Integrating Domain Knowledge in Supervised Machine Learning to Assess the Risk of Breast Cancer

Aniket Bochare, Aryya Gangopadhyay, Yelena Yesha, Anupam Joshi, Yaacov Yesha
University of Maryland Baltimore County, Baltimore, USA
(aniketb1, gangopad, yeyesha, joshi, yayesha)@.umbc.edu

Mary Brady
National Institute of Standards and Technology, Gaithersburg, USA
mary.brady@nist.gov

Michael A. Grasso
University of Maryland, School of Medicine, Baltimore, USA
mgrasso@umem.org

Napthali Rishe
Florida International University, Florida, USA
rishen@cis.fiu.edu

Abstract—Breast cancer is the most common form of cancer in women. Breast cancer comprises 22.9% of invasive cancers in women and 16% of all the female cancers. Our study takes into consideration data of postmenopausal women of European descent and their single nucleotide polymorphism (SNP) information to assess the risk of developing breast cancer. We have used various supervised machine learning and data mining techniques to generate a model for predicting risk of breast cancer in postmenopausal women using genomic data, family history, and age. In this paper we propose an approach to select 9 best SNPs using various feature selection algorithms and evaluate binary classifiers performance. We evaluated the performance of binary classifier by adding the domain knowledge of 11 SNPs into the training set and performing classification based on most informative features obtained from feature selection technique. We have also designed an algorithm to incorporate domain knowledge into our machine learning model. Our observations revealed that the machine learning model generated using both the domain knowledge and the feature selection technique performed better compared to the naive approach of classification. It is also interesting to note that, in addition to selecting 9 best SNPs, feature selection resulted in removing age from the set of features to be used for cancer risk assessment, and the machine learning model generated using both feature selection and domain knowledge provided improved performance without using age for prediction, compared with the naive method that did use both age and family history among the features. We could observe improvement in the accuracy and sensitivity values using domain knowledge learning which could be beneficial for initial screening.

Keywords—Breast Cancer, SNP, Genome-Clinical Classification, Domain Knowledge, Feature Selection

I. INTRODUCTION

According to Fletcher [1] each year in United States about 210,000 women are diagnosed with breast cancer. The risk of developing breast cancer varies from person to person depending on risk factors. Breast cancer can also occur in women with no observable signs of risk factors. Moreover, the risk of getting breast cancer is higher for women with strong family history. In addition, a breast cancer gene increases the likelihood of getting breast cancer more than any other risk factors. There are many environmental and clinical factors such as older age, family history, race, radiation exposure, density of breast, nulliparity, breast feeding, hormone replacement therapy, weight, etc. which increase a person's risk of developing breast cancer [1].

Cancer is a complex and a deadly disease, and its detection in early stages could help to improve the probability of survival. Therefore it is imperative to research the contribution of SNPs in early disease prediction. This will assist doctors in assessing the likelihood of developing breast cancer and in deciding whether order further testing.

In this study we have used various data mining and supervised machine learning techniques for generating a prediction model capable of distinguishing between cases and controls for initial screening. We present statistical analysis of 3 different methods named Naive SNP Selection Approach, Feature Selection Approach and Domain Knowledge Integration Approach. From our observation we could conclude that addition of domain knowledge of SNPs in machine learning procedures was beneficial.

A. SNPs and Personalized Medicine

According to Kong, et al. a single nucleotide polymorphism (SNP) is a location in the human genome which differs from one person to another and may affect the functions of the gene in which it is found. Researchers are trying to understand SNPs due to varying susceptibility of...
individuals to various diseases. Much attention was received by SNPs as genetic markers since different patients responded differently to various drugs. Hence, researchers are exploring SNPs to provide personalized drugs to individuals depending on their genetic makeup [2].

Engel, Simpson, and Landers address the contribution of SNPs to cancer development [5]. Onay, et al. highlight the fact that SNPs belonging to certain genes increases the susceptibility to breast cancer [3]. Our goal is to use breast cancer associated SNPs as genetic markers for classifying an individual as case or control. But it has been observed in the past that the use of SNPs only as features to develop a prediction model has not yielded satisfactory performance. In this paper we have identified 22 SNPs from SNPedia [6] and use domain knowledge of SNPs to come up with an improved prediction model.

Khoury M. J., et al. have shown that in complex diseases the disease susceptibility may vary with gene-environmental interactions and genes originating from diverse demography [7]. Therefore, we need to select features wisely for diagnosis of such diseases. Hence, we chose 17 SNPs for classification algorithms after initial filtering and pre-processing.

McCarthy, et al. focus on the importance of combinations of SNPs instead of a single SNP in the development of Type II Diabetes [8]. Due to cumulative effect of SNPs, we selected a set of risk associated SNPs for determining genomic risk of an individual. We used 3 different methods in our experiments and considered the cumulative effect of these SNPs to generate a prediction model. In the naive SNP selection approach, we cumulatively used 17 SNPs for classification. In the second method we used feature selection to extract 9 most informative SNPs and used them cumulatively for generating a classification model. In the third method we used 11 SNPs which had risk values associated with them in SNPedia to incorporate domain knowledge into the model.

II. BACKGROUND AND RELATED WORK

A. Genetics and Breast Cancer

SNPedia provides references to different studies conducted by researchers on breast cancer patients and SNPs associated with breast cancer [9]. Therefore, we hope that by delving into the genomic patterns of a population one can detect risks at an early stage and assist physicians.

B. Bioinformatics and Medicine

Bioinformatics is a field where various data mining and machine learning techniques are used for disease prognosis and drug interactions. Many classification approaches have been proposed earlier for disease prediction, such as: Decision Trees (J48), k Nearest-Neighbor (kNN), Naive Bayes (NB), Random forest (RF) and Support vector machine (SVM). These methods had been applied in various fields such as: decisions involving judgment, screening images, load forecasting, marketing and sales and medical diagnosis [10].

C. Domain Knowledge and Machine Learning

According to Yu, et al. auxiliary information about a learning task which can be obtained from credible sources or domain experts is called domain knowledge. They explain that prior domain knowledge helps in selection, initial sanitization and pre-processing tasks involved in machine learning. This is not limited to removal of noise or redundancy but also transforming data using domain knowledge for inputting into our machine learning system. Adding virtual samples to the training set has gained much attention in recent times and it is important when there are not enough training examples to learn from [11]. Partha Niyogi, et al. discusses incorporating domain knowledge by using virtual examples [23] in the learning task [12].

We know that traditional machine learning algorithms do not take into consideration the knowledge about data for training classifiers. Sun, et al. point out that combining prior domain knowledge with training set aids in machine learning. They have demonstrated a novel approach of combining domain knowledge into support vector machine for better efficiency [13]. There are various ways to integrate domain specific information depending on its context and type. Positive domain knowledge is said to have occurred when the use of domain specific information results in a more accurate hypothesis compared to the use of just training examples [11]. We designed an algorithm to incorporate the risk associated with each SNP into the training set for integrating domain knowledge in our model and add virtual examples based on some rules.

III. CLINICAL-GENOMIC DATA

A. Data Description

In our experiments we have used the 'Nurses Health Study (NHS)-GCEMS Stage-1’ breast cancer data from dbGaP. The dataset contains post menopausal women of European ancestry out of which 1145 are cases and 1142 controls.
handle database. We have used the popular machine learning tools WEKA [14] and MATLAB [15] for our experiments. Figure 1 shows the framework of our project.

IV. CLASSIFICATION AND VALIDATION

A. Classification Algorithms

1) Decision trees: Decision trees are classification trees used in statistics and machine learning to predict a target value of a class based on the attributes or feature space. The leaves represent the classification and the branches represent sets of features which lead to classification. [16] Demonstrates use of the C4.5 decision tree algorithm for classifying patients based on genes and clinical information. J48, a java open source implementation of C4.5 algorithm available in WEKA, was used in our experiments.

2) Naive Bayes: Naive Bayes classifier is used to solve a classification problem based on a probabilistic framework that classifies new samples assuming conditional independence among features under consideration. To classify a new sample, one uses Bayes Rule:

\[
P(\text{class}=Y | \text{data}=X) = \frac{P(\text{data}=X | \text{class}=Y) \times P(\text{class}=Y)}{P(\text{data}=X)}
\]

Naive Bayes has been used before in predicting the risk of breast cancer susceptibility from multiple SNPs. [17] demonstrates accuracy of 56% as compared to the baseline of 50%.

3) Bayesian Networks: Bayesian Network is a probabilistic model of relationships and predictions. Bayesian networks are used widely in the medical field to support prognosis and diagnosis by experts for predicting the outcome of an unknown event. We chose Bayesian Network for our experiments because it analyses dependencies among all the variables through relationships. Bayesian Networks are very powerful and were used in the medical domain for diagnosis and treatment of breast cancer in the past [24].

4) Support Vector machine - SVM: Cortes and Vapnik developed the SVM, a supervised learning approach which helps to predict the labels of the test samples from set of positive and negative training samples [18]. SVM attempts to establish a maximum margin for finding the best hyper plane to separate positive and negative samples in Euclidean space. The problem of handling real world data where samples are not linearly separable is handled by choosing a kernel [19]. We used an open source SVM library called LibSVM [20] for our experiments. In this research we have used polynomial kernel of degree 3 and Radial basis function (RBF) for classifying unlabelled instances into cases and controls.

B. Validation and Accuracy

1) 10-fold cross validation: The testing and validation was carried out using 10-fold cross validation [10].

2) Receiver Operating Characteristics: Area under ROC is used widely in machine learning and data mining. ROC plots the true positive (TP) rate against the false positive (FP) rate [21]. It is a measure used in machine learning to predict binary classifier’s performance.

C. Sensitivity and Specificity

The sensitivity (SN) and specificity (SP) are two of the statistical measures used to evaluate a binary classifier. Sensitivity is used in machine learning to measure the proportion of positive instances classified correctly by the model. Similarly, specificity is a proportion of negative instances classified correctly by the model. The following formulae can be used to calculate the specificity and sensitivity.

\[
\text{Sensitivity} = \frac{TP}{TP+FN} \\
\text{Specificity} = \frac{TN}{TN+FP}
\]

V. METHODS AND PROCEDURES

A. Classification Using Naive SNP Selection from literature

Table I shows 17 SNPs considered for our experiments which were obtained from SNPedia. Table II shows the results obtained after performing classification using 17 SNPs as features. This method is called naive approach since the SNPs were neither prioritized nor assigned any weights. We have used 10-fold cross validation to test our model.

<table>
<thead>
<tr>
<th>Gene</th>
<th>SNP</th>
<th>Risk Allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>rs1799966, rs15942</td>
<td>G</td>
</tr>
<tr>
<td>BRCA2</td>
<td>rs144488</td>
<td>T</td>
</tr>
<tr>
<td>BRCA2</td>
<td>rs3817198, rs4987117</td>
<td>T</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>rs3731239</td>
<td>T</td>
</tr>
<tr>
<td>FGFR2</td>
<td>rs1981579, rs242046</td>
<td>T</td>
</tr>
<tr>
<td>TNRC9</td>
<td>rs3803662</td>
<td>T</td>
</tr>
<tr>
<td>CENPF</td>
<td>rs438034</td>
<td>G</td>
</tr>
<tr>
<td>RH1</td>
<td>rs2054344</td>
<td>G</td>
</tr>
<tr>
<td>LUM</td>
<td>rs2298578</td>
<td>T</td>
</tr>
<tr>
<td>TCF7L2</td>
<td>rs1255372</td>
<td>T</td>
</tr>
<tr>
<td>LSP1</td>
<td>rs3817198</td>
<td>T</td>
</tr>
<tr>
<td>CCNE1</td>
<td>rs597669</td>
<td>A</td>
</tr>
<tr>
<td>CDKNB1</td>
<td>rs34330</td>
<td>T</td>
</tr>
<tr>
<td>2q35</td>
<td>rs13387042</td>
<td>A</td>
</tr>
</tbody>
</table>
B. Classification Using Feature Selection

Feature selection and feature extraction are dimensionality reduction techniques which are mostly used to preprocess the data. They help to reduce the number of features under consideration by eliminating irrelevant ones. Many times it is necessary to narrow down the number of features under consideration for efficient classification. The removal of irrelevant features helps to improve classification accuracy in most of the cases. G. Ustunkar, et al. highlight the importance of selecting a subset of the available SNPs for conducting association studies [22]. Hence, we provide data of 17 SNPs, family history, and age as input data to feature selection techniques for selecting a subset of informative features. Feature Selection (FS) approach is used here to find a subset of the features to improve data quality and remove noisy data.

We used 3 techniques named Filtered Attribute Evaluation, Gain Ratio Attribute Evaluation and Information Gain Attribute Evaluation available in WEKA to extract 9 most informative SNPs from the dataset for binary classification. Table III shows the 9 SNPs obtained using 3 feature selection techniques and table IV shows the results of binary classification using these 9 SNPs. We have used 10-fold cross validation in method I and method II.

### Table III

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Accuracy</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 - Decision Tree</td>
<td>52.30%</td>
<td>0.538</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>55.05%</td>
<td>0.557</td>
</tr>
<tr>
<td>LibSVM (Radial Basis)</td>
<td>53.41%</td>
<td>0.530</td>
</tr>
<tr>
<td>LibSVM (Polynomial)</td>
<td>52.95%</td>
<td>0.530</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>54.27%</td>
<td>0.566</td>
</tr>
</tbody>
</table>

### Table IV

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Accuracy</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 - Decision Tree</td>
<td>54.92%</td>
<td>0.559</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>56.39%</td>
<td>0.571</td>
</tr>
<tr>
<td>LibSVM (Radial Basis)</td>
<td>56.94%</td>
<td>0.573</td>
</tr>
<tr>
<td>LibSVM (Polynomial)</td>
<td>54.57%</td>
<td>0.562</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>55.83%</td>
<td>0.571</td>
</tr>
</tbody>
</table>

C. Classification Using Domain Knowledge Addition

We designed the following Algorithm for adding domain knowledge:

**Algorithm I:** Add Virtual Instances to the Original Dataset.

**Require:** <SNP₁…SNPₙ, age, familyhistory, case> {Original Training Set}. n > 0 {Risk associated SNPs}. R₁ₓ & R₂ₓ {risk values for medium & high risk SNPs from SNPedia}.

1: Assign ‘0’ - norisk, ‘1’ - mediumrisk, ‘2’ - highrisk SNP.
2: ’SNP’ₓ = (0; 1; 2) {depends on number of risk alleles}.
3: for SNP₁ to SNPₙ do
   4: Let C₁ₓ & C₂ₓ <-row-count and R₁ₓ & R₂ₓ<-risk-value when SNPₓ = 1 & SNPₓ = 2 respectively.
   5: Add total of C₁ₓ *(1- R₁ₓ) & C₂ₓ *(1- R₂ₓ) virtual instances where SNPₓ = 1 & SNPₓ = 2 respectively.
   6: Add a random row vector Vₓ as follows: < SNP₁…SNPₙ, age, familyhistory > where SNPₓ = 1 & SNPₓ = 2 queried from the original training set.

Let X₁ and X₂ be number of virtual controls assigned to class-label of row-vector Vₓ where SNPₓ = 1 & SNPₓ = 2 respectively.

\[ X₁ + R₁ₓ * X₁ = C₁ₓ \]  \hspace{1cm} (4)
\[ X₂ + R₂ₓ * X₂ = C₂ₓ \]  \hspace{1cm} (5)

Solve for X₁ and X₂. The number of virtual cases assigned to the class-label of row-vector Vₓ are R₁ₓ * X₁ & R₂ₓ * X₂ where SNPₓ = 1 & SNPₓ = 2 respectively.

7: end for

We repeat this procedure for all the 11 SNPs selected from SNPedia which have risk associated values. 8 SNPs out of these 11 SNPs were selected, which overlap with the SNPs obtained using feature selection technique along with family-history for classification purpose. We train the classifier using the combination of original and virtual training samples. The validation is conducted on randomly
selected 20% test samples from original dataset. The mean results of classification after addition of virtual instances to the dataset across 10 trials are shown in table V. By comparing table II and V we can see around 6-8% increase in prediction accuracy which shows that domain knowledge was helpful. The deviation in the accuracy seen in table V is around ±2.5% across 10 trials.

TABLE V
CLASSIFICATION RESULTS USING BOTH DOMAIN KNOWLEDGE AND FEATURE EXTRACTION

<table>
<thead>
<tr>
<th>Classification Algorithm</th>
<th>Accuracy</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48 -Decision Tree</td>
<td>60.56%</td>
<td>0.591</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>60.12%</td>
<td>0.574</td>
</tr>
<tr>
<td>LibSVM(Radial Basis)</td>
<td>58.93%</td>
<td>0.53</td>
</tr>
<tr>
<td>LibSVM(Polynomial)</td>
<td>59.79%</td>
<td>0.535</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>59.85%</td>
<td>0.588</td>
</tr>
</tbody>
</table>

From our observations we can conclude that although there is an improvement in accuracy, there isn’t significant improvement in ROC area using domain knowledge integration to the machine learning model. In this experiment, we also calculated values for statistically important parameters like specificity and sensitivity for all the 5 algorithms used to evaluate the performance of binary classifier. The table VI shows the values of sensitivity (SN) and specificity (SP) for all the 3 methods.

TABLE VI
SENSITIVITY AND SPECIFICITY RESULTS ACROSS 3 METHODS USING 10FOLD CROSS VALIDATION

<table>
<thead>
<tr>
<th>Sensitivity and Specificity Comparison</th>
<th>Method I</th>
<th>Method II</th>
<th>Method III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Algorithm</td>
<td>SP</td>
<td>SN</td>
<td>SP</td>
</tr>
<tr>
<td>J48 -Decision Tree</td>
<td>0.566</td>
<td>0.536</td>
<td>0.579</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.579</td>
<td>0.519</td>
<td>0.603</td>
</tr>
<tr>
<td>LibSVM(Radial Basis)</td>
<td>0.531</td>
<td>0.530</td>
<td>0.609</td>
</tr>
<tr>
<td>LibSVM(Polynomial)</td>
<td>0.664</td>
<td>0.414</td>
<td>0.750</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td>0.548</td>
<td>0.549</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Comparing specificity and sensitivity values of Method III with Method I or Method II, we can observe a marked difference in the sensitivity and specificity values. This demonstrates that addition of virtual instances or domain knowledge into the model helps to increase the sensitivity of a test. Breast cancer is a deadly disease and a highly sensitive test is considered very important. Along with an increase in sensitivity there is a simultaneous decrease in specificity using the method III prediction model. The results obtained using domain knowledge model demonstrates a balanced tradeoff between sensitivity and specificity in case of J48, Naive Bayes and Bayesian Network classifiers. This fact, and the fact that using domain knowledge resulted in improved accuracy, are both important.

VI. CONCLUSION

In this study 3 analytical methods were compared across 4 classification algorithms. Validation tests were performed to evaluate the classifier’s performance using 10-fold cross validation for method I and method II. Percentage split method was used to validate the classifier developed using method III. The methods were evaluated based on performance parameters like area under ROC, accuracy and statistically important attributes like specificity and sensitivity.

The initial method I called naive SNP selection was explored and the results obtained were unsatisfactory. Secondly, we could see marginal improvement in the accuracy by carrying out binary classification using just feature selection technique. We observed improvement in performance using domain knowledge of 11 SNPs in the prediction model. We could see around 6-8% increase in the accuracy and marginal improvement in the area under ROC values using domain knowledge. These experiments were conducted across 10 iterations and the observed deviation in the accuracy was around ±2.5%. Interestingly, in addition to improved accuracy, high sensitivity and lower specificity values were observed in the model developed using domain knowledge. For example, when J48 Decision Tree was used, the model developed using domain knowledge had improved accuracy (60.56%), 0.798 sensitivity, and 0.301 specificity, which can be useful for initial screening. Hence, we can conclude that the model generated using domain information of SNPs can be helpful for assessing the risk of breast cancer in European women.

VII. ACKNOWLEDGMENT

The authors of this paper would like to thank The National Institute of Standards and Technology (NIST) for funding the project. We would also like to thank Database of Genotypes and Phenotypes (dbGaP) for providing access to the breast cancer data for this research.

VIII. REFERENCES

http://dx.doi.org/10.1016/j.neucom.2010.05.007
http://doi.acm.org/10.1145/1102351.1102460.
[16] Shweta Kharya: Using data mining techniques for diagnosis and prognosis of cancer disease, CoRR abs/1205.1923; (2012)