Critical Analysis of Six Frequently Used Classification Algorithms

Taram Nayab Shah, Muhammad Zakir Khan, Mumtaz Ali, Bilal Khan, Hammad Muhammad
City University of Science and Information Technology, Peshawar.
E-mail: Taramshah82@gmail.com

Abstract: *Classification is the technique used to categorize the data into a given number of classes. The main goal of classification is to identify the category to which a new data will fall under. In other words, we can say that classification is the process of generalizing data according to different instances. This paper puts together the most frequently used classifications algorithms. The algorithms include are Logistic Regression, Linear Discriminant Analysis, Naïve Bayes, Decision Tree, K-Nearest Neighbors and Support Vector Machine. These six algorithms on classification problems are the starting point to explore the classification. We applied these six algorithms on the Ionosphere and diabetes dataset for binary classification. Out of six, the decision tree has surprisingly given better results as compared to others. That is 89.46% on the Ionosphere and 77.47% on diabetes respectively.*

Keywords: *Ionosphere Diabetes Problem, Machine Learning Algorithms, Classification Algorithms*

1. Introduction

Data mining is the process of sorting large data set into meaningful form to improve understanding and establish a relationship to solve a problem using analysis of data [1]. Nowadays data is increasing exponentially and to deal with such a huge amount of data needs data mining techniques to search, analysis in order to find the meaningful patterns within that data [1]. Classification is describing as supervised learning in which the class label is known in advance while clustering is describing as unsupervised learning. Classification techniques are most frequently used by machine learning and data mining problems [4]. A number of algorithms are available in classification like Decision Tree [3], Logistic regression [2], Neural Network [5], etc. Among these techniques, the decision tree algorithm is used mostly in research because it is easy to implement and understand due to its flow-chart-like tree structure. In this paper, the main focus on six top machine learning algorithms named Logistic Regression, Linear Discriminant Analysis, Naïve Bayes, Decision Tree [3], K-Nearest Neighbor and Support Vector Machine that can frequently use on classification problems.

2. Literature Review

2.1 Most Frequently Used Classification Algorithms

These six algorithms on our classification problem are the starting point. We will apply these six algorithms on the Ionosphere and diabetes problem of binary classification. This is chosen because of numerical in nature and having two classes to discriminate. Each instance describes the properties of radar returns from the atmosphere and the task is to predict where the structure of the ionosphere is according to ionosphere or not? The data set is taken from UCI, a total of 35 numerical attributes, 351 instances and having 98% accuracy.

2.1.1 Logistic Regression

Logistic regression analysis is one of the most preferred regression methods that can be implemented in modeling binary dependent variables. Logistic regression is a mathematical modeling approach used to define the relationship between such independent variables as X1, X2, …, Xn and Y binary dependent variable which is coded as 0 or 1 for two possible categories. The independent variables may be continuous, discrete, binary or a combination of them. This function is also called sigmoid function. Inputs values are combined using weight or coefficient values to predict the outputs.it is different than other methods as the output value being modeled is binary instead of continuous [16]. Logistic regression is a very simple and fast technique but sometimes very effective for some problems. In the results, it is observed that Logistic Regression achieves an accuracy of 88%.

2.1.2. Linear Discriminant Analysis

It is a very common technique for dimensionality reductions pre-processing steps for machine learning and pattern classification application. At the same time, it is usually used as a black box but sometimes not well defined [17]**.** It is a generalization of Fisher's linear discriminant, a method used in machine learning to find the linear combination of features that characterizes two or more classes. LDA is closely related to Principal Component Analysis in which LDA explicitly attempts to model the difference between the classes while PCA does not take any account any difference in class. The result shows that Linear Discriminant Analysis achieves an accuracy of 88%.

2.1.3. Naïve Bayes

Naive Bayes is a classification algorithm that is based on Bayes theorem with strong and naïve independence assumptions. It simplifies learning by assuming that features are independent of the given class [18]. The theorem states that "the probability of an event occurring giving the probability of another event that has already occurred". In this theorem, the calculation probability for each hypothesis is simplified for own calculation. It is also a classification algorithm and traditionally it assumed that the attribute taken is nominal. It gives a prediction of the next highest probability class. It can support both binary and multiclass classification problems. You will see in the result that Naïve Bayes achieves an accuracy of 82%.

2.1.4. Decision Tree

A decision tress starts with the single node which branches into possible outcomes that leads to additional nodes, which branch off into other possibilities. This gives a tree-like shape. They can be used to understand the non-linearity and map out an algorithm that predicts the best choice mathematically. Decision Tree can support both classification and Regression problems. It takes to start from the root and then narrows down to the leaf. The depth of the tree can be defined through Weka by Max. Depth attribute. You will see in the result that the Decision Tree achieves an accuracy of 89%.

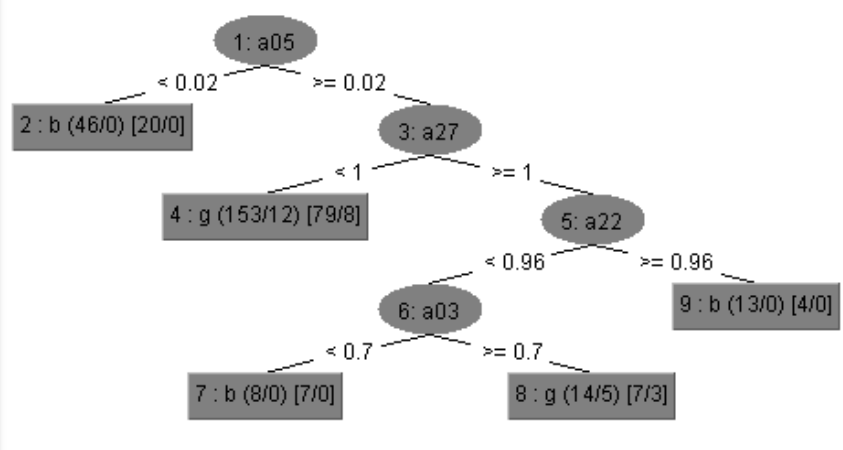

Figure 1. Tree Visualization

2.1.5. K-Nearest Neighbors

KNN support both Classification and Regression Problems. It works in a way to sort all the data and then locate the k nearest pattern during prediction. This paper presents a KNN text categorization method based on the shared nearest neighbor, effectively combining the BM25 similarity calculation method and the Neighborhood Information of samples. The effectiveness of this method has been fully verified in the NTCIR-8 Patent Classification evaluation [19]. The size of the neighbor is control by K. Where K=1, then predictions are made using single most similar training instances to the given new pattern for which prediction is requested. A common value for k is 3,7,11 and 21 for larger datasets. It is the laziest algorithm out there in machine learning. It works in a simple way by taking into account the distance from known data points.

2.1.6 Support Vector Machine

Support vector machines (SVMs), with their roots in Statistical Learning Theory (SLT) and optimization methods, have become powerful tools for problem-solving in machine learning. SVMs reduce most machine learning problems to optimization problems and optimization lies at the heart of SVMs. Lots of SVM algorithms involve solving not only convex problems, such as linear programming, quadratic programming, second-order cone programming, semi-definite programming, but also non-convex and more general optimization problems, such as integer programming, semi-infinite programming, bi-level programming and so on [20]. It accepts numerical attributes but can convert nominal to numerical automatically. SVM works by finding a line that best separates the data into two groups. In almost all problems of interest, a line cannot be drawn to neatly separately the classes therefor a margin is added. You will see in the result that the Support vector Machine achieves an accuracy of 88%.

In [10] the use of neural networks in classification is not uncommon in the machine learning community. In [11], [12] neural networks give a lower classification error rate than the decision trees but require longer learning time. In [13] Traditional classification techniques such as neural networks, logistic regression, and decision trees have been used in order to find the suitability of support vector machines, gradient boosting and random forests for loan default prediction. In [14] Different attempts are taken to improve Naïve Bayes for classification. In [15] SVM carries out nonlinear classification efficiently.

4. Result and Discussion:

For comparison of various classification algorithms, five datasets have been selected and taken from two different repositories that are KEEL [6] and UCI [7].

Table 1. Datasets Information

Dataset Name	No. of Instances	No of Attributes	No. of Classes	Test Method
Ionosphere	**351**	**35**	**34**	**10-CV**
Diabetes	**768**	**9**	**8**	**10-CV**

4.1. Using Weka:

Table 2. Resultant Table after comparison of Six most frequently ML algorithms on Ionosphere Dataset

Algorithm Name	**Accuracy**	**Mean Absolute Error**	**Precision**	**Recall**	**F-Measure**	**ROC**	**Time Taken**
Logistic Regression	88.8889%	0.1283	0.889	0.889	0.887	0.870	0.11
Decision Tree	**89.4587%**	0.158	**0.894**	0.895	0.894	0.891	0.03
K-Nearest Neighbor	86.3248%	0.139	0.871	0.863	0.857	0.825	**0.1**
Naïve Bayes	82.6211%	0.1736	0.842	0.826	0.829	0.935	0.02
Support Vector Machine	88.604 %	0.114	0.891	0.886	0.883	0.853	0.06
Linear Discriminant Analysis	88.3191%	0.1496	0.893	0.883	0.878	0.916	0.1

Table 3. Resultant Table after comparison of Six most frequently ML algorithms on Diabetes Dataset

Algorithm Name	Accuracy	Mean Absolute Error	Precision	Recall	F-Measure	ROC	Time Taken
Logistic Regression	77.21%	0.3094	0.767	0.772	0.765	0.832	0.1
Decision Tree	**77.47%**	0.3175	0.770	0.775	0.766	0.831	0.48
K-Nearest Neighbor	70.18%	0.2988	0.696	0.702	0.698	0.650	0.01
Naïve Bayes	76.32%	0.2841	0.759	0.763	0.760	0.819	0.01
Support Vector Machine	77.34%	0.2266	0.769	0.773	0.763	0.720	0.04
Linear Discriminant Analysis	71.224%	0.3448	0.706	0.712	0.708	0.773	0.08

4.2. Accuracy comparison:

Table 4. Resultant table after comparison of six most frequently ML algorithms Accuracy

Data set Size	Algorithms Name and Accuracy					
	Logistic Regression	Decision Tree	K-Nearest Neighbor	Naïve Bayes	Support Vector Machine	Linear Discriminant Analysis
351	88.89%	**89.46%**	86.32%	82.62%	88.60%	88.31%
768	77.21%	**77.47%**	70.18%	76.32%	77.34%	71.24%

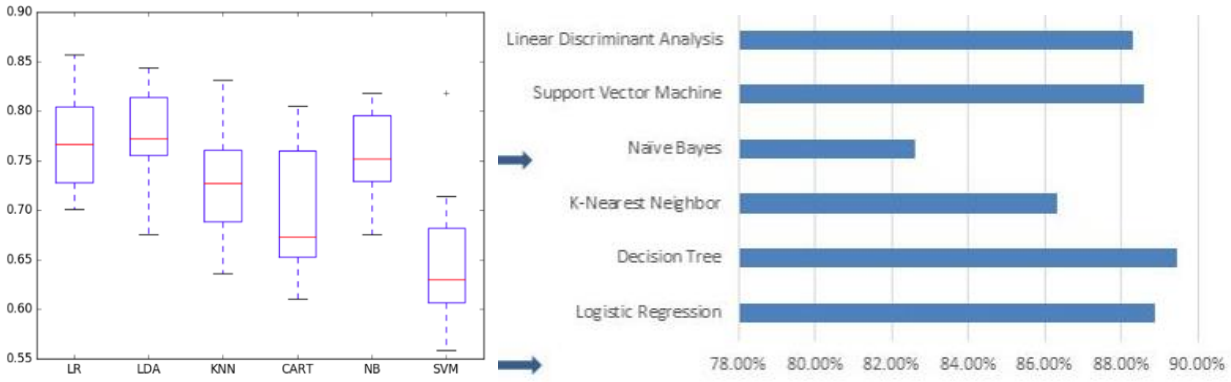

Figure 2. Comparison of Algorithms Plot diagram

From the above six frequently using algorithms results, the decision tree shows better results with respect to Accuracy, Precision, and Recall but lies in between the time taken. It was also concluded that both Linear and Logistic Regression discriminate analysis and would be worthy of further problems as well.

References

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001

[2] Khoshgoftaar, t. M., & Allen, e. B. (1999). Logistic regression modeling of software quality. International Journal of Reliability, Quality and Safety Engineering, 06(04), pp.303–317. Doi:10.1142/s0218539399000292

[3] Quinlan, J. R. (1986). Induction of decision trees. Machine Leaning, vol (1), pp.81-106 [4] Brijesh Kumar baradwaj and Saurabh pal (2011) "Mining educational data to analyze students' performance", (IJACSA) International Journal of Advanced computer science and applications.Vol. 2 no.6.

[5] Lippmann, R. (1987). An introduction to computing with neural nets. IEEE ASSP Magazine, 4(2), pp.4–22. doi:10.1109/massp.1987.1165576

[6] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L.Sánchez, and F. Herrera (2011), "KEEL Data - Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework.," J. Mult.-Valued Log. Soft Comput., vol. 17.

[7] A. Asuncion and D. Newman (2007), UCI machine learning repository Irvine.

[8] E. Frank, M. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, and L. Trigg (2005), "Weka," in Data Mining and Knowledge Discovery Handbook, Springer, pp. 1305–1314.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009), "The WEKA data mining software: an update," ACM SIGKDD Explor. Newsl., vol. 11, no. 1, pp.

[10] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood Series in Artificial Intelligence, 1994.

[11] J.R. Quinlan, "Comparing Connectionist and Symbolic Learning Methods," S.J. Hanson, G.A. Drastall, and R.L. Rivest, eds., Computational Learning Theory and Natural Learning Systems, vol. 1, pp. 445-456. A Bradford Book, MIT Press, 1994.

[12] J.W. Shavlik, R.J. Mooney, and G.G. Towell, "Symbolic and Neural Learning Algorithms: An Experimental Comparison," Machine Learning, vol. 6, no. 2, pp. 111143, 1991.

[13] A.Beque, K. Coussement, R. Gayler, S. Lessmann, "Approaches for credit scorecard calibration: An empirical analysis", Knowledge-Based systems, Vol. 134, pp. 213-227, 2017

[14] L. Jiang, D. Wang, Z. Cai, X. Yan, "Survey of improving naive bayes for classification", Lecture Notes in Computer Science, Vol. 4632, Springer, Berlin, Heidelberg, pp. 134-145, 2007

[15] C. Cortes, V. Vapnik, "Support-vector networks", Machine learning, Vol. 20, pp. 273-297, 1995

[16] The importance of logistic regression implementation in the Turkish livestock sector and logistic regression implementation/fields.

[17] Linear discriminant Analysis Alaa Tharwat, *Tarek Gaber, Abdelhameed Ibrahim and Aboul Ella Hassanien.*

[18] Short Survey on Naive Bayes Algorithm in International journal of advanced research in computer science 04(11). November 2017 with 1,685 Read

[19] KNN Research Paper Classification Method Based on Shared Nearest Neighbor Yun-lei Cai, Duo Ji ,Dong-feng Cai.

[20] Recent advances on support vector machines research article in technological and economic development of economy 18(1). March 2012.