www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 7 Issue 2 February 2018, Page No. 23628-23631
Index Copernicus Value (2015): 58.10, 76.25 (2016) DOI: 10.18535/ijecs/v7i2.17

Performance Analysis of KNN on Different Types of Attributes

R.Nancy Beaulah [1]
[1] Assistant Professor, Department of Computer Applications
V.V.Vanniaperumal College for Women, Virudhunagar, Tamilnadu, India

Abstract:
Data Mining is an inter-disciplinary promising field that focuses on access of information useful for high level decisions and also includes Machine Learning. Data Miners evaluate and filter the data as a result and convert the data into useful information. The useful information is converted into knowledge by performing some techniques. The K-Nearest Neighbor (K-NN) algorithm is an instance based learning method that has been widely used in many pattern classification tasks due to its simplicity, effectiveness and robustness. This paper presents a performance comparison of KNN algorithm in various data sets that includes different types of attributes. The results of this paper are achieved using WEKA tool.
Keywords: Data Mining, Classification, KNN, WEKA

1. Literature Review

In [1] Tina R. Patil, Mrs. S. S. Sherekar attempted to make comparative evaluation of classifiers NAIVE BAYES AND J48. The performance compared on the basis of classification accuracy, sensitivity and specificity. In [2] Satish Kumar David, Amr T.M. Saeb, Khalid Al Rubeaan compared Decision tree J4.8 classification algorithm, Bayesian Network, and a Naïve Bayes algorithms in Medical Bioinformatics. The evaluation is based on their accuracy, learning time and error rate. In [3] Mahendra Tiwari, Manu Bhai Jha, OmPrakash Yadav proposed a methodology for comparing the accuracy of different data mining algorithms on various datasets. The performance analysis depends on many factors encompassing test mode, different nature of data sets, and size of data set. In [4] Yogita Rani, Manju, Harish Rohil used BIRCH and CURE data mining algorithms for comparative analysis on Iris Plant dataset. In [5] Kavitha C.R, Mahalekshmi T used toxicity dataset of aliphatic carboxylic acids to make a comparison of different classification algorithms and to find out the best algorithm out of the five chosen algorithm which gives the most accurate result.

2. Proposed Methodology

For this study, the experiments and observations are carried out by using data mining tool i.e. WEKA (Waikato Environment for Knowledge Learning). It was developed by the University of Waikato, New Zealand. WEKA supports many data mining tasks such as data pre-processing, classification, clustering, regression and visualization. The workflow of WEKA would be as follows:

In this paper k-nearest neighbor classification algorithm is used and it is tested with six different data sets with different types of attributes. Before classifying the data, the data sets should be preprocessed. Preprocessing is done to clean the data, to remove noise and inconsistency. In WEKA, to remove missing values in the dataset, ReplaceMissingValues filter is used. k-nearest neighbor algorithm is implemented using IBk algorithm with k=5 neighbors. The test mode used is percentage split i.e. 50% of the data set is considered as Training set and the remaining 50% as Test set.

2.1 Data Selection:

In this paper datasets have been collected from UCI Machine Learning Repository website. The dataset contains different attributes and instances. The complete description of dataset is shown in Table 1.

Table 1 – Attribute Types

S. No	Data Set	No. of Instances	No. of Attributes	Types of Attributes
1.	Auto Imports Database	205	26	15 - Continuous 1 - Integer 10 - Nominal
2.	Ionosphere database	351	34	34 - Continuous
3.	King+Rook versus King+Pawn (kr-vs-kp) Data set	3196	36	36 - Discrete
4.	Letter Image Recognition Data Set	20000	16	16 - Integer
5.	Mushroom Database	8124	22	22 - Nominal
6.	Vehicle Silhouettes Data Set	846	18	18 - Real

3. Experimental Works and Results

An experimental comparison of *k*-NN classification technique with six different datasets is carried out in WEKA. Each of the datasets involved contains different data types as well as varied number of attributes. The computation results of *k*-NN with six datasets are listed in Table 2. The accuracy of *k*-NN is tabulated in Table 3. The comparison of error is tabulated in Table 4.

Table 2 – Results of *k*-NN

S. No	Dataset	Correctly Classified	Incorrectly Classified	Kappa Statistic
1	Auto Imports Database	54 52.9412 %	48 47.0588 %	0.3865
2	Ionosphere database	145 82.8571 %	30 17.1429 %	0.6044
3	kr-vs-kp Data set	1491 93.3041 %	107 6.6959 %	0.8652
4	Letter Image Recognition Data Set	9274 92.74 %	726 7.26%	0.9245
5	Mushroom Database	4059 99.9261 %	3 0.0739%	0.9985
6	Vehicle Silhouettes Data Set	277 65.4846 %	146 34.5154 %	0.5398

Table 3 – Accuracy

Data Set	Accuracy (%)
Auto Imports Database	52.9
Ionosphere database	82.9
kr-vs-kp Data set	93.3
Letter Image Recognition Data Set	92.7
Mushroom Database	99.9
Vehicle Silhouettes Data Set	65.5

Table 4 – Error Comparison

S. No	Data Set	Mean Absolute Error	Relative Absolute Error
1.	Auto Imports Database	0.1525	68.42%
2.	Ionosphere database	0.1957	42.43%
3.	kr-vs-kp Data set	0.1771	35.49%
4.	Letter Image Recognition Data Set	0.0094	12.74%
5.	Mushroom Database	0.0007	0.13%
6.	Vehicle Silhouettes Data Set	0.1982	52.76%

The performance of *k*-NN classifier is evaluated by using parameter such as TP (True Positive) rate, FP (False Positive) rate, TN (True Negative) rate, FN (False Negative) rate. TP is the proportion of positive cases that were correctly identified. FP is the proportion of negatives cases that were incorrectly classified as positive. TN is the proportion of negatives cases that were classified correctly. FN is the proportion of positives cases that were incorrectly classified as negative. Based on these values Figure 1 shows the comparative results of correctly classified instances with incorrectly classified instances. Figure 2 represents the accuracy of *k*-NN with these six datasets. Figure 3 shows the kappa statistic values. Figure 4 shows the error comparison results.

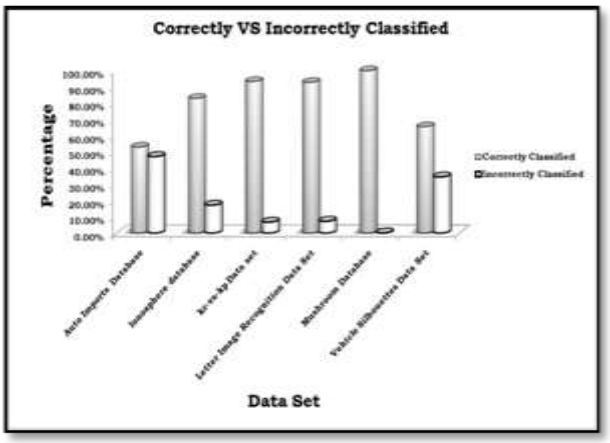

Figure1: Correctly VS Incorrectly classified

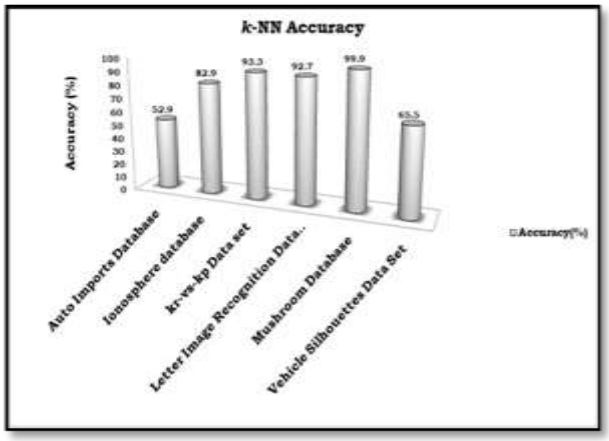

Figure 2: k-NN Accuracy

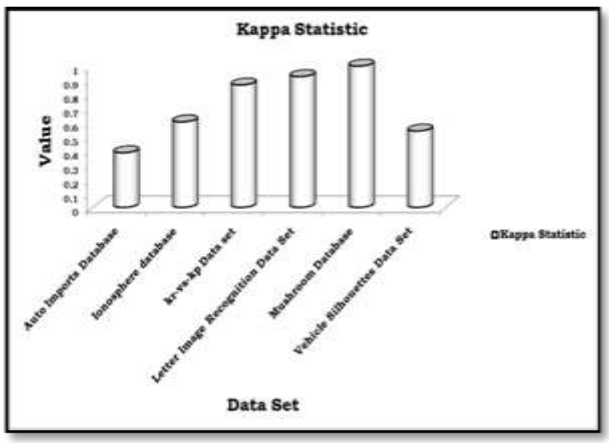

Figure 3: Kappa Statistic

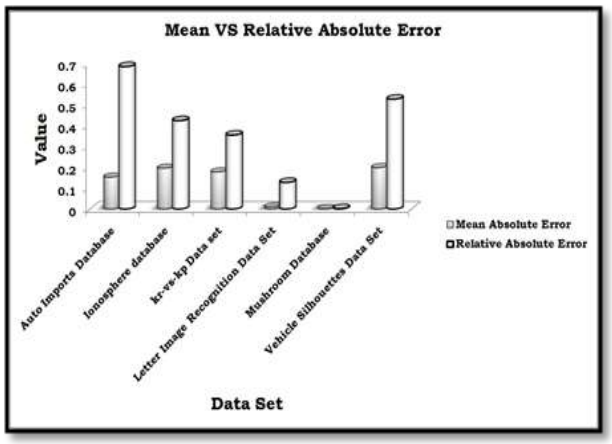

Figure 4: Error Comparison

4. Conclusion

The above results clearly show that the highest accuracy is achieved for Mushroom dataset that contains nominal attributes. The next highest accuracy is achieved for discrete and integer attributes. The comparative results of error show a minimum value for mushroom dataset. In fact, in this experimental comparison k-NN algorithm is more suitable for nominal type of attributes.

References

[1] "Performance Analysis of Naive Bayes and J48Classification Algorithm for Data Classification", Tina R. Patil, Mrs. S. S. Sherekar. International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011

[2] "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics", Satish Kumar David, Amr T.M. Saeb, Khalid Al Rubeaan. Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online) Vol.4, No.13, 2013

[3] "Performance analysis of Data Mining algorithms in Weka", Mahendra Tiwari, Manu Bhai Jha, OmPrakash Yadav. IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 6, Issue 3 (Sep-Oct. 2012), PP 32-41

[4] "Comparative Analysis of BIRCH and CURE Hierarchical Clustering Algorithm using WEKA 3.6.9", Yogita Rani, Manju, Harish Rohil. The SIJ Transactions on Computer Science Engineering & its Applications (CSEA), Vol. 2, No. 1, January-February 2014

[5] "A Comparative Study of Classification Algorithms on Aliphatic Carboxylic Acids Data Set using WEKA", Kavitha C.R, Mahalekshmi T. International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-3 Issue-7, May 2015