January 2020

Distributed multi-bucket sampling

Richard Steven Farnsworth

Distributed multi-bucket sampling

ABSTRACT

In many contexts, e.g., training of machine learning models, there is a requirement to sample a large dataset that includes data points that are classified as either positive or negative. Per techniques of this disclosure, distributed sampling is performed such that the dataset is read just once and a minimum number of each type of data point is captured without changing the ratio between positive and negative data points. This disclosure describes techniques to sample a large dataset using a mapreduce strategy. During mapping, a data point is one-to-one mapped to a partial result that includes counts of positive and negative data points and sets of sampled positive and negative data points. During the reduce phase, two partial results are combined into one in an iterative manner until only one partial result remains, which becomes the final result. Both the map and reduce phases are performed in a distributed manner. It is not necessary that the ratio of the positive and negative data points in the dataset be known in advance.

KEYWORDS

- Distributed sampling
- Multi-bucket sampling
- Statistical sampling
- Single pass sampling
- Machine learning
- Training data
- Mapreduce

BACKGROUND

In many contexts, dataset sampling is performed. For example, machine learning models are often trained on large datasets that include positive and negative training data based on the presence or absence of an attribute such as a specific feature label. For the purposes of training, such large datasets often need to be sampled. Sampling needs to be performed such that the dataset is read just once and a minimum number of each type of data is captured without changing the ratio between positive and negative data points. For example, if ten data points of each type are needed, a dataset with twenty positive and forty negative data points would be sampled to ten positive and twenty negative data points. If a dataset contains less than the minimum number of positive or negative data points, then the dataset is not sampled.

DESCRIPTION

Due to the large size of datasets, it is computationally efficient to read the dataset just once and to perform the sampling in a distributed manner. However, if the dataset is read just once, the ratio between positive and negative data points is not known in advance. In turn, the number of data points of each type to be kept is not known in advance. Techniques described herein address this problem.

In particular, this disclosure describes techniques to sample a large dataset using a mapreduce strategy. During mapping, each data point is one-to-one mapped to a partial result. A partial result, R, contains at least the following data.

- A count of positive data points represented by the partial result, R_{cp}
- A count of negative data points represented by the partial result, R_{cn}
- A set of sampled positive data points, R_{Sp}
- A set of sampled negative data points, R_{Sn}

A positive data point maps to a partial result, *Rp*, such that $Rp_{cp} = 1$, $Rp_{cn} = 0$, Rp_{Sp} comprises the data point associated with a random key, and Rp_{Sn} is empty. A negative data point maps to a partial result, *Rn*, such that $Rn_{cp} = 0$, $Rn_{cn} = 1$, Rn_{Sp} is empty, and Rn_{Sn} comprises the data point associated with a random key.

During the reduction phase, two partial results are combined into one. This process is iterated until only one partial result remains, which becomes the final result. Partial results can be combined in any order. The procedure for combining two partial results, *R1* and *R2*, into an output partial result, *Rout*, is as follows:

1. $Rout_{cp} = R1_{cp} + R2_{cp}$ (the positive counts are summed)
2. $Rout_{cn} = R1_{cn} + R2_{cn}$ (the negative counts are summed)
3. The cardinality of $Rout_{Sp}$ and $Rout_{Sn}$ are determined from $Rout_{cp}$ and $Rout_{cn}$.
    - If either $Rout_{cp}$ or $Rout_{cn}$ is smaller than the target minimum, *K*, then the cardinality of $Rout_{Sp}$ and $Rout_{Sn}$ are $Rout_{cp}$ and $Rout_{cn}$ respectively, e.g., sampling is not performed in this combination.
    - Otherwise, if $Rout_{cp} > Rout_{cn}$, then $Rout_{Sn}$ has a cardinality of *K* and $Rout_{Sp}$ has a cardinality of $K * Rout_{cp} / Rout_{cn}$
    - The reverse is true if $Rout_{cp} <= Rout_{cn}$: $Rout_{Sn}$ will have a cardinality of $K*Rout_{cn} / Rout_{cp}$ and $Rout_{Sp}$ will have a cardinality of *K*
4. $Rout_{Sp}$ is formed by merging $R1_{Sp}$ and $R2_{Sp}$ such that $Rout_{\mathrm{Sp}}$ is sorted on the random key assigned to each data point. Merging stops when the target cardinality of $Rout_{Sp}$ (calculated in step 3) is reached.

5. $Rout_{Sn}$ is formed by merging $R1_{Sn}$ and $R2_{Sn}$ such that $Rout_{Sn}$ is sorted on the random key assigned to each data point. Merging stops when the target cardinality of $Rout_{Sn}$ (calculated in step 3) is reached.

Both the mapping and reduction phases of this algorithm can be performed in a distributed manner. The ratio between positive and negative data points does not need to be known in advance; it is calculated as part of sampling. An example use case of this technique is to perform to obtain training examples for machine learning models such that a minimum number of positive and negative examples are included in the sample. The overall procedure for distributed multi-bucket sampling is summarized in Fig. 1.

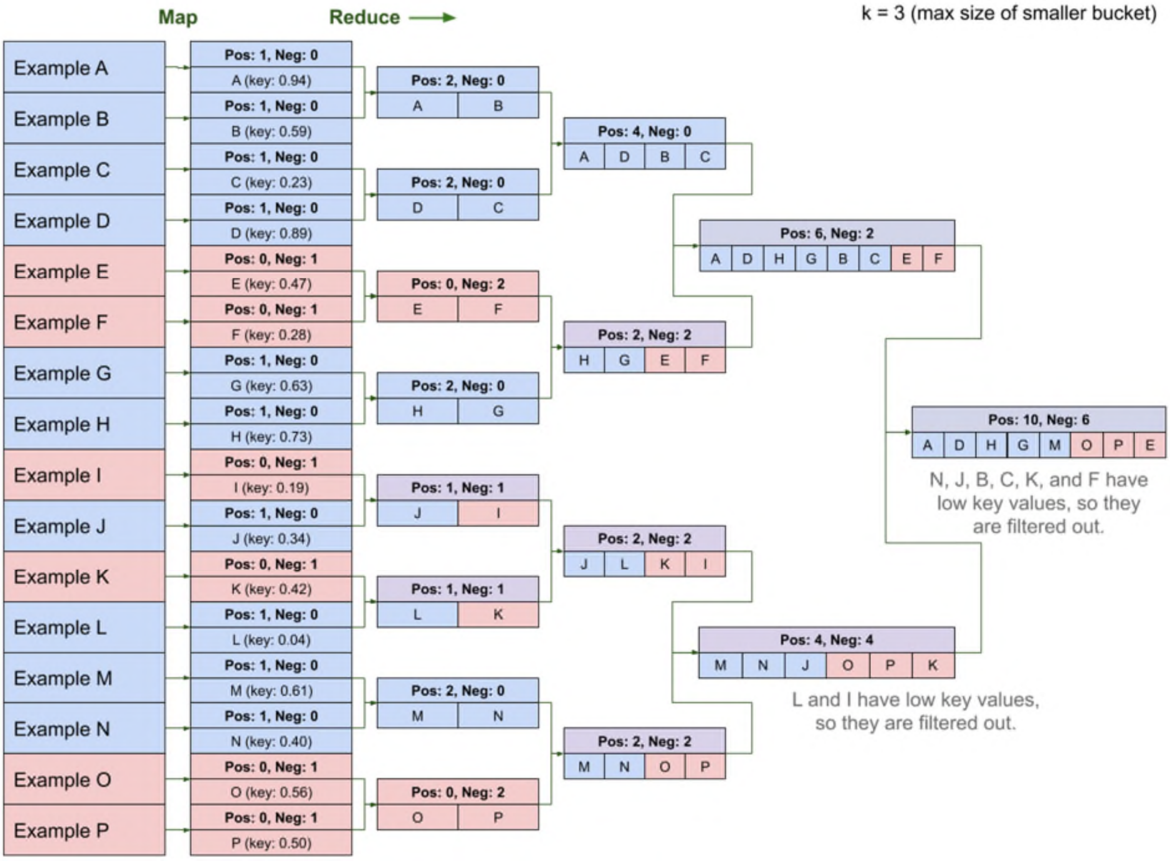

Fig. 1: Mapreduce procedure for distributed multi-bucket sampling

The described techniques can be used for any large dataset that includes positive and negative data points, such as training data for modeling.

CONCLUSION

This disclosure describes techniques to sample a large dataset using a mapreduce strategy. During mapping, a data point is one-to-one mapped to a partial result that includes counts of positive and negative data points and sets of sampled positive and negative data points. During the reduce phase, two partial results are combined into one in an iterative manner until only one partial result remains, which becomes the final result. Both the map and reduce phases are performed in a distributed manner. It is not necessary that the ratio of the positive and negative data points in the dataset be known in advance.