NVEO
Natural Volatiles &
Essential Oils

Predictive Modeling For Classification Of Breast Cancer Data Set Using Feature Selection Techniques

S. Leena Nesamani[1] , S. Nirmala Sugirtha Rajini[2] , Ibeth Catherine Figueroa Sánchez[3] , María Del Pilar Melgarejo Figueroa[4] , Digna Amabilia Manrique De Lara Suárez[5] , Oscar Felipe Carnero Fuentes[6]

[1,] Research Scholar,Department of Computer Applications. Dr. M.G.R. Educational and Research Institute, Chennai, India

[2,] Professor, Department of Computer Applications. Dr. M.G.R. Educational and Research Institute, Chennai, India

[3]UniversidadNacional Hermilio Valdizan, Huánuco – Perú cfigueroa@unheval.edu.pe https://orcid.org/0000-0002-0440-2504

[4]Universidad Nacional Hermilio Valdizan, Huánuco – Perú mapimefi@gmail.com https://orcid.org/0000-0003-2837-2386

[5]Universidad Nacional Hermilio Valdizan, Huánuco – Perú dmanrique@uhneval.edu.pe https://orcid.org/0000-0003-4488-252X

[6]Universidad católica de Santa María -Arequipa –Perú ocarnero@ucsm.edu.pe https://orcid.org/0000-0002-4532-9710

Abstract:

Predictive modeling or predictanalysis is the process of trying to predict the outcome from data using machine learning models. The quality of the output predominantly depends on the quality of the data that is provided to the model. The process of selecting the best choice of input to a machine learning model depends on a variety of criteria and is referred to as feature engineering. The work is conducted to classify the breast cancer patients into either the recurrence or non-recurrence category.A categorical breast cancer dataset is used in this work from which the best set of features is selected to make accurate predictions. Two feature selection techniques namely the Chi squared technique and the Mutual Information technique have been used. The selected features were then used by the Logistic Regression model to make the final prediction. It was identified that the Mutual Information technique proved to be more efficient and produced higher accuracy in the predictions.

Keywords: Predictive modeling, Machine Learning, Feature Selection, Logistic Regression, Chi squared, Mutual Information, Breast Cancer, Categorical

Introduction:

Machine learning is the task of solving a problem based on the available data patterns instead of being programmed directly.Machine learning models are deployed in various processes such asclassification, regression, clustering, anomaly detection, ranking, and recommendations and forecasting.

Predictingthe class to which an instance of data falls into is known as classification. Classification algorithms operate on labeled data where each label determines the class or the category to which the data belongs to. Classification may be either binary classification or multi class classification. The former predicts the unlabeled data into either one of the two available classes and the later makes predictions among Nclass or category of labels.Regression on the other hand is the process of trying to predict a label which is a continuous value from related set of features. The regression algorithm works on a set of labeled features and uses a function to predict the value of an unlabelled data.

Clustering is the task of grouping instances of data into different groups based on their similarity. The individual groups are called clusters and the members of the cluster share similar characteristics which are specific to a particular cluster.Anomalies are rare or infrequent events or observations that are misleading and dissimilar from the rest of the observations. Anomaly detections help in identifying fraudulent transactions, finding abnormal clusters, identify patterns that exhibit network intrusion, outlier identification, etc.In Ranking,the labeled data are grouped into instances and assigned scores which are used by the ranker to assign ranks for the unseen instances.Recommendation refers to the task of recommending products or services to the user based on their historical data. Making future predictions based on the past time-series data is known as Forecasting.

Machine learning model are mostly based on predictive modeling. In predictive modeling the model is trained on historical data in-order to make predictions on the new unseen data. The performance of the machine learning model depends on the efficiency of the algorithm that is chosen to handle the problem. Machine learning modelsperform well when they are provided with the right data. Feature Selection procedures help a lot in this aspect. They help not only to reduce the computational cost but also improve the performance of the model as well. Feature selection or variable selection is the process of selecting a subset of variables or features from the total dataset to build machine learning models. It is the key to construct faster, simpler and reliable Machine Learning models. Simpler models are easier to interpret and have shorter training time. It is easier to understand the model that uses ten variables rather than a model that uses hundred variables. Reducing the number of variables also reduces the computational cost and speeds up model building. Feature Selection also enhances generalization and hence improves model overfitting. Often many of the variables are the noise with no or very little predictive value. The Machine learning models learn from this noise reducing generalization and causing overfitting. By eliminating this noise we can substantially improve generalization and reduce overfitting. Reducing the number of variables also reduces the data errors that may be incurred during data collection. Variable redundancy could be removed by selecting only the necessary features and removing the highly correlated feature without losing important information.

Feature Selection Methods:

Feature selection procedure involves a combination of search procedures which selects different subset of features and an evaluation measure that scores each subset of features[1]. This is computationally expensive and different subset of features may produce optimal performance for different machine learning models. This means that there is no one set of optimal features but different sets of optimal features based on the machine learning algorithm that is intended to be used.

Feature selection methods may be classified into supervised or unsupervised. This classification is based on whether the target variable is considered or not in the process of feature selection. Unsupervised methods does not consider the target variable in removing redundant variable using the correlation method. Supervised methods are used to remove the variables that irrelevant to the target variable. Methods like filter methods and wrapper methods are known to be supervised in nature.

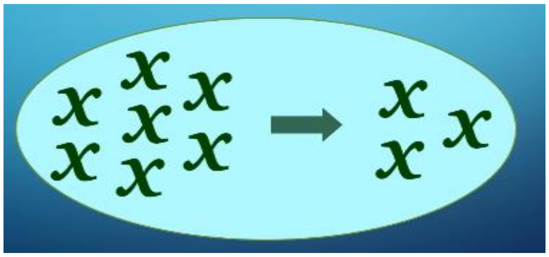


Fig 1: Feature Selection

Filter methods relies on the characteristics of the features themselves in the selection of variables. They do not involve any machine learning algorithm but simply rely only on the feature characteristics. Filter methods are model agnostic and are computationally inexpensive. But they tend to produce lower performance when compared to the other feature selection methods. On the other hand they are very well suited for quick screening and fast removal of irrelevant features from a data set.

Wrapper methods use a predictive Machine Learning algorithm to select an optimal feature subset. Wrapper methods build a machine learning algorithm for each of the subset and select the subset of features that produce the highest performance. This makes them computationally very expensive but tends to produce the best performing subset of features for the given Machine Learning algorithm. It also implies that the subset of selected features may not produce the optimum result for a different machine learning algorithm.

Embedded method is an unsupervised method that performs feature selection as part of the model creation process. The models contain built-in feature selection procedures that select and include only those variables that produce maximum accuracy. The embedded method considers the interaction between the features and the model.

Dimensionality reduction differ from the feature selection method by creating a new projection of data with a completely new set of variables in contrast to the feature selection methods discussed above which remove the variables from the dataset.

Choosing the best method:

Selecting the best method for the feature selection procedure is a million dollar question. There are no such rules as to select one. On the contrary we may move about in the selection process by looking at the variable type that we are intended to handle for the problem. The variables include both the input and output variables. This scenario is depicted in Fig 2, where the methods to be chosen for feature selection is based on the variable types involved.

When both the input and the output variables are numerical in nature the predictive modeling problem is a regression. The Pearson correlation coefficient technique is employed for a linear correlation. When the correlation appears to be nonlinear, the Spearman’s rank coefficient technique is deployed.When the input variable is numerical and the output variable is categorical in nature the predictive modeling problem is a classification. ANOVA correlation coefficient technique is employed when the correlation is linear and the Kendall’s rank coefficient is used in case of nonlinear correlation.

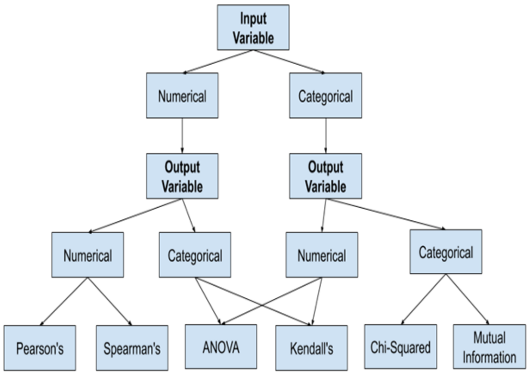


Fig 2: Choosing the correct method for feature selection

In some rare cases of regression where the input is a categorical variable and the output is numerical, the ANOVA method is used in the case of nonlinear coefficient and the Kendall's technique is for a linear correlation. When the predictive modeling problem employs categorical data for the input as well as the output the Chi-Square test and the Mutual Information techniques comes handy.

Literature Review:

Asim et al., [2] have extracted the texture features from the thermal images for the classification of the breast cancer. The texture features were extracted using Gabor filters at various orientation and scale levels. To extract the necessary features from the texture features the Gaussian modulated sinusoid that uniformly covered the spatial domain was used. The Gabor filters are a counterpart to performing wavelet transforms on an image at given spatial frequency domain. As a result twenty features were selected for the classification for breast cancer into cancerous or non cancerous.

In the work carried out by Gayathri et al [3] the authors have thoroughly studied the feature selection techniques and have concluded that online feature selections techniques will be the best method suited for high dimensional data that follow a sequential training strategy. Online classification Applications involve high dimensional data where batch learning feature selections methods cannot be employed directly. Since the available online feature selection techniques suffer due to the scalability issues in terms of the high dimensional data, this has inspires the author to explore various feature selection methods like the Scalable and Accurate Online Approach (SAOLA), the Online Streaming Feature Selection (OSFS) technique and the Scalable and Accurate Online Approach (SAOLA) for feature selection in huge datasets and a few of Hadoop based classifiers. The results obtained are far more better than the existing online feature selection techniques but have not catered to the scalability measures of high dimensional data which throws forth new research ideas relating to identifying new online feature selection technique that addresses the scalability issue on big data.

Satyabrata et al,[4] in their research of finding the quality of red and white wine, employed two techniques namely the Genetic Algorithm based feature selection technique and the Simulated Annealing based feature selection technique to select the important attributes from the original dataset that contribute to increase the quality of prediction. The selected features were input to a set of classifiers that were probabilistic, linear and nonlinear in nature. It was observed that the SVM classifier performed well with an accuracy of 95% to 98% on the simulated annealing based features selection technique.

Dhanya et al,[5] in their work of trying to maximize the prediction accuracy of breast cancer, have deployed two datasets namely the Wisconsin and the WDBC datasets. Recursive feature elimination, sequential feature selection, f-test and correlation were the four feature selection algorithms have been employed for selecting the optimal number of features that would maximize the accuracy of prediction. Naïve Bayes, Logistic regression, and Random forest were the list of classifiers used in this research work. The selected features were applied to each of the classifier and their performances being measured. Both the datasets were applied to the classifiers with and without feature selection. It was observed that the classifiers performed better with the reduced number of features rather than the original datasets. It was also experimented that Random forest outperformed the other classifiers in both the datasets. The final conclusion made was the filter method - f-test, improved the accuracy of the classifiers when compared with the rest of the feature selection methods and was best suited for smaller datasets (Wisconsin). And the wrapper method – sequential forward selection, made the classifiers perform well than the other feature selection methods and was best suited for larger data sets (WDBC).

In the breast cancer prediction work carried out by Quang etal, [6] as an initial step, data was prepared and made ready for classification through various preprocessing steps such as checking for missing values, class imbalance, normalization, correlation and the train/test split. In order to ensure better generalization of the solutions, feature selection techniques such as scaling and principal component analysis have been employed. This ensures that only the essential features are fed into the classifiers. Various classifiers were employed to evaluate the model and it was observed that four models namely, Logistics Regression, Ensemble Voting Classifier, SVM Tuning, AdaBoost gave accuracy over 98%. It was concluded that ensemble models gave the best performance in terms of recall, precision, F1 score, ROC-AUC and computational time.

Proposed Methods:

1. **Data Set:** The dataset used for this research work is the Breast Cancer dataset which isa multivariate categorical dataset containing a total of 286 instances with 9 attributes, among which some of them are linear in nature and the rest are nominal in nature. This dataset is given by the Oncology Institute of Oncology and is available at the UCI Machine Learning Repository.Patients who have recovered from breast cancer can be classified into two categories namely the recurrent and non recurrent based on whether they will be affected again or not.

   All the variables in the dataset are categorical in nature, and a few among them are ordinal and the rest are not.

2. **Methods :**

   A. Pearson's Chi-Squared Method:

   In the breast cancer dataset both the input variables and the target variable are categorical data and the problem here is a classification predictive modeling where it tries to classify the data into either the recurrent or a non recurrent class. The test for independence between the input variable and the target variable is obtained using the Pearson's Chi squared statistical method. A test statistics χ^2 is calculated between the observed and the theoretical values. The chi-squared test is used to test whether the distribution of the categorical observed variablesand the expected variables differ from each there. The value of the test statistics is obtained from the following formula:

(1)

Where,

χ^2 is the cumulative test statistics

O is the observed frequencies

E is the expected / theoretical frequencies

The Value of the statistics is interpreted as below:

If the value χ² is greater than a critical value then the null hypothesis is rejected and the result is significant which also interprets the variable to be dependent. Otherwise the value is insignificant and do not reject the null hypothesis in which case the variable is independent.The value can also be interpreted in terms of the p-value and a significant value (alpha) where variable is said to be dependent if the p-value is less than or equal to alpha and independent otherwise.

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)$$

B. Mutual Information Method:

Mutual information is the name given to Information Gain when it is deployed in the procedure of variable selection.In probability theory it is calculated as the statistical dependence between any two random variables.If (X,Y) are therandom variables then mutual information is given by

(2)

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

P(X,Y) is their joint distribution

P_X and P_Y are the marginal distributions

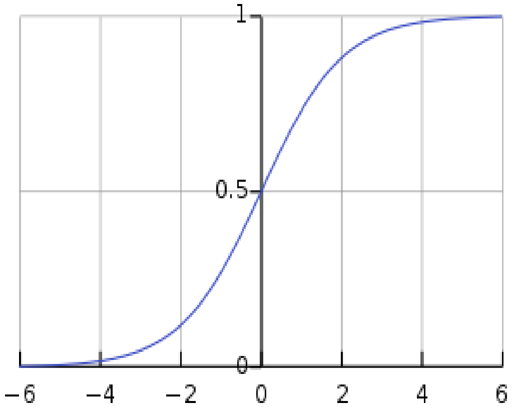
D_{KL} is the reference probability distribution or divergence

X and Y are independent when the information gain is zero and a non-negative value indicates they are dependent.

C. Logistic Regression:

Logistic regression is a machine learning technique that claims it origin from the field of statistics. It is a statistical model which uses a logistic function to estimate the values of a logistic model. The logistic function here is a sigmoid function which takes a real value (t) as input and outputs the value between 0 and 1.

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



(3)

Fig 3: A logistic function σ(t)

Regression analysis refers to the process of calculating the relation between a dependent variable and one or more independent variables by calculating the probabilities with a logistic function [7]. For a model having two predictor variables ($x_1$and $x_{2)}$ and one response variable Y, the relationship between them is given by the formula:

(4)

Where,

p, is the probability of the event

β_i is the model parameters

Once the values of the β_i are fixed then probability, Y=1 or Y=0 can be calculated.

Result and Discussion:

The breast cancer data set that is considered for this work consists of categorical data. The dataset is split into train and test sets to fit and evaluate the model. 67% of the data is used for training and 33% is used for testing. The categorical variables are encoded to integer using ordinal encoding. The target variable is label encoded to enable the binary classification. As a part of feature engineering the most relevant features are selected first using the chi-squared statistics method. It was identified that the third feature was the most relevant one. The top four features were selected for this work.

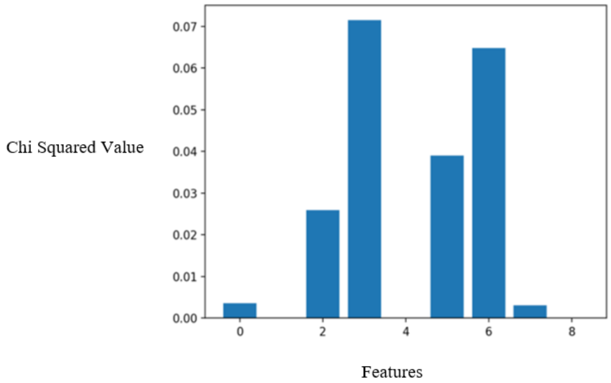


Fig 4 : Bar chart of features Vs the Chi squared value.

The mutual information method was applied next to select the most relevant features. It was observed that features 2,3,5 and 6 were the most relevant features.

A Logistic regression models was created to classify the patients into the recurrence and non recurrence class by including all the features in the dataset. Later two other logistic regression models were built using the features selected from the chi squared method and the mutual information method.

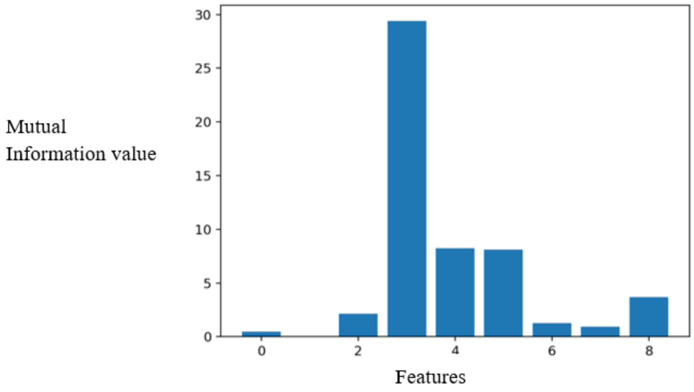


Fig 5: Bar chart of features Vs the Mutual Information value.

The results of the three models were compared. It was observed that the model which used all the features gave an accuracy of 75%, the model created with the features selected from the chi squared method yielded an accuracy of 74% and final model built the features extracted using the mutual information method produced an accuracy of 76%.

A Logistic regression models was created to classify the patients into the recurrence and non recurrence class by including all the features in the dataset. Later two other logistic regression models were built using the features selected from the chi squared method and the mutual information method. The results of the three models were compared. It was observed that the model which used all the features gave an accuracy of 75%, the model created with the features selected from the chi squared method yielded an accuracy of 74% and final model built the features extracted using the mutual information method produced an accuracy of 76%.

Table I. Comparison of feature selection methods

Classifier	Feature selection method	No. of features used	Accuracy %
Logistic Regression	No method	All	75
	Chi squared	2,4,5,8	74
	Mutual Information	2,3,5,6	76

Conclusion:

Feature engineering play a major role in predictive modeling. It is one of the major aspects that affect the accuracy of model. Best feature selection methods tend to produce good models. The choice of selecting the best method depends on multiple factors which include the characteristics of the features and the model being used. In this work, the focus was on using a categorical dataset and trying to employ the best feature selection method to obtain higher accuracy of prediction. It was observed that the best feature selection method for a model can only be identified by examining different subset of features from the dataset for the machine learning algorithm. The mutual information technique proved to give a better result for a categorical dataset on a regression problem. The work could be further extended to study the behavior of different feature selection techniques on numerical datasets and for classification problems by examining different subsets of features to fit different models.

References:

1. Azhar M.A., Princy Ann Thomas. Comparative Review of Feature Selection and Classification modeling, COMP-118-241-Ver-2©2019 IEEE
2. Asim Ali Khan1, AjatShatru Arora, Classification in Thermograms for Breast Cancer Detection using Texture Features with Feature Selection Method and Ensemble Classifier, 2019 2nd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), 978-1-7281-1772-0 ©2019 IEEE
3. S.Gayathri Devi, M.SabrigirirajFeature Selection, Online Feature Selection Techniques for Big Data Classification: - A Revie, Proceeding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India, 978-1-5386-3702-9/18 © 2018 IEEE 1
4. SatyabrataAich, Ahmed Abdulhakim Al-Absi, Kueh Lee Hui, Mangal Sain, Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques, ICACT Transactions on Advanced Communications Technology (TACT) Vol. 7, Issue 3, May 2018, ISBN 979-11-88428-02-1
5. Dhanya R, Irene Rose Paul, Sai Sindhu Akula, MadhumathiSivakumar,Jyothisha J NairA Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection, Proceedings of the International Conference on Intelligent Computing and Control Systems,(ICICCS 2019), IEEE Xplore Part Number: CFP19K34-ART; ISBN: 978-1-5386-8113-8
6. Quang H. Nguyen, Trang T. T. Do, Yijing Wang, Sin Swee Heng, Kelly Chen, Wei Hao Max Ang, Conceicao Edwin Philip, Misha Singh, Hung N. Pham, Binh P. Nguyen, Matthew C. H. Chua Breast Cancer Prediction using Feature Selection and Ensemble Voting, 2019 International Conference on System Science and Engineering (ICSSE), 978-1-7281-0525-3/19/$31.00 ©2019 IEEE
7. Liu Lei, Research on Logistic Regression Algorithm of Breast Diagnose Data by Machine Learning, 2018 International Conference on Robots & Intelligent System978-1-5386-6580-0/18/$31.00 ©2018 IEEEDOI 10.1109/ICRIS.2018.00049