействия на ROI;
- удобство работы с большим объемом данных

2.4. Weka[6]

Weka (Waikato Environment for Knowledge Analysis) — свободное программное обеспечение. Предоставляет пользователю возможность предобработки данных, решения задач классификации, регрессии, кластеризации и поиска ассоциативных правил, а также визуализации данных и результатов. Программа может быть дополнена новыми алгоритмами, средствами предобработки и визуализации данных.

Weka имеет большой спектр возможностей и использует методы “Data Mining”. Эта программа имеет открытый код, что позволяет создавать ПО для конкретных целей.

Возможности:

- позволяет импортировать данные из базы, CSV файла и т.д., и применять к ним алгоритмы фильтрации, например, переводить количественные признаки в дискретные, удалять объекты и признаки по заданному критерию;
- позволяет применять алгоритмы классификации и регрессии (в Weka они не различаются и называются classifiers) к выборке данных, оценивать предсказательную способность алгоритмов, визуализировать ошибочные предсказания, ROC-кривые, и сам алгоритм, если это возможно (в частности, решающие деревья);
- позволяет выявить все значимые взаимосвязи между признаками;
- позволяет решить определенные задачи кластеризации данных;
- имеет методы отбора признаков;
- позволяет построить матрицу графиков разброса (scatter plot matrix), позволяет выбирать и увеличивать графики и т.д.

Weka предоставляет прямой доступ к библиотеке реализованных в ней алгоритмов. Это позволяет легко использовать уже реализованные алгоритмы из других систем, реализованных на Java. Например, эти

алгоритмы можно вызывать из MATLAB. В частности, интерфейс доступа к алгоритмам Weka из MATLAB реализован в некоторых алгоритмических пакетах машинного обучения таких, как Spider и MATLABArsenal. Для использования Weka из систем, реализованных на других платформах, возможен вызов алгоритмов через интерфейс командной строки.

2.5. Обоснование выбора WEKA

Weka, на мой взгляд, обладает более удобными встроенными графическими интерфейсами и позволяет быстро визуализировать большие объемы данных. Реализованные алгоритмы из других систем можно легко использовать для решения текущей задачи. Благодаря открытому коду, можно создавать сложные проекты.

3. Постановка модельной задачи

- Пусть имеется набор показаний от 6 датчиков с временным расстоянием между отсчетами: D (дальность до аппарата), V (скорость передвижения), UKP (курсовой угол по первой линии связи), UTP (тангажный угол по первой линии связи), UKA (курсовой угол по второй линии связи) и UTA (тангажный угол по второй линии связи).
- Требуется определить места сбоев и выделить группу датчиков, показания которых не верны.

3.1. Методика обнаружения сбоев

Получим все комбинации пар (a,b) специальных векторов при выполнении следующих условий: каждый элемент пары есть набор значений $x_1^1 x_2^1 ... x_m^1 ... x_1^n x_2^n ... x_m^n$, где x_j^i- i-е показание j-го датчика, являющийся результатом комбинирования по n отсчетов с m разных датчиков. Число таких комбинаций равно $\frac{k!}{(k-2m)!}$. Для каждой комбинации (a,b) в Weka находим среднюю корреляцию, используя сканирующее окно. Варьируя размером сканирующего окна, определяем его оптимальный размер, при котором средневзвешанная корреляция для комбинации (a,b) будет максимальной.

Для нахождения группы датчиков с максимальной средневзвешенной корреляцией показаний используем ранговую корреляцию Спирмена[7]. Вычисление включает следующие этапы:

(1) сопоставление каждому из признаков их порядкового номера (ранга) по возрастанию (или убыванию);
(2) определение разности рангов каждой пары сопоставляемых значений (d);
(3) вычисление коэффициентов корреляции рангов по формуле: $r = 1 - \frac{6 \sum d^2}{p(p^2-1)}$, где $\sum d^2$ - сумма квадратов разностей рангов, а p - число парных наблюдений.

В табл. 1 приведены результаты эксперимента по нахождению взаимосвязанных наборов датчиков по файлу телеметрии (2240 отсчетов при t равном 60 миллисекундам). Диапазон поиска оптимального размера окна – от 5 до 45 отсчетов.

Таблица 1. Коэффициенты корреляции наборов датчиков

Набор датчиков	Корреляция
{UKA} и {UTA}, 40 отсчетов	0.732575
{D, V} и {UTP, UTA}, 15 отсчетов	0.776135
{D, UTP, UKA} и {V, UTA, UKP}, 15 отсчетов	0.651810

В результате, получены наборы датчиков, показания которых имеют высокую корреляцию. График корреляции для группы датчиков {D, V} и {UTP, UTA} изображен на Рис. **??**.

Рис. 3. График корреляции для группы датчиков D, V и UTP, UTA

На рисунке видно, что в некоторых моментах коэффициент корреляции резко падает. Это сбой.

Заключение

Рассмотрены программные пакеты для интеллектуального анализа данных. Обоснован выбор Weka как инструмента для анализа сырых данных. Проведены экспериментальные исследования, показавшие эффективность разработанного алгоритма в задаче обнаружения сбоев.

Список литературы

[1] А. А. Талалаев, В. П. Фраленко. Контроль и диагностика датчиков положения космического аппарата. Искусственный интеллект и принятие решений, 2009. — 49–52 с. https://docs.google.com/uc?-export=download&id=0B-Qay3kEFxqfazR1ZkFQLXJDVEU. ↑ 101.

[2] Data Mining. http://www.inftech.webservis.ru/it/database/datamining/ar2.html. ↑ 103.

[3] Deductor. http://www.basegroup.ru/. ↑ 104.

[4] Theano. http://habrahabr.ru/post/173819/. ↑ 104.

[5] STATISTICA Data Miner. http://www.statsoft.ru/products/STATISTICA_Data_Miner. ↑ 105.

[6] А. Г. Дьяконов. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab. http://www.machinelearning.ru/wiki/images/7/7e/Dj2010up.pdf. ↑ 106.

[7] The Proof and Measurement of Association between Two Things: University of Illinois Press, 1904. — 72–101 с. https://explorable.com/spearman-rank-correlation-coefficient. ↑ 107.

Специфика статьи: Развитие информационно-вычислительных технологий, Развитие авиационно-космических технологий, *Алгоритм*, *Подпрограмма или библиотека программ*, Средства компьютерной алгебры, *Языки программирования*, Вычислительный эксперимент.

Научный руководитель:

к.т.н. В. П. Фраленко

Об авторе:

Дмитрий Владимирович Бакро

УГП имени А. К. Айламазяна, 4И11

e-mail: bakrodmitry@mail.ru

Пример ссылки на эту публикацию:

Д. В. Бакро. «Прогнозирование состояния технического оборудования методами Data Mining.». *Наукоёмкие информационные технологии: Труды XIX Молодежной научно-практической конференции* SIT-2015. УГП имени А. К. Айламазяна. — Переславль-Залесский: Изд-во «Университет города Переславля», 2015 с. 101–110.

URL https://edu.botik.ru/proceedings/sit2015.pdf

Dmitry Bakro. *State of technical equipment prediction by data mining methods.*

Abstract. The article deals with the software packages for the analysis of raw data. Solves the problem of choosing an appropriate tool to predict the state of the technical equipment.

Key Words and Phrases: software packages, failure of the sensor, Programming.

Sample citation of this publication:

Dmitry Bakro. "State of technical equipment prediction by data mining methods". *Science-intensive information techlnologies: Proceedings of XIX Junior R&D conference SIT-2015.* Ailamazyan Pereslavl University. — Pereslavl-Zalesskiy: Pereslavl University Publishing, 2015 pp. 101–110. (*In Russian.*)

URL https://edu.botik.ru/proceedings/sit2015.pdf