Minimization of Datasets : Using a Master Interlinked Dataset

* *Syed Ausaf Haider*
** *N. S. Patil*

Abstract

We all know there are a lot of datasets. Each data set corresponds to the contents of a single statistical database. Datasets have several properties based on statistical measures applicable to the number and type of attributes or variables. Here, the focus is mainly on statistics i.e., sampling of data based on observation and analysis. Each data of a dataset is sampled quantitatively by doing binary encoding. Sampling of a dataset using a predictor can often result in error. However, these errors can have a trend that might be related to one or more datasets. This can differentiate every variable of one dataset from remaining datasets. All these datasets can be unified into a single master dataset based on user requirements.

Keywords: Controlled datasets, dataset binary encoding, machine learning, master datasets, progressive sampling of data

I. INTRODUCTION

Master dataset is a huge collection of datasets having different variety. Datasets can either be public that are accessible to everyone such as GitHub data, crime data, hacker news, open images, diseases data, international debts etc., or they can be commercial datasets belonging to an institution such as of Accuweather, Dow Jones, Xi guide – norlot data cloud etc. These datasets can be combined together into a single master dataset.

Dataset can be distinguished from each other on the basis of:

a. Dataset Id
b. Dataset location
c. Expiry date (never or days (Integer))

These datasets can be viewed, updated (addition or deletion of new fields), and tabularized i.e., these datasets are totally editable.

The objective of linked datasets is to make data searchable and reusable. However, despite the fact that extensive lists of open datasets are available as per their respective catalogs, most of the data publishers typically link their datasets only to popular ones, such as DBpedia2, Freebase3, and Geonames4. Although the linkage to popular datasets allows the exploration of external resources, it fails to cover more specialized data. Specialized data means data featuring only a particular instruction based on a given parameter.

Datasets can be spaced into p dimensional field. This means that each dataset can be of the order (x_1, y_1) to (x_j, y_j) if the function considered is of the form $f(x)=\Sigma^{p}_{j=1} x_j \ y_j + \beta$.

Here, y is the output corresponding to the input x. x can be qualitative input such as hot, cold, tall, short, weak, fast, fat, sick, healthy etc., or x can be quantitative such as trigonometric function, complex number etc.

β is the considered error in the function f(x). Each dataset can be regressed linearly, logistically or using a classifier to find k nearest neighbor or any other regression algorithms.

However, regression can produce errors. Similarly, there can be n datasets in a master dataset. There can be such f_n(x) data functions [1].

II. PROPOSED SYSTEM

Consider any one of the n datasets, having function $f(x) = mx + x_0$

Now apply any one of the regression algorithms to this function f(x). Let us for simplicity consider these f_n(x) to be linear [2].

Manuscript received July 28, 2018; revised August 12, 2018; accepted August 14, 2018. Date of publication September 6, 2018.

* S.A. Haider is a student of B.Tech (Computer Engineering), Bharati Vidyapeeth Deemed University College of Engineering, Pune, India - 411043. (email: jdevyani99@gmail.com)

** N. S. Patil is Assistant Professor with Computer Engineering, Bharati Vidyapeeth Deemed University College of Engineering, Pune, India - 411043. (email: 98ausaf@gmail.com)

DOI: 10.17010/ijcs/2018/v3/i5/138778

Fitting of data of this dataset is done using least squares.
Least squares = $\beta = \Sigma^{N}_{i=1} (y_i - \mathrm{f}(x_i)$; writing this equation into native form $\sqrt{\mathrm{X}}$.
X is matrix of N x (p+1).
Differentiating twice least squares with respect to ?

$$\frac{d\sqrt{\beta}}{d\beta} = -2X^{T}(Y - X\beta\hat{Y}) \qquad (1)$$

$$\frac{d\sqrt{\beta}^{2}}{d\beta^{2}} = -2X^{T}X \qquad (2)$$

Equation (2) states function is invertible and positive definite. If this equation is equated to zero error, can be minimized.

$$XT(Y - X\beta) = 0 \qquad (3)$$

$$\hat{\beta} = (X^{T}X)^{-1}X^{T}Y \qquad (4)$$

$$\hat{Y} = X\hat{\beta} \qquad (5)$$

$$\hat{Y} = X((X^{T}X)^{-1}X^{T}Y) \qquad (6)$$

From equation (6) we can conclude that actual output Y depends on estimated output $\hat{Y}$.
Such series of variates are regressed using a multivariate regression.Here a univariate variable (x,y) is considered.

$$\mathrm{Y} = \mathrm{X}\beta + € \qquad (7)$$

$$\hat{\beta} = \frac{\Sigma^{N}_{i=1} x^{i} y^{i}}{\Sigma^{N}_{i=1} x^{i}} \qquad (8)$$

In Normalized form, the equation can be written as

$$\hat{\beta} = \frac{<X,Y>}{<X,x>} \qquad (9)$$

<> denotes inner products i.e. can be complex numbers also.

$$\gamma = Y - X\hat{\beta} \qquad (10)$$

$$r^{i} = y_i - x_i \beta \qquad (11)$$

Here, r^{i} is residual error.

III. ERROR AS A DATASET

The main idea is to collect errors of the same kind. The error or noise of one dataset can be a data of another dataset depicting the same behavior and properties of that dataset.

Master dataset can act as a controller to predict error as new data of another dataset (belonging to master dataset). Thus, it adds or updates the datasets of a master dataset every time an error occurs.

As master dataset is totally editable when the user can collect errors of the same kind that might show a new unique property each time datasets are regressed into a new unique dataset.

The f_n(x) data functions (of different datasets) may have r_n^{i} residual errors.

There r_n^{i} residual errors can either be regressed iteratively and simultaneously verified with the parameters of all the remaining datasets.

IV. INTERLINKING OF DATASETS ACCESSING ENCODED DATASETS

A rank score function inspired from conditional probabilities that induces the ranking of the datasets in D (from the largest to the smallest score), can be defined as follows:
score $(D_i$, t) = X j=1…n
log (P $(F_j|D_i)$)! + log(P(D_i)) (12)

Based on the maximum likelihood estimate of the probabilities [8] in a training set of datasets, the above probabilities can be estimated as follows:
score(D_i,t) = X j=1…n
log(P($Fj|Di$))! + log (P(D_i)) (13)

Based on the maximum likelihood estimate of the probabilities in a training set of datasets, the above probabilities can be estimated as follows:
P($F_j|D_i$) =count(F$_j$di). Pn j=1 count(F_j,D_i); (14)
P(D_i) =count(D_i) Pm i=1 count(D_i) (15)

where,

count(F_j, D_i) is the number of datasets in the training set that have feature F_j and are linked to D_i,

count(D_i) is the number of datasets in the training set that are linked to di, disregarding the feature set.

For the score function computation, some auxiliary functions help to avoid computing log(0) replacing this value by *c*, which is a constant small enough to penalize the datasets di that do not have datasets with features F_j linked to them or that do not have links from other datasets. Thus, the idea is that if the set of features of t are very often correlated with datasets that are linked to D_i and *t* is not already linked to di, then it is recommended to try to link t to D_i.

V. BINARY ENCODING

8-bit binary encoding of dataset is one in which all the data along with its bias and errors are encoded such that each of the datasets have their individual own identity such that every entity set is regressed throughout. Datasets need not show any relations or any resemblance to each other for binary encoding as entities once encoded would be independent of each other. Encoded data whether it is an error in a dataset or an important feature of it can be structured and ready to organize even if the datasets are disjoints. Searching a particular attribute or an entity can be very easy from one dataset to another. Encoding enables automation of data throughout the dataset.

VI. ACCESSING ENCODED DATASETS

Search metadata: The metadata describing the dataset can be searched by entering any string into the text box provided by the interface.

Visualize data: When there is data suitable for visualization, a "Visualize Data" button under the files section of the assay pages launches a Genome browser track hub for visualization.

Download data: All released data are publicly available for download. Bulk downloads of data and metadata associated with the files can be performed by programmatic access of the encoded API.

VII. DIAGRAMMATIC EXPLANATION

A. Unorganized Data

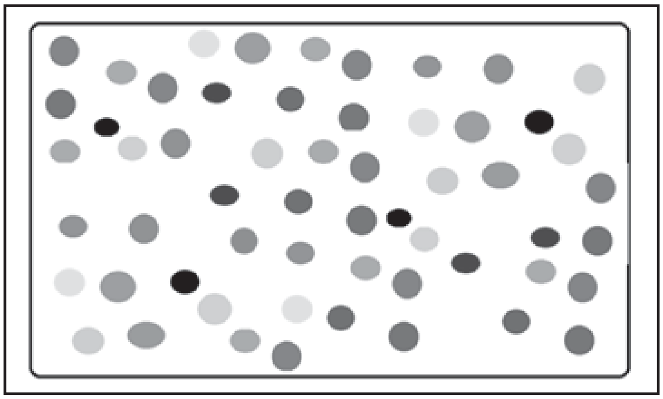


Fig. 1. Unorganized Data (A Garbage Dataset)

From Fig. 1 it can be inferred that unstructured data having no predefined model in the affinity of unclosed space. These data act as N raw information that act as garbage.

B. Binary Encoding

From Fig. 2 it can be inferred that an 8-bit binary encoding is performed such that there is no interference of data variables. The non-interference of data variable enables structuring of data and allow search of each entity independently. Binary encoding avoids ambiguity and provides a linkage between other entity. Binary encoding avoids randomness of data variables.

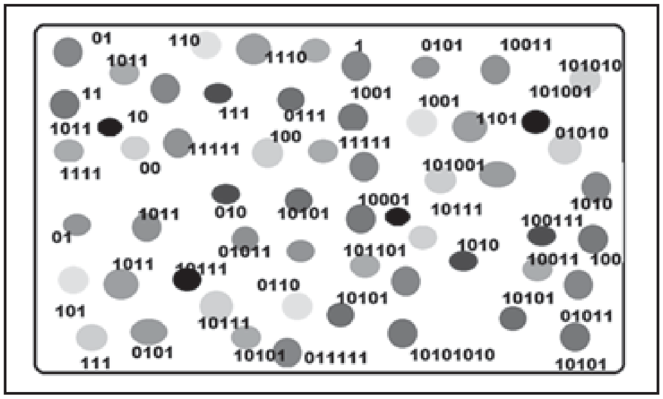


Fig. 2. Binary Encoding of Datasets

C. Structuring Datasets

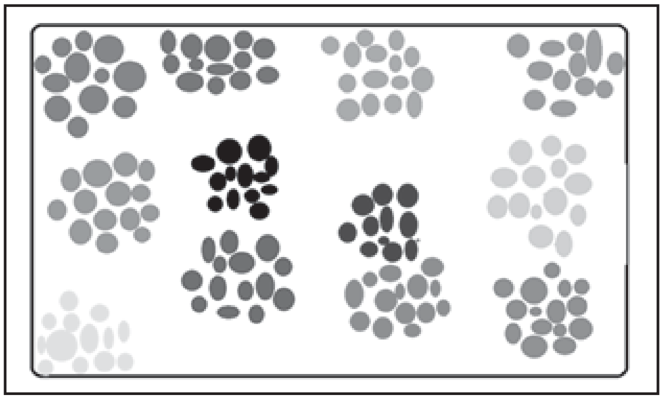


Fig. 3. Organization into Datasets

From Fig. 3 it can be inferred that encoded data are now clustered together as per the required domain. Each cluster shows identical attributes tending to show a specific behavior and functionality.

D. Labelling Clusters

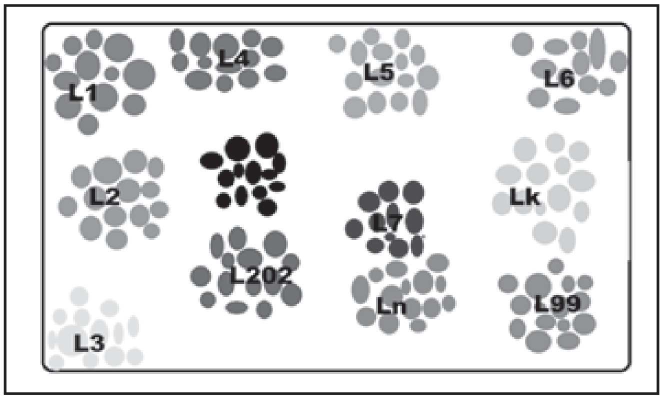


Fig. 4. Labeling of Datasets

From Fig. 4 it can be inferred that every cluster is now

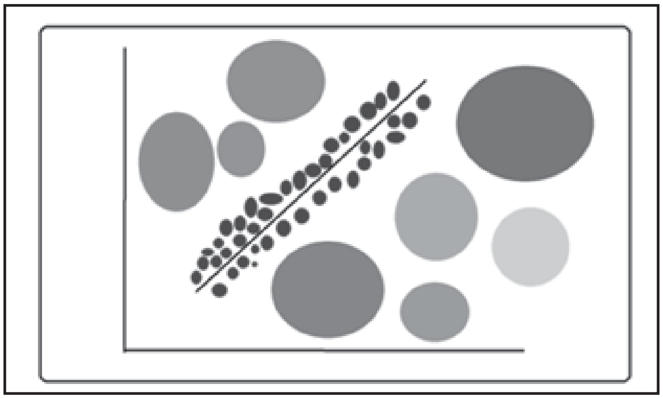
labeled as per the requirement. Labeling is performed on the basis of the following aspects:
a. Dataset id
b. Dataset Location
c. Expiry Date (Never or days (Integer))

E. User Access

Now there are symmetrically organized labeled datasets that are arranged as per the binary encodings. User can focus now on the required datasets avoiding the unnecessarily datasets. This can save execution memory space and unnecessary regression of datasets.

Encoding allows the user to search for an entity even if it lies in any other datasets. Thus, it creates a new relation (link) very easily among the datasets.

Suppose, now we want to focus on any one of the labels or Dataset ID (consider L7). We shall be using K nearest neighbor regression (as per graphical convenience).



. Fig. 5. K nearest neighbor (Regression only on Dataset L7)

From Fig. 5 it can be inferred that we see only Dataset ID L7 is regressed symmetrically while the uniformity of other datasets is maintained. This is only possible through binary encoding. Now, if we append a dataset with another dataset and then regress again, we achieve the graph as shown in Fig. 6.

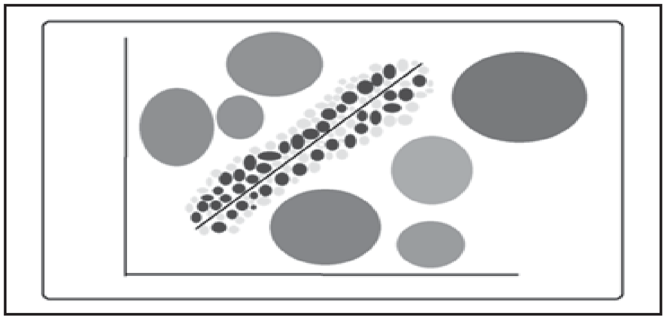


Fig. 6. Regression of Appended Datasets L3 and L7

VIII. ADVANTAGE OF USING ENCODED DATASETS

❖ Creating custom hierarchical data.
❖Encoded datasets can easily be navigated and organized.
❖ Affordable system.
❖ Avoiding interference between different datasets.
❖ Master dataset acts as controller to the other datasets.
❖ This system can also be associated with a large number of applications in various fields.

IX. CONCLUSION

Now a days there are a lot of datasets available and many are being created every day. Interlinking these datasets together can relate them with each other even if they are disjoint.

User can focus only on the required dataset among the all other datasets or the user can focus on a particular entity among all datasets.

Master dataset removes ambiguity and is involved in the creation of a new dataset either from given errors of one or more datasets or creation of new datasets from a resultant of two or more datasets in a controlled way.

REFERENCES

[1] Dataset [Online]. Available: https://en.wikipedia.org/wiki/Data_set

[2] Linear Regression.[Online]. Available: https://en.wikipedia.org/wiki/Linear_regression

[3] Encoding (memory) [Online]. Available: https://en.wikipedia.org/wiki/Encoding_(memory)

About the Authors

Syed Ausaf Haider is a student of B.Tech, Computer Engineering at Bharati Vidyapeeth Deemed University College of Engineering.

N.S. Patil is Assistant Professor (Computer Engineering) with Bharati Vidyapeeth Deemed University College of Engineering. She completed B.E. Computer Science and Engineering from Walchand College of Engineering, Sangli, Shivaji University in 2000. She completed M.E. (Computer Engineering) from Bharati Vidyapeeth Deemed University College of Engineering, Pune in 2008. She is pursuing Ph.D. She has teaching experience of 12 years.