The Trick of the Tail: Segmenting Heavy-Tailed Distributions

Jonathan Dunne, Sonya Leech
IBM Ireland
Dublin, Ireland
Jonathan_Dunne,
LEECHSY@ie.ibm.com

Markus Muller
IBM Germany
Munich, Ireland
markus_mueller@de.ibm.com

Irene Manotas, Mary Swift
IBM Research
New York, United States
Irene.Manotas,
mdswift@ibm.com

***Abstract*—Many systems generate heavy-tailed data sets within the Site Reliability Engineering (SRE) domain. Such datasets are composed of many small and few large observations. Fitting such datasets to known continuous distributions can be challenging due to the pronounced 'head' and long 'tail' of said datasets. This study considers a novel technique to split a dataset into two parts (head and tail) to allow for subsequent data modelling using existing fitting techniques. Using two test system datasets, we address whether a dataset can be modelled by its distribution 'head' and 'tail'. Our framework can aid SRE teams in modelling their datasets without resorting to non-parametric approaches such as Kernel Density Estimation (KDE).**

I. Introduction

II. Background and Related Research

A. Distribution Fitting

Distribution fitting is determining whether empirical data fits a known distribution type. The main interest of this technique is to predict the probability or to forecast the frequency of occurrence of an event at a distinct interval.

There are many techniques to determine the goodness of fit (GoF) of a distribution to empirical data. The two most commonly used techniques are discussed briefly.

The Cramér–von Mises criterion [7] [8] is a non-parametric approach that compares the goodness of fit of a cumulative distribution function (CDF) to that of an empirical density function (EMF). Using a significance test, we can test a hypothesis of whether a data set is from a probability distribution.

The Anderson–Darling test [9] [10] is a statistical test of whether a given sample of data is taken from a probability distribution. This test is an improvement of the Kolmogorov–Smirnov test as it gives more weight to the tails of data. As a result, such a test may be more sympathetic to heavy-tailed data.

B. Histogram Binning

A histogram is an approximate graphical representation of a univariate dataset. Karl Pearson is credited with introducing the term in 1895 [11]. A histogram arranges data in a series of 'bins' along the x-axis, while the frequency or density of each bin is aligned to the y-axis.

There are many ways to represent the histogram regarding the number of bins. Too few bins will provide a coarse view of the data distribution, while many bins provide a fine-grained view. We consider three of the most common ways to calculate the number of bins.

In 1926 Sturges provided one of the first techniques to represent the number of bandwidths using the following formula:

$$k = [log_2 n] + 1 \tag{1}$$

Sturges' formula is based on a binomial distribution which can also be used to approximate the normal distribution. Indeed Sturges' formula works best when a normal distribution is assumed. Otherwise, the histogram may provide an over-smooth histogram shape [12].

In 1981 Freedman and Diaconis [13] provided an additional technique designed to minimize the difference between the area under the empirical probability distribution and the area under the theoretical probability distribution.

David Scott proposed a method most suited to normally distributed data that minimises the integrated mean squared error in the bin [14].

C. Data Clustering Segmentation

Data Clustering arranges objects of a similar type into a defined shape known as a cluster. There are multiple ways to cluster data. The two most common types are centroid-based and hierarchal-based. Centroid-based clustering involves selecting a series of centroids to fix a series of clusters. Algorithms such as k-means arrange objects within a dataset around these predefined centroids based on Euclidian distance estimation [15].

Hierarchal-based clustering (HCA) relates to the idea that similar objects are more related based on their proximity. HCA can work in two ways: the Agglomerative (Bottom-up) approach. Each object starts as an individual cluster, and similar clusters are merged until the top of the hierarchy is reached. The Divisive (Top-down) approach works oppositely. All observations start as a single cluster; splitting occurs until the bottom of the hierarchy is reached [16].

Kneedling can be defined as identifying a point at which a computer system transitions from one state to another [24] in terms of an observable measurement. One such transition could be a series of short and long inter-arrival times observed between application server jobs. Determining such transition points known as 'knees' or 'elbows' is a source of much prior work, which we consider in the next subsection.

D. Mixture Distributions

A mixture distribution is a collection of two or more probability distributions, which can be used to express the density of measured observations more accurately [18]. A mixture distribution can be a collection of either homogeneous (e.g. Two Weibull distributions) or heterogeneous (e.g. A Normal and a Cauchy) distributions.

Mixture distributions are used where a subpopulation of data has a distinct series of characteristics that cannot be modelled effectively with a single distribution type.

E. Studies related to Data Segmentation

Salvador and Chan [19] introduce a technique known as the "L-Method" to estimate the number of clusters as part of a hierarchal clustering process. The authors construct an evaluation graph where the x-axis is the number of clusters, and the y-axis is the value of the similarity function. The 'knee' of this evaluation graph determines the number of clusters to return.

Zhao et al. [20] extend the L-Method using an angle-based approach. Their angle-based approach calculates the local minima of successive differences for a triple of points:

$$y1 + y3 - 2y2 \tag{2}$$

If the sum of the differences is 0, the line is straight; if the sum is > 0 or < 0, a positive or negative knee (elbow) is detected.

Milligan and Cooper evaluate thirty algorithms to determine the number of clusters in synthetic datasets [21]. The authors found that Calanski and Harabasz technique performed best. However, they caution that the results may be data-dependent.

Baxter and Oliver [22] provide a technique to determine the minimum message length for stating the region boundaries of one and two-dimensional examples. Through a simulation process, the authors could identify single and multiple boundaries.

In summary, data segmentation and distribution fitting are useful techniques to determine the underlying structure of a dataset. Additionally, where inferences cannot be made on the dataset, segmentation and clustering techniques have proven useful in dividing the data more meaningfully.

III. Dataset and Method

IV. Results

V. Discussion

Conclusion

References

[1] Mahalanobis, Prasanta Chandra. "On the generalised distance in statistics." Proceedings of the national Institute of Science of India. Vol. 12. 1936.

[2] Markovitch, Natalia M., and Udo R. Krieger. "Nonparametric estimation of long-tailed density functions and its application to the analysis of World Wide Web traffic." Performance Evaluation 42.2-3 (2000): 205-222.

[3] Dunne, Jonathan, and David Malone. "Different every time: A framework to model real-time instant message conversations." 2017 21st Conference of Open Innovations Association (FRUCT). IEEE, 2017.

[4] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." ACM computing surveys (CSUR) 31.3 (1999): 264-323.

[5] Pearson, Karl. "LIII. On lines and planes of closest fit to systems of points in space." The London, Edinburgh, and Dublin philosophical magazine and journal of science 2.11 (1901): 559-572.

[6] Chu, Chia-Shang James. "Time series segmentation: A sliding window approach." Information Sciences 85.1-3 (1995): 147-173.

[7] Cramér, Harald. On the composition of elementary errors: Statistical applications. Almqvist and Wiksell, 1928.

[8] Von Mises, Richard. "Statistik und wahrheit." Julius Springer 20 (1928).

[9] Anderson, Theodore W., and Donald A. Darling. "A test of goodness of fit." Journal of the American statistical association 49.268 (1954): 765-769.

[10] Anderson, Theodore W., and Donald A. Darling. "Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes." The annals of mathematical statistics (1952): 193-212.

[11] Pearson, Karl. "X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material." Philosophical Transactions of the Royal Society of London.(A.) 186 (1895): 343-414.

[12] Sturges, Herbert A. "The choice of a class interval." Journal of the American statistical association 21.153 (1926): 65-66.

[13] Freedman, David, and Persi Diaconis. "On the histogram as a density estimator: L 2 theory." Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57.4 (1981): 453-476.

[14] Scott, David W. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, 1992.

[15] Lloyd, Stuart. "Least squares quantization in PCM." IEEE transactions on information theory 28.2 (1982): 129-137.

[16] Sibson, Robin. "SLINK: an optimally efficient algorithm for the single-link cluster method." The computer journal 16.1 (1973): 30-34.

[17] Satopaa, Ville, et al. "Finding a" kneedle" in a haystack: Detecting knee points in system behavior." 2011 31st international conference on distributed computing systems workshops. IEEE, 2011.

[18] Robertson, C. A., and J. G. Fryer. "Some descriptive properties of normal mixtures." Scandinavian Actuarial Journal 1969.3-4 (1969): 137-146.

[19] Salvador, Stan, and Philip Chan. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms." 16th IEEE international conference on tools with artificial intelligence. IEEE, 2004.

[20] Zhao, Qinpei, Ville Hautamaki, and Pasi Fränti. "Knee point detection in BIC for detecting the number of clusters." Advanced Concepts for Intelligent Vision Systems: 10th International Conference, ACIVS 2008, Juan-les-Pins, France, October 20-24, 2008. Proceedings 10. Springer Berlin Heidelberg, 2008.

[21] Milligan, Glenn W., and Martha C. Cooper. "An examination of procedures for determining the number of clusters in a data set." Psychometrika 50 (1985): 159-179.

[22] Baxter, Rohan A., and Jonathan J. Oliver. "The kindest cut: minimum message length segmentation." ALT. 1996.

[23] Zhao, Qinpei, Ville Hautamaki, and Pasi Fränti. "Knee point detection in BIC for detecting the number of clusters." Advanced Concepts for Intelligent Vision Systems: 10th International Conference, ACIVS 2008, Juan-les-Pins, France, October 20-24, 2008. Proceedings 10. Springer Berlin Heidelberg, 2008.

[24] Satopaa, Ville, et al. "Finding a" kneedle" in a haystack: Detecting knee points in system behavior." 2011 31st international conference on distributed computing systems workshops. IEEE, 2011.

[25] Salvador, Stan, and Philip Chan. "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms." 16th IEEE international conference on tools with artificial intelligence. IEEE, 2004.

[26] Delignette-Muller M. L. and Dutang C., "fitdistrplus: An R package for fitting distributions," Journal of Statistical Software, vol. 64, no. 4, pp. 1–34, 2015. [Online]. Available: http://www.jstatsoft.org/v64/i04/

[27] Stephens, Michael A. The Anderson-Darling statistic. STANFORD UNIV CA DEPT OF STATISTICS, 1979.