A Comparative study of machine learning models for breast cancer prediction

Journal of Physics: Conference Series **1716** (2021) 012052 doi:10.1088/1742-6596/1716/1/012052

A Comparative study of machine learning models for breast cancer prediction

Afaf Athar[1], A K Ilavarasi[2*]

[1]School of Computer Science and Engineering,
Vellore Institute of Technology, Chennai, India

[2]Division of Healthcare Advancement, Innovation and Research,
Vellore Institute of Technology, Chennai, India
Email:*ilavarasi.ak@vit.ac.in[1]

Abstract. The enormous growth in the medical technologies and the availability of clinical data has motivated researchers to progress much towards predictive analytics. Integrating machine learning techniques to healthcare domain has a progressive outcome called Computer Aided Diagnosis. A comparative analysis is presented to study the suitability of machine learning algorithms for benchmark breast cancer data. Statistical non-parametric evaluation is also carried out to indicate the integrity of the framework.

1. Introduction

The early detection and accurate prediction of breast cancer increase the survival rate of women. Most breast cancer symptoms are not visible in the early stages and are detected only with a mammogram based professional screening. The national breast healthcare foundation [1] brings out a fact that only 5 percent of the affected women exhibit visible symptoms. Therefore, it is big challenge for the physicians which necessitate incorporating predictive machine learning models into healthcare. Danton et.al [2] addresses the ethical challenges involved while supplementing clinical diagnosis with machine learning results. The authors discuss the significance of quality indicators for assessing the machine learning models and the human biases in the decision making process. While the impact of machine learning in health care domain is prevalent , a false diagnosis has bigger implications on the reliability of the underlying model. A positive case called as “malignant” could be falsely predicted as negative (“benign”) and vice versa [3].

Characterization of data features aids in meta learning which identifies the potential learning model to be incorporated [4]. Therefore, identifying relevant feature subset is an important process in supervised learning. A feature subset selection method [5] focus on searching for good feature subset that contributes significantly towards the classification process. The vital contributions of feature subset selection methods are dimensionality reduction and increased computational efficiency. These methods tend to optimize feature subset and the degree of predictive relevance.

There are ample feature subset selection methods and predictive models available. This study emphasizes a guided procedure for comparing the performance of predictive models with respect to the data characterized feature selection methods. The section II focus on the related work, section III discusses the proposed methodology and section IV is organized with results based on data characterization and predictive performance.

2. Related work

As the proposed methodology is built around the feature subset selection method and its implication on the suite of machine learning models, a brief review on the relevant literature is presented.

[1] * Dr. Ilavarasi A K is associated with VIT Chennai as a teaching and research faculty. She is the corresponding author of this article

2.1 Data characterization and feature selection methods

According to [4], identification of meta features that will contribute to the specific learning task is the vital step for building a predictive model. Dash & Liu [6] have elucidated two criteria for evaluating a feature subset. The first artifact is that the classification accuracy obtained with the subset should not be less than the one obtained with the entire set of features. The second criterion is that the resulting class distribution from a feature subset should be as close as the original class distribution obtained with all features.

In general, the feature selection methods can be broadly classified into wrapper and filter models. The wrapper model is based on feedback from an induction algorithm to select a feature subset and filter model is independent of an induction algorithm for feature subset selection. Kohavi & John [7] have proposed the wrapper approach in which an optimal feature subset selection is intended for a specific learning algorithm. The search procedure exists as a wrapper around the inducer. It is a feedback method in which an induction algorithm like K-Nearest Neighbor (KNN) gets executed on candidate feature subset. The classification error rate obtained is used for evaluating the worth of feature subsets. Liu & Setiono [8] introduced a filter solution via chi-square correlation for feature subset selection. The filter method is a heuristic approach that makes an assessment on the merits of the features based on the general characteristics of data. Filters assess the worthiness of the feature subset independent of an induction algorithm. Far ahead, a novel feature selection algorithm based on correlation filters is presented in [9]. Rokach [10] has presented a feature set partitioning approach based on genetic search guided by KNN wrappers.

2.2 Machine learning models

Decision Trees is a popular model in supervised learning due to its simplicity. It is based on recursive partitioning approach, where the training instances are separated based on the splitting criterion on the attributes [11]. Different algorithms have been proposed based on decision tree induction, which include c4.5,ID3, SLIQ and SPRINT [12]. The difference lies in the splitting measures and tree pruning strategies.

Logistic Regression is the statistical model that uses the logistic functions to model the dependent variable in binary form. It models probability of output in terms of input using the sigmoid function given by equation (1),

$$f(x) = \frac{1}{1+e^{-c*(x-c)}} \tag{1}$$

It can determine the presence of the (event is existing) class variable. It is used when our dependent variable is binary. There are many applications employing this model, one such is the modelling of urban expansion pattern in metropolitan cities presented in [13].

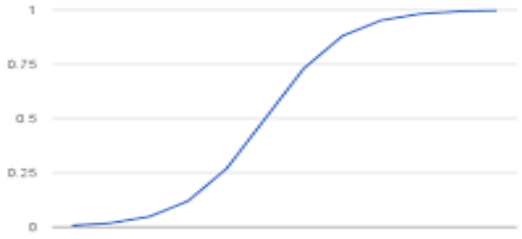

Figure 1 A sample Logit model

Support Vector machine is another powerful machine learning model widely applied in the diagnostic and prognostic analysis of breast cancer [14]. Recently, an advanced research on SVM brings out an ensemble SVM based on the weighted area under the Receiver Operating Characteristic Curve method [15].

3. Proposed methodology

The original dataset is divided into training and test set with a 5X2 cross validation procedure. Each partition serves as a training set in one iteration and test set in one iteration resulting in the evaluation of the algorithm twice. The 5×2 CV is preferably used rather than K-fold CV due to the acceptable range of Type-I error [16]. The training data is considered for the feature selection process. It is a two-step procedure where a Correlation based filter method is employed to select the features with greater predictive ability and are eligible candidates for Subset selection method. A Genetic K-Nearest Neighbour (KNN) wrapper validation is performed to evaluate the worthiness of a feature subset. Predictive models are built on algorithms like C4.5, logistic regression and Radial basis function Support Vector Machines (SVM). The accuracy of classifiers is assessed with the test set.

The KNN wrapper serves as the induction algorithm to evaluate a feature subset based on the classification accuracy. Genetic algorithms (GA) proposed by Holland [17] is a search technique derived from biological theory of evolution. A genetic based search is used in optimizing the results of KNN algorithm.

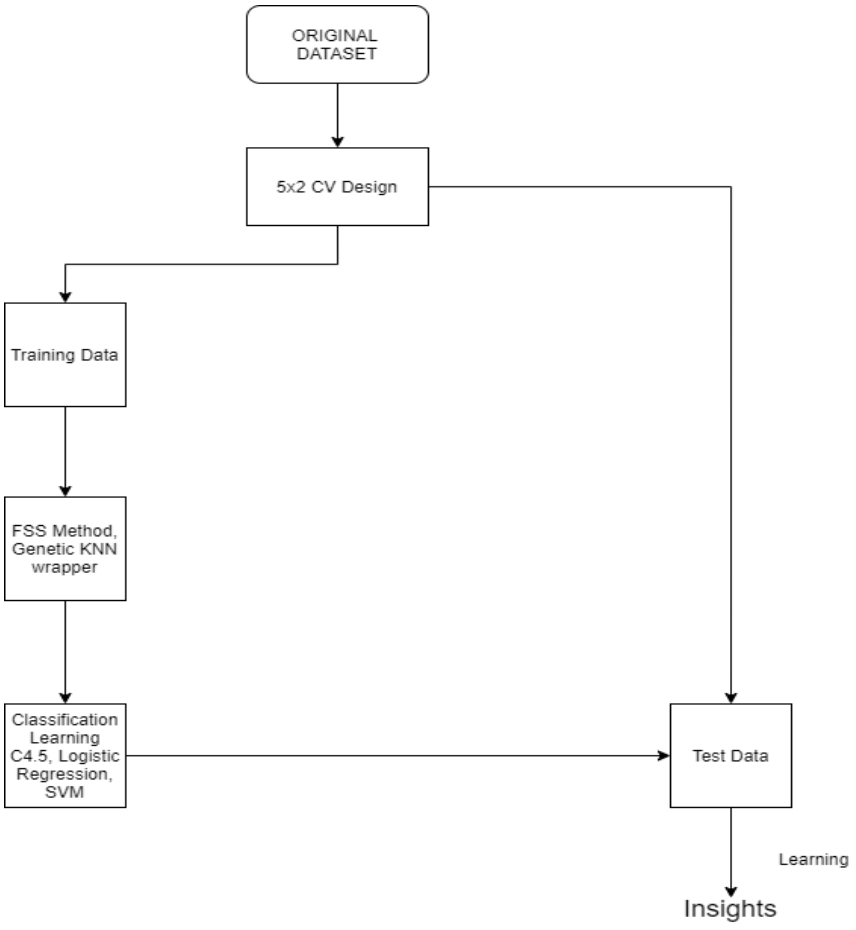

Figure 2 Proposed methodology

In general, GA requires a fitness function that assigns a fitness score to each candidate in the current population sample. The fitness of a candidate depends on the ability of the candidate to solve the problem at hand. Selection of candidates is performed randomly with a bias towards those with the highest fitness value. To avoid locally optimal solutions, crossover and mutation operators are introduced to produce new solutions along the whole search space.

Freitas [18] has identified GA, as a powerful tool for solving optimization problems through a series of genetic operations. Major challenges involved in using genetic algorithms are the number of details to define in run settings, such as the size of the population and the probabilities of crossover and mutation, and the convergence criteria of the algorithm. Specific values of parameters depend on the application employing a GA. The computational cost of GA might be controlled by appropriately choosing population size and stopping criterion.

A Steady State Genetic algorithm (SSGA) is employed in conjunction with KNN wrappers. The SSGA is a modest version of a generational Genetic algorithm which has rapid convergence properties. In SSGA procedure, two parents are selected from the population and the Selection of candidates from the population is based on their fitness scores with respect to accuracy of a specific feature set returned by the wrapper procedure. In order to circumvent locally optimal solutions, cross over and mutation operator is applied resulting in two best offspring.

4. Results and discussion

4.1 Dataset description

The dataset considered is the Wisconsin breast cancer data from the UCI machine learning Repositories [19]. It is described by 699 instances and 9 independent attributes which are transformed into discretized intervals and a class attribute with binary outcomes as Malignant(1) or benign(0):

ClumpThickness
cellSize
cellShape
marginalAdhesion
epithelialSize
bareNuclei
blandChromatin
normal Nucleoli
mitoses
class

4.2 Data characterization

A chi square correlation analysis is performed to identify the categorical features that are relevant in terms of predicting the target feature. The chi square statistics assesses the association between each of the independent features and the class feature via correlation ranking. It is observed from (1) that all the attributes have a correlation coefficient of above 0.7 and hence all the nine features participate in the subset selection process. The scatterplots are best to visualize the degree of correlation as indicated in Fig.3. The null hypothesis stating that there is no relationship between independent and dependent variables is rejected with a confidence interval of 0.95.

Table 1 Correlation statistics

Features	**Chi-square Correlation statistics with class feature**
clumpThickness	0.922
cellSize	0.837
cellShape integer	0.822
marginalAdhesion	0848
epithelialSize	0.886
bareNuclei	0.994
blandChromatin integer	0.972
normalNucleoli	0.721
mitoses integer	0.756

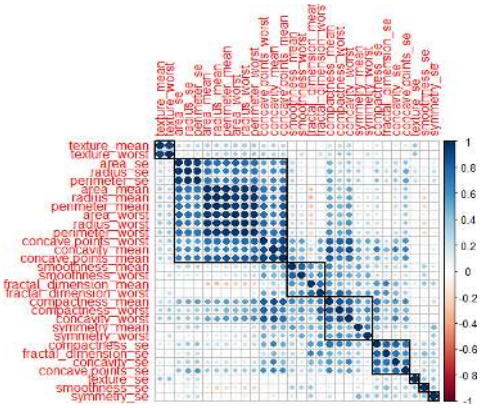

Figure 3. Scatter plot visualization

4.3 Genetic KNN wrapper procedure

The prime objective of this work is to embed the feature selection method with the classification task. An individual feature may be less relevant to the class feature whereas a subset of features may collectively contribute significantly to the classifier learning. By coupling the KNN induction with the genetic search, the classification process is naturally tied with the feature subset selection method. This implies that there is no post processing overheads to comply with the classifier models.

The goal of experimental evaluation is twofold:

- To investigate the robustness of the procedure with a broad suite of classification algorithms with the GA parameters (2). The objective function is to maximize the classification accuracy
- To empirically investigate the suitability of the framework on the candidate classification algorithms by non-parametric tests for ranking the classifier performance.

Table 2. Parameter setting for SSGA

Parameters	**values**
nEval	5000
Pop length	100
No.of features	3
ProbCrossover	0.6
ProbMutation	0.01
alfa	0.1

The classification workloads under study are C 4.5, logistic regression and RBF SVM. One representative algorithm is selected from each of the families of decision tree, Statistical classifiers and support vectors.

Table 3. Global average training results

Classifier models	**Accuracy (%)**	**Variance (%)**
C4.5	97.2	0.004
Logistic Regression	96.85	0.009
SVM (RBF)	97.55	0.002

Table 4. Global average test results

Classifier models	**Accuracy (%)**	**Variance (%)**
C4.5	93.25	0.04
Logistic Regression	95.26	0.01
SVM (RBF)	94.61	0.01

The global results (3,4) indicate the average accuracy across the 5 folds and that the logistic regression model exhibits the highest Predictive accuracy with the test data whereas SVM yields the highest accuracy with training data. As the cross validation is designed prior to the application of the proposed framework, the model variance is assessed with original data as the baseline.

A non-parametric evaluation is made by performing statistical tests of significance. Friedman test (5) was conducted to rank the performance and suitability of the proposal with respect to multiple classifiers and the null hypothesis is rejected with a confidence level of 0.95. The non-parametric results indicate the global average of classification accuracy and variance across the 5 folds and present a ranking order of logistic regression , SVM and C4.5 classifiers.

Table 5. Friedman's test for multiple classifiers

Classifier	**Global classification error (%)**
Logistic regression	4.73
SVM	5.38
C4.5	6.74

5 Conclusion

It is observed from (5-6) that there is a consensus between the prediction results and non-parametric statistical tests. This exhibits a positive synergy of the proposed framework towards the comparative

study of prediction models for breast cancer. The future scope of this work would be to integrate the knowledge obtained from diagnostic and prognostic clinical data for predicting the likelihood of breast cancer occurrence. This leads to early detection of the tumours improving the survival rate of affected women.

6. References

[1] National Breast Cancer Foundation Inc., http://www.nationalbreastcancer.org/about-breast-cancer
[2] Char DS, Shah NH, and Magnus D. 2018 Implementing machine learning in health care — addressing ethical challenges. *The New England journal of medicine*. **378(11)** 981.
[3] K. Al-Mashouq and Z. Nawaz, 2001 Characterization of machine learning benchmarking data sets 2001 *IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236), Tucson, AZ, USA*, pp. 3415-3419
[4] Castiello, Ciro, Giovanna Castellano, and Anna Maria Fanelli. "Meta-data: Characterization of input features for meta-learning." In International Conference on Modeling Decisions for Artificial Intelligence, pp. 457-468. Springer, Berlin, Heidelberg, 2005.
[5] Kira, K & Rendell, LA 1992 The feature selection problem: Traditional methods and a new algorithm *In Aaai,* vol. 2, pp. 129-134
[6] Dash, M & Liu, H 1997 Feature selection for classification *Intelligent data analysis*, **1(1-4),** pp. 131-156.
[7] Kohavi, R & John, GH 1997 Wrappers for feature subset selection', *Artificial Intelligence*, l. 97(2) pp. 273-324.
[8] Liu, H & Setiono, R 1995 Chi2: Feature selection and discretization of numeric attributes *In Tools with artificial intelligence, Proceedings Seventh international conference on IEEE*, pp. 388-391.
[9] Blessie, E. C., & Karthikeyan, E. 2012. Sigmis: A feature selection algorithm using correlation based method. *Journal of Algorithms & Computational Technology* **6(3),** 385-394
[10] BĂDULESCU, L. A. 2006. Data mining algorithms based on decision trees. Annals of the Oradea University. Fascicle of Management and Technological Engineering, 1583-0691.
[11] Alsharif, A. A., & Pradhan, B. (2014). Urban sprawl analysis of Tripoli Metropolitan city (Libya) using remote sensing data and multivariate logistic regression model. *Journal of the Indian Society of Remote Sensing*, ***42*(1),** 149-163.
[12] Sweilam, N. H., Tharwat, A. A., & Moniem, N. A. 2010. Support vector machine for diagnosis cancer disease: A comparative study. *Egyptian Informatics Journal*, **11(2)**, 81-92.
[13] 17. Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research*, **267(2),** 687-699.
[14] Alpaydin, E & Mayoraz, E 1999, 'Learning error-correcting output codes from data'.
[15] Holland, JH 1992 Genetic algorithms. *Scientific american*, **267(1)**, pp. 66-73.
[16] Freitas, AA 2005, 'Evolutionary algorithms for data mining', In *Data mining and knowledge discovery handbook* (pp. 435-467). Springer, Boston, MA
[17] Dua, D. and Graff, C. 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.