ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DATA ANALYSIS BASED ON DATA MINING ALGORITHMS USING WEKA WORKBENCH

Layla Safwat Jamil*
* University of Baghdad, Iraq

DOI: 10.5281/zenodo.59630

ABSTRACT

In this paper, machine learning algorithms and artificial neural networks classification from instances in the breast cancer data set are applied by Weka (Data Mining Workbench). The data set were donated in 1988. This is one of three domains provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. The different algorithms and their results were compared with received calculations on Weka.

KEYWORDS: artificial neural networks, machine learning algorithms, breast cancer data set, Weka, classification.

INTRODUCTION

Machine learning is all about learning rules from the data set. In this paper, I use classification and analysis processes on the breast cancer dataset. Naïve Bayes classifier, SMO (support vector machine), decision tree, KStar (Instance-based classifier), artificial neural networks (ANNs) have been used in order to analyze the results.

Artificial neural networks (ANNs) supply a general, practical method for learning real valued examples. Algorithms such as backpropagation use gradient descent to adjust network parameters to be best fitted. ANN learning minimizes the errors in the training data and has been successfully applied to the problems such as interpreting visual demonstrations [1]. In this paper, machine learning algorithms determined above and ANNs used on retrieving medical data from breast cancer patients.

METHODS

This breast cancer data was attained by the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklic for providing the data. Breast cancer data and attributes are fetched from breast cancer patients given in the data set. This data set includes 201 instances of one class (no-recurrence events) and 85 instances of another class (recurrence events). The instances are described by 9 attributes, and one class attribute, some of which are linear and some are nominal.

So as to run of machine learning algorithms, WEKA (The Waikato Environment for Knowledge Analysis) software is used. WEKA workbench aids to apply machine learning techniques for myriads of real world problems [2]. The WEKA machine learning workbench provides an environment for automatic classification, regression, clustering and common data mining problems in bioinformatics research. It has a user friendly graphical interface to compare the various algorithm results [3].

ISSN: 2277-9655
Impact Factor: 4.116

The data set description with attributes, and their values are given below.

Attributes	Values
age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause	lt40, ge40, premeno
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,45-49, 50-54, 55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39
node-caps	yes, no
deg-malig	1, 2, 3
breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiat	yes, no
Class	no-recurrence-events, recurrence-events

Table (1): Attributes and values of breast cancer data [4]

The visualization of age and tumor-size attributes can be shown at the Figure (1) below.

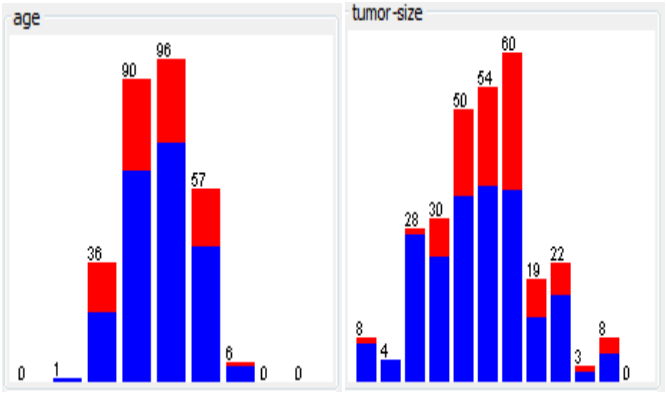

Figure (1): Age and tumor-size attributes on the breast cancer data set.

MACHINE LEARNING ALGORITHMS

Data mining, an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets methods at the intersection of machine learning such as in breast cancer data set. The main aim of the data mining process is to retrieve the data from data set, and transform into more meaningful form with the help of the algorithms.

In this paper, machine learning algorithms developed for data mining is used. These five algorithms are determined to analyze the results. The algorithms are Naïve Bayes classifier, SMO (support vector machine), decision tree, KStar (Instance-based classifier), artificial neural networks (ANNs). The WEKA workbench assists to retrieve the breast cancer data [5] for running the algorithms. Thus, we can easily realize the difference between the algorithm results. To sum up, the best classification on the breast cancer data set is understandable.

Naïve Bayes Classifier: Naïve Bayes is a statistical learning algorithm that applies a simplified version of Bayes rule in order to compute the posterior probability of a category given the input attribute values of an example situation. Prior probabilities for categories and attribute values conditioned on categories are estimated from frequency counts computed from the training data. Naïve Bayes is a simple and fast learning algorithm that often outperforms more sophisticated methods. The Bayesian classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the

model in a principled way by determining probabilities of the outcomes. It can solve diagnostic and predictive problems.

SMO *(Support Vector Machine):* SMO implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (In that case the coefficients in the output are based on the normalized data, not the original data, this is important for interpreting the classifier.)

Decision Tree: Decision tree analysis on J48 algorithm is applied to Weka. A decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values.

KStar: Kstar is an instance-based classifier that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.

Artificial Neural Networks (ANNs) Classifier

In this paper, multilayer perceptron networks (MLPs) classifier developed for data mining is used. The kind of multilayer networks learned by the backpropagation algorithm are capable of expressing a rich variety of nonlinear decision surfaces. To illustrate, a typical multilayer network and decision surface is demonstrated in the Figure (2). In a feedforward network information always moves one direction; it never goes backwards.

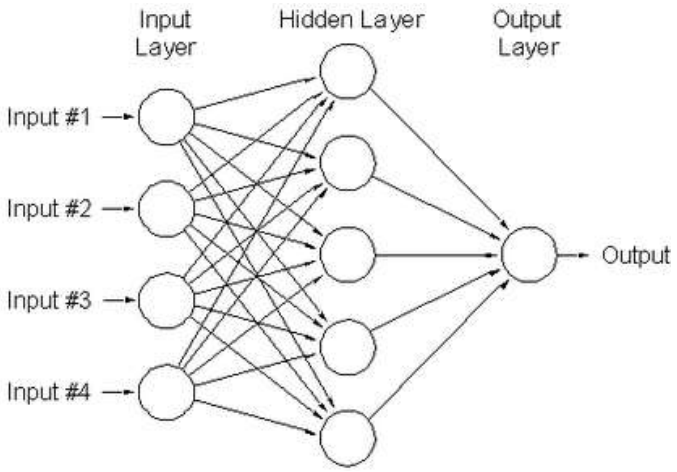

Figure (2): A feedforward Multilayer Perceptron (MLP)

Like other machine learning methods, neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming. An MLP network is composed of a number of identical units called neurons organized in layers, with those on one layer connected to those on the next layer, except for the last layer or output layer [6]. Indeed, MLPs architecture is structured into an input layer of neurons, one or more hidden layers and one output layer. Neurons belonging to adjacent layers are usually fully connected and the activation function of the neurons is generally linear. In fact, the various types and architectures are identified both by the different topologies adopted for the connections and by the choice of the activation function.

STUDY RESULTS

In this paper, Naïve Bayes classifier, SMO (support vector machine), decision tree, KStar (Instance-based classifier), and artificial neural networks (ANNs) have been analyzed on the Weka workbench.

ISSN: 2277-9655
Impact Factor: 4.116

Here, using machine learning algorithms and ANNs, analysis made with obtained experimental results from classifications on the breast cancer data set. Every machine learning algorithms have been applied separately on all data set, and classification results have denoted in Table (2).

The kappa statistic is used as a means of classifying agreement in categorical data. KS (Kappa Statistic) is used as a means of classifying agreement in categorical data. A kappa coefficient of 1 means a statistically perfect modeling whereas a 0 means every model value was different from the actual value. KS values for each algorithm have been calculated separately with the help of Weka functions.

Statistical results of the algorithms for breast cancer data set are given at Table (2).

Algorithms	Correctly Classified Instances (%)	Kappa Statistics	Mean Absolute Error
Naïve Bayes Classifier	71.67	0.2857	0.3272
SMO	69.58	0.1983	0.3042
J48 Decision Tree	75.52	0.2826	0.3676
KStar	73.52	0.2864	0.3354

Table (2): Performance analysis results of machine learning algorithms on breast cancer data set.

1- *Naïve Bayes classifier*, the first algorithm results are:

The correctly classified instances are 71.67% and incorrectly classified instances are 28.33%, the Kappa statistics is 0.2857, and the mean absolute error is 0.3272.

How to generate the results with "Naïve Bayes classifier" can be shown at the Figure (3) below.

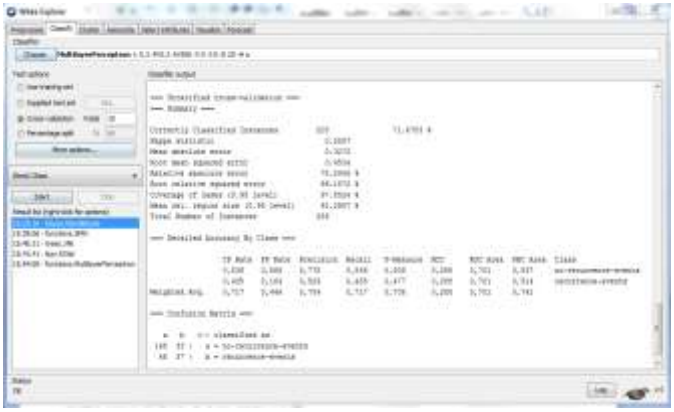

Figure (3): Naïve Bayes classifier results on Weka

2- *SMO*, the second algorithm results are:
The correctly classified instances are 69.58% and incorrectly classified instances are 30.42%, the Kappa statistics is 0.1983, and the mean absolute error is 0.3042.

3- *J48*, the third algorithm on decision tree, results are:
The correctly classified instances are 75.52% and incorrectly classified instances are 24.48%, the Kappa statistics is 0.2826, and the mean absolute error is 0.3676.

The decision tree visualization is shown at the Figure (4) below.

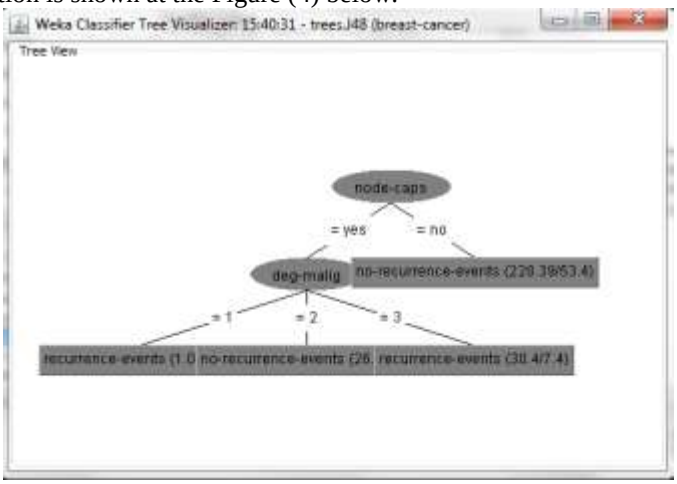

Figure (4): J48 algorithm result to get decision tree view.

4- *KStar*, the fourth algorithm results are:
The correctly classified instances are 73.52% and incorrectly classified instances are 26.48%, the Kappa statistics is 0.2824, and the mean absolute error is 0.3354.

In conclusion, the best and worst conditions are taken, regardless to applied machine learning algorithms. KStar algorithm realized best modeling with KS (Kappa Statistics) = 0.2864 value in making classification on data set. SMO (support vector machine) algorithm realized worst modeling with KS = 0.1983 value in making classification on data set.

J48 algorithm with % 75.52 accuracy classification percent rate realized to best classifications on data set. SMO algorithm with % 69.58 accuracy classification percent rate realized to worst classifications on data set.

Mean absolute error of J48 algorithm has fixed to maximum value (0.3676). Mean absolute error of SMO algorithm has fixed to minimum value (0.3042).

5- *Artificial Neural Networks (ANNs) Classifier*
The algorithm has two more parts, the first one is “epoch”, the second one is “learning rate”.
“Epoch” defines as the number of iterations over the data set in order to train the neural network. If we change the epoch number, this means that we have more rigid results in different conditions.

At each training step the network computes the direction in which each bias and link value can be changed to calculate a more correct output. The rate of improvement at that solution state is also known. A “learning rate” is user-designated in order to determine how much the link weights and node biases can be modified based on the change direction and change rate. The higher the learning rate (max. of 1.0) the faster the network is trained [7]. However, the network has a better chance of being trained to a local minimum solution. A local minimum is a point at which the network stabilizes on a solution which is not the most optimal global solution. Thus, I prefer to adjust learning rate, regardless to diverse epoch numbers.

ISSN: 2277-9655
Impact Factor: 4.116

Hidden Layer	Epoch	Learning Rate	Correctly Classified Instances (%)	Kappa Statistics	Mean Absolute Error
1	500	0.3	71.32	0.2637	0.3402
1	1000	0.3	72.02	0.2816	0.3397
2	500	0.3	73.07	0.2851	0.317
2	1000	0.3	73.42	0.2971	0.3194
2	1000	0.2	73.07	0.3258	0.3188
3	500	0.3	72.37	0.2775	0.3133
3	1000	0.3	72.37	0.2828	0.3138
3	1000	0.2	73.77	0.3386	0.3198

Table (3): Performance analysis results of ANN (Artificial Neural Network) algorithm on breast cancer data set.

For realized the best classification rate of ANNs algorithm was made by changing the classification parameters of ANNs algorithm. The best accuracy classification percentage of ANNs (Multi Layer Perceptron) algorithm realized with % 73.77 on data set in Table (3). Here, the number of hidden layer of ANNs algorithm is 3, the number of epoch of ANNs algorithm is 1000, and learning rate of ANNs algorithm is 0.2 value.

RESULTS AND DISCUSSIONS

When the number of instances decreased, falling in performance of algorithms is observed. Performances of algorithms have been increased on large number instances. Data mining indicates high performance on large dimension databases. Thus, applied machine learning algorithms have been indicated high performance on large dimension databases.

Their ability to learn by example makes artificial neural networks very flexible and powerful. There is no need to devise an algorithm in order to perform a specific task; i.e. there is no need to understand the internal mechanisms. Along various other advantages of artificial neural networks, there are disadvantages too. They cannot be programmed to perform a specific task; the examples must be selected carefully, otherwise, useful time is wasted or even worse the network may be functioning absolutely incorrectly.

While some of the algorithms show high performance, and some of them show poor performance. While some of the algorithms make the best modeling and some of them make the worst modeling.

REFERENCES

[1] Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994. *Machine Learning, Neural Statistical Classification*, Ellis Horwood.
[2] Witten, I.H. & Frank, E. 2000. *Weka Machine Learning Algorithms in Java*, in Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, pp. 260-320.
[3] Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I.H. *Data Mining in Bioinformatics using Weka*, Bioinformatics Applications Note, pp. 2475-2481, (2004)
[4] The data set are retrieved from the website below: *http://archive.ics.uci.edu/ml/datasets/Breast+Cancer*
[5] The arff file data was taken from the website below: https://www.lri.fr/~antoine/Courses/Master-ISI/TD-TP/breast-cancer.arff
[6] Yusuf, U. & Gülay, T., *Rule learning with Machine Learning Algorithms and Artificial Neural Network.* Journal of Selçuk Unıversıty Natural and Applied Science, vol.1 no.2, 2012.
[7] Tan, M. & Eshelman, L. 1988. *Using Weighted Networks to Represent Classification Knowledge in Noisy Domains.* Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.