BlogFeedback Data Set

Download: Data Folder, Data Set Description

Abstract: Instances in this dataset contain features extracted from blog posts. The task associated with the data is to predict how many comments the post will receive.

Data Set Characteristics:	Multivariate	**Number of Instances:**	60021	**Area:**	Social
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	281	**Date Donated**	2014-05-29
Associated Tasks:	Regression	**Missing Values?**	N/A	**Number of Web Hits:**	67990

Source:

Krisztian Buza
Budapest University of Technology and Economics
buza '@' cs.bme.hu
http://www.cs.bme.hu/~buza

Data Set Information:

This data originates from blog posts. The raw HTML-documents of the blog posts were crawled and processed.
The prediction task associated with the data is the prediction of the number of comments in the upcoming 24 hours. In order to simulate this situation, we choose a basetime (in the past) and select the blog posts that were published at most 72 hours before the selected base date/time. Then, we calculate all the features of the selected blog posts from the information that was available at the basetime, therefore each instance corresponds to a blog post. The target is the number of comments that the blog post received in the next 24 hours relative to the basetime.

In the train data, the basetimes were in the years 2010 and 2011. In the test data the basetimes were in February and March 2012. This simulates the real-world situtation in which training data from the past is available to predict events in the future.

The train data was generated from different basetimes that may temporally overlap. Therefore, if you simply split the train into disjoint partitions, the underlying time intervals may overlap. Therefore, the you should use the provided, temporally disjoint train and test splits in order to ensure that the evaluation is fair.

Attribute Information:

1...50:
Average, standard deviation, min, max and median of the
Attributes 51...60 for the source of the current blog post
With source we mean the blog on which the post appeared.
For example, myblog.blog.org would be the source of
the post myblog.blog.org/post_2010_09_10
51: Total number of comments before basetime
52: Number of comments in the last 24 hours before the
basetime
53: Let T1 denote the datetime 48 hours before basetime,
Let T2 denote the datetime 24 hours before basetime.
This attribute is the number of comments in the time period
between T1 and T2
54: Number of comments in the first 24 hours after the
publication of the blog post, but before basetime
55: The difference of Attribute 52 and Attribute 53
56...60:
The same features as the attributes 51...55, but
features 56...60 refer to the number of links (trackbacks),
while features 51...55 refer to the number of comments.
61: The length of time between the publication of the blog post
and basetime
62: The length of the blog post
63...262:
The 200 bag of words features for 200 frequent words of the
text of the blog post
263...269: binary indicator features (0 or 1) for the weekday
(Monday...Sunday) of the basetime
270...276: binary indicator features (0 or 1) for the weekday
(Monday...Sunday) of the date of publication of the blog
post
277: Number of parent pages: we consider a blog post P as a
parent of blog post B, if B is a reply (trackback) to
blog post P.
278...280:
Minimum, maximum, average number of comments that the
parents received
281: The target: the number of comments in the next 24 hours
(relative to basetime)

Relevant Papers:

Buza, K. (2014). Feedback Prediction for Blogs. In Data Analysis, Machine Learning and Knowledge Discovery (pp. 145-152). Springer International Publishing.

Citation Request:

Buza, K. (2014). Feedback Prediction for Blogs. In Data Analysis, Machine Learning and Knowledge Discovery (pp. 145-152). Springer International Publishing.