Applied Medical Informatics
Original Research Vol. 37, No. 4 /2015, pp: 11-22

Prediction of Breast Cancer using Rule Based Classification

Nagendra Kumar SINGH

Department of Biological Science & Engineering, Maulana Azad National Institute of Technology Bhopal, 462003, India
E-mail: nagendravns@gmail.com

* Author to whom correspondence should be addressed; Mob. No.: +91-9752864687

Received: September 3, 2015 /Accepted: November 28, 2015/ Published online: December 15, 2015

Abstract
The current work proposes a model for prediction of breast cancer using the classification approach in data mining. The proposed model is based on various parameters, including symptoms of breast cancer, gene mutation and other risk factors causing breast cancer. Mutations have been predicted in breast cancer causing genes with the help of alignment of normal and abnormal gene sequences; then predicting the class label of breast cancer (risky or safe) on the basis of IF-THEN rules, using Genetic Algorithm (GA). In this work, GA has used variable gene encoding mechanisms for chromosomes encoding, uniform population generations and selects two chromosomes by Roulette-Wheel selection technique for two-point crossover, which gives better solutions. The performance of the model is evaluated using the F score measure, Matthews Correlation Coefficient (MCC) and Receiver Operating Characteristic (ROC) by plotting points (Sensitivity V/s 1- Specificity).

Keywords: Chromosome; Mutation; Cancer; Classification; Data Mining; Bioinformatics

Introduction

Cancer is amongst the most dreaded diseases affecting humankind. In 2012, around 14.1 million new cancer cases were found, 8.2 million cancer - related deaths occurred and 1.7 million women were diagnosed with breast cancer [1]. Sometimes mutations occur in the genetic information which causes uncontrolled growth and division of cells. This unlimited growth and division of cells will lead to the formation of the tumor in the human body [2]. Cancer is nothing but unlimited growth and division of cells. There are various factors such as genetic, hormonal, environmental, socio - biological and physiological which cause somatic mutations. The somatic mutation in breast cell is mainly responsible for breast cancer [2, 3]. Breast cancer is mostly found in milk producing glands (lobules) or ducts (surrounding blood vessels, connective tissue of lobules).

Breast cancer is of mainly two types: invasive (infiltrating ductal carcinoma (IDS) [4], medullary carcinoma [5], infiltrating lobular carcinoma (ILC) [6], tubular carcinoma [7] and inflammatory breast cancer (IBC) [8]) and non–invasive (ductal carcinoma in-situ (DCIS) [9] & Paget's disease) [10]. Invasive breast cancer has tendency to spread over surrounding tissues whereas the non – invasive breast cancer does not have the tendency to spread over surrounding tissues [11, 12]. The mutation in various genes such as AR, ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, *CHEK2*, DIRAS3, ERBB2, NBN, PALB2, RAD50, RAD51, RAD51 STK11and TP53 causes the Breast cancer [13-15].

The most important step from the viewpoint of clinical treatment and possible recovery is early diagnosis of breast cancer which is usually done by self - examination, mammogram, MRI, breast

ultrasound, etc. [11, 12, 16, 17] but none of these methods can be regarded as an absolutely flawless early – prediction tool. Besides, almost all these methods are cost – intensive [18]. The current work has focused on a data mining approach for predicting the relationship between various factors causing breast cancer. Data mining is an autonomous or semi - autonomous tool for extraction of relevant knowledge from a huge amount of data or data sets [19-21]. There are various types of task primitive such as association rules, classification, clustering, and predictions for knowledge discovery. Here, genetic algorithm is used for knowledge discovery. The genetic algorithm (GA) employed is a meta - heuristic global search technique that gives excellent computational performance for searching IF - THEN classification rules in a large search space. The classification is the process of predicting the class label of unknown data based on known class label [19-21]. In this method, IF - THEN classification technique is defined for knowledge discovery. GA discovers the intelligent rules or knowledge from stored historical data based on predicted attributes or dimensions [22, 23]. The discovering rules help us to predict whether the patient has breast cancer risk or not.

The genetic algorithm (GA) is one type of the evolutionary computing, which is based on the Darwin principle of survival of the fittest and genetics. John Holland introduced GA in 1975 [24]. GA is a guided random search for optimization and searching solution in problem state space. It works best for searching the steady state, multipoint or multimodal search and high dimensional problem in a search space. GA is also called greedy and adaptive parallel search technique [22, 23].The Genetic algorithms for classification gives better accuracy than rules discovered by other classification algorithms [25-32]. M. V Fidelis et al.[33] proposed GA based methods for IF-Then rules discovery from different data sets (dermatology data set and breast cancer dataset). In this work, the author used two points cross over (100%) with mutation rate (30%) along with sensitivity and specificity measures for generating and validating the solution. Korkut Koray et al. [34] proposed non –random and uniform operator based GA for rules discovery that remove the demerit of random population generation and give better performance. Later Basheer et al. [35] describes variable gene encoding mechanism for rules discovery form multidimensional data set. Mutaher et al. [36] developed a GA based tool for knowledge acquisition. Here, they used Michigan style encoding mechanism, uniform population generations and one point crossover along with precision, coverage, simplicity, contribution measure for evaluation of solution. Ayad & Anar [37] classified breast cancer into two types of classes called risk or safe by back propagation algorithm on the basis of mutation in breast causing gene BRCA1 & BRCA2. They used alignment techniques for comparing the normal vs. abnormal gene sequence to detect the mutation and training with algorithms.

Above proposed classification breast cancer data and breast cancer detection methods by various researchers included behavior of breast cancer or mutation in BRCA1 or BRCA2 gene only. Many more genes take part in breast cancer named as above. The proposed genetic algorithms based model used for breast cancer detection with the help of classification (Risk or safe)by including thirteen parameters of risk factor and symptoms of breast cancer.

Material and Method

The projected method has included seventeen breast cancer causing genes along with symptoms of breast cancer and risk factors. Thirteen attributes have been screened from different resources including breast cancer causing risk factor, symptoms and genes that take part in breast cancer [11-13]. The following steps are included in computations.

1. Alignment of normal vs. abnormal gene sequence.
2. If both match then generate protein sequence and align the both sequences.
3. Searching the rule using Genetic algorithms
    a. Mapping of symptoms, risk, and gene of breast cancer in form of chromosomes by variable gene coding mechanism.
    b. Generate the uniform population, calculate the fitness of each individual, and create initial population.

c. While termination criteria satisfy
   i. Select two chromosomes from initial population
   ii. Two point Crossover performed and offspring generates
   iii. Calculate the fitness and mutation of offspring
   iv. If offspring support minimum fitness, then select it and include in population for next generation reproduction.

d. End

4. End

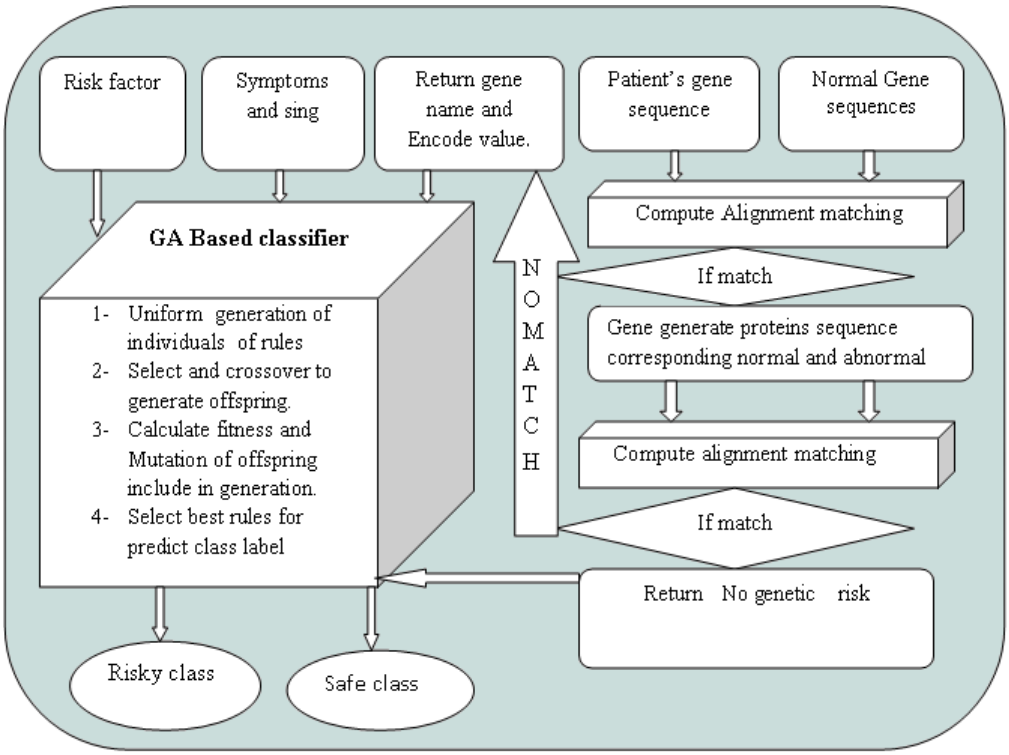

Figure 1. Diagrammatic representation of computation

The following steps are used for training the breast cancer data using genetic algorithms:

Chromosome Encoding

The Michigan style encoding mechanism for individual representation has been used. This method is better for finding small set of classification rules. The projected method included thirteen parameters as risk factors and symptoms of breast cancer for class label training (Table 1-3). Encoding mechanism has set of dimensions or predicted attributes as gene represented on a chromosome (Figure 2). The gene value is changed in individuals according to domain of attributes that belong to specified rules. A zero value for a gene means predicted attribute does not have domain attribute in the rule.

Table 1. The encoding of causing Risk factors for breast cancer encoded into genetic algorithm

Risk not controllable by human beings	Short Code	Range of Dimension	Gene code value
Age	Ag	A>=60, A= >40, A=<40	1,2, 3
Gender	Ge	Female, Male	1,2
Genetic Mutation in Gene	GMg	AR, ATM, BRAD1, BRCA1, BRCA2, BRIP1, CHD1, CHEK2, DIRAS3, HER2, ERBB2, NDN, PAIB2, RAD50, RAD1, STK11, TP53,	1,2,3,4,5,6,7,8,9,10,11, 12,13,14,15,16,17
Family History	FH	Sibling , Mother, Father, Ancestor	1,2,3,4
Previous chest Therapy	PTh	Radio therapy in young age, Hormone Therapy	1,2

Table 2. Life style related risk factors responsible for causing breast cancer encoded into short-code and corresponding each domain of risk factor define code value for chromosome formation

Life style factor that can reduce the risk of breast cancer	Short code	Domain of Dimension	Gene code
Maternal status	MTs	Married, Single, have child, Birth control, Breast feeding, Others.	1, 2, 3,4 ,5
Hormones Therapy	HT	Post-menopausal hormone therapy (PHT), Hormone replacement therapy (HRT), and Menopausal hormone therapy (MHT).	1,2,3
Physical	Ph	Heavy weight, Drinking alcohol, etc.	1,2,3

Table 3. The symptoms & signs of breast cancer encoded into short-codes and assigned code value to each domain of symptoms & signs for chromosome formation

Symptoms & Signs	Short code	Range of Dimension	Gene Code
Lump	Lmp	Thickened tissues, mass	1,2
Swelling	Swe	Breast, Nipple, Both	1,2,3
Pain	Pan	Breast, Nipple, Both	1,2,3
Skin color	SCo	Rash nipple, Pink or scaly breast, Black around nipple, Nipple retraction, dimpling, Itching, Burning sensation, All	1,2,3, 4,5,6,7,8
Discharge	Dic	Breast, Nipple, Both	1,2,3

Ag	Ge	GMg	FH	PTh	MTs	HT	Ph	Lmp	Swe	Pan	Sco	Dic
1	1	4	4	1	2	3	2	1	2	1	3	1

Figure 2. The mapping of short code and corresponding domain gene value of risk factor & symptoms of breast cancer to form a chromosome (where Ag = age, Ge = gender, GMg = Genetic Mutation in Gene`, FH = Family History, PTh = Previous chest Therapy, MTs = Maternal status, HT = Hormones Therapy, Ph = Physical, Lmp = Lump, Swe = Swelling, Pan = Pain, Sco = Skin color, Dic = Discharge)

The IF part of rules such as: If **(**Age>=60) ∧ **(**Gender= "Female") ∧ **(**Gene Mutation= "BRCA2") ∧ **(**Maternal = "Married") ∧ (Physic= "Drinking alcohol") ∧ (Skin color= "Pink") ∧ (Swelling= "Both") ∧ (Pain= "Both") ∧ (Discharge= "Both"), will be represented in the form of individual rule as given in Figure 3.

Ag	Ge	GMg	FH	PTh	MTs	HT	Ph	Lmp	Swe	Pan	Sco	Dic
1	1	5	0	0	1	0	2	0	3	3	2	2

Figure 3. The Representation of rule in the form of chromosome having short gene code along with code value of risk factor &symptoms of breast cancer,Where zero code value shows absence of factor in the rule (where Ag = age, Ge = gender, GMg = Genetic Mutation in Gene`, FH = Family History, PTh = Previous chest Therapy, MTs = Maternal status, HT = Hormones Therapy, Ph = Physical, Lmp = Lump, Swe = Swelling, Pan = Pain, Sco = Skin color, Dic = Discharge)

Fitness Function Formulation

GA is based on survival of the fittest principle. The best fit chromosome survives and has the possibility to give more desirable solution or fit individual after the crossover operation. The Fitness function f(x) measures the efficiency or validation of the individual for the reproduction or next level solution. The goal of fitness function is to maximize the fitness of an individual. The projected method definition of fitness function is based on IF G_1… G_n THEN M rule. Here G1…$G_n$are antecedent parts that may be satisfied or unsatisfied set of predicted dimensions or attributes and M, the consequent part, has class label. The rule based classification technique gives rise to four situations, TP, TN, FP, and FN as defined bellow.

		Breast Cancer Patients		
		Positive	**Negative**	
Class Test	**Positive**	True Positive(TP)	False Positive(FP)	TP + FP
	Negative	False Negative(FN)	True Negative(TN)	FN+TN
		TP + FN	FP+TN	

TP = predicted rule says patient has breast cancer and it is true patient has breast cancer
FP = predicted rule says patient has breast cancer but patient does not have breast cancer
TN= predicted rule says patient does not have breast cancer and patient also does not have breast cancer
FN= predicted attribute says patient does not have breast cancer but patient has breast cancer

The discovered rule If G then M

TP = tuples satisfy both G and M
TN = tuples satisfy neither G nor M
FP = tuples satisfy G but not M
FN = tuples not satisfy G but satisfy M

Definitions of some measures:

Confidence. In rule, (If G then M) confidence can be formulated as:

$$Confidence = \frac{|G \& M|}{|G|} = \frac{TP}{TP + FP} \quad (1)$$

where |M| is the number of attributes in conjunction condition that satisfy antecedent part in the rule and |G&M| number of tuples that satisfy antecedent G and consequent M.

Support. In the above rule, support can be formulated as:

$$Support = \frac{|G \& M|}{|D|} = \frac{TP}{TP + FN} \quad (2)$$

where |D| is number of instance or cardinality in data set D.

Support is ratio of tuples covered by rules in antecedent having predicted class M.

Cover. The cover can formulated as

$$\text{Cover} = \frac{\textit{The predicted attribute present in rule}}{\textit{Total number of predicted attribute}} \quad (3)$$

This is a modified measure for improving the solution because the rules having more number of attributes has better accuracy than those having less number of attributes [21].

Example: In the Fig. 3 above, individuals have total 13 predicted attributes and 9 attributes are present in the rule.

Cover = 9/13 = 0.069

Fitness of individual calculated by this function:

Fitness (R_i) = Confidence × Support ×Cover × 100 (4)

Initial Population Generation

The above-discussed individual encoding mechanism is used to produce uniformly population. For n predicted attributes, maximum 2^n individuals are produced in population. The uniform technique of population generation removes the demerit of random generated population.

Selection Mechanism

Selection mechanism helps us in deciding which technique is to be used for selection of two individual chromosomes from the population pool to improve the solution. The strategies of selection of individuals conceptually consider premature converses and diversity. The Roulette-Wheel Selection technique is used here for selecting individuals. In this technique individual selection is proportional to its fitness. If M number of individual have fitness $F_i>0$ (i= 0, 1,2,.....M), then selection probability of i^{th} chromosome is:

$$Si = \frac{Fi}{\sum_{i=1}^{M} Fi} \quad (5)$$

where (i = 1 , 2, . . . , M)

If Roulette-Wheel sectors size is proportional to Fi (i=1,2,3....M) of individuals then selection of a chromosome is equivalent to randomly selecting a point on the wheel.

Crossover or Recombination

Producing offspring from the parent is called recombination. In crossing over process, two chromosomes are swapping gene features on specific positions to create a better individual. There are many methods for recombination; two point crossovers (exchange of gene on two points of parent chromosome) have been used in the current study (Figure 4).

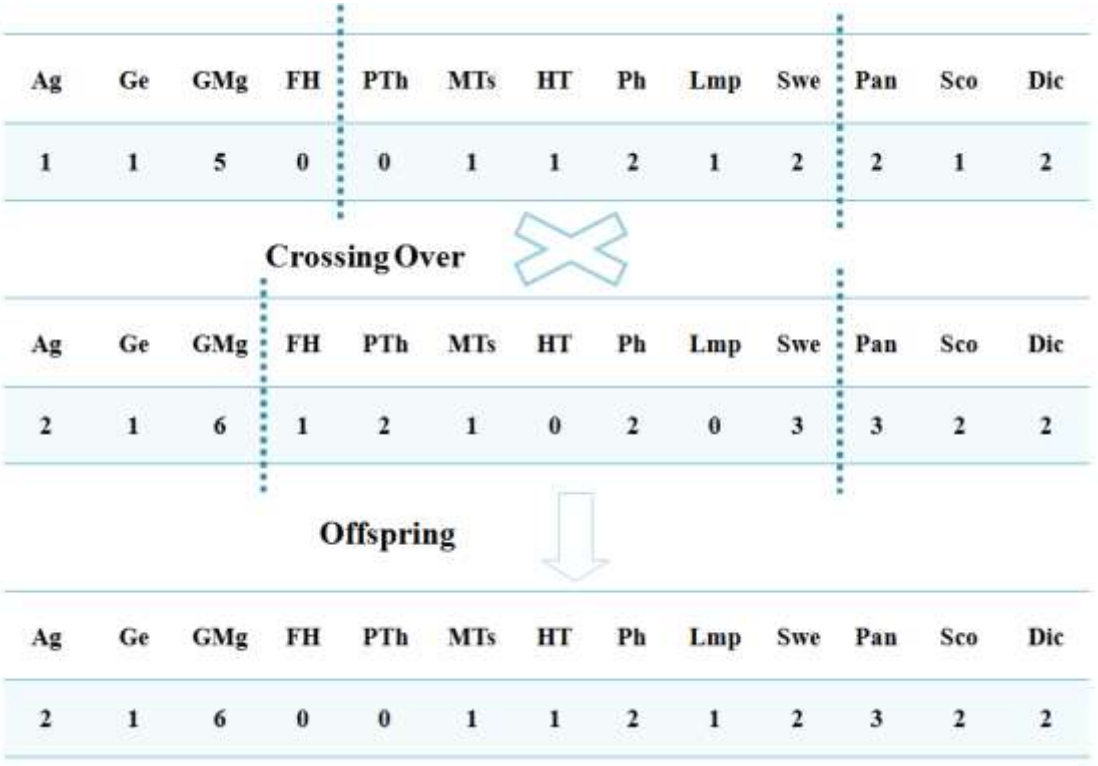

Figure 4. Creation of offspring by two-point crossing over of chromosome

Mutation

Here one point mutation (1%) is used for improving solution. In mutation, changes in gene value of individual maintain diversity in population for improving the solution.

The leave-one-out cross validation is used for validating the solution. Leave – one - out is a special type of M-fold cross validation technique in which sample artificial data set partitions into 1 (one for testing) and n-1 (for training) sets [21]. The generating rules using test set and training sets for discovering rules support the minimum defined fitness threshold. The next time selects another tuple as a test set of validation and repeats this process 'n' times. The overall performance analysis of the model is based on defining fitness threshold (2.0) of discovering rules in validation supporting negative and positive breast cancer using F-score [38], Mathews Correlation Coefficient (MCC)[39] and Plotting of receiver operating characteristic (ROC) on sensitivity V/s 1 – specificity [40].

$$\text{F-Score} = [2\times\text{Sensitivity}\times\text{Specificity}]/(\text{Sensitivity}+\text{Specificity}) \quad (6)$$

$$\text{Sensitivity} = (\text{True Positive})/[(\text{True Positive})+(\text{False Negative})] \quad (7)$$

$$\text{Specificity} = (\text{True Negative})/[(\text{True Negative})+(\text{False Positive})] \quad (8)$$

$$\text{MMC} = [(TP\times TN)-(FP\times FN)]/\sqrt{[(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)]} \quad (9)$$

where TP = predicted rule says patient has breast cancer and it is true patient has breast cancer; FP = predicted rule says patient has breast cancer but patient does not have breast cancer; TN = predicted rule says patient does not have breast cancer and patient also does not have breast cancer; FN = predicted attribute says patient does not have breast cancer but patient has breast cancer

The proposed method is validated with MATLAB (R2010a) in two steps [41]. Due to lack of data set results are evaluated based on generated artificial dataset (supplementary file). In the first step, the Gene sequence responsible for causing breast cancer was retrieved from Gene Bank Database [42] and gene sequence was aligned with the patient's gene sequence. The protein

sequence is generated using a bioinformatics toolbox with the help of Open Reading Frame (ORF) extracted form both gene sequences (normal as well as abnormal). The generated protein sequence forms a specific gene sequence and is compared through global alignment. In the second step, discovering classification rules using GA based technique is implemented with uniform population, roulette wheel selection, crossing over (100%) and mutation rate (1%) on given artificial training data set.

Result and Validation

The result presented in Figure 5 shows diagonal dense line dot plot between normal and patient gene sequence as matching (identical/similar) pattern. The point between 0.0 - 0.5 (Red Box) matching line shows some sparse gap. This gap denotes the occurrence of mutation in gene between normal and patient.

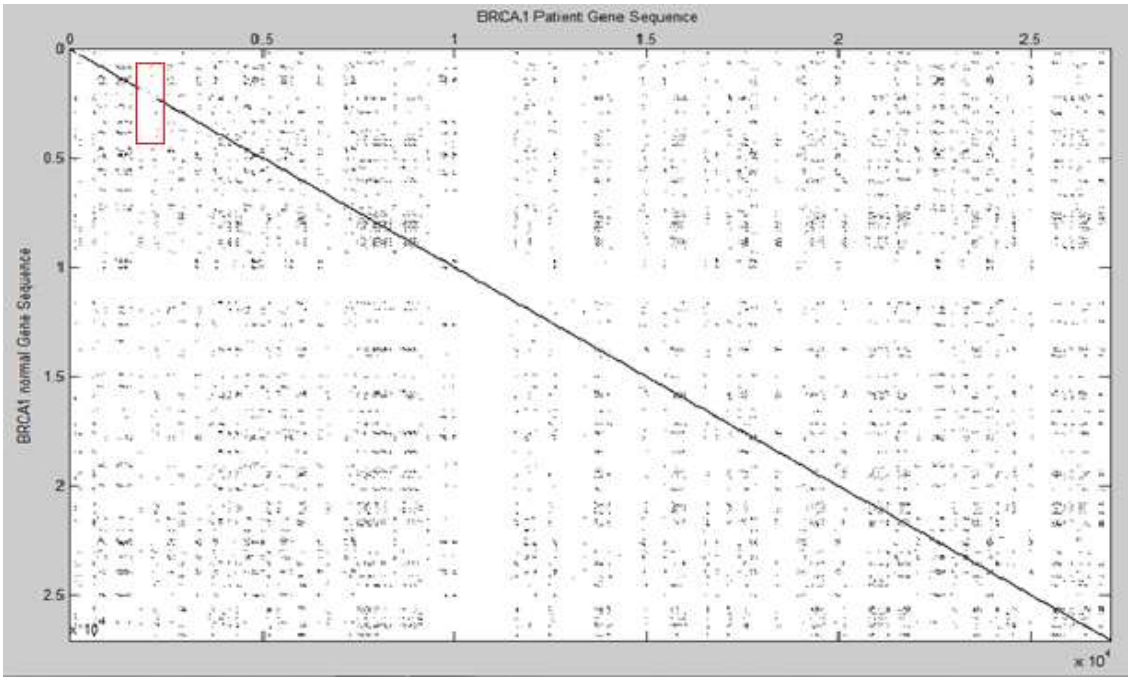

Figure 5. Gene sequence dot plot for normal BRCA1 gene and patient BRCA1 gene showing mutation

The results (Figure 6) show that the both sequences share 98% identity and 99% positivity between normal and patient protein sequence. The 2% dissimilarity is because of mutation in BRCA1 protein, which causes the risk of breast cancer.

Figure 6. Comparison of generated protein sequence from normal gene and patient gene showing identity and positivity

Table 4. Best discovered rules from artificial data set along with class label and fitness of rules

Rule	If Part of discovered rules	Class label	support	conf	Cover	F(Ri)
R1	If(Age=>60)∧(Family history= 'Mother')∧(Previous therapy='Radio')∧(Maternal= 'Married'∧(Hormone therapy='HRT')∧ (swelling='Both')∧(Pain= 'Both')	Risky	0.30	1	0.54	16.2
R2	If(Age= '>=60')∧(Gender= 'Female')∧(Mutation= 'BRCA1')∧(Swelling= 'Both')∧(Pain= 'Both') ∧ (Discharge= 'Both')	Risky	0.20	1	0.47	9.4
R3	IF(Age=>40) ∧ **(**Mutation= 'TP53') ∧ **(**Family history= 'Mother')∧ (Physical= 'Alcoholic')∧ **(**Swelling= 'Both')∧('Pain= 'Both') ∧ (Skin color= 'Pink')	Risky	0.10	1	0.54	5.4
R4	IF(Age=>40) ∧ **(**Maternal status= 'Married') ∧(Physical= 'Alcoholic')∧ **(**Swelling= 'Both')∧('Pain= 'Both') ∧ (Skin color= 'Pink') ∧(discharge= 'Both')	Risky	0.10	1	0.54	5.4
R5	IF**(**Maternal status= 'Married') ∧ (Physical= 'Alcoholic')∧(Lump= 'Mass')∧ **(**Swelling= 'Both')∧('Pain= 'Both') ∧ (Skin color= 'Pink') ∧(Discharge= 'Both')	Risky	0.10	1	0.54	5.4
R6	IF (Age=>40) ∧ **(**Therapy = 'Hormone')∧ **(**Maternal status= 'Married') ∧(Physical= 'Heavy weight')	Safe	0.10	1	0.31	3.1
R7	IF(Age=>40) ∧**(**Maternal status= 'Married') ∧ (Physical= 'Alcoholic')	Safe	0.20	0.50	0.24	2.4

The value of the F-score is maximized towards 1 showing the best performance. The MCC value belongs between -1 to +1. A sub - zero MCC value means worse performance than random solution, whereas a greater than zero MCC value means better prediction than the random solution. The calculation of Sensitivity=87%, specificity=62%, F Score=0.727 and MCC=0.501 is performed using above formula and Plotted ROC shows better performance than random.

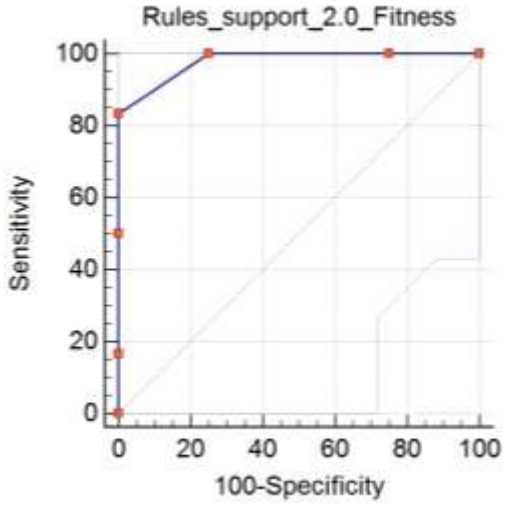

Figure 7. Graph showing sensitivity vs. (1- Specificity)

Conclusions

The proposed method accurately predicts the chance of breast cancer disease because it verifies mutation in the gene and simulates relationship between various risk factors as well as symptoms. Computational approach for classification to discover knowledge gives faster results because genetic algorithms have a global and an adoptive search strategy in high dimensional search spaces. The projected method performs parallel search operation on thirteen predicted attributes for extracting comprehensible knowledge to predict breast cancer. The computational methods in GA, such as - uniform population generation selection, two point crossing over, mutation, and fitness function avoid premature convergence and maintain population diversity. Based on sensitivity = 87%, specificity = 62%, F-Score = 0.727, MCC = 0.501 and ROC, it can be concluded that this is a better prediction. It is hoped that this method will help researchers hit upon rules focusing on identification of gene mutation causing breast cancer and predict the associated causing factors.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. Int. J. Cancer 2015;136:E359-E386.
2. Jiao X. Somatic Mutations in Breast Cancer Genomes: Discovery and Validation of Breast Cancer Genes [Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty

of Medicine]. Acta Universitatis Upsaliensis; 2012 [cited 2015 June] Available from: URL:https://www.diva-portal.org/smash/get/diva2:559456/FULLTEXT01.pdf.
3. Reintjes N, Li Y, Becker A, Rohmann E, Schmutzler R, Wollnik B. Activating Somatic *FGFR2* Mutations in Breast Cancer. PLoS ONE 2013;8(3):e60264.
4. Ruan X, Liu H, Boardman L, Kocher J-PA. Genome-Wide Analysis of Loss of Heterozygosity in Breast Infiltrating Ductal Carcinoma Distant Normal Tissue Highlights Arm Specific Enrichment and Expansion across Tumor Stage. PLoS ONE 2014;9(4):e95783.
5. Malyuchik SS, Kiyamova RG. Medullary Breast Carcinoma. Exp Oncol. 2008;30:96-101.
6. Pestalozzi BC, Zahrieh D, Mallon E, Gusterson BA, Price KN, Gelber RD, et al. Distinct Clinical and Prognostic Features of Infiltrating Lobular Carcinoma of the Breast: Combined Results of 15 International Breast Cancer Study Group Clinical Trials. J Clin Oncol. 2008;26(18):3006-3014.
7. Emad AR, Andrew HSL, Andrew JE, Sindhu M, Nancy YA, Zsolt H, et al. Tubular Carcinoma of the Breast: Further Evidence to Support Its Excellent Prognosis. J Clin Oncol. 2010;28(1):99-104.
8. Nouh MA, Mohamed MM, El-Shinawi M, Shaalan MA, Cavallo-Medved D, Khaled HM, et al. A potential prognostic marker for inflammatory breast cancer. J Trans Med. 2011;9(1):1. doi:10.1186/1479-5876-9-1.
9. Zhou W, Jirstrom K, Amini R M, Fjallskog ML, Sollie T, Lindman H, et al. Molecular subtypes in ductal carcinoma in situ of the breast and their relation to prognosis: a population-based cohort study. BMC Cancer 2013;13(1):512.doi:10.1186/1471-2407-13-512
10. Ling H, Hu X, Xu XL, Liu ZB, Shao ZM. Patients with Nipple-Areola Paget's Disease and Underlying Invasive Breast Carcinoma Have Very Poor Survival: A Matched Cohort Study. PLoS ONE 2013;8(4):e61455. doi :10.1371/journal pone. 0061455.
11. Breast Cancer symptoms [Internet] [cited 2015 June] Available from: URL: http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-diagnosis.
12. Breast Cancer early diagnosis [Internet] [cited 2015 June] Available from: URL: http://www.nationalbreastcancer.org/breast-cancer-diagnosis.
13. National Library of Medicine (US). Genetics Home Reference [Internet]. Bethesda (MD) The Library; 2013. Cystic fibrosis [reviewed 2012 Aug; cited 2013 Sep 19]. Available from: URL :http://ghr.nlm.nih.gov/condition/cystic-fibrosis.
14. Parshad R, Price FM, Bohr VA, Cowans KH, Zujewski JA, Sanford KK. Deficient DNA repair capacity, a predisposing factor in breast cancer. Brit. J. Can. 1996;74(1):1-5.
15. McPherson K, Steel CM, Dixon JM.ABC of Breast Diseases Breast cancer-epidemiology, risk factors, and genetics. BMJ 2000;321(7261):624-628.
16. Berg WA, Blume JD, Cormack JB, Mendelson EB, Lehrer D, Bohm-Vélez M. Combined Screening with Ultrasound and Mammography vs. Mammography Alone in Women at Elevated Risk of Breast Cancer. JAMA 2008;299(18):2151-2163.
17. Morrow M, Waters, Morris E. MRI for breast cancer screening, diagnosis, and treatment. The Lancet 2011;378(9805):1804-1811.
18. Miller AB, Wall C, Baines CJ, Sun P, To T, Narod SA. Twenty five year follow-up of breast cancer incidence and mortality of the Canadian National Breast Screening study: Randomized Screening trial. BMJ 2014;348:g366. doi:10.1136/bmj.g366.
19. Aggarwal CC, Philip SY. Data mining techniques for associations, clustering and classification [In Methodologies for Knowledge Discovery and Data Mining]. Springer Berlin Heidelberg Lecture Notes in Computer Science 1999;1574:13-23.
20. Fu AW. Data mining. Potentials IEEE 1997;16(4):18-20.
21. Han J, Kamber M, Pei J (Eds). Data mining concepts and techniques. (3rd Edition) The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 2011.
22. Ngan PS, Wong ML, Lam W, Leung KS, Cheng JC. Medical data mining using evolutionary computation. Art Intell Med 1999;16(1):73-96.
23. Freitas AA. A review of evolutionary algorithms for data mining. Soft Computing for Knowledge Discovery and Data Mining 2010;371-400.

24. Holland JH. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, Michigan; re-issued by MIT Press, 1975.
25. Keshavamurthy BN, Khan AM, Toshniwal D. Improved Genetic Algorithm Based Classification. Int J Compu Sci Info. 2010;1(3):2231-5292.
26. Assadi A, Zade SH. UGA: A New Genetic Algorithm-Based Classification Method for Uncertain Data. Mid-Est J Scient Res. 2014;20(10):1207-1212.
27. Pei M, Goodman ED, Punch WF, Ding Y. Genetic algorithms for classification and feature extraction. In: Classification Society Conference. 1995 [cited 2015 June] Available from: URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.4377&rep=rep1&type=pdf
28. Yeganeh M., Ahmadabadi MES, Madadi Y. An Accurate Classification Algorithm with Genetic Algorithm Approach.Int J Comp & Info Tech. (IJOCIT) 2013;1(3):198-210.
29. Smith SL, Cagnoni S. Genetic and Evolutionary Computation: Medical Applications, Wiley: 2010.ISBN: 978-0-470-74813-8.
30. Laetitia J, Dhaenens C, Talbi EG. A genetic algorithm for feature selection in data-mining for genetics. Proceedings of the 4th Meta-heuristics International Conference Porto (MIC' 2001)2001;1:29-34.
31. Yang L, Widyantoro DH, Ioerger T, Yen J. An entropy-based adaptive genetic algorithm for learning classification rules. IEEE Evolutionary Computation 2001;2:790-796.
32. Ahmed ABED, ElarabyIS. Data Mining: A prediction for Student's Performance Using Classification Method. World J Comp Appl Tech. 2014;2:43-47.
33. Fidelis MV, Lopes HS, Freitas A. A Discovering comprehensible classification rules with a genetic algorithm. Evolutionary Computation, Proceedings of the 2000 Congress on 2000;1:805-810.
34. Gundogan KK, Alatas B, Karci A. Mining Classification Rules by Using Genetic Algorithms with Non-random Initial Population and Uniform Operator. Turk J Elect Eng& Comp Sc 2004;12(1):43-52.
35. Al-Maqaleh, Basheer M., Shahbazkia H. A genetic algorithm for discovering classification rules in data mining. Int J Compu Appli 2012;41(18):40-44.
36. Ba-Alwi MF. Knowledge Acquisition Tool for Classification Rules using Genetic Algorithm Approach. Int J CompuAppl 2012;60(1):0975-8887.
37. Ismaeel AG, Ablahad AA. Novel Method for Mutational Disease Prediction using Bioinformatics Techniques and Back propagation Algorithm. arXiv preprint arXiv:1303.0539, 2013.
38. Yutaka S. The truth of the F-measure. Teach Tutor Mater. 2007;1-5.
39. BekkarM,Djemaa HK, Alitouche AT. Evaluation Measures for Models Assessment over Imbalanced Data Sets. J Info Eng and Appl 2013;3(10):27-38.
40. Karimollah TH. Receiver Operating Characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspi J Intern Med 2013;4(2):627-635.
41. MATLAB version 7.10.0.499 (R2010 a) Natick, Massachusetts. The MathWorks Inc. Available From: URL: http://www.mathworks.in/products/matlab/.
42. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank**.** Nucl. Acids Res. 2012; 40(Database issue): D48–D53.doi:10.1093/nar/gkr1202