Learning with aggregation and correlation in the presence of large fluctuations

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

Publisher's version / Version de l'éditeur:

Proceedings of New Frontiers in Mining Complex Patterns (NFMCP 2012), pp. 12-23, 2012-09

Learning with Aggregation and Correlation in the Presence of Large Fluctuations

Eric Paquet[1,2], Herna L. Viktor[2], and Hongyu Guo[1]

[1] National Research Council, 1200 Montreal Road, Ottawa, Ontario, K1A 0R6, Canada
[2] School of Information Technology and Engineering, University of Ottawa, 800 King Edward, Ottawa, Ontario, K1N 6N5, Canada
eric.paquet@nrc-cnrc.gc.ca, hlviktor@site.uottawa.ca, hongyu.guo@nrc-cnrc.gc.ca

Abstract. Consider a scenario where one aims to learn models from dynamic and evolving data being characterized by very large fluctuations that are neither attributable to noise nor outliers. This may be the case, for instance, when predicting the potential future damages of earthquakes or oil spills, or when conducting financial data analysis. If follows that, 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 present an analysis of data aggregation and correlation in such scenarios. To this end, we introduce the Lévy, or stable, distribution which is a generalization of the Gaussian distribution. Our theoretical conclusions are illustrated with various simulations, as well as against a benchmarking financial database. We show which specific strategies should be adopted for aggregation, depending on the stability exponent of the Lévy distribution. Our results firstly show scenarios where it may be impossible to determine the mean and the standard deviation of an aggregate. Secondly, we discuss the case where an aggregate may have to be characterized with its largest fluctuations. Thirdly, we illustrate that the correlation in between two attributes may be underestimated if a Gaussian distribution is erroneously assumed.

Keywords: Aggregation in Relational Learning, Correlation-based Analysis and Covariance, Lévy Distribution, Stable Distribution.

1 Introduction

Aggregation is an important step when pre-processing data, prior to building a data mining model. This step is crucial when considering complex data that represents interactions between several potentially heterogeneous entities. For instance, in social network analysis the frequency of a particular relationship is often represented by an aggregation based on the number of occurrences. The same observation holds in multi-relational database mining and in spatial data exploration, where aggregation is needed to link multiple tables together [1, 2]. Similarly, data obtained from data streams are frequently summarized into manageable sized buckets or windows, prior to mining.

Often, during a data mining exercise, it is implicitly assumed that large-scale data fluctuations must be either associated with noise or with outliers, or that a concept drift has occurred. 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, or assumed to indicate that a model has become outdated. However, there are many categories of data which are characterized by large-scale fluctuations. For instance, surprisingly, supermarket ketchup sales have been shown to be typified by such large-scale fluctuations [4]. Further, financial data and earthquake-related data are also examples of data exhibiting this behaviour [3]. This issue is highly relevant when aiming to build models that predict the potential damages caused by catastrophic events, such as financial market turbulences and tsunamis. The large-scale fluctuations do not origin from noise or outliers, but constitute an intrinsic and distinctive feature of the data. 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 paper studies the aggregation of data presenting large-scale fluctuations, to determine their properties, the best approaches for their aggregation and the impact of such behaviour on their correlation.

Our main contributions are as follows. Firstly, we provide a theoretical analysis which shows the importance of taking the data distribution characteristics, when learning involves aggregation and correlation, into account. Secondly, we introduce the stable distribution as a mechanism to allow machine learning methods to compute meaningful aggregated information and to correctly evaluate the correlation when learning from data presenting large fluctuations. Thirdly, we demonstrate the proposed method's applicability in typical machine learning and data mining problems. Here, we discuss the analysis of financial data and relational learning involving aggregation. We also mention, in Section 4, that our approach is highly relevant for correlation-based privacy preservation data mining.

This paper is organized as follows. In Section 2, we review the fundamental assumptions behind aggregation, namely the central limit theorem and the Gaussian distribution. In Section 3, we introduce a more general distribution for the aggregate, the Lévy or stable distribution, for which the Gaussian distribution is a particular case. We explain how this distribution may be estimated from the empirical data and present some useful properties of the Lévy distribution. Then, we study the rank ordering statistics of the Lévy distribution in order to determine if there are some dominant terms in the distribution. We introduce the multivariate stable distribution in order to generalize, in Section 4, the concepts of covariance and correlation to stable distributions. In Section 5, we present various simulations in order to illustrate the theoretical results obtained in the previous sections, as well as their consequences for aggregation. We show that our methodology is applicable to real world financial data. The last section present our conclusions and directions for future work.

2 Aggregation, Central Limit Theorem and Gaussian Distribution

In this section, we review the basic assumptions on which aggregation is based. Despite the fact that these assumptions are quite general, they do not cover all possible data distributions, for instance, the Lévy distribution. Importantly, as will be shown in Section 5, data associated with catastrophic events such as oil spill, stock market crashes and earthquakes often follows the Lévy distribution and need special care during data pre-processing and model building. Consequently, we aim to understand their strengths as well as their limitations in order to be able to address them in the following sections. Aggregation is based on the standard central limit theorem which may be stated as follow: 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. That means that aggregation, in the sense of a sum of real numbers, has a Gaussian distribution irrespectively of the original distribution of its individual data. This is a very powerful theorem because the Gaussian distribution may be characterized with solely two numbers: its mean μ and its variance σ^2 which are the first two moments of the distribution. 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. In the next section, we consider a generalization of the central limit theorem which requires the introduction of the Lévy or stable distribution.

3 Aggregation with an Underlying Lévy Distribution or Stable Distribution

In this section, we review the Lévy distribution, we show how it may be estimated from the underlying empirical data and we analyze its properties. Finally, we introduce the multivariate Lévy distribution in order to extend the notions of covariance and correlation to data distributed according to a stable distribution. One may associate to a probability distribution $L(x)$ its Fourier transform or characteristic function $L(k)$. Stable or Lévy distributions are distributions for which the individual data as well as their sum are identically distributed [5]. This fact 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. The Lévy distribution does not have a closed form and is more easily defined from its characteristic function:

$$L_{\alpha,\mu,\beta,\gamma}(k) = \mathbf{E}\left(\exp\left(ik\ X\right)\right) = \exp\left[-i\mu\ k - \gamma^{\alpha}\left|k\right|^{\alpha}\left(1 - i\beta\,\mathrm{sgn}\left(k\right)W\left(\alpha,k\right)\right)\right] \tag{1}$$

where $\mathbf{E}(\bullet)$ is the expectation and where

$$W\left(\alpha,k\right) \triangleq \begin{cases} \tan\dfrac{\pi\alpha}{2} & \alpha \neq 1 \\ -\dfrac{2}{\pi}\ln\left|k\right| & \alpha = 1 \end{cases} \tag{2}$$

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 α:

$$L_\alpha\left(x\right) \sim \left.\frac{C_\pm}{\left|x\right|^{1+\alpha}}\right|_{x\to\pm\infty} \tag{3}$$

Eq. (3) shows that 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 mean, 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. This behaviour is due to the tail power decay law. It should be noted that the Lévy distribution reduces to the Gaussian distribution when $\alpha = 2$ and when the asymmetry parameter is equal to zero; then one has $\sigma^2 = \frac{1}{2}\gamma$. Finally, the moments of the Lévy distribution, $m_n = \int_{-\infty}^{\infty} x^n \; P\left(x\right) dx$, may be finite if $n \leq \alpha$ while they are infinite if $n > \alpha$. That means that 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 tremendous consequences from the aggregation point of view.

We explain how the parameters of the Lévy distribution may be estimated from the empirical data and how the validity of the Lévy distribution hypothesis may be asserted. Although various approaches have been proposed in the literature, one of the most efficient is the one presented by Paulson, Holcomb and Leitch (PHL) [6] in which the following objective function is minimized against the parameters of the Lévy distribution

$$\min_{\alpha,\beta,\gamma,\mu} \left\|\hat{L}\left(k\right) - L_{\alpha,\mu,\beta,\gamma}\left(k\right)\right\|_{\text{PHL}} \tag{4}$$

where the PHL-norm is defined as

$$\left\|\hat{L}\left(k\right) - L_{\alpha,\mu,\beta,\gamma}\left(k\right)\right\|_{\text{PHL}} \triangleq \int_{-\infty}^{\infty} \left|\hat{L}\left(k\right) - L_{\alpha,\mu,\beta,\gamma}\left(k\right)\right|^2 \exp\left(-k^2\right) dk \tag{5}$$

Because the integration domain is not bounded, Eq. (5) is more readily solved with a Gauss-Hermite quadrature.

We address the relative importance of the maximum with respect to the sum or aggregation, in order to demonstrate that the sum is potentially dominated by a small

number of very large terms. The expectation of this ratio, for any distribution, is given by

$$\mathbf{E}\left[\frac{X_1 + \ldots + X_N}{\max\left\{X_1,\ldots,X_N\right\}}\right] = N\int_0^1 dz\; z\; G\left(z\right) \tag{6}$$

where

$$\begin{aligned} G\left(z\right) &= \frac{1}{N}\delta\left(z-1\right) + \left(N-1\right)\int_0^\infty dy\; P\left(z\; y\right) P\left(y\right)\left[\mathcal{P}_{\leq}\left(y\right)\right]^{N-2} \\ \mathcal{P}_{\leq}\left(x\right) &= \int_{-\infty}^{x} P\left(y\right) dy \end{aligned} \tag{7}$$

If the Lévy distribution is substituted in this expression, one obtains, for a stability parameter smaller than one, that the sum is largely dominated by a single term, the maximum. In practice, that means that the aggregation must be performed in terms of the maximum and *not* in terms of the sum or the variance. For example, when aiming to build a data mining model to predict the damages associated with a tsunami or oil spill, which has been shown to follow a Lévy distribution, it is more appropriate to use the maximum values rather than the mean and the standard deviation.

More insight about the Lévy distribution may be obtained from its rank ordering statistics

$$F\left(y_n\right) dy_n = \left(N-n+1\right)\binom{N}{n}\left(1-\int_{y_n}^{\infty} P\left(x\right) dx\right)^{N-n} P\left(y_n\right) dy_n \left(\int_{y_n}^{\infty} P\left(x\right) dx\right)^{n-1} \tag{8}$$

which gives the probability that the largest value of order n be y_n; for instance, y_1 is the maximum. From Eq. (8), it may be demonstrated that the maximum of likelihood for the statistics of order n associated with the stable distribution is:

$$y_n^{\mathrm{ML}} = \left[\frac{\left(\alpha N + 1\right)\gamma}{\alpha\; n + 1}\right]^{1/\alpha} \tag{9}$$

When the stability exponent is inferior to one, the rank ordering statistics exhibit a strong hierarchical behaviour up to the point that the ordering statistic of order one (the maximum) completely dominates over all the other rank ordering statistics. This behaviour shall become more evident with the experimental results, presented in Section 5. A practical consequence associated with this behaviour is that the aggregation should be based on the maximum value which completely dominates the ordering statistics; this is the case, for instance, when a Stock Market Index crashes. The information obtained from the rank ordering statistics may be exploited in order to group the elements of the hierarchy according to their scale, or order of magnitude. Then, it may be shown that each scale is characterized by its own Gaussian distribution. Consequently, the Gaussian paradigm is applicable to the Lévy distribution in a multiscale framework and the Lévy distribution might be thought of as a multiscale generalization of the Gaussian distribution.

We extend the Lévy distribution to the multivariate case, i.e. when we have more than one dimension or feature. Such a multivariate distribution is required in order to

study the correlation in between two stable stochastic variables. The multivariate Lévy characteristic function [5] of dimension d is defined as follow

$$L_{\alpha,\boldsymbol{\mu}}\left(\mathbf{k}\right) = \exp\left(-I_{\mathbf{X}}\left(\mathbf{k}\right)\right) = \mathbf{E}\left[\exp\left(i\left\langle \mathbf{k},\mathbf{X}\right\rangle\right)\right] = \exp\left[-i\left\langle \boldsymbol{\mu},\mathbf{k}\right\rangle - \int_{S^d} \psi_\alpha\left(\left\langle \mathbf{k},\mathbf{s}\right\rangle\right)\Delta\left(ds\right)\right] \tag{10}$$

where as usual $\mathbf{E}\left(\bullet\right)$ is the expectancy, where

$$\psi_\alpha\left(u\right) \triangleq \begin{cases} \left|u\right|^\alpha\left(1 - i\,\mathrm{sgn}\left(u\right)\tan\dfrac{\pi\alpha}{2}\right) & \alpha \neq 1 \\ \left|u\right|\left(1 + i\dfrac{\pi}{2}\mathrm{sgn}\left(u\right)\ln\left|u\right|\right) & \alpha = 1 \end{cases} \tag{11}$$

and where the Euclidian inner product, the frequency vector and the stochastic data vector are defined as

$$\left\langle \mathbf{k},\mathbf{X}\right\rangle \triangleq \sum_{i=1}^{d} k_i X_i, \quad \mathbf{k} = \left[k_1,\ldots,k_d\right]^{\mathrm{T}}, \quad \mathbf{X} = \left[X_1,\ldots,X_d\right]^{\mathrm{T}} \tag{12}$$

As opposed to the univariate case, only two parameters are required: a stability exponent α and a localisation vector $\boldsymbol{\mu}$. The information about the scale and the asymmetry is encapsulated in $\Delta\left(ds\right)$ which is a measure, or a partition, defined on the hypersphere S^d (or the sphere in two dimensions). In order to estimate such a distribution from the empirical data, we follow an approach introduced by Nolan et al. [7]. At first, the means and the stability exponents associated with each dimension are estimated independently. Then, the estimated mean vector and the stability exponent of the multivariate distribution are given by:

$$\hat{\boldsymbol{\mu}} = \left(\hat{\mu}_1,\ldots,\hat{\mu}_d\right)^{\mathrm{T}}, \quad \bar{\alpha} = \frac{1}{d}\sum_{k=1}^{d}\hat{\alpha}_k \tag{13}$$

The empirical multivariate characteristic function is obtained from a discrete formulation of Eq. (10). The discrete equation associated with the expectation is:

$$L\left(\mathbf{k}\right) = \mathbf{E}\left[\exp\left(i\left\langle \mathbf{k},\mathbf{X}\right\rangle\right)\right] \quad \Rightarrow \hat{L}\left(\mathbf{k}_i\right) = \frac{1}{M}\sum_{j=1}^{M}\exp\left(i\left\langle \mathbf{k}_i,\mathbf{X}_j\right\rangle\right) \tag{14}$$

where $\left\{\mathbf{X}_j\right\}_{j=1,\ldots,M}$ is the set of all the empirical multivariate data while the discrete equation associated with the right part of Eq. (10) is:

$$\hat{I}\left(\mathbf{k}_i\right) = -\ln\hat{L}\left(\mathbf{k}_i\right) = \sum_{j=1}^{n}\psi_{\bar{\alpha}}\left(\left\langle \mathbf{k}_i,\mathbf{s}_j\right\rangle\right)\Delta_j \tag{15}$$

If a symmetric grid is assumed for both the hypersphere and the frequency domain, the weights on the hypersphere $\left\{\Delta_i\right\}_{i=1,\ldots,n}$ may be estimated from the following constrained objective function

$$\min \quad \left\|\mathbf{c} - \mathbf{A}\,\boldsymbol{\Delta}\right\|^2 \quad \therefore \quad \boldsymbol{\Delta} \geq 0 \tag{16}$$

where

$$\mathbf{c} = \begin{bmatrix} \mathrm{Re}\left(\hat{I}\left(\mathbf{k}_1\right)\right),\ldots,\mathrm{Re}\left(\hat{I}\left(\mathbf{k}_m\right)\right), \\ \mathrm{Im}\left(\hat{I}\left(\mathbf{k}_1\right)\right),\ldots,\mathrm{Im}\left(\hat{I}\left(\mathbf{k}_m\right)\right) \end{bmatrix}^{\mathrm{T}} \qquad (17)$$
$$\mathbf{\Delta} = \left[\Delta_j\right], \quad \mathbf{A} = \left[A_{i,j}\right]$$

and where

$$A_{i,j} = \begin{cases} \mathrm{Re}\left(\psi_{\bar{\alpha}}\left(\left\langle \mathbf{k}_i, \mathbf{s}_j \right\rangle\right)\right) & i,j = 1,\ldots,m \\ \mathrm{Im}\left(\psi_{\bar{\alpha}}\left(\left\langle \mathbf{k}_i, \mathbf{s}_j \right\rangle\right)\right) & i,j = m+1,\ldots,2m \end{cases} \qquad (18)$$

For instance, in two dimensions, the symmetric grid is equal to:

$$\left\{\mathbf{s}_i\right\}_{i=1,\ldots,n} = \left\{\mathbf{k}_i\right\}_{i=1,\ldots,n} = \left\{\left[\cos\left(\frac{2\pi\left(i-1\right)}{n}\right), \sin\left(\frac{2\pi\left(i-1\right)}{n}\right)\right]^{\mathrm{T}}\right\}_{i=1,\ldots,n} \qquad (19)$$

Then, we use this grid in order to define the generalization of the covariance for the stable distribution.

4 Generalization of the Covariance: the Covariation

We extend the concept of covariance to stable distributions. This is important, in practice, because we often need to determine if two variables are correlated or not. This is the case, for instance, when aiming to protect data privacy, where the goal is to determine if an attribute may be inferred from another [8]. If the data are distributed according to a Lévy distribution, the concept of covariance must be generalized with the concept of covariation [9]. The covariation in between two stochastic stable variables is defined as

$$\left[\!\left[X_1, X_2 \right]\!\right]_\alpha = \int_{S^2} k_1\ k_2^{\langle \alpha-1 \rangle} \Delta\left(ds\right) \quad \therefore \quad x^{\langle \alpha-1 \rangle} \triangleq \left|x\right|^{\alpha-1} \mathrm{sgn}\left(x\right) \qquad (20)$$

where the measure on the bidimensional sphere is associated with the bivariate distribution of the vector formed from the concatenation of the two stochastic variables involved in the convolution $\Delta\left(ds\right) \quad \Leftrightarrow \quad \mathbf{X} = \left[X_1, X_2\right]^{\mathrm{T}}$.

Such a measure may be estimated with the method presented in the previous section and with the grid introduced in Eq. (19). From the covariation, it is possible to define a norm which is also related to the scale parameter γ_X of the variable:

$$\left\|X\right\|_\alpha = \left(\left[\!\left[X, X \right]\!\right]_\alpha\right)^{\frac{1}{\alpha}} = \gamma_X \qquad \left[\!\left[X_1, X_2 \right]\!\right]_2 = \frac{\mathbf{Cov}\left(X_1, X_2\right)}{2} \qquad (21)$$

The covariation reduces to the covariance when the distribution is Gaussian, i.e. when the stability exponent is equal to two and the asymmetry is zero. The correlation belongs to the interval $\left[0,1\right]$ where zero indicates an absence of correlation while the unity indicates a strong correlation. As shown in the next section, the covariance and the correlation tend to be much stronger if the stability exponent is less than two.

Practically, if one incorrectly assumes a Gaussian distribution from the start, one may strongly underestimate the real correlation between two variables. Such an underestimation might have severe consequences [10]. For instance, consider the scenario where private attributes should be identified and protected. Suppose that an absence of correlation between two attributes is assumed, due to an erroneous assumption that the data distribution is Gaussian. However, in reality, these two attributes may be highly correlated. This incorrect assumption would result in an absence of protection for sensitive attributes, when they do indeed need to be protected against induced attacks.

5 Experimental Results

In this section, we present various simulations and experiments against real-world financial databases which illustrate our previous theoretical results. All experiments were performed using Mathematica 8.0 on a Dell Precision M6400. In the following, one should keep in mind that $\alpha = 2$ corresponds to a Gaussian distribution.

Table 1. Estimation of the exponent of the Lévy distributions associated with various Stock Exchange. Excerpted from [11].

Index	**Period**	α
FTA W Jap	86.01-93.09	1.808
TOPIX	75.01-91.02	1.519
MSCI Japan Net	80.01-93.09	1.463
Nikkei 225	80.01-93.09	1.626

We analyze some results on financial data as reported by Lévy Véhel and Walter (LVW) [11] as well as from the PKDD 1999 discovery challenge financial database [12] in section 5.1. The importance of stable distribution is not only theoretical; as a matter of fact, it has far reaching consequences for financial 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 1 which shows the results obtained for various Stock Market Indexes in Europe and in Japan by LVW [11]. The stability exponent was estimated with the method presented in Section 3 and the null hypothesis was asserted with the Kolmogorov-Smirnov test. As shown by the data, all these indexes clearly have a stable distribution (confidence level of 99%) and the value of the stability exponent is typically in between 1.6 and 1.8, which is clearly not in the Gaussian regime.

Table 2. Covariation of two shares (Thompson and Michelin) and the CAC 40 Index for various values of the stability exponent. and the acquisition period is from 87.07.09 to 95.05.31. Excerpted from [11].

$[\![X_1, X_2]\!]_\alpha$	α			
	2.0	**1.7**	1.5	1.3
THOMPSON, CAC 40	0.042	**0.157**	0.390	0.975
MICHELIN, CAC 40	0.042	**0.159**	0.326	0.993
CAC 40, CAC 40	0.036	**0.128**	0.300	0.750

Table 2 shows the covariations in between the Michelin and the Thompson titles as well as with the CAC 40 Stock Exchange Index. The covariations may be estimated with the method presented in Sections 4. The calculation was repeated for various values of the stability exponent; the real one being around 1.7. Table 2 shows that, if a Gaussian distribution is incorrectly assumed, the correlation (covariation) tends to be underestimated. For instance, the covariation in between the Michelin share and the CAC 40 Index is 0.042 with the false assumption of a Gaussian distribution for the data while in reality it is 0.159 for a stability exponent of 1.7. Here, the covariation tends to be stronger, when the stability exponent is smaller.

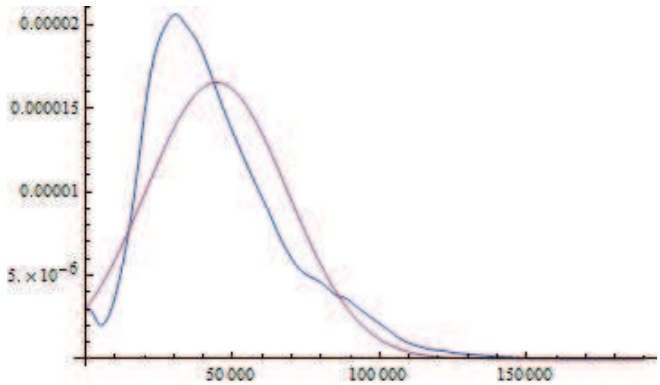

Fig. 1. Distribution of the Balance attribute (PKDD 1999 discovery challenge financial database) in blue and of the best fitting normal distribution in red; the later is obtained with the maximum of likelihood method.

5.1 PKDD 1999 Discovery Challenge Financial Database

Next, we consider the PKDD 1999 discovery challenge financial database, which has been widely used as a benchmark in the multi-relational classification domain. This database was offered by a Czech bank and contains data describing the level of risk of a customer to default on a loan [12]. The database consists of eight tables. The Account table contains the account number, as well as the information regarding the district a person falls in and the frequency of payment. Other tables include the Demographic profile, the client's Disposition in terms of type, Credit Card information and Client descriptions, including gender and the location in which they reside. The Order table details the number of money transfers and the Loan table describes the payments of loans. Very often, this database is used to classify whether a Loan is at risk or not.

We turn our attention to the Transaction table, which describes all transactions against an account. This table contains, amongst others, the Balance attribute that contains 54694 entries. We apply a number of feature selection algorithms to this database, including the Gain ratio, Chi Squared and Correlation based Feature Selection (CFS) measures [8]. Our results indicate that the Balance attribute is always

selected as being one of the features that are strongly related to the outcome of a Loan, with and without aggregation.

Figure 1 shows the distribution of the Balance attribute as well as the best fitting normal distribution. The parameters of the normal distribution are obtained with the maximum likelihood method. It follows that the normal distribution offers a poor fit to the Balance attribute distribution. We attempted to fit numerous distributions to the Balance such as the Student distribution, the Weibull distribution, amongst others. However, the best fit was obtained with the stable distribution, as illustrated in Fig. 2. The parameters corresponding to this distribution, which were obtained with the maximum likelihood method, are shown in Table 3. The tails of the Balance attribute distribution and the associated stable distribution are shown in Fig. 3. The tail of the fitted stable distribution seems thicker than the tail of the Balance attribute distribution: nevertheless, they both have the same order of magnitude. As may be observed from Fig. 3, the statistics of the tail is quite poor. This means that, in order to draw definitive conclusions about the tails, a database with more entries would be required.

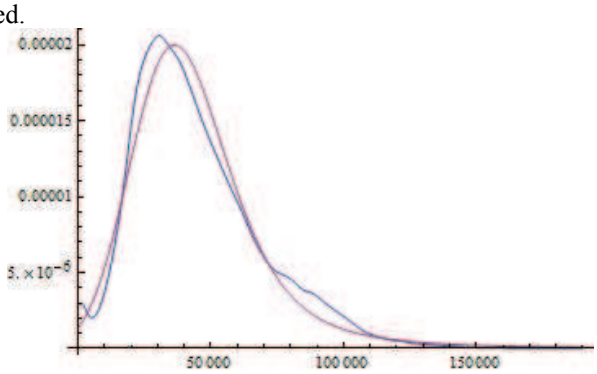

Fig. 2. Distribution of the Balance attribute in blue and of the best fitting stable distribution in red; the later is obtained with the maximum of likelihood method.

Table 4 shows the mean, the standard deviation, the skewness and the kurtosis calculated from the normal and the stable distributions associated with the Balance attribute distribution. Since the stability exponent α is smaller than two, it is not possible to evaluate the standard deviation and the skewness from the stable distribution, because the statistical moments needed for the calculation are infinite.

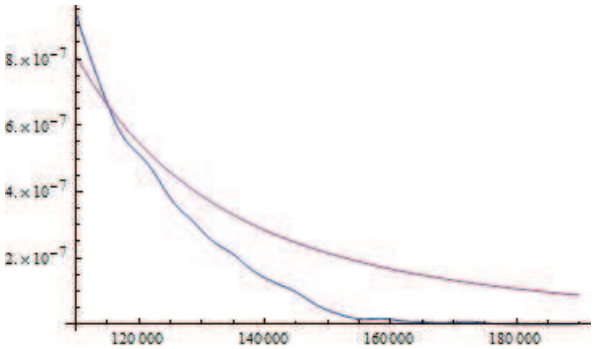

Fig 3. Tail of the Balance attribute distribution in blue and of the stable distribution in red.

TABLE 3. Parameters of the fitted stable distribution associated with the Balance attribute.

α	β	μ	γ
1.6232	1.0	47321.3	14013.9

Nevertheless, the parameters of the stable distribution provide a measure of the standard deviation and of the skewness through the scale parameter γ and β the asymmetry parameter. This implies that, when the underlying distribution is stable, the scale and the asymmetry should not be estimated directly from the data, but from the parameters of the fitted distribution.

Table 4. Mean, standard deviation, skewness and kurtosis as obtained from the fitted normal and stable distributions associated with the Balance attribute.

Distribution	Mean	Std Dev.	Skewness	Kurtosis
Normal	44534.2	24109.7	0	0
Stable	47321.3	∞	∞	∞

In order to further stress the importance of not directly estimating these parameters from the data, we have generated three data sets. These datasets consisted of 54694 entries as in the original bank database with the stable distribution parameterized by Table 4. Table 5 shows the results. Since the stability exponent is greater than one (1.62), the estimation of the mean from the data is consistent from one set to the next. However, because the stability exponent is less than two (1.62), it is not possible to estimate any moment greater or equal than two which means that the standard deviation, the skewness and the kurtosis are not consistent from one synthetic data set to the next. In other words, any measure based on the statistical moments greater or equal than two is a random number which will vary from one realization of the data set to the other as shown in Table 5. These findings suggest that when correlation evaluation is sensitive, say in privacy preservation learning, one should carefully select the right equation for correlation computation [8].

Table 5. Mean, standard deviation, skewness and kurtosis as obtained from data generated from the fitted stable distribution associated with the Balance attribute. Each generated data set consists of 54694 entries.

Stable	Mean	Std Dev.	Skewness	Kurtosis
1	47419.2	99020.1	145.408	27467.7
2	47996.1	158948	168.972	33756.5
3	47041.7	46051.2	18.1492	677.382

6 Conclusions

The development of new data mining models for catastrophic event prediction, such as the damages caused by oil spills, stock market crashes, tsunamis and other natural disasters, are an important and urgent research topic. In this communication, we have analyzed data aggregation and the data covariance (covariation), of such data, where the underlying distributions are not Gaussian, but stable. We have shown that such data may be aggregated with the mean, but not with the variance. This is due

to the fact that the variance becomes infinite and its estimate tends to fluctuate randomly when evaluated on a finite size aggregate. We have also shown that the estimation of the mean converges rather slowly when the stability exponent is small. In this case, both the mean and the variance are infinite and their estimate on a finite size aggregate tends to fluctuate randomly. In these circumstances, the aggregate is better characterized with its maximum which tends to dominate by many orders of magnitude over the other elements of the aggregate, both from a sum and rank ordering statistics point of view. We have shown that financial data may be characterized with stable distributions with a stability exponent typically around 1.7. The calculation of the covariations in between Stocks and Stock Market Indexes has shown that the covariance (covariation) tends to be underestimated if a Gaussian distribution is wrongly assumed. We also shown that, for a well-known benchmarking financial database, some attribute values follow a stable distribution rather than the normal distribution. This work thus has implications when pre-processing and mining many highly imbalanced data that are typified with large-scale fluctuations, such as earthquake and oil spill data, which would be worth investigating further. We also mentioned, in Section 4, that our approach is highly relevant for correlation-based privacy preservation data mining. We aim to explore this research issue in our future work.

References

1. Knobbe, A., Siebes, A., Marseille, B.: Involving Aggregate Functions in Multi-Relational Search, Principles of Data Mining and Knowledge Discovery. LNCS, vol. 2431, pp. 145--168 (2002)
2. Malerba, D.: A relational perspective on spatial data mining. Int. J. Data Mining. Modelling and Management 1 (1), pp. 103--118 (2008)
3. Walter, C.: 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), pp. 37--56 (1999)
4. Groot, R. D.: Lévy distribution and long correlation times in supermarket sales. Physica A: Statistical Mechanics and its Applications 353, pp. 501--514 (2005)
5. Samorodnitsky, G., Taqqu, M. S.: Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance. Chapman & Hall, New York (1994)
6. Paulson, A. S., Holcomb, E., Leitch, R.: The estimation of the parameters of the stable law. Biometrica 62 (1), pp. 163--170 (1977)
7. Nolan, J. P., Panorska, A. K., McCulloch, J. H..: Estimation of spectral measures. Mathematical and Computer Modelling 34 (9-11), pp. 1113--1122, (2001)
8. Guo, H., Viktor, H. L., Paquet, E.: Privacy Disclosure and Preserving in Learning with Multi-relational Databases. Journal of Computing Science and Engineering 5 (3), pp. 183-196 (2011)
9. Cheng, B., Rachev, S.: Multivariate Stable Future Prices. Mathematical Finance 5, pp. 133--153 (1995)
10. Tao, Y., Pei, J., Li, L., Xiao, X., Yi, K., Xing, Z.: Correlation hiding by independence masking. IEEE 26th International Conference on Data Engineering (ICDE). pp. 964--967, (2010)
11. Lévy Véhel, J., Walter, C.: Les marchés fractals ("The fractal markets"). Presses Universitaires de France, Paris (2002)
12. Berka, P.: Guide to the Financial Data Set. In A. Siebes and P. Berka, editors, PKDD 2000 Discovery Challenge (2000)