**A Data Analytics Methodology to Visually Analyze the Impact of Bias and Rebalancing**

Link for this Journal:<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10132458>

**Introduction:**

This paper is completely about the problems faced by many machine learning models or AI models in performing data analytics and efficient data models for prediction. The authors tried to solve the problem of data biasing and its effects.

The problem of biasing can be raised from the given data itself, for example if we can consider an attribute from a table then we can check the class distribution for that particular attribute and how the data is distributed for a particular class of attribute. In day-to-day life we can observe that real time data is biased based on several reasons. It is this data which leads to biasing due to the uncontrollable generation of the data.

Biasing can drastically decrease the efficiency and accuracy of algorithmic models. There are several types of biases but out of them all this paper tried to focus on two types in particular because the author felt these are the most influential. They are carefully addressed by the author in their experimentation.

The authors tried to find a new approach from their observations and experience with the data as they mentioned about their previous work in this paper. The authors mentioned about other techniques in their previous papers,

1. They introduced a model to address users with their visual requirements.
2. They developed this Model Driven Architecture (MDA) to customize these visual requirements.
3. They developed a bias detection process which detects bias automatically.

The main aim of this paper is to reduce biasing by developing an efficient algorithm for the data which will decide the biasing factor of each attribute and which attributes have more bias in designing an efficient model and also show the bias distribution over a specified attribute visually.

**Technical Contributions:**

The solution for the problem prescribed is given by the authors by including a bias detecting algorithm for their previous work to check biasing factor.

This can be explained as the following process: First the data from the dataset goes inside the Data Profiling Model and then it combines with the User Requirement Model obtained by the user and these customizations are then sent inside Data Visualization Model and then the output from this is checked if it satisfies the user. If this satisfies the user, then it is later sent into the Bias Detection Process for checking the bias within the data.

The algorithm mentioned inside this Bias Detection Process is that the bias for each individual is calculated by using the formula:

Bias attribute = (max – min/max) \*10

max – represents the maximum number of repetitions of a particular attribute class.

min – represents the minimum number of repetitions of a particular attribute class.

This value is compared with a biasing factor value which scales from 0 to 10 (0 – equally distributed and 10 - strongly biased).

The authors tried to calculate Accuracy, False Positive rate, and False Negative rate of 7 different machine learning models by including certain attributes as key metrices for their training and testing data. The authors divided the dataset into three categories 1) Normal dataset, 2) Over sampled dataset and 3) Under sampled dataset. The results are much favored for the normal dataset compared to the other two. And also, from these tables the author was able to clearly distinguish between the key attributes based on their biasing factor. This method which is prescribed by the author also shows us how biased those attributes were in the form of visualization which is the main aim of this paper.

According to me, I definitely agree this can be a reliable method in determining the biasing, but I also think that the deciding metrics for the bias attribute can also be decided by using other strategies of calculation not only using the percentage calculation as described by the author. If there exists any other method of calculation, the author should also include the information by conducting experiments and compare those results with this method.

**Improvisations:**

Now according to this paper and as I said earlier, the author’s main aim in this paper is to represent biasing through visualization and check for biasing automatically. My main concern is that the author should have also worked on the solution for this biasing by giving a much more efficient way rather than using over-sampling or under-sampling of the given data.

I think that the paper can be more knowledgeable after including additional information such as considering several other types of calculation metrices to decide the biasing factor apart from the one included by the author. I would have worked on other methods and tried to do the experimentation to determine which method can give the accurate biasing coefficient value and then later on also try to give the solution to handle that biasing.

If I had a chance to meet the author in person and talk about the paper, I would have definitely asked him about these two problems which I stated above and also asked him, if he had performed on all types of datasets using those models. And also coming to the point on performing the analysis on these datasets, the authors should also have considered the fact that in order to perform the data analytics it is necessary to not mistake majority discrimination of data with biasing.

The side effects of this bias discrimination are that it basically tries to reduce gaining insights from the data because, if we place all the attribute classes on the same sample level then how can we determine the difference between reality and biasing. For example, if we consider a soap company to publish their new soap over old soap and from the data if we try to consider the analytical visualization and if the sales of new soap are greater than old soap it doesn’t mean that tha data is biased over new soap. This is a very small detail but yet it can have a huge impact on research.