Exploring the Influence of Sampling on Pattern Support Distribution

Luofeng Xu, Stephen Marsland, Ruili Wang
School of Engineering and Advanced Technology, Massey University
Private Bag 11 222, Palmerston North, New Zealand
{l.xu, s.r.marsland, r.wang}@massey.ac.nz

Abstract

Identifying the pattern support distribution (PSD) in datasets is useful for many data mining tasks, such as market basket analysis. The support of a pattern is the frequency of its occurrence in a dataset. Calculating the distribution of these supports over an entire dataset is computationally expensive; this cost can be reduced by sampling from the dataset and computing the PSD on a relatively small sample. However, this may miscount patterns and cause significant changes in the distribution identified. Based on the fact that the PSD shows a power-law relationship, in this paper we investigate the influence of sampling on the characteristics of the power-law relationship in the pattern support distribution. We consider sampling effect on this relationship under two assumptions: uniform distribution of pattern supports, and independent identically distributed (i.i.d.) distributions. We experimentally evaluate the influence on data from four real-world transaction datasets.

1. Introduction

Market baskets and similar datasets are a staple feature of data mining problems. From a set of m possible items, each basket consists of a subset of these items denoting, for example, the purchases of one shopper from all of the items it is possible to buy. Analysis of these baskets is used to find items that are frequently bought together to inform shop layout and allow for targeted marketing, amongst other applications [10].

One of the important characteristics of a market basket is the pattern support distribution (PSD). The support of a pattern is the frequency of its occurrence in a dataset, and the distribution of patterns against their corresponding supports in a transaction dataset is known as the pattern support distribution. From the view of probability theory and statistics, the PSD is a discrete probability distribution that expresses the probability that the support of an arbitrary pattern in a dataset is equal to some particular value. An alternative view is that the PSD indicates the connection between the cohesion of a pattern in a pattern set and the size of the pattern set in a dataset. The PSD is of use in many data mining tasks, such as providing a way of determining an appropriate minimum support for mining frequent patterns [4].

To identify the PSD in a dataset requires finding the corresponding support for every pattern in the dataset. For large datasets, this computational cost is generally prohibitive. One way to alleviate this cost is to sample from the dataset and compute the supports of patterns based on this sample (or set of samples). However, the results are not as accurate as those based on the entire dataset [10, 4]. Consequently, the pattern support distribution based on samples can differ from the one based on the entire dataset.

In this paper we are interested in the influence of sampling on the characteristics of the power-law relationship in the pattern support distribution. We will compute and analyze the difference between the PSDs from a dataset and its samples. This study can lead to methods of identifying the PSD in a dataset based on that found in its samples.

2. Background

We begin by defining the pattern support distribution, and then introduce related work.

2.1. Pattern support distribution (PSD)

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of m distinct literals called items. A transaction dataset comprises a set of transactions $D = \{t_1, t_2, \ldots, t_n\}$, where each transaction $t \in D$ is a nonempty subset of I. The length of a transaction $|t|$ describes the number of items in t. Each of the possible nonempty subsets of I is known as a pattern. The absolute support of a pattern z in dataset D, $S_A(z, D)$, is the total number of transactions containing z in D, while the relative support is $S_R(z, D)$ $(= S_A(z, D)/|D|)$. The pattern support distribution in D is discrete and comprises a set of

978-0-7695-3242-4/08 $25.00 © 2008 IEEE
DOI 10.1109/CIT.2008.Workshops.91

points (S, y) with an absolute/relative support value S and a number y of patterns with support value S.

2.2. Sampling

Sampling has been successfully used in many applications where large datasets make computations expensive and approximate results are acceptable. The goal of sampling here is to select a sample (or set of samples) from the original large dataset so that the pattern support distribution in the dataset can be estimated cheaply from the sample.

Stratified sampling according to transaction length is considered here. Transactions are randomly selected so that the number of patterns of different lengths matches the proportions in the original dataset. This is required because the longer a transaction, the more patterns it is likely to contain.

2.3. Power-law relationships in PSDs

The discovery of power-law relationships (distributions of the form $f(x) \propto x^{-\alpha}$ for some positive constant α, also known as Zipf's law or heavy tails) in empirical observations of many man-made and natural phenomena has been incredibly popular in recent times [7, 1, 5, 6]. Their ubiquity makes them of interest, but they are often favoured for the fact that they appear as straight lines on log-log plots, which makes them easy to identify.

Two recent papers [3, 4] used empirical studies on various retail transaction datasets to demonstrate that power-law relationships exist in the pattern support distributions of real datasets. It is suggested that this is caused by self-similarity within the pattern support distribution.

Since power-law relationships show up as straight lines in log-log space, a PSD can be described in the form:

$$\ln(y_i) = a \ln(S(i)) + b \ \ (i = 1, 2, \dots, k), \qquad (1)$$

where (S, y) is a point in the pattern support distribution, and k is the maximum support value in D. The distribution is parameterized by the slope a and Y-intercept b. The values for a and b can be roughly calculated by linear regression[1], as in [4].

$$a = \frac{\sum_{i=1}^{k} \ln(S(i)) \ln(y_i)}{\sum_{i=1}^{k} \ln^2(S(i)) - \frac{1}{k}\left(\sum_{i=1}^{k} \ln(S(i))\right)^2} - \frac{\frac{1}{k}\left(\sum_{i=1}^{k} \ln(S(i))\right)\left(\sum_{i=1}^{k} \ln(y_i)\right)}{\sum_{i=1}^{k} \ln^2(S(i)) - \frac{1}{k}\left(\sum_{i=1}^{k} \ln(S(i))\right)^2}. \qquad (2)$$

$$b = \frac{1}{k}\sum_{i=1}^{k} \ln(y_i) - \frac{a}{k}\sum_{i=1}^{k} \ln(S(i)). \qquad (3)$$

3. The influence of sampling

The distribution of support values for the patterns in a sample typically deviates from the underlying distribution in the original dataset. This is sometimes known as the support-deviation phenomenon. It can be minimized by determining an appropriate sample size in order to control the bias of the estimated support for patterns in the dataset. For example, progressive sampling [8] aims to keep increasing the sample size and testing the corresponding model accuracy until the model accuracy stops increasing notably. Statistical approaches include using the Central Limit Theorem [10] and Chernoff bound [9]. Another approach [4] used a relatively large sample size and quantized the support distribution in order to reduce the support deviation.

We focus on another problem with sampling – known as the shape deviation phenomenon – which is that the shape of the pattern support distribution differs between the entire dataset and its sample. In this paper, we attempt to discover the relationship between the sample size chosen and the shape deviation of the pattern support distribution, when the stratified sampling with replacement is used.

3.1. Uniformly distributed item supports

In this section, the influence of sampling on pattern support distribution is analyzed based on the assumption that the items in the targeted transaction dataset are uniformly distributed. In other words, each transaction in the dataset has a uniform probability of containing each item in the dataset. Even though this assumption is apparently far from reality – especially in a transaction dataset where the pattern support distribution is a discrete distribution – the discussion can still highlight some interesting phenomena and provide preliminary understanding of the PSD in datasets where each item is treated independently.

Suppose that the pattern support distribution in a transaction dataset D satisfies a power-law relationship, which can be written in the form:

$$\ln(f_D(X = x)) = a \ln(x) + b, \qquad (4)$$

where $f_D(X = x)$ denotes the number of the patterns whose support is equal to x in D. A sample ζ is randomly drawn from D. Treating the pattern support distribution as a continuous distribution enables the computation of absolute support for a pattern z in ζ as:

$$\frac{S_A(z, \zeta)}{|\zeta|} \approx \frac{S_A(z, D)}{|D|}. \qquad (5)$$

[1] More accurate methods using a maximum likelihood framework can be derived, but for the sake of simplicity the linear regression estimator is used here.

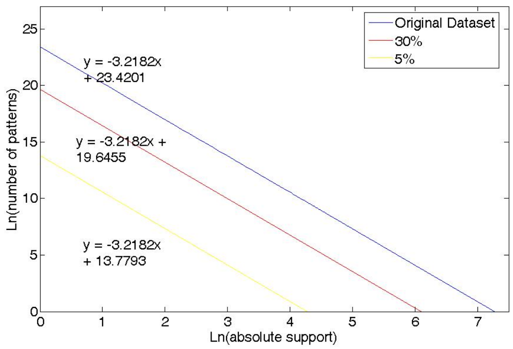

Figure 1. The influence of sampling on the PSD in a dataset where the support of each pattern is continually uniformly distributed

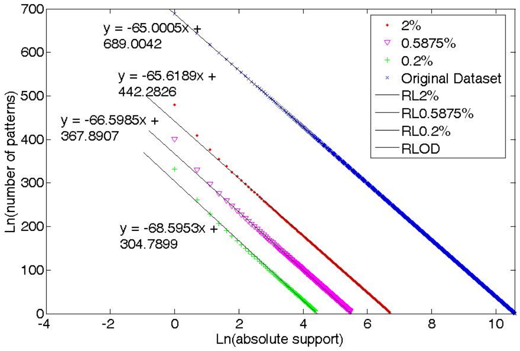

Figure 2. The influence of sampling on the PSD in a dataset where the support of each pattern is discretely uniformly distributed

Under the assumption of uniformly distributed items, the distribution of support values in ζ will approximate those in D, with the match getting stronger as the sample size increases. We can therefore make the approximation that:

$$f_D(S_A(z, D)) = f_\zeta(S_A(z, \zeta)). \tag{6}$$

Accordingly, the pattern support distribution in ζ can be described by:

$$\ln(f_\zeta(X = x)) = a\left(\ln(x) + \ln\left(\frac{|D|}{|\zeta|}\right)\right) + b, \tag{7}$$

where the values of a and b will match those computed in Equations 2 and 3 for D. Figure 1 shows the effect of this based on the samples drawn from an artificially created dataset where the support of each item is independent uniformly distributed. It clearly demonstrates that the influence of sampling on the pattern support distribution can be precisely predicted by Equation 7, namely an affine shift along the X-axis by $\ln(|D|/|\zeta|)$. However, in reality the PSD of a transaction dataset is discrete and the value of the absolute support of a pattern in a transaction dataset has to be a positive integer. As a result, each point in the PSD in ζ is aggregated by (on the average) $|D|/|\zeta|$ corresponding sequential points in the PSD of D. Therefore, the number of patterns whose absolute support is equal to a positive integer S'_A in ζ, approximately satisfies:

$$f_\zeta(S'_A) \approx \sum_{i=1}^{\frac{|D|}{|\zeta|}} f(S_{Ai}) = e^b \sum_{i=1}^{\frac{|D|}{|\zeta|}} (S_{Ai})^a. \tag{8}$$

We can put bounds on this approximation very simply since $S'_A/|D| \approx S_A/|\zeta|$, and so:

$$\begin{aligned}\ln(f_\zeta(S'_A)) &= b + \ln\left(\sum_{i=1}^{\frac{|D|}{|\zeta|}} f(S_{Ai})\right)\\ &\approx \left(e^b \sum_{i=1}^{\frac{|D|}{|\zeta|}} (S_{Ai})^a\right)\\ &= a\ln(S'_A) + (a+1)\ln\left(\frac{|D|}{|\zeta|}\right) + b.(9)\end{aligned}$$

Since $S'_A - 0.5 \leq \frac{|\zeta|}{|D|} S_{Ai} < S'_A + 0.5$ and $a < 0$,

$$\begin{aligned}&\ln\left(\sum_{i=1}^{\frac{|D|}{|\zeta|}} (S_{Ai})^a\right) - \ln\left(\sum_{i=1}^{\frac{|D|}{|\zeta|}} \left(\frac{|D|}{|\zeta|} S'_A\right)^a\right)\\ \geq\ &\ln\left(\frac{|D|}{|\zeta|}\right) + a\ln\left(1 - \frac{1}{2S'_A}\right)\ ,\end{aligned} \tag{10}$$

and

$$\begin{aligned}&\ln\left(\sum_{i=1}^{\frac{|D|}{|\zeta|}} (S_{Ai})^a\right) - \ln\left(\sum_{i=1}^{\frac{|D|}{|\zeta|}} \left(\frac{|D|}{|\zeta|} S'_A\right)^a\right)\\ <\ &\ln\left(\frac{|D|}{|\zeta|}\right) + a\ln\left(1 + \frac{1}{2S'_A}\right)\ .\end{aligned} \tag{11}$$

As can be seen in Figure 2, when the pattern support distribution is treated as a discrete distribution, the influence of sampling approximately implements an affine transformation in log-log space that shifts the PSD of D by $\ln(|D|/|\zeta|)$ along the X-axis, then shifts it by $\ln(|D|/|\zeta|)$ along the Y-axis. Note that, as shown by Equations 10 and 11, the real behaviour of the influence of sampling is slight different to the one shown by Equation 9. With a certain sampling ratio, the closer S'_A is to 1, the steeper the slope in S'_A is.

Of course, in reality the support of each item is seldom independent uniformly distributed in transaction datasets. While there is very likely to be some correlation between items within a dataset, and therefore possible correlations within support distributions, this is not usually known to the data miner; indeed this may be one of the features they want to discover. We therefore consider the case of datasets where each item support can be treated as if it has an independent unknown distribution.

3.2. Independently distributed item supports

Suppose that a pattern z in D is randomly distributed with an unknown distribution. For a given transaction t, the pattern z is either a subset of t, or it is not. When randomly sampling with replacement, the probability of a pattern z being included in any given transaction in D is equal to its relative support in D, i.e., $S_R(z, D)$, and the probability that it is not included is $1 - S_R(z, D)$.

The absolute support of a pattern z in a random sample ζ will follow a binomial distribution, with probability mass function:

$$\begin{aligned} f(S_A(z,\zeta); |\zeta|, S_R(z,D)) = \\ C^{|\zeta|}_{S_A(z,\zeta)} (S_R(z,D))^{S_A(z,\zeta)} \times \\ (1 - S_R(z,D))^{|\zeta| - S_A(z,\zeta)}, \end{aligned} \quad (12)$$

where the absolute support $S_A(z, \zeta)$ runs over all values from 1 up to $|\zeta|$.

Figure 3 shows the effect of randomly sampling: the absolute support value $S_A(z, \zeta)$ of a pattern z in ζ may not equal the expected value, which is $S_R(z, D) \times |\zeta|$. In particular, the binomial distribution is a discrete distribution. In fact, these issues can significantly affect the values of the parameters of the pattern support distribution in a sample. Given a sample ζ randomly drawn with replacement from D, the patterns with support S in D will have the same probability that their support in ζ is equal to certain support value S'. Let (S_A, y) be a point in the PSD of D, where S_A denotes an absolute support value, and y is the number of patterns with that support. As the number of patterns with the same support value increases, the distribution will tend to be binomial. In this case the number y' of patterns that have absolute support S_A in D and S'_A in ζ is:

$$\begin{aligned} y' &\approx y\, f\left(S'_A; |\zeta|, \frac{S_A}{|D|}\right) \\ &= y\, C^{|\zeta|}_{S'_A} \left(\frac{S_A}{|D|}\right)^{S'_A} \left(1 - \frac{S_A}{|D|}\right)^{|\zeta| - S'_A}. \end{aligned} \quad (13)$$

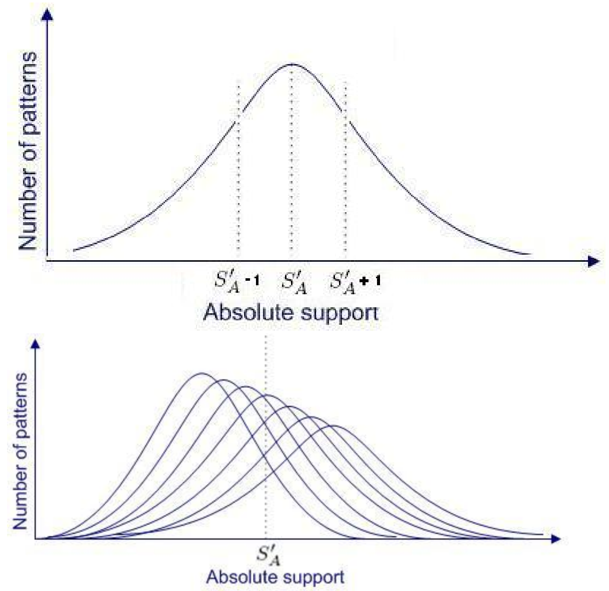

Figure 3. The influence of sampling on the absolute support value of a pattern

Let (S'_A, y') be a point in the PSD of a sample ζ drawn from a transaction dataset D. S'_A indicates an absolute support value in ζ. When the absolute supports of patterns in ζ ideally follow their corresponding binomial distributions, or y_i tends to a very large positive number,

$$\begin{aligned} f_\zeta(S'_A) &\approx \sum_{i=1}^{k} \left(f_D(S_A(i)) f\left(S'_A; |\zeta|, \frac{S_A(i)}{|D|}\right)\right) \\ &= \sum_{i=1}^{k} \left(f_D(S_A(i)) C^{|\zeta|}_{S'_A} \left(\frac{S_A(i)}{|D|}\right)^{S'_A} \times \right. \\ &\qquad \left. \left(1 - \frac{S_A(i)}{|D|}\right)^{|\zeta| - S'_A}\right), \end{aligned} \quad (14)$$

where $S_A(i)$ is a particular absolute support value in the original dataset D, and k is the maximum support value in D.

4. Experimental studies

Four real transaction datasets are used in this paper. All of them are widely used in data mining as benchmark datasets[2]. The parameters of these datasets are summarized in Table 1, where $|I|$ indicates the number of items in the

[2] Available from the Frequent Itemset Mining Implementations Repository (`http://fimi.cs.helsinki.fi/data/`) and the UCI Machine Learning Repository (`http://www.ics.uci.edu/~mlearn/MLRepository.html`).

dataset, $|D|$ indicates the number of transactions, $l_{\max}$ indicates the maximum length of patterns, and l_{avg} indicates the average length of patterns. The program for retrieving the patterns and their corresponding supports in a dataset is derived from the FP-growth algorithm[3] [2].

Table 1. Parameters of the datasets

Dataset	$\|I\|$	$\|D\|$	$l_{\max}$	l_{avg}
Retail	16,470	88,162	76	10.3
Adult	29,035	48,842	15	15
Accidents	468	340,183	51	33.8
Pumsb	2,113	49,046	74	74

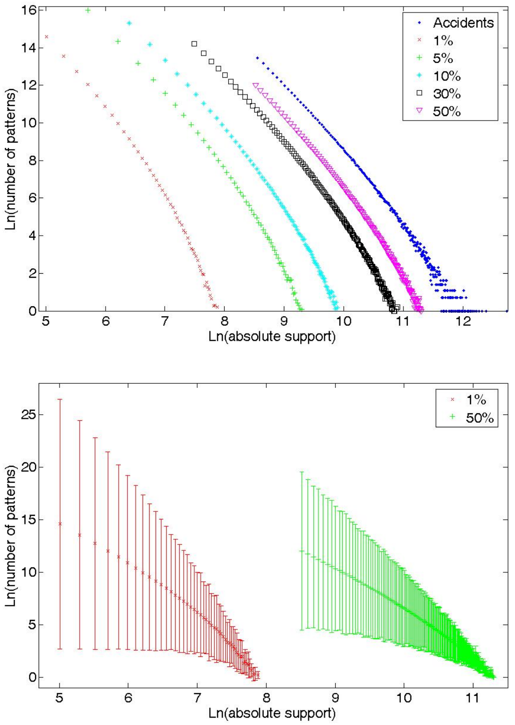

Figure 4. Observations on the PSDs in the Accidents dataset and its samples

Figure 4 shows the observations of the PSDs in the Accidents dataset and its samples. The first graph shows five different sampling ratios: 0.01 (1%), 0.05 (5%), 0.1 (10%), 0.3 (30%), and 0.5 (50%). Each PSD shown is the mean result of 30 trials. The log-log plots show the logarithm of absolute support value against the logarithm of the number of the corresponding patterns. The second graph shows the mean results and the corresponding standard deviations of the PSDs drawn by two different sampling ratios: 1% and 50%.

The retrieval process does not identify every pattern in any dataset, retrieving only those whose absolute support is greater than 30. However, the absolute support ranges of the pattern support distributions are different in a dataset and its samples. Therefore, a fixed minimum relative support value was used for the retrieval process, so that the shape deviation could be explored in the same relative support range for the PSDs in a dataset and its samples. For example, in Figure 4, to allow easy comparison the PSDs only contain the patterns whose relative support is greater than 0.044.

As can be seen, the shapes of the PSDs in the samples are different to the one in the original dataset. This clearly shows the existence of the influence of sampling and the effect of sample size on PSD. Moreover, it suggests that, with a certain sampling ratio, the part of the PSD with high-support value has a lower standard deviation than the part with low-support value. With the decrease in the sampling ratio, the corresponding standard deviation grows and the effect of the noise existing in the high-support value part of the pattern support distribution is reduced.

The datasets described in Table 1 are further examined for the purpose of practically verifying the analysis of the influence of sampling. Although the pattern support distributions in the samples drawn from a dataset can still be fitted into a straight line, the values of the slope and Y-intercept of the straight line are different to those of the original PSD. In addition, the influence of sampling varies with sample size.

Although the degree of accuracy of estimation by sampling mostly relies on the sample size taken, based on the theoretical analysis and experimental studies presented in this paper, a very large sample size may not be a necessary condition for estimating the pattern support distribution in the original dataset with a desired bias, as shown in Figure 5.

5. Conclusions

To identify the pattern support distribution in a dataset requires discovering all patterns and their supports in the dataset. This is very computationally expensive, since the number of patterns in a dataset is often large. For the sake of efficiency, sampling methods can be adopted to solve the problem. Sampling can reduce the cost of retrieving the patterns and their supports. However, at the same time, sampling can also miscount patterns. As a result, the shape of the PSD in a sample deviates from that of the original dataset. In this paper, the influence of sampling on pattern support distribution is investigated and evaluated. Studying the influence is useful for efficiently and effectively estimating the original PSD from a sample instead of the original

[3] Available from `http://fuzzy.cs.uni-magdeburg.de/~borgelt/fpgrowth.html`

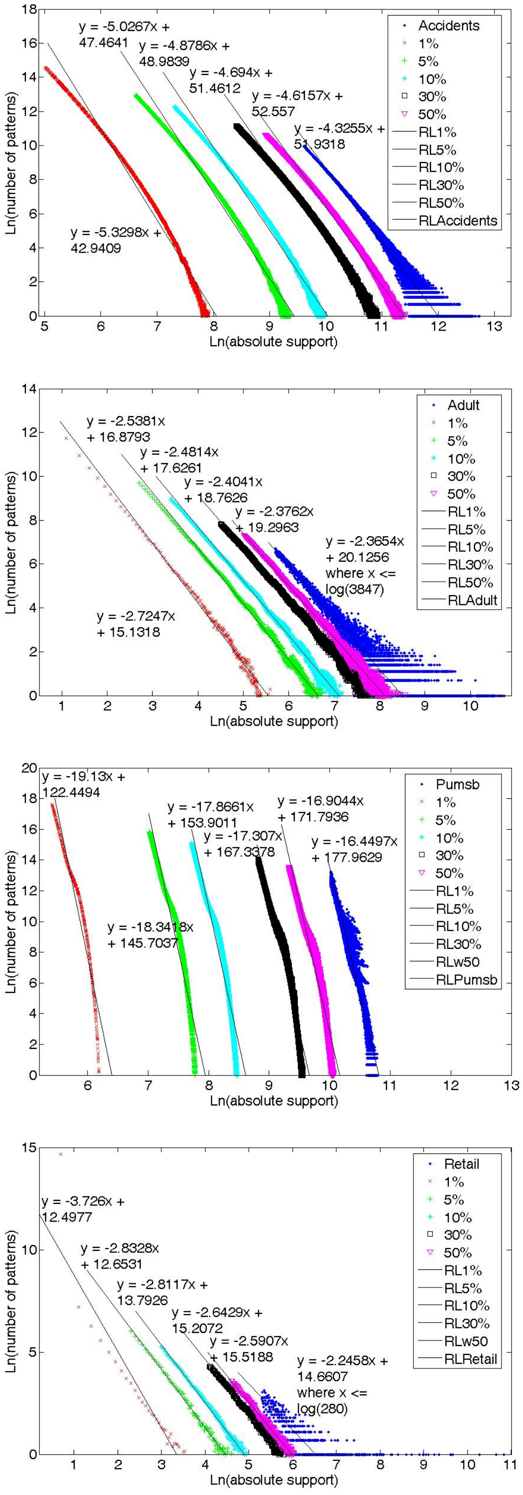

Figure 5. The influence of sampling on the characteristics of the power-law relationships in the PSDs in the four real datasets

dataset. As shown in the experimental studies, the influence of sampling on PSD has been demonstrated in four real transaction datasets. The points with high supports in PSDs in the four datasets have large support fluctuation and this phenomenon affects identifying the true characteristics of PSDs. The experimental studies clearly explores that the support fluctuation in the high-support part of PSDs is significantly reduced in samples. Thus, using sampling methods to estimate the characteristics of the power-law relationship in a PSD can also decrease the effect of noise on the high-support part of the PSD.

Acknowledgements

Thanks go to Dr Siva Ganesh and Associate Professor Mark Bebbington for their support and advice at the beginning of this study.

References

[1] L. Adamic and B. A. Huberman. The nature of markets in the world wide web. Computing in economics and finance 1999, Society for Computational Economics, May 1999.

[2] C. Borgelt. An implementation of the fp-growth algorithm. In *OSDM '05: Proceedings of the 1st international workshop on open source data mining*, pages 1–5, New York, NY, USA, 2005. ACM Press.

[3] K.-T. Chuang, J.-L. Huang, and M.-S. Chen. On exploring the power-law relationship in the itemset support distribution. In *EDBT*, pages 682–699, 2006.

[4] K.-T. Chuang, J.-L. Huang, and M.-S. Chen. Power-law relationship and self-similarity in the itemset support distribution: analysis and applications. *The International Journal on Very Large Data Bases*, July 2007.

[5] S. Miyazima, Y. Lee, T. Nagamine, and H. Miyajima. Power-law distribution of family names in Japanese societies. *Physica A: Statistical Mechanics and its Applications*, 278:282–288, April 2000.

[6] R. M. W. Musson, T. Tsapanos, and C. T. Nakas. A power-law function for earthquake interarrival time and magnitude. *Bulletin of the Seismological Society of America*, 92(5):1783–1794, June 2002.

[7] G. Neukum and B. A. Ivanov. Hazards due to comets and asteroids. chapter Crater size distributions and impact probabilities of Earth from lunar, terrestrial-planet, and asteroid cratering data, pages 359–416. Univ. of Ariz. Press, 1994.

[8] S. Parthasarathy. Efficient progressive sampling for association rules. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, page 354, Washington, DC, USA, 2002. IEEE Computer Society.

[9] M. J. Zaki, S. Parthasarathy, W. Li, and M. Ogihara. Evaluation of sampling for data mining of association rules. In *Proceedings of the 7th International Workshop on Research Issues in Data Engineering*, pages 42–50, April 1997.

[10] C. Zhang and S. Zhang. *Association rule mining: models and algorithms*. Springer-Verlag New York, Inc., New York, NY, USA, 2002.