Feature Selection Using Classification Methods

Ruyi Huang and Arturo Ramirez, University of California Los Angeles

Biological/clinical studies usually collect a wide range of information with the potential of (1) Identifying disease-associated features (symptoms, bio-markers, etc.), (2)Diagnosis, prognosis, and prediction of therapeutic responsiveness for existing disease; (3) Discovering new symptom or type of disease. These biological information could be presented in many different forms: microarray gene expression data, mass spectrometry data, functional magnetic resonance imaging for the whole brain, etc.. Since these datasets all have tens or hundreds of thousands of variables to look into, techniques in data analysis are needed to pick out the variable that can be used as the predictor to predict for the dataset (e.g. symptoms or bio-markers indicate the disease), improve the performance of the predictors (get faster and more cost-effective predictors) and provide a better understanding of the underlying process that generated the data (e.g., mechanism underlying the disease pathology). Variable and feature selection are focusing on constructing and selecting subsets of features that are useful to build a good predictor. In this paper review, we are going to use the analysis of locomotion behavioral data collected from an Alzheimer's Disease gait study as example to go through the variable and feature selection using different classification methods and strategies, including logistic regression, Classification and Regression Tree (CART), random forest. Support vector machine will be used to further test and refine the feature selection criteria.

Categories and Subject Descriptors: H.5.2 [**Survey**]: User Interfaces—*Evaluation/ methodology*

General Terms: Feature Selection

Additional Key Words and Phrases: Classification, Logistic Regression, CART, Random Forest, Support Vector Machine

ACM Reference Format:

Ruyi Huang and Arturo Ramirez. 2014. Feature Selection Using Classification Methods. *ACM Trans. Appl. Percept.* 0, 0, Article 0 (2014), 8 pages.
DOI:http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

1.1 Classification

Classification is one supervised learning analytical method used for pattern recognition. The training data are accompanied by labels indicating the class of the observations and once the training set is set up, new data is classified based on the training set. A classification problem occurs when an object needs to be assigned into a predefined group or class based on a number of observed attributes related to that object. The individual observations are analyzed into a set of quantifiable properties, known as explanatory variables, features, etc. These properties may variously be categorical, ordinal, integer-valued or real-valued. During classfication, given objects are assigned to prescribed classes. A classificer is a mathematical function, implemented by a classification algorithm, that maps input data to a category, which performs classification. The current developed classfiers including, Naive Bayesian Classifier, neural networks, K-nearest neighbor algorithms, Decision tress and Support vector machines, etc.. How to select the classifiers serve the problem the best would be important for getting the pattern well recognized and in biological application that means whether we can separate one phenotype (gene expression, behavioral observations, etc.) clearly from another.

However, one problem in classifying the biological data is the dimensionality. Because of the limitations in getting experimental samples, the number of varaibles/ features would be too large relative to the number of the samples to get trained for classification. One straight forward way to solve the data dimensionality problem is to perform dimensionality reduction by feature selection, which is to select the best subset from a given feature set to represent the whole dataset.

1.2 Feature Selection

As of 1997, when a special issue on relevance including several papers on variable and feature selection was published, few domains explored used more than 40 features. However, started from 2000, we have witnessed a surge of activity to develop analytical technologies to monitor ”global” changes in biological or clinical studies such as transcriptome, integrated behavioral study, etc.. Research focusing on the ”global” changes requires sophisticated instrumentation, ideally managed by expert operators, can be both costly in terms of research consumables and experimentation time. Usually these ”global” changes studies often contain tons of variables or features while the samples per class replicates usually are too few to allow adherence to an experimental statistical design that can cope easily with such high degree of biological and instrument-derived variance encountered when data modeling is undertaken. Since high efficient and accurate comprehensive analysis of datasets with hundreds of variances will facilitate the discover of new biological/ clinical markers (predictors), which could be detected faster and with high efficiency, help disease diagnosis, prognosis, prediction of therapeutic responsiveness and reveal the possible underlying mechanism of the disease/ biological phenomenon, it is important to search for a decent statistic strategy to solve this analytical task.

For all different biological or clinical datasets with a high dimensional data spaces, they mostly share a common subtask, which is feature/ variable selection. For predictive classification, only a subset of variables is used to avoid overfitting, where a classifier is known ”too well” to fit even irreproducible ”noisy” training patterns and, thus, to achieve predictive accuracy that generalizes well to unseen/ test data. A support vector machine, a computer algorithm that learns by example to assign labels to objects, is applied by many research groups in combination with the feature selection to use a limited number of already seen/ tested data to build up a classification model. Lastly, a separate objective is to identify the variables and their effect.

Although feature selection is integral to each of these analytical tasks, practical feature/ variable selection techniques are heuristic, with an inherent accuracy/complexity tradeoff. Moreover, while multivariate analysis methods based on complex criterion functions may reveal subtle joint marker effects, they are also prone to overfitting. Additionally, high dimensionality compromises the ability to validate marker discovery, which requires accurately measuring true and false discovery rates. These issues have prompted the development of a variety of statistical strategies for estimating and limiting false discoveries.

Animal behavioral tests data analysis could be a good example for the biological problems involving high-dimensional data spaces feature selection. For our study, we are focusing on the gait analysis for Alzheimers’ Disease model mice to find out whether there is any difference in the gaiting pattern of Alzheimers’ transgenic animals and their wildtype (normal) littermates. For each animal, 54 variables are monitored at same time and will be recorded by high frequency video-taper for at least 3200 frames. Going through all the recorded data to figure out a special gait pattern of Alzheimer’s Disease transgenic mice would be time-consuming as well as labor demanding to finish. To solve this problem, we choose to use feature/variable selection which belongs to pattern recognition problem. Since pattern recognition system is made up by two mode: Classification Mode and Training Mode

we plan to carry out the feature/variable selection in two steps: (1) Variable classification. In this step, we are aiming at further separating all the recorded variables into different subsets and then test the subsets to pick a combination separating the gait pattern of the Alzheimer’s animals from the normal wildtype animals the best (choose the classifier to separate Alzheimer’s gait from normal gait). (2) Test the classification with different classifiers such as Decision trees, decision lists or Support

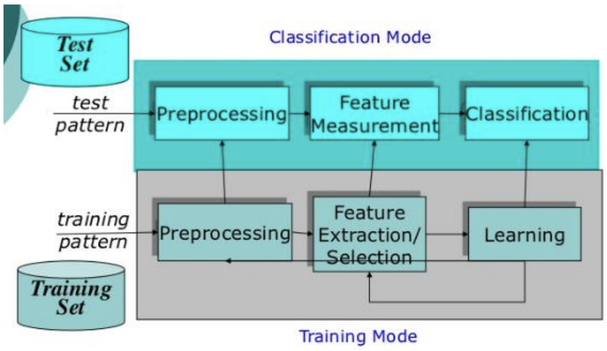

Fig. 1. Medical Example

vector machines (SVM). In order to deal with more upcoming data collected from the experiments, we are going to establish classifiers based on the seen/ tested data to train the classifiers and pick out the fittest classifiers with the new data.

2. LOGISTIC REGRESSION

Logistic regressio is a form of Generalized Linear Model (GLM) that uses a logit link function in order to assign probability scores. In our case, we would assign the probability of being in one of the two output classes ($Y = 1$ for transgenic and $Y = 0$ for wild type). The model is constructed by taking a linear combination of the input features :

$$f = B_0 + B_1X_1 + B_2X_2 + ... + B_nX_n = B_0 + \sum_{i=1}^{n} B_iX_i$$

We can then assign a probability score as follows:

$$P(Y = 1|f) = \frac{1}{1 + e^{-f}} \text{and } P(Y = 0|f) = 1 - P(Y = 1|f)$$

If we let $X = [X_0X_1...X_n]^T$ and $B = [B_0B_1...B_n]^2$, then $f = B^TX$. We can then estimate B using maximum likelihood estimation. The log likelihood function can be represented as:

$$l(B) = logL(B) = \sum_{i=1}^{n}(1 - Y^i)(-B^TX^i) - log(1 + e^{-B^TX^i})$$

No closed form solution exists to this problem, but B can be estimated using a variety of numerical methods. A common approach for this problem is using Newton's Method, which in this case can be outlined as follows:

$$l'(B) = XT(Y - p)\text{and } l''(B) = -X^TWX$$

where W is a diagonal weight matrix. The update equation then becomes :

$$B^{new} = B^{old} + (X'WX)^{-1}X'(y - p)$$

We continue this process of estimating p and updating B until convergence to obtain our estimates for B.

2.1 Feature Selection Methods

Once a model is constructed following the described outline, we can then use test statistics to evaluate variable significance and derive a parsimonious model. Individual variable significance can be derived using the Wald statistic. From Newton's method we have $Var(B) = (X'WX)^{-1}$. It follows that for variable j, the Wald statistic is:

$$W_j = [\frac{B_j}{SE(B_j)}]^2 \sim \ \chi_1^2$$

Alternatively, we may want to use the likelihood ratio test to determine variable significance, as **?**] states that this test is more reliable than the Wald test in small sample sizes. The likelihood ratio test is computed in the normal fassion:

$$-2log\frac{L_0}{L_1} = -2[log(L_0) - log(L_1)]\chi_{p-q}^2$$

where L_1 is the full model likelihood, L_0 is the reduced model likelihood, p is the number of estimated parameters in the full model, and q is the number of estimated parameters in the reduced model. The likelihood ratio test statistic can also be used as an importance score for performing backward elimination to identify significant variables. The end result is a model consisting of the smallest number of significant variables necessary to provide a relatively adequate fit. Logistic regression also provides a simple probabilistic interpretation of variable significance in the form of odds ratios.

3. RANDOM FOREST

Random Forest classification is a classification method that was first proposed by **?**] which involves the construction of multiple classification trees at the training stage and outputting the class model that occurs most often among the multitude of individual classification trees. Therefore, it can be thought of as an extension of the Classification and Regression Tree methods (CART) outlined in **?**].

3.1 Classification Tree

An individual classification tree can be thought of as a heirarchical partitioning of the instance space where the entire space is used initially and then we recursively divide the space into smaller regions. The end result is that each region is assigned with a class label. In our case we are dealing with a binary outcome, wild type or transgenic. The tree consists of nodes that form a directed path, or rooted tree, with a root node. This root node has no incoming edges. All other nodes in the tree have one incoming edge. Nodes with edges going down the tree are known as internal nodes. Nodes without outgoing edges are know as leaves or decision nodes. Each internal node of the tree splits the instance space into two or more sub spaces based on a discrete boundary rule of the input attributes values. In this way we can see that categorical inputs have a natural splitting rule for internal nodes, while continuous input variables must be discretized, or partitioned into a categorical variable based on ranges of the continuous input. Each leaf at the bottom of the tree can represent the class most representative of the target value or more commonly, a probability vector indicating the classification probabilities associated with that particular path down the tree. A simple example can be found in figure 2 where we examine the risk of death among patients admitted to the hospital based on measurements taken during the initial 24 hour period of admittance.

In this example we have binary classification outcomes where the two classes are low risk and high risk. The root node in this tree is based on the minimum systolic blood pressure within the initial 24 hours where the continuous values are split by if the measurement is above 91. If it is below 91 the patient is classified as high-risk. None of the other measurements are needed in order to reach this classification decision. If it is above 91, then we move down the edge to the next node. The classifier then checks the age of the patient. If it is below 62.5 years old, the patient is classified as low risk. If the patient is over 62.5 years old, we move down the edge into the next node. Here we check whether sinus tachycardia is present. If it is absent, the patient is classified as low risk while if it is present, the patient is classified as high risk.

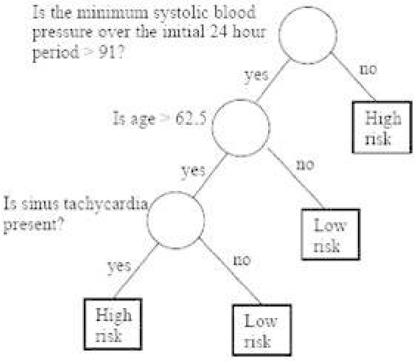
Is the minimum systolic blood pressure over the initial 24 hour period > 91?
yes
no
High risk
Is age > 62.5
yes
no
Low risk
Is sinus tachycardia present?
yes
no
High risk
Low risk

Fig. 2. Medical Example

The classification tree is constructed through top down induction where we select the ”best” input variables to use for each node based on a purity measure. A purity score measure is defined and the input variable which realizes the best score among possible(remaining) input variables is selected. A variety of score measures can be used, but each score measure is designed to favor input variables which best discriminate between observations belonging to different classes. In this way, the score typically evaluates the ability of the input variable to reduce the classification error in the sample or sub sample. Two common measures of impurity include Shannon’s entropy :

$$I(S) = -\sum_{c=1}^{m} \frac{N_c}{N} log \frac{N_c}{N}$$

and Gini’s entropy:

$$I(S) = \sum_{c=1}^{m} \frac{N_c}{N} (1 - \frac{N_c}{N})$$

where N is the size of the training sample S and N_c is the number of observations from output class c. Once a measure is chosen, we construct a tree starting with the input variable that provides the best discrimination, split the tree, and recalculate the measure on the remaining variables. We run this process until we reach a leaf node.

3.2 Random Forest Methodology

The above procedure describes the process of creating an individual classification tree. A random forest is an extension that creates a classification tree based on a random permutation of the data. This random permutation is used the training set for growing the tree. Each ”best split” is calculated using a random subset of all candidate input variables. This process is repeated at each edge to produce the nodes of the tree until a leaves are reached. A classification tree is constructed for each random permutation of the data set. The collection of each of these trees constitutes the forest. For each observation, we assign classification based on the class label assigned by a majority of the trees in the forest.

3.3 Feature Selection Methods

We can extract relevant features from an individual tree using the out of bag error rate. The out of bag error rate is the proportion of observations not used in training that are misclassified by the tree. Because each tree is constructed using a permutation of the data set, each tree has its own natural testing set that was not used in the construction of the tree. To evaluate the significance of a variable j, we first find the out of bag error rate for the tree. We then randomly permute the values of variable j in the out of bag data set. Under the assumption that the out of bag error rate is not influenced by j :

$$\frac{OOB_s - OOB_(j^-)}{SE(OOB_(j^-)} \sim \ N(0,1)$$

where OOB_s is the overall out of bag error rate and $OOB_(j^-)$ is the out of bag error rate under randomly permuted j values. We can then aggregate the out of bag error increase across all trees to determine the significance of variable j. We can also determine the smallest subset to adequately classify observations by performing backwards elimination based on out of bag error rate at each splitting stage.

The main advantages of this classification method with respect to our goal are that it runs efficiently, can handle many input variables without deletion, has classification validation and variable importance measures built in, and is relatively simple to explain and understand. But, we must be aware of the inherent biased to favor inputs with more levels.

4. SUPPORT VECTOR MACHINES

Support Vector Machines (SVM's) is a non-probabilistic binary linear classifier that operates by mapping all observations into a representation space where observations of one class are as "seperate" as possible from the observations of another class. In essence we attempt to construct a hyperplane that maximizes the distance between it and the closest point on both sides of the boundary plane. **?**] show that this is an optimization problem, because as figure 3 shows, there are a multitude of ways of constructing this hyperplane.

Given a our x inputs and a binary outcome ($Y_i = -1, 1$) we can construct the optimal hyperplane using the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$$

Input vectors x_i are mapped to a higher dimension space by function ϕ. C is a penalty parameter of ξ_i, the degree of misclassification. This optimization problem can be solves through the use of Lagrange multipliers. If we want to construct a non-linear function to represent the hyperplane, we can use the "kernel trick", where we define a kernel function, $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$. The goal is to transform the feature space in a way that provides a linear classifier on this transformed spaces. Figure 4 shows an example.

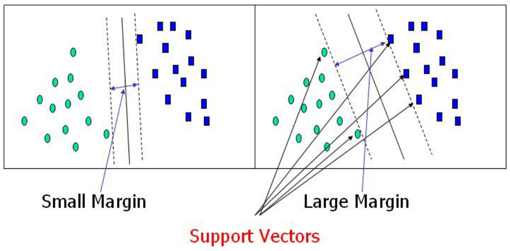

Fig. 3. Two possible hyperplanes (lines in 2 dimensions)

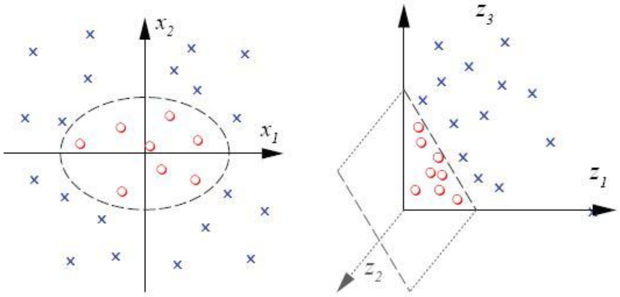

Fig. 4. Kernel Transformation

From this fomulation we can see that any categorical inputs must be transformed to numeric calues. We would also need to scale the input variables, in order to avoid having inputs with larger ranges dominate over inputs with smaller ranges. Scaling will also avoid numerical calculation issues, as kernel values typically depend on an inner product of feature vectors.

4.1 Feature Selection Methods

Once we construct the optimal hyperplane, we can evaluate individual input variable significance using the F-score proposed by **?**]. Given training vectors x_i, if the number of positive and negative observations are n_+ and n_- respectively, the F-score of the ith feature is :

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{j=1}^{n_+}(x_{j,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{j=1}^{n_-}(x_{j,i}^{(-)} - \bar{x}_i^{(-)})^2}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, and $\bar{x}_i^{(+)}$ are the average of the ith feature for the whole, positive, and negative respective sets and $x_{j,i}^{(+)}$ and $x_{j,i}^{(-)}$ are the ith feature of the jth positive and negative observations respectively. In this score, numerator indicates the discrimination between the positive and negative sets, and the denominator indicates the one within each of the two sets. Therefore, the larger F-scores are associated with features that are more likely to be discriminative.

The F-score allows us to evaluate input variables individually. **?**] introduces the concept of recursive feature elimination for SVM's. Here, criteria other than the F-score are introduced, but the backward feature elimination process is the same:

1. Construct the classifier
2. Compute the selected ranking criterion for each feature
3. Remove features with smallest ranking criterion

The last step can remove either individual features or a group of features. Removing a set of features at each iteration can be computationally efficient, but the paper notes that it introduces the possibility of classification performance degredation. By controlling this process, we can designate stopping points based on the desired number of features or classification error rate.

5. DISCUSSION

The goal of our project is to determine features that can discriminate between binomial classes. The proposed course of action is to develop a classification technique to differentiate between classes and evaluate feature significance within this classifier. We have looked at three possible candidate classification techniques in logistic regression, random forests, and support vector machines. Within each method, we explored a few of the techniques for evaluating variable significance and for identifying the best smallest subset of features to adequately classify the data. Also within each technique are various methods for ranking the significance of an input. For example, random forest metrics are typically based on information measure while SVM metrics can be based off of risk minimization measures, although information criteria have also been proposed **?**]. The goodness of fit of a particular classifier can often be evaluated through classification error rates or ROC curves, which will allow us to compare the different methods. But, our focus is on the selected features. It is still unknown as to how large of an effect using different classification methods, or even ranking criteria within a specific method, will have on the subset of features selected as significant. Exploration of these and other methods under the guidance of biological fundamentals of the research will be carried out as the data set is finalized.

REFERENCES

A. Agresti. 1996. *An Introduction to Categorical Data Analysis*. John Wiley and Sons, Inc.

B. E. Boser, I. Guyon, and V. Vapnik. 1992. A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (1992), 144–152.

L. Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

L. Breiman, J. Friedman, R. Olsen, and C. Stone. 1984. *Classification and Regression Trees*. Wadsworth International (California).

Gerda Claeskens, Christophe Croux, and Johan Van Kerckhoven. 2008. An information criterion for variable selection in Support Vector Machines. In *JMLR, SPECIAL TOPIC ON MODEL SELECTION*. 541–558.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. 2000. Gene selection for cancer classification using support vector machines. *Machine Learning* 3 (2000).

Yi wei Chen and Chih jen Lin. 2005. Combining svms with various feature selection strategies. In *Taiwan University*. Springer-Verlag.