IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-5 E-ISSN: 2347-2693

Enhancing Classification Performance with Subset and Feature Selection Schemes

Aniket G. Meshram[1*], K. Rajeswari[2] and V. Vaithiyanathan[3]

[1*,2]*Department of Computer Engineering, PCCOE, University of Pune, India*
[3]*Associate Dean, Research, CTS Chair Professor,SASTRA University,* Tanjore, India.

www.ijcseonline.org

Received: Apr/22/2015 Revised: May/02//2015 Accepted: May/16/2015 Published: May/30/ 2015

***Abstract*—**Classification is one of the important steps in data mining for categorizing huge amount of data. Different classifiers are in use today for the classification of large data sets. Some classifiers have shown better performance than the others. Though these classifiers have proven better than others, there is still a chance for improvement in the classification process. This improvement can be considered in terms of selecting the important or rather the features that affect most in the classification process. Thus, here we focus on using a set of five attribute selection techniques applied on two classifiers to show how attribute selection affects classification performance. We compare and discuss two well-known classification schemes – MultiLayer Perceptron and Simple Logistics based on the application of these five attribute selection techniques. The aim of this study is to demonstrate and understand the behaviour of these classifiers once subjected to attribute selection schemes.

***Keywords*—**MultiLayer Perceptron, Simple Logistics, Attribute Selection Techniques, Subset Evaluation, Weka.

I. INTRODUCTION

Data around us, data on our systems, data on the servers, they are all growing at an exponential rate. Data mining as we very well know by now, today is one of the important process which helps in processing and understanding these huge data and datasets to make vital decisions. In data mining, classification is one of the most popular steps for categorizing data. There exist many classification schemes with their merits and demerits. Two widely used classification schemes are adopted here to understand the effect of attributes selection mechanisms on them. Five different datasets are used which are not related to each other, to demonstrate a clear difference in characteristics of selecting the attributes of these five datasets.

Feature selection is very useful in reducing the dimensionality of attributes in the datasets [3]. Any attributes that are not relevant, or noisy or is repetitive is removed. We use feature selection techniques for both classifications, and compare each other based on different selections. Five subset attribute selection technique have been used here for feature selection. In some cases selecting attributes may degrade the classification performance. However, it can be shown that in many cases, the use of feature selection contributes to the improvement of the efficiency of classification.

II. CLASSIFICATION SCHEMES

A. MultiLayer Perceptron (MLP)

It is one of the artificial neural network model based on the workings on neurons in a human brain. It is a kind of model where the flow of data is in one direction from the input to the output, which is the basic feedforward mechanism. This network helps to create a model that correctly maps the input data to the output data using the historical data, so that one can get a view of the output when the desired output is unknown. This type of network uses the backpropagation learning algorithm for training its model. In the training process, the input data is repeatedly fed to the neural network. The output of the neural network is then compared to the desired output, which results into some error. This error is then given back (backpropagated) to the network model for further processing. The adjusted weights(as a result of such errors) help to decrease further errors to get close to the desired output.

Basically, the neural network model (Fig.1) consists of three layers, viz. the Input Layer, the Hidden Layer, and the Output Layer, which applies the feed forward mechanism for processing.

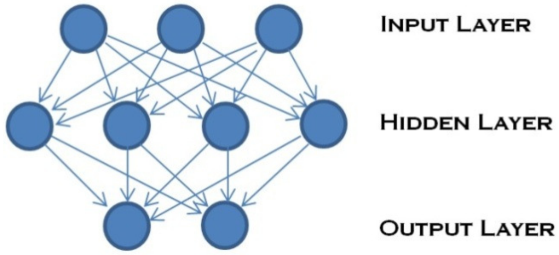

Fig. 1.A Neural Network Model depicting a three layer process.

B. Simple Logistics

Simple logistic regression is used to explore associations between one (dichotomous) outcome and one (continuous, ordinal, or categorical) exposure variable. Simple logistic regressionis used in situations where we have two variables, a nominal variable with two values (for example, left/right, dead/alive, male/female, etc.) and a measurement variable [9]. The nominal variable is the dependent variable, while measurement variable is the independent one. One canconsider comparing Simple logistic regression with linear regression which are analogous, however, Simple logistic regression finds the equation that best predicts the value of the Y variable for each given value of the X variable. Consider the example of people who have heart attacks. Here the nominal variable values would be "did have a heart attack" vs. "didn't have a heart attack". The Y variable would then be the probability of having a heart attack, which is used in logistic regression.

III. FEATURE SELECTION SCHEMES

A. CFS Subset Evaluation with RankSearch

The central hypothesis of the Correlation-based Feature Selection (CFS) is that good feature sets contain features that are highly correlated with the class, yet uncorrelated with each other. A subset evaluation heuristic is considered at the heart of the algorithm which takes into account the usefulness of individual features for predicting the class along with the level of inter-correlation among them. The heuristic assigns high scores to subsets containing attributes that are highly correlated with the class and have low inter-correlation with each other.

$$Merit(S) = \frac{k\,\overline{R_{cf}}}{\sqrt{k+k(k-1)\overline{R_{ff}}}} \qquad (1)$$

Where,
Merit(S), is the heuristic "merit" of a feature subset S containing k features,
$\overline{R_{cf}}$, is the average feature-class correlation, and,
$\overline{R_{ff}}$, is the average feature-feature inter-correlation.

B. Classifier Subset Evaluation with RankSearch

It evaluates attribute subsets on training data or a separate hold out testing set. It also uses a classifier to estimate the 'merit' of a set of attributes.

C. Consistency Subset Evaluation with RankSearch

In consistency-based feature selection, consistency measures are used to evaluate relevance of feature subsets. A consistency measure can be defined as a metric to measure the distance of a feature subset from the consistent state. A feature set {F1 ,...,Fn} is said to be consistent, when,

$$\Pr(C = c \mid F1 = f1, \ldots, Fn = fn) = 0 \text{ or } 1$$

holds forall $c\ ,f1\,,\ldots,fn$. When a feature subset is consistent, the inconsistency value is 0, and as an inconsistent feature subset approaches the consistent state, the measure approaches 0.

D. InfoGain Attribute Evaluation with Ranker

Entropy is commonly used in the information theory measure, since it characterizes the purity of an arbitrary collection of examples. It is the foundation of the Information Gain (IG) attribute ranking methods. The entropy measure is considered as a measure of system's unpredictability. The entropy of Y is given as,

$$H(Y) = -\sum_{y\in Y} p(y) \log_2(p(y)) \qquad (2)$$

where, $p(y)$is the marginal probability density function for the random variable *Y*.

The entropy of *Y* after observing *X* is:

$$H(Y/X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p\left(\frac{y}{x}\right) \log_2\left(p\left(\frac{y}{x}\right)\right)(3)$$

where, $p(y/x)$, is the conditional probability of *y* given *x*. We can define a measure reflecting additional information about *Y* provided by *X* that represents the amount by which the entropy of *Y* decreases. This measure is known as IG. It is given by,

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y)(4)$$

E. ChiSquared Attribute Evaluation with Ranker

Feature Selection via chi square (X^2) test is another, very commonly used method. The X^2 method evaluates features individually by measuring their chi-squared statistic with respect to the classes.

The Chi-square test is also called as Goodness of Fit Test or Independence Test, which calculates the test statistics of the sum of squares of differences between observed and expected frequencies. This can be given mathematically as,

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{-}OF_i - EF_i)^2}{EF_i} \qquad (5)$$

where,
χ^2 is the Cumulative test statistics,
OF_i is the Observed Frequency,
EF_i is the Expected (or Theoretical) Frequency,
n is the number of table values.

IV. LITERATURE SURVEY

Here we choose two classification schemes as mentioned earlier, MultiLayer Perceptron, and Simple Logistics, with five attribute selection schemes. The study conducted by Mark Hall, et. al [3], have explained the use of attribute selection in a quiet vivid way. They suggested the use of Ranker search to be an important factor for minimizing noisy data in the classification process, which may lead to an improvement in the Correctly Classified Instances. Malay Mitra and R. K. Samanta [1], have shown, how the use of attribute selection helps in a Cardiac Arrhythmia Classification. JasminaNovaković, et al [5], have also showed in their study, how the use of Ranking procedure can help improve classification evaluation. As the dimensionality of a domain expands, the number of features 'N' increases. Feature selection has thus proven in both theory and practice to be effective in enhancing efficiency of learning, increasing prediction and reducing complexity of results [2]. There are various techniques available for feature reduction, notable among them are techniques like Information Gain [6,7], ChiSquared test [8], and Subset evaluators. These techniques consider various parameters for improving classification performance, by eliminating redundant and inconsistent features. Reducing the dimensions of data can help to select and understand which features contribute more to the classification and which ones create redundancy. Some attribute may not contribute at all in the classification process.These inconsistent attributes can be removed to focus only on those that are important.

The tool that we have used for performing these evaluations is best suited for various types of operations like clustering, feature selection and classification as well. Understanding how WEKA helps in classification is important when it comes to extracting features that are important since these extracted 'important' features will be subjected to the classification [4]. One can also find the applications of using the Neural Network and Logistics schemes for classification in the electroencephalograph (EEG) signals for prediction of epileptic disorders [12]. This is one of reason why MLP and SimpleLogistics were chosen, since their use in many fields has been vital for decision making and prediction.

V. UCI DATASET REPOSITORY

Five Dataset were chosen for processing which can be described as below:

1. Balance-Scale Dataset: Consisting of 625 instances with number of attributes equal to 5.
2. Breast Cancer Dataset: Consisting of 286 instances with number of attributes equal to 10.
3. Credit – A Dataset: Consisting of 690 instances with number of attributes equal to 16.
4. Heart Risk Dataset: Consisting of 712 instances with number of attributes equal to 16.
5. Hepatitis Dataset: Consisting of 155 instances with number of attributes equal to 20.

These dataset were selected randomly considering some datasets with large number of instances while considering others with more attributes [10].

VI. WEKA

Waikato Environment for Knowledge Analysis (WEKA) is a comprehensive suite of most of the data mining and machine algorithms for decision making[11]. One of the major advantages of using WEKA is that it contains implementations of these algorithms with a simple GUI one can interact with and deduce conclusions analyzing huge data sets. The main features that WEKA includes is, Data-Preprocessing, Classification, Clustering, Attribute Selection, etc. which makes it a pretty good data mining tool. It also includes support for association rule mining, comparing classifiers and data set generations. Another reason for using WEKA was that in this software we get all the algorithms we need to implement packed together.

One of biggest advantage of using WEKA is that it is an Open Source yet powerful application, which is available freely for processing large datasets. Another advantage is that researchers can easily implement new algorithms for data manipulation and scheme evaluation, caring less of having to be concerned with supporting infrastructure.

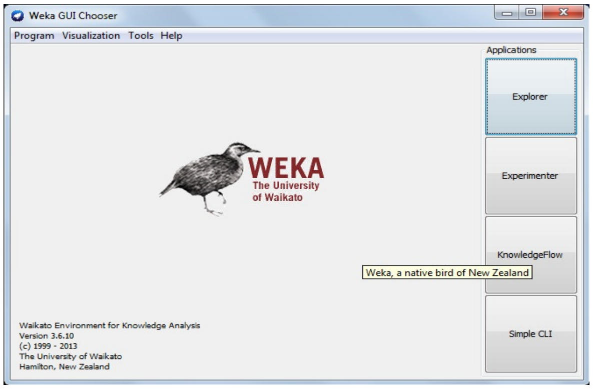
Weka GUI Chooser
Program Visualization Tools Help
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI
WEKA
The University of Waikato
Weka, a native bird of New Zealand
Waikato Environment for Knowledge Analysis
Version 3.6.10
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Fig. 2. WEKA Main Window (Snapshot)

VII. EXPERIMENTATION PROCESS

The study involves experimentation on five datasets taken from the UCI Dataset Repository to analyse the behaviour of classification, when applied by feature selection schemes. At first, the classification algorithms were applied on the datasets and result were noted for later comparisons. Then feature selection techniques were applied one-by-one on the datasets and the subjected to the respective classification. The process of selecting features included a "reverse approach", where the ranked attributes were considered for evaluation.

After applying each attribute selection scheme to the respective datasets, the attributes which ranked low were removed from the evaluation process, while considering the remaining attributes for further processing. Thus the method that we applied can be said as a feature removal process with respect to features that are inconsistent or unimportant (those that contribute less in the classification process).

The overall process can be summarized as below:

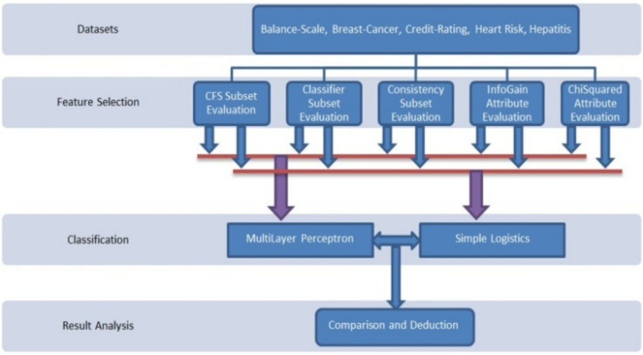
Datasets
Balance-Scale, Breast-Cancer, Credit-Rating, Heart Risk, Hepatitis
Feature Selection
CFS Subset Evaluation
Classifier Subset Evaluation
Consistency Subset Evaluation
InfoGain Attribute Evaluation
ChiSquared Attribute Evaluation
Classification
MultiLayer Perceptron
Simple Logistics
Result Analysis
Comparison and Deduction

Fig. 3. Architecture describing the overall Classification process with attribute selection

VIII. Result Analysis

Initially the approach was to test the behaviour of MLP on different datasets. When different feature selection schemes were applied on datasets like Breast-Cancer, the classification carried out after feature selection resulted in an increase by nearly 5.44%. Further, the feature selection schemes also showed a significant increase when applied to other datasets as well (Fig. 4). Though the improvement may seem small but this can further be enhanced by removing the inconsistent attributes by considering the attributes whose presence or absence has little effect on the classification process.

While MLP shows a significant improvement over the Heart-Risk Dataset, one cannot fail to notice the improvement of SimpleLogistics over MLP considering the Hepatitis Dataset. However here we see that if the dataset consists of only few attributes, there is a higher probability that all or nearly all the attribute contribute to the classification process. Hence feature selection may not be applied in this scenario since nearly all the features play a role in the classification process and hence may be important. Our results thus, support this theory for the Balance-Scale dataset where the correctly classified instances are reduces when subjected to different feature selection schemes.

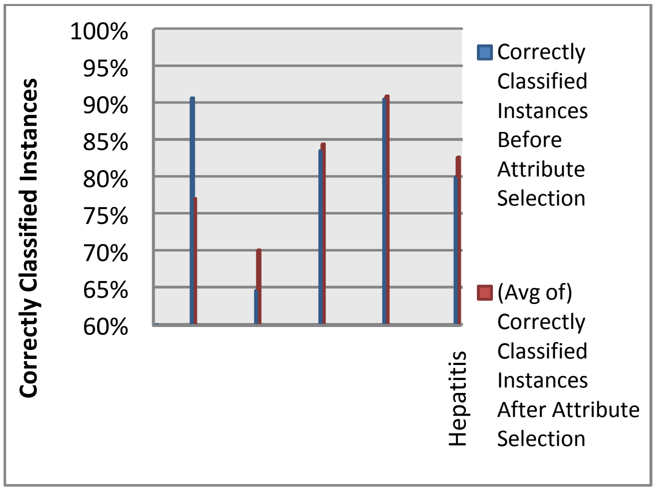

Fig. 4. Graph showing results of applying feature selection on MLP.

Simple Logistics also shows significant improvement when preprocessed with feature selection schemes. The graph below (Fig. 5) depicts the average instances that were correctly classified when feature selection was applied comparing it the correctly classified instances without attribute selection on SimpleLogistics Classification scheme.

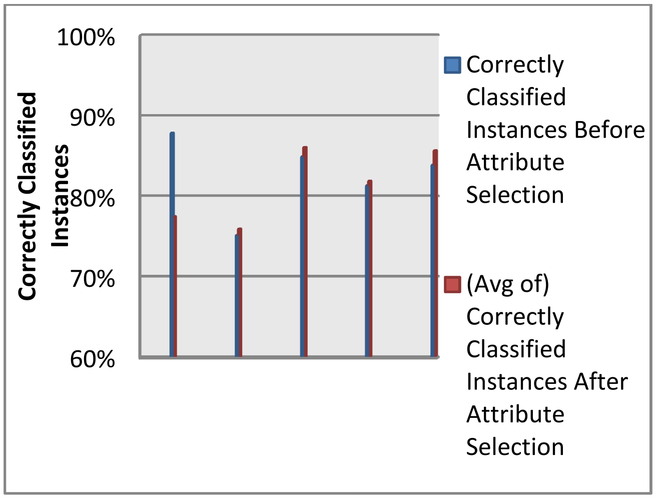

Fig. 5. Graph showing results of applying feature selection on Simple Logistics.

IX. Conclusion

By observing the behaviour of attribute selection schemes on MLP, we see that Simple Logistics has shown better performance in the Hepatitis dataset as compared to the MLP classification. The results have thus shown that selecting appropriate feature from the datasets is one of the important tasks one needs to consider for enhancing the performance of large datasets that required decision making.

This evaluation can thus help make a better decision support for understanding and analysing complex systems with complex data in their respective area.

REFERENCES

[1] Malay Mitra and R. K. Samanta, "Cardiac Arrhythmia Classification Using Neural Networks with Selected Features", International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) (2013).

[2] K. Rajeswari, RohitGarud and V. Vaithiyanathan, "Improving Efficiency of Classification using PCA and Apriori based Attribute Selection Technique", Research Journal of Applied Sciences, Engg. and Technology 6(24): 4681-4684, (2013).

[3] Mark Hall and Geoffrey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", Department of Comp. Science, The University of Waikato 2002.

[4] Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, April (2013).

[5] JasminaNovaković, PericaStrbac, DusanBulatović, "Toward Optimal Feature Selection Using Ranking Methods and Classification Algorithms", Yugoslav Journal of Operations Research 21 (2011).

[6] B.Azhagusundari, Antony SelvadossThanamani, "Feature Selection based on Information Gain", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-2, Jan (2013).

[7] JasminaNovakovic, "Using Information Gain Attribute Evaluation to Classify Sonar Targets", 17th Telecommunications forum TELFOR (2009).

[8] PhayungMeesad, PudsadeeBoonrawd and VatineeNuipian, "A Chi-Square-Test for Word Importance Differentiation in Text Classification", International Conference on Information and Electronics Engineering IPCSIT vol.6 (2011).

[9] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression", 10.1093/bioinformatics/btg308 pg. 2246–2253 Vol. 19 no. 17 (2003).

[10] Datasets available at the UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets.html.

[11] WEKA,an Open Source software freely available at http://www.cs.waikato.ac.nz/.

[12] AbdulhamitSubasi, Ergun Ercelebi, "Classification of EEG signals using neural network and logisticregression", Computer Methods and Programs in Biomedicine 78, 87—99 (2005).