Classification of Breast Cancer Data Using Machine Learning Algorithms

Burak Akbugday
Department of Biomedical Technologies
Izmir Katip Celebi University
Izmir, Turkey
burak.akbugday@gmail.com

***Abstract*—Machine learning has become an increasingly popular subject of study over the last two decades. Different algorithms have been applied to different types of datasets including medical ones. The literature already has numerous studies that involve medical data classification on different platforms. In this study, accuracies of three different machine learning algorithms, k-Nearest Neighbors (k-NN), Naïve Bayes (NB) and Support Vector Machine (SVM), have been investigated with Weka software. Results have been generated using different appropriate parameters for each algorithm respectively. k-NN and SVM algorithms were found to be the most accurate ones with identical confusion matrices and accuracy values.**

Keywords—machine learning; weka; breast cancer; classification

I. Introduction

According to the United States Centers for Disease Control and Prevention (CDC), cancer is the second leading cause of death in the United States, being responsible for 21.8 percent of overall fatal cases. CDC's records also indicate that among different cancer types, breast cancer is the tenth type of cancer that leads to death and it's more commonly seen in women.

One of the definitions for machine learning is it can be defined as a computational strategy which can be used to determine optimal solutions to a given problem without explicitly being programmed into a computer program by a programmer or an experimenter [1]. Over the last two decades, the use of machine learning algorithms has spread to several fields including medicine. With the help of advanced processing units, now it's possible to analyze medical data, which is very hard to analyze manually, via machine learning algorithms. There has been an increase in such studies over the last decade and more and more effective means to analyze medical data is introduced to the academic literature every day.

Breast Cancer Wisconsin (Original) [2] dataset is a public, widely used dataset in machine learning studies. There have been numerous studies published in the last couple of years that make use of this dataset. Vikas Chaurasia and Saurabh Pal compared the performance of various supervised learning classifiers [2] . Their results show that Support Vector Machine – Radial Basis Function (SVM-RBF) kernel is the most accurate classifier with 96.84% accuracy. S. Aruna and L. V. Nandakishore have made a similar comparison and found out that with 96.99% accuracy, Support Vector Machine (SVM) classifier was the best [3].

978-1-7281-2420-9/19/$31.00 ©2019 IEEE

In this study, Breast Cancer Wisconsin (Original) dataset has been used to investigate the effectiveness of Naïve Bayes (NB), k-Nearest Neighbors (k-NN) and Support Vector Machine (SVM) machine learning algorithms. The dataset contains 699 instances, 10 attributes and a class attribute which classifies instances as either malignant or benign cancer which indicate if a tumor is cancerous or non-cancerous respectively. All the classifications were done using Weka 3 machine learning software.

II. Methodology

A. Data Acquisition and Preprocessing

Dataset has been acquired from the UCI Machine Learning Repository website [2] for free as it's a public database donated for studies in 1992. After the acquisition, the dataset has been converted into *.csv file format so that it can easily be imported into Weka software via appropriate import tools. After successful import, the first attribute in the dataset has been deleted since it only contains ID numbers for the dataset. Also, the last attribute, the class attribute has been changed from "numerical" to "nominal" via a text editor to enable correct classification. After these steps, remaining attributes from 1 to 9 (9 included) have been used in classification.

B. Analysis Software

The classification has been done via an open-source software that is written Java called Weka. The latest version of the program is used at the time of classification. Weka has a collection of machine learning algorithms and it has tools for data preparation, classification, regression, clustering, association rules mining and visualization.

C. Classification Process

For classification, built-in algorithms in Weka software is used for the k-NN, NB and SVM classifiers. Relation of accuracy with k-values in k-NN algorithms is investigated, so, the dataset is classified using k-values from 1 to 10. For NB, default settings of the classifier are used. Finally, for SVM, different cost (C), gamma and different kernel functions are used to investigate their effect on accuracy. For all classifiers, k-Fold Cross-Validation is used for 10-folds. This has been done to prevent overfitting of data so that the model built by the classifier would classify a brand-new data correctly when presented.

III. RESULTS

Results that are related to the investigation of the effect of k-value in k-NN classifier are presented in Table I. According to those results, k = 3 condition has the highest accuracy.

TABLE I. CLASSIFICATION RESULTS AND ACCURACY FOR DIFFERENT VALUES OF K IN K-NN CLASSIFIER.

	Correct Classification	Incorrect Classification	Accuracy (%)
k = 1	665	34	95.14
k = 2	661	38	94.56
k = 3	667	22	96.85
k = 4	676	23	96.71
k = 5	676	23	96.71
k = 6	673	26	96.28
k = 7	675	24	96.57
k = 8	672	27	96.14
k = 9	673	26	96.28
k = 10	674	25	96.42

Table II is the confusion matrix for the k-NN classifier, k-values changing from 1 to 10, respectively. As expected, k = 3 condition has the lowest number of misclassifications in total in both class attributes.

TABLE II. CONFUSION MATRIX FOR DIFFERENT VALUES OF K IN K-NN CLASSIFIER.

k-Values	Classified As			
	a	*b*		
k = 1	443	15	**a**	**Classes**
	19	222	**b**	
k = 2	448	10	**a**	
	28	213	**b**	
k = 3	445	13	**a**	
	9	232	**b**	
k = 4	448	10	**a**	
	13	228	**b**	
k = 5	444	14	**a**	
	9	232	**b**	
k = 6	447	11	**a**	
	15	226	**b**	
k = 7	445	13	**a**	
	11	230	**b**	
k = 8	446	12	**a**	
	15	226	**b**	
k = 9	446	12	**a**	
	14	227	**b**	
k = 10	447	11	**a**	
	14	227	**b**	

Accuracy data of NB classifier can be found in Table III and the confusion matrix of the same classifier is presented in Table IV. Since there aren't many parameters to change in NB classifier, there's only a single result.

TABLE III. CLASSIFICATION RESULTS AND ACCURACY OF NB CLASSIFIER

Correct Classification	671
Incorrect Classification	28
Accuracy (%)	95.99

TABLE IV. CONFUSION MATRIX FOR NB CLASSIFIER

a	**b**	
436	22	**a**
6	235	**b**

For SVM classifiers, two different SVM classifiers are investigated, C-SVM and nu-SVM. Classification results and confusion matrices of C-SVM and nu-SVM classifiers can be found in Table V, VI, VII and VIII, respectively. According to Table V, the most accurate C-SVM result with 96.85% accuracy is obtained when $C = 2^{15}$ and $\gamma = 2^{-15}$ parameters are used.

TABLE V. RESULTS FOR C-SVM ALGORITHM FOR DIFFERENT COST (C), GAMMA PARAMETERS AND KERNEL FUNCTIONS.

	Correct Classification	Incorrect Classification	Accuracy (%)
$C = 1$ $\gamma = 2^{-3}$ **Polynomial Kernel**	652	47	93.28
$C = 2^{15}$ $\gamma = 2^{-15}$ **Radial Basis Kernel**	677	22	96.85
$C = 2^{10}$ $\gamma = 2^{-6}$ **Sigmoid Kernel**	226	473	32.33
$C = 2^{10}$ $\gamma = 2^{-9}$ **Sigmoid Kernel**	670	29	95.85

TABLE VI. CONFUSION MATRIX OF OUTPUT FROM C-SVM CLASSIFIER WITH DIFFERENT PARAMETERS

	a	*b*	
$C = 1$ $\gamma = 2^{-3}$ **Kernel = Polynomial**	441	17	**a**
	30	211	**b**
$C = 2^{15}$ $\gamma = 2^{-15}$ **Kernel = Radial Basis**	445	13	**a**
	9	232	**b**
$C = 2^{10}$ $\gamma = 2^{-6}$ **Kernel = Sigmoid**	225	233	**a**
	240	1	**b**
$C = 2^{10}$ $\gamma = 2^{-9}$ **Kernel = Sigmoid**	443	15	**a**
	14	227	**b**

And from Table VII, it can be seen that the most accurate nu-SVM classifier is generated with 94.85% accuracy when $C = 2^{15}$ and $\gamma = 2^{-15}$ parameters are used.

TABLE VII. RESULTS FOR NU-SVM ALGORITHM FOR DIFFERENT COST (C), GAMMA PARAMETERS AND KERNEL FUNCTIONS

	Correct Classification	Incorrect Classification	Accuracy (%)
$C = 2^{-1}$ $\gamma = 2^{-6}$ Polynomial Kernel	617	82	88.27
$C = 2^{15}$ $\gamma = 2^{3}$ Polynomial Kernel	617	82	88.27
$C = 2^{1}$ $\gamma = 2^{3}$ Radial Basis Kernel	514	185	73.53
$C = 2^{5}$ $\gamma = 2^{1}$ Radial Basis Kernel	600	99	85.84
$C = 2^{15}$ $\gamma = 2^{-15}$ Sigmoid Kernel	663	36	94.85
$C = 2^{-5}$ $\gamma = 2^{0}$ Sigmoid Kernel	351	348	50.21

TABLE VIII. CONFUSION MATRIX OF OUTPUT FROM NU-SVM CLASSIFIER WITH DIFFERENT PARAMETERS

	a	b	
$C = 2^{-1}$ $\gamma = 2^{-6}$ Polynomial Kernel	455	3	a
	79	162	b
$C = 2^{15}$ $\gamma = 2^{3}$ Polynomial Kernel	455	3	a
	79	162	b
$C = 2^{1}$ $\gamma = 2^{3}$ Radial Basis Kernel	273	185	a
	0	241	b
$C = 2^{5}$ $\gamma = 2^{1}$ Radial Basis Kernel	359	99	a
	0	241	b
$C = 2^{15}$ $\gamma = 2^{-15}$ Sigmoid Kernel	449	9	a
	27	214	b
$C = 2^{-5}$ $\gamma = 2^{0}$ Sigmoid Kernel	230	228	a
	120	121	b

IV. DISCUSSION

In this study, k-NN, NB and SVM classification algorithms have been applied to Wisconsin Breast Cancer dataset using Weka 3 machine learning software and the detailed results of the individual classification processes are provided in various tables in the Results section. The first subject of interest was the investigation of an optimal k-Value for a k-NN classifier, so the dataset is applied to a k-NN classifier with different k-Values and as a result, the most accurate result is obtained when k is set to be 3 with 96.85% accuracy (Fig. 1). When k = 3 it means that the classifier will look at 3 neighbors of a data point to determine its class, in this case, cancer or not cancer. k-NN with 3 neighbors were able to classify 445 out of 458 non-cancerous, benign, cases correctly and 232 out of 241 cases of cancerous, malignant cases correctly stated as confusion matrix in Table II. Considering k-NN is a lightweight, lazy learning algorithm with very short build times (0 seconds in this study), this level of accuracy is good in comparison to other classifiers.

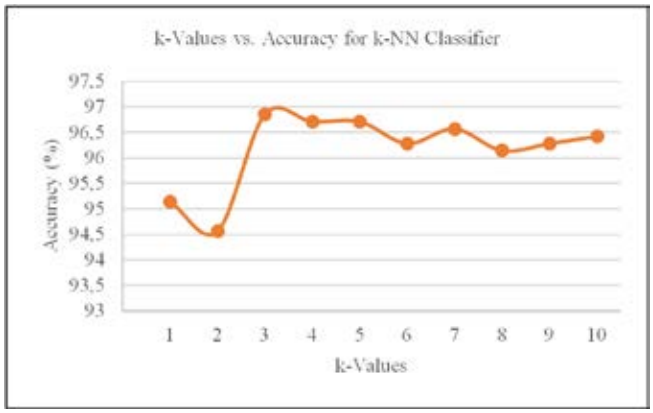

Fig. 1. The effect of k-values to accuracy in k-NN algorithm.

Naïve Bayes, another classifier that is used in this study has the accuracy of 95.99% (Table III) and were able to correctly classify 436 out of 241 benign and 235 out of 232 malignant cases (Table IV). Since this method doesn't have many parameters available to change in Weka program, further investigation was not possible but almost 96% of accuracy can also be considered as good.

Finally, Support Vector Machine algorithms have several sub-classifiers and numerous parameters available to change. In this study, the effect of cost (c) and gamma parameters to accuracy was the second subject of interest as well as the effect of different kernel types. Considering the results presented in Tables V to VIII, the best result was acquired from C-SVM classifier, with the same accuracy as 3-NN classifier, 96.85% (Table V). This result also aligns with the existing findings in the literature [2, 4]. Finding the most suitable parameters for SVM classifiers is a problem of optimization and is a subject of interest in many machine learning studies of their own. So, in this study, only values that are already known in the existing literature have been used [5]. Some accuracy values in Tables V and VII are low and those values indicate either the starting or the endpoint of a suitable range for those particular parameters, since in some studies C and ɣ values are used as a range and the results are plotted as a surface to visualize the best and the worst possible selections for those parameters.

The use of k-Fold Cross-Validation is almost a standard in machine learning studies alongside with (67-33%) split, so, a 10-Fold Cross-Validation option is preferred in all instances of classification to reduce possible biases that might occur while the algorithm is building a model to classify the dataset. Methods like k-Fold and splitting prevent especially overfitting which is a serious problem in machine learning studies and real-world applications as in the cases of overfitting, the model might fail to classify a new dataset when presented, so it should be prevented as much as possible.

V. Conclusion

Machine learning algorithms can be applied to numerous types of large datasets including medical ones. Application of machine learning algorithms to medical data is an emerging topic and, in this study, the effectiveness of k-NN, NB and SVM algorithms in breast cancer data classification has been investigated. Taking into consideration of all the results, k-NN with k = 3 condition and C-SVM with parameters $C = 2^{15}$ *and* $\gamma = 2^{-15}$ and that has Radial Basis Kernel is found to be the most accurate classification algorithms with 96.85% accuracy.

In future studies, same algorithms with same criteria can be implemented using different programming languages or environments such as Python or R. Since Weka is based on Java, and has all the algorithms built-in, use of another platform with better-implemented algorithms may lead to more accurate classifiers with better programming practices and/or platform-specific advantages.

Acknowledgment

The author would like to thank Aysenur Pamukcu, MSc, for her continued support during this study.

References

[1] M. Walter *et al.*, "Translational machine learning for psychiatric neuroimaging," *Prog. Neuro-Psychopharmacology Biol. Psychiatry*, vol. 91, pp. 113–121, 2019.

[2] W. H. Wolberg, "Breast Cancer Wisconsin (Original) Data Set", Center for Machine Learning and Intelligent Systems, Machine Learning Repository. [Accessed on May 5, 2019], Available from: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[3] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability," *Int. J. Comput. Sci. Mob. Comput. IJCSMC*, vol. 3, no. 1, 2017.

[4] S. Aruna and L. V Nandakishore, "K Nowledge B Ased a Nalysis of V Arious S Tatistical T Ools in D Etecting B Reast," pp. 37–45, 2011.

[5] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016.

[6] H. You and G. Rumbe, "Comparative Study of Classification Techniques on Breast Cancer FNA Biopsy Data," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 1, no. 3, p. 5, 2012.