Data Mining Algorithms for Classification of Diagnostic Cancer Using Genetic Optimization Algorithms

Rafaqat Alam Khan
Lahore Garrison University,
Lahore, Pakistan

Taseer Suleman
Lahore Garrison University,
Lahore, Pakistan

Muhammad Sajid Farooq
Lahore Garrison University,
Lahore, Pakistan

Muhammad Hassan Rafiq
Lahore Garrison University,
Lahore,Pakistan

Muhammad Arslan Tariq
Lahore Garrison University,
Lahore,Pakistan

ABSTRACT
The breast tumor is the primary driver of female casualty everywhere throughout the world and the real area of study from a long time but with slighter development than anticipated. Numerous establishments and associations are working in this field to prompt to a conceivable arrangement of the issue or to prompt to additionally comprehension of the issue. Numerous past inquiries about the said were contemplated for improved comprehension of the issue and the research performed previously was to reduce dimensionality and to contribute to the betterment in the field of cancer, Wisconsin-Madison Diagnostic Breast cancer (WDBC) dataset was taken from learning repository of UCI database with 569 distinct instances for training by choosing finest features out of 32 different attributes. Different feature selection algorithms were used with data mining algorithms for better classification. Numerous enhancements in classification accuracy of WDBC were discovered by utilizing distinctive methodologies than the prior reviews directed in a similar field. The Logistic Regression, Linear Regression, and SVM algorithms showed better classification accuracy i.e. 98.24 %, 98.24 % and 98.07 % than the previous outcome results known for the said classification algorithms. The results were generated using 10 fold cross validation, by using different classification algorithms with feature selection and generation algorithms.

General Terms
Machine Learning, Data Mining, Classification, Genetic algorithms, Feature Selection, Algorithms, and Cancer.

Keywords
SVM, Logistic Regression, Linear Regression, Accuracy, Benign, Malignant.

1. introduction

Motivation: Breast tumor is the 2nd [1] top reason of demise in the female, by the number of new cases analyzed. Breast cancer has two different classes i.e. benign and malignant. A benign tumor, usually known as non-cancerous is left alone by a doctor without removing it because it is not that much aggressive towards the other tissue, but occasionally it may grow causing pain or other problems. On the other end, Malignant or known as cancerous is too much aggressive and has the capability to damage the surrounding tissues and if the patient is diagnosed with malignant than doctor performs a biopsy to find out the aggressiveness or severity of the tumor.

In this paper, numerous types of classification algorithms are used with feature optimization algorithms to differentiate between two types of breast tumor i.e. Benign and Malignant. The result has been found out with feature selection and without feature selection algorithms. There are 10 different data mining algorithms used with 4 distinct genetic algorithms. The result to find out the accuracy of the tumor data set is almost improved as compared to previous research performed in this similar field. In almost all cases the accuracy is improved but in the random forest, it is improved up to 5 % with feature selection algorithm while in some cases it has slightly improved.

1.1 Naïve Bayes

This classification algorithm is a regulated learning machine. In the RM tool example set is used as an input and produces output as a Boolean value which is by default true. In that case, to lessen the impact of zero probabilities, Laplace revision is used. Estimated normal distribution is used to return to the desired classification algorithm.

Bayesian Rule

$$P(S \mid A) = \frac{P(A \mid S)P(S)}{P(A)}$$

$$Posterior = \frac{prior \times likelihood}{Evidence}$$

1.2 Logistic Regression

The rapid miner tool for Logistic Regression takes the example to set as an input and as an output, it returns a

Manuscript received December 5, 2017
Manuscript revised December 20, 2017

Boolean value. False is the default value for the output. To determine whether to include an intercept, add intercept is used and to find out the performance the performance model is used. (Number; 1-+1; default: 10000), after evaluation population is started and stopped; the generation is stopped if no improvement occurs. In logistic Regression when (- 1: improves until max iterations). (Number; - 1-+1 default: 300) for keeping the example set improvement in the input object would occur. The other different parameters used are time, apply count, mutation, local random seed, selection, loop time and cross over.

1.3 Support Vector Machine

In Rapid miner, this learning algorithm used mySVM by Stefan RA ¼ ping and its implementation is in JAVA. This learning algorithm tool is used for both classification and regression and performed better and fast results for different learning task examples.

It takes input as an example set and on the output, it gives the model, estimated performance, weights and an example set. There are different parameters for SVM like Kernel type, Kernel Gamma, Kernel Sigma, calculate weights and much more.

1.4 Linear Regression

In Rapid miner, this classification algorithm used Akaike criterion for the model selection. The different parameters used for linear regression are feature selection which is used during regression; eliminate collinear features used to indicate that this algorithm is trying to delete the collinear features during regression, use bias, min standard coefficient, and ridge.

2. literature review

In this paper [2] different classification algorithms are applied on WDBC data set i.e. Naïve Bayes, Logistic Regression, and Decision Tree Algorithms along with feature selection algorithm Pearson Correlation Coefficient (PCC) to find out the accuracy of the cancer dataset. The different classification accuracy results for Naïve Bayes, Logistic Regression, and Decision tree are 94.40 %, 97.90% and 96.50 respectively.

In this [3] authors has found out accuracy of WDBC data set using Binning technique. It had to find out the accuracy of data set with binning and without binning concept. The feature extraction algorithm used for the extraction of features was PCA algorithm. The PCA algorithm is used with Classification algorithms Naïve Bayes, SVM, and ensembles and found the accuracy as 95.16 %, 95.53 %, and 95.9108 % respectively. This PCA algorithm is also used with these classification algorithms using binning technique and reveals that out of these three algorithms Naïve Bayes algorithm performed the best with a maximum accuracy of 97.3978 % with only five features and time complexity of 0.102 milliseconds which are far better than the other two classification algorithms.

In [4] three different classification algorithms i.e. ANN, PSO classifier and GA-Classifier are applied on three different Wisconsin data set i.e. WDBC, WBC and WPBC to find the accuracy of these three datasets along with feature selection algorithms and without feature selection algorithms. For WBC data set PS classifier performed better than the others two, While for WDBC and WPBC ANN performed better than the remaining two applied.

In the paper [5], UCI machine learning repository data set for WDBC is taken. They have found out the accuracy of data set by using SVM classifier with feature selection algorithms. The WDBC dataset accuracy with training test partitions with highest accuracy classification i.e. 98.5 % (50-50 %), 99.02 % (70-30 %) and 99.51 % (80-20 %) for training test partition.

In [6], the objective of this paper was a comparative analysis of different classification algorithms i.e. Bayesian Network, SVM, Back Propagation Neural Network and linear programming was applied on WDBC dataset having 569 instances with 357 malignant and 212 benign cases. Each dataset instance consists of 30 features using 10 cross fold validation. In this study, Naïve Bayes classifier achieved an accuracy of 89.55 %.

In [7] authors have implanted approach, choosing the best subgroup of elements is carried out amid the model development prepare. Ant colony algorithm, decent measure of research on breast cancer disease data sets utilizing highlight determination techniques is found in writing, for example, such as ant colony algorithm, In [8] swarm optimization technique of discrete particle, In [9] genetic algorithm along with wrapper method, In [10] SVM and linear discriminate is used with support vector based feature selection, In [11] FCBF multi thread based feature selection and DDC- DIC.

3. data analysis

The dataset used for analysis was taken from the UCI Machine Learning Repository [12][3] i.e. WDBC (Wisconsin Diagnostic Breast Cancer) Dataset.

The dataset consists of 569 observations having 32 attributes, divided into two classes. The two classes malignant and benign have 357 and 212 cases respectively. The data is gathered by Dr. Wolberg [13][14] since 1984. For classification and regression analysis of the data, this data set

is widely used. The dataset consists of instances of patients suffering from both benign (non-cancerous) and malignant (cancer with high risk).

Table 1: Wisconsin Diagnostic Breast Cancer Dataset

Dataset	Attributes	Instances	Class
WDBC	32	569	2

For the computation of every cell nucleus, different feature values are taken i.e. ten real-valued

Table 2: Features of WDBC Data set

01) Radius (points on perimeter from center of mean distances)
02) Texture (gray-scale values standard deviation)
03) Smoothness
04) Area
05) Concavity
06) Perimeter
07) Concave points
08) Fractal Dimension
09) Symmetry
10) Compactness

Rapid Miner 5.5 [7] tool is used for the examination of data set. The input data given to the Rapid miner tool is in the form of CSV format and it gives the output result in the form of the accuracy of the dataset. The tool has distinct classifiers for classification of data and also has different genetic algorithms, for feature selection and each of them has the unique technique to select the best feature for data classification.

The information provided is utilized to anticipate the precision of the location utilizing diverse classifier. For the prior examination the element determination is not utilized and just classifiers are utilized to get the required exactness for each of the classifiers. The information is, of course, arranged yet this time diverse component choice systems are utilized for the upgrade of the outcomes or to check for any conceivable enhancements. The utilization of highlight determination methods likewise helps in dimensionality decrease as highlight diminishment. This also results in the optimization of memory and time efficiency. The classifiers tried for precision were the same as those utilized as a part of a prior review.

Distinctive element determination strategies from the ones earlier utilized were tried here for development in precision. The grouping precision now and again was not the same as the one determined in the earlier study so all things considered the change was dealt with as a rate change in exactness to check for advancement. The Naïve Bayes and calculated relapse characterization demonstrated changes from the earlier study.

The different genetic algorithms used for feature reduction and generation algorithms are AGA, GGA, YAGGA, and YAGGA2. The classifiers utilized were Random Forest, Linear Regression, Logistic Regression, Decision Stump, KNN, Decision Tree, Rule Induction, ID3, Decision Tree (Weight Based), Naïve Bayes and Random Tree.

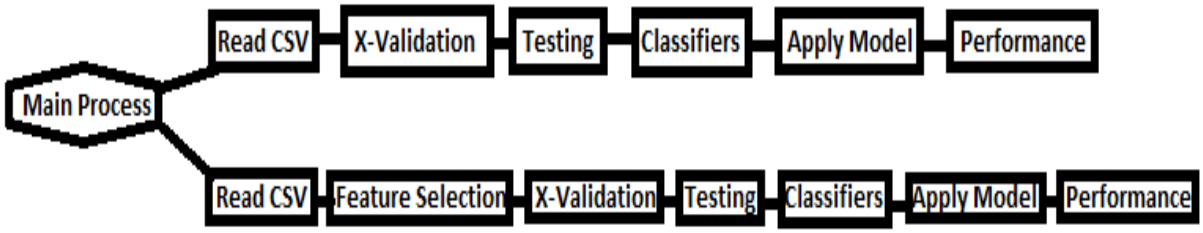

Fig. 1 Proposed Diagnostic Breast Cancer Model [15]

Table 3.WDBC Dataset Description [8]

Attribute Name	Attribute ID
Patient id	A1
Outcome	B1
TTR	C1
RADIUS1	D1
TEXTURE1	E1
PERIMETER1	F1
AREA1	G1
SMOOTHNESS1	H1
COMPACTNESS1	I1
CONCAVITY1	J1
CONCAVEPOINTS1	K1
SYMMETRY1	L1
FRACTALDIMENSION1	M1
RADIUS2	N1
TEXTURE2	O1
PERIMETER2	P1
AREA2	Q1
SMOOTHNESS2	R1
COMPACTNESS2	S1
CONCAVITY2	T1
CONCAVEPOINTS2	U1
SYMMETRY2	V1
FRACTALDIMENSION2	W1
RADIUS3	X1
TEXTURE3	Y1
PERIMETER3	Z1
AREA3	AA1
SMOOTHNESS3	AB1
COMPACTNESS3	AC1
CONCAVITY3	AD1
CONCAVEPOINTS3	AE1
SYMMETRY3	AF1
FRACTALDIMENSION3	AG1

4. Feature Selection Algorithms

The genetic algorithms on WDBC data set for feature selection and generation are discussed below.

4.1 GGA

The Generating Genetic Algorithm (GGA) is practiced on WDBC data set for feature selection and generation. The length of each attribute is modified after new attributes are produced and due to which inimitable mutation and crossover operators are used. For the selection of randomly, generators generator lists are used with Boolean parameters. An operator having no algorithm the example set is bound to only single attribute to extract it from value series. Ingo Mierswa is carrying out for the auto selection of features.

The GGA used Exa as an input example set i.e. to be classified and the exam is used to generate and select features. The three parameters used for the output are as per, Exa, and Att. Out of this 'per' is used for performance i.e. dataset accuracy, 'Att' for finding out the weight of attributes and 'Exa' for the output of Exa in.

4.2 AGA

This algorithm used for feature selection and generation and it is the modified version of GGA. It utilizes the same administrator as GGA yet this calculation includes extra generators and some essential intron aversion procedures are utilized to improve fundamental GGA. This administrator gives unmistakable outcome as the past one yet for the most part lower as a difference to YAGGA2.

4.3 YAGGA

The YAGGA algorithm does not change the individual length of the shorter or the drawn out one of them end up being better in wellness. This calculation unique in relation to the over the two methodologies utilized by creating new characteristics, with various probabilities.

- Feature vectors are added with newly generated attributes.
- The feature vector is added with original randomly selected attributes.
- Feature vector removes randomly selected attributes.

From that, we come to know that the length of the feature vector would increase and diminish. The original length would be obtained in such way if the smaller or larger length of the individual works as the best fit. From that, it ends into the increase and finishing of feature vector length.

4.4 YAGGA2

This calculation is same as YAGGA calculation yet in a demonstrated variant of highlight choice and era. These element choice and era calculations concede more choice for creating of components and render extensive strategies for the anticipation of the intron. This brings about to the fewer case sets and diminishment of elements.

5. Experimental Results

The dataset used in this study for classification was trained and then tested to find out the accuracy of the dataset and to differentiate them into two classes i.e. Benign and Malignant. After finding out the accuracy of the dataset the same classifiers are used with four different genetic algorithms for dimensionality reduction of the dataset to select the best features out of 32 different features keeping accuracy of the dataset into account. The features selected for different classifiers are different in number for different

classification algorithms are shown in the accuracy column of table 7.

The Table 4 performed comparative analysis between different classification algorithms with feature selection algorithm. The table shows that there are slight improvements in the accuracy except for the decision tree algorithm.

Table 4: Comparative analysis of proposed method with related work on WDBC dataset [1]

Classifiers (Reference)	Naïve Bayes	Log Regression	Decision Tree	This Study		
				Naïve Bayes	Log Regression	Decision Tree
Accuracy	94.40	97.90	96.50	96.66	98.24	96.31

Table 5: Comparative analysis of proposed method with related work on WDBC dataset [2]

Classifiers (Reference)	Naïve Bayes	Support Vector Machine	This Study	
			Naïve Bayes	Support Vector Machine
Accuracy	95.167	95.5390	96.66	98.07

Table 5 performed a comparative analysis of classification accuracy on the same data set i.e. UCI Machine learning repository. From the analysis of the proposed model performed in this paper provides a better result for Naïve Bayes and Support Vector machine than the previous research performed in the same field. For Naïve Bayes it has improved from 95.167 % to 96.66 % and that of for Support Vector Machine the accuracy is increased from 95.5390 % to 98.07 %, which is quite appreciable.

Table 6: Comparative analysis of proposed method with related work on WDBC dataset [4]

Classifiers (Reference)	ANN	PS Classifier	GA-Classifier	This Study		
				Linear Regression	Log Regression	SVM
Accuracy	97.3	97.2	96.6	96.66	98.24	98.07

Table 7 has the data of 10 different classification algorithms used along with 4 distinct genetic algorithms i.e. feature selection algorithms. The table shows the result of the accuracy of classification algorithms with and without feature selection algorithms. There are different types of genetic algorithms used for the proposed model was GGA, AGA, YAGGA, and YAGGA2. From the result, it can be concluded that the accuracy of the dataset is improved in each case when classification algorithm is used with feature selection algorithms i.e. Genetic Algorithms.

Table 7. Classifiers Result with and without Feature Selection Algorithm

S.No	Classifiers	Accuracy without Feature Selection Algorithms	Accuracy with Feature Selection Algorithms			
			GGA	AGA	YAGGA	YAGGA2
1	Naïve Bayes	93.32	96.66(6)	95.77(12)	95.60(12)	96.66(14)
2	Log-Regression	97.18	97.36(15)	97.54(16)	98.24(15)	98.07(11)
3	KNN(k=8)	93.31	94.01(11)	93.85(7)	94.91(6)	93.85(7)
4	Decision Tree	94.03	96.31(5)	95.76(14)	95.96(10)	95.95(10)
5	Decision Stump	89.98	91.91(3)	91.38(10)	91.73(13)	91.38(8)
6	Random Tree	88.56	93.65(9)	93.84(4)	93.66(10)	92.78(8)
7	Random Forest	92.63	95.95(5)	95.24(12)	94.55(10)	95.41(8)
8	Rule Induction	92.44	94.90(13)	95.08(12)	95.43(14)	95.78(6)
9	Linear Regression	95.61	96.83(9)	95.95(17)	98.24(15)	96.83(11)
10	SVM	97.36	97.54(8)	97.89(17)	98.07(15)	98.06(10)

6. Analysis and conclusion

Ten distinct classifiers with genetic algorithms were used. A large number of these classifiers were used as a part of past reviews with various feature selection algorithms. The accuracy of different classifiers with 4 different genetic algorithms is given in the above Table 7. The result of this study is much appreciable as compared to previous studies performed in the same area. From the results, it has been

noticed that three classifier i.e. Logistic Regression, Linear Regression and SVM its accuracy is much better than the other classification algorithms.

The feature selector and accuracy calculation classifiers used with genetic algorithms are YAGA, AGA, YAGAA2, and GGA. These genetic algorithms determine systems enhanced the element choice by choice of the better elements and they brought about change in the general precision of the grouping.

From the examination of the dataset it has been concluded that by using of feature selection algorithms along with different classification algorithms help in dimensionality reduction i.e. Decision Stump used with GGA select only three features, Random tree with AGA select four features and Decision tree with GGA used five features for decision making, this results in taking little time for training and testing of data set. The number of the feature selected by each genetic algorithms used with classification algorithm are also given in the above table 7.

7. Summary

Dataset used in this research was taken from Wisconsin Breast Cancer with 569 examples. The diagnostic breast cancer data set consists of two classes i.e. malignant and benign. In this study, four different feature selection algorithms are used along with ten distinct classification algorithms. For finding the accuracy of the data set genetic algorithms i.e. feature selection algorithms used with classification algorithms were different as compared to the previous research performed in this area. Results show a lot of improvement in the dataset accuracy which laid down the base for further research in cancer domain. Moreover, from the study results, it has been concluded that different classification algorithms used with genetic algorithms result in more dimensionality reduction i.e. selected fewer features for decision making and which in result would take less time for training and testing of the dataset.

References

[1] Wolberg, Street, Heisey, and Mangasarian. “Computer-derived nuclear ``grade" and breast cancer prognosis, Analytical and Quantitative Cytology and Histology”, 1995 Vol. 17, PP 257-264.

[2] S. K. Mandal 2017 “Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree” International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 2 Feb. 2017, Page No. 20388-20391 Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.40

[3] J.D. Malley, J. Kruppa, A. Dasgupta, K.G. Malley and A. Ziegler. 2012 “Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms” International Journal of Computer Applications (0975 – 8887) Volume 145 – No.2, July 2016.

[4] William H. Wolberg, Olvi Mangasarian, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA.

[5] A. Sh, Shahraki. H, Rowhanimanesh. AR, Eslami. S. “Feature selection using a genetic algorithm for breast cancer diagnosis: an experiment on three different datasets”. Iran J Basic Med Sci 2016; 19:476-482.

[6] Mehmet Fatih Akay,“Support vector machines combined with feature selection for breast cancer diagnosis.” Expert Systems with Applications 36 (2009) 3240–3247.

[7] Mierswa, Ingo and Wurst, Michael and Klinkenberg, Ralf and Scholz, Martin and Euler, Timm: YALE: Rapid Prototyping for Complex Data Mining Tasks, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006.

[8] A.d. MH, Ghasem. A. N, Ehsan B. M, editors.“Application of ant colony optimization for feature selection in text categorization”. Evolutionary Computation, 2008 CEC 2008(IEEE World Congress on Computational Intelligence) IEEE Congress on; 2008: IEEE.

[9] Murat.A, Under. A, “A discrete particle swarm optimization method for feature selection in binary classification problems”. 2010; 206:528-539.

[10] Kare Gowda. A. G, Jayaram. M, Manjunath. A. “Feature subset selection problem using wrapper approach in supervised learning”. 2010; 1:13-17.

[11] Youn. E, Koenig. L, Jeong. M. K, Baek. S. H. “Support vector-based feature selection using Fisher’s linear discriminant and Support Vector Machine. 2010; 37:6148-6156.

[12] Daisy. C, Subbulakshmi. B, Baskar. S, Ramaraj. N, editors. Efficient dimensionality reduction approaches for feature selection. Conference on Computational Intelligence and Multimedia Applications, 2007 International Conference on; 2007: IEEE.

[13] ftp.cs.wisc.edu cd math-prog/cpo-dataset/machine-learn/WDBC/

[14] Canadian Cancer Society’s Steering Committee on Cancer Statistics. Canadian Cancer Statistics 2012.Toronto, ON: Canadian Cancer Society; 2012. May 2012 ISSN 0835-2976.

[15] Dr. K.Usha, D.Lavanya “ Analysis of feature selection with classification: Breast Cancer Datasets” 2011 ISSN: 0976-5166 Vol. 2 No. 5 Oct-Nov 2011.

[16] Rafaqat Alam Khan, Nasir Ahmed, Nasru “Classification and Regression Analysis of the Prognostic Breast Cancer using Generation Optimizing Algorithms” International Journal of Computer Applications (0975 – 8887) Volume 68– No.25, April 2013

[17] Gouda. I. Salama1, M.B. Abdelhalim2, and Magdy AbdelghanyZeid3.2012 “Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers” International Journal of Computer and Information Technology (2277 – 0764) Vol 01– Issue 01, September 2012

[18] W. N. Street, Olvi L. Mangasarian, William H. Wolberg and Dennis M. Heisey “Computer-Derived Nuclear Features Distinguish Malignant from Benign Breast Cytology” 1995. vol 26, Num 7.