Optimized Sampling Strategy For Big Data Mining Through Stratified Sampling

Kamlesh Kumar Pandey, Diwakar Shukla

Abstract: The revolution of digital and communication technology changes the nature and behavior of data to big data respect to volume, variety, and velocity. When applying any data mining techniques for these types of data, it suffers the various challenges with execution time. Divide and conquer, parallel execution, and sampling techniques give a platform for fast execution in the big data environment. This paper used data reduction techniques as sampling for fast execution and identifying which sampling technique is appropriate for big data mining based on existing research. From a practical examination, and theoretical, and existing research perspective, this paper describes big data and their characteristics, big data mining, sampling techniques, sampling for big data mining and design the optimized sampling strategy algorithm through stratified sampling for big data mining. This optimized algorithm describes the new data mining technique based on stratified sampling. The optimized sampling strategy is explained by using the partitioning based K-means clustering algorithm, which is known as the SSBKM.

Index Terms: Big Data, Big Data Dimension, Big Data Mining, Big Data Sampling, Sampling, Stratified Sampling, Data Reduction, Stratified Sampling based K-means (SSBK).

1. INTRODUCTION

The revolution of digital and communication technology is given to an easier life of human and changed behaviors of data to big data in term of Volume, Velocity, and Variety. Nowadays various sources are available for big data such as social media, internet of things, smart grid, sensor network, cloud computing, medical applications, collective intelligence, data center, retail and enterprise management, and public sector administration. Laney (2001) presents a common framework of big data through Volume, Variety, and Velocity. These three V's describes the base dimensions of constricting the big data. After that IBM (2012) presents Veracity, SAS defined Variability and Oracle represents Value dimensions inside of Laney dimensions. The last V's is Visualization is based on the user view expectation [1]. The rise of 7 V's is defined what is big data and their popularity, difference between traditional data and big data, and limitation of relational database management system [2]. Figure 1 summarized an integrated view of big data dimensions.

The Volume is first V's of big data, which is represented to the huge quantity of data generated through sensors, machines and real-time applications such as social media, internet of things, cloud and mobile computing, and another's machine depend on the device. For example, a jet engine machine-generated 10 terabytes of the data in half an hour. The second V's, Variety is showing a variety of data types and sources. Traditional databases captured only well-defined data formats such as structured, but the big data captured the heterogeneous data as structured, unstructured, and semi-structured data. These data types are not captured by homogenous sources, but it is captured by heterogeneous sources. The third V's, velocity is showing the speed of data creation and mining. For example, the internet of things, and social media generated a huge amount of structured, unstructured, and semi-structured data in very less time. The fourth V's, Veracity has represented the uncertainty of data and definitions objectivity, truthfulness, unreliability, untrustworthiness, and credibility for improving the accuracy of the sources. The fifth V's, Variability represents those data where meaning is continued changes due to time, machine and human interactions. For example, the share market data value is continuously changed due to time. The sixth V's, the value is identifying the valuable data or defining constrained attribute during data mining and data analysis. The last V's Visualization visualizes the mined result as user expectation such as a graph, figure, table and so on [2-4]. In general, big data are dealing with large volumes of structured or unstructured or semi-structured data with heterogeneous sources and high velocity. Big data researcher reduced big data challenges and tried to develop better to better techniques for capturing, processing, analyzing, and visualizing the huge amounts of data within a limited time span in various disciplines such as mathematics and statistics, data mining and machine learning, optimization, social network analysis, signal processing, and visualization [5-6]. Determining the patterns and extracting knowledge is one of the challenging issues inside of the big data [7]. Data sampling is one of the techniques that extract subsets of the entire data set for analyzed or mining [8] and data mining one of the techniques for big data analysis [9]. Data sampling is a good method for utilizing the memory and time, reducing the size of data for mining and analyzing fast processing, and it has capability increases performance with accuracy for any data mining or analyzed algorithms of the large dataset [10-11].

* Kamlesh Kumar Pandey, Research Scholar, Department of Computer Science and Applications, Dr. HariSingh Gour Vishwavidyalaya, Sagar, India.
* Prof. Diwakar Shukla, Dean & HoD, Department of Computer Science and Applications, Dr. HariSingh Gour Vishwavidyalaya, Sagar, India.
The main purpose of sampling is given data reduction schemes for the large scale data sets. A general sampling framework for performing big data mining shows in Figure 2. According to this framework firstly takes a sample of the large
dataset through the sampling techniques and after that applies
data mining techniques in this sample and data mining result
applies to the un-sample data. This concept applies to predict
the better to better results through sample data [12]. The
objective of this paper is identifying the scalable and suitable
sampling techniques for big data mining such as clustering,
classification, and associated mining and proposed new
sampling model for big data mining.

2 RELATED WORKS
Shirkhorshidi et al. distributed the big data clustering
algorithms into two groups as a single machine clustering
multiple machine clustering. Single machine-based clustering
using sampling with dimension reduction and multiple
machines clustering with the parallel and MapReduce
approach [13]. Tsai et al. summarized the sampling method is
very helpful for big data analytics respect to the reduction
method for the computation time and volume [14]. Rezvanian
et al. presented the weighted networks for real data points
and proposed six networks based sampling algorithms and
simulated these algorithms through real and synthetic
weighted networks respect to a relative error, Pearson’s
correlation coefficient, skew divergence, and Kolmogorov
Smirnov statistic [15]. Rojas et al. reviewed the random
sampling and showing the results the combined multiple
sampling techniques for helping to the data scientists respect
to filters out with and without uncertainty outliers, density
and QBC sampling [16]. Hu et al. focuses on the sampling problem
of the big data and gives efficient heuristic Cluster Sampling
Arithmetic (CSA). CSA algorithms give the idea of clustering
through a minimum sample set of big data by the random
sampling method [17]. Kim et al. analyzed the selection of bias
in the big data through two methods. The first methods
describe various sampling techniques and the second method
gave the framework for combining the big data sample [18].
Boicea et al. discussed Hash Mod, Random Sort, and
Systematic, R, Z, and D sampling techniques and their evaluation
due as execution time and accuracy, and test environment in the large set of structured data [7]. Liang et al.
proposed stratified sampling for large-scale data with fuzzy
clustering on the basis of three steps. The first step defines
sensitive hashing technique for obtaining the number of samples from different strata, the second step partitioning the
sample into various clusters through fuzzy c-means and third
steps assigning the un-sample object in a closed cluster [12].

3 SAMPLING STRATEGY FOR BIG DATA
ENVIRONMENT
A good sampling strategy always reduces the data size and
gives the guarantee sample of data must behold, all
characteristics of the sample and gives high-quality mining
results [17]. In order to achieve this goal, this section describes many sampling methods for the big data
environment [7][10][19]

3.1 Uniform Random Sampling or Random Sampling
Random sampling is a sampling design which selected n
distinct data points from the N data points in the large dataset
and selection of every n distinct data points have the same
probability of selection. This sampling technique works only
when every data points are numbered between 1 to N so that
easily find out probability \( \pi_i = \frac{n}{N} \) of ith data points of the N
dataset with the n sample [19]. Some random sampling based
clustering algorithms are CURE (Clustering Using
REpresentatives), CLARANS (Clustering Large Applications
based on Randomized Sampling), RSEFCM (Random
Sampling plus Extension Fuzzy C-Means) [12].

3.2 Progressive Sampling
Progressive sampling starts from a small size sample for data
analysis or other purpose and after calculating the efficiency
and accuracy of result data. When selected sample size not
given appropriate result as efficient and accurate than sample
size continuous increases until desired efficiency and accuracy
are not met. In Progressive sampling describe as schedule S =
\{n0,n1 \ldots \ldots \ldots nk \} where each nk describe the sample size
which size is contiguous increases through an induction
algorithm [20]. Some Progressive sampling based sampling
algorithm is GOFCM (geometric progressive fuzzy c-means)
[12].

3.3 Reservoir Sampling
Reservoir sampling is based on randomized algorithms and
memory efficient, which are taking on sample huge dataset
where the size is unknown. Reservoir sampling takes sample
S >=n , where n is the size of a random sample. This sampling
technique takes first n data for data analysis and the rest of
the data takes sequentially [21]. Some Reservoir sampling
techniques is R, Z, and D [7].

3.4 Stratified Sampling
Stratified sampling is working under two steps first is dividing a
whole data in strata S0,S1,......Sn group and the second step
are doing random sampling in each S strata for the creation of
the sample[12]. Inside of strata relative homogeneous data
object are grouped together according to needed conditions
and after that takes a sample for data analysis through
random sampling [22].

4 PROPOSED OPTIMIZED SAMPLING
STRATEGY FOR BIG DATA MINING
Sampling is given solutions during some problem raised in the
big data analysis, such as “bag of little bootstraps” [23] and
sampling strategy gave a more effective evaluation and
accuracy of datasets [16]. Under the big data mining, good
sampling techniques must be working under Volume with
Variety and/or velocity. This paper dis-cussed under the third
section four sampling strategies for the big data mining based
on various literature reviews which are described in section
two. According to sampling behaviors Uniform random
sampling suitable on volume dimensions, Progressive
sampling suitable on velocity and Volume dimensions,
Reservoir sampling suitable on velocity dimensions and
Stratified sampling suitable on Volume and Variety. Big data is
a collection of heterogeneous data from various sources, and
when applied various data mining techniques for these types of
data, then data mining needs to the similar type or a special
type of data, such as numerical data, categorical data,
continuous data, text data, video data, and so on.
Categorization of the data with sampling is more challenging
issues under big data mining. This section tries to resolve this
issue through Stratified based sampling and proposed
optimized sampling strategy through. Stratified sampling with
Uniform random sampling. In this section paper firstly presents
the optimized sampling Strategy and describes these
algorithms shows in stratification, sample size, and sample
allocation.

4.1 Algorithm of Optimized Sampling Strategy

1. According to the research problem or research condition whole dataset are divided into various groups or categories or L strata through stratified sampling.
2. Identify the sample size n on each group through eq.
3. If sample size n > N * MAX(Wl) then sample size n need to normalized according to n=[N* MAX(Wl)] .
4. Identify the number of data objects from each strata l through proportional allocation and optimal allocation. If the data mining techniques used various attributes for mining than proportional allocation is useful otherwise optimal allocation used. In generally optimal sample allocation technique used to create sample data Xs in each strata l.
5. Used to data mining techniques according to the research problem or re-search condition on these sample data Xs and get partial mine result and this data mining algorithm known as stratified sampling based data mining (algorithm name).
6. Measures and evaluated the partial mined data through sampling and data mining evaluate criteria such as accuracy, use case Set, error, variances, precision, relative recall, speedup ratio, statistical test, and other data mining algorithm related criteria.
7. When the partial mine result fulfills the evaluate criteria, then this partial mine result is applied to un-sample data which is not selected through proportional allocation and optimal allocation and get final mine result.
8. Apply to again data mining evaluate criteria to final mine result if need.

4.2 Stratification

Stratified sampling manages two techniques for big data analysis such as stratification and sample allocation [12][18][19][24]. Stratification dividing heterogeneous data object in homogenous data object according to data analysis condition and sample allocation determine sample size and take data object from each stratum or needed stratum for big data mining. In this section, the paper describes the performance of the proposed optimized sampling Strategy. Any Experimental Analysis of any algorithms depends upon to Experiment environment (tools), dataset, existing algorithm, evaluation criteria, and results. Suppose dataset X is consists of N heterogeneous data object. According to stratification dataset X is dividing the total L strata and every stratum have N homogenous data object. The number of data objects in stratum 1 to L is denoted by N1, N2…… Nl therefore total size of dataset X is N1 + N2 + ....... + Nl. Let μ for dataset mean which is defining averages of homogenous or heterogeneous data objects. µl means say of the ith stratum which is expressed by xil where il is data value in the ith stratum and let Wl for the fraction of the dataset in the ith stratum where Wl = Nl / N.

\[
\mu = \frac{1}{N} \sum_{l=1}^{L} \sum_{i=1}^{N_l} x_{il} = \frac{1}{L} \sum_{l=1}^{L} N_l \mu_l = \frac{1}{N} \sum_{l=1}^{L} W_l \mu_l
\]

(1)

After Stratification applies random sampling for each stratum and takes nl sample size. The mean Xi of the sample data object in lth stratum define as to where by Xil where ilth sample data value in the lth stratum.

\[
\bar{X}_l = \frac{1}{n_l} \sum_{i=1}^{n_l} X_{il}
\]

(2)

4.3 Sample size

The size of the sample for big data mining is depending on data object selection. The size of the sample is formulated as eq3 if the size of sample n is higher than 5% of the whole data set size N then n need to adjust. In this formula, σ2 is denoted as the variance of whole data set X and σ, are denoted as the standard deviation of the data object of the stratum Sp.

\[
n = \frac{(\sum_{l=1}^{L} w_l)\sigma_l^2}{\sigma^2 + \sum_{l=1}^{L} n_l \sigma_l^2}
\]

(3)

4.4 Sample allocation

Determining the sample size from each stratum is related to the variance of the sample object Xl and variance of the sample object X is always belongs to minimization. Variability is one of the dimensions of big data and variance is useful to measure the variability in the form of mean and average. In big data mining, every stratum are independent to each other in the samples taken time for reduce the error and easily calculated to variance and random sampling or stratum mean must be unbiased so stratified estimate of the sample object \( \bar{x} \) is equal to \( \mu \). The variance \( \sigma^2 \) of stratified estimate \( \bar{x} \) is express by \( var(\bar{x}) \) and minimize the variance of the sample object is solved by \( min var(\bar{x}) \).

\[
var(\bar{x}) = \sum_{l=1}^{L} W_l^2 \left( \frac{1}{n_l}(1 - \frac{n_l}{N_l}) \right) \sigma_l^2
\]

(4)

\[
min var(\bar{x}) = \sum_{l=1}^{L} W_l^2 \sigma_l^2
\]

(5)

By using the Lagrangian method for the chosen optimal \( n_l \) sample size for under the optimal sample allocation strategy because High number of data objects inside of strata define heavily sample size with \( W_l \). A large \( W_l \) define a large fraction of object and large \( \sigma_l \) define quite spread of data value inside of l stratum.

\[
n_l = \frac{n_l w_l \sigma_l}{\sum_{l=1}^{L} w_l \sigma_l}
\]

(6)

Optimal sample allocation scheme depend upon the individual variance of each stratum and big data have various attributes and mining techniques may depend upon the various attributes so it is very difficult to find out optimal sample allocation to each attribute. The solution of these problem is used to proportional allocation. The formulation of the dataset mean based on proportional allocation is
Optimal sample allocation and Proportional allocation is also ignoring the fix sized dataset and it is useful for velocity (real time) dimensions of big data. The formulation of optimal sample allocation and proportional allocation under the velocity dimensions with Stratified sampling are as

\[\text{var}(\bar{X}_{sp}) = \frac{(\sum_{i=1}^{n} W_i \sigma_i)^2}{n}\]

(8)

\[\text{var}(\bar{X}_{so}) = \frac{1}{n} \sum_{i=1}^{n} W_i \sigma_i^2\]

(9)

The differences between the variance of proportional allocation and optimal allocation under the velocity dimensions with Stratified sampling are as

\[\text{var}(\bar{X}_{sp}) - \text{var}(\bar{X}_{so}) = \frac{1}{n} \sum_{i=1}^{n} W_i (\sigma_i - \bar{\sigma})^2\]

(10)

Where \(\bar{\sigma} = \frac{1}{n} \sum_{i=1}^{n} W_i \sigma_i\)

When applying uniform random sampling after Stratification process the variance is formulated as

\[\text{var}(\bar{X}) = \frac{\sigma^2}{n}\]

\[\text{var}(\bar{X}) = \frac{1}{n} \frac{1}{\sum_{i=1}^{n} W_i} (X_i - \bar{X})^2\]

\[\text{var}(\bar{X}) = \frac{1}{n} \frac{1}{\sum_{i=1}^{n} W_i} \sigma_i^2 + \frac{1}{n} \frac{1}{\sum_{i=1}^{n} W_i} (\mu_i - \bar{X})^2\]

(11)

The differences between the variance of the mean uniform random sampling and proportional allocation define as eq. 12 with velocity.

\[\text{var}(\bar{X}) - \text{var}(\bar{X}_{sp}) = \frac{1}{n} \frac{1}{\sum_{i=1}^{n} W_i} (\mu_i - \bar{X})^2\]

(12)

This paper observes according to the above formulation, the proportional allocation with stratified sampling every time provides smaller variance as compared to uniform random sampling. If the means of strata is quite variable then the proportional allocation with stratified sampling is highly acceptable as compare to uniform random sampling and if the standard deviations of strata are variable then the optimal allocation is better than proportional allocation under stratification.

5 EXPERIMENTAL ANALYSIS OF OPTIMIZED SAMPLING STRATEGY

In this section, the paper describes the performance of the proposed optimized sampling Strategy. Any Experimental Analysis of any algorithms depends upon to Experiment environment (tools), dataset, existing algorithm, evaluation criteria, and results.

5.1 Experiment Environment (tools) and Dataset

The experiment environment of optimized sampling Strategy is used R computing environment and Anuran Calls (MFCCs) Data Set. The experiment environment is configured with an Intel I3 processor, 320 GB hard disk, 4 GB DDR3 RAM with windows 7 Operating System. Anuran Calls (MFCCs) data set consist of 22 dimensions with 7195 real data points [25].

5.2 Exiting Algorithm for Comparing

The paper [26] [27] describes K-means, BIRCH, CLARA, CURE, CLARANS, DBSCAN, and DENCLUE more satiable under the high dimensions, velocity, and volume data. This paper takes the partitioning based K-Means (KM) algorithm with Euclidean distance for cluster constructions using the optimized sampling strategy because K-Means data mining techniques is known as the top second algorithm [28].

5.3 Evaluation Criteria for Validation

Accuracy and running time is the external evaluation criteria of clustering based research for big data mining. This paper used to cluster algorithm for optimized sampling Strategy so that clustering accuracy is identified the correctly classified unlabeled data points and running time gives the total execution time of sampling based clustering and without sampling based clustering (traditional clustering algorithm). The formulation of the clustering algorithm shows in eq. 13.

\[CA(\%) = \frac{\sum_{i=1}^{K} \max_{j=1}^{N} N_{ij} \times 100}{N}\]

(13)

In the formulation of CA, N is total data object, k (undefined class) and \(k^\prime\) (pre-defined class) the total number of cluster and \(N_{ij}\) the common object.

5.4 Results

According to the algorithm of optimized sampling Strategy, this paper divide the Anuran Calls dataset into three strata based on categorical, numeric and char data type and used to optimal sample allocation scheme according to clustering data mining technique. Here K-Means algorithms used as a data mining technique which is known as Stratified Sampling based K-Means (SSBKM) under optimized sampling Strategy. The result of an optimized sampling Strategy is shows in Table 1 and Table 2 according to evaluation criteria. Table 1 shows the clustering accuracy in and Table 2 shows the average running time between SSBKM and KM.

<table>
<thead>
<tr>
<th>TABLE 1 CLUSTERING ACCURACY OF SSBKM AND KM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
</tr>
<tr>
<td>KM</td>
</tr>
<tr>
<td>SSBKM (overall)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE 2 AVERAGE RUNNING OF SSBKM AND KM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
</tr>
<tr>
<td>KM</td>
</tr>
<tr>
<td>SSBKM (overall)</td>
</tr>
</tbody>
</table>

6 CONCLUSION

This paper present an optimized sampling strategy for big data mining through Stratified sampling and reviewing the big data with their characteristics, big data mining, sampling techniques, and sampling for the big data mining. The first section presents an introduction about big data, characteristics, big data mining, and data reduction technique as sampling for big data. The second section gave the related work of sampling techniques which are related to big data characteristics. This section is identifying how to sampling techniques suitable for big data mining. The third section describes the sampling strategy for big data environments.
such as uniform random sampling, progressive sampling, reservoir sampling, and stratified sampling. This section identifies uniform random sampling is suitable under the volume dimensions, progressive sampling suitable under the velocity and volume dimensions, reservoir sampling suitable under the velocity dimensions and stratified sampling suitable under the volume and variety through existing research perspective and related work section. The fourth section proposed the algorithm of optimized sampling strategy for big data mining through stratified sampling. This algorithm describes how to define Stratification, Sample size, sample allocation and how to validate sampling and data mining algorithms. The fifth section evaluates the proposed algorithms through the R computing environment, Anuran Calls data set, K-means clustering. This section proved the optimized sampling Strategy (SSBKKM) clustering accuracy is higher as compared to the K-mean algorithm but SSBKM takes more time as compared to the K-mean algorithm.

ACKNOWLEDGMENT
I would like to express my very great appreciation to Prof. Diwakar Shukla, Department of Computer Science and Application, Dr. HariSingh Gour Vishwavidyalaya, Sagar for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated.

REFERENCES