Attribute Selection using Machine Learning Technique

R. Samya
Department of Computer Science
Periyar University
Salem, India
samyacs93@gmail.com

***Abstract* - A Central problem in machine learning is to identify a representative set of attributes from which to construct a classification model for a particular task. Attribute selection is a well-known problem in the field of machine learning technique. It allows probabilistic classification and shows promising results on several benchmark problems. Attribute Selection is a task of choosing a small subset of features/attributes that is sufficient to predict the target labels well. Attribute Selection reduces the computational complexity of learning and prediction algorithms and saves computational the cost spent for measuring irrelevant features. This work addresses the problem of attribute selection for machine learning through Regression Analysis with different attribute selection methods like Forward Selection, Backward Elimination and Quick Reduct algorithm. The performance of the proposed approaches is studied based on the AIC measure. Further the classification accuracy of the proposed approach is analyzed by comparing it with the benchmark classification algorithm like K-Nearest Neighbor approach and Decision Tree approach. The result shows that accuracy of the classification algorithm without attribute selection. The proposed approach greatly improves the efficiency of the classification algorithms and the prediction accuracy is also remains satisfactory. So the Quick Reduct based attribute selection is better for machine learning techniques.**

Keyword: Attributes, Benchmark, Classification, Machine learning Techniques, Prediction

1. INTRODUCTION

Data Mining is a term coined to describe the process of sifting through large databases for interesting patterns and relationships. With the declining cost of disk storage, the size of many corporate point where analyzed by anything but parallelized machine learning algorithms running on special parallel hardware is infeasible (Padmavathi, 2012). Two approaches that enable standard machine learning algorithms to be applied to large databases are feature selection and sampling. Both reduce the size of the database – attribute selection by identifying the most salient attributes in the data; Sampling by identifying representative examples. This thesis focuses on the attributes selection – a process that can benefit learning algorithms regardless of the amount of data available to learn.

Attributes selection is the process of identifying and removing attributes from a training data set as much irrelevant and redundant attributes as possible. This reduces the dimensionality of the data. Many factors affect the success of machine learning on a given task. The representation and quality of the instance data is first and foremost. If there is much irrelevant and redundant information present or the data is noisy and unreliable, then knowledge discovery during the training phase is more difficult. In real world data, the representation of data often uses too many attributes, but only a few of them may be related to the target concept. (Amir Navot, 2006)

Generally attributes are characterized as

- Relevant: These are attributes which have an influence on the output and their role cannot be assumed by the rest.
- Irrelevant: Irrelevant attributes are defined as those attributes not having any influence on the output whose values are generated at random for each example.
- Redundant: A redundant exists, whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

A “feature” or “attribute” or “variable” refers to aspects of the data. Usually before collecting data attributes are specified or chosen. Attributes selection is a process commonly used in a subset of the attributes available from the data are selected for application of a learning algorithm (Heba Abusamra, 2013). The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining i.e. insignificant dimension. This is an important stage of preprocessing and is one of two ways to avoid the curse of dimensionality (the other is feature extraction). The main aim of attribute selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original attributes. In many real world problems attribute selection is a must due to the abundance of noisy, irrelevant attributes, etc., by removing these factors, learning techniques can be benefited. It evident that the attribute selection is an ideal approach for testing all the enumerations of attributes subsets, which is infeasible in most cases as it will result in 2n subset of n attributes (Hongton Sun, 2015). Attribute selections have been an active research area in pattern recognition, statistics and data mining communities.

2. LITERATURE SURVEY

In the literature survey, the existing research contributions are tabulated, there are lot of machine learning methods for different dataset were studied. The list of related research works are listed in the table 2.1.

Table 1. Literature Survey

Author	Dataset	Machine Learning Techniques	Description
Amir Novat et.al,	Cortical neural dataset	KNN, Regression	They proposed a non-linear, simple, yet effective feature subset selection method for regression and use it in analyzing cortical neural activity. Accuracy rate: 95%
RaghavendraB.K et.al,	Pima diabetes, Hepatitis, Heart-c, Heart-h, Statlog-Heart, Bupa Liver Disorders, Spect Test, Wiscosin Breast cancer, Haberman, postoperative patient Dataset	Logistic regression	The attribute selection algorithm using forward selection and backward elimination is applied on the dataset and the selected features from these algorithms are used to develop a predictive model for classification using logistic regression. Accuracy rate: 90%
Baranidharan Raman et.al,	Medical Dataset.	KNN, Decision Tree, Naïve Bayes	They invented SCRAP and LASER algorithm, and compare with these three enhancing learning for attribute selection and then resulted in better prediction accuracy rate is Naive Bayes learner. Accuracy rate: 80%
K.Anitha et.al,	Leukemia, Prostate cancer, Breast cancer and Lung cancer	Quick Reduct Algorithm	In their paper Quick Reduct Algorithm is used to reduce the number of genes from gene expression data.
Waked Yamany et.al,	Breast cancer, M-of-N, Exactly, Exactly2, Vote, Zoo, Lymphography, Led, Soybean-small, Lung and DNA.	Rough set, Flower pollination optimization.	They proposed an innovative use of an intelligent optimization method, namely the flower search algorithm (FSA), with rough sets for attribute reduction. Accuracy rate: 100%
Hongtan Sun	Exp_10567_11346.mat	KNN, SVM classifier	In this paper, they explore the methods of performing data separation, implement feature extraction, and implement feature selection using KNN and SVM classifier after the dimensionality

			reduction. Accuracy rate: 46.15%
D.Lavanya et.al,	Breast cancer, Breast cancer, Wisconsin (original), Breast cancer Wisconsin (Diagnostic)	Decision Tree	This paper analyzes the performance of Decision Tree classifier _CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast cancer dataset. Accuracy rate: 96.99%
A.Suresh	Rating system.com dataset.	Decision Tree, Naïve Bayes classifier	In this paper, the important center is on feature selection for sentiment analyzes and utilizing decision tree. Accuracy rate: 75%
JeevanadamJotheeswaran et.al,	IMDB dataset	Decision Tree, Manhattan Hierarchical cluster measure.	In this paper the main focus on feature selection for opinion mining using decision tree based feature selection. Accuracy rate: 75%
HebaAbusamra	Brain tumor dataset.	KNN, Random forest, SVM	This thesis aims on a comparative study of state -of-the-art feature selection methods, classification methods and the combination of them, based on gene expression data. Accuracy rate: 94.59%
Sofia visa et.al,	The Tomato Fruit dataset	CART, KNN	This paper introduces a new technique for feature selection and the method uses information a confusion matrix and evaluates one attribute at a time. Accuracy rate: 98%
Femina B et.al,	Hepatitis dataset	SVM, GA-SVM, KNN, RS-KNN	In this paper they proposed classifier with rough set based feature selection and k-Nearest Neighbor classifier. Accuracy rate: 84.52%

From the study, the different methods like SVM, KNN, Naïve Bayes classifier, Decision Tree etc., are used for attribute selection. These selected attributes are used for classification purpose to get high accuracy rate. It is evident that Decision Tree method has high classification accuracy.

In this paper, various analyses of the attribute selection methods using machine learning techniques in the literature are summarized briefly.

3. PROPOSED WORK

The aim of the attribute selection is to determine the attribute subset as small as possible. The mining performance is improved by reducing data dimensionality. Even though there exists a number of feature selection algorithms, it is an active research area in data mining, machine learning and pattern recognition communities. It selects the subset of original attributes, without any loss of useful information. The main problem focused in this paper is attribute selection and provide an overview of the existing methods that are available for handling several problems. There is lot of methods in Data Mining and refine the methods for comparative studies on attribute selection, in order to investigate which method perform best for specific tasks. Machine learning investigates how computers can learn (or improve their performance) based on data. A main theme of the machine learning is to automate the systems to train and to recognize complex patterns then make intelligent decisions based on patterns. For example, a typical machine learning model is to instruct the system, so that it can automatically recognize handwritten postal codes on mail, after learning from a set of examples. In many research works, the Machine learning techniques are related to Data Mining (Amir Navot, 2006).

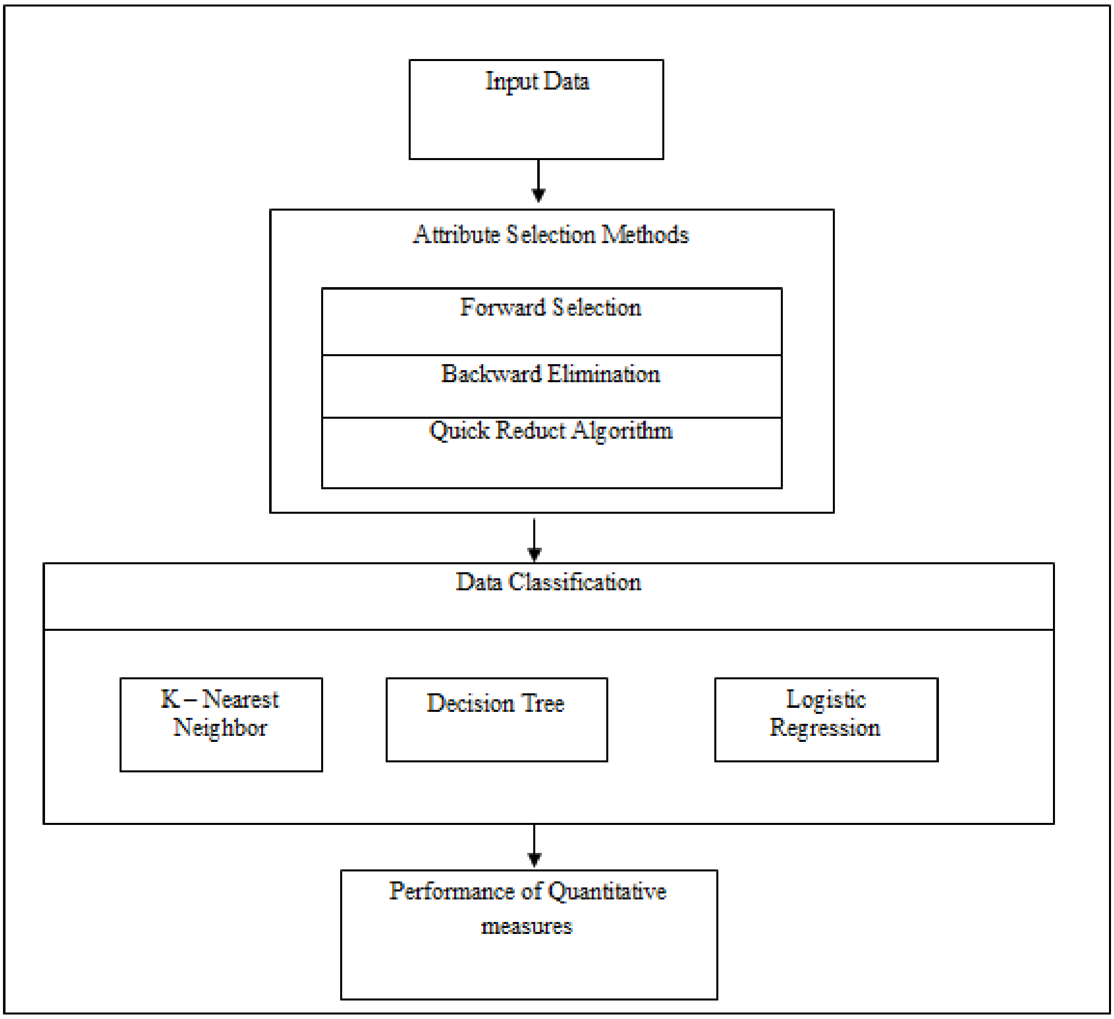

Figure 1. An Overview of the Proposed Approach

In the proposed work, Figure 1 describes the attributes selection for machine learning; first process the input data for attribute selection process using Forward Selection and Backward Elimination methods, Quick reduct algorithm. Then the supervised learning process using Logistic Regression, K – Nearest Neighbor and Decision Tree for classification. It is used to find the Quality measurement for the better accuracy rate. Using Decision Tree is the best for accuracy than K – Nearest Neighbor. The figure 3.1 shows an overview of attribute selection for machine learning technique in the proposed work

Algorithm 1: Attribute Selection for Machine Learning Technique

```
Step 1: ‘x’ is the given set of attributes
Step 2: For each set of attributes
        a. Apply MLR and find the set of optimal attributes given as $\mathbf{As_1}$.
        b. Apply MLR with Variable Selection Method given as $\mathbf{As_2}$.
        c. Apply Quick Reduct Algorithm given as $\mathbf{As_3}$ .
           End For
Step 3: For each set of selected attributes $\mathbf{i} = \{\mathbf{As_1}, \mathbf{As_2}, \mathbf{As_3}\}$
```

```
a. Find the Quality measures for accuracy rate.
b. Compare the accuracy rate for KNN, Decision Tree.
For each set of selected attributes in i
        i.    Apply LR
        ii.   Apply KNN
        iii.  Apply Decision Tree
End For
c. Select the best classifier.
End For
```

4. EXPERIMENTAL RESULT

A novel approach proposed in this paper is the Combination of attribute selection method with Logistic Regression; here the Quick Reduct Algorithm is used for attribute selection. The Comparative study of proposed approach with classification algorithms like KNN and Decision Tree is performed.

4.1. Dataset Description

Dataset describes the Diabetes and Weather data. The diabetes dataset is a supervised dataset with the parameters like “Number of times pregnant”, “plasma glucose concentration a 2 hours in an oral glucose tolerance test”, “Diastolic blood pressure”, “Triceps skin fold thickness”, “2-Hour serum insulin”, “Body Mass Index”, “Diabetes pedigree function”, “Age” and “Class Variable”. Diabetes dataset is taken from http://onlinecourse.science .psu.edu/stat857/node/45 and the weather dataset is a unsupervised data which is taken from www.indiawaterportal.org for the entire period of 1901 to 2002.This dataset briefly explain the weather data with the parameters like “Minimum Temperature”, ”Maximum Temperature”, ”Precipitation”, ”Rainfall”, ”Average Temperature”, “Cloud cover”, “Diurnal Temperature”, ”Ground frost frequency”, ”Potential Evapotranspiration”, ”Reference Crop Evapotranspiration”, “Vapor pressure”, “Wet day frequency”. The table 2 describes the Dataset in the proposed work

Table 2. Dataset Description

Dataset	No of observation	No of attributes
Weather Dataset	102	12
Diabetes Dataset	768	9

4.2. Attribute Selection

In the proposed work, the different attribute selection approaches defined for Logistic Regression, they are logistic regression with Backward Elimination, Logistic Regression with Forward Selection and Logistic Regression with Quick reduct Algorithm, and the results elaborates the Intercept, Coefficient and AIC measures for the Diabetes dataset. Using the dimensionality reduction, the number of attributes refined based on target attribute. The table 3 describes the comparison of different approaches for attribute selection model (LR, Backward elimination for LR, Forward selection for LR, Quick Reduct Algorithm) for supervised dataset.

Table 3. Logistic Regression with Attributes Selection for Diabetes Dataset

Types of Model	Intercept	Coefficient	AIC	No of Attributes	Selected Attributes
Logistic Regression without Attribute Selection	-8.4046964	0.1231823 0.0351637 -0.0132955 0.0006190 -0.0011917 0.0897010 0.9451797 0.148690	781.45	9	All Attributes

Logistic Regression with Backward Elimination	-8.0676572	0.1176812 0.0349682 -0.0008293 -0.0008293 0.0928813 0.0159594	747.62	9	1,2,3,5,6,8
Logistic Regression with Forward Selection	-5.6574046	0.1089282 0.0370769 -0.0049367 -0.0001639 -0.0001639	794.81	9	1,2,3,5,8
Logistic Regression with Quick Reduct algorithm	-6.17842	0.130232 0.036407 1.001286	739.27	9	1,2,7

The results show that AIC measure of the three proposed approaches is minimal than the LR without Attribute Selection. Similarly, the LR with Quick Reduct approach performs well than the other proposed approaches, which is evident from the AIC value. sThe classification accuracy of three different proposed approaches are analyzed using various quality measures like Accuracy, Sensitivity, Specificity, Precision, Mean Absolute Error and F-measure. They are discussed in the table 4. The Accuracy rate of the LR with Quick Reduct outperforms the other approaches.

Table 4. Performance of Logistic Regression with Attribute Selection

Types of Model	**Confusion Matrix**			**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Mean Absolute Error**	**F - measure**
Without AS		0	1	0.7825	0.89	0.582	0.89	1.67	0.89
	0	445	55						
	1	112	156						
Backward Elimination with LR		0	1	0.2317	0.118	0.4216	0.2757	1.67	0.3937
	0	59	441						
	1	155	113						
Forward Selection with LR		0	1	0.7513	0.878	0.5149	0.7715	1.91	1.3547
	0	439	61						
	1	130	138						
Quick Reduct with LR		0	1	0.8578	0.884	0.5223	0.7754	1.86	0.8257
	0	442	58						
	1	128	140						

The results of KNN and Decision Tree with three attribute selection methods are tabulated in the table 5 and 6. The performance measures of KNN show that the attribute selection with Quick Reduct algorithm show better performance. Similarly for Decision Tree also the Quick reduct based approach performs better than other approaches.

Table 5. Performance of KNN with Attribute Selection

Types of Model	**Confusion Matrix**			**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Mean Absolute Error**	**F-measure**
Without AS		0	1	0.648	0.9166	0.5086	0.8020	0.2119	1.5022
	0	231	24						
	1	57	59						
Backward Elimination with LR		0	1	0.7153	0.9043	0.5258	0.8171	0.2031	0.8826
	0	226	24						
	1	55	61						

Forward Selection with LR	 0 1	0 228 5 3	1 24 60	0.7653	0.9047	0.5258	0.8172	0.2031	0.8526
Quick Reduct with LR	 0 1	0 224 56	1 28 60	0.8717	0.8888	0.5172	0.80	0.2282	0.8380

The graphical representation of the Accuracy rate of the classification process with attribute selection approaches are represented in the figure. It shows that the LR with Quick Reduct performs alike the benchmark approaches with attribute selection. This approach is further applied to another dataset to study its performance on unsupervised dataset. As the Logistic regression is meant for supervised data, the Multiple Linear Regression (MLR) is chosen for this purpose.

Table 6. Performance of Decision Tree with Attribute Selection

Types of model	**Confusion Matrix**			**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**Mean Absolute Error**	**F Measure**
Without AS	 0 1	0 57.8 8.3	1 7.2 26.5	0.7437	0.88799	0.76119	0.874015	0.15625	0.8809
Backward Elimination with LR	 0 1	0 59.6 13.9	1 5.4 20.9	0.8059	0.91601	0.6007	133.202	0.194	1.195
Forward Selection with LR	 0 1	0 50.9 12.2	1 14.9 22.6	0.7356	0.7820	0.6492	0.8061	0.264	0.7938
Quick Reduct with LR	 0 1	0 60.6 17.5	1 3.42 17.3	0.8599	0.932	0.4962	0.7753	0.22005	1.819

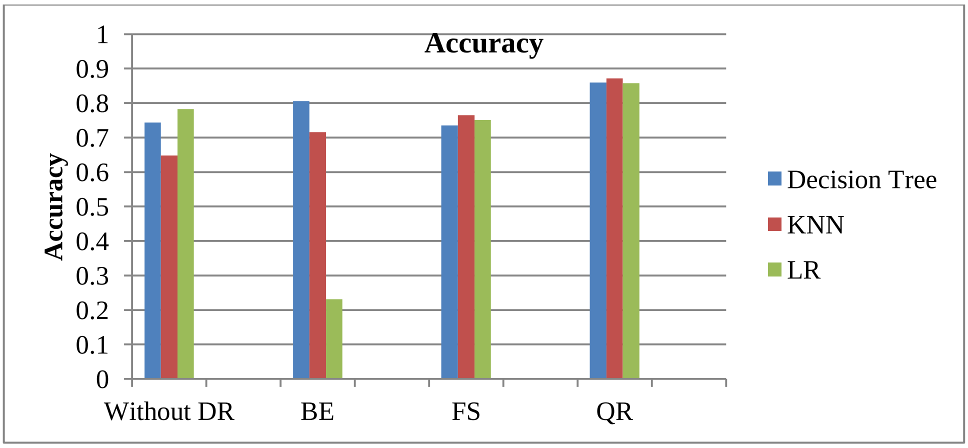

Figure 2. Accuracy rate of Proposed Approaches with Attribute Selection

The results of the MLR with attribute selection (Backward elimination, Forward selection and Quick Reduct) on the weather dataset are tabulated in the table 6. The different models derived for weather dataset is given in the table 7.

Table 7. Multiple Linear Regressions with Attribute Selection for Weather Dataset

Types of Model	Intercept	Coefficients	No of attributes	Selected attributes
Multiple Linear Regression without Attribute Selection	1.745e-13	2.089e-14 0.000e+00 1.000e+00 -1.75e-14 -6.350e-13 -1.378e-14 -6.270e-13 -3.073e-14 -2.371e-14 -9.579e-16 -3.524e-16	12	All Attributes
Multiple Linear Regression with Backward Elimination	1.359e-13	-3.189e-14 1.000e+00 3.402e-14 -9.726e-16 -5.053e-14	12	1,3,5,6,9.
Multiple Linear Regression with Forward Selection	4.503e-14	1.000e+00	12	3.
Multiple Linear Regression with Quick Reduct algorithm	-125.796	5.956	12	2

5. CONCLUSION AND FUTURE WORK

The different combinations of attribute selection approaches for machine learning Techniques are proposed to provide a computational solution for various classification problems with large data. Analyzing from different combination of these approaches, the results show the best attribute selection method which can handle a classification problem with irrelevant attributes. The analysis based on the attribute selection over regression analysis shows that the Regression with Quick Reduct performs well. Similarly the analysis based on the classification accuracy over the benchmark classification approaches like K-Nearest Neighbor and Decision Tree shows that the regression with QR performs alike those approaches. The future work is aimed at a better understanding of attribute selection for the machine learning techniques through combination of some other attribute selection methods.

REFERENCE

A. Suresh , & C.R Bharathi. (2016). Sentiment Classification using Decision Tree Based Feature Selection . International Science Press, 419-425.

Amir Navot. (2006). On the Role of Feature Selection in Machine Learning.

Amir Navot, L. N. (2006). Nearest Neighbour Based Feature Selection for Regression and its Application to Neural Activity.

Amir Navot, Lavishpigelman, Naftali Tishby, & Eilon Vaadia. (2005). Nearest Neighbor based Feature Selection for Regression and its Application to Neural Activity.

Andreas G.K.Janeek, Wilfried N. Gansterer, Micheal A.Damel, & Gerhand F.Ecker. (2008). On the Relationship Between Feature Selection and Classification Accuracy. JMLR: Workshop and Conference Preceedings, 90-105.

Arif Salekin, & John Stankovic. (2006). Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes.

Baranidharan Raman, & Thomas R.Ioerger. (2003). Enchancing Learning using Feature and Example Selection. Journal of Machine Learning Research.

D. Lavanya. (2011). Analysis of feature selection with classification: Breast Cancer Datasets. Indian Journal of Computer Science and Engineering, 2.

Daniel T.Larose, & Chantal D. Larose. (2015). Data Mining and Predictive Analytics. John Wiely&Sons.

Darrel R.Massive, & Mark A.Rose. (1997). Predicting Daily Maximum Temperature using Linear Regression and Eta Geopotential Thickness Forecasts.

Femina B, & Anto S. (2015). Disease Diagnosis using rough set based feature selection and Knearest neighbor classifier. International Journal of Multidisciplinary Research and Development, 2(4), 664-668.

Govind P Gupta, & Manish Kalariya. (2016). A Framework for Fast and Efficient Cyber Security Network Intrusive Detection using Apache Spark. International Conference on Advances in Computing & Communication, 824-831.

Gulden Kaya Uyanik, & Nese Guler. (2013). A Study on Multiple linear Regression Analysis. International Conference on New Horizons in Education, 234-240.

Heba Abusamra. (2013). A Comparitive Study of Feature Selection and Classification methods for Gene Expression Dta.

Hongton Sun. (2015). K-Nearest Neighbor and SVM classifier with feature extraction and feature selection .

Intan Martina MdGhani, & Sabri Ahmad. (2010). Stepwise Multiple Regression Method to Forecast Fish Landing. International Conference on Mathematics Education Research, 549-554.

Isabelle Guyon, & Andre Elisseeff. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 1157-1182.

Jeevanandam Jotheeswaran. (2013). Opinion Mining using Decision Tree Based feature selection through Manhattan Hierarchial cluster Measure. Journal of Theoritical and Applied Information Technology, 58.

Jennifer G.Dy, & Carla E.Brodley. (2004). Feature Selection for Unsupervised Learning. Journal of Machine Learning Research, 845-889.

Jiawei Han , Micheline Kamber, & Jian Pei. (2012). Data Mining Concepts and Techniques (3rd ed.).

Jiye Liang, Feng Wang, Chuangyin Dang, & Yuhua Qian. (2012). An efficient rough feature selection with a multi- granulation view . International Journal of Approximation Reasoning, 912-926.

K. Ming Leung. (2007). K-Nearest Neighbor Algorithm for Classification.

k.Anitha. (2014). Feature Selection by Rough Quick Reduct Algorithm. International Journal of Innovative Research in Science, Engineering and Technology, 2(8), 2319-8753.

Kenjikira, & Larry A.Rendell. (2013). A Practical Approach to Feature Selection.

LI Dan, & Xiaofa Shi. (2009). Estimates of Pedestrian Crossing delay based on Multple linear regression and Application. COTA International Conference of Tranportation Professionals.

M. Akhil jabbar, B.L Deekshatulu, & Priti Chandra. (2013). Classification of Heart Disease using K- Nearest Neighbor and Genetic Algorithm. International Conference on Computational Intelligence: Modeling Techniques and Applications, 85-94.

M. Karagiannopoulos, D.Anyfantis, S.B Kotsiantis, & P.E. Pintelas. (2006). Feature Selection for Regression Problems.

Meigfeng Sun, Jangtao Chen, Yun Zhang, & Shangzhe Shi. (2012). A new method of Feature Selection for Flow Classification. 1729-1736.

Min Li, Shaobo Deng, & Jianping Fan. (2013). Quick Attribute Reduction Based on Approximation Dependency Degree. Journal of Computers, 8.

Mukesh Kumar, & Nitish Kumar Rath. (2015). Feature Selection and Classification of Microarray Data using Mapreduce based ANOVA and K-nearest Neighbor. international Multi-Conference on Information Processing .

Osiris Villacampa. (2015). Feature Selection and Classification Methods for Decision Making: A Comparitive Analysis.

P Kalyani. (2011). A new implementation of Attribute reduction using Quick Relative Reduct Algorithm. International Journal of Internet Computing, 1(1).

Padmavathi. (2012). Logistic regression in feature selection in Data mining. International Journal of Scientific & Engineering Research, 3(8), 2229-5518.

Paras, & Sanjay Mathur. (2016). A Simple Weather Forecasting Model using Mathematical Regression. Indian Research Journal of Extension Education .

Phivos Mylonas, Manolis Wallace, & Stefanos Kollias. (2004). Using K- Nearest Neighbor and Feature Selection as an Improvement to Hierarchial Clustering . 191-200.

Radek Silhavy, Petr Silhavy, & Zdenka Prokopova. (2017). Analysis and Selection of a regression model using a stepwise approach. The Journal of Systems and Software, 0164-1212.

Raghvendra B.K. (2011). Evaluation of Logistic Regression Model with Feature Selection Methods on Medical Dataset. International Journal of Advanced Engineering Technology, 2, 228-223.

Rashmi Agrawal. (2016). A Modified K- Nearest Neighbor Algorithm using Feature Optimization. International Journal of Engineering and Technology, 8, 0975-4024.

Sabita Mahapatra, SreeKumar S.S, & Mahapatra. (2010). Attribute Selection in Marketing : A rough set approach. Indian Institute of Management Bangalore, 0970-3896.

Sanyam Gupta, Indhumathy K, & Govind Singhal. (2016). Weather Prediction using Normal Equation Method and Linear Regression Techniques. International Journal of Computer Science and Information Technologies, 7, 1490-1493.

Sofia visa, Brian Ramsay, Anca Ralescee, & Esther van. (2014). Confusion Matrix based Feature Selection.

Stephan Dreiseitt, & Lucila Ohno Machado. (2003). Logistic regresion and artificial neural network classifiation models: a methodology review. Journal of Biomedical Informatics, 352-359.

Waleed Yamany, Eid Emary, Aboul Ella Hassanien , Gerald Scharfer, & Shao Ying Zhu. (2016). An Innovative Approach for Attribute Reduction using Rough Sets and Flower Pollination Optimisation. International Conference on Knowledge Based and Intelligent Information and Engineering Systems , 403-409.

Xiaobo Zhou, Kuanga Yu Liu, Stephen, & T.C Wong. (2004). Cancer Classification and Prediction using logistic regression with Bayesian gene selection. Journal of Biomedical Informatics, 249-259.