Sampling and Evaluating the Big Data for Knowledge Discovery

Andrew H. Sung[1], Bernardete Ribeiro[2] and Qingzhong Liu[3]
[1]School of Computing, The University of Southern Mississippi, Hattiesburg, MS 39406, U.S.A.
[2]Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
[3]Department of Computer Science, Sam Houston State University, Huntsville, TX 77341, U.S.A.

Keywords: Data Analytics, Knowledge Discovery, Sampling Methods, Quality of Datasets.

Abstract: The era of Internet of Things and big data has seen individuals, businesses, and organizations increasingly rely on data for routine operations, decision making, intelligence gathering, and knowledge discovery. As the big data is being generated by all sorts of sources at accelerated velocity, in increasing volumes, and with unprecedented variety, it is also increasingly being traded as commodity in the new "data economy" for utilization. With regard to data analytics for knowledge discovery, this leads to the question, among various others, of how much data is really necessary and/or sufficient for getting the analytic results that will reasonably satisfy the requirements of an application. In this work-in-progress paper, we address the sampling problem in big data analytics and propose that (1) the problem of sampling the big data for analytics is "hard"–specifically, it is a theoretically intractable problem when formal measures are incorporated into performance evaluation; therefore, (2) heuristic, rather than algorithmic, methods are necessarily needed in data sampling, and a plausible heuristic method is proposed (3) a measure of dataset quality is proposed to facilitate the evaluation of the worthiness of datasets with respect to model building and knowledge discovery in big data analytics.

1 INTRODUCTION

As the Internet of Things (IoT) connects an ever increasing variety of devices, sensors, and other physical objects to the Internet, the big data will only get bigger and usher in the era of the "data economy", where businesses and organizations generate, buy, and sell data much like commodity (Datanami, 2016).

For many applications in data analytics, such as business intelligence, the "value" of a dataset is directly proportional to its volume. However, for applications such as knowledge mining or scientific discovery, where the results of data analytics needs to be deeper than merely descriptive (e.g. finding frequent items sets in shopping baskets or correlation among features in patient datasets), the value of a dataset may not necessarily be proportional to its volume. This is because to extract the heretofore unknown knowledge from the available data (say, by first building a learning machine model, then validate it), the available data must be of sufficiently good quality to allow model building without resulting in overfitting or underfitting, etc., and the quality of the data has little to do with its quantity. Worse, there are problems where data is plentiful and clearly provides hope for knowledge discovery that will lead to new insights or solutions, but the quality of the data is insufficient. For examples: important features may have been overlooked and not measured and collected at all, such as in a patient dataset; or the features are not even directly available, such as in the cybersecurity problem of detecting steganography or the multimedia forensics problem of detecting image tampering (Liu et al., 2015); for another perspective, see (Wiederhold, 2016).

This leads to two questions in dealing with the big data analytics for knowledge discovery: 1. how to sample a given dataset to facilitate good model building for knowledge discovery; 2. how to evaluate the quality of datasets with respect to their potential for sound knowledge discovery?

The second question is interesting and very challenging: in view of the rapidly evolving landscape of the data economy, an effective and practicable measure of data quality will be extremely helpful; however, the question seems too application dependent to allow universally applicable measures to be developed such as the common statistics-based formulas, e.g. sample mean vs. population mean, confidence intervals, etc. (Hastie et al., 2009).

Sung, A., Ribeiro, B. and Liu, Q.
Sampling and Evaluating the Big Data for Knowledge Discovery.
DOI: 10.5220/0005932703780382
In *Proceedings of the International Conference on Internet of Things and Big Data (IoTBD 2016)*, pages 378-382
ISBN: 978-989-758-183-0

This position paper addresses the first question, namely how to sample a dataset for performing analytics in knowledge discovery. It is noted that, in addition to sampling, feature selection is also an important task in preprocessing datasets for analytics (and it likewise contributes to reducing the volume by eliminating useless features), and many feature selection methods have been proposed and studied experimentally to evaluate their performance, e.g. (Guyon and Elisseeff, 2003) (Liu et al., 2006). In contrast, there are relatively fewer studies on sampling, even though the problem is equally–if not more–important, especially for data analytics tasks in the current era of big data.

We show that the sampling problem is intractable in a formal sense, thereby necessitates heuristic methods for solution. We then propose a practical method for sampling as a promising method deserving experimental study for further refinement. Finally, based on our ongoing study of the critical sampling problem, and in conjunction with our previous study on the concomitant problem of feature selection, a metric for dataset quality–with respect to its capacity for knowledge discovery through data mining–is proposed as a first step in developing more practically useful metrics for dataset quality.

2 CRITICAL SAMPLING

In this section we outline a proof that the problem of selecting an optimal sample for model building using learning machines is intractable. To analyze the problem in a formal setting, assume a given dataset is represented as an n by p matrix $D_{n,p}$ with n points or patterns, each represented as a p-dimensional vector. In other words, the relevant datasets from the big data have been selected and fully integrated into a uniform format before the data analytics task commences, this assumption may be simplistic but makes the proof easier without losing generality.

The concept of the *Critical Sampling Size* of a dataset with n points is that there may exist, with respect to a specific learning machine M and a given performance threshold T, a unique number $\nu \leq n$ such that the performance of M exceeds T when some suitable sample of ν data points is used; further, the performance of M is always below T when any sample with less than ν data points is used. Thus, ν is the critical (or absolute minimal) number of data points required in any sample to ensure that the performance of M meets the given performance threshold T.

Formally, for dataset D_n with n points (the number of features in the dataset, p, is considered fixed here when we are concerned only with sampling, and therefore dropped as a subscript of the data matrix $D_{n,p}$), ν (an integer between 1 and n) is called the *T-Critical Sampling Size* of (D_n, M) if the following two conditions hold:

1. There exists D_ν, a ν-point sampling of D_n (i.e., D_ν contains ν of the n vectors in D_n) which lets M achieve a performance of at least T, i.e., $(\exists D_\nu \subset D_n)\ [P_M(D_\nu) \geq T]$, where $P_M(D_\nu)$ denotes the performance of M on dataset D_ν.
2. For all $j < \nu$, a j-point sampling of D_n fails to let M achieve performance of at least T, i.e., $(\forall D_j \subset D_n)\ [j < \nu \Rightarrow P_M(D_j) < T]$

In the above, the specific meaning of $P_M(D_\nu)$, the performance of machine (or algorithm) M on sample D_ν, is left to be defined by the user to reflect a consistent setup of the data analytic (e.g. data mining) task and the associated performance measure. For examples, the setup may be to train the machine M with D_ν and define $P_M(D_\nu)$ as the overall testing accuracy of M on a fixed test set which is distinct from D_ν; alternatively, the setup may be to use D_ν as training set and use $(D_n - D_\nu)$ or a subset of it as the testing set. The value of threshold T, which is to be specified by the user as well, represents a reasonable performance requirement or expectation of the specific learning machine that is used for the data analytic task.

To determine whether a critical sampling size exists, for a D_n and M combination, is a very difficult problem. Precisely, the problem of deciding, given D_n, T, k $(1 < k \leq n)$, and a fixed M, whether k is the T-critical sampling size of (D_n, M) belongs to the class $D^P = \{L_1 \cap L_2 \mid L_1 \in NP, L_2 \in coNP\}$, where it is assumed that the given machine M runs in polynomial time (in n). In fact, it can be shown to be D^P-hard, as presented next.

2.1 Critical Sampling Problem is Hard

The problem is first restated formally: Let CSSP (the Critical Sampling Size Problem) be the problem of deciding if a given k is the T-critical sampling size of a given dataset D_n when a learning machine M is used. Then we show that the problem belongs to the class D^P under the assumption that, for any $D_i \subset D_n$, whether $P_M(D_i) \geq T$ can be decided in polynomial (in n) time, i.e., the machine M can "process" D_i and has its performance measured against T in polynomial time. Otherwise, the problem may belong to some possibly larger complexity class, e.g., Δ^P_2. Note here

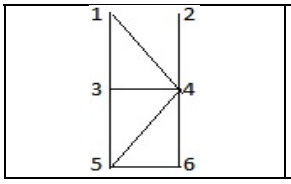
that NP ⊆ (NP ∪ coNP) ⊆ D^P ⊆ Δ^P_2 in the polynomial hierarchy of complexity classes (Garey and Johnson, 1979).

To prove that the CSSP is a D^P-hard problem, we take a known D^P-complete problem and transform it into the CSSP. We begin by considering the maximal independent set problem. In graph theory, a Maximal Independent Set (MIS) is an independent set that is not a subset of any other independent set; so the size of an MIS of the graph is between 1 and n, the number of nodes; also, a graph may have multiple MIS's.

EXACT-MIS Problem (EMIS) – Given a graph with n nodes, and $k \leq n$, decide if there is a maximal independent set of size exactly k in the graph. This is a problem proved to be D^P-complete (Papadimitriou and Yannakakis 1984). Now we describe how to transform the EMIS problem to the CSSP.

Given an instance of EMIS (a graph G with n nodes, and integer $k \leq n$), construct an instance of the CSSP such that the answer to the given instance of EMIS is Yes iff the answer to the constructed instance of CSSP is Yes, as follows: let dataset D_n represent the given graph G with n nodes (e.g., D_n is made to contain n data points, each with n features, representing the symmetric adjacency matrix of G); let T be the value "T" from the binary range {T, F}; let $\nu = k$ be the value in the given instance of EMIS; and let M, the learning machine, be simply an algorithm that decides if the dataset represents a graph containing an MIS of size exactly ν, if yes P_M = "T", otherwise P_M = "F"; then a given instance of the D^P-complete EMIS problem is transformed into an instance of the CSSP.

To explain the transformation in the proof, three examples are shown for the given instance of EMIS in Figure 1.



0	0	1	1	0	0
0	0	0	1	0	0
1	0	0	1	1	0
1	1	1	0	1	1
0	0	1	1	0	1
0	0	0	1	1	0

Figure 1: A 6-node graph with multiple MIS of 3 nodes: {1,2,5}, {1,2,6}, {2,3,6}.

Example 1: k=3. Threshold T = "T" from the binary range {T, F} to mean true, $\nu = 3$, and an MIS of size 3 exists in D_6 as highlighted in the adjacency matrix of G above. So, algorithm M that decides if the dataset D_6 contains an MIS of size exactly 3 (or M "verifies" that some D_3 corresponds to a MIS of size 3) succeeds; i.e., $P_M(D_3)$ = "T" for some D_3. Since the solution to the instance of EMIS problem is Yes, solution to the constructed instance of the CSSP is also yes, as required for a correct transformation.

Example 2: k=4. The constructed instance of CSSP has T = "T" and $\nu = 4$. From D_6 it can be seen that there does not exist any independent sets of size 4, so no exact MIS of size 4 exists. Let M be an algorithm that decides if the dataset D_6 represents a graph containing a maximal independent set of size 4. In this instance M fails to find an exact MIS of size 4 and thus P_M = "F", i.e., $P_M(D_4)$ = "F" for all possible D_4. So the solution to the constructed instance of CSSP is No, as is the solution to the given instance of EMIS.

Example 3: k=2. The constructed instance of CSSP has T = "T" and $\nu = 2$. Independent sets of size 2 exist but they are not MIS's, so algorithm M that decides that some $D_2 \subset D_6$ correspond to an MIS of size exactly 2 fails. The solution to the constructed instance of CSSP is No, as is the solution to the given instance of EMIS, as required.

The D^P-hardness of the Critical Sampling Size Problem indicates that it is both NP-hard and coNP-hard; therefore, it's most likely to be intractable (that is, unless P = NP).

In mining a big dataset $D_{n,p}$ the data analyst is naturally interested in obtaining $D_{\nu,\mu}$ (a ν-point sampling with μ selected features, and hopefully $\nu << n$ and $\mu << p$) to achieve high accuracy in model building or knowledge extraction. From the above analysis of the CSSP and the analysis of the Critical Feature Dimension Problem in (Suryakumar, Sung, and Liu, 2014), this is clearly a highly intractable problem and therefore calls for heuristic solutions.

Discussions: the proof outlined above is rather general and its validity depends on a loose definition of "learning machines". The conclusion that determining the critical sampling size for a given learning machine with respect to a given performance criteria is highly intractable, however, is hardly surprising, as many other optimality problems pertaining to machine learning have been proved to be intractable.

2.2 Heuristic Method for Finding Critical Sampling

Due to the symmetry or similarity of the CSSP and CFDP problems and the convincing examples of using simple heuristic methods to solve the CFDP problem (Suryakumar et al., 2013), we are naturally led to believe that simple heuristic methods can be developed to be sufficiently useful for practical purposes in solving the CSSP problem. Therefore,

proposed in the following is a heuristic method for finding a critical sampling:

1. Apply a clustering algorithm like *k*-means to partition D_n into *k* clusters. (The choice of *k* may be determined, for example, by the number of classes that the learning machine is intended to be trained for performing classification.)
2. Select, say randomly, *m* points from each cluster to form a sampling with $m \cdot k$ points. (The value *m* is set to be fairly small.)
3. Supplement the sampling with additional $d \cdot k$ (for some *d*) points, selected randomly from the whole dataset D_n, to form a sampling *D*.
4. Apply *M* (learning machine, analytic algorithm, etc.) on the sample, then measure performance $P_M(D)$.
5. If $P_M(D) \geq T$, then *D* is a critical sampling, and its size ν is the critical sampling size for (D_n, M). Otherwise enlarge *D* by randomly select another *m* points from each cluster, then supplement with additional, randomly selected points from the entire dataset, and repeat the above procedure until a critical sampling is found, or until the whole D_n is exhausted and procedure fails to find ν.

The values of the parameters k and *m* are to be decided in consideration of the size and nature of the dataset, the specific data analytic problem or task being undertaken, and the amount of resource available. As usual in all data analytic problems, prior knowledge and domain expertise are always helpful in designing the experimental setup. Likewise, whether the random sampling is done with or without replacement is a decision to be made according to the dataset and the problem. Also, experiments may need to be performed repeatedly and adaptively (with regard to k and *m*) to obtain good results.

The authors are conducting experiments on many large datasets that are publicly available to observe if the "critical sampling size" indeed exists for most datasets, and if so whether it is generally much smaller than the size of the whole dataset.

3 DATASET QUALITY METRICS

As more and more networked physical objects become components of the Internet of Things and techniques of big data analytics advance (NRC, 2013), the society is experiencing the transformation into a "data economy" where individuals, businesses, and organizations alike are contributing to both the generation and consumption of the big data. Further, various types of data may be traded in the market place much like commodity.

There are many techniques of data mining for different purposes; an ultimate goal of data analytics, however, is knowledge discovery for problems that have thus far defied solutions through conventional methods of study. Many believe that in the big data one can find the answers to complex problems, if only there is sufficient amount of data (that contains sufficient amount of knowledge) and that proper analytic techniques are applied. This belief, in fact, is one of the driving forces of the big data phenomenon: the exponential growth of data in nearly all areas of human activity and interest.

In view of the above, a quantitative measure for the quality of datasets would be very useful for all concerned in big data analytics.

With regard to the capacity or potential for knowledge discovery or knowledge extraction from a dataset, we propose the following quality metrics for a given dataset $D_{n,p}$ (the dataset is represented as a matrix with *n* points, each represented as a *p*-dimensional vector).

From the concept of the critical sampling size discussed in Section 2, the *Sample Quality* of $D_{n,p}$ is defined as

$$Q_s = \nu / n$$

where ν is the critical sampling size of $D_{n,p}$. So, $\nu \leq n$; and $\nu = 0$, by definition, if the critical sample does not exist (or, equivalently, heuristic methods for finding ν fail), in which case $Q_s = 0$ as well. In the optimal case, $\nu = n$ and so $Q_s = 1$, indicating that all data in the dataset are essential for the data mining task for knowledge discovery.

Likewise, the *Feature Quality* of $D_{n,p}$ is defined as

$$Q_f = \mu / p$$

where μ is the critical feature dimension of $D_{n,p}$ (Suryakumar, Sung, and Liu, 2014). So, $\mu \leq p$; and $\mu = 0$, by definition, if no critical feature sets exist, in which case $Q_f = 0$ as well, indicating that the dataset is insufficient due to the lack of certain features that are necessary. In the optimal case, $Q_f = 1$ when $\mu = p$, indicating that all the features are essential for the data mining for knowledge discovery task.

Q_D, the O*verall Quality* of the dataset $D_{n,p}$ (with respect to a specific learning machine *M* that is applied in mining it for model building and knowledge discovery), can therefore be defined as

$$Q_D = Q_s \times Q_f = \frac{\nu \times \mu}{n \times p}$$

Note that $0 \leq Q_D \leq 1$. $Q_D = 0$ when either the critical feature set or critical sample does not exist (or cannot

be found experimentally, with the respective heuristic methods employed to find them), indicating that the dataset is inadequate for the purpose of model building to achieve acceptable performance. At the other extreme, $Q_D = 1$ when $\nu = n$ and $\mu = p$, indicating that the dataset $D_{n,p}$ is indeed optimal, in terms of both the number of features and the number of data points, when it is evaluated with respect to the data mining task of model building and knowledge discovery for the problem under study.

4 CONCLUDING REMARKS

This position paper addresses the dataset sampling problem in model building for knowledge discovery. The issue of data mining and association rule extraction, etc. from small samples of large datasets have been studied by many authors before (Domingo et al., 2002) (Provost et al., 1999), and sampling techniques have been studied extensively. However, the problem of the critical sampling size of a dataset has not been studied thoroughly. It is shown in this paper that the problem of searching for an optimal sampling is, unsurprisingly, intractable, as in many other optimality problems encountered in machine learning and data mining. Inspired by previous successful results of using heuristic methods to find optimal feature set, we similarly propose a heuristic method for finding an optimal sample of a dataset.

Finally, a quality metric for a dataset, which is computed experimentally by finding a critical feature set and a critical sample, is proposed. This measure indicates the percentage of data in the dataset that is essential for model building in knowledge discovery to meet performance requirements. A low value (close to 0) indicates that the dataset contains much "unimportant" data; an exact zero (0) value indicates that the dataset is in fact inadequate for model building; while a high value (close to 1) indicates that the dataset is "knowledge-intensive" and contains very little unimportant data.

It is our position that the proposed quality metric–perhaps combined with some of the statistics based "static" measures (i.e. those computed from the dataset alone without having to carry out experiments using learning machines)–holds great promise to be refined into practically useful quality measures for datasets, which will be very helpful to the big data industry in the emerging "data economy".

The authors' ongoing study includes investigating the possible interrelation between critical feature dimension and critical sampling, as well as conducting experiments on different datasets for proof of concept.

REFERENCES

Datanami, 2016. *Finding Your Way in the New Data Economy* (by A. Woodie).
http://www.datanami.com/2016/01/25/finding-your-way-in-the-new-data-economy/

Liu, Q., Sung, A.H., Chen, Z. and Chen, L., 2015. *Exploring Image Tampering with the Same Quantization Matrix*, in *Multimedia Data Mining and Analytics–Disruptive Innovation* (Editors: Baughman, A.K., Gao J., Pan J-Y., and Petrushin V.) Springer, pp.327-343.

Wiederhold, G., 2016. *Unbalanced Data Leads to Obsolete Economic Advice*, Communications of the ACM, Vol. 59 No. 1, pp.45-46.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The Elements of Statistical Learning, Data Mining, Inference, and Prediction*. 2nd Edition, Springer, 2009.

Guyon, I., Elisseeff, A., 2003. *An Introduction to Variable and Feature Selection*, Journal of Machine Learning Research, Vol 3, pp.1157-1182.

Liu, Q., Sung, A.H., Xu, J., Liu, J. and Chen, Z., 2006. *Microarray Gene Expression Classification based on Supervised Learning and Similarity Measures*. Proceedings of 2006 IEEE International Conference on Systems, Man, and Cybernetics, Vol. 6, pp.5094-5099.

Garey, M.R., Johnson, D.S., 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W. H. Freeman and Compnay.

Suryakumar, D., Sung, A.H. and Liu, Q., 2014. *The Critical Dimension Problem: No Compromise Feature Selection*. Proceedings of eKNOW 2014, The Sixth International Conference on Information, Process, and Knowledge Management, pp.145-151.

Domingo, C., Gavaldà, R. and Watanabe, O., 2002. *Adaptive sampling methods for scaling up knowledge discovery algorithms*, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Vol. 6 No. 2, pp.131-152, 2002.

National Research Council, 2013. *Frontiers in Massive Data Analysis*, The National Academies Press.

Papadimitriou, C.H.,Yannakakis, M., 1984. *The complexity of facets (and some facets of complexity)*, Journal of Computer and System Sciences 28:244-259.

Provost, F., Jensen, D. and Oates, T., 1999. *Efficient Progressive Sampling*. Proceeding of the Fifth International Conference on Knowledge Discovery and Data Mining, ACM KDD-99, pp.23-32.