IJITKM Volume 7 • Number 2 • Jan– June 2014 pp. 32-36 (ISSN 0973-4414)

Impact of attribute selection on the accuracy of Multilayer Perceptron

Niket Kumar Choudhary[1], Yogita Shinde[2], Rajeswari Kannan[3], Vaithiyanathan Venkatraman[4]

[1,2]Dept. of Computer Engineering, Pimpri-Chinchwad College of Engineering, Pune University, Pune, Maharashtra,
[3]Associate Professor, Dept. of Computer Engineering, PC College of Engg, Pune University, Maharashtra, India.
[4]Associate Dean Research, SASTRA University, Thanjavur, Tamilnadu, India.
niket_choudhary@ymail.com, shindeyd2@gmail.com, raji.pccoe@gmail.com, vvn@it.sastra.edu.

Abstract: Data classification is an essential operation in any data analysis process. The purpose of classification techniques is to accurately classify the data. It has been found that whenever attribute selection techniques are applied before classification, accuracy is improved significantly. Here we present the performance analysis of Multi-Layer Perceptron classification technique when applied before and after six mostly used attribute selection techniques. The experiment showed significant improvements. The experiments are conducted using WEKA 3.6.10 tool.

Keywords: Data Mining, Classification, Multi-Layer Perceptron, Attribute Selection, WEKA.

1. Introduction

In present world abundant data is collected in large volume. This large volume of data contains valuable information, which is most of the time hidden. Data Mining is one such domain used to extract valuable information from large data sets. In Data Mining, Classification is one such technique which is used to assign a class label to a set of unclassified instances. It predicts the target class for each instance in the data set. Classification is divided into two categories- supervised and unsupervised. In supervised classification the set of possible classes is known in advance whereas in unsupervised classification the set of classes is not known in advance, a name is predicted to be assigned after classification. Many times a data set may contain irrelevant or redundant attributes which badly affects the accuracy of classification techniques. Attribute Selection is a domain in Data Mining for selecting a subset of relevant attributes. It removes redundant or irrelevant attributes from the dataset. Through this proposed work we indent to analyze the impact of some most commonly used feature selection techniques namely Principal Component Analysis, Correlation based feature selection, Consistency based subset evaluation, RelieF, Wrapper subset evaluation, Information Gain on a classification technique namely Multi-Layer Perceptron.

(Hall and Holmes, 2003) [1] carried out similar experiment on Naive Bayes and C4.5 classification techniques. (Ozer, 2008) [2] showed improved efficiency and pruning by decision tree in fewer computations. (Phyu, 2009) [3] compared the accuracy of decision tree and bayes classification tree and also explains how these are suitable techniques for good accuracy. (Atlas LE et al.., 1989) [4] demonstrated the importance of Multi-Layer Perceptron in classification. (Rajeswari K et al.., 2012) [5] discussed the feature selection using hashing and apriori and analyzed them with Naive Bayes, J48 and Multi-Layer Perceptron.

2. Multi-Layer-Perceptron (MLP)

MLP is a classification algorithm based on a supervised learning technique called back propagation used for training the network [5]. MLP consists of multiple layers of sigmoid processing elements or neurons that interact using weighted connections. The neurons output signals which are a function of the sum of the inputs to the node modified by a simple non-linear transfer or activation function. It is the superimposition of many simple non-linear transfer functions that enables the MLP to approximate extremely non-linear functions [6].

3. Attribute selection

Attribute selection is a technique of selecting a subset of attributes from a relation which is more important to describe the relation. The selected attributes are used for model construction and the remaining attributes are considered irrelevant or redundant and thus are discarded from the relation. Following are some mostly used Attribute Selection techniques:

3.1 Correlation based Feature Selection (CFS)

CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficient is used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation.
Equation for CFS is given in equation 1.

$$r_{zc} = \frac{k r_{zi}}{\sqrt{k + k(k-1) r_{ii}}} \quad (1)$$

where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, r_{zi} is the average of the correlations between the subset features an the class variable, and r_{ii} is the average inter-correlation between subset features [7].

3.2 Consistency based subset evaluation (CNS)

Consistency measure is defined by inconsistency rate which is calculated as follows:
(1) A pattern is considered *inconsistent* if there exists at least two instances such that they match all but their class labels; for example, an inconsistency is caused by instances (0 1, *1*) and (0 1, *0*) where the two features take the same values in the two instances while the class attribute varies which is the last value in the instance.
(2) The *inconsistency count* for a pattern of a feature subset is the number of times it appears in the data minus the largest number among different class labels. For example, let us assume for a feature subset S a pattern p appears in np instances out of which $c1$ instances has class label 1, $c2$ has label2, and $c3$ has label3 where $c1 + c2 + c3 = np$.
If $c3$ is the largest among the three, the inconsistency count is $(n–c3)$. Notice that the sum of all n_ps over different patterns p that appear in the data of the feature subset S is the total number of instances (P) in the dataset, i.e.,$\sum_p n_p = P$.
(3) The *inconsistency rate* of a feature subset S ($I_R(S)$) is the sum of all the inconsistency counts over all patterns of the feature subset that appears in the data divided by P.
The consistency criterion does not incorporate any search bias relating to a particular classier enables it to be used with a variety of different learning algorithms [8].

3.3 Information Gain (IG)

Entropy is commonly used information theory measure, which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG attribute ranking methods. The entropy measure is considered as a measure of system's unpredictability. The entropy of Y is given by following equation,

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

where $p(y)$ is the marginal probability density function for the random variable Y. If the observed values of Y in the training data set S are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X. Then the entropy of Y after observing X is given by following equation (3):

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

where $p(y \mid x)$ is the conditional probability of y given x. Given the entropy as a criterion of impurity in a training set S, 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 following equation (4):

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

IG is a symmetrical measure (refer to equation (3)). The information gained about Y after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative [9].

3.4 Wrapper (WRP)

Wrapper attribute selection uses a target learning algorithm to estimate the worth of attribute subsets. Cross-validation is used to provide an estimate for the accuracy of a classifier on novel data when using only the attributes in a given subset. Wrappers generally give better results than filters because of the interaction between the search and the learning scheme inductive bias [1].

3.5 Principal Component Analysis (PCA)

PCA is one of the most widely used multivariate data analysis techniques and is employed primarily for dimensional reduction and visualization. PCA extracts a lower dimensional feature set that can explain most of the variability within the original data. The extracted features, PC_i's (Yi) are each a linear combination of the original features with the loading values (α_{ij}, i, j = 1, 2, …, p). The Yi's can be represented as follows:

$$Y1 = \alpha_{11}X1 + \alpha_{12}X2 + \ldots + \alpha_{1p}Xp; \quad (5)$$

$$Y2 = \alpha_{21}X1 + \alpha_{22}X2 + \ldots + \alpha_{2p}Xp; \quad (6)$$

$$\ldots$$

$$Yp = \alpha_{p1}X1 + \alpha_{p2}X2 + \ldots + \alpha_{pp}Xp. \quad (7)$$

The loading values represent the importance of each feature in the formation of a PC. For example, α_{ij} indicates the degree of importance of the jth feature in the *i*th PC [10].

3.6 RelieF

Relief algorithm is a feature ranking algorithm first proposed by Kira and Rendell and was enhanced to 'ReliefF' by Kononenko. The key idea of 'Relief' is to rank the quality of features according to how well their values distinguish between the cases that are near to each other. It is reasonable to expect that a useful feature should have different values between cases from different classes and have the same value for cases from the same class [11].

4. Proposed methodology

In our experiment seven standard datasets from University of California, Irvine (UCI) Machine Learning Repository are used. Experiments are performed with and without prior use of attribute selection techniques. The characteristics of datasets are shown in Table 1. The classification technique used is MLP and attribute selection techniques used are CFS, CNS, IG, WRP, PC and RLF. MLP is carried out with 10-fold cross validation. To carry out the experiment we used WEKA 3.6.10. The proposed methodology is shown in Figure.

5. Results and discussions

The results are shown and compared in Table 2. Significant improvements in accuracy are shown in bold and underline. It can be seen from the table that improvement in accuracy is gained with the use of attribute selection techniques. Different attribute selection techniques produced almost similar kind of ranking of attributes and thus when MLP is applied after them their values are not changed much. However, significant improvement is found when MLP is applied after attribute selection techniques.

6. Conclusion and future work

Data sets may contain irrelevant or redundant attributes which badly affects the accuracy of classification. It is important to remove irrelevant or redundant data from the data set. This paper has presented an analysis of MLP when used before and after six attribute selection techniques and it is found that attribute selection is beneficial for improving the accuracy of MLP.

There are still many other classification techniques remaining whose accuracies needs to be analyzed with and without the prior use of attribute selection techniques.

Table 1. Characteristics of Datasets used in the Experiment

Data sets	No. of Attributes	No. of Instances
Credit-g	21	1000
Diabetes	9	768

glass	10	214
Heart-c	14	303
labor	17	57
vote	17	435
zoo	18	101

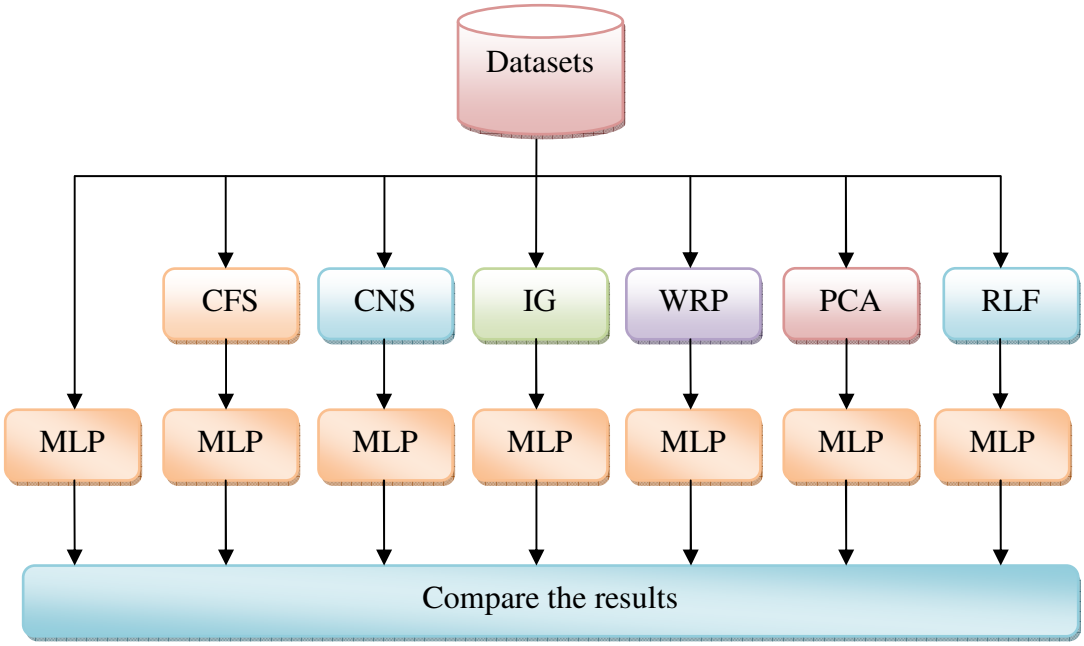

Figure1. Proposed methodology

Data sets	MLP	CFS	CNS	IG	WRP	PCA	RLF
Credit-g	71.6	**73.8**	**73.9**	**73.9**	**73.9**	**73.4**	**73.6**
Diabetes	75.39	75.78	75.78	75.52	75.78	76.17	75.91
Glass	67.76	**70.09**	**70.09**	**70.09**	**70.09**	67.76	**70.56**
Heart-c	80.86	**83.49**	**83.49**	**83.49**	**83.49**	**81.85**	**84.16**
Labor	85.96	85.96	**87.72**	85.96	84.21	85.96	85.96
Vote	94.71	94.94	**96.09**	**96.09**	**96.09**	**95.71**	95.17
Zoo	96.03	96.04	96.04	96.04	96.04	96.04	96.04

Table 2. Comparison of accuracy (in percentage) before and after selection of attributes with MLP

REFERENCES

[1] Hall M, Holmes G. Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE transactions on knowledge and data engineering 2003; vol. 15, no. 3; 1-16.

[2] Ozer P. Data Mining Algorithms for Classification. B.Sc. Radboud University, Nijmegen, Netherlands, 2008.

[3] Phyu T. Survey of Classification Techniques in Data Mining. In: IMECS 2009 International MultiConference of Engineers and Computer Scientists; March 18 - 20, 2009, Hong Kong: IMECS. pp. 1-5.

[4] Atlas LE, Conner J, Park DC, El-Sharkawi MA, Marks II RJ, Lippman A, Cole R, Muthusamy Y. A performance comparison of trained multi-layer perceptrons and trained classification trees. In: IEEE 1989 International Conference on Systems, Man and Cybernetics; 14-17 Nov. 1989; Hyatt Regancy, Cambridge, Massachusetts, USA: IEEE. pp. 915-920.

[5] Rajeswari K, Vaithiyanathan V, Tonge S, Phalnikar R. Mining Association Rules using Hash Table. International Journal of Computer Applications 2012; Vol. 57, No. 8: 7-11.

[6] Gardner M, Dorling S. Artificial neural network (the multilayer perceptron) - A review of applications in atmospheric sciences. Atmospheric Environment 1998; Vol. 32, No. 14/15: 2627-2636.

[7] Karegowda A, Manjunath A, Jayaram M. Comparative study of attribute selection using gain ratio and correlation based feature selection. International Journal of Information Technology and Knowledge Management 2010; Vol. 2, No.2: 271-277.

[8] Dash M, Liu H. Consistency-based search in feature selection. Artificial Intelligence 2003; 151: 155-176.

[9] Novakovic J. Using Information Gain Attribute Evaluation to Classify Sonar Targets. 17th Telecommunications forum TELFOR 2009; 1351-1354.

[10] Kim S, Rattakorn P. Unsupervised feature selection using weighted principal components. Expert Systems with Applications 2011; 38: 5704-5710.

[11] Liu Y, Schumann M. Data mining feature selection for credit scoring models. Journal of the Operational Research Society 2005; 1-10.