Point Biserial Correlated Feature Selection of Weather Data

Pooja S.B, R.V Siva Balan

Abstract: ***Big data is said to be huge amount of data. These data's may be either structured or unstructured data. It is used for performing prediction in many fields and one among them is weather forecasting. Many feature selection techniques has been introduced but all these techniques failed to get accurate result. In order to improve weather prediction with less complexity, a Point Biserial Correlated Feature Selection (PBCFS) technique is introduced. The big weather dataset comprises the 'n' numbers of features. Initially, the PBCFS technique uses a point biserial correlation coefficient to determine relevant feature or irrelevant features among the several features. These relevant features which is selected with the help of this feature selection method can be used for clustering, classification or any other method to perform prediction. The polytomous (i.e. different classes) regression function analyzes the input data with the selected features to provide the significant results as output. Experimental evaluation of proposed PBCFS technique and existing methods are carried out using a big weather dataset. The result shows that we get the output with high feature selection accuracy.***

Index Terms: ***Big data, weather forecasting, feature selection.***

I. INTRODUCTION

The term big data means large amount of data it differ from the normal data. It consist of both structured and unstructured data and it is difficult to handle. To predict the future weather condition weather forecasting is used [1][2]. Big data plays a major role in most of the applications. In this work input data is taken with the help of remote sensing device which is said to be remote sensed data. In order to select the relevant features for more accurate prediction, typicality-and-eccentricity-based evolving intelligent method was introduced in [3] using Spearman rank correlation. The method has a higher false positive rate in the weather time series prediction. Genetic algorithm based feature selection was introduced in [4] for predicting the very short-term deep rainfall. The approach does not use efficient machine learning approaches for enhancing the prediction accuracy. Hadoop Map Reduce technique was developed in [5] for improving the weather prediction with big data. The technique does not use any feature selection algorithm for minimizing complexity[6][7]. Inorder to improve the feature selection with high complexity Point Biserial Correlated Feature Selection (PBCFS) is introduced. The main contributions of the proposed PBCFS-LPBC technique are described as follows. In this Point Biserial Correlated Feature Selection (PBCFS) technique feature selection is done. Feature selection is an attribute selection from the large dataset for predictive analytics with less complexity. The various data mining techniques such as clustering, classification and so on are used for predictive analytics. The relevant feature from the big dataset is selected before the predictive analysis. In first, the number of features in the big weather dataset is selected for solving the complexity problem in the data classification and improving the prediction performances. The Point Biserial Correlation separates the features into two subsets as relevant and irrelevant by computing the mean and deviation.

II. POINT BISERIAL CORRELATED FEATURE SELECTION

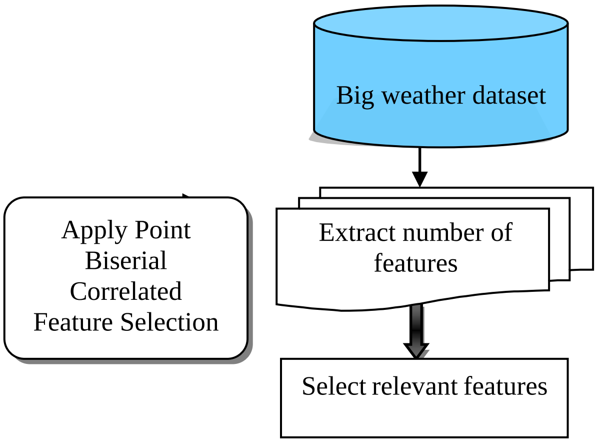
Big weather dataset
Apply Point Biserial Correlated Feature Selection
Extract number of features
Select relevant features

Figure 1 Flow process of PBCFS technique

A Point-Biserial Correlation Coefficient is used for measuring the linear correlation to identify the relevant features. Flow process of LPBC technique is illustrated in figure 1. The process of the PBCFS technique is the feature selection from the big weather dataset. The big weather dataset consists of hundreds and thousands of instances and each of which is represented through the number of features (i.e. attributes). While processing more instances with the feature set, time complexity arises [8][9]. Conventional feature selection algorithm was developed to improve the stability of choosing relevant features under the small sample size [10]. However, accurate and effective feature selection was not performed. Therefore, a feature selection is required for minimizing the complexity in data classification.

Revised Manuscript Received on October 30, 2019.
* Correspondence Author

Pooja S.B Department of Computer Science, Noorul Islam Centre for Higher Education, Kumaracoil 629180, Tamil Nadu, India

R.V Siva Balan*, Department of MCA, Noorul Islam Centre for Higher Education, Kumaracoil 629180,Tamil Nadu, India

NOTATIONS	DEFINITION
f_i	Set of features
D^W	Big weather dataset
ρ_{pb}	Point-biserial correlation coefficient
m_1	Mean value on the features in the subset 1
m_2	Mean value on the features in the subset 2
p_1	Number of features in subset 1
p_2	Number of features in subset 2
n	Total number of features
s_d	Standard deviation

Table 1: Notations and its definition

Table 1 shows the notation and its definition used in the below equation.

The PBCFS technique uses the Point Biserial Correlation coefficient for finding the statistical correlation between the two variables such as features and dichotomous variable. A Dichotomous is a separation of an entire set into two subsets namely relevant set and irrelevant set. The correlation between the subset and features are measured to find the relevant features. Therefore, PBCFS technique performs better than the least absolute shrinkage and selection operator (LASSO). Because, LASSO is used for variable section but it is computationally slow. Let us consider the 'n' number of features in the big weather dataset D^W

$$f_i = \{f_1, f_2, f_3, \dots, f_n\} \in D^W \quad (1)$$

From (1), f_i denotes a set of features $\{f_1, f_2, f_3, \dots, f_n\}$, D^W represents a big weather dataset. To calculate correlation, the dataset is divided into two subsets where the first set received the binary values as '1' and the other one is received as '0'. The binary value '1' indicates features in the subset are more relevant for weather forecasting. The other binary value '0' indicates a feature in the subset is not relevant for further processing. In addition, standard deviation is calculated for dividing the dataset in two subsets. If the feature is highly diverged from standard deviation (i.e., weather data), then the dataset is divided into irrelevant set. Otherwise, the feature is slightly diverged from output of standard deviation, then the dataset is divided as relevant. Based on the concept, the point-biserial correlation coefficient is calculated as follows:

$$\rho_{pb} = \frac{m_1 - m_2}{s_d} * \left(\frac{p_1 * p_2}{n^2}\right)^{1/2} \quad (2)$$

From (2), ρ_{pb} denotes a point-biserial correlation coefficient, m_1 denotes a mean value on the features in the subset 1 and m_2 represents a mean value on the features in the subset 2. p_1 denotes a number of features in subset 1 and p_2 denotes a number of features in subset 2.'n' denotes a total number of features in the dataset. s_d represents the standard deviation. Standard deviation is used to identify how much the members of a subset diverge from the mean value for that subset. The standard deviation is mathematically formulated as follows,

$$s_d = \sqrt{\frac{\sum_{i=1}^{n}(f_i - m_f)^2}{n-1}} \quad (3)$$

From (3), s_d represents the standard deviation, f_i denotes features in the particular set, m_f denotes mean values of the features in the particular subset (subset 1 or subset 2), 'n' denotes a total number of features in the dataset. Based on the standard deviation and mean value, the features are separated as relevant or irrelevant into two subsets. The correlation coefficient selects relevant features in the subset 1 based on the mean and deviation value. The algorithm of PBCFS technique is described as follows.

Algorithm 1 describes the point biserial correlated feature selection to improve the weather prediction accuracy. Initially, the weather big data comprises the number of features and the data. Before classifying the data, the feature selection is carried out based on the correlation measure. The high correlation between the features and the mean values of the subset used to select the relevant features.

```
Input: Weather big dataset$D^W$, number of
features$f_1, f_2, f_3, \dots, f_n$, Number of data $d_1, d_2, d_{3,\dots} d_n$, $c_j$
denotes a different classes.
Output:  Improve feature selection
Begin
\\ Feature selection
    1.  for each            then
    2.      Measure the correlation
    3.      Select high correlated  features
    4.    end for
    End
```

Algorithm 1 Point Biserial Correlated Feature Selection

III. EXPERIMENTAL SETTINGS

Experimental evaluations of proposed Point Biserial Correlated Feature Selection technique and existing methods namely hybrid neural model [11] and SVR [12] are implemented using Java language. The Atlantic hurricane database is used. The database is taken from https://www.kaggle.com/noaa/hurricane-database. Proposed PBCFS uses holdout method for cross validation. The input dataset is separated into two sets such as training set and testing set. Most of training set is used for training i.e., 60 percentage of data and smaller portion of the data is taken for testing i.e., 40 percentage of data. Besides, response variable i.e., prediction of cyclone is considered for experimental analysis.

Response variable (dependent variable) is defined as the result variable in the experiment. While studying the impact of the cyclone prediction from weather dataset, the response variable is prediction of cyclone (Y) and the independent variables are feature selection accuracy (x_1), This result provides the feature selection accuracy (x_1) as 75%. Out of 1000 data,

Impact feature selection accuracy

Feature selection accuracy is defined as the number of features that are more relevant for weather prediction is selected correctly to the total number of features. The feature selection accuracy is mathematically calculated as follows,

$$feature\ selection\ accuracy = \frac{Number\ of\ features\ selected\ correctly}{n} * 100$$

(4)

From (4), 'n' represents the umber of features taken for the experimental evaluation. The feature selection accuracy is measured in terms of percentage (%).

Sample calculation for feature selection accuracy:

Proposed PBCFS technique: Number of features selected correctly is 3 and the total number of features is 4. The feature selection accuracy is computed as follows,

$$feature\ selection\ accuracy = \frac{3}{4} * 100 = 75\%$$

Existing hybrid neural model: Number of features correctly selected is 3 and the total number of features is 4. The feature selection accuracy is computed as follows,

$$feature\ selection\ accuracy = \frac{2}{4} * 100 = 50\%$$

Existing SVR: Number of features correctly selected is 3 and the total number of features is 4. The feature selection accuracy is computed as follows,

$$feature\ selection\ accuracy = \frac{2}{4} * 100 = 50\%$$

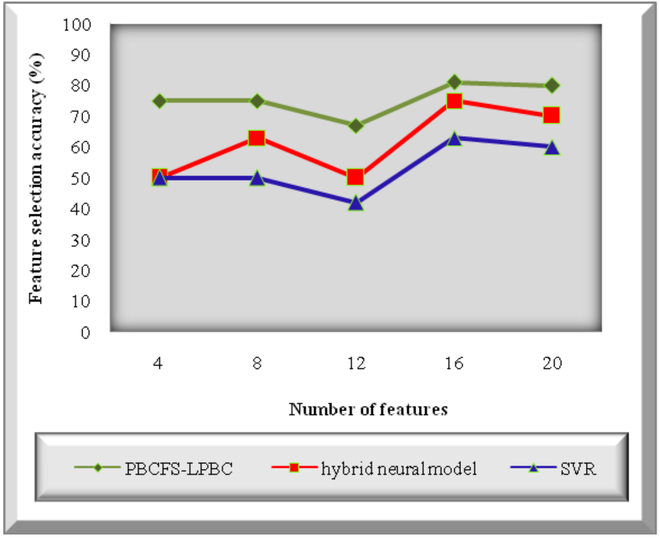

Figure 2 performance results of feature selection accuracy

Figure 5 illustrates performance results of feature selection accuracy with three different methods namely of LPBC technique and hybrid neural model [11] and SVR [12]. As shown in figure 5, the numbers of features are taken as input varied from 4 to 20 for computing the feature selection accuracy. Totally five different runs are carried out and provide the various feature selection accuracy results for three different methods. The above graphical result shows that the PBCFS technique accurately selects the features for predicting the different tropical cyclones in the Atlantic Ocean when compared to existing methods. This significant improvement is achieved by measuring the Point Biserial correlation. The Atlantic Ocean dataset is divided into two subsets. Let us consider 4 features in the first run. The feature selection accuracy of LPBC technique is 75% whereas the feature selection accuracy of the hybrid neural model [11] and SVR [12] are 50% and 50% respectively. Similarly, the four remaining runs are carried out and compare the performance results of proposed and existing methods. The comparison results prove that the feature selection accuracy is considerably improved using PBCFS technique by 25% when compared to the existing hybrid neural model [11]. In addition, the PBCFS technique increases the feature selection accuracy by 44% when compared to SVR [12].

IV. CONCLUSION

The main goal of this research work is to perform feature selection with high accuracy in a minimum amount of time so that it can be used as the input of predictive analysis in many fields such as clustering classification and so on [13]. To improve the feature selection PBCFS technique is used. It helps to selects the relevant features for weather prediction from the dataset by using point-biserial correlation. The correlation coefficient separates the relevant features and irrelevant features to improve the feature selection accuracy and minimizes the time complexity. The performance of the proposed method is determined by comparing it with the existing method and shows the it is more accurate than the existing methods.

REFERENCES

1. S. B. Pooja and R. V. S. Balan, "An Investigation Study on Clustering and Classification Techniques for Weather Forecasting", Journal of Computational and Theoretical Nanoscience vol. 16, no. 2, pp. 417–421, 2019.
2. S. B. Pooja and R. V. S. Balan, "Iterative Gradient Ascent Expected Maximization Clustering for Weather Forecasting", International Journal of Recent Technology and Engineering no. 6, pp. 412–418, 2019.
3. Eduardo Soares, Pyramo Costa Jr, Bruno Costa and Daniel Leite, "Ensemble of evolving data clouds and fuzzy models for weather time series prediction", Applied Soft Computing, Elsevier, Volume 64, March 2018, Pages 445-453
4. Jae-Hyun Seo,Yong Hee Lee, and Yong-Hyuk Kim, "Feature Selection for Very Short-Term Heavy Rainfall Prediction Using Evolutionary Computation", Advances in Meteorology, Hindawi Publishing Corporation, Volume 2014, January 2014, Pages 1-15
5. Basvanth Reddy and B.A Patil, " Weather Prediction Based on Big Data Using Hadoop Map Reduce Technique", International Journal of Advanced Research in Computer and Communication Engineering, Volume 5, Issue 6, 2016, Pages 643-647
6. Prasanta Rao Jillella S.S, P Bhanu Sai Kiran, P. Nithin Chowdary, B. Rohit Kumar Reddy, Vishnu Murthy, "Weather Forecasting Using Artificial Neural Networks and Data Mining Techniques", International Journal Of Innovative Technology And Research, Volume 3, Issue.6, 2015, Pages 2534 – 2539

7. Mumtaz Ali , Ravinesh C.Deo, Nathan J.Downs, Tek Maraseni, "Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting", Atmospheric Research, Elsevier, Volume 213, 2018, Pages 450-464
8. Gunasekaran Manogaran, Daphne Lopez, Naveen Chilamkurti, "In-Mapper combiner based MapReduce algorithm for processing of big climate data", Future Generation Computer Systems, Elsevier, Volume 86, 2018, Pages 433-445
9. P. Samuel Quinan and Miriah Meyer, "Visually Comparing Weather Features in Forecasts", IEEE Transactions on Visualization and Computer Graphics, Volume 22, Issue 1, January 2016, Pages 389-398
10. Yue Han and Lei Yu, "A Variance Reduction Framework for Stable Feature Selection", Statistical analysis and data mining, Wiley, Volume 5, Issue 5, 2012, Pages 428-445
11. Tanzila Sabar, Amjad Rehman, Jarallah S. AlGhamdi, "Weather forecasting based on the hybrid neural model", Applied Water Science, Springer, Volume 7, Issue 7, 2017, Pages 3869–3874
12. Xiongxin Xiao, Tingjun Zhang, Xinyue Zhong, Wanwan Shao, Xiaodong Li, "Support vector regression snow-depth retrieval algorithm using passive microwave remote sensing data", Remote Sensing of Environment, Elsevier, Volume 210, 2018, Pages 48–64
13. S. B. Pooja and R. V. S. Balan, "Weather Data and Its Future Selection Using Principal Component Regression Technique", Journal of Advance Research in Dynamical & Control Systems, Vol. 11, 04-Special Issue, 2019.

AUTHORS PROFILE

Pooja S.B After completing her post graduation degree she joined as a research scholar in Noorul Islam centre for higher education. Her area of interest is cloud computing, big data, data mining, and remote sensing.,

Dr .R.V Siva Balan he is working as a associate professor in MCA department. In the year 2013 from Anna University Coimbatore he took his doctorate degree. He guided more than 13 students in different fields. He has nearly 30 paper publications in Scopus and many SCI indexed journals. His interested fields are image processing, software engineering,, cloud computing, data mining and network security