Student Id- 18186432
Name- Raj R. Kupekar

Data Mining and Machine Learning

Project Proposal

Motivation

In chess game, it is challenging to anticipate the moves of the opponent and accordingly figure out the chess board position for winning the game. In times, a rook can be better than bishop and sometimes a pawn can take the game. Thus, evaluating the definite chess position is now been the area of interest for data scientist who can solve the complex problem by making the use of predictive modelling. The first dataset is the Chess End Game Dataset. This dataset is associated with different end game moves between the king & rook versus the king & pawn where the pawn is on the A7 position, i.e. it is 1 square away from queening. There are 3196 records with 36 features, where each feature represents a position on the chess board. The dependent attribute is classification having two outcomes ; white can win or white can't win. In this dataset the class distribution for win is about 52% whereas for no win is 48%. Thus, eliminating the biasness of the data.

Letter identification is basically the ability to recognize the attributes, shapes as well as patterns of the 26 alphabets. It can be used to identity various alphabets irrespective of the way in which the same letter is been written in different manners. This dataset consists of large number of black and white rectangular image pixels entries each representing one of the 26 English Alphabets. The image pixels records in this dataset are scaled down to integer values ranging from 0-15, corresponding to the character images from 20 different fonts. There are in total 20000 records with 16 independent integer variables and 1 dependent variable. Like the chess dataset the class distribution for the independent variable is uniformly distributed.

Finally, the third dataset is the Tesco Marketing Content Dataset. The dataset consists of a behavior for each customer towards the marketing content on Tesco website in order to provide them the appropriate content card based on their spending and previous history of viewing the content card. In this dataset, there are 9 different marketing content history with following entries in it, 1 means the customer clicked on the content card, 0 means the customer haven't clicked on the content card and NA means the card was never shown to the customer. Affluency is the predictor where the customer is categorized into five categories (very low, low, mid, high, very high) such that the marketing team can come up with best suited content for each customer. The dataset consists of around 10000 rows and 27 columns.

Research Questions

Dataset 1 : Here, depending upon the different possible position for king + rook (white) and king + pawn (black), we are going to predict whether the white is going to win the game or not, considering white's turn to play.

Dataset 2 : The use case of this dataset is to predict the English alphabets depending upon the integer values corresponding to the equivalent image representation of the character.

Dataset 3 : In this dataset, our aim is to categorize the customers into different categories based on their spending and content card history.

Literature Review

In (Yusuff1, et al., 2012) Logistic Regression is used to predict the breast cancer from mammograms. Data was gathered by the radiologist with the help of survey during their observation of the patients. The aim was to retrieve the patient with breast cancer with the help of given potential factors. It was achieved using cross tabulation with accuracy result up to 91.5%. The factors that highly contributed are presence of mass, architectural distortion, skin thickening and calcification.

In (Wang, et al., 2010) paper, various KNN algorithms are studied and new KNN approach is been evaluated based on the evidence theory. They overcome the limitations of Evidence based theory KNN by introducing Density Based EKNN. In this paper the problem associated with Imbalance data is solved using DEKNN algorithm more effectively as compared to resampling method. Thus, reducing expense for generating new data points. This is achieved using Global and Local frequency estimation of prior probability across the training data space and neighborhood data space respectively.

This paper (Durgesh & Bhambhu, 2009), introduces a novel Support Vector Machine learning methodology applied on various datasets having two or multi-level classes. In this experiment, SVM algorithm is applied on datasets to get comparative results using different kernel functions (linear, polynomial, sigmoid and rbf). The tuning of cost and kernel parameters are done in order to obtain an efficient model. It makes use of grid search with 5-fold cross validation to select the best parameters.

Paper (Himani & Sunil, 2016) focuses on different Decision Tree concepts. This paper examines the characteristics, challenges and advantages and disadvantages compared with various decision tree algorithms model like ID3, C4.5 & CART. The efficiency of algorithms is determined based on the execution time and their accuracy. This conference paper helps us to conclude that each decision tree algorithms have their own importance based on the different problem cases.

In (Kulkarni & Sinha, 2014), depicts classification study using Random Forest algorithm. This research work undertakes five different methods to improve the accuracy and performance of the ensemble model algorithm. In this paper three techniques disjointed portioning approach, weighted hybrid decision tree and optimal subset of random forest are used to evaluate the efficient model for classification. In interest to improve the accuracy hybrid decision tree with weighted voting is used whereas an attempt to boost the time for learning and classification is achieved by reducing the number of base decision tree.

Data Sources

Kaggle is a storehouse, which host thousands of datasets such that data mining and machine learning techniques can be implemented. It is also a good platform to get involve in competitions related to predictive and analytic modelling. Tesco Marketing Content dataset was in one of the Kaggle competition. Another immense data resource is the UCI machine learning repository. It is having a variety of datasets which are readily available for Machine learning application with a brief description for each dataset. Chess End Game and Letter Recognition are the collection of UCI repository.

Machine Learning Methods

Logistic Regression is a Machine Learning technique which makes use of logistic function to perform prediction. In regression, Linear Regression is used when the response variable is quantitative. Thus, in order to work with categorical response variable, Logistic Regression is introduced. It makes use of Sigmoid function which is S-curved in shape and can take any value in between 0 and 1. It takes probability into account and sets a decision boundary to decide the class of the data point. It works on the principle of Maximum Likelihood which estimates the values for coefficient that minimizes the error in predicting the probabilities of the model.

K Nearest Neighbor is a Supervised Machine Learning method which is used for categorization based on the shared nearest neighbor. KNN can be used for both Regression and Classification predictive modelling. It is a Non-Parametric and Lazy algorithm which makes no assumption on the underlying data and does not imply any rationalization on the training data points i.e. it has no training phase. It works with K factor to classify a data point where; K is the number of nearest neighbors. Depending upon the value of K, KNN makes decision boundaries between different classes thus; categorizing the given data point.

Another Classification technique is a flowchart-based tree structure called Decision Tree, where a class label is represented by a leaf node. A path is tracked down from root node which have all the attributes, to the leaf node having only the class for the given data point. Attributes selection is done using Information Gain/ Entropy and Gini Index. Lowest error rate is obtained by performing a cost complexity pruning as a function of a tuning parameter α. Further, a K-Fold Cross Validation is used for choosing the α to minimize the average error of the model.

Decision Trees are the elementary unit of Random Forests to design an efficient predictive model. Random Forest is an ensemble modelling technique which builds several decision trees by using Bootstrap sample set with replacement on the given dataset and provides low variance with decorrelated trees output. It also overcomes the Out of Bag Error. In case of Regression, the output from each Decision Tree is averaged and delivered as the final outcome whereas; for classification the class having maximum number of outputs is given as the final class.

Support Vector Machine is yet another powerful Supervised learning algorithm which can perform both Regression as well as Classification. It works on the principal of maximum marginal classifier. It generates a linear as well as non-linear hyperplane in an N dimensional space which can directly classify the data points. SVM is also robust to outlier and can fits a generates a hyperplane having highest margin using support vectors i.e. nearest data points.

Evaluation of Methods

For Chess End Game dataset, the best suited algorithm would be the Logistics Regression. As because, this dataset is having a binomial dependent variable thus; Logistics Regression would be the best fits for the model with two classes. Initially, after cleaning and outlier's treatment ; Logistics Regression model can be implemented by using the glm() function with family of the class equals to binomial in R. The tuning of the model can be done by adjusting the threshold value depending upon the odds ratio. The equation to evaluate the model is given as,

$$p(y) = \frac{e^{\beta_0+\beta_1 x_1+\cdots\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\cdots\beta_n x_n}}$$

where, β_0, β_1, β_n are the coefficients of the variables.

The working principle of KNN algorithm can be applied on the Letter Recognition Dataset to gain a powerful predictive model. The model is carried out by installing 'class' package which comprises of knn() function. After choosing the appropriate K value, Euclidean Distance is used to predict the class of the given data point. Using Confusion Matrix, the model statistics & accuracy is printed and accordingly the model can be tuned simply by altering the K value i.e. the number of nearest neighbors.

Similarly, Support Vector Machine Classification can be implemented on the same dataset which draws hyperplanes in the between the different classes. The required library is the 'e1071' library for svm()

function. By amending the gamma and cost values, the tune.svm() function can be used to tune the model using K-fold Cross Validation method.

Decision Trees performs better when the response variable is having more than two classes. Thus, for Tesco Marketing Content dataset with multiclass response variable Decision Tree Classification algorithm can be implemented. The Decision Tree Classifier can be implemented by installing the 'knit' package and using the tree library. The Cross Validation can be done using cv.tree() function with function equal to prune.misclass. Similarly, pruning can be done using prune.misclass and setting the appropriate best cost complexity pruning in order to select a sequence of trees.

Furthermore, Random Forest can outperform as compared to Decision Tree. The deficiency in Decision Tree such as correlated output and high variance can be overcome by Random Forest. By calling the random forest () library; Random Forest model can be implemented. The tuning of the model can be done by altering the default values for number of trees to grow (i.e. ntree= 500) and the number of features randomly sampled at each split (i.e. mtry= 2). Comparing the error rate for each modification in the above two parameters; an efficient Random Forest model can be evaluated.

Bibliography

Durgesh, K. S. & Bhambhu, L., 2009. DATA CLASSIFICATION USING SUPPORT VECTOR MACHINE. *Journal of Theoretical and Applied Information Technology,* p. 7.

Himani , S. & Sunil, K., 2016. A Survey on Decision Tree Algorithms of Classification in Data Mining. *International Journal of Science and Research (IJSR)* , 5(4), p. 4.

Kulkarni, V. Y. & Sinha, P. K., 2014. Effective Learning and Classification using Random Forest Algorithm. *International Journal of Engineering and Innovative Technology (IJEIT)* , 3(11), p. 7.

Wang, L., Khan, L. & Thuraisingham, B., 2010. *An Effective Evidence Theory based K-nearest Neighbor (KNN) classification.* Texas, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

Yusuff1, H., Mohamad, N., Ngah, U. & Yahaya, A., 2012. BREAST CANCER ANALYSIS USING LOGISTIC REGRESSION. *IJRRAS,* 10(1), p. 9.