ISSN: 0067-2904
GIF: 0.851

Best Way to Detect Breast Cancer by UsingMachine Learning Algorithms

Nahla Arabi Hamdo*
Shaqlawa Technical Institute, Polytechnic University, Erbil, Iraq.

Abstract
Breast cancer is the second deadliest disease infected women worldwide. For this reason the early detection is one of the most essential stop to overcomeit dependingon automatic devices like artificial intelligent. Medical applications of machine learning algorithmsare mostly based on their ability to handle classification problems, including classifications of illnesses or to estimate prognosis. Before machine learningis applied for diagnosis, it must be trained first. The research methodology which isdetermines differentofmachine learning algorithms,such as Random tree, ID3, CART, SMO, C4.5 and Naive Bayesto finds the best training algorithm result. The contribution of this research is test the data set with missing value and without missing value, where the missing value is one attribute is missing from one sample for data set. The test result is show SMO is the best algorithm, especiallywhen the research removes the samples that contained the missing value.

Keywords: Breast Cancer Detection, Machine Learning Algorithms, Data Mining

أفضل طريقة لأكتشاف سرطان الثدي بأستخدام خوارزميات تعليم الالة

نهلة عربي حمدو*
المعهد التقني شقلاوة، جامعة بولي تكنك، اربيل، العراق.

الخلاصة
سرطان الثدي هو ثاني أخطر مرض يصيب النساء في جميع أنحاء العالم. لهذا السبب الكشف المبكر هو واحد من المحطات الأكثر أهمية للتغلب عليه اعتمادا على الأجهزة الآلية مثل الذكاء الصناعي. التطبيقات الطبية في خوارزميات تعليم الآلة تعتمد في الغالب على التعامل مع مشاكل التصنيف، بما في ذلك التصنيفات للأمراض أو لتقدير او التكهن. قبل تطبيق التشخيص، يجب تدريب الالة اولا. وفي هذا البحث يتم استخدام منهجية البحث لخوارزميات مختلفة مثلا Random tree وID3 وCART وSMO و C4.5 و Naive Bayes لإظهار أفضل نتيجة لتدريب الخوارزمية. المساهمة العلمية في هذا البحث هو اجراء الاختبار على مجموعة بيانات يوجد ضمنها عينات احد عناصرها مفقود ثماجراء الاختبار بعد حذف تلك العينات. نتيجة التجربة تظهر ان SMOهو أفضل خوارزمية خاصة عند ازالة العينات التي تحتوي على القيم المفقودة.

Introduction

Breast cancer is the second deadliest disease in women worldwide [1]. Breast cancer affects not only women but also men,only 1% of all the cases are found in men. Detection of the disease at an earlier stage can save precious lives [2]. The automatic diagnosis of breast cancer is an important real-world medical problem. A major class of problems in medical science involves the diagnosis of disease, based upon various test performed with the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized diagnostic tools, intended to aid physicians in making sense out of the confusion of data [3].Machine learning algorithms are systems inspired by the human brain [4]. They are densely

*Email: nahlahamdo@yahoo.com

interconnected networks of Processing Elements together with rule to adjust the strength of the connections between the units in response to externally supplied data [3]. Like other machine learning methods - systems that learn from data - Machine learning algorithms have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and recognition.

Medical applications of Machine learning are mostly based on their ability to handle classification problems, including classifications of illnesses or to estimate prognosis [3]. Before the network can be applied for diagnosis, it must be trained first. The training process consists of applying to the network, a subset of the data from Wisconsin Diagnostic Breast Cancer (WDBC) web site that includes the Fine-Needle Aspiration feature parameters and the corresponding classification results,the result will be Benign or Malignant.

Related Work

Waikato Environment for Knowledge Analysis (WEKA) which is a popular suite of machine learning software written in Java, developed at the Universityof Waikato, New Zealand. WEKA is free software available under the GNU General Public License. The WEKA Experiment Environment enables the user to create, run, modify, and analyses experiments in a more convenient manner than is possible when processing the schemes individually. For example, the user can create an experiment that runs several data to determine the relation between every group of data. [5]. However, there is another tool called Tanagara, it is open source data analysis software for academic and research purposes, which proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and database's area. The main purpose of Tanagra is to provide researchers and students to use data mining software in an easy way by conforming to the present norms of the software development and allowing to analyses either real or synthetic data [6].

The researchers [7] conducted a research on comparison of four data mining tools namely WEKA, Tanagra and other techniques for classification purpose. The results concluded WEKA toolkit was the best one in terms of classifiers applicability issue, so that this research focused on it.The researchers [8] comparison results show that Sequential Minimal Optimization (SMO) has higher prediction accuracy than Instance Based learning with parameter k (IBK) and Best First Decision Tree (BF Tree) methods. However, other researchers [9] have investigated three data mining techniques: the Naive Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, the result shows that C4.5 algorithm has a much better performance than the other two techniques. A modest AdaBoost algorithm was proposed by researchers [10] to extract breast cancer survivability patterns using K-means, Relief and modest adaboost. The performance measures analyzed were accuracy, sensitivity and specificity. Furthermore, the researchers [11] made a comparative study of different learning models used in data mining and provided some practical guidelines to select an algorithm for a specific medical application. Many classification algorithms were applied for breast cancer, diabetes and iris data. Among various classification algorithms, Bayesian classification and SMO served with highest accuracy.

The problem of this research is to test which algorithm is best one to detect breast cancer, where researchers found SMO has higher prediction accuracy than IBK and BF Tree methods [8], but another researchers [9] found C4.5 algorithm is a better than Naive Bayes and the back-propagated neural network. The objective of this research is specifying which one is best one to detect breast cancer. This research chose C4.5 and SMO with a different other algorithms to betested. The other algorithms are Random Tree, ID3, CART and Naive Bayes. There are 16 instances in data set for breast cancer and some contain a single missing (i.e., unavailable) attribute value, now denoted by "?" and some researchers removes the 16 instances [12, 13] and some keep it. The contribution of this research tested the dataset with missing value and without missing value, to find the best accuracy to detect breast cancer. Table 1 shows the attributes of WDBC. The first nine attribute take number between 1 to 10. The class attribute takes two numbers, 2 for benign and 4 for malignant.

Table 1- WDBC Attributes

	Attribute	Domain
	Sample code number	id number
1	Clump Thickness	1 - 10
2	Uniformity of Cell Size	1 - 10
3	Uniformity of Cell Shape	1 - 10
4	Marginal Adhesion	1 – 10
5	Single Epithelial Cell Size	1 – 10
6	Bare Nuclei	1 – 10
7	Bland Chromatin	1 – 10
8	Normal Nucleoli	1 – 10
9	Mitoses	1 - 10
10	Class	2 for benign, 4 for malignant

Machine Learning Algorithms

The work in this research focused on six algorithms depends on related work; the first algorithm is Sequential Minimal Optimization (SMO), which an algorithm is that used for training Support Vector Machines (SVMs). The Sequential Minimal Optimization (SMO) algorithm proposed by John Platt in 1998, is a simple and fast method for training a SVM. The main idea is derived from solving dual quadratic optimization problems by optimizing the minimal subset, including two elements at each iteration. The advantage of SMO is that it can be implemented simply and analytically. Training a support vector machine requires the solution of a very large quadratic programming optimization problem. SMO breaks this large quadratic programming problem into a series of smallest possible quadratic programming problems. These small quadratic programming problems are solved analytically, which avoids using a time-consuming numerical quadratic programming optimization as an inner loop. The amount of memory required for SMO is linear in the training set size, which allows SMO to handle very large training sets. Because matrix computation is avoided, SMO scales somewhere between linear and quadratic in the training set size for various test problems, while the standard chunking SVM algorithm scales somewhere between linear and cubic in the training set size. SMO‘s computation time is dominated by SVM evaluation; hence SMO is fastest for linear SVMs and sparse data sets. SMO is an algorithm for solving the Quadratic Programming (QP) problem that arises during the training of support vector machines [14].

The second algorithm is random trees have been introduced by Leo Breiman and Adele Cutler. The algorithm can deal with both classification and regression problems. Random trees classification works as follows: the random trees classifier takes the input feature vector, classifies it with every tree in the forest, and outputs the class label that received the majority of “votes”. In case of a regression, the classifier response is the average of the responses over all the trees in the forest [15].

The third algorithm is Classification, and Regression Trees (CART) is the ultimate classification tree that has revolutionized the entire field of advanced analytics and inaugurated the current era of data mining. CART, which is continually being improved, is one of the most important tools in modern data mining. Others have tried to copy CART but no one has succeeded as evidenced by unmatched accuracy, performance, feature set, built-in automation and ease of use. Designed for both non-technical and technical users, CART can quickly reveal important data relationships that could remain hidden using other analytical tools [16].

The fourth algorithm is Naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (Naive) independence assumptions between the features. The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes models are also known under a variety of names in the literature, including simple Bayes and independence Bayes [17].

The fifth algorithm is the C4.5 algorithm; it used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan’s earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason C4.5 is often referred to as a statistical classifier. C4.5 algorithm uses information gain as splitting criteria. It can accept data with categorical or numerical values. To handle continuous values it generates threshold and then divides attributes with values above the threshold and values equal to or below the threshold. C4.5 algorithm can easily handle missing values. As missing attribute values are not utilized in gain calculations by C4.5 [18].

The sixth algorithm is Iterative Dichotomiser 3 (ID3). It is an algorithm invented by Ross Quinlan used to generate a decision tree from a data set. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains [19]. ID3 begins by choosing a random subset of the training instances. This subset is called the window. The procedure builds a decision tree that correctly classifies all instances in the window. The tree is then tested on the training instances outside the window. If all the instances are classified correctly, then the procedure halts. Otherwise, it adds some of the instances incorrectly classified to the window and repeats the process. This iterative strategy is empirically more efficient than considering all instances at once. In building a decision tree ID3 selects the feature which minimizes the entropy function given below and thus best discriminates among the training instances [20].

Setup and Test Results:

The research test different algorithms. The result of the research focused on correctness of the algorithms in the training. The research depended on WDBC. The research chose C4.5 and SMO, Random Tree, ID3, CART and Naive Bayes,the result specified which one is best one to detect breast cancer. The research used WEKA Version 3.6.10 for test C4.5 and SMO, Random Tree, ID3, CART and Naive Bayes and Tanagra Version 1.4.41 for test ID3. This research wastest the data set with missing value and without missing value. The test result shows that the SMO is the best algorithm. The best way for better correctness was when the research removed the sample for missing value in training for C4.5, SMO, CART ,Naive Bayes and ID3 as shown in Table 2 and Table 3.Because the samples that include missing value make the algorithm not to make the right decision in the case of the sample is incomplete information.However, Random Tree result was better when keeps the sample for missing value, because may be include different strategic for decision making.

Table 2- The Training Correctness Result when Removed Missing Value

Algorithm	Correctly Classified	Incorrectly Classified
SMO	97.0717	2.9283
Naive Bayes	96.3397	3.9531
C4.5	96.0469	3.9531
CART	95.1684	4.8316
Random Tree	94.1435	5.8565
ID3	92.6793	7.3206

Table 3- The Training Correctness Result with Missing Value

Algorithm	Correctly Classified	Incorrectly Classified
SMO	96.9957	3.0043
Naive Bayes	95.9943	4.0057
CART	94.8498	5.1502
Random Tree	94.5637	5.4363
C4.5	94.5637	5.4363
ID3	92.4177	7.5822

Conclusion

The research test different algorithms. The result of the research focused on correctness of the algorithms in the training. It depended on WDBC data set. This work chose C4.5 and SMO, Random Tree, ID3, CART and Naive Bayes for test; the result specified which one is the best to detect breast cancer. The test result shows that the SMO is the best algorithm. The best way was when the research removed the sample for missing value in training for C4.5, SMO, CART , Naive Bayes and ID3. However, Random Tree result was keep better correctness when keeps the sample for missing value.

References

1. Beg, M. M. and Jain, M. **2012**. An Analysis of the Methods Employed for Breast Cancer Diagnosis. *International Journal of Research in Computer Science*, 2(1), pp:25-29.

2. Coleman, W. B. **2013**. Breast Cancer Personalized Medicine: Challenges and Opportunities. *The American Journal of Pathology*, 183(1), 1036-1037.

3. Ahmed, H. A. A. and Hamza, M. H. **2011**. Breast Cancer Diagnosis Using Artificial Intelligence Neural Networks. *Journal of Science and Technology*, 12, pp:159-171.

4. Høysæter, L. S. **2014** ." Sentiment Analysis for Financial Applications", Norwegian University of Science and Technology.

5. Scuse, D. and Reutemann, P. **2007**. WEKA Experimenter Tutorial.University of Waikato.

6. Begum, S. H. **2013**. Data Mining Tools and Trends. An Overview. *International Journal of Emerging Research in Management &Technology*, 2(1), pp:6-12.

7. Wahbeh, A. H., Radaideh, Q. A. A., Kabi, M. N. A. and Al-Shawakfa, E. M. **2011**. A Comparison Study between Data Mining Tools over some Classification Methods. *International Journal of Advanced Computer Science and Applications*.

8. Chaurasia, V. and Pal, S. **2014**. A Novel Approach for Breast Cancer Detection using Data Mining Techniques. *International Journal of Innovative Research in Computer and Communication Engineering,* 2(1), pp: 2456-2465.

9. Bellaachia, A. and Guven, E. **2006**. Predicting Breast Cancer Survivability using Data Mining Techniques, Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining.

10. Thongkam, J., Xu, G., Zhang, Y. and Huang, F. **2008**. Breast Cancer Survivability via AdaBoost Algorithms. In Second Australasian Workshop on Health Data and Knowledge Management, Australia.

11. Andreeva, P., Dimitrova, M. and Radeva, P. **2004**. Data Mining Learning Models and Algorithms for Medical Application. In Proceedings of the International Conference on Systems for Autom.

12. Aruna, S. and Rajagopalan, S. P. **2011**. Knowledge Based Analysis of Various Statistical Tools in Detecting Breast Cancer. *Computer Science & Information Technology*, 1(12), pp: 37–45.

13. Muhic, I. **2013**. Fuzzy Analysis of Breast Cancer Disease Using Fuzzy C-Means and Pattern Recognition. *Southeast Europe Journal of Soft Computing*, 2(1), pp: 50-55.

14. Platt, J. C. 1998. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Research MSR-TR-98-14.

15. Breiman, L. and Cutler, A. 2014.Available: http://www.stat.berkeley.edu/ ~breiman/RandomForests/cc_papers.htm.

16. Loh, W.Y. **2011**. Classification and Regression Trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*,1(1), pp: 14-23.

17. Karlık, B. and Öztoprak, E. *2012*. Personalized Cancer Treatment by Using Naive Bayes Classifier.*International Journal of Machine Learning and Computing*, 2(1), pp: 339-344.

18. Patel, B. R. and Rana, K. K., **2014**. A Survey on Decision Tree Algorithm for Classification. *International Journal of Engineering Development and Research,* 2(1), pp: 1-5.

19. Rashmi, G. D., Maheswari, K. U. and Narayani, V. **2014**. Analytical Research in the Geographical Area for Classifying Childhood Obesity Using ID3 Algorithm. *International Journal of Computer Science*, 2(1), pp:13-17.

20. Vasudevan, P. **2014**. Iterative Dichotomiser-3 Algorithm in Data Mining Applied to Diabetes Database. *Journal of Computer Science*, 10(1), pp:1549-3636.