International Journal of Research in Engineering, Science and Management
Volume-2, Issue-1, January-2019
www.ijresm.com | ISSN (Online): 2581-5792

Particle Swarm Optimization for Breast Cancer Recurrence Prediction Using Data Mining Techniques

K. Vinitha[1], R. Kavitha[2]
[1]*M.E. Student, Dept. of Computer Science and Engg., Parisutham Inst. of Tech. and Sci., Thanjavur, India*
[2]*Associate Professor, Dept. of Computer Science and Engg., Parisutham Inst. of Tech. and Sci., Thanjavur, India*

***Abstract*: Women who have recovered from breast cancer always fear its recurrence. The fact that they have endured the painstaking treatment makes recurrence their greatest fear. However, with current advancements in technology, early recurrence prediction can help patients receive treatment earlier. The availability of extensive data and advanced methods make accurate and fast prediction possible. This research aims to compare the accuracy of a few existing data mining algorithms in predicting breast cancer recurrence. It embeds particle swarm optimization as feature selection into three renowned classifiers, namely, naive Bayes, K-nearest neighbor, and fast decision tree learner, with the objective of increasing the accuracy level of the prediction model.**

***Keywords*: Breast cancer; recurrence; feature selection; REPTree; naïve Bayes; K-nearest neighbor; particle swarm optimization.**

1. Introduction

Today, cancer is the primary cause of death around the globe. As stated by Siegel, breast cancer (BC) will continue to be the most prevalent cancer in women. Every woman is at risk for breast cancer. If she is 85 years old, there is a one in eight chance (12%) that she will develop breast cancer once during her life. In 2010, breast cancer was ranked the ninth leading cause of death in the Kingdom of Saudi Arabia (KSA). Before that, in 2009, it was reported that there were 1,308 new breast cancer cases, representing 25% of registered new cancer cases among Saudi women. It was forecasted that this disease incidence would increase over the coming decades in KSA as the population grows and ages. It is also notable that obesity; having a first child at a late age, a young age at menarche, a short period of lactation, or an unhealthy lifestyle; and geographical, racial and ethnic characteristics are risk factors contributing to the cause of breast cancer . Previous research indicated that the characteristics of this disease are high aggressiveness, poor clinic pathologic features, and early onset among the Saudis. Studies also reported that the advanced stage of breast cancer disease was found to be more prevalent in younger women with a median age of 47 years than in older women with a median age of 63 years in industrialized nations.

From the perspective of breast cancer behavior, recurrence of breast cancer refers to the reoccurrence of breast cancer in a patient whose previous cancer had gone into remission. Remission is the desired result of chemotherapy and continual treatment by oncologists. Recurrence of breast cancer or any other cancer is among the most significant fears faced by a cancer patient. Consequently, it becomes one of the concerns that affect their quality of life. Regardless of its relevance, it is infrequently recorded in most breast cancer datasets, which makes research into its prediction more problematic. In addition to the obvious mortality ramifications of recurrence, BC patients also confront severe treatment-related intricacies, which increases their risk of death from causes irrelevant to breast cancer itself .Accurate prediction of BC behavior assumes an essential role in this situation, as it helps clinicians in their decision-making process, supporting a more personalized treatment for patients. Methods such as knowledge discovery in databases (KDD) provide an exciting avenue to investigate such data driven problems. Data mining, a subset of KDD, is an iterative process in the search for new, valuable, and non-trivial information in large volumes of data .The data mining and machine learning approaches have been successfully used in diagnosing and predicting various health-related diseases. These include breast cancer, oral cancer, cardiovascular diseases, and diabetes. The results from these successful studies are used as a motivator to apply data mining technologies as a predictive tool for breast cancer recurrence prediction. Thus, utilizing data mining in these specific forms is the basis of this research. With the advancement of high-throughput technologies, various types of high-dimensional data have been generated in recent years, specifically those related to disease occurrence or management of cancer recurrence. The high dimensionality of the data makes it more difficult to obtain insights from them. There is an urgent need to convert high dimensional data to low dimensional data by using dimensionality reduction methods. Dimensionality reduction facilitates the classification, visualization, communication, and storage of high dimensional data. This study proposed the particle swarm optimization (PSO) algorithm as the feature selection method in reducing the high dimensionality of the Wisconsin Prognosis Breast Cancer

dataset, with three renowned classification algorithms as the classifiers, in an effort to analyze the accuracy level of these three different prediction models. The algorithms are the naive Bayes, K-nearest neighbor (IBK), and fast decision tree learner (REPTree). We also conducted a comparative analysis of the performance metrics between the original dataset and the dataset that has selected features or attributes only. The remainder of this paper contains Section II, which corresponds to a review of all related works within the study domain. The review includes the background information on breast cancer research, prognosis factors, uses of ranking algorithms, several data mining techniques for breast cancer estimation, and a comparison of their accuracies. Section III describes the information about the Wisconsin Prognosis Breast Cancer Dataset that was used to experiment with the three algorithms and various other testing processes. Section IV explains the performance evaluation method. Section V then discusses the experimental results of this study, mainly focusing on the performance evaluation per the aim of the study. Finally, Section VI presents the conclusions of the study and highlights the scope of future work

2. Related work

A. Dimensionality issues

Dimension reduction is defined by Burgess as the mapping of data to a lower dimensional space by removing uninformative variance in data such that a subspace in which the data reside is then detected. Dimension reduction can be divided into feature extraction and feature selection. Feature extraction is the process of distinguishing and disregarding irrelevant, less relevant, or redundant attributes of dimensions in a given dataset. With feature selection, it is possible to identify and remove as much irrelevant and redundant information as possible to build robust learning models. Thus, feature selection not only reduces the computational and processing costs but also improves the model developed from the selected data. A number of existing works have been performed using the feature selection method on healthcare data. The feature selection methods can be categorized into three types of algorithms: filters, wrappers, and embedded approaches. Dimension reduction, as explained by Burgess, is the mapping of data to a lower dimensional space such that uninformative variance in the data is removed such that a subspace in which the data reside is detected. We can further divide dimension reduction into instance selection or reduction and feature selection techniques. Instance reduction is the process of reducing the irrelevant instances from the dataset to increase the classification accuracy, while feature selection is the selection of a subset of the relevant features used in the model construction. These irrelevant instances are not beneficial for classification and may reduce the classification performance. Feature selection helps in removing irrelevant, redundant, and noisy features that are not instrumental to the accuracy of the model. Therefore, it becomes easier to determine only the useful and relevant features for classification rather than using all of them. This results in a fewer number of features, which is desirable, as it simplifies the model and makes it easier to understand. Implementing feature selection in healthcare data will reduce the number of tests required to identify a disease, saving time and money for the patient undergoing tests. In general, we can broadly divide traditional feature selection algorithms into three classes, which are filter approaches, wrapper approaches, and embedded approaches.

B. Data mining in healthcare

Every day, the size of data is increasing; therefore, the need to understand large and complex data is also growing in varied fields, including business, medicine, science, and many others. The ability to extract useful information hidden in this vast amount of data and act on the information is becoming an increasingly important challenge in today's competitive world. The data mining standard is grounded in disciplines such as machine learning, artificial intelligence, probability, and statistics. There are two kinds of data mining models: predictive models and descriptive models. A predictive model is usually applied to supervised learning functions to predict unknown or future values of the variables of interest. Meanwhile, unsupervised learning functions use a descriptive model in finding patterns to describe the data that can be interpreted by humans.

3. Methodology

The overall research methodology for this study was adapted based on the knowledge discovery process. The data acquisition phase was the first phase of this methodology, in which we obtained the relevant data for the study. The second phase was the data pre-processing stage, in which the collected information was integrated, cleaned, and transformed such that the datasets were suitable for classification prediction. After this, still in the second phase, we carried out feature extraction. The data from the pre-processing stage were then carried over to Phase 3 for classification prediction. In Phase 4, the enhanced prediction algorithm, based on the particle swarm, was designed, trained, and tested on the data for classification prediction. In the final phase of the research, we performed a comparative analysis of the models without feature selection and the models that used feature selection.

A. Data acquisition

In the data selection phase, we collected breast cancer data from the UCI public database. The Breast Cancer Wisconsin Breast Cancer Prognostic Dataset has 198 instances and 34 attributes.

B. Data pre-processing

1) Data cleaning

The integrated database went through the data cleaning process, in which we removed improper data entries, such as

those that provided an irrelevant answer, in the database. To smooth noisy data, the tuples with improper data entry were eliminated or filled with the most probable value, as this is one of the most popular strategies to counter this issue. Additionally, the find and replace function was used to handle inconsistency in the format of data from the survey.

2) Data splitting

In the data splitting phase, we divided the data into two datasets: the training dataset and the test dataset. The standard proportion of the splitting process is 60% training and 40% testing. The purpose of splitting the data is to ensure that the model is not over fitted during the model testing with the testing dataset.

C. Classification evaluation without feature selection

In this phase, we constructed three models by using three renowned algorithms, namely, naïve Bayes, REP Tree, and KNN (IBK), as the classifier with the test option of 10-fold cross-validation. The training dataset with all the features/attributes was used for the evaluation

D. Classification evaluation with feature selection

In this phase, the process of feature selection was performed by using PSO in an effort to acquire the best-fit features. Then, we used the selected features to build the three models by performing the same process described in the above paragraph.

E. Phase 5 – comparative analysis

In this phase, we compared the three models without PSO feature selection and the three models with PSO feature selection. Before performing this analysis, we tested all the models for fitness. The method of testing the fitness of the prediction model was examining the confusion matrix; the confusion matrix contained information about the actual and predicted classification obtained by the proposed classifier. The proposed model was validated and benchmarked with the help of an oncologist to evaluate the classification and verify the correctness of the prediction model. The other measures assessed for effectiveness were classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and ROC curves .The proposed method was evaluated against existing data mining methods for accuracy, correctness, and effectiveness.

4. Dataset description

For this research, which focused on the methods and techniques previously discussed, the study leveraged the available dataset provided by the UC Irvine machine learning repository, acquired from the Wisconsin Prognostic Breast Cancer sub-directory with 198 instances.

5. Classification algorithm descriptions

In this study, three renowned [61] classification algorithms for the prediction model, namely, naive Bayes, fast decision tree learner, and K-nearest neighbor, were evaluated in the prediction of breast cancer recurrence by using the Wisconsin Prognostic Breast Cancer Dataset. The following paragraph briefly describes each of the algorithms.

A. Naive Bayes algorithm

Bayesian classification represents a supervised learning method as well as a statistical method for classification. It assumes an underlying probabilistic model and allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems. This classification is named after Thomas Bayes (1702-1761), who proposed the Bayes theorem. Bayesian classification provides practical learning algorithms, in which prior knowledge and observed data can be combined. Bayesian classification presents a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for a hypothesis, and it is robust to noise in input data.

B. Fast decision tree learner algorithm

The reduced error pruning (REP) tree classifier is a quick "decision tree learning algorithm and is based on the principle of computing the information gain with entropy and minimizing the error arising from variance". This algorithm was first recommended in REP Tree applies regression tree logic and generates multiple trees in altered iterations. Afterward, it selects the best tree from all spawned trees. This algorithm constructs the regression/decision tree using variance and information gain. Additionally this algorithm prunes the tree with reduced-error pruning using a back fitting method. It sorts the values of numeric attributes once at the beginning of the model preparation. Additionally, as in the C4.5 Algorithm, this algorithm also addresses missing values by splitting the corresponding instances into pieces.

C. K-nearest neighbors algorithm

K-nearest neighbors (KNN) is a supervised classification algorithm in which the k nearest neighbors of a point are chosen, found by minimizing a similarity measure. To determine the class of an unlabeled example, KNN computes its distance to the remaining examples and determines its k-nearest neighbors and respective labels. The unlabeled object is then classified either by majority voting the dominant class in the neighborhood or by a weighted majority, where greater weight is given to points closer to the unlabeled object.

6. Feature selection algorithm descriptions

Classification involves conscientious consideration of the dataset before assigning the data to a classifier. The recommendation is to consider only necessary features to make the classification process much easier, rather than adding many irrelevant features. Therefore, it is beneficial to have sufficient techniques that are capable of selecting the relevant and significant features. Moreover, if feature selection is adopted in classification, it helps in finding the significant feature and

reducing the workload of the classifier, which also improves the classification accuracy. Based on the review of the existing literature, particle swarm optimization enjoys better selection, in terms of classification accuracy, compared to other existing feature selection techniques.

7. Conclusion

In this paper, we focused on investigating the effect of integrating the feature selection algorithm with classification algorithms in breast cancer prognosis. We proposed that we can improve most classification algorithms by using feature selection techniques to reduce the number of features. Some features have more importance and influence over the results of the classification algorithms compared to other features. We have presented the results of our experiments on three popular classifying algorithms, namely, naïve Bayes, IBK, and REP Tree, with and without the feature selection algorithm, particle swarm optimization (PSO). To conclude, naive Bayes produced better output with and without PSO, whereas the other two techniques improved when used with PSO.

References

[1] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically Optimized Neural Network model," Expert Syst. Appl., vol. 42, pp. 4611–4620, 15 June 2015.

[2] K. Polat, S. Sahan, H. Kodaz, and S. Gunes, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism," Expert Syst. Appl., vol. 32, pp. 172–183, January 2007.

[3] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," Expert Syst. Appl., vol. 41, pp. 1476–1482, March 2014.

[4] E. D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," Expert Syst. Appl., vol. 33, pp. 1054–1062, November 2007.

[5] Cheng and Y. Shi, "Diversity control in particle swarm optimization," in Proceedings of 2011IEEE Symposium on Swarm Intelligence (SIS 2011), Paris, France, April 2011, pp. 110–118.

[6] "A study of normalized population diversity in particle swarm optimization," International Journal of Swarm Intelligence Research (IJSIR), vol. 4, no. 1, pp. 1–34, January-March 2013.

[7] S. Cheng, Y. Shi, and Q. Qin, "Population diversity of particle swarm optimizer solving single and multi-objective problems ,"International Journal of Swarm Intelligence Research (IJSIR), vol. 3, no. 4, pp. 23–60, 2012.

[8] A. Farr, R. Wuerstlein, A. Heiduschka, C. F. Singer, and N. Harbeck, "Modern risk assessment for individualizing treatment concepts in early-stage breast cancer," Rev. Obstet. Gynecol., vol. 6, no. 3/4, pp. 165–173, 2013.

[9] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," Expert Syst. Appl., vol. 36, pp. 3240–3247, March 2009.

[10] N. Sharma and H. Om, "Data mining models for predicting oral cancer survivability," Netw. Model. Anal. Health Inform. Bio inform., vol. 2, pp. 285–295, December 2013.