East African Scholars Journal of Engineering and Computer Sciences
Abbreviated Key Title: East African Scholars J Eng Comput Sci
ISSN 2617-4480 (Print) | ISSN 2663-0346 (Online) |
Published By East African Scholars Publisher, Kenya
Volume-2 | Issue-12 | Dec-2019 | DOI:10.36349/EASJECS.2019.v02i12.03

OPEN ACCESS

Review Article

Comparative Analysis of Various Data Mining Classification Algorithms

Er. Sandeep Kaur[1*], Dr. Mary Amirtha Sagayee G[2], Er. Arshdeep Singh[3], Prince Jindal[3]

[1]Research Scholar, Department of CSE, BGIET Sangrur, India
[2]Professor, Department of CSE, BGIET Sangrur, India
[3]Assistant Professor, Department of CSE, BGIET Sangrur, India

*Corresponding Author
Er. Sandeep Kaur

Abstract: Data mining is the process of digging through and analyzing various sets of data and then extracting the meaning of the data. Classification is a data mining method used to predict the class of objects whose class label is not known. There are many classification mechanisms used in data mining such as KNearest Neighbor (KNN), Bayesian network, Cross validated parameter selection (CVPS), Naive Bayes Multinominal Updatae- ble (NBMU) Algorithm, Fuzzy logic, Support vector machines etc. This paper presents a comparison on four classification techniques which are K-Nearest Neighbor, User Classifier, Cross validated parameter selection and Naive Bayes Multinominal Updataeble Algorithms. The goal of this research is to enumerate the best technique from above four analyzed under a given data set and provide a fruitful comparison result which can be used for further analysis or future development.

Keywords: KNN, CVPS, NBMU, User Classifier.

1. INTRODUCTION

Data mining concept is growing very fast in popularity, it is a technology that involving methods at the intersection of (Ar- tificial intelligence, Machine learning and database system), the main goal of data mining process is to extract information from a large data into a form which could be understandable for further use. Classification is a data mining technique based on machine learning [1]. Basically, it is used to classify each data item in a set of data into one of a predefined set of classes or groups. The classification technique makes use of mathematical techniques such as decision trees, linear programming, neural network etc. In classification, we make the various types of software that can learn how to classify the data items into groups. This research has conducted a comparison study between a number of available data mining software and tools depending on their ability for classifying data correctly and accurately. The accuracy measure which represents the percentage of correctly classified instances that is used for judging the performance of the selected tools and software. The rest of the paper is organized as follows: Section 2 summaries related works on data mining, data classification. Section 3 summaries the various types of data classification techniques used. Section 4 provides a general description of the tools and software under test and dataset used. Section 5 reports experimental results and compares the results of the different algorithms. Finally, I close this paper with a summary and an outlook for some future work.

2. LITERATURE SURVEY

Oliver, *et al.*, (2012) proposed Introduction to k Nearest Neighbor Classification and Condensed Nearest Neighbor Data Reduction; k Nearest Neighbors (KNN) [2] algorithm is to use a database in which the data points are separated into several separate classes to predict the classification of a new sample point. The process of choosing the classification of the new observation is known as the classification problem and there are the various ways to tackle it. Here we consider choosing the class of the new observation based on the classes of the observations in the database which it is most similar" too.

Thair, *et al.*, (2009) suggested that Classification is a data mining or machine learning technique used to predict group membership for data instances. In this, he presents the basic classification mechanisms. Several major kinds of classifica- tion techniques including decision tree induction, Bayesian

Quick Response Code

Journal homepage:
https://easpublisher.com/journal/easjecs/home

Article History
Received: 01.12.2019
Accepted: 14.12.2019
Published: 30.12.2019

Copyright @ 2019: This is an open-access article distributed under the terms of the Creative Commons Attribution license which permits unrestricted use, distribution, and reproduction in any medium for non commercial use (NonCommercial, or CC-BY-NC) provided the original author and source are credited.

networks, k-nearest neighbor classifier [3], case- based reason- ing, genetic algorithm and fuzzy logic techniques. The goal of this survey is to provide a comprehensive review of different classification mechanisms in data mining.

Delveen, *et al.,* (2013) proposed, Data mining concept is growing very fast in popularity, it is a technology that in- volving large no. Of methods at the intersection of (Machine learning, database system and Statistics), the main goal of data mining method is to extract information from a large data into a form which could be understandable for further use. Some algorithms for data mining are used to give solutions to various classification problems in the database. In this a comparison among three classifications algorithms will be studied, these are (K- Nearest Neighbor classifier, Decision tree [4] and Bayesian network) algorithms. In this he demonstrates the strength and accuracy of each algorithm for classification in term of performance efficiency and time complexity required.

Sohil, *et al.,* (2013) suggested that there are several methods of data mining like classification, clustering, association rule, outlier analysis, etc. That is used for uncovering hidden patterns from the data. There are various algorithms of above techniques are developed by various researchers. In this he tried to examine and investigate various techniques of classifications like Decision Trees, Naive Bayes, k-Nearest Neighbor [5], Feed Forward Neural Networks and Support Vector Machine to identify the best fit methods among them. All the above mentioned algorithms were implemented using WEKA which consists of a collection of machine learning algorithms for data mining tasks.

Abdullah, *et al.,* proposed that huge amount of data and info. Are available for everyone, Data can now be stored in many different kinds of databases, besides being available on the Internet. With such amount of data, there is a need for powerful methods for better interpretation of these data that exceeds the human's ability for making decision in a better way. In order to get the best tools for classification task that helps in decision making [6]. He has shown a comparative study between a number of freely available data mining and knowledge discovery tools. He has shown the results that the performance of the tools for the classification task is affected by the kind of dataset used and by the way the classification techniques were implemented within the toolkit.

Yogesh, *et al.,* (2013) suggested that network security needs to be concerned to provide secure information medium due to increase in potential network attacks. In today's era detection of various security threats that are commonly referred to as intrusion [7], has become a very critical issue in the network. Highly secured data of large organizations are present over the network so in order to prevent that data from unauthorized users or attackers a strong security technique is required. Intrusion detection system plays a major role in providing security to computer networks. Intrusion Detection System is a valuable tool for providing security to computer networks. In this paper a comparative analysis of different feature selection mechanisms is presented on the KDD dataset and their performance are evaluated in terms of computational time, detection rate and ROC.

3. TECHNIQUES USED

3.1. K-Nearest Neighbor Classifiers (KNN)

K-NN is a type of lazy learning where the function is only approximated locally and all computation is deferred until classification. The k-NN is a lazy learning technique [8], and instead of estimating the target function once for the whole instance, they delay processing until classification. This algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the most common class amongst its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor. The neighbors are taken from a collection of objects for which the correct classification is known. In this no training step is required. The k-nearest neighbor algorithm is sensitive to the local structure of the data. K-Nearest Neighbors (K-NN) algorithm is a nonparametric method in that no parameters are estimated.

For eg: To classify an unknown object:

- Compute distance to other training objects
- Identify total no. Of k nearest neighbors Use classification of nearest neighbors to determine the class label of unknown record Algorithm of KNN: Consider k as the desired number of nearest neighbors and S: = p1,..., pn is the set of training samples in the form p1 = (xi, ci), where xi is the dimensional feature vector of the point pi and ci is the class that pi belongs to. For each p'= (x', c')
- Compute the distance d (x', xi) between p' and all pi belonging to S using Euclidean distance formula: d(p,q)=$\sqrt{\sum_i (p_i - q_i)2}$
- Sort all points pi according to the distance d (x', xi).
- Select the first k training samples from the sorted list, those are the k closest training samples to p'. Assign classification to p' based on majority vote of classification.

3.2. Decision tree

Decision trees are trees that classify objects by sorting them based on feature values. Each node in a tree represents a feature value in [9] an object to be classified, and each branch represents a value of the node. Objects are classified starting at the root node and

sorted based on their feature values. Decision Trees offer many benefits of data mining technology like:

- Easy to follow when compacted.
- The ability of handling a variety of input data: numeric and text etc.
- High performance in a relatively small computational effort.
- Useful for various techniques, such as classification, clustering and feature selection etc.

3.3 Naive Bayes Multinomial Updateable Algorithm

The task of text classification can be approached from a Bayesian [10] learning perspective, which predicts that the word distributions in documents are generated by a specific parametric model, and the parameters can be estimated from the training data. There is an option available for NaiveBayes-Multinomial. When the option is debugging and if it's set to true, classifier may output additional info to the console. This is the incremental version of the naive Bayes multinomial algorithm. This uses the Bayes rule theory as its core equation.

3.4. User Classifier

Interactively classify through visual means. You are Presented with a graph of the data against two user selected attributes [11], as well as a view of the decision tree. By creating polygons around data plotted on the scatter graph, You can create binary splits, as well as by allowing another classifier [12] to take over at points in the decision tree should you see fit. There is an option available for User Classifier. When the option is debugging and if it's set to true, classifier may output extra info to the console.

3.5. CV Parameter Selection Algorithm

This is a class for performing parameter selection by cross validation for any classifier. There are various types of the options available for CV Parameter Selection. When the option is CV Parameters, then it sets the scheme parameters which are to be set by cross-validation. When the option is a classifier, then the base classifier is to be used. When the option is debugging and if it's set to true, classifier may output extra info to the console. If the option is num Folds, then it gets the no. Of folds used for cross validation. If the option is a seed, then the random no. seed to be used.

4. THE COMPARATIVE STUDY

The methodology of the study consists of collecting a set of data mining and knowledge discovery tools to be tested, specifying the data set to be used, and selecting a various set of the classification algorithm to test the tools' performance.

4.1 Tools Description

Weka 3.6 is a collection of machine learning algorithms for data mining tasks. Weka stands for Waikato Environment for Knowledge Analysis [13]. The algorithms can either be applied directly to a dataset or called from the Java code. Weka contains various tools for data pre-processing, classification, regression, association rules, clustering, and visualization. The Weka GUI Chooser (class weka.gui.GUIChooser) provides a starting point for launching Weka's main GUI applications and supporting tools. The GUI Chooser consists of four buttons: one for each of the four major Weka applications and four menus. The buttons can be used to start the applications that are explained as follows:

- **Explorer:** It is an environment used for exploring data with WEKA (the rest of this documentation deals with this application in more detail). Experimenter: It is an environment for performing experiments and conducting statistical tests between learning schemes.
- **Knowledge Flow:** This environment supports essentially the same functions as the Explorer, but with a drag-and drop interface. It supports incremental learning.
- **Simple CLI:** It provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

4.2 Data Set Description

To verify the efficiency of KNN algorithm with other classifi- cation algorithm, I have used KDD dataset. This dataset contains 39 features and is labeled with exact one specific attack type i.e., either normal or an attack. Each vector is labeled as either normal or an attack, with exactly one specific attack type. Deviations from normal behavior, everything that is not normal, are considered attacks. Attacks labeled as normal are records with normal behavior. The training dataset has 53.18% normal and 46.81% attack connections. KDD CUP 99 has been most widely used in attacks on network. The simulated attack falls in one of the following four categories: Denials-of Service (DoS) attacks [7] have the goal of limiting or denying services provided to the user, computer or network. A common tactic is to severely overload the targeted system (E.g. apache, smurf, Neptune, Ping of death, back, mailbomb, udpstorm, SYN flood, etc.).

- Probing or Surveillance attacks have the goal of gaining knowledge of the existence or configuration of a computer system or network. Port Scans or sweeping of a given IP address range typically fall in this category (e.g. saint, portsweep, mscan, nmap, etc.).
- User-to-Root (U2R) attacks have the goal of gaining root or super-user access to a particular computer or system on which the attacker previously had user level access. These are attempts by a non-privileged user to gain administrative privileges (e.g. Perl, xterm, etc.).

- Remote-to-Local (R2L) attack is an attack in which a user sends packets to a machine over the internet, which the user does not have access to in order to expose the machine vulnerabilities and exploit privileges which a local user would have on the computer.

Features of data set are grouped into four categories: Basic Features: Basic features can be derived from packet headers without inspecting the payload.

- **Content Features:** Domain knowledge is used to assess the payload of the original TCP packets. This includes features such as the number of failed login attempts.
- **Time-based Traffic Features:** These features are designed to capture properties that mature over a 2 second temporal window. One example of such a feature would be the number of connections to the same host over the 2 second interval. Host-based Traffic Features: Utilize a historical window estimated over the number of connections in this case100 instead of time. Host based features are therefore designed to assess attacks, which span intervals longer than 2 seconds.

In order to test the classifiers, I randomly selected 4973 connection records as a training data set and 1000 connection records as a testing data set. Below Table 1 shows the detail of connection records in these both datasets. KDD dataset contains symbolic as well as continuous features.

Table 1: Details of connection records in used dataset

Label	Training set	Testing set
Normal	2645	269
Probe	114	114
DOS	2147	550
U2R	21	21
R2L	46	46
Total	**4973**	**1000**

5. EXPERIMENTS AND EVALUATIONS

5.1 Result Evaluation Parameters

1) The correctly and incorrectly classified instances show the percentage of test instances that were correctly and incorrectly classified. The percentage of correctly classified instances is often called accuracy or sample accuracy.
2) Root Mean Squared Error (RMSE): The RMSE is a quadratic scoring rule which measures the average magnitude of the error.

RMSE= sqrt $((p_1\text{-}a_1)^2+\ldots..+ (p_n\text{-}a_n)^2/n)$

3) Relative Absolute Error (RAE): It is just the total, absolute error, with the same kind of normalization.

RAE= $(|p_1\text{-}a_1|+\ldots+|p_n\text{-}a_n|)/(|\bar{a}\text{-}a_1|+\ldots+|\bar{a}\text{-}a_n|)$

(4) Root Relative squared error (RRSE): The root relative squared error takes the total root of squared error and normalizes it by dividing the total squared error of the default predictor. Root relative squared error Ei of an individual program i is evaluated by the equation:

$$E_i = \sqrt{\sum_{j=1}^{n}(P_{ij} - T_j)2 / \sum_{j=1}^{n}(T_j - \bar{T})\,2}$$

Where P(ij) is the value predicted by the individual program i for sample case j (out of n sample cases); Tj is the target value for sample case j; and is given by the formula:

$$\bar{T} = 1/n \sum_{j=1}^{n} T_j$$

(5) Mean absolute error (MAE): The mean absolute error is less sensitive to outliers than the mean squared error. The error rates are used for numeric prediction rather than classification.

MAE=$|p_1\text{-}a_1|+\ldots..+|p_n\text{-}a_n|/n$

5.2. Result for KNN algorithm

In this I have taken upper defined KDD dataset as a training set and a testing set. By implementing the KNN algorithm on this training set and testing set by using console application, I have found the % of correctly classified instances, incorrectly classified instances, Mean absolute error, Root mean squared error, Root Relative squared error, Relative absolute error by using majority vote classification among the class of the K objects. Finally, I got the result given by the KNN algorithm as shown in below tables 2, 3 and 4.

Table 2: Output given by knn algorithm

Label	Testing set	Output set
Normal	269	10
Probe	114	220
DOS	550	550
U2R	21	0
R2L	46	220

Table 3: Correctly classified instances given by knn algorithm

Attacks	Frequency
Normal	10
Probe	114
DOS	550
U2R	0
R2L	46

Table 4: Results of knn algorithm

Parameters	Result
% of correctly classified instances	72.00
% of incorrectly classified instances	28.00
Mean absolute error	0.56
Root mean squared error	0.3302
Root Relative squared error	75%
Relative absolute error	66%

5.3: Result of different classification algorithms on Weka

In this I have taken upper defined KDD dataset as a training set and a testing set in the weka. By implementing different algorithms on this training set and testing set, I have found the % of correctly classified instances, incorrectly classified instances, Mean absolute error, Root mean squared error, and Root Relative squared error, a Relative absolute error that is shown in below table 5.

Table 5: Performance of different algorithms on weka

Parameter	User Classifier	NBMU	CVPS
% of correctly classified instances	52.33	57.89	52.33
% of incorrectly classified instances	47.6641	42.1053	47.664
Mean absolute error	0.2123	0.1687	0.2125
Root mean squared error	0.3265	0.4104	0.3265
Root Relative squared error	100%	125.68%	100%
Relative absolute error	99.92%	79.39%	100%

5.4: Comparison of Results obtained by KNN, User Classifier, CVPS and NBMU Algorithms

The below table no. 6 and figure no. 1 enable us to analyze the different algorithm results with better perception.

Table 6: Result analysis of KNN, User Classifier, NBMU, and CVPS Algorithms

Parameter	KNN	User Classifier	NBMU	CVPS
% of correctly classified instances	72.00	52.33	57.89	52.33
% of incorrectly classified instances	28.00	47.6641	42.1053	47.664
Mean absolute error	0.56	0.2123	0.1687	0.2125
Root mean squared error	0.3302	0.3265	0.4104	0.3265
Root Relative squared error	75%	100%	125.68%	100%
Relative absolute error	66%	99.92%	79.39%	100%

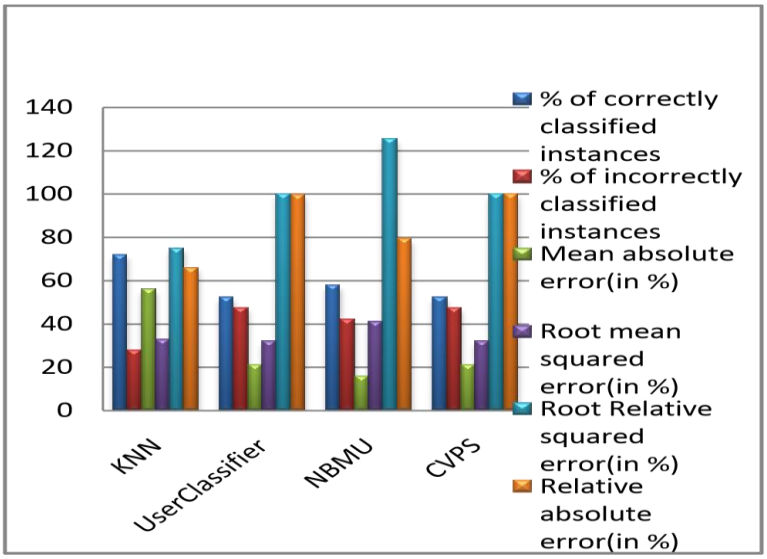

Fig. 1: Chart for comparison of various classifiers

From the results of these experiments, K-Nearest Neighbor algorithm proved to have better results of finding the 72 % of correctly classified instances from the KDD dataset. While having a % of correctly classified instances of 52.33 of User Classifier with the minimum error rate than CVPS algorithm as shown in table 6 had the second best algorithm.

6. CONCLUSION

In this work, I compare the basic classification algorithms. The goal of this study is to provide a comprehensive review of different four techniques k-nearest neighbor, User Classifier, NBMU, and CSPV Algorithms in data mining. In order to compare these four algorithms based on the Correctly classified instances, Relative absolute error, Relative squared error, Mean absolute error, Mean squared error, Root mean squared error parameters, we came to the conclusion which algorithm is more efficient to use. The performance of the each algorithm is tested on a KDD data set. After the execution of each classification algorithm, I got the numbers of correctly classified instances and the incorrectly classified instances. This gave the accuracy of the algorithm. Other important factors, Mean squared error, Root mean squared error, Relative absolute error, Relative squared error, Mean

absolute error, and describe the error rate of an algorithm. The overall evaluation shows that K-nearest neighbor algorithm is far better than User Classifier, NBMU, and CSPV Algorithms. In future studies, we can enhance the accuracy of the KNN algorithm to achieve better results than the previous methodology that I have discussed.

7. ACKNOWLEDGMENT

I highly grateful to the Dr. Tanuja Srivastava, Director, Bhai Gurdas Institute of Engineering & Technology (BGIET), Sangrur, for providing this opportunity to carry out the present thesis/work. The constant guidance and encouragement received from Er. Amandeep Kaur, Head, Department of Computer Science & Engineering, BGIET, Sangrur has been of great help in carrying out the present work and is acknowledged with reverential thanks. I would like to express a deep sense of gratitude and thanks profusely to Er Yogesh Kumar, Asstt. Prof., Department of Computer Science & Engineering, BGIET, who is the research work Supervisor. Without the wise counsel and able guidance, it would have been impossible to complete the research work in this manner. The help rendered by Er Yogesh Kumar, AP, BGIET, for the literature and their associates for experimentation is greatly acknowledged. I express gratitude to other faculty members of Department of Computer Science & Engineering, BGIET, for their intellectual support throughout the course of this work. Finally, I indebted to all whosoever have contributed in this research work and friendly stay at BGIET.

REFERENCES

1. Saabith, A. L. S., Sundararajan, E., & Bakar, A. A. (2014). Comparative study on different classification techniques for breast cancer dataset. *Int. J. Comput. Sc. Mob. Comput*, *3*(10), 185-191.
2. Sutton, O. (2012). Introduction to k nearest neighbour classification and condensed nearest neighbour data reduction. *University lectures, University of Leicester*, *1*.
3. Phyu, T. N. (2009). Survey of classification techniques in data mining. *In Proceedings of the International Multi Conference of Engineers and Computer Scientists*, 1, 18– 20.
4. AL-Nabi, D. L. A., & Ahmed, S. S. (2013). Survey on classification algorithms for data mining: (comparison and evaluation). *Computer Engineering and Intelligent Systems*, 4(8), 18–24.
5. Pandya, S. D., & Virparia, P. V. (2013). Comparing the application of classification and association rule mining techniques of data mining in an indian university to uncover hidden patterns. *In Intelligent Systems and Signal Processing (ISSP), 2013 International Conference on. IEEE*, pp. 361–364.
6. Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., & Al- Shawakfa, E. M. (2011). A comparison study between data mining tools over some classification methods. *International Journal of Advanced Computer Science and Applications, Special Issue*, 18–26.
7. Kumar, K., Kumar, G., & Kumar, Y. (2013). Feature selection approach for intrusion detection system. *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, *2*(5), 47-53.
8. Zheng, W., & Tropsha, A. (2000). Novel variable selection quantitative structure-property relationship approach based on the k-nearest-neighbor principle. *Journal of chemical information and computer sciences*, 40(1), 185–194.
9. Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., & Wolpaw, J. R. (2006). A comparison of classification techniques for the p300 speller. *Journal of neural engineering*, 3(4), 299.
10. Sebe, N. (2005). Machine learning in computer vision. *Springer*, v29.
11. Han, J., & Kamber, M. (2006). Data Mining, Southeast Asia Edition: Concepts and Techniques. Morgan kaufmann.
12. Horton, P., & Nakai, K. (1997). Better prediction of protein cellular localization sites with the it k nearest neighbors classifier. *In Ismb*, 5,. 147–152.
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), 10–18.