IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

A Comparative Analysis on Evaluation of Classification Algorithms Based on Ionospheric Data

Chandrika[1*], Divya. C[2], Gowramma. G. S[3], Varun. C. R[4]

[1] Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India
[2] Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India
[3] Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India
[4] Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru, India

Available online at: www.ijcseonline.org

Accepted: 21/May/2018, Published: 21/May/20182018

Abstract— Data mining technique is an application of the regular process for analyzing the huge size of existing data, excavating valuable information to support the decision-making process. The Earth's upper atmosphere consists of an ionized part referred to as the ionosphere. It lies between eighty kilometre to one thousand kilometer height above the sea level, an area which comprises the parts of the thermosphere, mesosphere as well as the exosphere. The ionosphere is a shell of electrons and electrically stimulated atoms that ambiances the Earth. The target for Weka tool classification are these free electrons in the ionosphere. The performance analysis and experimental results carried out for five classifiers such as Naive Bayes, SVM, ANN, K-NN, and J48 are compared and evaluated in this study. The overall performance of these algorithms is analyzed based on the classification accuracy in which decision tree algorithm has achieved best performance compared to other algorithms. The above accuracy in ionospheric data classification is the focal idea of assessing the performance in data mining algorithms.

Keywords— Data mining, Naive Bayes, SVM, ANN, K-NN, J48

I. Introduction

Data mining is a dominant technology with the notable competence to assist the organization in present-days. The data mining tools forecast the forthcoming implications and knowledge driven decisions. It is a process of excavating the valuable information from an enormous amount of data. The data mining is classification based on different data models and data types. Data mining routines refine mathematical algorithms for the segmentation of the data and assess the probability of upcoming events. Data mining is also referred to as Knowledge Discovery in Databases (KDD) [1]. The Earth's upper atmosphere consists of an ionized part referred to as the ionosphere. It lies between eighty kilometer to one thousand kilometer height above the sea level, an area which comprises the parts of thermosphere, mesosphere as well as exosphere. The solar radiations help in ionizing the ionosphere. It plays a key role in atmospheric electricity. It has practical prominence because, it impacts radio propagation to distant places on the Earth [2].

The Earth's atmosphere is stroked by solar radiations at a power density of 1300 watts per square meter (known as Solar Constant) leading to the formation of the ionosphere. This radiation level is intense and the spectrum range lies from radio frequencies to infrared and visible light to X-rays. Solar radiation is considered as "ionizing" at ultraviolet and shorter wavelengths because an electron is dislodged from neutral gas atom at these frequencies when the collision occurs. This solar radiation falls on a gas atom. During this process, the atom absorbs a part of this radiation producing a free electron and a positively charged ion. At the atmosphere's highest level, there are very fewer atoms to interact because of very strong solar radiation so the occurring of ionization is in a small amount. Ionization increases as there is a decrease in altitude because of the presence of more gas atoms. In the meantime, a free moving electron is caught by a positive ion when it travels closer to it, this process is called as recombination. At a lower altitude, there is a rise in gas density where recombination speed up because molecules and ions are close together. The presence of the degree of ionization is determined when there is a point of balance between the two processes. The ionosphere consists of a shell of electrons and electrically stimulated atoms that ambiances the Earth. The target for Weka tool classification are these free electrons in the ionosphere.

Classification techniques are controlled learning techniques that catalog data element into the already defined class label. Data mining helps in building classification models from an input data set which has turned out to be the most beneficial technique in this field. There are numerous algorithms for classification such as Bayesian classifier, Functional classifier, Lazy classifier, Decision tree classifier and Rule-Learner classification.

II. RELATED WORK

Sigillito V G., Wing S P, Hutton L V and Baker K B in 1989 performed a Classification on radar returning from the ionosphere by means of neural networks. The investigation was carried out using a backdrop as well as the perceptron training algorithm on the database. Here the first 200 instances were used for training, which was wisely distributed 50% positive and 50% negative, it was discovered that a "linear" perceptron reached 90.7%, a "non-linear" perceptron reached 92%, and backdrop attained an average accuracy of 96% on the rest 150 test instances, comprising of 123 "good" and only 24 "bad" instances [3].

Marie Fernandes has stated in her work that data mining a procedure that inspects information from alternate view points and wraps it into supportive data [4].

P. Rutravigneshwaran in his research has evaluated the efficiency of machine learning methods in intrusion detection system using classification tree and support vector machine. The comparison depicts that C4.5 outperforms the SVM in accuracy and detection [5].

III. METHODOLOGY

Information of the Data Set used:
The radar dataset is collected via a system in Goose Bay, Labrador. It comprises 16 high-frequency antenna of a phased array having the total transmission of power is in the order of 6.4 kilowatts. For this study, free electrons in the ionosphere are considered as the targets. To show indication of some type of structure in the ionosphere “Good” radar is returned. “Bad" returns are those that do not; their signals pass through the ionosphere. Pulse number for this Goose Bay system was 17. Database instances are marked by 2 attributes for pulse abode, to the degree that esoteric values reverted by the field causing from the complex electromagnetic signal.

Attribute Information:
In ionospheric dataset, it consists of 351 instances and 34 attributes plus a class attribute. All 34 predictor attributes are continuous in nature. The 35th attribute is defined as either "good" or "bad". This task is carried out as binary classification. Missing values are not found in the dataset.

A. Naive Bayes Classifier

The Naive Bayes classifier is a simple and powerful technique. It also comparatively overtakes other classification algorithms. Thus, it is one of the famous algorithms. Naive Bayes independently observes and analyzes the variables in the data sample. Classification problems can be solved after the testing process. Here the models are built fast by giving improved predictions. Naive Bayes algorithm plays a major role in finding the missing data. By characterizing the problem in Naive Bayes, the unseen data can be easily predicted. The attributes are separated during the construction and prediction time. During the isolation process, only the sufficient data is required by the probability of each attribute. Thus, the collection of more data is not required. The performance of this classification will be degraded if the data contains highly associated features.

B. SVM (Support Vector Machines)

Support vector machines come under supervised learning and discriminative classifiers. It is also called as SMO (Sequential Minimal Optimization). For the best performance, SVM makes use of kernels for transforming the problem from linear classification techniques to non-linear data. Non-traditional data such as strings and trees are used as input for SVM classification. Hence, both the type of datasets such as small and large can be applicable. The data instances in multi-dimensional space is arranged by the application of kernel equations where the classes are separated by the hyperplane where categorization is done. The target variable value is decided by the hyperplane. The maximized margin among the support vectors on each side of the plane has to be selected. Support vectors will be on each side of the separating planes or a slightly on the wrong side where Kernel equations are found to be linear, quadratic or anything that accomplishes the tenacity. In SVM, the data that has to be separated should be in binary. The data is treated as binary by the machine, even though it is not a binary value and sequence of binary assessment is performed.

C. ANN (Artificial Neural Network)

Artificial Neural Network is a data processing algorithm which is an interconnected cluster of nodes inspired by the human brain. There are two feedforward networks in ANN namely Single Layer Perceptron and Multi-Layer Perceptron. ANN is also referred to as Multi-Layer Perceptron in Weka tool. It has three layers: input, hidden and the output layer. Each circular node signifies a data structure called as neuron and a line connects where the output of a neuron is an input to the other neuron. Input neurons are present in the first layer (input layer) which help in sending data to the next layer (hidden layer) and then to the third layer (output layer). To handle the data processing huge amount of small processors are included in the system. In order to solve a problem, the processors behave as an interconnected network parallel to one another.

D. k-NN (K-nearest neighbour)

K-Nearest Neighbour is a basic Machine Learning classification algorithm. It belongs to supervised learning domain and is non-parametric. The intention of using this algorithm is to foresee the classification of new sample point by separating the data points present in the database. This algorithm is based on feature similarity. Applications include pattern recognition, data mining and intrusion detection. The training phase of k-NN is very fast. It comes under the type of lazy learning or instance-based learning, where there is a local approximation of the function. The major disadvantage of the k-NN algorithm is that its accuracy can be corrupted by the existence of noisy or extraneous features. Each point in n-dimensional space represents an instance creating pattern space of training tuples. This algorithm looks for the training tuples which are close to the new instance, whenever an unknown instance arrives. The algorithm decides which points from the training set is selected for deciding on the class to foresee new observation in order to select the k closest data points. This is why it is called the k Nearest Neighbours algorithm. It can be summarized as follows, with a new sample a positive integer k is specified. The k entries which is close to the new sample in the dataset are selected. The entries with most common classification are identified. Thus the new sample is classified.

E. J48 (Decision Tree)

Decision tree classifiers are one among the widespread tools under classification techniques. It is a predictive machine-learning algorithm which elects the dependent variable target value based on several attributes of the data that is available. Commonly, decision tree classifiers are represented in the form of tree-like structure starting from root attributes and ending with leaf nodes. It has four partitions such as Decision node, leaf node, edge and path [6]. Decision node consists of a solo attribute. The target attribute is defined by the leaf node. Splitting of one attribute is edge and the path is a final decision. The algorithm for J48 Decision tree classifier is as follows. A decision tree is generated based on the attributes of the training data that is available. So that when a training set is encountered, various instances are discriminated by that attribute. These data instances give us a clear idea so the classification is done in a way that the best- classified data has the highest information gain. If there is any value having no ambiguity then the branching is terminated and is consigned to the obtained target value. Decision tree has several benefits - It can handle a variety of input data (Nominal, Numeric and Text), it can process erroneous datasets or missing values. This can be implemented on data mining packages over multiple platforms [7].

Table 1. Description of Algorithms [8]

Algorithms	*Purpose*	*Limitations*
Naive Bayes	Text Classification	Existence of dependencies among variables
	Spam Filtering	Naive Bayes Model cannot classify these dependencies
	Hybrid Recommender System	
SVM	It is also effective in text classification	To achieve best classification, key parameters are needed.
	Capturing of essential data characters	High complexity for classification
	Has high accuracy classification	To solve parameter model, interpretation is difficult
ANN	Fast testing process	Training process is slow
	Shows good result for complex domains	
	Shows better result for continuous domains	
K-NN	Simple, non-parametric and easy to implement	Long classification time
	Low error rate is found during training process	Finding optimal value is difficult
J48	Variable screening or feature selection is performed implicitly	There are possibilities of spurious relationships
	It is easy to interpret and explain	Limited to one output per attribute

IV. Results and Discussion

In this segment, we present the performance analysis and experimental results carried out for this study. Here five classifiers such as Naive Bayes, SVM, ANN, K-NN and J48 are conducted for analysis. The accuracy measurements are checked using cross-validation with 10 folds of a training set. We use Weka tool application for implementing this work.

Table 2. Performance of Algorithms

Algorithms Used	*Correct Classification of Instances (%)*	*Incorrect Classification of Instances (%)*
J48	91.453	8.547
ANN	91.168	8.831
SVM	88.604	11.396
K-NN	86.324	13.675
Naïve Bayes	82.621	17.378

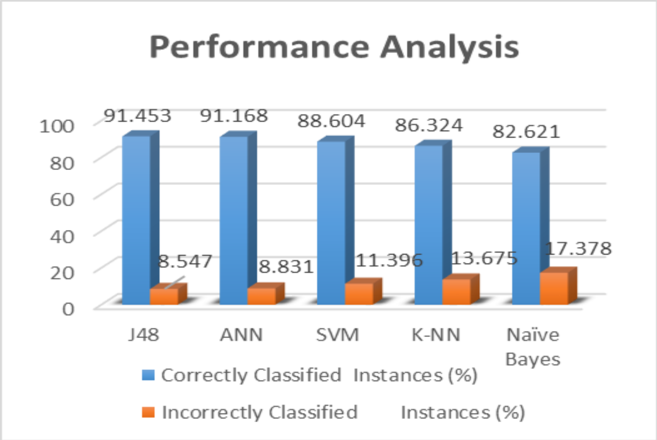

Figure 1. Graphical form of Performance Analysis

The above chart describes the behavior of different data mining algorithms used to identify the classification accuracy for the ionosphere dataset. From this study, Decision Tree (J48) having performance accuracy of 91.45% and ANN having performance accuracy of 91.17% are considered to be comparatively better than the other algorithms.

Table 3. Classification Accuracy

Sl No.	*Algorithms*	*Classification Accuracy*
1	J48	91.45%
2	ANN	91.17%
3	SVM	88.60%
4	K-NN	86.32%
5	Naive Bayes	82.62%

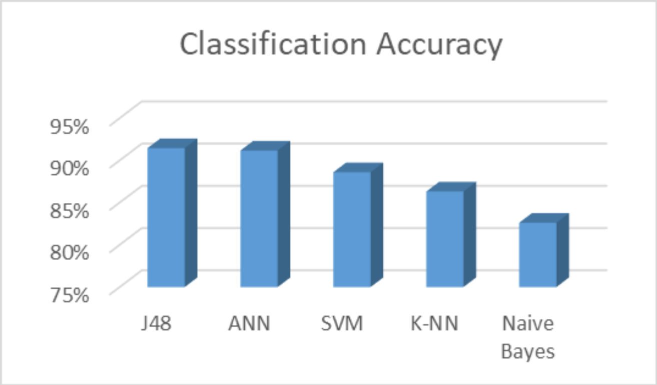

Figure 2. Graphical form of Classification Accuracy

The above figure displays the classification accuracy of different data mining algorithms. The Decision Tree (J48) algorithm shows the highest accuracy in the above classification.

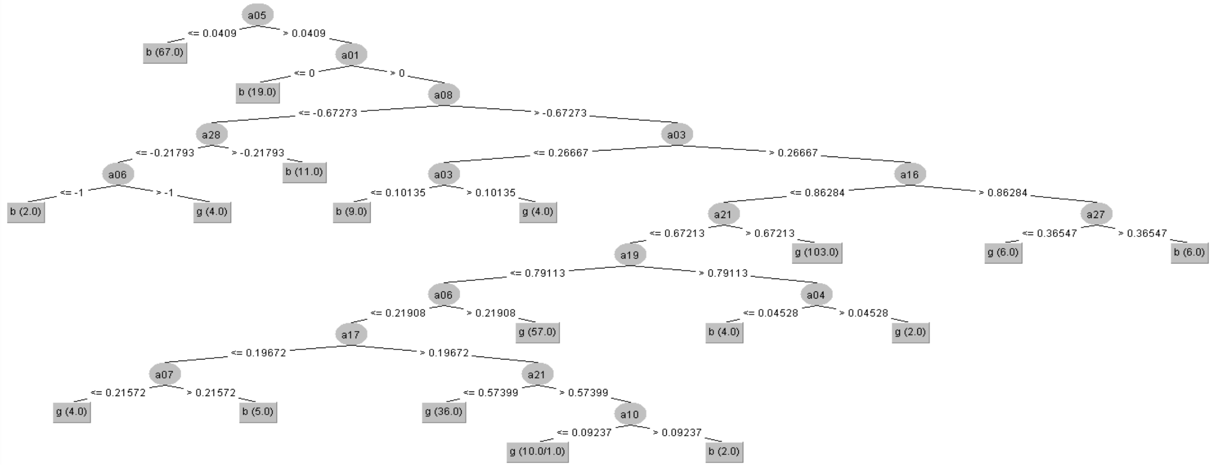

Figure 3. Decision Tree build using J48 classifier

The above decision tree created using Weka tool shows that "Good" radar returns shows the presence of some type of structure in the ionosphere. "Bad" radar returns are those that do not have any structure in the ionosphere.

V. Conclusion and Future Scope

In this study, Naïve Bayes, SVM, ANN, KNN and J48 were applied to Ionospheric dataset. Here, the overall performance of several algorithms is analyzed based on the accuracy. These chosen algorithms found to have accuracy as follows: Naïve Bayes provides 82.62%, SVM is 88.60%, ANN shows 91.17%, K-NN gives 86.32% and finally, Decision tree shows 91.45%. Based on the above investigation Decision Tree (J48) has achieved the best performance. The above accuracy in ionospheric data classification is the focal idea of assessing the performance in data mining algorithms. The whole result shown in the paper is stepping into additional development in imminent technology. For future study, we can put on different data mining methods on the data set to get more precise results. Aside from classification, data mining practices like clustering algorithms can be applied to the dataset to get knowledge from it.

References

[1] Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996), "*From Data Mining to Knowledge Discovery in Databases*"

[2] K. Rawer, "*Wave Propagation in the Ionosphere*". Kluwer Acad.Publ., Dordrecht 1993. ISBN 0-7923-0775-5

[3] Sigillito V G., Wing S P, Hutton L V and Baker K B, "*Classification of radar returns from the ionosphere using neural networks*" Johns Hopkins APL Technical Digest, 10, 262-266.

[4] Marie Fernandes , "*Data Mining: A Comparative Study of its Various Techniques and its Process*", International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.19-23, 2017.

[5] P. Rutravigneshwaran, "*A Study of Intrusion Detection System using Efficient Data Mining Techniques*", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.6, pp.5-8, 2017.

[6] P.Keerthana et al, "*Performance Analysis of Data Mining Algorithms for Medical Image Classification*" International Journal of Computer Science and Mobile Computing, Vol.5 Issue.3, March- 2016.

[7] Rokach, Lior, and Oded Maimon. "*Decision Trees*" 28. Web. 1 Feb. 2013.

[8] P Thamilselvana, Dr. J. G. R. Sathiaseelanb, "*A Comparative Study of Data Mining Algorithms for Image Classification*" Published Online June 2015 in MECS. DOI: 10.5815/ijeme.2015.02.01.

Authors Profile

Ms. Chandrika pursuing Bachelor of Engineering in Computer Science from Don Bosco Institute of Technology, Bengaluru, affiliated to Visvesvaraya Technological University.

Ms. Divya C pursuing Bachelor of Engineering in Computer Science from Don Bosco Institute of Technology, Bengaluru, affiliated to Visvesvaraya Technological University.

Mrs. Gowramma G S is an Electronics and Communication Engineering graduate from Bapuji Institute of Engineering and Technology, Davangere and has Master Degree in Engineering from Dr. MGR University, Chennai specialized in Computer Science and Engineering. She is currently pursuing Ph.D. on Data Mining and Artificial Intelligence from Visvesvaraya Technological University, Belgaum and also working as Associate Professor in Department of Computer Science and Engineering, Don Bosco Institute of Technology having 13 years of teaching experience and 5 years of Industry Experience. She has published 2 research papers in international and 5 research papers in National Conferences/Journals.

Mr. Varun C R pursued Bachelor of Engineering from Visvesvaraya Technological University in the year 2011 and M.Tech from M S Ramaiah Institute of Technology, Bengaluru in the year 2014. He is currently pursuing Ph.D. on Data Mining and also working as Assistant Professor in Department of Computer Science and Engineering, Don Bosco Institute of Technology, Bengaluru since 2014 and has 4 years of teaching experience.