International Journal of Engineering & Technology, 7 (4.6) (2018) 223-226

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Prediction of Breast Cancer Using Big Data Analytics

K. Shailaja[1]*, B. Seetharamulu[2], M.A. Jabbar[3]

[1]M. Tech Scholar, Centre for Data Science, Vardhaman College of Engineering, Hyderabad, Telangana
[2]Professor, Department of CSE, Vardhaman College of Engineering, Hyderabad, Telangana
[3]Professor, Centre for Data Science, Vardhaman College of Engineering, Hyderabad, Telangana
**Corresponding author E-mail: kshailajasai@gmail.com*

Abstract

Big data is a phrase which is used to report collection of data that vast in size and still growing exponentially with time. It covers structured unstructured and semi-structured data. Now a day's big data is widely used in healthcare for prediction of diseases. Breast cancer is one of top cancer that occurs in a woman. It is the second main leading reason for the death of a woman in the United States and in Asian countries. If we identify this disease in early stages there is a better chance for curing. For this experiment, we used K nearest neighbor (KNN) algorithm for finding classification accuracy and it is implemented on R tool. We consider Wisconsin breast cancer (original) dataset taken from UCI machine learning repository.

Keywords: *Big data; Healthcare; Breast cancer; KNN; Wisconsin dataset.*

1. Introduction

Big data is a term utilized for maintaining a group of information that is large in size and it is growing continuously. It is represented by five 5v's they are volume, variety, velocity, variability, and veracity. Big data handle structured, unstructured and semi-structured data. Structured data contains words and numbers, so it can be easily identified and analyzed. Unstructured data contains difficult information and it can't be analyzed easily, because the data is represented in the form of images [1].
Big data analytics provides the best solution for handling, storing, and analyzing a large number of mammographic pictures. Big data life cycle has a leading prediction structure. It improves the clinical adversity style by stimulating prediction structure, estimating statistical improvements and determining various disease patterns. It also provides better solutions for handling healthcare data.

1.1 Breast Cancer:

Breast cancer is one of the most dangerous diseases due to this most of the women's died every year. Some tumors present in the breast they may be cancerous (malignant) and noncancerous (benign). Benign tumors can't be extended to remaining components of the body and also these tumors are not harmful to the body. After removing these tumors they don't grow again. Malignant tumors are very dangerous and these are spread to the remaining components of the body, after removing this tumor it will grow again.
This paper is systematized as follows: 2nd section describes literature review, the 3rd section describes related work, and section 4 presents the proposed work and section 5 deals with the experimental results which are achieved. Finally conclusion of this paper described in section 6.

2. Literature Review

Big data could be an extensive word for datasets so huge or complicated that traditional processing applications are insufficient. To avoid diseases, spot business trends, and conflict crime etc. we will analyze datasets to realize the new correlations. Governments, scientists, and hospitals will face many difficulties by utilizing complex datasets. By using different techniques of machine learning and data mining we can build efficient and powerful classifiers for huge databases [1].

2.1 Breast Cancer

Cancer is one of the dangerous diseases produced by uncontrolled partition of aberrant cells in the portion of the body. Different types of cancer are present in the earth. Breast cancer is one of the top cancer that occurs in a woman and it is the second main leading reason for woman in the United States and in Asia countries. Breast cancer begins once cells within the breast start to spread out of regulation. These cells typically kind a tumour and it can be observed on x-ray [2]. If the tumour is deadly (malignant) then the cells will expand into close tissues or expand to different elements of the body.

2.2 Signs & Symptoms

Some of the major symptoms of breast cancer are:

- Lumps present in the breast with any volume, texture, lineament, and with soft or hard edges.
- Pain in the breast.
- The complete or some portion of the breast suffered from swelling, redness
- There may be some changes occur in the nipple that is nipple retraction, ulceration, itching.
- Bleeding from the nipple that may be in any color.

- Thickening of the nipple.

2.3 Stages of Breast cancer:

Breast cancer has four stages. Stage 0: Tumors are not expanded to the neighboring tissue of the breast. Stage 1: Small tumor with the size of 2cm but it is not expanded to lymph glands. Stage 2 splits into 2a,2b; In stage 2a small tumor which has the 2cm size that is spread to the lymph glands and tumor from 20mm (2cm) to 50mm (5cm) is not expanded to the axillary lymph glands. Stage 2b: Tumor greater than 5cm or tumor from 2cm to greater than 5cm then spread into the one or three axillary lymph glands. Stage 3a: Tumor greater than 5cm expanded to 5 to 10 axillary lymph glands that are knitted with each other or with the neighboring tissues. Stage 3b: tumor expanded to the breast wall, skin or internal lymph glands. In this stage, it is not expanded to the remaining components of the body. Stage 3c: Tumor with any size that has expanded to more lymph glands. Stage 4: any size of the tumor has been expanded to the remaining organs such as the distant lymph nodes, bones, brain, lungs, liver and chest wall [2].

2.4 Feature Selection

Feature selection is nothing but it is the method of removing features from the given dataset which are unrelated [3]. It is used in various fields like pattern recognition, data mining, machine learning. If the dataset contains a large number of features it is very difficult to understand such a huge amount of data in some situations. Due to this most of the researchers utilized various techniques of feature selection. The major aim of this process is eliminating the redundancy, increase accuracy by removing unrelated and probably unnecessary features. Hence by removing unrelated data, it decreases the time complexity [4]. It is split into the wrapper, filter, and hybrid approaches.

I. Filter:
These methods heavily depend on the mathematical measures which achieve the characteristics between different features [6]. These approaches are light-weight and it is having less computational expenditure when compared to different methods. In this approach, rank is assigned to every feature depend on the data value of the given feature. Few algorithms of filter approaches are the fast correlation-based feature selection (FCBF), correlation-based feature selection (CBF), symmetrical uncertainty (SU) and so on [5]. The benefits of filter approaches are it is having a simple structure and in this strategy, mining algorithms are independent hence it performs the feature selection only once. But these methods cannot communicate with the classifier [7].

II. Wrapper:
In wrapper method for every attribute in the dataset classification algorithm applied and it uses the classification outcome; by applying algorithm it evaluates the attribute subspace [8]. Its computational cost is very expensive and also every fresh subset of options must build a hypothesis [7]. To calculate the feature significance in advance over testing dataset it uses different learning algorithms. The benefit of wrapper approach is it provides the communication between model selection and attribute subset search. Limitation of wrapper approach is it is very slow compared to filter method.

III. Hybrid:
The hybrid approach uses both wrapper and filter methods. Hence it takes the benefits of two approaches [6]. To determine the finest subsets hybrid technique utilizes the independent measure and to pick out the finest subsets it utilizes the learning algorithm.

2.5 Supervised Learning Algorithms

1. K-Nearest Neighbor (KNN):

KNN algorithm is also called as Instance-Based Learning. K-NN is the simplest approach for classification of samples. Here different distance measures are used for classifying samples. K-nearest neighbor finds the number of samples from training data which is near to the test samples and assigns to the frequent class label.
In this algorithm, training samples generate the classification rules without considering extra information. It has high probability when related instances belonging to the identical class. Based on K training samples KNN algorithm identifies the test samples. For all situations, K value will be a positive integer. In our dataset (WDBC) all the instances are present in between one (1) and ten (10), hence no need to calculate normalization between those instances. Because no any attribute will influence the others in the distance calculation of KNN. KNN widely used in the various decision support system for biomedical applications [9-11].

2. Support Vector Machine

Support Vector Machine (SVM) which is designed in 1990's. To achieve machine learning (ML) tasks support vector machine (SVM) is used, and it is a simple and prominent process. During this technique, a collection of training samples is given each sample is divided into different categories. Support vector machine (SVM) mainly used for classification and regression problems [12].

3. Decision tree

For classification issues, a decision tree is the most suitable algorithm. To arrange a tree structure this algorithm utilizes rules that are derived from the training dataset, and also these rules are enforced on validation data. It classifies an unknown label from a list of decisions. The Decision tree is very simple and it is easy to implement. Some of the well-known decision tree algorithms are C4.5, J48 and ID3 etc. The Decision tree is divided as regression tree and classification tree [13].

4. Naive Bayes

Naive Bayes shows better performance in classification because of simple relations. It performs only one scanning of data and hence classification is easy [14]. Naive Bayes performs conditional probability based on given class label.

5. Sequential Minimal Optimization (SMO)

This algorithm is used for solving the problems of quadratic instructions and these problems have occurred while training data of Support Vector Machine [15]. This algorithm splits the problem into a set of subproblems later that are resolved analytically. Hence Sequential Minimal Optimization is frequently utilized for training the SVM.

6. Expectation Maximization algorithm (EM)

EM algorithm is used to calculate the maximum likelihood to evaluate the existence of absent or invisible data. It is an iterative process that contains two steps that is expectation step (E-step) and a maximization step (M- step). In the E-step hidden information is evaluated for the given observed information [17]. In the M-step for the known of missing data assumptions are maximized for the likelihood function.

3. Related Work

K. Sivakami [15] uses Decision tree and Support Vector Machines (DT-SVM) both of these are hybrid methods. To introduce a disorder status prognosis they employ DT-SVM methods. The experiment was performed through Weka tool. The authors have considered the Wisconsin breast cancer dataset that includes 699 instances; in that 458 instances belong to not cancer (benign) class and other 241 instances belong to cancer (malignant) class. Finally, the author compared the outcomes of the DT-SVM model with Naive Bayes (NB), instance-based learning (IBK), and sequential minimal optimization (SMO) and conclude that DT-SVM gives better accuracy i. e 91% compared to NB, IBK, and SMO.

G. Sumalatha and S. Archana [16] have used j48 decision tree algorithm for the classification of breast cancer patients. The authors have utilized Weka tool for they experiment and the dataset contains 238 instances with 10 attributes along with the class label. They conclude j48 decision tree the gives accuracy (95.37%), error rates, recall, and precision.

D.R. Umesh and B. Ramachandra [17] have utilized Expectation Maximization (EM) algorithm for identifying the breast cancer recurrence. To find out the classification accuracy they have used SEER dataset (surveillance, epidemiology and end results) which contains 2,20,811 instances with 17 attributes. The authors have performed their experiment through Amazon cloud computing environment (EC2) and declare expectation maximization algorithm gives 88.54% of accuracy.

Hiba Asri et al. [18] performed this experiment to determine the efficiency and effectiveness of various algorithms like Support Vector Machine (SVM), K Nearest Neighbor (K-NN), Decision Tree (C4.5), and Naive Bayes (NB). They utilized Wisconsin breast cancer (original) dataset taken from UCI machine learning repository contains 699 instances with 11 attributes. The experiment is performed on WEKA tool and outcomes show that the SVM gives higher accuracy 97.13% compared to K-NN, C4.5 i.e 95.27%, 95.13%.

4. Proposed Work:

For this experiment, we used KNN algorithm for predicting breast cancer risks. We used breast cancer Wisconsin (original) dataset that is considered from the UCI machine learning algorithm. The dataset includes 699 instances and 10 attributes along with the class label and it contains missing values (?) which are replaced by the mean values of the attributes. The distribution of class will be 458 (65.5%) instances belong to the benign class and other 241 (34.5%) instances belong to the malignant class.

Dataset description:

Table 1: Summary of breast cancer dataset [19]

s.no	Attribute Name	Domain
1	sample code number	Id number
2	Clump thickness	1-10
3	Uniformity of cell size	1-10
4	Uniformity of cell shape	1-10
5	Marginal adhesion	1-10
6	Single epithelial cell size	1-10
7	Bare nucleli	1-10
8	Bland chromatin	1-10
9	Normal nucleoli	1-10
10	Mitoses	1-10
11	Class	Benign-2 Malignant-4

Sample code number indicates id number and it is not useful for the experiment hence we removed from the dataset.

Clump thickness determines whether it contains single or multilayered cells.

Uniformity of cell size means in the given samples it determines the size of cells which are consistency.

Uniformity of cell shape: It recognizes marginal differences and determines the cell shapes.

Marginal adhesion: It evaluates how many cells present on the external of the epithelial and they are stick together.

Single epithelial cell size: It identifies the epithelial cells that are necessarily expanded and it also describes the uniformity of cells.

Bare nuclei: it computes the hypothesis of the bunch of cells that are not encircled by the cytoplasm.

Bland chromatin: it ranks the pattern of a nucleus from admirable to rude.

Normal nucleoli: it identifies either the nucleoli are tiny, hardly apparent or huge, most clearly visible.

Mitoses: mitoses illustrate the level of the mitotic state.

4.1 Evaluation methods:

For this experiment, we utilized K nearest neighbor (KNN) algorithm and it is implemented on R tool. R is widely used for the execution and it is free open source software. For computing statistics, graphics it is most frequently used software environment. The first version of R tool was developed by the Ross Ihaka and Robert Gentleman in the 1990's, for our experiment we used R 3.4.2 version. It is integrated software that includes many facilities they are for storing and handling the data is a very effective tool. It also performs the array and matrix calculations. For analyzing data it contains a huge group of intermediate tools [20]. In R tool there is a package available called as 'neighbr' for KNN algorithm. By applying KNN algorithm in R tool it provides better accuracy 97.65% compared to other methods.

To find attribute ranking we performed feature selection on WEKA tool. We treated feature selection successful when the accuracy of the algorithm increases or no change or remain same. We used symmetrical uncertainty (SU) attribute evaluation which is one of the most commonly used feature selection methods [21]. SU finds the correlation between the class and the attributes. It is defined as

$$SU= (H(Y)+H(Z).H(Y/Z))/(H(Y)+H(Z))$$

Where H(Y), H(Z) are the entropies of Y and Z. These entropies are associated with every class value and attribute value. SU takes the values in between 0 and 1, where zero demonstrates two attributes that are unrelated and one demonstrates this single attribute can predict remaining attributes completely.

In Weka, we choose full training set for the attribute ranking. The list of ranking attributes for WDBC dataset given in below

=== Attribute Selection on all input data ===

Search Method:

Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 class):

Symmetrical Uncertainty Ranking Filter

Ranked attributes:

0.419 2 cell size
0.375 3 cell shape
0.366 6 bare nucleli
0.32 8 normal nucleoli
0.319 5 epithetical cell size
0.296 7 bland chromatin
0.286 4 marginal adhesion
0.234 1 clump thickness
0.206 9 mitoses

Selected attributes: 2,3,6,8,5,7,4,1,9 : 9

Ranking attributes are shown according to the selection of attributes that 0.419 is with the first rank displayed in second attribute i.e cell size, 0.206 is the last rank displayed in 9th attribute name as mitoses with least rank. To improve the accuracy we removed the least ranked attribute that is mitoses. After removing LRA, KNN algorithm gives the better accuracy of 98.14%.

5. Experimental Results:

This section describes the results of data analysis. Table 2 describes the accuracy of before feature selection and after feature selection for various k-values. Table 3 describes the performance measure which includes accuracy, precision, recall, and F-measure. Table 4 describes the confusion matrix for all k=5 values.

Table 2: Accuracy of various K-values

K-value	Accuracy before Feature selection	Accuracy after Feature selection
1	94.85	95.31
2	96.97	95.79
3	97.65	96.77
4	96.97	98.14
5	97.65	97.65

Table 3: Performance measure for all K-values

K-Value	Accuracy	Precision	Recall	F-Measure
1	95.31	97.97	95.27	96.34
2	95.79	97.22	96.52	96.81
3	96.77	97.22	97.91	97.98
4	98.14	98.61	98.61	98.61
5	97.65	98.61	97.93	98.29

Table 4: Confusion Matrix for All K-Values

K-Value	**Confusion Matrix**		
1	Prediction	Benign	Malignant
	Benign	141	7
	Malignant	3	63
2	Prediction	Benign	Malignant
	Benign	140	5
	Malignant	4	65
3	Prediction	Benign	Malignant
	Benign	140	3
	Malignant	4	67
4	Prediction	Benign	Malignant
	Benign	142	2
	Malignant	2	68
5	Prediction	Benign	Malignant
	Benign	142	3
	Malignant	2	67

The following figure shows the comparison of our approach with different models

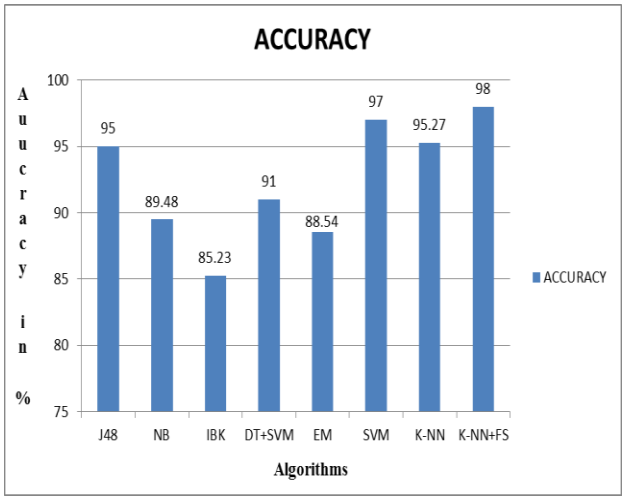

Figure 1: Comparative analysis of accuracy over reviewed & proposed method.

6. Conclusion

This paper deals with KNN algorithm to classify cancer tumors as either benign or malignant. We applied feature selection on the dataset to remove duplicate and irrelevant features. we applied symmetrical uncertainty attribute evaluation in WEKA for feature selection. Our proposed approach is evaluated and compared using Wisconsin breast cancer dataset. The experimental result showed that accuracy, precision, recall, and F-measure are increased by our proposed method when compared with different models. In future, we will work on feature selection techniques to improve the accuracy of the model.

References

[1] K. Shailaja et al., "Applications of Big Data Analytics: A Systematic Review", International Journal of Engineering Research in Computer Science and Engineering, volume 5, 2018.
[2] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. http://www.cancer.org/.
[3] Ms. Shweta Srivastava et al., "A Review Paper on Feature Selection Methodologies and Their Applications", International Journal of Engineering Research and Development, Volume 7, PP. 57-61, 2013.
[4] Abdur Rahman Onik et al., "An Analytical Comparison on Filter Feature Extraction Method in Data Mining using J48 Classifier, International Journal of Computer Applications, volume 13, 2015.
[5] Mitushi Modi et al., "An evaluation of filter and wrapper methods for feature selection in classification", International Journal of Engineering Development and Research, volume 2, 2014.
[6] Syed Imran Ali et al., "A Feature Subset Selection Method based on Symmetric Uncertainty and Ant Colony Optimization", International Journal of Computer Applications, volume 11, 2012.
[7] Sai Prasad Potharaju et al., "A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Datasets", Journal of Engineering Science and Technology Review, volume 6, pp.154-162, 2017.
[8] Bangsuk Jantawan et al., "A Comparison of Filter and Wrapper Approaches with Data Mining Techniques for Categorical Variables Selection", International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, 2014.
[9] MA Jabbar, "Prediction of heart disease using k-nearest neighbor and particle swarm optimization", Biomedical Research , volume 28, 2017.
[10] M Akhil Jabbar, et al., "Heart disease classification using nearest neighbor classifier with feature subset selection", Anale. Seria Informatica, volume 11 , 2013.
[11] M Akhil Jabbar et al., Classification of heart disease using k-nearest neighbor and genetic algorithm, Procedia Technology, volume 10, 85-94, 2013.
[12] K .P Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.
[13] A.Priyanga, "Effectiveness of Data Mining - based Cancer Prediction System", International Journal of Computer Applications, volume 10, 2013.
[14] .Animesh et al., "Study and analysis of Breast cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms", International Journal of Computer Applications, vol.2, 2016.
[15] K.Sivakami et al., "Mining Big Data: Breast Cancer Prediction using DT - SVM Hybrid Model", International Journal of Scientific Engineering and Applied Science, volume 1, 2015.
[16] G. Sumalatha et al., "A Study on Early Prevention and Detection of Breast Cancer using Data Mining Techniques", International Journal of Innovative Research in Computer and Communication Engineering, volume 5,2017.
[17] D.R Umesh et al., "Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach", International Journal of Computer Applications, volume 7, 2016.
[18] Hiba Asri, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis", The 6th International Symposium on Frontiers in Ambient and Mobile Systems, pp.1064-1069.
[19] Asuncion, A. & Newman, D.J. (2007). UCI Machine learning repository, http://www.ics.uci.edu/~mlearn/MLRepository.html, Irvine, CA: University of California, School of Information and Computer Science.

[20] https://www.r-project.org/
[21] Sai Prasad Potharaju et al., "A Novel M-Cluster of Feature Selection Approach Based on Symmetrical Uncertainty for Increasing Classification Accuracy of Medical Data sets", Journal of Engineering Science and Technology Review, volume 6, pp. 154-162, 2017.