Differential diagnosis of breast cancer and its types by Data mining algorithms

Hamidreza Asemi Zavareh

[1.] Faculty of Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
h.asemizavareh@yahoo.com

Abstract: This study aims to diagnose breast cancer by appropriately accurate data mining on Mammographic Images. In this case, diagnosing a patient with cancer should be initially done followed by predicting malignant candidates, i.e. a single spot (white) on mammographic image; the most optimal results are obtained if error decreases in diagnosing error of diagnosing a patient with cancer. As the cancerous candidates are low about 650 of total 102,000 candidates, balancing data must be initially done to decrease errors. Then, effective variables are identified among 117 variables per candidate (of its nobles) using PGA and Wilcoxon. Finally, decision tree META COST is used to model neural network algorithms. 70% data is used as training data and the rest 30% is used for validation. The model result is based on the surface beneath the function FROC. The error of predicting unhealthy class to healthy is almost 11%. As a patient is with 11 malignant candidates among his candidates, error of diagnosing a patient with cancer as a healthy one is 11% in power of 11 and it will be approximately zero.

[Hamidreza Asemi Zavareh. **Differential diagnosis of breast cancer and its types by Data mining algorithms**. *Life Sci J* 2013;10(3s):98-103] (ISSN:1097-8135). http://www.lifesciencesite.com. 13

Keywords: Differential diagnosis, breast cancer, Data mining, variable selection.

1. Introduction

Current improvements of different sciences in obtaining, storage and maintaining data caused in increased volume and dimensions of databases. There is always some information in databases to extract and exploit. On the other hand, information hidden in these databases is very useful and functional. The information can be effectively used in different fields. For example, data underlies creation of marketing, investing, producing, manufacturing and trading in global business. This data can an important source of knowledge. Scientific research is the other field in which database and obtaining information can be beneficial.

Medical industry is among several industries which involve with large scale data and information. Stored information, data and knowledge of these industries are increasingly growing. A study revealed that 5 trillion bytes of data are annually generated and stored in a hospital. Ability to use and extract knowledge of this data is very important and deterministic. For, it can help the industries prevent error and support decisions. This necessitates making use of complicated powerful methods as a robust tool to be used in medical database including medical information systems such as tests results and information, medical histories, estimates and medical surveys, etc. These data sets often face with two problems: structure diversity and disintegration. Different data structures lead to dissimilarity in data sets, examples of which are scans images, X ray images, text information describing details, medical history, psychological reports and or illness description with different signals such as ECG and EEG. On the other hand, there is no integrated data; since an individual's medical information is not always stored in one site but is indeed in different places. The disintegration of data formats and dissimilarity of its site cause in a very complicated process encountered in restoring information from a database. In this regard, a pre-process in medial data sets is necessary using telecommunication due to its flexibility, compliancy and usefulness.

A pre-processor used in medical databases is called Telemedicine. The concept is defined as follows: searching, reviewing, patient management, patient training and using system staff who are allowed to access patients' information and practitioners' views, regardless that where is the patient now and where can his information be found out.

Systems like these consistently consider two rules; first, to use it completely professional in medical research to diagnose conditions and second, to create an information source for patients with special illnesses. Telemedicine systems encounter several simple but major problems. First problem is restoring information in different formats. As noted earlier, medicine is a filed in which there is different formatted data. Second, there is incomplete, lost or wrong data in these databases. Finally, extracting knowledge from these databases is difficult. Various techniques of extracting knowledge process, as well as large number of data, different formatted data, incomplete information and large scale databases make it difficult to extract data.

Data mining is a new filed in relation to marketing and extracting information which attempts to extract knowledge and useful interesting features from database sets. In other words, data mining is used as a step to extract knowledge from a database and it is connected to database management, typography, statistics and machine learning. Data mining techniques are developed extensively to decrease dimensions. These techniques can be used in linear and non-linear systems. Some data mining techniques use covariance structure to decrease dimensions and extract hidden structures. They include Principal Component Analysis (provided that hidden characteristics extracted without dependence between found characteristics are orthogonal) factor analysis (not considering orthogonality of hidden characteristics with possibility of dependent characteristics).

Many activities have been done in medicine using data mining techniques. Witten and Frank (2005) used SVP in machine learning. Khedr, A. E., & Mohmed (2012) applied some image mining techniques such as neural networks and association rule mining techniques to detection early liver Cancer using and helping physicians to decide an important decision on a particular patient state. Shahbaz et al. (2012) uses data mining classification tools such as k-nearest neighbors, Naïve bayesian, and SVM to make a decision support system to identify different types of cancer on the Genes dataset. Dudoit et al. (2002) applied the different discrimination methods for the classification of tumors based on gene expression data. The methods include nearest-neighbor classifiers, linear discriminant analysis, and classification trees. Wang et al. (2006) compared the performance of the three classification algorithms, as described by Dudoit *et al*. (2002), for disease classification in the five publicly available cDNA microarray datasets. Troyanskaya et al. (2001) presented a comparative study of several methods for the estimation of missing values in DNA microarray data. Kim et al. (2005) proposed a local least squares imputation method (LLSimpute) in missing value estimation for DNA microarray. Rhodes et al. (2004) presented *ONCOMINE*, a cancer microarray database and web-based data-mining platform aimed at facilitating discovery from genome-wide expression analyses. George and Raj (2011) presented a review of feature selection techniques that have been employed in micro array data based cancer classification and also the predominant role of SVM for cancer classification. Keerin et al. (2012) introduced a Cluster-based KNN missing value imputation for DNA microarray data. Hall and Miller (2009) used correlation based variable selection for discrete data mining to select modeling components. Mitchell (1997) used entropy based feature selection for machine learning. Other studies analyzed gene structure and discovered hidden structures by dimension reduction techniques. Bura and Pfeiffer (2003) introduced simple graphical classification and prediction tools for tumor status using gene-expression profiles using dimension reduction techniques on DNA microarray data. For more studies, see Baldi and Brunak (2001) for various the machine learning approaches in bioinformatics.

This study aims to differentially diagnose breast cancer using data mining algorithms. Cancer begins when cells of an organ uncontrollably grow, split, invade to different organs and spread in whole body. A series of these uncontrollable cells (any unnatural cell increasing and proliferating) is called ratomur. A most common cancer is breast cancer, especially among women. Breast cancer typically begins in lobules namely breast tracts; it can then penetrate through tracts and gland septum and invade peripheral adipose tissues even organs. Treatment and consequences are determined by illness degree or development and spreading rate of cancer in the body. Breast cancer is generally classified into 4 steps:

First step: cancer is limited to breast.

Second step: cancer has spread to other nearby tissues; for example, lymph nodes below the arm of women (usually associated with locally advanced cancers of the breast).

Third step: Cancer has spread to tissues placed below the chest wall (related to locally advanced cancers).

Forth step: Cancer has spread to other parts of the body (related to advanced cancer)

In 2002, N. Mascio et al. examined micro-calcifications in digital mammography as automatic analysis. In this method, gray-scale morphology is used for detecting micro-calcifications in digital mammography and a reduction in false-negative percent. Besides, the study used a number of features to distinguish between revealed cases. An extensive survey has been initially done on the characteristics of an appropriate digitizer. It should be noted that the type of digitizer is very important in an image analysis by CAD in terms of resolution and bit depth. In addition, using appropriate hardware and software is important to minimize analysis speed as low as possible.

2. Research Objectives

According to available information two main objectives is defined in this paper.

2.1. First objective

Given that the incidence of patients with breast cancer is very low in vitro (On average 5 to 10 cases in 1000 are diagnosed with breast cancer). Therefore, participants can judge an image based on the area under the FROC curve in zone 0.3 to 0.2

false positive. For this, participants must send a file containing Confidence Score for each candidate of test data (from infinite negative to infinite positive) which determines the level of confidence in classifier algorithm diagnosing malignant candidate. Infinite positive score shows that considered candidate is malignant with full confidence and infinite negative indicates a benign candidate.

2.2. Second objective

Second phase aims to reduce the workload of radiologists so that they only examine a subset of cases diagnosed by algorithm, at least slightly, as suspected. Therefore, this phase aims to evaluate a fraction of normal patients (who do not need a radiologist examine their image) so that CAD algorithms are 100% sensitive to malignant patients. For this purpose, participants should send a file with a binary decision classifier about whether a patient must be examined by a radiologist in test data.

According to above, the objectives of this problem can be summarized as follows:

- To predict each malignant candidate
- To diagnose a patient with cancer

As noted in previous section on CAD systems, it can be categorized in 4 stages:

1. To determine candidates and suspected regions of being defective in medical images (Figure 1)

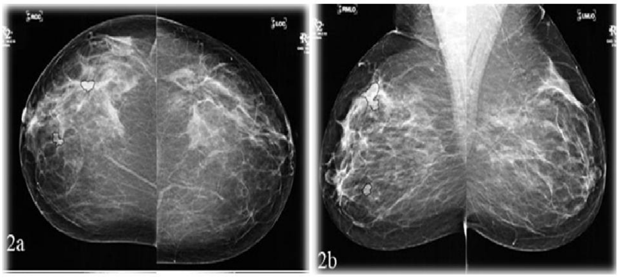

Figure 1: A mammographic images and candidate sites identified on

2. To extract traits so that each candidate is described using values of traits set (using image processing algorithms, Figure 2).

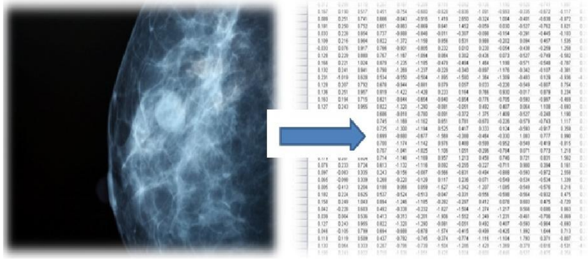

Figure 2: extracting traits

3. To classify candidates as benign and malignant regions using traits values (Figure 3)

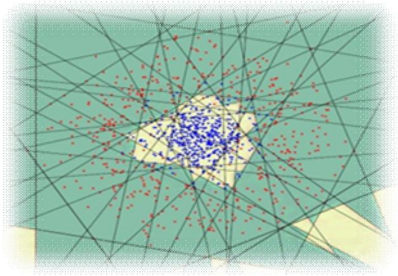

Figure 3: Classifying candidates

4. To provide a comprehensible classification results for Radiologists.

According to objectives of the project, the third phase of CAD systems is considered.

It is worth noting that two errors can be appeared in the second objective:

- Error 1: failure to properly diagnose a patient with cancer (FP)
- Error 2: failure to properly diagnose a non-cancerous patient (FN)

Error 1 means a patient with cancer is mistakenly diagnosed healthy by the system. Since there is usually relatively long interval between medical examinations such a mistake can lead to patient death; whereas error 2 can be determined only by some additional experiments. Thus, one of the major limitations of second objective must be to achieve zero error in error 1.

3. Materials and methods

3.1. Data

The samples studied in this project are related to the KDD Cup 2008 competition. This competition is the oldest and most prestigious data mining competition on the sidelines of a valid KDD conference which is held in Las Vegas, United States in 2008. Training data are related to 118 malignant patients (a patient with at least one injured malignant mass) and 1594 benign patients. Most (but not all) patients have four images; two different images based on imaging angle (MLO, CC) of left and right breast, that is total four different images called MLO Left, MLO Right, CC Left, CC Right. Several candidate points are stored in data set for each image. Therefore, training data set contains 102,294 candidate lines whose a very small part is malignant.

3.2. Data Preparation

In this phase, data is initially studied. Table 1 shows the distribution of target variable for available data in four types of images.

As it is obvious, malignant candidates in each of four groups are close together; in general data is very low. Also, descriptive study on extracted traits obtained following results:

- All 117 traits are normalized with mean zero and variance one.
- Data do not contain missing and nonconforming values .
- There are no irrelevant Data in variables.

Histogram of three traits is shown in Figure 4.

Table 1: Data distribution in different images according to class

Type	Class	Present	Count
CC Left	-1	99.42	**26592**
	1	0.58	**156**
CC Right	-1	99.43	**26392**
	1	0.57	**150**
MLO Left	-1	99.32	**24832**
	1	0.68	**171**
MLO Right	-1	99.39	**23855**
	1	0.61	**146**
TOTAL	**-1**	**99.39**	**101671**
	1	**0.61**	**623**

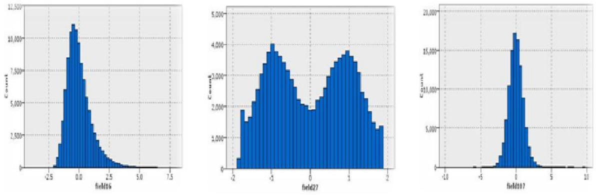

Figure 4: Histogram of three input variables

Figure 5 shows distribution diagram of coordinates X, Y in CC image of left breast (Left CC) prior to standardization. In this image, benign or malignant spots are separated. it clearly indicates distribution of malignant spots.

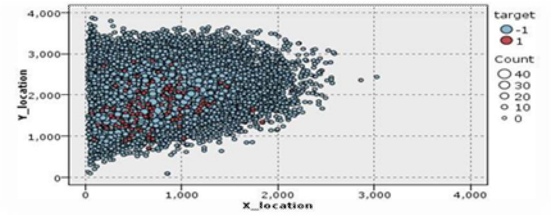

Figure 5: Scatter plot of X and Y coordinates (pre-standardization)

Scatter plot of X, Y (Left CC) post-standardization relative to coordinates of the nipple is as shown in Figure 6, in which red dots are candidate spots. In new coordinates point (0,0) represents the nipple and the other candidates are located relative to it.

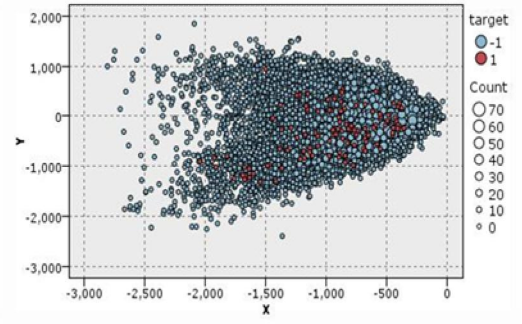

Figure 6: Scatter plot of X and Y (post-standardization)

Following preliminary studies on data and understanding the data, their nature and characteristics, two cases were identified which must be done before modeling phase:

- Balancing data
- Selecting effective variables

Before these steps, 30% of data is initially isolated to assess created models; all further steps is done on 70% of the data and the final results is tested on first 30%. All the above steps as well as modeling and evaluating are done in four separate files separated by image type and position of breast.

3.4. Balancing data

Many studies have been done to reduce imbalance in data. This has been discussed in numerous papers. Strategies proposed in these papers are mainly divided into two categories:

a. allocating costs in proportion to each class so that the model can predict sensitively to cost (Cost Sensitive)
b. Using sample to reduce larger class and increase smaller class. (Over-sampling & Under-sampling)

In this paper both second strategies were performed on data. Given that Smote is mostly used in the literature and the algorithm is implemented in different applications, it was chosen for Over-sampling in the competition.

3.5. Selecting effective variables

As noted earlier, 117 traits derived from image processing algorithms with two developed variables related to standard coordinates of candidates are available to explain variables. However limited number of them will have significant effects on the target variable. Therefore, to select variables influencing target variables is important before entering the modeling phase to reduce dimensions of data. Many algorithms have been introduced to select variables:

- using Entropy

- using Genetic Algorithm
- using decision tree
- using regression (stepwise, backward, forward)
- using Wilcoxon statistic

Different algorithms were used in the study; finally, results derived from PGA algorithm were adopted due to its validity in various references and applicability in similar cases.

4. Results

Various algorithms were used for modeling, including neural network, decision tree, Meta Cost etc. algorithms available in SPSS, Clementine and Weka packages. 70% data of a training dataset was used to develop the model and rest 30% was applied as model validation. Finally, the model was evaluated based on 30% initial row data (isolated earlier) and final models were selected.

The surface beneath the curve is obtained in the range of 0.2 to 0.3 using FROC function by MATLAB to answer the first objective following performance of selected models.

The result desired by selected models is allocating a number in the range of zero and one to each line of datasets which indicates which candidate is malignant and which is benign. The best model is obtained by comparing the surface beneath the curve so that the best prediction model has the most surface beneath the curve (table 2).

Table 2: results of model predicting first objective

inst.#	Predict
27	0.02
28	0.02
29	0.08
30	0.1
31	0.06
32	0.07
33	0.25
34	0.02
35	0.18
36	0.06
37	0.06

Results below were obtained for four classes of studied data by calculating the surface beneath the curve using FROC function by MATLAB (table 3).

Table 3: surface beneath the curve using FROC function

Type	Model	AUC
mloright	M1	0.096803934
mloleft	M2	0.100030731
ccright	M3	0.097326835
ccleft	M4	0.099975743

Models M1 to M4 are defined as follows using WEKA software:

M1 = meta / bagging / reptree
M2 = meta / bagging / random forest
M3= meta / bagging / random forest
M4 = meta / bagging / random forest

Comparison between above results and results of winner team in KDD Cup 08 competition, Research Company IBM indicates that obtained algorithms are highly optimal. Average surface obtained by winner team was 0.96.

In response to the second objective, according to the problem, the system must automatically decide about cancerous patients based on studied candidates. As noted earlier, FP error, that is invalid diagnosis of patients with cancer, should be zero. Since the model is used to diagnose a candidate and considering that occurred error is for invalid diagnosis of a candidate not a patient, the following strategy is used to control errors in diagnosing individual illness.

According to available data in training dataset, there are 11 malignant candidates per patient in his total four images as average. Thus, an individual prediction error can be estimated by the error of model per candidate (table 4).

Table 4: average number of candidates and malignant patients per images

Type	No. Of Patient	No. Of Candidate per patient
Cc Left	56	2.8
Cc Right	58	2.94
Mlo Left	63	2.4
Mlo Right	61	2.4
Total		**10.54**

Final results are provided in Table 5.

Determined models M5 to M8 built in WEKA, are defined as follows:

M5=meta/stacking {meta classifier:metacsot/J48, classifiers: reptree, randomforest}

M6 to M8 = meta / stacking {meta classifier:J48, classifiers: reptree, randomforest}

According to the results, average total validity of prediction per candidate is as follows:

Yes: 89. 18%
No: 98%

Table 5: final results for the second objective

Type	Class	Present	Model
Cc left	Yes	89.5	**M5**
	No	97	
Cc right	Yes	98.5	**M6**
	No	98.8	
Mlo left	Yes	79.2	**M7**
	No	99.2	
Mlo right	Yes	89.5	**M8**
	No	97	

Given that prediction error of unhealthy class is almost 0.11, individual invalid prediction error can be estimated as follows:

P (a cancerous patient is mistakenly diagnosed as healthy)~(0.11) power (11) ~0

As a result, the model is able to diagnose patients with cancer by 100% sensitivity.

It is worth noting that above probability was obtained based on the number of patient candidates. A program for decision system is written in MATLAB based on following algorithm in order to consider dispersion of candidates per patient:

```
X = the number of patient candidates
Y = the number of malignant candidates
If X<150 and Y>= 1 then 1
If X<150 and Y< 1 then -1
If X<300 and Y>= 2 then 1
If X<300 and Y< 2 then -1
If X>300 and Y>= 3 then 1
If X>300 and Y< 3 then -1
```

Based on above algorithm, patients are attempted to diagnose.

5. Conclusion

A consulting system for prediction of breast cancer was provided using the new science, data mining, during the study through creating a learning system, discovering effective input variables on malignancy of studied candidates, developing a prediction model for malignant new candidates. According to the results, the system is able to diagnose patients with cancer by a high validity; its error in mistakenly diagnosis of healthy patients is reasonable. Thus, prediction and diagnosis of cancer can be performed using data analyzing algorithms by higher accuracy and lower error than human diagnosis; in this way, it is possible to save many financial and time costs.

Corresponding Author:
Hamidreza Asemi Zavareh
Faculty of Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.
Email: h.asemizavareh@yahoo.com

References

1. Baldi, P. F., & Brunak, S. (2001). Bioinformatics: the machine learning approach. Adaptive computation and machine learning. Retrieved from http://library.wur.nl/WebQuery/clc/1667873
2. Baldi, P., & Brunak, S. (2001). Bioinformatics: The Machine Learning Approach. MIT Press.
3. Bura, E., & Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. Bioinformatics, 19(10), 1252–1258.
4. Dudoit, S., Fridlyand, J., & Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association, 97(457), 77–87.
5. George, G., & Raj, V. C. (2011). Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification using Gene Expression Profile. arXiv preprint arXiv:1109.1062. Retrieved from http://arxiv.org/abs/1109.1062
6. Hall, P., & Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. Journal of Computational and Graphical Statistics, 18(3), 533–550.
7. Kim, H., Golub, G. H., & Park, H. (2005). Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics, 21(2), 187–198.
8. Mitchell T. M. (1997). Machine Learning. McGraw-Hill.
9. Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Chinnaiyan, A. M. (2004). ONCOMINE: A cancer microarray database and integrated data-mining platform. Neoplasia (New York, NY), 6(1), 1-6.
10. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525.
11. Wang, D., Lv, Y., Guo, Z., Li, X., Li, Y., Zhu, J., Yang, B. (2006). Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules. Bioinformatics, 22(23), 2883–2889.
12. Witten I, Frank E (2005) Data Mining. Practical Machine Learning Tools and Techniques with Java Implementations. Second edition. Morgan Kaufmann Publishers.
13. Keerin, P., Kurutach, W., & Boongoen, T. (2012). Cluster-based KNN missing value imputation for DNA microarray data. In 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 445 –450). Presented at the 2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC). doi:10.1109/ICSMC.2012.6377764.
14. Khedr, A. E., & Mohmed, A. E. G. A. M. (2012). A proposed image processing framework to support Early liver Cancer Diagnosis. Life Sci J, 9(4). 3808- 3813.
15. Shahbaz, M., Faruq, S., Shaheen, M., & Masood, S. A. (2012). Cancer Diagnosis Using Data Mining Technology. Life Science Journal, 9(1), 308-313.

2/2/2013