International Journal of Engineering & Technology, 7 (2.32) (2018) 233-236

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



Data Analytics for Cardiotocography Data Using Principal Component Analysis

Pratuisha.K[1], Rajeswara Rao .D[2], J.V.R.Murthy[3]

[1,2]Dept of CSE, K L E F, Vaddeswaram, Guntur, Andhra Pradesh
[3]Computer Science Enginering, J.N.T.U,Kakinada, Andhra Pradesh, India
**Corresponding author E-mail:kpratuisha@gmail.com*

Abstract

With growing congenital anamelies in recent years detection of heart problems in fetus has become critical. Cardiotocography test assists doctors in such dignosis followed by cure. Here analytics of cardiotocography data is presented in details.Understanding ,cleaning and preprocessing the data is one of the the foremost part for any researcher,In this work data is cleaned,preprocessed,normalized, Also the attributes are selected by using the Chi-square test. Colinearity problem is addressed using Principle component analysis.Such analytics and prepro-cessing will help in machine learning or allied models for predict-ing precise diagnosis at an early stage

Keywords: *DataAnalysis,Chisquare,Normalization,PCA,*

1. Introduction

Cardiotocography is a specialized methods for recording the fetal pulse and the uterine constrictions in pregnancy. The machine used to play out the observing is known as a cardiotocograph[1], Many researchers have worked on the adult heart diseases,and only few on babies. So Cardiotocography data set is considered.Using analysis and perprocessing presented in this work better machine learning techniques [2]In this work the data analysis is done by the preprocessing of data is done by cleaning,Normalization and co-relation analysis and multicollinearity problem is addressed by applying Principle component analysis In the following section the literature data analysis is discussed and finally the conculsion of work.

2. Literature

The authour Johannes L. Grabmeier et.al describes the gini and pearson's chi-square measure,entropy and calualating the chisquare test for the decision tree[3]
The authour Meesad et.al Content characterization is the principle issue keeping in mind the end goal to help quests of advanced libraries and the Web. Most methodologies experience the ill effects of the high dimensionality of feature space, e.g. word recurrence vectors. To conquer this issue, feature selection technique is based on the use of the chi square test is utilized.[4]
The authour Meesad et.al Content characterization is the principle issue keeping in mind the end goal to help quests of advanced libraries and the Web. Most methodologies experience the ill effects of the high dimensionality of feature space, e.g. word recurrence vectors. To conquer this issue, feature selection technique is based on the use of the chi square test is utilized.[4]
The author Monsen et.al describes the Exploratory Data Analysis (EDA) as an approach for increasing comprehension and understanding about a specific dataset,keeping in mind the end goal to help and approve factual discoveries and furthermore to possibly produce, recognize, and make new speculations in view of examples in information. Cases of EDA are given and understandings are talked about. EDA might be utilized at any phase in the information investigation process from cleaning through change and elucidating examination, and also utilizing outcomes from each stage. Disclosure of examples may move another bearing inmediation adequacy explore, and additionally supporting or approving existingactivities. Information perception abilities are basic for people working with hugedatasets.[5]
In this study the author Teychene et.al has described about the five ground water resources,mainly the drinking water by scaling their chemical,physical properties and filter performance,the pca has been used for the feature selection of the water like organic matrix and fouling propensity the clustering analysis of water is done as three catagerioes like strong,low,and intermediate fouling.[6]
In this paper the author the author Saidi et.al has done the Principal Component Analysis on e-nose dataset and applied for Support Vector Machines (SVMs),Hierarchical Cluster Analysis (HCA) and Partial Least Squares-regression (PLSregression),and diagnosis to distinguish between breath of Chronic Kidney Disease, Diabetes Mellitus patients and healthy controls based on breath Volatile Organic Compounds s analysis.[7]

3. Data Analytics

The data analysis is the important part in developing the model,the analysis is done to clean up the data and make it in a structured format so that it make some sense for the model to learning about the data.The data analysis of two types one is Exploratory analysis and other is Confirmatory analysis ,Importantly the exploratory analysis is used for seeing what the data can tell us beyond the formal modeling or hypothesis testing task is an approach to analysis the data sets to summarize their main characteristics, often with visual methods.like histogram,box plots scatter plots,In this paper we are concerting on the Exploratory analysis[8].Outliners

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

should to be examined deliberately. Regularly they contain significant data about the procedure under scrutiny or the information assembling and recording process. Before thinking about the conceivable end of these focuses from the information, one should endeavor to comprehend why they showed up and whether it is likely comparative esteems will keep on appearing. Obviously,anomalies are regularly bad information points .Here in this paper the data is normalized and then we can find the outliners the dots above the box lines are the outlines.

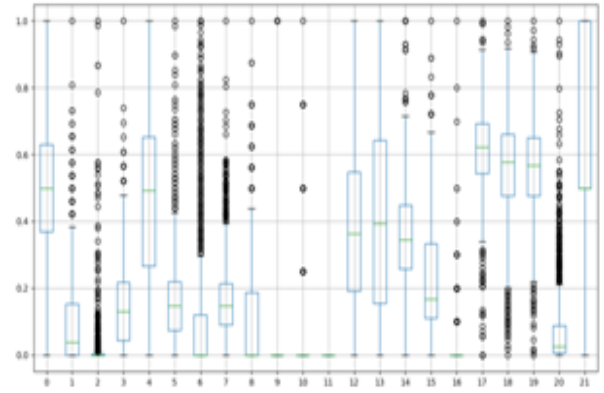


Fig 1: Outliners in data

3.1 Data Collection

In general , the clinicial data is collected in the standard format of CRF forms(Case Report Forms) i.e a structured format. The data is collected manually from reports of patients through retrospective approach, The Data that we considered is of Cardiotocography Data set which is of the featel heart beat ,we have obtained the data is from UCI website [9].They are 2126 instances and the 22 attributes and one classattribute which is a multi class classication label as Normal,Suspect,Pathologic.

3.2 Cleanng & Handling Missing Values

This is a very signicant and important part to handled for the real datasets .They are some columns and rows that the values will not be populated,The data can be cleanedl manually or by ignoring the segment or else by calculating the mean of the attribute [10],The data set that we considered consit of no missing values so the cleaning is not required for the the datasetconsidered.

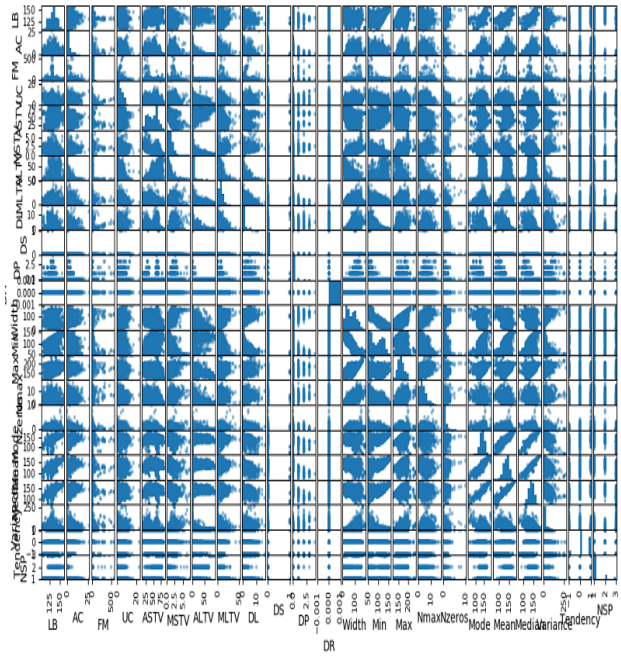


Fig:2: Attributes description using the scatter plot

Here The histogram for each attribute as shown in figure 2.After the data cleaning process is done as stated above, The data relational need to be observe between the data varibles,so by using the scatter plot ,The realtions are able to determined between them by scatter plot In the scatter plot we can observe the data having both the continious data and the categorical data the variables AC,ALTV,ASTV,DL are the continuous data ,DR is categorical data.

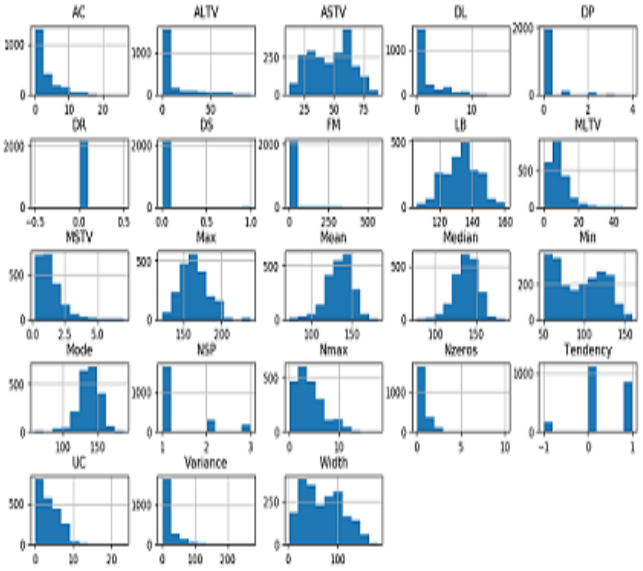


Fig 2: Cardiotocography Data Visualization

3.3 Data Analysis

A basic descriptive analysis of data involves calculation of simple measures of composition and the distribution of variables. Depending upon the type of data, these measures may be proportions, averages., normalization the attributes data is scaled to t in range from 0 to 1. it retains attribute's specific patterns [11]. Description of the data is in the table 1. There are high value attributes like LB ,Max ,Mode ,Mean ,Median. Also there are very low value attributes like Tendency, Nzeros, DP,DS. so because of the difference in the values the variance is high,so data is normalized , Data is well distributed and will form a good patterns

Attr	count	mean	std	min	25Per	50Per	75Per	max
LB	2126	133.3038570085	9.8408442577	106	126	133	140	160
AC	2126	2.7224835372	3.5608502149	0	0	1	4	26
FM	2126	7.2412982126	37.1253090398	0	0	0	2	564
UC	2126	3.6599247413	2.8470935097	0	1	3	5	23
ASTV	2126	46.9901222954	17.1928137186	12	32	49	61	87
MSTV	2126	1.332784572	0.8832413341	0.2	0.7	1.2	1.7	7
ALTV	2126	9.8466603951	18.3968796752	0	0	0	11	91
MLTV	2126	8.1876293509	5.6282466041	0	4.6	7.4	10.8	50.7
DL	2126	1.570084666	2.4992288114	0	0	0	3	16
DS	2126	0.0032925682	0.0572998389	0	0	0	0	1
DP	2126	0.1260583255	0.4643611013	0	0	0	0	4
DR	2126	0	0	0	0	0	0	0
Width	2126	70.4459078081	38.955692958	3	37	67.5	100	180
Min	2126	93.5794920038	29.5602122563	50	67	93	120	159
Max	2126	164.0253998119	17.9441831101	122	152	162	174	238
Nmax	2126	4.0682031985	2.9493856219	0	2	3	6	18
Nzeros	2126	0.3236124177	0.7060593732	0	0	0	0	10
Mode	2126	137.4520225776	16.3812892734	60	129	139	148	187
Mean	2126	134.6105362183	15.5935963326	73	125	136	145	182
Median	2126	138.0903104421	14.466588856	77	129	139	148	186
Variance	2126	18.8080903104	28.9776360084	0	2	7	24	269
Tendency	2126	0.3203198495	0.6108286301	-1	0	0	1	1
NSP	2126	1.3043273754	0.6143768486	1	1	1	1	3

Fig 3: Data analysis of an attributes

The description of data is seen in the fig:3 visulizing the data is more effective so we have made scatter plot for the data visualization , fron the fig:4 It is observed that their are both continoius and categorical varibles in the data and the data is widely distributed in the patterend form.

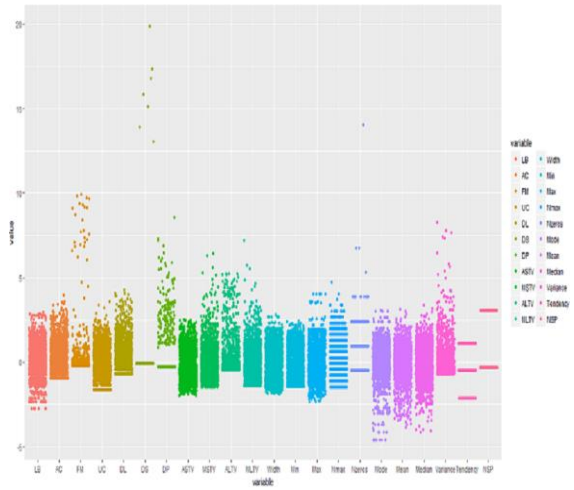


Fig 4: Cardiotocography Scatterplot

3.3.1 Feature Selection with Chi-Square(χ)

Feature selection will have a high impact on the developing models or the frameworks mainly with respective to the performance,Considering the Irrelevent or partiall relevent features in the data set can lead to the negative impact of the model performance,so while considering the feature's of the data we have see whether the variables are correlated to each other,The main outcome for doing the feature se-lection before the modeling the dataset are avoiding overting,Increasing accuracy and decreasing the training time.[12][13]The Chi-Square test for feature selection is applied on this data. Out of 21 independent attributes

Table1: Feature selection with chi-square with attributes Ranking

Attribute	Score Index	Rank
DP	306.223	1
ALTV	196.945	2
ASTV	58.56	3
DS	55.307	4
AC	51.983	5
Variance	43.87	6
DL	33.651	7
MIN	29.718	8
MSTV	21.839	9
Mean	16.849	10
LB	16.327	11
Width	13.327	12
Median	12.68	13
Tendency	12.092	14
Mode	11.932	15
UC	11.171	16
MLTV	10.055	17
FM	3.379	18
Nmax	2.846	19
Nzeros	0.676	20
Max	0.325	21

3.3.2 Co-Relation Analysis with Principle Component Analysis

To understand relation between different attributes for finding best suited ones is the prior thing. So, correlation analysis was carried out and their is the large varience is ploted in correlation, So the variance is to be reduced to zero for that Principle Component Analysis(PCA).Principle Component Analysis method is widely used dimensional reduction of compunets tahat are desired by the user, or principal components in the transformed result[14]. These components are combination of multiple features and covers maximum variance in dataset. Thus components is used to represent information contained in dataset with less number of dimensions in given data. First component always represents maximum variance. Later components are arranged in decreasing order of variance.the figure explains of the positive varience between the attributes that are between 0 and 1.,Here the analysis reveals interdependency between independent attributes termed as multicollinearity problem. In such case Principle Component Analysis provides best factors. Each factor involves multiple attributes with diferent weights

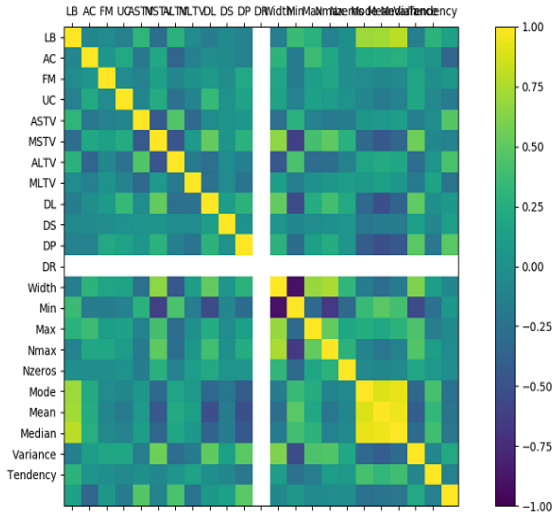


Fig:5 Co-Relation for all attributes

So the variance is to be reduced to zero for that Principle Component Analysis(PCA).Principle Component Analysis method is widely used dimensional reduction of compunets tahat are desired by the user, or principal components in the transformed result[14]. These components are combination of multiple features and covers maximum variance in dataset. Thus components is used to represent information contained in dataset with less number of dimensions in given data. First component always represents maximum variance. Later components are arranged in decreasing order of variance.the figure explains of the positive varience between the attributes that are between 0 and 1.,Here the analysis reveals interdependency between independent attributes termed as multicollinearity problem. In such case Principle Component Analysis provides best factors. Each factor involves multiple attributes with diferent weights. Factors generated from PCA are orthogonal to each other, so no correlation exists between them. shows attributes with their weights involved in first two factors of PCA [15]on given dataset. The figure describes how input attributes are related withgenerated factors from PCA, The graph explains corelation between the attributes of the componet one and componet two, were the most attributers are corelated

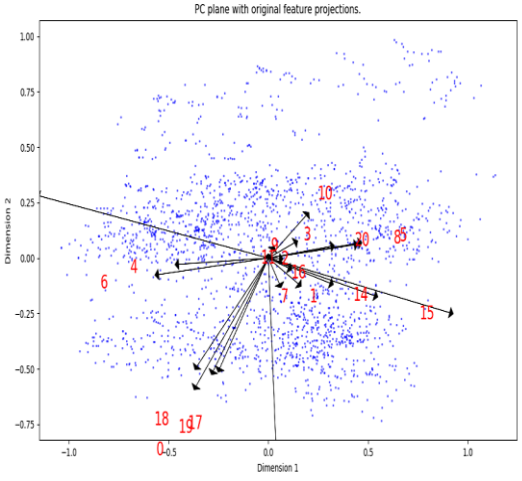


Fig:6: Bipolar graph

4. Conclusions

In this paper the data analysis is done on Cardiotocography data set by cleaning the data set also the data is been described by their mean, standard error , min ,max, and also the statatical calucation in done for the better understaning of the data. By the observation of the data their are difference in the correlation of attribute values so the Principle Component Analysis is used to reduce the varience among the attributes,the data is preprocessed and ready to feed to the any kind of classifer as an input for the classication.with out analysing the data developing the models lead to the imperfect results and also some times leads to the failure of the model.Hence before developing the model it is necessry that do the analysis of the data completly.

References

[1] Z.Alrevic, D. Devane, G. Gyte,et al.,"Continuous cardiotocography (ctg) as a form of electronic fetal monitoring (efm) for fetal assessment during labour,"Cochrane Database Syst Rev, vol. 3, no. 3, 2006..

[2] R. Li, \Data mining and machine learning methods for dementia research," in Biomarkers for Alzheimers Disease Drug Development, pp. 363-370, Springer, 2018..

[3] J. L. Grabmeier and L. A. Lambe, \Decision trees for binary classication variables grow equally with the gini impurity measure and pearson's chi-square test,"International Journal of Business Intelligence and Data Mining, vol. 2, no. 2, pp. 213-226, 2007.

[4] P. Meesad, P. Boonrawd, and V. Nuipian, \A chi-square-test for word importance differentiation in text classication," in Proceedings of International Conference on Information and Electronics Engineering, pp. 110-114, 2011..

[5] K. A. Monsen,Exploratory Data Analysis, pp. 77-85. Cham: Springer International Publishing, 2018..

[6] B. Teychene, A. Touet, J. Baron, B. Welte, M. Joyeux, and H. Gallard, \Predicting of ultraltration performances by advanced data analysis," Water research, vol. 129, pp. 365-374, 2018.

[7] T. Saidi, O. Zaim, M. Moud, N. El Bari, R. Ionescu, and B. Bouchikhi, \Exhaled breath analysis using electronic nose and gas chromatography{mass spectrometry for non-invasive diagnosis of chronic kidney disease, diabetes mellitus and healthy subjects," Sensors and Actuators B: Chemical, vol. 257, pp. 178-188, 2018.

[8] V. Cox, \Exploratory data analysis," in Translating Statistics to Make Decisions, pp. 47-74, Springer, 2017..

[9] D. Dheeru and E. Karra Taniskidou, \Uci machine learning repository," 2017..

[10] E. Rahm and H. H. Do, \Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3-13, 2000.

[11] J. F. Hair, W. C. Black, B. J. Babin, R. E. Anderson, R. L. Tatham,et al.,Multivariate data analysis, vol. 5. Prentice hall Upper Saddle River, NJ, 1998.

[12] J. R. Ummadi, B. V. R. Reddy, and B. E. Reddy, \A novel statistical feature selection measure for decision tree models on microarray cancer detection," in Proceedings of International Conference on Computational Intelligence and Data Engineering, pp. 229-245,Springer, 2018.

[13] T. Niu, J. Wang, K. Zhang, and P. Du, \Multi-step-ahead wind speed forecasting basedon optimal featureselectionandamodied bat algorithm with the cognition strategy,"Renewable Energy, vol. 118, pp. 213-229, 2018.

[14] Y. Du, J. A. Kitzmiller, A. Sridharan, A. K. Perl, J. P. Bridges, R. S. Misra, G. S.Pryhuber, T. J. Mariani, S. Bhattacharya, M. Guo,et al.,\Lung gene expression analysis(lgea): an integrative web portal for comprehensive gene expression data analysis in lung development,"Thorax, pp. thoraxjnl-2016, 2017.

[15] E. Mooi, M. Sarstedt, and I. Mooi-Reci, \Principal component and factor analysis," in Market Research, pp. 265{311, Springer, 2018.