Center for Machine Learning and Intelligent Systems

IDA2016Challenge Data Set

Download: Data Folder, Data Set Description

Abstract: The dataset consists of data collected from heavy Scania trucks in everyday usage.

Data Set Characteristics:	Multivariate	**Number of Instances:**	76000	**Area:**	Computer
Attribute Characteristics:	Integer	**Number of Attributes:**	171	**Date Donated**	2017-01-17
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	13983

Source:

-- Creator: Scania CV AB
VagnmakarvÃ¤gen 1
151 32 SÃ¶dertÃ¤lje
Stockholm
Sweden
-- Donor: Tony Lindgren (tony **'@'** dsv.su.se) and Jonas Biteus (jonas.biteus **'@'** scania.com)
-- Date: September, 2016

Data Set Information:

This file is part of APS Failure and Operational Data for Scania Trucks.

---

1. Title: APS Failure at Scania Trucks

2. Source Information
-- Creator: Scania CV AB
VagnmakarvÃ¤gen 1
151 32 SÃ¶dertÃ¤lje
Stockholm
Sweden
-- Donor: Tony Lindgren (tony '@' dsv.su.se) and Jonas Biteus (jonas.biteus '@' scania.com)
-- Date: September, 2016

3. Past Usage:
Industrial Challenge 2016 at The 15th International Symposium on Intelligent Data Analysis (IDA)
-- Results:
The top three contestants | Score | Number of Type 1 faults | Number of Type 2 faults
---
Camila F. Costa and Mario A. Nascimento | 9920 | 542 | 9
Christopher Gondek, Daniel Hafner and Oliver R. Sampson | 10900 | 490 | 12
Sumeet Garnaik, Sushovan Das, Rama Syamala Sreepada and Bidyut Kr. Patra | 11480 | 398 | 15

4. Relevant Information:
-- Introduction
The dataset consists of data collected from heavy Scania
trucks in everyday usage. The system in focus is the
Air Pressure system (APS) which generates pressurised
air that are utilized in various functions in a truck,
such as braking and gear changes. The datasets'
positive class consists of component failures
for a specific component of the APS system.
The negative class consists of trucks with failures
for components not related to the APS. The data consists
of a subset of all available data, selected by experts.

-- Challenge metric

Cost-metric of miss-classification:

Predicted class | True class |
| pos | neg |
---
pos | - | Cost_1 |
---
neg | Cost_2 | - |

Cost_1 = 10 and cost_2 = 500

The total cost of a prediction model the sum of 'Cost_1'
multiplied by the number of Instances with type 1 failure
and 'Cost_2' with the number of instances with type 2 failure,
resulting in a 'Total_cost'.

In this case Cost_1 refers to the cost that an unnessecary
check needs to be done by an mechanic at an workshop, while
Cost_2 refer to the cost of missing a faulty truck,
which may cause a breakdown.

Total_cost = Cost_1*No_Instances + Cost_2*No_Instances.

5. Number of Instances:
The training set contains 60000 examples in total in which

59000 belong to the negative class and 1000 positive class.
The test set contains 16000 examples.

6. Number of Attributes: 171

7. Attribute Information:
The attribute names of the data have been anonymized for
proprietary reasons. It consists of both single numerical
counters and histograms consisting of bins with different
conditions. Typically the histograms have open-ended
conditions at each end. For example if we measuring
the ambient temperature 'T' then the histogram could
be defined with 4 bins where:

bin 1 collect values for temperature T < -20
bin 2 collect values for temperature T >= -20 and T < 0
bin 3 collect values for temperature T >= 0 and T < 20
bin 4 collect values for temperature T > 20

| b1 | b2 | b3 | b4 |

-20 0 20

The attributes are as follows: class, then
anonymized operational data. The operational data have
an identifier and a bin id, like 'Identifier_Bin'.
In total there are 171 attributes, of which 7 are
histogram variabels. Missing values are denoted by 'na'.

Attribute Information:

The attribute names of the data have been anonymized for
proprietary reasons. It consists of both single numerical
counters and histograms consisting of bins with different
conditions. Typically the histograms have open-ended
conditions at each end. For example if we measuring
the ambient temperature 'T' then the histogram could
be defined with 4 bins where:

bin 1 collect values for temperature T < -20
bin 2 collect values for temperature T >= -20 and T < 0
bin 3 collect values for temperature T >= 0 and T < 20
bin 4 collect values for temperature T > 20

| b1 | b2 | b3 | b4 |

-20 0 20

The attributes are as follows: class, then
anonymized operational data. The operational data have
an identifier and a bin id, like 'Identifier_Bin'.
In total there are 171 attributes, of which 7 are
histogram variabels. Missing values are denoted by 'na'.

Relevant Papers:

Industrial Challenge 2016 at The 15th International Symposium on Intelligent Data Analysis (IDA)

-- Results:
The top three contestants | Score | Number of Type 1 faults | Number of Type 2 faults
---
Camila F. Costa and Mario A. Nascimento | 9920 | 542 | 9
Christopher Gondek, Daniel Hafner and Oliver R. Sampson | 10900 | 490 | 12
Sumeet Garnaik, Sushovan Das, Rama Syamala Sreepada and Bidyut Kr. Patra | 11480 | 398 | 15

Citation Request:

If you have no special citation requests, please leave this field blank.
