International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

A Research Travelogue on Classification Algorithms using R Programming

S.Nagaparameshwara Chary, B Rama

Abstract: ***Classification is a machine learning task which consists in predicting the set association of unclassified examples, whose label is not known, by the properties of examples in a representation learned earlier as of training examples, that label was known. Classification tasks contain a huge assortment of domains and real world purpose: disciplines such as medical diagnosis, bioinformatics, financial engineering and image recognition between others, where domain experts can use the model erudite to sustain their decisions. All the Classification Approaches proposed in this paper were evaluate in an appropriate experimental framework in R Programming Language and the major emphasis is on k-nearest neighbor method which supports vector machines and decision trees over large number of data sets with varied dimensionality and by comparing their performance against other state-of-the-art methods. In this process the experimental results obtained have been verified by statistical tests which support the better performance of the methods.***

In this paper we have survey various classification techniques of Data Mining and then compared them by using diverse datasets from "University of California: Irvine (UCI) Machine Learning Repository" for acquiring the accurate calculations on Iris Data set.

Keywords: Classification, Data Mining, Decision Table, kNN, SVM

I. INTRODUCTION

The keyword "Data mining (DM)" is a universal term that merges various computer-based procedures for performing analysis over the possible data over huge datasets with chunks of data. Most of the researchers consider this to be an interdisciplinary research area that includes areas such as: "statistics, computer science, mathematics etc" where all often work together with experts from various application areas. Almost all the DM techniques are typically deployed for identifying novel and useful patterns without with the data may always refrain to be unknown. Mostly cited definition by researchers is given by Decker and Focardi [3]: "Data mining is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data". In contrast to the traditional statistical methods whose task is to verify and approve or disprove a benchmarked hypothesis as the data mining provides us with the opportunity to dynamically generate new hypotheses.

Manuscript published on November 30, 2019.
* Correspondence Author
S.Nagaparameshwara chary*, Research Scholar in Department of Computer Science, Kakatiya University, Warangal,
B Rama, Assistant Professor in Department of Computer Science, Kakatiya University, Warangal,

The major task of data mining is to efficiently and selectively extracts information and knowledge from distinct data sets that should make sense of the data as the knowledge must tend to exhibit various essential attributes that must be understandable or valid or novel and utilizable[4]. By definition: "data mining is a part of a super ordinate process called knowledge discovery in databases (KDD)" where the idiom "database" represents distinct data storage by not only solely comprising data that is stored in databases. "KDD" is defined by Fayyad et al. [6] as: "Knowledge discovery in databases describes the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data".

The classification of machine learning algorithms was calabelorized into two distinct groups: supervised or predictive and unsupervised or descriptive learning methods. Where the supervised learning algorithms use a set of labeled training examples where each feature vector is survey by a unique class label. In utilization of training data most of the algorithms produce a classifier model that maps every object with a unique class label. In contradiction to this various unsupervised learning algorithms aim to identify similarities among non labeled objects that tend to work without class information over a sample space (IV (Iris)). In this paper we mainly focus on the applications that implement supervised learning algorithms used to classify and compare then evaluate the classification algorithm and the performance is assessed by using various data subsets that are generated.

Major classification models related to our problem statement are illustrated briefly in "Related work" that is in Part II of this paper are explained and discussed thoroughly. And on the same hand we have compared all the algorithms to identify classification accuracy measure that is we compared k-nearest neighbor classifiers with the generated decision trees then with support vector machines generated and several ensemble methods that exists.

II. RELATED WORK

KNN is the most researched topic in the present day era by most of the researchers where the major aspect is to perform a detailed study over survey applications that are performed by implementing introductory data mining books and the reports of surveys that are performed as specified in the survey article documented by by Jiang et al. [7] which proposed many improvements of KNN algorithms for implementing classification of data. Another interesting publication is performed on the weighted kNN classification algorithm that is based on various symbolic features that are published by Cost and Salzberg [9] in which the distance are measured and calculated then depicted in the form of tables to produce real-valued distances from symbolic domains that also represents features.

The authors propose that the proposed algorithm is superior when compared with the existing algorithm like KNN as it is implemented on three distinct application domains whose major advance label is the possession of training speed and simplicity in implementation.

Han et al. [1] publishes his study that focuses on a adjustment of weight while implementing KNN for identifying optimum weighted vector by using an optimization function the is based on the "leave-out-out cross-validation" technique and "greedy hill climbing technique" and on the same hand the work introduced three major "decision tree algorithms" [13] and many other studies provide comprehensive surveys being performed on the applications that are based on distinct decision tree algorithms in the fields of machine learning and data mining approaches. References [8-9] represents the review methods that are based on the tree simplification method which improves the comprehensibility of the decision trees. And in the very recent many publications focused on parallel and distributed decision tree algorithms, and in the similar approach a study by Yildiz and Dikmen [14] has analyzed the parallelization potential of C 4.5 algorithm and provided three major ways to implement parallelization to improve the algorithm based on the classification feature or on the data that is used or on the nodes that are generated in decision tree among several slave process. Experimental results depict node based parallelization implementation that demonstrates the best speedup process among the three approaches that exists. References [6] and [14] focus on distributed algorithms for implementing decision tree induction over huge peer-to-peer environments or data sets.

SVMs and their mathematical background are illustrated in C. Cesarano, A. d'Acierno, and A. Picariello[12] where the survey is implemented in the comprehensive book by Smola and Scholkopf [1] that provides illustration in support vector machines that comprises of self lerarning kernels and the othere publication that has cited major aspects of SVMs are applied over implementation of text categorization in [8]. Huang et al. [6] references focuses on kernel based algorithms over massive dataset and Ivanciuc [I1] illustrates applications of SVMs in the field of chemistry and mathematical background information is used to represent the SVMs and multi class SVMs which are already listed in the appropriate paragraphs where the literature and further references are represented to implement support vector regression tools that are denoted as illustrated by Smola and Schollkopf [SS04] as a tutorial.

III. CLASSIFICATION ALGORITHMS

These algorithms comprises of a outcome or target variables or dependent variables that are used to perform prediction over a given set of independent variables or predictors by using which the function that generates mapped inputs with the desired outputs where the training process continues till the model obtain desired level of accuracy over the training data set. Some of the examples of supervised learning are: "Regression, Decision Tree, Random Forest, KNN, Logistic Regression etc,."

Classification is the process of performing data processing that assigns various classes over a set of data that is aided towards accurate prediction over analyzed data set [2]. Conjointly referred to as typically refer to as a Decision Tree, classification is one amongst many strategies supposed to form the analysis of terribly giant Data sets effective to form a good set of classification rules that answers a query, makes call supported the query and predicts the performance. to start with a collection of training in order sets square measure created with sure set of attributes or outcomes. The most objective of the classification algorithmic program is to extract however that set of attributes reaches its conclusion.

K-Nearest Neighbor is considered to be a non-parameterized lazy learning algorithm that performs classification on the tested instances of data that is based on computed distances measures over instances of labeled training set. The distance that is attained represents the data set with the nearest neighbors that are required to perform voting towards the generated predicted class by verifying the known classes in the vicinity of a new test instances that are available. Yu et al. (2001) use SVMs as their primary classification algorithm of choice but they do also experiment with the use of a K-Nearest Neighbor algorithm. This algorithm simply records training data and then uses a distance measure from a test instance to known training instances to predict which class the test instance should belong to by examining its nearest neighbors. The KNN method can still produce desirable results. This study compares the two approaches by studying four emotions i.e. anger, happiness, sadness, and neutral from movie and teleplays. Given that the KNN rule relies on a distance metric to perform classification, Yu et al. explain how different distance metric can yield different results. Lee et al. (2001) also use the KNN method of classification to classify between negative and positive sentiment.

SVMs are another classification approach which functions by analyzing a feature space and attempting to construct a hyper plane to separate data points belonging to different classes. They operate by mapping data onto a higher dimensional space using a kernel function and defining the hyper plane there. Although SVMs are inherently binary classifiers they can be modified for multiclass problems by using pair wise classification, which tackles a problem as a series of binary problems.

Decision Trees are another approach which many researchers employ in tackling the problem of automatic affect recognition. This is a simplistic classifier which makes observations on data and maps these observations to decisions on class ownership. It functions by constantly querying a test instance to gain more information about which class it may belong through a combination of if-then rules. Hoque et al. (6) use a J48 Decision Tree to classify a number of utterances from movies under four emotion categories.

IV. EXPERIMENTS USING R PROGRAMMING

classification strategies are applied on "IRIS dataset to gauge the performance of every algorithmic rule on testing dataset". The coaching dataset consists of one hundred rows that is employed to perform coaching the training model as the testing dataset contains fifty plus rows that are employed to perform calculation of performance attained through the model on unseen dataset used[4].

PACKAGES FOR R: R uses collections of packages to perform completely different functions [4]. Cubic measure project Views offer varied packages to completely different users in keeping with their style. R package contain completely different functions for data processing approaches.

```
→data(IRIS)
→summary(IRIS)
 Sepal.Length   Sepal.Width    Petal.Length   Petal.Width        Species
Min  :4.350  Min  :2.050  Min  :1.050  Min  :0.150  setosa    :60
1st Qu :5.150  1st Qu :2.850  1st Qu :1.650  1st Qu :0.350  versicolor:60
Median :5.850  Median :3.050  Median :4.350  Median :1.350  virginica :60
Mean  :5.845  Mean  :3.059  Mean  :3.759  Mean  :1.219
3rd Qu.:6.405  3rd Qu.:3.305  3rd Qu :5.105  3rd Qu.:1.805
Max  :7.905  Max.  :4.405  Max  :6.905  Max  :2.505
```

Figure 1: Input parameters considered in R Application

A. KNN

```
library(class)
cl=train$Species
set.seed(1234)
preds.knn=knn(train[1:4],test[1:4],cl,k=3)
CrossTable(preds.knn,test$Species,chisq=F,prop.r=F,prop.c
=F,prop.t=F,prop.chisq=F)
```

Figure 2: R Application trace for KNN Implementation

Total Observations in Table: 50

Table 1 : Accuracy calculations for kNN using iris dataset

Preds.knn	Test$species Setosa	Versicolor	Virginica	Row Total
Setosa	15	0	0	15
Versicolor	0	15	2	17
Virginica	0	2	17	19
Column Total	15	17	19	51

kNN be able to exist used for along classification and regression drawback. kNN considers the foremost alike different objects clear in terms of their instances, examine their labels, and provides the unassigned item the bulk vote. once a replacement datum is introduced, its similarity (using geometrician distance during this case as all variables ar continuous) is measured from every of the assemblage and species of the take a look at datum is assigned in step with the closest (distance-wise) grouping. Hence, kNN is simply used for classification of testing information points wherever we are able to simply determine the clusters of coaching information points.

B. Support Vector Machine

- **model.svm=svm(Species~.,data-train)**
- **preds.svm=predict(model.svm,newdata=test)**
- **CrossTable(preds.svm,test$Species,chisq=F,pro p.r=F,prop.c=F,prop.t=F,prop.chisq=F)**

Figure 3: R Application trace for SVM Implementation

Total Observations in Table: 50

Table 2:Accuracy calculations for SVM using iris dataset

Preds.svm	Test$ species Setosa	Versicolor	Virginica	Row Total
Setosa	14	0	0	14
Versicolor	0	16	2	18
Virginica	0	3	15	18
Column Total	14	19	17	50

Support vector machines (SVMs) are helpful once there are terribly several input variables or once input variables move with the end result or with one another in sophisticated (nonlinear) ways in which. By perceptive the plots we can we will we are able to clearly see that some variables are non-linearly associated with one another. Hence, victimization SVM could be a smart choice on the Iris dataset. Since our information is linearly separable, SVM would be an honest alternative for classification purpose of Iris dataset.

C. Decision Tree

```
model.rpart=rpart(Species~.,data=train)
preds.rpart=predict(model.rpart,newdata=test,type="cl
ass")
CrossTable(test $
Species,preds.rpart,chisq=F,prop.r=F,prop.c=F,prop.t=F
,prop.chisq=F)
```

Figure 4: R Application trace for decision tree implementation

Total Observations in Table: 50

Table 3: Accuracy calculations for Decision Tree using iris dataset

Preds.decission tree	Test$species Setosa	Versicolor	Virginica	Row Total
Setosa	13	2	1	16
Versicolor	0	15	0	15
Virginica	2	1	16	19
Column Total	15	18	17	50

The decision trees are supervised classification algorithmic rule helpful once all the input variables tend to move with the output in

"if-then" styles of ways in which the data are additionally appropriate once inputs have an "AND" relationship to every dissimilar or redundant input variables correlate.

By marking the data plots from "Exploratory information Analysis", we are able to clearly identify a positive relationship or correlation among distinct variables of "IRIS dataset" for creating procedure calls of ideal trees where the classification of the species additionally implement the "if-then" relation between various variables of "IRIS dataset" denoted by d where the accuracy comparisons for the above three approaches are illustrated as:

```
models=data.frame(Technique=c("DecissionTree","kNN","SVM")
Accuracy_Percentage=c((92,91,93),models)
```

Figure 5: R Application trace for comparision

Table 4 : Accuracy Comparisons of classification Approaches

S.No	Classification Approach	Percentage of Accuracy
1	kNN	92.1
2	SVM	90.2
3	Decision Tree	92.2

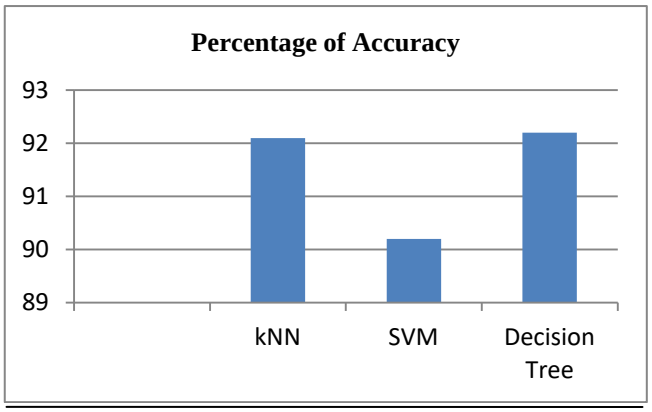

Figure 6: Accuracy Comparisons of classification Approaches

Decision Tree performed effectively and with efficiency than different algorithms because the range of various observations made over distinct variables randomly selected from the dataset are considerably small and additionally due to this not all variables are part of observation in IRIS information hence is non-linearly dependent which is higher than observation support vector machine performed poorer than kNN.

V. CONCLUSION

Thus in this paper we have compared the performance of various classifiers. Iris data set (UCI) is used for experimentation. Numbers of cross-folds in each case are 10. In terms of overall performance that is if we consider Accuracy of kNN, SVM and Decision Tree. The Algorithms Decision tree and kNN performs comparatively better than SVM. Thus we found that finally different classifiers are also compared with the help of Confusion Matrix.

REFERENCES

1. Haiyang, Z., 2011. A Short Introduction to Data Mining and Its Applications, IEEE.
2. Witten, I.H., E. Frank and M.A. Hall, 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann.
3. Yi~Shang, H. Shi, and S.S. Chen, "An Intelligent Distributed Environment for Active Learning", in Proc. ACM Journal of Educational Resources in Computing, 2001.
4. https://rpubs.com/
5. Praveen P., Jayanth Babu C. (2019) Big Data Clustering: Applying Conventional Data Mining Techniques in Big Data Environment. In: Saini H., Sayal R., Govardhan A., Buyya R. (eds) Innovations in Computer Science and Engineering. Lecture Notes in Networks and Systems, vol 74. Springer, Singapore
6. Discovering Interesting Classification rules with Genetic Programming: De Falco, A. Della Cioppa, E.Tarantino.
7. Agrawal, R., T. Imielinski and A.N. Swami, 1993. Database Mining: A Performance Perspective, IEEE Trans. Knowledge and Data Engineering, 5(6): 914-925.
8. P. Praveen, B. Rama, Uma N. Dulhare, "A study on monothetic Divisive Hierarchical Clustering Method", *International Journal of Advanced Scientific Technologies Engineering and Management Sciences*, vol. 3, no. 1, March 2017, ISSN 2454-356X.
9. Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2nd .
10. Mohammed Ali Shaik, P. Praveen , R. Vijaya Prakash.(2019) "Novel Classification Scheme for Multi Agents" ,Asian Journal of Computer Science and Technology ISSN: 2249-0701 Vol.8 No.S3, 2019, pp. 54-58.
11. UCI Machine Learning Repository: http://archive. ics. uci.edu/ml/datasets.html
12. R. Ravi Kumar, M. Babu Reddy and P. Praveen, "A review of feature subset selection on unsupervised learning", Proc. 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), pp. 163-167, Chennai, 2017.
13. B. Rama, P. Praveen, H. Sinha and T. Choudhury, "A study on causal rule discovery with PC algorithm," *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, Dubai, 2017, pp. 616-621.doi: 10.1109/ICTUS.2017.8286083
14. Comparison of different classification techniques using different datasets: V.vaithiyanathan, K.Rajeswari, Rahul Pitale.
15. A survey on the Application Of Genetic Programming to Classification: Pedro G. Espejo, Sebastian Ventura, and Francisco Herrera