

Preprocessing of Various Data Sets Using Different Classification Algorithms for Evolutionary Programming

R. Karthick¹, Dr. A. Malathi²

¹Research Scholar, Karpagam University, Coimbatore

²Assistant professor, Government Arts College, Coimbatore

Abstract: *The performance analysis and preprocessing of data is very important in data mining. Here, the performance is analyzed on various classification algorithms using different data sets so as to identify which results the best accuracy of different data. The preprocessing helps us to improve data efficiency and remove the noisy data which helps to identify the survival of the fittest. This paper presents the efficient data based on the various classification algorithms and preprocessing is performed. Here, various data sets are chosen, which results in comparison of data set based on the classification algorithm.*

Keywords: Classification, Data Mining, Preprocessing, Genetic algorithm

1. Introduction

Data Mining is essential step in the processing the data in data base. Data pre-processing takes several forms, including data cleaning, data transformation and data reduction. Data cleaning is done to remove noise in the data. Data transformation is to normalize the data. Finally data reduction is to reduce the amount of data by aggregating values or removing and clustering redundant attributes.[1] Data Mining refers to process or method that extracts or mines interesting knowledge or patterns form large amounts of data. It searches for relationship and patterns that exist in large database but are hidden among vast amount of data. It is referred as knowledge extraction and data dredging. [2]

2. Data Mining Methods

A. Data Clustering

Clustering is a task whose goal is to determine a finite set of categories to describe a data set according to similarities among its objects.[3] Clustering techniques are classified as partitioned, hierarchical and non-exclusive i.e., over lapping methods. When we use the machine learning data set that is when we are in an unsupervised nature, the clustering technique is a challenging one and plays a vital role over here. Clustering is a meaningful and useful technique in data mining, in which it groups cluster of same objects using an automated tool. Clustering is based on similarity, In clustering analysis it is compulsory to compute the similarity or distance. So when data is too large or data arranged in a scattered manner it is quite difficult to properly arrange them in a group.[4] Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone.[5] Clustering is meant by dividing of large data sets into smaller ones and clustering all the homogeneous data into single. The following diagram depicts the stages in clustering. Clustering is also called as Data Segmentation [5], where the cluster is used for group of large homogeneous data.

B. Genetic Algorithm

They are class of randomized search procedure capable of adaptive search over a wide range of search space topology. They have ability to solve problem in parallel, so it is a powerful tool for data mining. Genetic algorithm search uses a set of solutions during each generation. [2]. The Genetic Algorithm consists of the following operations they are, randomly initialization of population, identifying the fitness values based on fitness function, and finally performing selection, mutation and cross over. The iterations are performed until the final best solution is reached. GA generates a set of population of solution by applying the survival of the fittest to produce better and better approximation solutions.[6]. Genetic algorithm reproduces the process of the evolution of survival of fittest. The Genetic algorithm is mainly used for the purpose of solving multi- objective optimization problems. In the Cross over operation, pair of individuals is chosen at random and an exchange process is carried out to obtain the results. In the selection (reproduction), an individual is selected randomly for its better performance. In the mutation operation, a individual is chosen randomly, so as a slight modification is done in the value with a difference of one.

C. Decision Tree

Decision trees are a way of representing a series of rules that lead to a class or value [7]. It is a hierarchical group of relationship organized into tree like structure. It determines which data and in which order should be collected to achieve effective decision that represent knowledge with minimal cost [2].

3. Related Work

In this section, some related work is done based on the data preprocessing and here the need for data is focused based on the data set. Data preprocessing has received far less attention that it deserves. Methods for user identification, session zing, and path completion are presented [8]. Here,

various data sets have been taken for preprocessing such that for a set of genetic algorithm the response of the data set has been tested under various classification algorithms. The classification algorithm which gives best accuracy for the data set is taken as input to the genetic algorithm for the next process of identifying the survival of the fittest. The authors employ the data warehousing technology as a preprocessing step to apply piecewise regression as a predictive data mining technique that fits a data model which will be used for prediction [8].

4. Data Mining Process Based On Preprocessing And Classification Of Data In Genetic Algorithm

In this we perform a data preprocess techniques and classify and cluster the data based on the attributes of data.

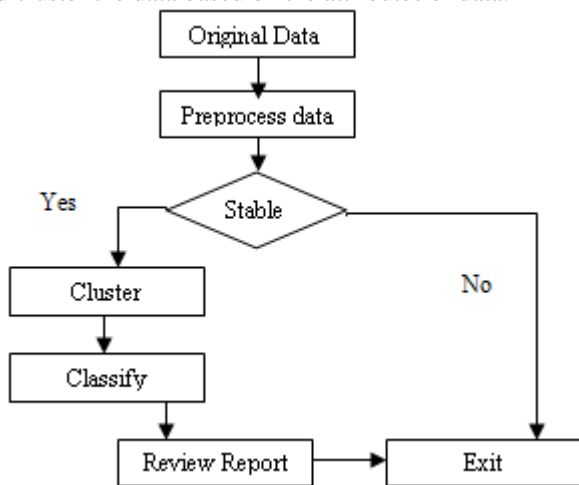


Figure1: Flow of Preprocessing a Original Data

A. Data Preparing

It is a process to make the data get prepared for various data mining methods. The various data mining methods includes

i) Data Preprocessing

The preprocessing is to clean the original data and extract the useful information from the data set.

ii) Data cleaning

Its routine work, it is to clean the data and fill the missing values and remove the noisy data from the data set [2]. Data cleaning means eliminate the irrelevant information from the original file [8].

iii) Data Option

It is used to select data arrange [2].

5. Experimental Setup And Result

The experiment was conducted based on the data set from UCI repository and IRIS. Here WEKA 3.7 tool is used to perform the preprocessing of data. Here the ZERO-R classification algorithm is compared with the dataset to find the performance of the IRIS-Fisher Dataset, Hospitality Data Set and Breast Cancer Dataset.

A. Data set Selected

The data set taken here for preprocessing is given below

Irish-Fisher Data set

The Iris data set is shown in the below table in which it length and width of sepals and petals. All the attributes in the data set are numeric. It is differentiated based on the number of runs and cross- validation.

Table1: Iris Data set Parameters after selection attributes

Parameters	Value	
No of data	150	
Cross validation folds	10	
Type	0,1	
Best fit	Type 1: 100%	
Average Merit	Type	0
	PW	0
	PL	0.003
Average Rank	Type1	1.3
	PW	1.7
	PL	3

In this data set we have type of 0 and 1 when the type attribute is removed from the table the mean and standard deviation gets increases. The below table depicts the mean and standard deviation before and after removing the type attribute.

Table2: Iris Dataset Before and after removing an attribute

Statistics	Value	
	Before removing Type attribute	After removing Type attribute
Minimum	0	1
Maximum	2	25
Mean	1	11.927
SD	0.819	7.569

B. Attribute Selection and Simple K- Means Cluster Analysis

To ensure that the patterns derived are as accurate as possible, it is essential to improve the quality of the data set in pre-processing stage [1]. The main goal of this method is to remove the redundancy in the data set and here the table shows the attribute selection based on the merit of best subset found and identified the squared errors based on the K-means. Here various types of the data set like IRIS, Hospitality, Diabetics, Breast cancer data set has been taken and the comparison is done with both the best fit and squared errors of Simple K-means.

Table 3: Selection attribute and K-means result.

Data Set	Best Fit Accuracy %	K-Means (Squared Errors)
Fishers Dataset (IRIS)	0.622	14.50
Diabetics	0.133	1777.0
Breast Cancer	0.744	260.42
Hospitality	0.941	856.95

C. Dataset Classification

Here to perform the classification among different data set the open source data mining tool Weka 3.7 has been used for the analysis. Four different data set has been given and the performance of each classification of each dataset has been analyzed and the best data set has been taken.

i) Data Set 1

The second dataset is a medium dataset which is a breast cancer with 9 attributes apart from class and it has 700 instances.

ii) Data Set 2

The Third dataset is similar to data set 2 but here, the diabetes dataset has chosen and it has 9 attributes and 769 instances.

iii) Data Set 3

The first dataset is a small dataset which is an IRIS- Fisher Dataset with 5 attributes and 150 instances.

iv) Classification used

Here 12 classification algorithms are used on the data set to perform a comparative study. The algorithm based on classification is categorized into various sections in WEKA here Bayesnet, NaiveBayes (from Bayes), SGD, SMO (from meta), Classificationviaregression, Filtered Classifier, Multiclass classifier (from rules), ZeroR, OneR (from trees), J48, Randomforest, Random tree (from trees) are taken for the analysis.

Table 4: Comparison of different Classifiers for Breast Cancer Data Set

Classification	Time Taken	Correctly classified Instances	Incorrectly Classified Instances	Kappa Instances	Mean Absolute Error	Root Mean Squared Error	Confusion Matrix
BayesNet	0.2	679 (97.13%)	20 (2.86%)	0.9374	0.0286	0.1614	442 16 4 237
NaiveBayes	0.3	671 (95.99 %)	28 (4.00%)	0.9127	0.0407	0.1991	436 22 6 235
SGD	0.15	676 (96.70%)	23 (3.29%)	0.9274	0.0329	0.1814	445 13 10 231
SMO	0.14	677 (96.85%)	22 (3.14%)	0.9306	0.0315	0.1774	445 13 9 232
Classification via Regression	0.75	670 (95.85%)	22 (4.14%)	0.9777	0.0666	0.1774	446 12 17 224
Filtered Classifier	0.06	658 (96.56%)	41 (5.86%)	0.8703	0.797	0.2299	437 21 20 221
Multiclass Classifier	0.12	675 (96.56%)	24 (3.43%)	0.924	0.0473	0.1646	446 12 12 229
ZeroR	0	458 (65.52%)	241 (34.47%)	0	0.452	0.4753	458 0 241 0
OneR	0.02	648 (92.70%)	51 (7.29%)	0.8348	0.073	0.2701	444 14 37 204
J48	0.08	656 (93.84%)	43 (6.15%)	0.8642	0.0735	0.2406	435 23 20 221
RandomForest	0.24	668 (95.56%)	31 (4.43%)	0.9017	0.0597	0.1795	443 15 16 225
RandomTree	0	658 (94.13%)	41 (5.86%)	0.8698	0.0587	0.2422	439 19 22 219

Table 5: Comparison of different Classifiers for Diabetes Data Set

Classification	Time Taken	Correctly classified Instances	Incorrectly Classified Instances	Kappa Instances	Mean Absolute Error	Root Mean Squared Error	Confusion Matrix
BayesNet	0.02	44 (57.42%)	327 (42.57%)	0.0797	0.4429	0.5102	115 153 174 326
NaiveBayes	0	463 (60.28%)	305 (39.71%)	0.0312	0.4455	0.4297	59 209 96 404
SGD	3.18	442 (57.55%)	326 (42.44%)	0.034	0.4245	0.6215	86 182 144 356
SMO	1.16	454 (59.11%)	314 (40.88%)	0.0332	0.4089	0.6394	71 197 117 383
Classification via Regression	6.07	435 (56.64%)	333 (43.35%)	0.0432	0.4346	0.5887	100 168 165 335
Filtered Classifier	0.03	500 (65.10%)	268 (34.89%)	0	0.4544	0.4766	0 268 0 500
Multiclass Classifier	13.21	419 (54.55%)	349 (45.44%)	0.0059	0.4523	0.663	97 171 178 322
ZeroR	0	500 (65.10%)	268 (34.89%)	0	0.4545	0.4766	0 268 0 500
OneR	0	363 (47.26%)	405 (52.73%)	0.0425	0.5273	0.7262	188 80 325 175
J48	0.09	500 (65.10%)	268 (34.89%)	0	0.4544	0.4766	0 268 0 500

Random Forest	0.35	457 (59.50%)	311 (40.49%)	-0.0212	0.4476	0.5314	39 229 82 418
Random Tree	0.02	451 (58.72%)	317 (41.27%)	-0.0186	0.4456	0.557	47 221 96 404

Table 6: Comparison of different Classifiers for IRIS - Fisher Data Set

<i>Classification</i>	<i>Time Taken</i>	<i>Correctly classified Instances</i>	<i>Incorrectly Classified Instances</i>	<i>Kappa Instances</i>	<i>Mean Absolute Error</i>	<i>Root Mean Squared Error</i>	<i>Confusion Matrix</i>
BayesNet	0.03	148 (98.66%)	2 (1.33%)	0.9446	0.0488	0.1484	128 0 2 20
NaiveBayes	0.02	128 (85.33%)	22 (14.66%)	0.5623	0.1442	0.3613	108 20 2 20
SGD	0.05	143 (95.33%)	7 (4.66%)	0.8024	0.0467	0.216	126 2 5 17
SMO	0.11	132 (88%)	18 (12%)	0.275	0.12	0.3464	128 0 18 4
Classification via Regression	0.08	150 (100%)	0	1	0.0641	0.1223	128 0 0 22
Filtered Classifier	0.02	150 (100%)	0	1	0	0	128 0 0 22
Multiclass Classifier	0.02	146 (97.33%)	4 (2.66%)	0.8935	0.0272	0.1619	126 2 2 20
ZeroR	0	128 (85.33%)	22 (14.66%)	0	0.2541	0.354	128 0 22 0
OneR	0	150 (100%)	0	1	0	0	128 0 0 22
J48	0.02	150 (100%)	0	1	0	0	128 0 0 22
Random Forest	0.02	148 (98.66%)	2 (1.33%)	0.9446	0.016	0.0879	128 0 2 20
Random Tree	0	149 (99.33%)	1 (0.66%)	0.9239	0.0067	0.0816	127 1 0 22

6. Results And Discussion

The results from the Table 1 and Table 2 have been analyzed very efficiently and the right classification algorithm is chosen. Here most of classifiers are not preformed the expected results. But the Filtered classifier, SMO and ZeroR classifier performs better on the above two data sets. The other values like kappa instances, Mean absolute error and Root mean squared error gives some more or less similarity results for the classifiers in both the data set. But the maximum accuracy for the data set is got in breast cancer data set.

7. Conclusion

In this Paper, we have analyzed and preprocessed the various data sets by removing the attributes and finding the accuracy of the data set. This paper mainly focused on identifying the best classification algorithm. Even though some of the classifier algorithms may not give correct accuracy for data set the processing may get differ for various kinds of data set and the number of instances.

References

- [1] K.C.Tan, E.J. Teoh, Q. Yu, K.C. Goh, "A Hybrid evolutionary algorithm for attribute selection in data mining", Expert systems with applications 36(2009) PP. 8616-8630
- [2] Ankita Agarwal, "Research of data mining based on neural network", Int. Journal of Comp. Science and Information Technology & security Vol2. No.5. Oct. 2012. PP. 893-898
- [3] Eduardo R. Hruschka, Ricardo, G.B. Campello, Alex A. Freitas, Andre C.P.L.F. de Carvalho, "A Survey of Evolutionary Algorithms for Clustering"
- [4] Ravindra Jain, "A Hybrid Clustering Algorithm for Data Mining", Computer Science & Information Technology (CS & IT), PP.387-393.
- [5] Chittu. V.N. Sumathi, "A Modified Genetic Algorithm Initializing K-Means Clustering", Global Journal of Computer Science and Technology Vol.11 Iss.2 Ver. 1.0 PP. 55-62.
- [6] Avneet Kaur, Mohsineen Wazir, Aparna Garg and AK Bakhshi "Designing of type -I polymeric superlattices using genetic and ant algorithms: A comparative study", Indian Journal of Chemistry. Vol 48 A Dec. 2009, PP.1652-1656.
- [7] M. Vijayalakshmi and Mulugu Narendhar, "A Novel approach for WEKA & Study on Data mining tools", Int. Journal of Engineering and Innovative Technology Vol 2. Iss.2 Aug.2012.
- [8] Suneetha K. and Dr.R. Krishnamoorthi, "Data preprocessing and Easy Access retrieval of data through data warehouse", proceedings of the world congress on engineering and computer science, 2009 Vol.1. Oct. 2009.