Application of Data Mining and Machine Learning in Sports

Sankara Subramanian Venkatraman
School of Computing
National College of Ireland
Dublin, Ireland
x18179541@student.ncirl.ie

***Abstract*—This project deals with 3 different sports datasets and 5 unique data mining and machine learning techniques.**

Logistic regression and k-nearest neighbor (k-NN) models will be applied for 1st dataset. c5.0 decision trees, and regression and model trees will be applied for the 2nd dataset. The 3rd dataset will use Naive Bayes text mining model. R programming language will be extensively used for data preparation, processing, cleaning (ETL) and evaluating model performance.

***Index Terms*—logistic regression; k-NN; c5.0 decision trees; model and regression trees; naive bayes; and R.**

I. Motivation

Data mining serves the world with useful insights from ubiquitous knowledge. Knowledge discovery in databases (KDD) and Cross-Industry Standard Process for data mining (CRISP-DM) are the methodologies used for incorporating data mining applications in real-time. It is used for pattern recognition, identify opportunities, process improvement and increase the business revenue.

The first dataset has information about skaters that played in each game. The challenge of this dataset is that only 62% of the time the best model predicts the winner. So, I would like to re-engineer this dataset with certain changes. I will train the model using the (k-NN) algorithm and alter the attributes to obtain a high-performance model. I would also like to apply a logistic regression model on the dataset to find what are the predictor variables that predict the target variable (decision) which is dichotomous. This model also helps us to visualize the relationship among different attributes using a scatter plot with a correlation matrix.

The second dataset has information about Hong Kong Horse Racing Results 2014-2016 seasons for 1561 local races. The interesting part of this dataset is, a gambling model will be developed for upcoming races. In this dataset, decision tree and regression and model tree algorithms will be applied. First, the decision tree to predict the finishing position of horses based on factors like a trainer, jockey, actual weight, running positions and win odds, etc. Regression and model trees algorithm will also be applied to predict the final position. Finally, compare which model performs well for the above-mentioned attributes.

The third dataset contains information about tweets captured during Cleveland Cavaliers Vs Golden State Warriors, 3rd game of the 2018 NBA finals. The tweets are in 39 different languages. But, this project scope is limited to the English language. Naive Bayes a text-mining classifier is applied to this dataset. The interesting part of this model is, it works well for noisy and missing data. The algorithm in this dataset will help us to classify whether the tweet is from an android user or iPhone users based on the tweets captured.

II. Research Question

1. Compare the result (Win or Lose) of the NHL game using the K-nearest neighbor algorithm and logistic regression using predictor variables.
2. Analogy of C5.0 decision trees and regression and model trees models for predicting the position of the horses in Hong Kong Horse Racing.
3. Filtering, Visualizing, and Evaluation of tweets from an android or iPhone of NBA match.

III. Initial Literature Review

Identifying suitable sports for beginners by calculating anthropometric measurements. Achievements in various sports can be predicted using this measurement. Using, normalization a scaling technique that discovers a new set of range from a current range called Min-Max Normalization. They have also used Euclidean distance for calculating similarities [3].

Decision model trees are simple and efficient in determining the result of a football match. It is a common two classification machine learning technique. The prediction of the winning team based on home-away conditions and the previous match played against each team of a football match [4].

This idea of using the C4.5 algorithm was extracted from the paper which predicts the success of the licensure examination for teachers (LET). They found the gain ratio for success and failure teachers, which is helpful to improve the quality of the education system and success rate in LET. Similarly, we could apply it in the C5.0 decision tree model [5].

Edu720 platform that predicts the success rate in the education system uses the k-NN and decision tree algorithm. The J48 decision tree is better than naive Bayes for classification problems. They have compared the values of k-NN and decision tree to attain optimum model. They have also used z-score normalization for re-scaling the attributes [6].

The objective of the multiple regression model is to predict the comfort of sports fabric using attributes such as thermal insulation, air permeability, moisture regain and evaporation rate, etc. The experiment was carried out in 4 stages and the essential predictors are decided to design sportswear [7].

In general, naive Bayes has 2 text-based classification models. Bernoulli model which uses conditional probability. The second one is a widely used word count vector which uses a multinomial model [8].

Bayesian using multinomial regression represents a variety of documents which has a wide-range collection of vectors containing attributes and feature value (0/1). The new class sample is selected from the class which has a maximum conditional probability [9].

NFL game's fantasy point of each player is determined using multiple regression. For a single week (w), they have estimated the points (Y) recorded for each player corresponding to the predictor variables (X). The correlation of dependent variables are calculated using Pearson correlation (r), a measure of relatedness between the variables. [10].

Predicting movie box office success uses multiple regression. They have calculated t-static, p-value and coefficients of predictor variables. They have also used evaluation methods such as coefficient of determination (R^2), adjusted (R^2), kappa statistic (κ), mean absolute error [13].

IV. Data Sources

A. NHL dataset

The first dataset [1] contains National Hockey League data which deals with game goalie status data source which has 19 attributes in it. It contains 24,647 records.

B. Hong Kong horse racing results dataset

The second dataset [2] has information about Hong Kong horse racing results. This dataset is obtained by joining two datasets, which has horse and race results. In race dataset, race_distance attribute is pulled and joined with horse dataset which contains 20 attributes together. It contains 30,190 records.

C. Tweets during Cavaliers vs Warriors dataset

The third dataset [3] has tweets information which was gathered during the Cavaliers vs Warriors NBA match. This dataset contains 43 attributes. It contains 51,426 records.

V. Identification of Machine Learning Methods

k-nearest neighbor:

The k-nearest neighbor is simple and effective. There are no underlying data distribution assumptions. In the first dataset, we have an attribute **decision**, which classifies whether the player won or lost the match. k-NN works well with numeric data. By choosing appropriate k value and other numeric attributes will try to predict the result of the game. Also, this method uses both the normalization technique min-max and z-score normalization. 80% of the records for training the model. 20% of the records are utilized for testing the model built.

Logistic Regression:

Regression technique that models the size and strength of numeric relationships. In the regression model, correlations between predictor variables are calculated using the Pearson product-moment correlation coefficient before applying those variables into the model. Assumption of "regression analysis includes:

1) The model is linear.
2) Error terms have constant variances (Homoscedasticity).
3) Error terms are independent
4) Error terms are normally distributed" [11, pp. 502].

In this dataset, **(decision)** is used as the dependent variable and all the numerical and categorical attributes are treated as predictor variables. Then test which variables and order are highly significant. Compare both models and choose the best model which fits the dataset.

C5.0 decision tree:

The decision tree utilizes a tree structure to model the relationships among attributes and the possible outcomes. It uses a heuristic approach called recursive partitioning, which uses divide and conquer. It can handle both numeric and nominal attributes and more efficient than other complicated models [2, pp. 121-123]. The second dataset which is about horse riding results makes use of the C5.0 decision tree to predict the position of the horses. 90% of the records are used for training the dataset and 10% is utilized to test the model built.

Regression and Model trees:

Regression and model trees will be used for numeric predictions by altering the tree-growing algorithm. Regression tree does not use linear regression methods instead it makes predictions on the average value. Homogeneity is measured by standard deviation, variance or absolute deviation from the mean [2, pp. 189-190]. In the second dataset, the finishing position of horses is classified using Classification and Regression Tree is widely known as (CART) algorithm (rpart) or (M5-prime) package in R. 75% of the dataset is used for training and 25% utilized for testing. Finally, compare the result of the classification and regression model to select the best model.

Naive Bayes:

Bayes' Theorem, "the probability that an event has happened given a set of evidence for it is equal to the probability of the evidence being caused by the event multiplied by the probability of the event itself" [12, pp. 252]. The third dataset will use the Naive Bayes algorithm, which is fast and effective for text analytics. The Document Term Matrix (DTM) is used for the data structure. Text mining processes such as stemming, lemmatization, tokenization, and punctuation mark removal

[1] https://www.kaggle.com/martinellis/nhl-game-data
[2] https://www.kaggle.com/edwardckw/notebookd02724dac3/data
[3] https://www.kaggle.com/xvivancos/tweets-during-cavaliers-vs-warriors

will be used for data preparation. This model will predict whether the tweet received is from an android or iPhone.

VI. Identification of Evaluation Methods

i) In the k-NN algorithm, CrossTable() Validation and predictive accuracy will be used to validate the results. Cross-Table, a proportion of values will fall into one of the 4 categories. True Negative, False Positive, True Positive and False Negative. False positive and False Negative predicts the test data disagrees with a true label. It has 3 main classifiers actual class values, predicted class values, and the estimated probability of the prediction. The actual value comes directly from the target. Predicted values will be obtained from the model we build and test. predict() function will predict the model's estimated probability.

ii) The regression model is different from other machine learning models, it will allow the users to alter the predictor variable to get the desired output. The values of t and pr($> |t|$) will help the user to evaluate the model. If the p-value is greater than 0.05, then the predictor variable is not a suitable attribute for predicting the target variable. If the p-value is less than 0.05 and the significance code is (*), then it is 95% statistically significant that the predictor variable will be the effective estimator of the target variable. The coefficient of determination (R^2) provides the overall model performance for the corresponding dependent variable.

iii) In Regression trees, while comparing summary statistics of prediction to the actual summary, if it falls into a narrow range, then we predict the model is not accurate for identifying extreme cases. Correlation cor() is also one of the best function to evaluate the relationship between actual and predicted values. Mean absolute error (MAE) will also be used to evaluate model performance. Accuracy is defined as a proportion of the number of true positive and true negative divided by the total number of predictions. (1-accuracy) is used to calculate the error rate [2, pp.300-301]. The caret package has a function to create a confusion matrix.

iv) C5.0 decision tree uses crossTable() function to compare actual and predicted values. Along with this, we can calculate the kappa statistic or Cohen's Kappa coefficient. (κ) can be calculated by
pr(a) - the proportion of actual agreement
pr(e) - the proportion of expected agreement between the classifier and the true value. Similarly, sensitivity, specificity, precision, and recall can also be calculated. Sensitivity and specificity are used to maintain a balance between predictions. This helps us to identify whether the model is over-fitting or under-fitting.

v) Naive Bayes, a text-mining algorithm will use F-score and Receiver Operating Characteristic (ROC) curve for evaluation. Like other models, CrossTable Validation will also be used to find the probabilities of True and False Positive and Negative. Using harmonic mean, F-measure can be calculated by combining precision and recall. It will be used to compare several models at a time. Visualization helps us to understand how well the model performs than numeric values. ROC evaluate the trade-off between sensitivity Vs (1-specificity). The perfect ROC value will be classified as the area under the ROC curve (AUC). The range of AUC is 0 to 1.

References

[1] Ian H. Witten, Eibe Frank, Mark A.Hall Christopher J.Pal/Data mining: practical machine learning tools and techniques – 4th ed.

[2] Brett Lantz - 2nd ed/Machine Learning with R.

[3] P. T. Amarasena, B. T. G. S. Kumara, S Jointion, "Data Mining Approach for Identifying Suitable Sport for Beginners", *2019 International Research Conference*, pp.57-62, Mar, 2019.

[4] Tang Xiaohu, Liu Zhifeng, Li Taizhao, Wu Wenbin, Wei Zhenhua, "The Application of Decision Tree in the Prediction ofWinning Team", *2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS)*, pp.239-242, Aug, 2018.

[5] Clarin, J.A., Sta Romana, C.L.C., Feliscuzo, L.S., "Academic analytics: Applying C4.5 decision tree algorithm in predicting success in the licensure examination of graduates", *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS*, pp.193-197, Feb, 2019.

[6] Dervisevic, Omar,Zunic, Emir,Eonko,Dzenana,Buza, Emir, "Application of KNN and Decision Tree Classification Algorithms in the Prediction of Education Success from the Edu720 Platform", *2019 4th International Conference on Smart and Sustainable Technologies (SpliTech)*, pp.1-5, Jun, 2019.

[7] Min Li,Dong-Ping Li,Wei-Yuan Zhang,Xiao-Zhong Tang, "A Multiple Regression Model for Predicting Comfort Sensation of Knitted Fabric in Sports Condition Based on Objective Properties",*2009 Second International Conference on Information and Computing Science*, pp.372-375, May, 2009.

[8] Nghia Nguyen, et al."Hierarchical Scheme for Assigning Components in Multinomial Naive Bayes Text Classifier",*2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*, pp.335-340, Dec, 2018.

[9] Madigan, David, et al. "Bayesian multinomial logistic regression for author identification." *Bayesian Inference and Maximum Entropy Methods in Science and Engineering* 803 pp.509-516, Nov, 2005.

[10] Landers, J.R and Duperrouzel, B,"Machine Learning Approaches to Competing in Fantasy Leagues for the NFL", *IEEE Transactions on Games IEEE Trans*, pp.159-172, Jun, 2019.

[11] Carlos Cortinhas, Ken Black/ Statistics for Business and Economics -1st European Edition.

[12] Aoife D'Arcy, Brian Mac Namee, and John D. Kelleher/ Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies.

[13] Subramaniyaswamy, V, et al."Predicting movie box office success using multiple regression and SVM", *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pp.182-186, Dec, 2017.