Population Prediction Using Machine learning

[1]R. Mahesh, [2]R. Priyanga, [3]S. Gomathi

[1,2]I MSc DA, Department of Computer Science (PG), PSGR Krishnammal College for Women, Coimbatore, India.

[3] Assistant Professor, Department of Computer Science (PG), PSGR Krishnammal College for Women, Coimbatore, India.

Abstract. The main objective of the paper is to find the best machine learning algorithm to predict the population outcome in the future. This paper discusses about the three algorithms, which are naïve Bayes, IBk and Random Trees. Machine learning tool used for running these algorithms is WEKA. In this test, WEKA is used to analyze the data and the three algorithms are added from the library. The data set is obtained from UCI repository and from number of instances available in the dataset, only limited number of instances is used to run the algorithm to keep the controlled environment for this test. Of all the three algorithms Naïve bayes algorithm shows the highest result.

Keywords: Machine learning, Naïve Bayes, Weka, Lazy IBk, Random trees, UCI repository.

1 Introduction

In the last few years Machine Learning and its application has increased exponentially and its real-life use cases are becoming more prevalent. Increasing amount of data and the capability to store and process this data has made mandatory of smart data analysis for businesses to thrive and succeed in this data driven, information technology rich world. Machine learning is actually an application of Artificial Intelligence and it enables the machines to learn without programming them explicitly. There are four general machine learning methods and they are:

1) supervised,
2) unsupervised,
3) semi-supervised,
4) reinforcement learning

The objectives of machine learning are to enable machines to make predictions, perform clustering, extract association rules, or make decisions from a given dataset.

1.1 WEKA- A Machine Learning tool

WEKA stands for Waikato Environment for Knowledge Analysis. It's a new technology with wide range of applications. It deals with real-world problems arising from agricultural and horticultural domains. The main emphasis of this tool is on providing a working environment for the domain specialist rather than the machine learning expert. It includes the necessity of providing a wealth of interactive tools for data manipulation, result visualization, database linkage and cross-validation and comparison of the rule sets, to complement the basic machine learning tools.

1.2 The WEKA Workbench

It currently runs on Sun Workstations under X-windows' It gives access to machine learning tools written in a variety of programming languages (C, C++ and LISP). It is not a single program, but rather a set of tools bound together by a common user interface. WEKA project is redressing the balance by applying standard machine learning techniques to a variety of agricultural and horticultural problems. The goal is to discover and characterize what is required for successful applications of machine learning.

2. Dataset Description

The Adult data set obtained from UCI Repository. Out of 3000 and above instances the paper present 100 instances.

Table 1. The Adult data set

Attributes	Data types	Description
Age	Numerical	Age of person.
Work class	Nominal	Private, State Government, Federal Government
Education	Nominal	Bachelor, 11th,12th.
Marital status	Nominal	Married, Unmarried, Divorced.
Occupation	Nominal	Farming, Transport, Sales.

3. Naïve Bayes Algorithm

The naïve Bayes classifier greatly simplify learning by assuming that features are independent given classes. The main aim of the algorithm is to understand the data characteristics that affect the performance of naïve Bayes. This algorithm reaches for the best performances in two opposite cases: completely independent features and functionally dependent features. Accuracy of naïve Bayes is not directly correlated with the degree of feature measured as the class conditional mutual information between the features. Based on analysis of the training data numeric estimator precision values are chosen. For this reason, the classifier is not an updateable classifier (which in typical usage are initialized with zero training instances)-if you need the updateable Classifier functionally, use the naive Bayes Updateable classifier. The naive Bayes Updateable classifier will use a default precision of 0.1 for numeric attributes when build classifier is called with zero training instances [3].

3.1 Formula

$$P(C/X) = P(X/C)\, P(C)/P(X)$$

Probability of outcome/evidence =
Probability of likelihood of evidence* Prior/ (Probability of evidence).

3.2 Capabilities

Class- Binary Class, Missing Class values, Nominal Class.
Attributes- Binary Attributes, Empty Nominal Attributes, Missing Values, Nominal Attributes, Numeric Attributes, Unary Attributes.
Interface- Weighted Attributes Handler, Weighted Instance Handler.

Table 2. Statistical analysis Result of Naïve Bayes

Statistical parameters	Result
Correctly classified instances	77%
Incorrectly classified instances	23%
Total number of instances	100
Mean absolute error	0.242
Root mean squared error	0.422

4. Lazy-IBk Algorithm

Consider k nearest neighbor's classifier. Select appropriate value of k based on cross-validation. Can also do distance weighting [4].

4.1 Capabilities

Class- Binary class, Date class, Missing class values, Nominal class, Numeric class.
Attributes- Binary Attributes, Date Attributes, Empty Nominal Attributes, Missing Values, Nominal Attributes, Numeric Attributes, Unary Attributes.
Interface- Updatable classifier, Weighted Instance Classifier.

Table 3. Statistical Analysis Result of IBk algorithm

Statistical Parameters	Result
Correctly classified instances	74%
Incorrectly classified instances	26%
Total number of instances	100
Mean absolute error	0.2652
Root mean squared error	0.5044

5. Random Tree

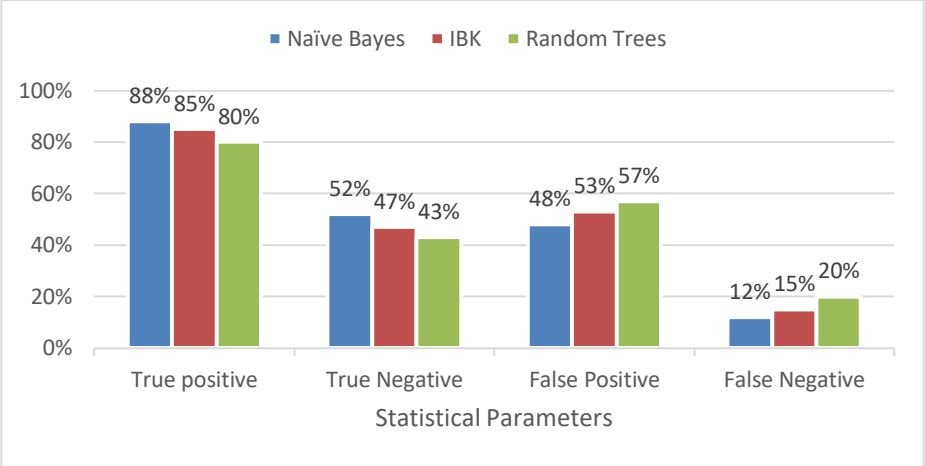
Considers K randomly chosen attributes at each node. It has an option to allow estimation of class probabilities (or target mean in the regression cases) based on a hold-out set [5].

5.1 Capabilities
Class- Binary Class, Missing class values, Nominal Class, Numeric Class.
Attributes- Binary attributes, Data attributes, Empty Nominal Attributes, Missing values, Nominal attributes, Numeric Attributes, Unary Attributes.
Interface- Drawable, Partition Generator, Randomizable, Weight Instance Handler.

Table 4. Statistical Analysis Result of Random Tree

Statistical Parameters	Result
Correctly classified instances	68%
Incorrectly classified instances	32%
Total number of instances	100
Mean absolute error	0.335
Root mean squared error	0.5408

6 Results & Discussion



In true positive, Naïve bayes showed greater result. In true negative, Naïve bayes showed greater result. In false positive, Random trees showed greater result. In false negative, Random trees showed greater result. Overall, Naïve bayes showed greater result.

Conclusion

Machine learning techniques are being widely used to solve real-world problems by storing, manipulating, extracting and retrieving data from large sources. Supervised machine learning techniques have been widely adopted however these techniques prove to be very expensive when the systems are implemented over wide range of data. This is due to the fact that significant amount of effort and cost is involved because of obtaining large labelled datasets. Thus active learning provides a way to reduce the labelling costs by labelling only the most useful instances for learning.

References

1. Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.s
2. Alexander J. Smola and Vishy Vishwanathan and Eleazar Eskin. Laplace Propagation. NIPS. 2003.
3. Rish, Irina. "An empirical study of the naive Bayes classifier." In IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, pp. 41-46. 2001.
4. Kasat, Neha R., and Sudeep D. Thepade. "Novel content based image classification method using lbg vector quantization method with bayes and lazy family data mining classifiers." Procedia Computer Science 79 (2016): 483-489.
5. Kalmegh, Sushilkumar. "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news." International Journal of Innovative Science, Engineering & Technology 2, no. 2 (2015): 438-446.