International Journal of Scientific Research and Engineering Development-– Volume 5 Issue 6, Nov- Dec 2022

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Risk Factor Analysis Associated with BRFSS Dataset

Tapan Bhagvanbhai Golakiya*, Ruchita Tapan Golakiya**, Shrey Gupta***

*(Software Development Engineer, Brillio LLC, Canton, MI, USA
Email: golakiyatapan@gmail.com)

** (Department of Computer Science, Wayne State University, Detroit, MI, USA
Email : hi3265@wayne.edu)

*** (Department of Computer Science, Emory University, Atlanta, GA, USA
Email : shrey.gupta@emory.edu)

----------**********----------

Abstract:

The Behavioral Risk Factor Surveillance System is a collaborative project between the states of the US which aims at collecting information about health-related risk behaviors. We have applied the unsupervised learning algorithm aka K-means clustering to determine the group with high and low risk of health issues. The BRFSS dataset has been put through some useful feature selection and reduction techniques.

Keywords ***—dimensionality, variance, correlation, K-means, clustering.***

----------**********----------

I. INTRODUCTION

The Behavioral Risk Factor Surveillance System is a collaborative project between all the states of the USA. The goal of the BRFSS is to collect state-specific data on health risk-behaviors, chronic diseases, and access to health services. This would help them determine the leading causes of death and disability in the United States. Using this data set we would try to determine is a person is at risk of death/disability to health risks. We are not aiming towards a particular disease and trying to predict that but, are looking for a red flag pointing towards a potential health risk. This type of analysis is also called risk factor analysis. Since, no labels have been provided in the data set, hence we would be applying unsupervised learning approaches.

II. METHODS

A. Data Pre-processing

The BRFSS dataset consisted of 450,016 samples of different people residing in the United States and 358 different features that contained health information of the people. Since, the feature space and the sample space were large, there were also missing values in the dataset. Multiple methodologies were applied to reduce the dimensionality of the dataset. In the following sections we would be covering those methodologies. In Pre-processing we were concerned with making the dataset robust. By this we mean that unwanted features should be removed, and all the null values should be replaced with the average value of the feature. We performed the following steps in pre-processing:

1) We removed all the features from the dataset which had all the values as null. These features did not contribute anything to the dataset and our analysis process.

2) We found that the following initial eight columns: _STATE, FMONTH, IDATE, IMONTH, IDAY, IYEAR, DISPCODE and SEQNO were not contributing any useful information which could be analyzed. These features had values which could be used to uniquely identify a sample. However, they had no real contribution towards the resultant grouping of samples.

3) We then found the count of the missing values for each feature. The number of missing features vs the feature names plot is as shown in Fig1. If a feature had more than 200,000 or about ~45% of values are missing, then the feature was removed. This is because it would be highly inappropriate to predict such a large no. of values for a feature.

4) Since the features were a mixture of categorical and numerical data, we filled the missing values with the mode of the feature samples. Finding the mean would not be the correct measure since some features might just contain three to four unique values.

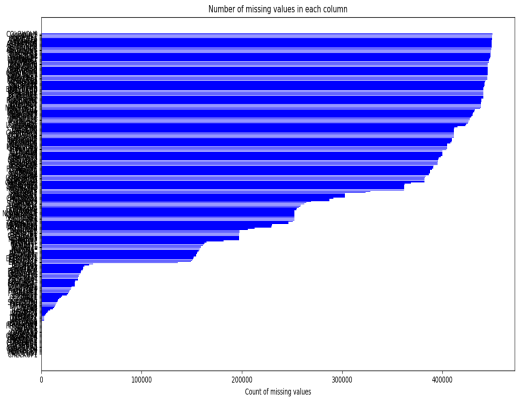

Fig 1. The y-axis represents the features and the x-axis represent the missing count for the features. The more the blue values, the higher the missing values in the sample

B. Low Variance Filter

We applied feature selection on the set of features obtained after the pre-processing step. One of the steps included finding the features with high variance. A high variance feature meant that number of samples varied in values and therefore contributed towards the resultant groups. This is called a Low variance filter. We first separated out the features that had unique values less than 10. Since applying a variance filter to such features has no utility. We then applied the variance filter on the remaining feature. The variance threshold was set at 0.65. This means that the features having a variance less than 0.65 were not selected.

C. High Correlation Filter

Another step towards feature selection was applying a high correlation filter. If the feature had samples that were highly correlated, then these samples behaved almost the same and did not contribute with different cases towards finding the resultant groups. Therefore, we applied a correlation filter to extract all the features with samples having low correlation among each other. The threshold for the correlation filter was kept being greater than 0.8.

D. Principal Component Analysis

We then joined the feature set that was removed before both the filter was applied with the feature set obtained after the application of filters. We applied the dimensionality reduction technique called the Principal Component Analysis (PCA). This technique helps in reducing the feature space by transforming the features into features with a high variance. The newly produced features or the principal components as we call them, are ranked with the first principal component being the best representation of all the features and so on. We chose first two principal components for our unsupervised learning technique.

E. Data Analysis using K-means Clustering

We applied K-means clustering on obtained principal components. We first normalized our principal components in the range [0,1]. This was done to ensure that there are no negative values, and the cluster shape are not arbitrary. If the range would have been kept from [-1,1], the cluster obtained were in the shape of a circumference of circle. This was because it had some values going into the negative region. We then applied K-means clustering for K=2,3...10 clusters. We used the pre-defined function provided in the python library to perform the clustering. Since, we had no pre-defined labels to compare our results with, therefore, we went with K=2 clusters. The clusters obtained when K=2 is shown in Fig 2. below.

III. RESULTS AND DISCUSSION

The results obtained after applying the unsupervised learning algorithm i.e., K-means clustering, shows that the resultant labels can be divided into two groups. The first group/cluster represented all the people who were at a minimal risk of death/disability. Similarly, the second group/cluster represented all the people who were at a high risk of death/disability due to health conditions.

Although it is difficult to determine which cluster belongs to which group/labels, but we can observe from the graph that one of the groups is varied along the first principal component and another group is varied along the second principal

component. This shows that the first principal component captures the essence of all those features which might belong to a certain group of people. The group might be with a high risk of death/disability or a low. But it is easily distinguishable by the two principal components and the clusters formed by them.

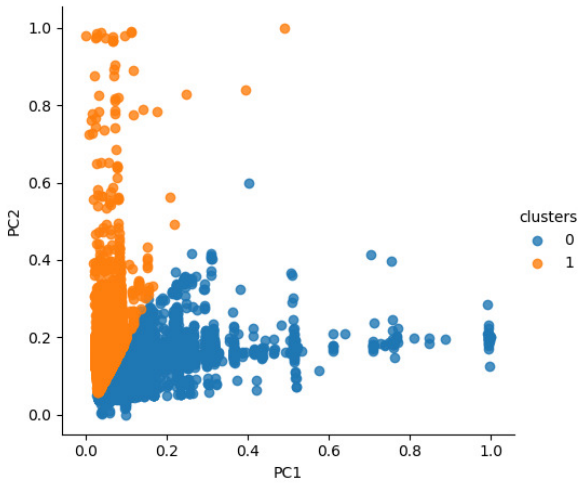

Fig 2. The clusters obtained after applying K-means clustering on the obtained principal components. Here, K=2. The two clusters are represented by distinct colors.

IV. CONCLUSION

Unsupervised learning is a complicated process especially when the data set is so sparse and has a lot of null values. BRFSS data set on health risk gives us a glimpse of that. Our feature set is large and so is the sample space. We show in our methodology on how to tackle such a dataset. We implement different feature reduction and selection technique to reduce the number of features. We finally use the K-means clustering algorithm to achieve the class labels. The labels obtained can be mapped to the original labels (if present) and tested for the accuracy of the procedure.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. SuhNew York: Academic, 1963, pp. 271–350.

[4] S. Na, L. Xumin and G. Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm," 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, 2010, pp. 63-67, doi: 10.1109/IITSI.2010.74.

[5] K. Elissa, “Title of paper if known,” unpublished.

[6] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: UniversityScience, 1989.

