ISSN:0975-9646

Efficient Attribute Evaluation, Extraction and Selection Techniques for Data Classification

Gaurang Panchal & Devyani Panchal

U & P U Patel Department of Computer Engineering
Chandubhai S. Patel Institute of Technology
Changa, India

Abstract— In machine learning and statistics, feature selection, also known as variable selection, feature reduction, attribute selection or variable subset selection, is the technique of selecting a subset of relevant features for building robust learning models. Feature selection also helps people to acquire better understanding about their data by telling them which are the important features and how they are related with each other. Attribute subset selection on the basis of relevance analysis is one way to reduce the dimensionality. Relevance analysis of attribute is done by means of correlation analysis, which detects the attributes (redundant) that do not have significant contribution in the characteristics of whole data of concern. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for applications. This paper shows various feature selection techniques for various dataset. We have taken Intrusion Detection Problem Dataset and Gas Consumption Dataset for testing. The comparison of various feature selection techniques discussed in this paper.

Keywords-Crossover, Genetic Algorithm, Mutation, Random Population

1. INTRODUCTION

If we think for a minute about how we classify common everyday objects such as people and cars, it's pretty clear that we are using features of those objects to do the job. People have legs that are a feature that cars don't have. Cars have wheels that are a feature that people don't have. By selecting the appropriate set of features, we can do a good job of classification. To make this kind of feature-based classification work, we need to have some knowledge of what features make good predictors of class membership for the classes we are trying to distinguish [1,2]. For example, having wheels or not distinguishes people from cars, but doesn't distinguish cars from trains. These are two different classification tasks.

Depending on the classification task we are facing, different features or sets of features may be important, and knowing how we arrive at our knowledge of which features are useful to which task is essential [1].

2. FEATURE SELECTION

a. Features in Data

Before getting into feature selection in more detail, it's worth making concrete what is meant by a feature in gene expression data. One gene, call it Gene A, clearly has an enhanced expression value around samples. This expression level 'bump' is a feature [3].

b. Probability

So far, it may seem as though a nice clean distinction between features that distinguish classes clearly and those that don't al-ways exists. In fact, this is rarely the case. Most of the time all we see is an enhanced correlation between a feature and a class. For example, tall people tend to be stronger than short people. There are several reasons for this: tall people have longer arms and legs, which gives their muscles more mechanical advantage; tall people tend to have bigger muscles, simply because they are bigger people; and tall people tend to be men, who have higher testosterone levels, which helps them build more muscle. The fact remains, however, that some short women can lift more weight than some tall men. So if we were to try to classify people into two groups, 'strong' and 'weak', without actually measuring how much they can lift, height might be one feature we would use as a predictor [3]. But, it couldn't be the only one if we wanted our classification to be highly reliable.

If a single feature is not a good class predictor on its own, the alternative is to look for one or more sets of features that to-gather make a good predictor of what class an object falls into. For example, neither height nor weight are particularly good predictors of obesity; but taken together, they predict it fairly well.

c. Learning

The general process by which we gain knowledge of which features matter in a given discrimination task is called learning. For those of us who are parents, one example of this type of learning (feature selection) involves teaching our children about types of animals. We (endlessly) point to animals and say words like dog or cat or horse. We don't generally give our children a feature list that a biologist might use to define. Instead, we present examples and expect our children to figure out for themselves what the important features are. And when they make a correct guess about an animal (a correct classification or prediction), we give copious amounts of positive feedback. This procedure is called supervised learning. We present our children or our computer programs with examples and tell them what category each example belongs to, so they learn under our supervision. This is in contrast to unsupervised learning. In unsupervised learning objects are grouped together based on perceptions of similarity (or more properly, relative lack of difference) without anything more to go on. While un-supervised learning is indispensable, supervised learning has a substantial advantage over unsupervised learning [4]. In particular, supervised learning allows us to take advantage of our own knowledge about the classification problem we are trying to solve. Instead of just letting the algorithm work out for it-self what the classes should be, we can tell it what we know about the classes: how many there are and what examples of each one look like. The supervised learning algorithm's job is then to find the features in the examples that are most useful in predicting the classes [5, 6].

The classification process with supervised learning always in-valves two steps:

Training (with assessment) - this is where we discover what features are useful for classification by looking at many pre-classified examples.

Classification (with assessment) - this is where we look at new examples and assign them to classes based on the features we have learned about during training.

3. Problem Statement

The problem is: 1. How do we find a set of features that is a good for predictor? 2. Having found a good set of features, how do we use it to predict?

4. Data Pre-Processing

There are a number of data preprocessing techniques. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse. Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance [3]. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining. It reduces the number of features, removes irrelevant, redundant, or noisy data, and brings the immediate effects for application [7–9]. The Value with min and max for all accepted numeric columns from your input data file. You can change the minimum and maximum values if you know that future data for forecasting will be lower than the minimum and higher than the maximum presented in your input data file. Scaling Numeric Columns: Numeric columns are automatically scaled during data pre-processing [10–13]. By default, numeric values are scaled using the following formula.

$$SF = \frac{(SR_{max}\quad SR_{min})}{(x_{max}\quad x_{min})} \qquad (1)$$

$$X_p = SR_{min} + (X\ \ X_{min})\ \ SF \qquad (2)$$

Where X is actual value of a numeric column, Xmin is minimum actual value of the column, Xmax is maximum actual value of the column, SRmin is lower scaling range limit, SRmax is upper scal-ing range limit, and SF is scaling factor and Xp - pre-processed value. We have scale inputs with scaling range[-11] and output col-umn with scaling range [01], all the Categorical column like Sex , Marital Status with two state, Children with 3 State, Education with two state and Retention Probability with two state. We scale the data:

5. Feature Selection Techniques

There are various feature selection techniques available. Using these techniques we could find the required attributes from the dataset and the important attributes from dataset as well. Few model-selection techniques are as follows:

(1) NONE :

Specifies that no selection. This method is the default and uses the full model given in the MODEL statement to fit the linear regression [1].

(2) FORWARD :

Specifies that variables be selected based on a forward-selection algorithm. This method starts with no variables in the model and adds variables one by one to the model. At each step, the variable added is the one that most improves the fit of the model. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for inclusion [2].

(3) BACKWARD :

Specifies that variables be selected based on a backward-elimination algorithm. This method starts with a full model and eliminates variables one by one from the model [14–16]. At each step, the variable with the smallest contribution to the model is deleted. You can also specify groups of variables to treat as a unit during the selection process. An option enables you to specify the criterion for exclusion [2].

(4) STEPWISE :

This technique Specifies that variables be selected for the model based on a stepwise-regression algorithm, which com-bines forward-selection and backward-elimination steps. This method is a modification of the forward-selection method in that variables already in the model do not necessarily stay there. You can also specify groups of variables to treat as a unit during the selection process. Again, options enable you to specify criteria for entry into the model and for remaining in the model [17–19].

(5) MAXR :

This technique Specifies that model formation be based on the maximum R2 improvement. This method tries to find the best one-variable model, the best two-variable model, and so on. The MAXR method differs from the STEPWISE method in that many more models are evaluated. The MAXR method considers all possible variable exchanges before making any exchange. The STEPWISE method might remove the Design a classifier using the features in the particular sub-set

(a) Use independent data to estimate its error rate

(b) Remember the subset giving the smallest error rate

(6) Stepwise Regression :

Statistics Solutions has written the following example write-up of justification for stepwise regression. Contact Statistics Solutions today if you need assistance around stepwise regression analysis and justification [5]. Forward stepwise selection begins with independent variables being entered into the regression equation one at a time, provided predictors meet the statistical significance criteria with the dependent variable. Se-lection of independent variable entry will be based on the descending order of the largest significant correlation coefficient. Independent variables will be entered into the regression until an independent variable does not uniquely influence the de-pendent variable. Tabachnick and Fidell (2001) state stepwise regression is a model-building rather than a model-testing procedure. As an exploratory technique, it may be useful for such purposes as eliminating variables that are clearly superfluous in order to tighten up future research"(p. 144). The regression will use the F test to investigate whether independent variable(s) if any, uniquely influence the dependent variable. The R2, or the multiple correlation coefficients, will be reported to decipher how much variance can be accounted in the de-pendent variable from the independent variable(s). In the event that more than one independent variable is entered into the regression equation, R2 will also be presented to show exclusive attributes to the model or additional accounted variance in the dependent criterion.

(7) Correlation (Multiple) :

Let Y be one variable, and (X_1, X_2….X_n) a set of other variables. Let X be a linear combination of the X_i's :

$$X = \overset{X}{\underset{i}{a_i}} \quad X_i \ (3)$$

And consider the correlation coefficient? (X; Y). When the coefficients a_i are made to vary in every possible way, the value of changes. It can be shown that, in general, there is a single set of values of the coefficients that maximizes. This largest possible value of (X; Y) is usually denoted R, and is called the Multiple Correlation Coefficient between Y and the set of variables (X1, X2… Xn) [8]. The Sample Multiple Correlation Coefficient, R, is a measure of the strength of the association between the independent (explanatory) variables and the one dependent (prediction) variable. Strength of the Association:

(a) The strength of the association is measured by the sample Multiple Correlation Coefficient, R.
(b) R can be any value from 0 to +1.
(c) The closer R is to one, the stronger the linear association is.
(d) If R equals zero, then there is no linear association be-tween the dependent variable and the independent variables.

(8) Fitness Function :

We required fitness function of calculating the important of at-tributes. We can say lower the correlation is higher the fitness value will be. The fitness function is taken over here is;

$$F\,(X) = 1 \quad \min(r_x) \qquad (4)$$

Where, min(rx) is minimum value of correlation coefficient corresponding to any attribute X.

6. Data Set And Analysis

We have taken two data set for the experiment. One data set contain student information having 14 rows and six attributes (Table 6) and second data set is downloaded from the KDD CUP Dataset's web site which contains which contain the more than thousand rows and 23 attributes. Encoding Representation: The Table 6 show the bi nary representation of attributes. We have taken binary encoding in genetic Algorithm. Correlation Matrix: Table 4 shows the min(r) (Minimum value of correlation coefficient corresponding to any at-tribute). The table shows the importance of age attributes with all other attributes. Its shows the possibility of number of rows with cross combination of attributes

Table 4. Sample Data Set For Feature Selection

RID	Age	Income	Student	Credit Rating	Buys PC
1	<=30	High	no	fair	no
2	<=30	High	no	excellent	no
3	3140	High	no	fair	Yes
4	>40	Medium	no	fair	Yes
5	>40	Low	yes	fair	Yes
6	>40	Low	yes	excellent	no
7	3140	Low	yes	excellent	Yes
8	<=30	Medium	no	fair	no
9	<=30	Low	yes	fair	Yes
10	>40	Medium	yes	fair	Yes
11	<=30	Medium	yes	excellent	Yes
12	3140	Medium	no	excellent	Yes
13	3140	High	yes	fair	Yes
14	>40	Medium	no	excellent	no

Table 5. Binary Representation of feature vector

Chromosome Label	Chromosome String
Age	00
Income	01
Student	10
Credit Rating	11

Table 6. Contingency Table

	Income	**Student**	**Credit**	**Rating**	**Buys PC**	**Min (r)**
Age	1	-0.42	0.17	0	0.18	0
Income	-0.42	1	-0.45	-0.28	-0.07	0.07
Student	0.17	-0.45	1	0	0.45	0
Credit Rating	0	-0.28	0	1	-0.26	0

7. Experiments And Result

We have discussed various feature selection techniques also generated results based on that techniques using Aluda NeuroIntelligence and Sipina Tool. Below table show the various architecture selection techniques like Correlation, Inverse Training Error, In-verse Testing Error, R-Square, Akaiks Information Criterion (AIC) and Inverse Test Error.

We have taken three different dataset. One the simple student dataset, second is Intrusion Detection dataset from KDD Cup dataset and third is Gas Consumption. The following Table 4 shows the various feature selection techniques for Gas-Consumption data. Its shows the number of attributes selected after applying those techniques. The backward stepwise selection select the three at-tributes, forward stepwise select only one, while exhaustive search use three attributes and Genetic Algorithm select two attributes. The genetic algorithm find most required attributes that is Year and Tmin (Minimum Temperature) which are important attributes to find the gas consumption.

The following Table 7 shows the importance of all the selected at-tributes of using different methods. Also these methods are used to find best neural network architecture for any dataset. Also its shows the network error generated using same method. Table 7 shows the selected attributes using Forward Selection technique. This method adds one by one all the attributes and finds the accuracy of sub-set of attributes. Whichever having the lower fitness value is best combination for training dataset.

The Contingency Table 7 shows the incident between class label and protocol for given dataset. It is an important role to select number of input attributes. We have discussed various feature selection methods and some methods also show the importance of all the attributes while training the dataset.

Table 7. Contingency Table (Protocol vs. Attack Type)

tcp	0.01	0	-0.08	0	0	0	0	0.1
icmp	-0.05	-0.02	0.33	0	0	0	0	0.41
udp	0	0	0	0	0.49	0	0	0.5
Sum	0.07	0.02	0.41	0	0.5	0	0	1

8. CONCLUSION

By observing behavior of various feature selection techniques, the Genetic Algorithm find the best attributes from the data set taken for the training. Also the Correlation technique found best techniques for feature selection. Its accuracy is more for when ap-plied for real data. It is very simple and light because GA is used to search the optimal subset of attributes besides being used for searching the optimal techniques for attribute selections among the available ones.

REFERENCES

[1] A. Ganatra, G. Panchal, Y. Kosta, C. Gajjar, *"Initial classification through back propagation in a neural network following optimization through GA to evaluate the fitness of an algorithm,"* International Journal of Computer Science and Information Technology, vol. 3, no. 1, pp. 98–116, 2011.

[2] G. Zhang, "Neural Networks for Classification: A Survey," IEEE Transactions on System, Man and Cybernetics - Part C: Application and Reviews, vol. 30, no. 4, NOVEMBER 2000.

[3] G. Panchal, A. Ganatra, *"Classification and Optimization to Evaluate the Fitness of an Algorithm"*. Lap Academic Publisher, Germany, 2012.

[4] G Panchal, A Ganatra, Y Kosta, D Panchal, *"Behaviour analysis of multilayer perceptron with multiple hidden neurons and hidden layers,"* International Journal of Computer Theory and Engineering, vol. 3, no. 2, pp. 332–337, 2011.

[5] G. Panchal, A. Ganatra, P. Shah, D. Panchal, *"Determination of over-learning and over-fitting problem in back propagation neural network,"* International Journal on Soft Computing, vol. 2, no. 2, pp. 40–51, 2011.

[6] G. Panchal, A. Ganatra, Y. Kosta, D. Panchal, *"Searching most efficient neural network architecture using Akaikes in-formation criterion (AIC),"* international Journal of Computer Applications, vol. 1, no. 5, pp. 41–44, 2010.

[7] G. Panchal, D. Panchal, *"Solving NP hard problem using Genetic Algorithm,"* in National Women Conference, CITC, Changa.

[8] G. Panchal, Y. Kosta, A. Ganatra,D. Panchal, *"Electrical Load Forecasting Using Genetic Algorithm Based Back Propagation Network,"* in 1st International Conference on Data Management, IMT Ghaziabad. MacMillan Publication, 2009.

[9] C. Gershenson, "Artificial Neural Networks for Beginners."

[10] K. C. V. Cheung, "An Introduction to Neural Networks," Sig-nal & Data Compression Laboratory, Electrical & Computer Engineering University of Manitoba, Winnipeg, Manitoba, Canada.

[11] MIT, "Artificial Neural Networks." [Online]. Available: ocw.mit.edu.

[12] G. Panchal, A. Ganatra, Y. Kosta, D. Panchal, *"Forecast-ing Employee Retention Probability using Back Propagation Neural Network Algorithm,"* IEEE 2010 Second International Conference on Machine Learning and Computing (ICMLC), Bangalore, India, pp. 248–251, 2010.

[13] G. Panchal, A. Ganatra, *"Optimization of Neural Network Parameter using Genetic Algorithm"*, in National Conference on Advanced Computer Application (NCACA), 2008.

[14] G. Panchal, A. Ganatra, P. Shah, Y. Kosta, *"Unleashing Power of Artificial Intelligence for Network Intrusion Detection Problem"*, International Journal of Engineering Science and Technology, vol. 2, no. 10, 2010.

[15] D. P. Y. K. G. Panchal, A. Ganatra, *"Performance analysis of classification techniques using different parameters"*, Springer Verlag-Germany (Springer-LNCS), vol. 6411, 2010.

[16] G. Panchal, A. Ganatra, D. Panchal, Y. Kosta, *"Employee Re-tention Probability using Neural Network Based Back Propagation Algorithm"* IEEE Xplore, vol. 248, 2010.

[17] T. Ludermir, A. Yamazaki, C. Zanchettin, "An Optimization Methodology for Neural Network Weights and Architectures," IEEE Transactions on Neural Networks, vol. 17, no. 6, 2006.

[18] E. Wang, W. Lu, A. Leung, S. Lo, Z. Xu, "Optimal feed-forward neural networks based on the combination of constructing and pruning by genetic algorithms," IEEE Transactions on Neural Networks, 2002.

[19] A. Fiszelew, P. Britos, "Finding Optimal Neural Network Architecture Using Genetic Algorithms," R.Software and Knowledge Engineering Center. Buenos Aires Institute of Technology .Intelligent Systems Laboratory. School of Engineering. University of Buenos Aires.

[20] G. Panchal, A. Ganatra, *"Optimization of Neural Network Parameter using Genetic Algorithm"*. Lap Academic Publisher, Germany, 2012.