To Aggregate or not to aggregate : that is the question

Paquet, Eric; Viktor, Herna L.; Guo, Hongyu

Publisher's version / la version de l'éditeur:
Proceedings

TO AGGREGATE OR NOT TO AGGREGATE: THAT IS THE QUESTION

Eric Paquet[1,2], Herna L Viktor[2] and Hongyu Guo[1]
[1]*Institute of IT, National Research Council of Canada, Ottawa, Ontario, Canada*
[2]*Department of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Ontario, Canada*
{*eric.paquet, hongyu.guo*}*@nrc-cnrc.gc.ca, hlviktor@site.uottawa.ca*

Keywords: Data pre-processing, aggregation, Gaussian distribution, Lévy distribution.

Abstract: Consider a scenario where one aims to learn models from data being characterized by very large fluctuations that are neither attributable to noise nor outliers. This may be the case, for instance, when examining supermarket ketchup sales, predicting earthquakes and when conducting financial data analysis. In such a situation, the standard central limit theorem does not apply, since the associated Gaussian distribution exponentially suppresses large fluctuations. In this paper, we argue that, in many cases, the incorrect assumption leads to misleading and incorrect data mining results. We illustrate this argument against synthetic data, and show some results against stock market data.

1 INTRODUCTION

Aggregation and summarization is an important step when pre-processing data, prior to building a data mining model. This step is increasingly needed when aiming to make sense of massive data repositories. For instance, online analytic processing (OLAP) data cubes typically represent vast amounts of data grouped by aggregation functions, such as *sum* and *average*. The same observation holds for social network data, where the frequency of a particular relationship is often represented by an aggregation based on the number of occurrences. Furthermore, data obtained from data streams are frequently summarized into manageable size buckets or windows, prior to mining (Han et al., 2006).

Often, during such a data mining exercise, it is implicitly assumed that large scale fluctuations in the data must be either associated with noise or with outliers. The most striking consequence of such an assumption is that, once the noisy data and the outliers have been eliminated, the remaining data may be characterized in two ways. That is, firstly, their typical behaviour (i.e. their mean) and secondly, by the characteristic scale of their variations (i.e. their variance). Fluctuation above the characteristic scale is thus being assumed to be highly unlikely. Nevertheless, there are many categories of data which are characterized by large scale fluctuations. For instance, supermarket ketchup sales, financial data and earthquake related data are all examples of data exhibiting such behaviour (Walter, 1999; Groot, 2005). The large scale fluctuations do not origin from noise or outliers, but constitute an intrinsic and distinctive feature.

Mathematically speaking, small fluctuations are modelled with the central limit theorem and the Gaussian distribution, while large fluctuations are modelled with the generalized central limit theorem and the Lévy distribution. This position paper discusses the aggregation of data presenting very large scale fluctuations, and argues that the assumption of the underlying Gaussian distribution leads to misleading results. Rather, we propose the use of the Lévy (or stable) distribution to handle such data.

2 AGGREGATION AND THE CENTRAL LIMIT THEOREM

Aggregation is based on the standard central limit theorem which may be stated as follows: The sum of N normalized independent and identically distributed random variables of zero mean and finite variance σ^2 is a random variable with a probability distribution function converging to the Gaussian distribution with variance σ^2 where the normalization is defined as in the following equation:

$$z = \frac{X - N\langle x \rangle}{\sqrt{N}\sigma}$$

This means that when we refer to aggregated data, in the sense of a sum of real numbers, we implicitly assume that such aggregated data has a Gaussian distribution. This distribution is irrespectively of the original distribution of its individual data. In practice, this implies that an aggregation, such as a sum, may be fully characterized by its mean and its variance; this is why aggregation is so powerful. All the other moments of the Gaussian distribution are equal to zero. Despite the fact these assumptions on which aggregation is based are quite general they do not cover all possible data distributions, for instance, the Lévy distribution.

Stable or Lévy distributions are distributions for which the individual data as well as their sum are identically distributed (Samoradnitsky and Taqqu, 1994; Véhel and Walter, 2002). That implies that the convolution of the individual data is equal to the distribution of the sum or, equivalently, that the characteristic function of the sum is equal to the product of their individual characteristic functions. Extreme values are much more likely for the Lévy distribution that they are for the Gaussian distribution. The reason being that the Gaussian distribution fluctuates around its means, the scale of the fluctuations being characterized by its variance (the fluctuations are exponentially suppressed) while the Lévy distribution may produce fluctuations far beyond the scale parameter because of the tail power decay law.

The Lévy distribution is characterized by four parameters as opposed to the Gaussian distribution which is characterized by only two. The parameters are: the stability exponent α, the scale parameter γ, the asymmetry parameter β and the localisation parameter μ. While the tail of the Gaussian distribution is exponentially suppressed, the tail of the Lévy distribution decays as a power law (heavy tail) which depends on its stability exponent, as the following equation shows:

$$L_\alpha(x) \sim \left. \frac{C_\pm}{|x|^{1+\alpha}} \right|_{x \longrightarrow \pm\infty}$$

It should be noticed that the Lévy distribution reduces to the Gaussian distribution when $\alpha = 2$ and when the asymmetry parameter is equal to zero. A Lévy distribution with $1 \leq \alpha < 2$ has a finite mean, but an infinite variance while a distribution with $\alpha < 1$ has both an infinite mean and an infinite variance. As we will see in the following sections, these properties have grave consequences from the aggregation point of view.

3 SIMULATION RESULTS

In this section, we present simulations which illustrate our previous observations. All simulations were performed using Mathematica 8.0 on a Dell Precision M6400. In the following, $\alpha = 2$ corresponds to a Gaussian distribution.

3.0.1 Simulations

Table 1 shows the mean and the standard deviation estimated from empirical data drawn from a stable distribution for various values of the stability exponent α and size N. One may notice that when $\alpha < 1$, the mean and the standard deviation are many orders of magnitude higher than those associated with the Gaussian distribution. This implies that the extreme values, associated with the tail of the distribution, dominate the mean and the standard deviation.

Table 1: Mean and standard deviation for the Lévy distribution for various values of the stability exponent and of the size of the aggregate.

α	N	Mean	Standard Deviation
2	100	-0.05	1.25
	1000	0.02	1.46
	10000	0.01	1.41
1.7	100	-0.07	1.26
	1000	0.23	3.73
	10000	-0.03	5.12
1.5	100	-0.01	2.01
	1000	-0.22	5.34
	10000	0.14	10.70
1.0	100	-0.17	12.93
	1000	0.12	13.97
	10000	11.37	1086.21
0.5	100	-1796.93	20136.90
	1000	340.02	7736.64
	10000	75756.40	5.59×10^{6}
0.1	100	4.31×10^{18}	1.43×10^{19}
	1000	6.03×10^{27}	1.91×10^{29}
	10000	1.10×10^{42}	1.10×10^{44}

Furthermore, the standard deviation does not converge when $\alpha < 2$ and the mean and the standard deviation do not converge when $\alpha < 1$; their estimate becomes a meaningless random number. Consequently, if the empirical data have a Lévy distribution, the aggregation with the standard deviation is meaningless if $\alpha < 2$ and the aggregation with the mean is meaningless if $\alpha < 1$. For instance, (Groot, 2005) has reported that supermarket sales of ketchup (tomato sauce) are characterized by a Lévy distribution with $\alpha = 1.4$.

We may understand this behaviour by considering the histograms of the empirical distribution. Fig. 1 shows the histogram with $\alpha = 2$ and Fig. 2 with $\alpha = 1.7$. One immediately notices that the maximum of the aggregation dominates over the other values and that the scale of the fluctuations for small values of α is many orders of magnitude higher than the one associated with a Gaussian distribution. Although the maximum of the distribution has a low probability, it totally dominates the mean and the variance if $\alpha < 1$.

More insight may be obtained by considering the cumulative sum of Lévy distributed data. Fig. 3 shows the cumulative sum for $\alpha = 2$ and Fig. 4 for $\alpha = 0.5$. Once more, one notices the importance of the maximum which eventually tends to completely dominates the cumulative sum when $\alpha = 0.5$. Consequently, Lévy distributions are suitable to characterize data for which the behaviour is mostly determined, depending on the value of α, by a limited number of extreme events.

For instance, the value of a share is usually dominated by a few large fluctuations and so are the damages associated with earthquakes and tsunamis. When $\alpha < 1$, the aggregation should be performed with the maximum function. In this particular case, the mean and the standard deviation are infinite which means that their estimations from a collection of empirical data are just meaningless random numbers. When $1 \leq \alpha < 2$, the aggregation may be performed with the mean but the standard deviation becomes infinite.

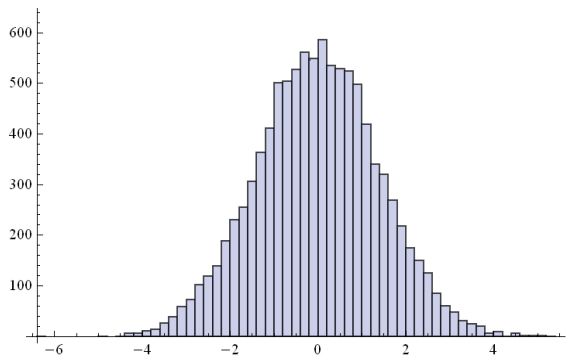

Figure 1: Histogram of a Gaussian distribution with $\alpha = 2$ and N=10000.

One should keep in mind that the closer is α to one, the slower is the convergence of the mean estimated on a collection of empirical data. In practice, that means that the mean should be estimated from a large number of data in order to obtain a meaningful result.

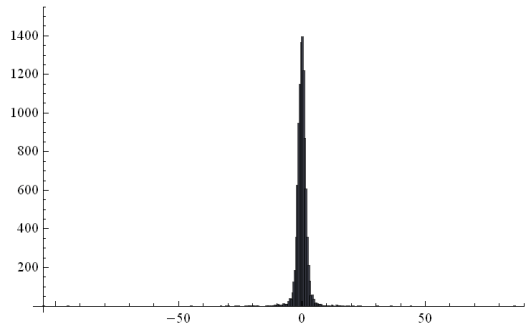

Figure 2: Histogram (notice the scale) of a Lévy distribution with α=1.7 and N=10000.

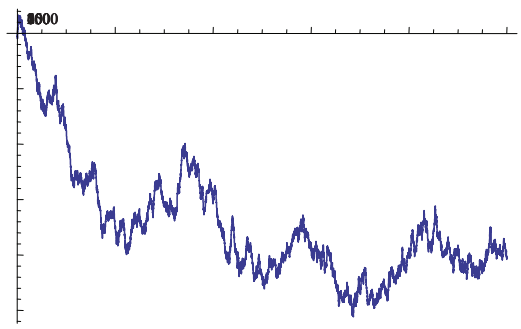

Figure 3: Cumulative sum for Gaussian distributed data with α=2 and N=10000.

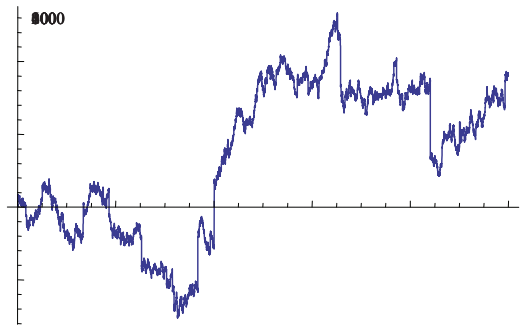

Figure 4: Cumulative sum for Lévy distributed data with α=1.7 and N=10000.

3.0.2 The Lévy Distribution and the Real World

The importance of Lévy distribution is not only theoretical. As a matter of fact, it has far reaching consequences for, amongst others, financial market data. With the pioneer work of Mandelbrot, it became increasingly apparent that financial data may be characterized with stable distributions. For instance, let us consider Table 2 which shows the results obtained for various Stock Market Indexes in Europe(Véhel and Walter, 2002). Here, γ is the scale factor and the threshold is 1% (confidence level of 99%). As shown

by the data, all these indexes clearly have a Lévy distribution and the value of the stability exponent is typically around 1.7, which is not in the Gaussian regime.

Table 2: Estimation of the parameters of the Lévy distributions associated with various Stock Exchange Indexes in Europe (Véhel and Walter, 2002).

Index	Currency	Period	N	α	γ	Threshold (1%)
FTA W Europe	GBP	86.01 -93.09	93	1.716	2.690	1.1408
MSCI Europe	USD	80.01 -93.09	165	1.713	2.936	0.1057
MSCI EUR ex UK	USD	80.01 -93.09	165	1.719	2.951	0.1057

Stock market data is not the only type of data that are suspect to such large data fluctuation that does not have a Gaussian distribution. As previously mentioned, the sales of ketchup are another example of such data. Also, the damages caused by natural disasters such as hurricanes, tornados and earthquakes, fall within this domain. Using the standard data preprocessing techniques, and incorrectly assuming that the standard limit theorem holds in such cases, has grave impact on the validity of the resultant models constructed. This is especially true in domains where the data are aggregated prior to model building. As mentioned earlier, the vast size of massive data mining repositories necessitates aggregation, due to the sheer size and complexity of the data being mined.

4 CONCLUSIONS

This position paper challenges the implicit assumption, which is often made during numerous data mining exercises, that the standard limit theorem holds and that the data distribution is Gaussian. We discuss the implications of this assumption, especially in terms of aggregated data that is characterised with large fluctuations. We show the nature of the differences between the Gaussian and Lévy distributions, on synthetic data and show an example from the real-world financial stock market data. We observe that the two sets of distributions are vastly different, and that it follows that, during any data mining exercise, that data with a Levy distribution should be treated with caution, especially during data pre-processing and aggregation.

The implications and applications of this observation are far-reaching in many domains. It has been shown that the value of a share is usually dominated by a few large fluctuations. Damages associated with earthquakes and tsunamis, such as those caused by the recent events in Japan, are also characterized by such large fluctuations. The same observation holds, e.g., when observing the sizes of solar flares or craters on the moon, as well as for the data obtained from many climate change studies. This fact needs to be taken into account, when aiming to create valid data mining models for these types of domains, which are becoming increasingly important for socio-economic reasons.

REFERENCES

Groot, R. D. (2005). Lévy distribution and long correlation times in supermarket sales. *Lvy distribution and long correlation times in supermarket sales*, 353:501–514.

Han, J., Kamber, M., and Pei, J. (2006). *Data Mining: Concepts and Techniques (2nd edition)*. Morgan Kauffman.

Samoradnitsky, G. and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, New York.

Véhel, J. L. and Walter, C. (2002). *Les marchés fractals (The fractal markets)*. Universitaires de France, Paris.

Walter, C. (1999). Lévy-stability-under-addition and fractal structure of markets: implications for the investment management industry and emphasized examination of matif notional contract. *Mathematical and Computer Modelling*, 29(10-12):37–56.