Rademacher Averages Bounded Progressive Sampling Algorithm for Big Data Analytics: A Novel Approach

Yathish Aradhya B C[1]
Assistant Professor
Dept of CSE
KIT Tiptur

Dr.Y.P.Gowramma [2]
Professor
Dept of CSE
KIT Tiptur

Abstract— Sampling of Big Data for its analytics is a tedious task. Progressive Sampling Algorithm(PSA) is a primary tool adopted elsewhere to produce minimal training data set for learning algorithm used in Big Data Analytics. PSA can be characterized by its underlying operations used such as initial sample size, sampling schedule and stopping criterion. Training data set is a determining factor of traing cost, computational cost and learning model accuracy. Rademacher Averages Bound of Sampling can be used to bound the sampling process. This paper suggests novel ways to underlying operations of PSA and scope for significant reduction of the cardinality of training dataset while retaining the behavior of Learning model's Accuracy within Probably Acceptable Correct(PAC) Framework using Rademacher Averages Bounds.

Index Terms – Progressive Sampling Algorithm (PSA), VC-Dimension, Rademacher Averages,Big Data, Staistical Optima Size Sample,Convergence, Rademacher penalty,Bounds

IJSER

1 INTRODUCTION

Big data is vast in size and multi-vectored. Analytics of such big data is a tedious task and impractical to use whole of such huge data to train a learning algorithm. However learning algorithms can be trained to analyze such data by using a samples set with minimal cardinality derived from the actual big data given in hand. To construct suitable training data, big data will be subjected sampling process. Sampling such data may need multiple scans through the huge database in order to arrive at sample set with minimum cardinality. Progressive Sampling Algorithms aim at producing the sample set of small sizes iteratively and finalize when cardinality of sample set meets sufficiency to converge a learning algorithm used to train for some big data. Each iterations adopts defined scheduling scheme to increase the sample size w. r. t. to the previous. Single scan of database to pick out samples and number of samples picked up at the termination of Progressive Sampling process along with initial size to start with play a vital role in deciding the computational and overall run time of learning algorithm. Additionally ensuring that behavior of trained learning algorithm to be within acceptable error bound depends on whether samples used for training are erroneous or not.VC (Vapnik-Chervoniks) dimension bound was used to find the Ɛ-Approximation of big data using its independent and identical distribution.VC dimension gives optimal results in scenario of worst distribution. Therefore generalization of error bounds based on VC dimension tends to be pessimistic. Hence Data dependent bound such as Rademacher averages is used in this work to derive bounds to Progressive Sampling Algorithm(PSA). Additionally novel approaches to initial sample size selection ,sampling schedules and stopping criterion are suggested along with process flow of PSA in generating adequate number of samples for training data set. Subsequent sections in the paper are organized as follows:

2) Background of The work.
3) Progressive Sampling Algorithm with novel approach inclusions
4) Rademacher Averages and Model Accuracy.
5) Conclusions.

2 . BACKGROUND OF THE WORK

To train a learning algorithm for a Big Data, it incurs huge cost in resources such as time and computation. However any learning algorithm can be trained to a given big data set using a samples of such a huge data. Sample set of minimal size can be derived using Progressive Sampling Algorithm(PSA). A progressive sampling Algorithm with several scheduling schemes and verification was provided by Provest et.al [1,2]. Efficient starting sample size estimation was not addressed in[1]. Since Staring size play vital role in

number of iterations to be carried out to arrive at final sample set, a method to determine efficient sample size was provided by Baohua et al [2]. Baough et al has suggested a method called statistical optimal size sample designing method.The at most number of samples beyond cardinality should not increase happens to be a bound.Vapnik and Chervonenkis has provided a projection based approximations on subsets of large sets . An ε-Approximation of range set of a distribution can be constructed provided VC dimension of the range set is known [3]. VC Dimension of Range set of given domain (Range is subset of Big data) used to sample the big data in randomized fashion [5,6,7,8]. Thus Roindtte provided a useful way to select a subset of a range having certain VC dimension as the cardinality of training samples set[5]. However Divergence properties of the selected cardinality set can be used to select the initial sample set for training the learning algorithm[2]. In addition to that it has been shown that VC dimension cannot offer better bounds when the distribution of the data is worst as a possible.On the other hand Rademacher Bounds provide acceptable bounds even when the distribution is in worst case. Therefore an novel approach to significantly reduce the number of training samples i.e cardinality of the training set is to employ the Rademacher bounds to approximations and statistical optimal size technique based on divergence of the range space of any given domain . Till to this time such an approach is never documented .So it can be regarded that this work can be novel, conceptually in theoretical computational sciences.

3. PROGRESSIVE SAMPLING ALGORITHM WITH NOVEL APPROACH

Progressive sampling algorithm presented in this section incoraporates novel approaches and also contributes estimated improvements in compuatation and runtime .This algorithm is designed to to be bounded by rademacher averages and a criteria to choose intial sample size thus expected to run on for fewer iteration when its predecessor methodoloagies are taken in to consideration.I.I.D of given big data set is used to identify range spaces with less value for rademacher penalty . Rademacher average is then computed for identified range space which happens to be the tight bound on the input to each subsequent itarion of PSA. ε-approximation is computed the given range space.intial sample size is computed using statistical divergence method.Stopping criterian to PSA is no improvements in the model acuracy when convergence of learning algorithm is met ; rademacher avaeargaes boundwould be a otherwise choice.Although Sampling big data is a non polynomial time problem optimal soultion can be generated using progresssive sampling algorithm and data bound such as Rademacher average. progressive sampliong Algorithm preseted in this section is expected to meet a poynomial runtime of the order $O(\varepsilon$ - *approximation+* n**Computation time of Learning Algorithm)*i.e where n is number of itreations which and also remarkable efficent usgae of Spcae .The methodology used in this algorithm also ensures one time scanning of database to find the samples and a prioiri stoppimg criterian An algorithm approach of PSA is given below:

Algorithm: PSA with Novel approach

Input:Big Data set and its Distribution
Output: Rademacher Average Bound and Sampled set.

Step 1: Identical and independent distribution of the big data is used to identify the range spaces.
Step 2: Compute Rademacher Average of any range space $\mathbb{R} \subseteq \mathbb{D}$

$$Argmin\{P_r\{sup_{h\in H} |\varepsilon_P(h)-\varepsilon_n(h)| \geq \varepsilon\} \leq \delta\}$$

Step 3: ε-approximation of range set $\mathbb{R}$ of any Domain $\mathbb{D}$ where $\mathbb{R} \subseteq \mathbb{D}$ within the Rademacher bound is computed.it will be designated as sample size and represented by S
Step 4: Calculate the Statistical optimal size based on divergence of S and it happens to be minimum cardinality training set. It is represented by $\mathbf{S_{optimal.}}$
Step 5: Compute Sampling Schedule $\mathbf{S_{schedule} = \{S_{optimal}, S_{o1}, S_{02}, S_{03}, \ldots\ldots\ldots S_{on}\}}$by geometric sample scheduling.
Step 6 : At first iteration Train the Chosen Learning Algorithm with $\mathbf{S_{optimal.}}$ check the performance for any unseen data .
While (until no more accuracy in the learning model)
Choose the subsequent sample size from $S_{Schedule}$.
Repeat the training of learning algorithm with new sample sets from $S_{Schedule}$ until no more progress is found in learning model or when sample set size exceeds empirical rademacher average bound.
Step 7 : The vary last Sample size happens to be training data set for any given big data set as whole although sample are selected from a Range set.

3.1 Progressive Sampling

Progressive Sampling refers to a process of incrementally selecting the instanaces to construct a training sample set, in order to achive convergence of a learning algorithm. Basically method starts with minimal instanaces and advances to sufficiency set by various mehodoligies which depends on the contexts and criteia.
Definition 1. Sampling procedure for any given da-

taset to train any learning algorithm , $\mathbf{S_{optimal}}$ is the minimal sufficient training set. Models built with smaller sufficient training sets of size than the $\mathbf{S_{optimal}}$ have smaller accuracy while also showing no higher accuracy with larger size training instances. Model accuracy remains stable after a certain sample size.Estimation of $S_{optimal}$ within Empirical Risk Minimization principle such that $S_{optimal}$ and its progressive incremental values will always be within ε-Approximation of range spaces.

Compute Sample Schedule **S={** $\mathbf{S_{optimal}}$, **S1,S2, ...$\mathbf{S_n}$ }** of Sample Sizes
N ← $\mathbf{S_{optimal}}$
M ← Model trained by N instances
While not met Convergence
Compute S again
N ← Next element of S of size > N in the Previous iteration.
M ← Model trained By **N**
End While
Return M

Fig2. Generic model of progressive sampling

Figure 2 is a generic algorithm that represents basic mechanism of progressive sampling Algorithm. Contributions to basic methodology of progressive sampling incorporate a series of procedures to select a schedule for determining convergence and for altering the scheduling adaptively.

3.2 Novel Approach to Progressive Sampling.

The Novel methodology suggested in this work is calculation of Rademchaer Averages bound of any given large data set. Rademacher bound defines upper limit on the input size of training data set and ensures the trained model to be within Probably Approximate Correct Framework(PAC).
Rademacher Averages bound is significantly is more confident than VC dimension when worst case distribution is considered. Rademacher averages are computed for highly dense range spaces of any given large data set. Statiscal optimal sample size based on information divergence is used to estimate the initial sample size $S_{(optimal)}$ .Empirical Rademacher Average will be the stopping criteria when expected convergence is not met , otherwise it serves as a tight bound on the sample complexity. This Empirical Rademacher average as a stopping criteria and statistical optimal sample size for the initial sample size selection can be accounted as a novel approach in progressive sampling techniques. Statistical optimal sample size reduces the unnecessary iteration of progressive sampling and estimated Sample size can be obtained in one scan of range space. The approach to find statistical optimal size of sample set is given below as pseudo procedure.

3.2.1 Algorithm for finding $s_{(optimal)}$

INPUT : A range space R ∈ D(given large data set)
OUTPUT: (S_i, Q_i)
STEP 1: Apply random sampling method to choose the samplesS_i={S_i|i=1,2,3,.4..5.....Ni=sizeof(Ɛ-Approximation) }
STEP2 : Foreach (S_i∈ R)
calculate r = RandomNum(0.0 ...0.1) for each reading of sample S_i.
Update the corresponding statistics for R
if [r<(S_i/N_i)]
update corresponding statistics for S_i
STEP 3: ForEach Si Calculate Qi and outpu (Si,Qi) pair.
STEP 4 : Plot (Si,Qi) and draw a quality curve . By applying linear regression on every consecutive points covering possibly sufficient points in the range space.
STEP 5 : The Size of the mid point of regressed line will be considerd as $\mathbf{S_{optimal}}$ which will be intial sample size for the progresive sampling algorithm. Since this novel approach is expected to reduces time complexity of Progressive sampling method it is perhaps possible outperform all predecessor methodologies. However choice of learning algorithm and calculation of Q_i accounts to significantly affect over all runtime.

4. Rademacher Averages and Model Accuracy

Within a PAC framework and statistical learning setting, it can be assumed that a learning algorithm chooses its hypothesis from some fixed hypothesis class H .Generalization of error analysis provides theoretical results bounding generalization of error of hypotheses h ∈ H which will be based on the sample and properties of hypothesis class. For any given hypothesis h ,its generalization error is the probability that a randomly drawn example is misclassified;

$$Ɛ_p(h)=P(h(x)\neq y)$$

Although goal of any learning algorithm is to find a hypothesis with a small generalization error, it cannot computed on sample per se , as it would also depend on probability distribution P. However an attempt can be made to approximate generalization error of hypothesis h , by its training error on n examples:

$$Ɛ_n(h)= 1/n \sum^{n}{}_{i=1} L(h(x_i),Z_i)$$

where L is 0/1 loss function

L(z ,z')= { (1,if z≠z') , (0, otherwise) }

Empirical Risk minimization (ERM) is a principle according to which only such hypothesis whose training error is minimal is to be chosen. To ensure that ERM yields hypothesis with small generalization error the bound $\mathbf{sup_{h∈H} | Ɛ_p(h)- Ɛ_n(h)|}$ can be used. The difference of training error of hypothesis h on n examples

and true generalization error converge to 0 in probability as n tends to infinity, provided examples are independent and identical in distribution while the hypothesis class $\mathbb{H}$ is not complex. A Rademacher Random Variable (Koltchinskii 2001) takes values +1 and -1 with probability 1/2 each. Let Rademacher random variables $r_1,r_2,r_3,.....r_n$ be i.i.d. Rademcaher variables independent of the data $(x_1,y_1),.....(x_n,y_n)$. The Rademacher Penalty of hypothesis class $\mathbb{H}$ is defined as follows

$$\mathbb{R}_n(\mathbb{H}) = \sup_{h\in\mathbb{H}} | \ 1/n \ \sum_{i=1}^{n} r_i \ L(h(x_i),y_i) \ |$$

By symmetrization inequality of the theory of empirical processes

$$E\{ \sup_{h\in\mathbb{H}} \ | \ \mathcal{E}_p(h)- \mathcal{E}_n(h)| \ \} \leq 2E\{ \ \mathbb{R}_n(\mathbb{H}) \ \} \qquad ..(1)$$

where expecations are taken over the choices of examples on the left and over the choice of examples and Rademacher random variables on the right.Using standard conecntartion of inequalities that with probability at least 1-δ

$$\mathcal{E}_p(h) \leq \mathcal{E}_n(h)+2 \ \mathbb{R}_n(\mathbb{H})+ \eta(\delta,n) \qquad (2)$$

where $\eta(\delta,n)=5\sqrt{(\ln(2/\delta)/2n)}$

is a small error term that takes care of fluctuations of analyzed random variable around their expectations.The usefulness of the inequality (2) is araised from the fact that its right hand side depends only on the training smaple and not on p directly. this means that $\mathbb{R}_n(\mathbb{H})$ can be compuated with algorithm used for ERM..Rademacher Penalities can be applied to provide approximate solution to progrssive sampling algorithm ,particularly in estimation of stopping time in terms data dependent upper bound (Ɛ-Appoximation) of sample set used for training the learning algorithm(Koltchinskii et al 2000).The minimal number of samples required to gauarantee ERM is within a distance of Ɛ form generalization error of h for every $h\in\mathbb{H}$:

$$Argmin\{P_r\{\sup_{h\in H} |\varepsilon_P(h)-\varepsilon_n(h)|\geq\varepsilon\} \leq\delta\}.$$

The rademacher stopping time $v(\mathcal{E},\delta)$ with parameters $(\mathcal{E},\delta)$ for hypothesis of class $\mathbb{H}$ is

$$v(\mathcal{E},\delta) = \min\{n_i=2^i n_0(\mathcal{E},\delta) \ | \ \mathbb{R}_{ni}(\mathbb{H}) < \mathcal{E}\}.$$

Koltchinskii et al (2000) derived data dependent resultsthat hold for any distribution that could have produced a sample S.

5. CONCLUSION

Sampling Big data is a tedious task. Progressive sampling algorithm bounded by Radmecher averages can be used to derive training data set from the given big data . In this paper we have proposed a combination of statistical optimal size and rademacher averages bound to progressive sampling algorithm. Novel strategies based on statistical divergence in choosing initial sample set for sampling schedule and Rademacher Averages based data dependent Ɛ-Approximation which serves as tight bound for learning process, are proposed. A theoretical analysis of Rademacher averages in computing stopping time for progressive sampling algorithm is presented within the scope of Progressive sampling.PSA with novel approaches presented can governed to be within runtime $\boldsymbol{O}$(Ɛ -approximation+n*Computation time of Learning Algorithm). Since this novel approach is expected to reduces time complexity of Progressive sampling method it is perhaps possible outperform all predecessor methodologies.

6. References

[1] Provost F., Jensen D., Oates T. (2001) Progressive Sampling. In: Liu H., Motoda H. (eds) Instance Selection and Construction for Data Mining. The Springer International Series in Engineering and Computer Science, vol 608. Springer, Boston, MA

[2] Gu B., Liu B., Hu F., Liu H. (2001) Efficiently Determining the Starting Sample Size for Progressive Sampling. In: De Raedt L., Flach P. (eds) Machine Learning: ECML 2001. ECML 2001. Lecture Notes in Computer Science, vol 2167. Springer, Berlin, Heidelberg.

[3] Static and Dynamic sampling by GH John - 1996 Aug 2, 1996 - Publication: KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data MiningAugust 1996.

[4] Aounallah M., Quirion S., Mineau G.W. (2004) Distributed Data Mining vs. Sampling Techniques: A Comparison. In: Tawfik A.Y., Goodwin S.D. (eds) Advances in Artificial Intelligence. Canadian AI 2004. Lecture Notes in Computer Science, vol 3060. Springer, Berlin, Heidelberg.

[5]Machine Learning By Tom M Mitchell published by Tata Mc-Graw Hill ISBN-13: 978-0070428072.

[6] An Overview of Statistical Learning Theory. Vladimir N. Vapnik. Abstract—Statistical learning theory was introduced in the late. 1960's. Until the 1990's it was a by VN Vapnik - 1999 - Cited by 5582 articles.

[7] Sampling-based Randomized Algorithms for Big Data Analytics" by Matteo Riondato, Ph.D., Brown University, May 2014.

[8] Mining frequent item through Progressive Sampling with Rademacher Averages **Publication:**KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data MiningAugust 2015 Pages 1005–1014https://doi.org/10.1145/2783258.2783265.

IJSER