Finding Critical Samples for Mining Big Data

Andrew H. Sung[1], Bernardete M. Ribeiro[2], Qingzhong Liu[3], and Divya Suryakumar[4]
[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, USA
[4]Apple, Inc., Sunnyvale, California, USA

Abstract – *To ensure success of big data analytics, effective data mining methods are essential; and in mining big data two of the most important problems are sampling and feature selection. Proper sampling combined with good feature selection can contribute to significant reductions of the datasets while obtaining satisfactory results in model building or knowledge discovery. The critical sampling size problem concerns whether, for a given dataset, there is a minimum number of data points that must be included in any sampling for a learning machine to achieve satisfactory performance. In this paper, the critical sampling problem is analyzed and shown to be intractable–in fact, its theoretical formulation and proof of intractability immediately follow that of the previously studied critical feature dimension problem. Next, heuristic methods for finding critical sampling of datasets are proposed, as it is expected that heuristic methods will be practically useful for sampling in big data analytic tasks*.

Keywords: Sampling, Mining Big Data, Machine Learning

1 Introduction

One of the many challenges of big data analytics is how to reduce the size of datasets without losing useful information contained therein. Many datasets that have been or are being constructed for intended data mining purposes, without sufficient prior knowledge about what is to be specifically explored or derived from the data and how to do it, likely have included measurable attributes that are actually insignificant or irrelevant, which results in large numbers of useless attributes (or features) that can be deleted to greatly reduce the size of datasets without any negative consequences in data analytics or data mining [1]. Likewise, many of these massive datasets conceivably already contain much more data points (or samples, vectors, patterns, observations, etc.) than necessary for knowledge discovery (model building, hypothesis validation, etc.), leading to the questions of what sampling size is sufficient (in, say, machine learning tasks) and how to generate the sample (or training dataset) to ensure successful data analytic results.

For dimension reduction, effective feature ranking and selection algorithms [2] can be utilized to reduce the size of the dataset by eliminating features that are insignificant, irrelevant, or useless. The authors have recently studied the feature dimension problem in general settings by consider the question: Given a dataset with *p* features, is there a *Critical Feature Dimension* (or the smallest number of features that are necessary) that is required for, say, a particular data mining or machine learning process, to satisfy a minimal performance threshold? That is, any machine learning, statistical analysis, or data mining, etc. tasks performed on the dataset must include at least a number of features no less than the critical feature dimension – or it would not be possible to obtain acceptable results. This is a useful question to investigate since feature selection methods generally provide no guidance on the number of features to include for a particular task; moreover, for many poorly understood complex problems to which big data brings hope of scientific breakthrough there is little prior knowledge which may be otherwise relied upon in determining this number (of critical feature dimension).

Similarly, the question about sampling size can be raised: Given a dataset with *n* points, is there a *Critical Sampling Size* (or the smallest number of data points) that is required for any particular data mining (or machine learning, etc.) process to satisfy a minimal performance requirement? This is also an important and practical question to consider since various sampling techniques provide no clue with regard to the critical sampling size for any specific dataset. When dealing with big data where the number of data points (the value of *n*) is huge, the question becomes more relevant.

In previous papers by these authors, the critical feature dimension was shown to be intractable; and yet a simple heuristic method based on feature algorithms was demonstrated to be able to find approximate critical dimensions for many datasets of various sizes, and therefore provides a practically useful solution to the problem.

This position paper shows that the critical sampling size problem, formulated in general, has the same complexity as the critical feature dimension problem. In fact, the same proof of the complexity of the critical feature dimension problem carries over to the critical sampling size problem.

In section 2, the critical sampling size problem is formulated in general terms and shown to be intractable. In section 3, a simple ad-hoc method is proposed as a first attempt to approximately solve the problem, and some discussions conclude the paper.

2 Critical Sampling Size

Assume the dataset is represented as the typical n by p matrix $D_{n,p}$ with n objects (or data points, objects, patterns, etc.) and p features (or measurements, attributes, etc.) The intuitive concept of the critical sampling size of a dataset with n points is that there may exist, with respect to a specific “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 in a sample that is required to ensure that the performance of M meets the given threshold T.

Formally, for dataset D_n with n points (the number of features in the dataset, p, is considered fixed here 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 to achieve a performance of at least T, i.e., $(\exists D_\nu \subset D_n)$ $[\mathrm{P}_M(D_\nu) \geq T]$, where $\mathrm{P}_M(D_\nu)$ 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., $(\forall D_j \subset D_n)$ $[j < \nu \Rightarrow \mathrm{P}_M(D_j) < T]$

Note that in the above, the specific meaning of $\mathrm{P}_M(D_\nu)$, 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 an associated performance measure. For examples, the setup may be to train the machine M with D_ν and define $\mathrm{P}_M(D_\nu)$ as the overall testing accuracy of M on a fixed test set distinct from D_ν, or the setup may be to use D_ν as training set and use $D_n - D_\nu$ as testing set. The value of threshold T, which is to be specified by the user as well, may represent a reasonable performance requirement or expectation.

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 $D^P = \{ L_1 \cap L_2 \mid L_1 \in \mathrm{NP}, L_2 \in \mathrm{coNP}\}$ [3], where it is assumed that the given machine M runs in polynomial time (in n). In fact, it is shown in the following that the problem is D^P-hard.

Since NP and coNP are subclasses of D^P (Note that D^P is not the same as NP ∩ coNP), the D^P-hardness of the Critical Sampling Size Problem (CSSP) indicates that it is both NP-hard and coNP-hard, and thus most likely to be intractable [3,4].

2.1 Proof CSSP is Hard

CSSP: The problem of deciding if a given k is the T-critical sampling size of a given dataset D_n belongs to the class D^P under the assumption that, for any $D_i \subset D_n$, whether $\mathrm{P}_M(D_i) \geq T$ 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 larger complexity class, e.g., $\Delta^p{}_2$. Note here that $(\mathrm{NP} \cup \mathrm{coNP}) \subseteq D^P \subseteq \Delta^p{}_2$ in the polynomial hierarchy of complexity classes [4].

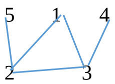
To prove that the CSSP is a D^P-hard problem, we take a known D^P-complete problem and transform it into the CSSP. We begin by considering the maximal independent set problem. In graph theory, a Maximal Independent Set (MIS) is an independent set that is not a subset of any other independent set; a graph 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 is a problem which is D^P-complete [3]. 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 be an algorithm that decides if the dataset represents a 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.

2.2 Explanation of Proof

Consider the 5-node graph given below, with its adjacency matrix:



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

This represents a graph with exactly one MIS of size 3, which is {1,4,5}, correspond to the shaded rows.

Example 1: k=3. Threshold T = “T“ from the binary range {T, F} to mean true, $\nu = 3$, and an exact MIS of size 3 exists in D_5 as highlighted in the adjacency matrix of G above. So, algorithm M that decides if the dataset D_5 contains a 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_5 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_5 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 $D_2 \subset D_5$ 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 CFDP and CSSP, this is clearly a highly intractable problem and therefore calls for heuristic solutions.

3 Heuristic Methods for CSSP

The authors of this paper have previously studied heuristic methods for solving the critical feature dimension problem –due to its theoretical intractability, heuristic methods for approximate solutions are clearly called for [5]. Among the findings of the large number of experiments on datasets:

- Simple methods (such as eliminating one feature at a time) produced successful results in finding a critical number of features that is necessary to ensure performance of M exceeds a threshold. The heuristic method used in [5] works in conjunction with a feature ranking algorithm and purports to identify the critical features.
- The critical feature dimension, as determined experimentally by the heuristic method, is in fact different from–but hopefully close to–the formally defined critical feature dimension.
- For datasets with large numbers of features, their critical feature dimension may be much smaller than the total number of features, as shown in Figure 1.
- Many datasets, of various sizes, exhibit the phenomenon of having a critical feature dimension.
- If the critical feature dimension indeed exists for a dataset, then the performance of M is largely preserved when only the critical features are used, as shown in Figure 2.
- The feature ranking algorithm employed in the heuristic method has more significant influence (than the learning machine) on the value of the critical feature dimension.

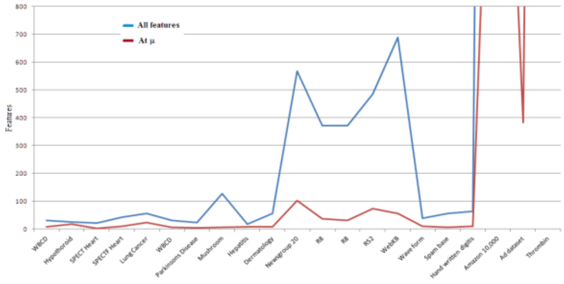


Figure 1. Reduction in feature size at the critical dimension

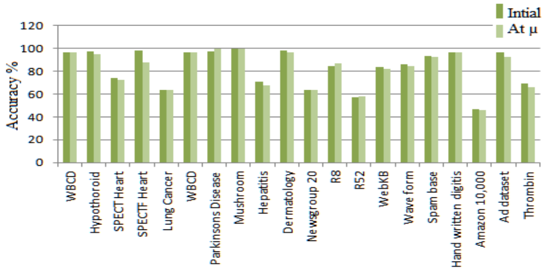


Figure 2. Accuracies with all features, and with critical features selected by the heuristic method

As the simple heuristic method is computationally feasible and appears to be quite sufficient (for many of the datasets studied in the experiments) in finding the critical feature dimension despite the problem's intractability, hope is raised that heuristic methods can be designed to approximately solve the critical sampling size problem satisfactorily as well. Proposed in the following as our position on the CSSP problem is such a heuristic method:

1. Apply a clustering algorithm such as k-means to partition D_n into k clusters.
2. Select, say randomly, m points from each cluster to form a sampling D with m·k points.
3. Apply M (learning machine, analytic algorithm, etc.) on the sample, then measure performance $P_M(D)$.
4. 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, and repeat until a critical sampling is found, or 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 be performed repeatedly and adaptively (with regard to k and m) to obtain good results.

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

4 Conclusions

The issue of data mining and association rule extraction, etc. from small samples of large datasets have been studied by many authors before [6,7,8,9], and formal sampling techniques have been studied extensively in e.g. [10]. However, the problem of the critical sampling size of a dataset has not been studied previously. Not surprisingly, a complexity analysis of the problem, in its most general formulation, shows that it is highly intractable (in the sense of being both NP-hard and coNP-hard), thus defying any attempt for exact solutions and calling for heuristic methods for approximation.

Encouraged by the success of simple heuristic methods in finding critical feature dimensions of datasets with large numbers of features [11], a heuristic method is proposed in this paper for finding the critical sampling size of large datasets, and experiments are underway to validate the concept. Even though simple enough, the heuristic method–if it turns out to be successful like the simple heuristic method for finding critical feature dimension–can serve to provide a practical solution for sampling in data mining, which should be highly useful in coping with some of the challenges of big data [12].

We conclude with this statement of our position: Under formally defined conditions of optimality, both the feature selection problem and the sampling problem easily become intractable; however, simple and practically useful heuristic solutions can often be developed to deal with the feature selection and sampling size problems in data mining.

5 References

[1] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features", Ninth National Conference on Artificial Intelligence, MIT Press, pp.547-552, 1991.

[2] A. Blum and P. Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, Vol. 97, 1997.

[3] C. H. Papadimitriou and M. Yannakakis, "The complexity of facets (and some facets of complexity)", Journal of Computer and System Sciences Vol. 28 No. 2, pp.244-259, 1984.

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

[5] Q. Liu, B. M. Ribeiro, A. H. Sung and D. Suryakumar, "Mining the big data: the critical feature dimension problem", Proceedings of 2nd International Conference on Smart Computing and Artificial Intelligence (ICSCAI 2014), August 2014.

[6] J. Kivinen and H. Mannila, "The power of sampling in knowledge discovery", Proceedings of PODS '94, the 13th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pp.77-85, 1994.

[7] H. Toivonen, "Sampling large databases for association rules", Proceedings of VLDB'96, 22th International Conference on Very Large Data Bases, pp.134-145, 1996.

[8] C. Goh, M. Tsukamoto and S. Nishio, "Fast methods with magic sampling for knowledge discovery in deductive databases with large deduction results", Proceedings of ER'98, the Workshops on Data Warehousing and Data Mining: Advances in Database Technologies, pp.14-28, 1999.

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

[10] J. S. Vitter, "Random sampling with a reservoir", ACM Transactions on Mathematical Software, Vol. 11 No. 1, pp.37-57, 1985.

[11] D. Suryakumar, "The Critical Dimension Problem – No Compromise Feature Selection", Ph.D. Dissertation, New Mexico Institue of Mining and Technology, 2013.

[12] National Research Council, "Frontiers in Massive Data Analysis", The National Academies Press, 2013.