Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 8 August 2021

Short Communication

Review of a Proposed Data Quality Measure

Andrew H Sung*

School of Computing Sciences and Computer Engineering, The University of Southern Mississippi, Hattiesburg, USA

***Corresponding Author:** Andrew H Sung, School of Computing Sciences and Computer Engineering, The University of Southern Mississippi, Hattiesburg, USA.

Received: April 20, 2021
Published: July 13, 2021

Introduction

The rapidly growing big data collected or generated by various sources including public and private organizations, connected sensors, devices, the web and social network platforms, etc., have stimulated the advancement of data science, which holds tremendous potential for problem solving in diverse domains. Various methods and techniques have been developed to carry out data analytics and data mining for different purposes, where an ultimate goal of data mining is knowledge discovery pertaining to problems that have thus far defied solutions through conventional analytical approaches.

How to properly utilize the data to obtain useful analytics or to build accurate models in knowledge discovery is a topic of great importance, and two important issues arise: given a dataset, how to select a critical subset of features and how to select a critical subset of data points for sampling, in order to obtain an accurate model, using machine learning or other methods?

Feature selection, as machine learning and data mining researchers are aware of, is a challenging problem, especially when the number of features is very large. Likewise, sampling is an important consideration in data analytics and model building, since all the points in the dataset may not be needed in the process of, say, training and testing of a learning machine, to obtain the best results.

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

Data quality measures

With regard to the capacity or potential for knowledge discovery or model building 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, i.e., the dataset comprises n points in the p-dimensional space), in terms of n and p.

The concept of critical sampling is that there exists a critical number v ($0 \leq v \leq n$), such that a minimum of v samples (or data points from $D_{n,p}$) will be required to construct a model, using a specific methodology, that satisfies the minimal performance requirement of the user. Note that if a critical sample does not exist, then $v = 0$; in this case the dataset is inadequate since there is no way to use a sample of the dataset, or even the whole dataset, to construct a useful model.

The Sample Quality of $D_{n,p}$ is defined as $Q_s = v/n$ where v is the critical sampling size of $D_{n,p}$. So, $v \leq n$; and in case $v = 0$ when critical samples do not exist, then $Q_s = 0$ as well. In the optimal case, $v = n$ and so $Q_s = 1$, indicating that all points in the dataset are essential for the data mining or model building task. In other words, omitting any datapoint would lead to unsatisfactory results.

Likewise, the Feature Quality of $D_{n,p}$ is defined according to the concept of critical features, that is, there exists a critical number μ ($0 \leq \mu \leq p$), such that a minimum of μ features (out of the p features from dataset $D_{n,p}$) will be required to construct a model, using a specific methodology, that satisfies the minimal performance requirement of the user. The Feature Quality of dataset $D_{n,p}$ is defined as $Q_f = \mu/p$ where μ is the critical feature dimension of $D_{n,p}$. 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 inadequate possibly 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 analytic or model building task.

Q_D, the Overall Quality of the dataset $D_{n,p}$ (with respect to a specific methodology that is applied for model building and knowledge discovery purposes), can therefore be defined as:

$$Q_D = Q_s * Q_f = \frac{(v * \mu)}{(n * p)}$$

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 in practice, with the respective methods employed to find them), indicating that the dataset is inadequate for the purpose of model building to achieve minimal acceptable performance. At the other extreme, $Q_D = 1$ when $v = n$ 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 analytic or model building task for the problem under study.

As more and more IoT devices and physical objects generate data at accelerated pace while various public and private organizations are collecting more data, the society, with the advancement of data science, 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. Therefore, developing a quantitative measure for data quality, as described above, would be very useful. Finding a critical sample and a critical feature set, however, is a challenging problem and depends on many factors, including the dataset, the methodology for data analytics, the subjective performance requirement, and the application. Simple heuristic methods for finding critical samples and critical feature sets are described in the reference, which was also the first to introduce the definitions for data quality as explained above in this review [1].

Bibliography

1. B Ribeiro, J Silva, A H Sung and D Suryakumar. "Critical Feature Selection and Critical Sampling for Data Mining". Computational Intelligence, Cyber Security and Computational Models. Models and Techniques for Intelligent Systems and Automation (Ganapathi, G. et al., Editors) Springer CCIS Series (2018): 13-24.

Volume 3 Issue 8 August 2021