SESUG Paper 200-2019

Summarizing some conventional methods to classify a binary target

Yida Bao, Dr. Philippe Gaillard, Auburn University

Abstract:

An average of about thousands of sports articles are published online every day. However, the quality of article varies, and a good article is always easy to be neglected. To elicit readers' interest and give user a better experience, World Wide Web hire experienced editors to manually (i.e., reading) classify articles as "Subjective" or "Objective". This classification procedure is tardiness, which may substantially hamper the efficiency of the website. We propose an automated way of classifying the sports articles, using several conventional methods to classify the sports articles and compare the misclassification error rate of each method. Each article has its own syntactic or semantic features, including parts of speech-grams, word level sentiment, and phrase-level sentiment. During our first step of classification, we use the **SAS ®** procedure **PROC** **HPCLUS** to explore 1000 sports article's cluster information based on these features. Later, **PROC** **DISCRIM** implements K-Nearest Neighbors and Discriminant Analysis. Also, we use SAS ® Enterprise Miner to apply several machine learning methods into this case, such as Logistic regression, random forest, decision tree, and neural network.

Key words: variables features ●Labels ● Classification ● quadratic discriminant analysis ●Random forest ● error rate ● KNN ●Neural Network ● decision tree● logistic prediction ● cluster ●

Introduction:

Classification stands an extremely vital position among statistical methods. The idea of trying to design and build a machine that recognizes difference modes would be the pursuit of science researchers. Many applications, from automatic speech recognition to fingerprint recognition, optical character recognition, DNA sequence analysis, etc., clearly demonstrate the essential role of a reliable and accurate pattern recognition system. When facing complex sensor data, classification is the first key processing step when extracting useful information for all intelligence systems.

A classifier uses part of training data to figure out how given variables relate to the group. In supervised classification setting, researchers have a set of training observations (x_1, y_1)…. (x_n, y_n) that they can use to build a classifier. Since the study already tags a label for each sample size, the purpose of the classification process is to establish a model that, given a sample of dataset and desired outputs, best approximates the relationship between input and output observable in the data. A good classifier will perform well not only on the training data but also on test observations that were not used to train the classifier. Note that "correct" output is determined entirely from the training data, so while statisticians do have an objective truth that our model will assume is true. Actually, data labels are not always correct in real life. Noisy, or incorrect, data labels will reduce the effectiveness of the model. There are many possible classification techniques, for example, ***logistic regression**, **linear discriminant analysis**, and **K nearest neighbors***, that one might use to predict a qualitative response. Also, there exist some more computer-intensive methods, such as ***random forests**, **neural networks**, **decision tree** , and **support vector machine**.*

Besides, within the study of pattern recognition, the other main type of tasks is unsupervised learning. Since the unsupervised learning does not have any labeled outputs, so its goal is to infer the natural structure present within a set of data points. Unsupervised learning is useful in the exploratory analysis because it can automatically identify structure in data. The most common methods within unsupervised learning are clustering. Some common algorithms include **K-means clustering**, **principal component analysis**.

Classification may refer to many areas, such as Business, Biological, and Chemical. In this study, we apply classification into the media market. The sports market is constantly going up during the last half of the century people love sports and treat their favorite sports team as part of their lives. People never more than now eager for sports information with high quality to better enjoy the sports itself, and they need an accurate daily source from experts with responsibility. How to recommend the reader appreciate sports articles? Indeed, the company could hire lots of employees to figure out which sports article is an object or subject. That's exactly what they do now. In this case, when the classifier is trained accurately, it can be used to detect subjective articles by the system. An automated solution with the ability to classify between objective and subjective articles become a desirable and marketable requirement.

Also, in our paper, we will focus on the different types of classifiers, and compare the misclassification for each method.

Data description:

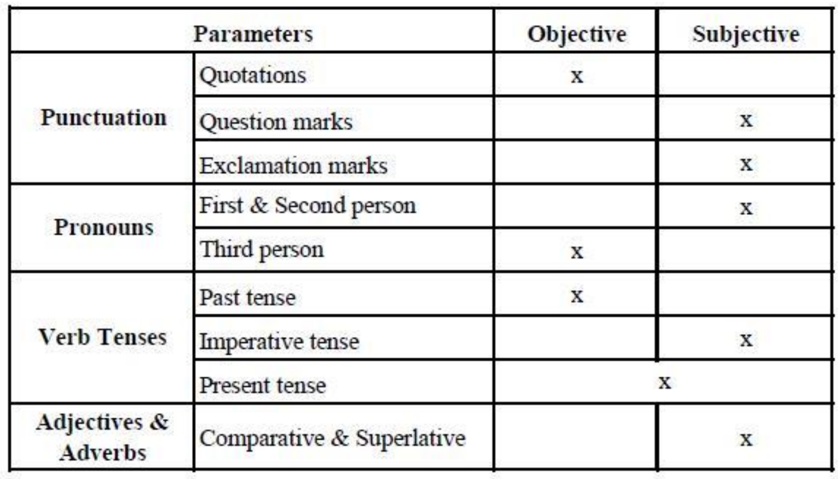
How to use the numeric character to describe the features of the article? According to some linguistic experts' opinion, subjectivity can be express in a multiple of ways that are "directly using statements such as "I believe and I think" or inferred. So during the process of creating a dataset, the researchers use syntactic to semantic features including total words count, n-grams, word-level sentiment, and phrase-level sentiment scores. A human reader can usually evaluate the subjectivity of an article by inspecting, in a glance, the presence of specific syntactic features and their frequency of occurrences such as the excessive usage of adjectives. Upon focused formal researchers’ work, **Table 1** gives a summary of basic character difference between objective and subjective articles.

- **Quotations** rather insinuate objectivity since they indicate that eh author is reporting something said by others;

- **Question marks and exclamation marks** show evidence of subjectivity since the author could be inquiring for information, expressing her/his surprise , or emphasizing some news.

- **Past tense verbs** point to objectivity since they involved events that occurred in the past and are being narrated by the author. However, it is important to also consider the pronoun used in conjunction with the verb since a past tense verb with the first or second person pronoun infer a sentence that is rather somewhat subjective.

- **Third-person pronouns** refer to objectivity, while **first and second-person pronouns** indicate subjectivity.



Parameters		Objective	Subjective
Punctuation	Quotations	x	
	Question marks		x
	Exclamation marks		x
Pronouns	First & Second person		x
	Third person	x	
Verb Tenses	Past tense	x	
	Imperative tense		x
	Present tense	x	
Adjectives & Adverbs	Comparative & Superlative		x

Table 1 basic feature information

There exist 1000 sports articles in the dataset, including 635 articles label as objective labels and 365 articles label as subjective. Based on different types of features, we got 48 variables (Originally, there exist 50 variables. However, variable WRB NNP‘s value are all equal to 0, we prefer to abandon them from the analysis).

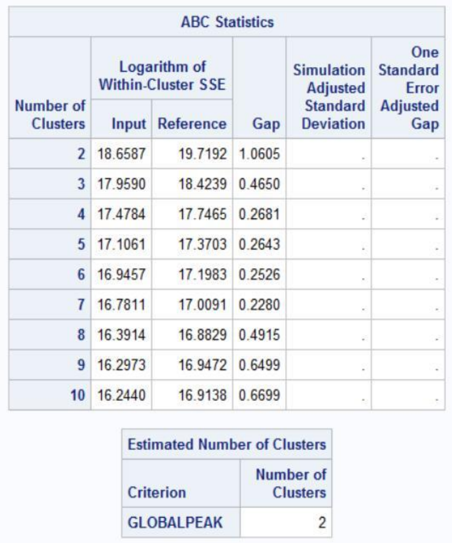
Proc Hpclus:

Normally, we don't code Cluster process when data already exist an order or categorical certain group. However, Some times, we want to solve the problem in reverse: we hide the labeled variable and directly train sample sets with the classifier, and compare the result with the original label. In principle, can we learn something useful from the unmarked samples? It totally depends on whether we are willing to accept some assumptions. This method is more suitable for data mining applications because these projects often do not understand the specifics of the data to be processed. We want to extract some basic characteristics of data with unsupervised learning, which will be useful for our further classification. More importantly, most unsupervised learning methods work in a data-independent manner, which provides very effective pre-processing for subsequent steps.

In this study, the label for sports article manually works .Sometimes, many expert labels "subject" or "object" tags by following their habits. Before the classification process, it better for us to confirm the original label is reasonable. We use **Proc Hpclus** as the pretreatment method, find the best K number and compare it with the label number. **Proc Hpclus** is the previous step of the K-Means process. At SAS, we have two methods for solving this problem: **cubic clustering criterion** (CCC), a method that estimates the number of clusters, and **Aligned Box Criterion** (ABC), which leverages parallel computing to evaluate the reference distribution based on the input data to estimate the number of clusters.

According to the opinion Ilknur Kabul, senior manager of machine learning algorithms group at SAS," as for ABC method goes, we can estimate how many clusters you have in a data set and how confident we are in these clusters—it's really based on your input data set." Since we already have hypothetical expectations for this dataset cluster, we use the **ABC** method to code the clustering process.

```
proc Hpclus data = sports maxcluster = 10 MAXITER=100 NOC=ABC;
Input[…Variable insert…] / level = interval;
Ods output ABCStats = ABC;
run;
```



ABC Statistics					
	Logarithm of Within-Cluster SSE				
Number of Clusters	Input	Reference	Gap	Simulation Adjusted Standard Deviation	One Standard Error Adjusted Gap
2	18.6587	19.7192	1.0605	.	.
3	17.9590	18.4239	0.4650	.	.
4	17.4784	17.7465	0.2681	.	.
5	17.1061	17.3703	0.2643	.	.
6	16.9457	17.1983	0.2526	.	.
7	16.7811	17.0091	0.2280	.	.
8	16.3914	16.8829	0.4915	.	.
9	16.2973	16.9472	0.6499	.	.
10	16.2440	16.9138	0.6699	.	.

Estimated Number of Clusters	
Criterion	Number of Clusters
GLOBALPEAK	2

Table 2 Gap value

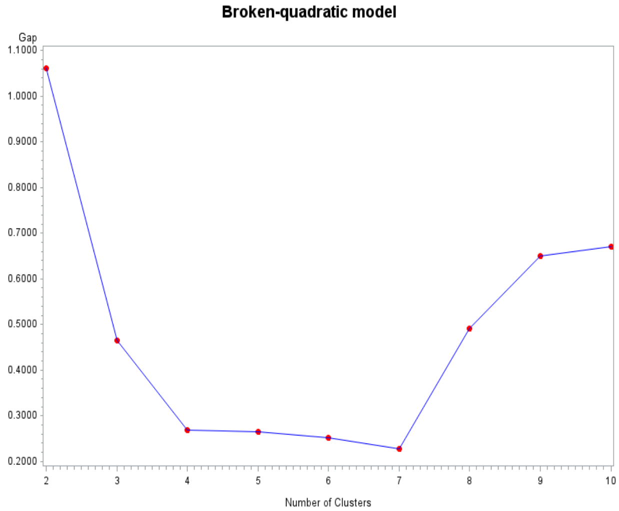
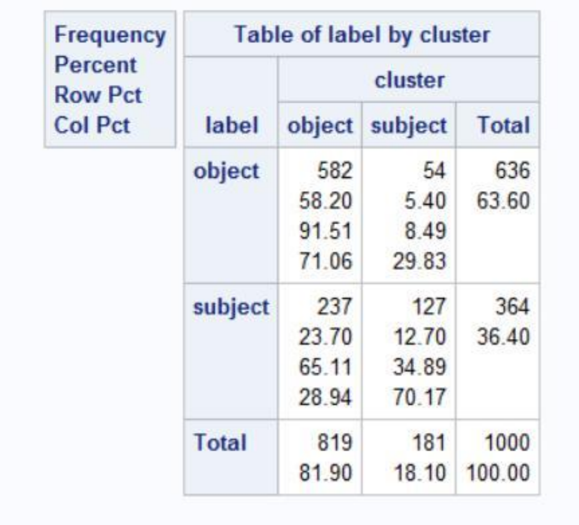


Figure 1 gap graph

During the **Proc Hpclus** process, **SAS** test the gap value under the different number of clusters. The **GLOBALPEAK** option selects the peak value that has the maximum value among the peak values in **Gap(k).** Gap(2) reaches the max value of 1.0605. SAS estimates the number of clusters is 2, which has same number as our label ---subjective or objective. Besides, it is easy for us to observe from the graph. When k = 2, the line chart reach the first peak. We will get the same result if use the **FIRSTPEAK** option. Then we proceed with **Proc Fastclus** to assign each sports article a specific cluster value, compare the result with the original label, calculate the frequency and generate the two by two frequency table.



Frequency
Percent
Row Pct
Col Pct

Table of label by cluster

	cluster		
label	object	subject	Total
object	582 58.20 91.51 71.06	54 5.40 8.49 29.83	636 63.60
subject	237 23.70 65.11 28.94	127 12.70 34.89 70.17	364 36.40
Total	819 81.90	181 18.10	1000 100.00

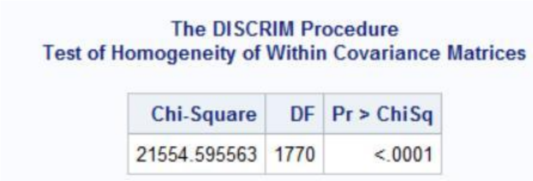
Table 3 Cluster versus label 2 by 2 table

It’s easy to obtain the error rate, which is equal to $\frac{237+54}{\text{Total}}$ = 29.1%. Indeed, 29.1 % is either a good or bad result, since the cluster is unsupervised analysis (or we call it Data-Driven), we can’t use a common standard to judge the result. According to the output, the original manually label seems plausible in a sense.

Discriminantanalysis and KNN:

From this part, we will introduce several different classification algorithms and based on an original label with supervised learning. The discriminantanalysis (**LDA & QDA**) is a generalization of Fisher's linear discriminant, a method used in statistics to find a linear combination of features that separate into subjective and objective. Discriminant analysis models the distributions of predictors X separately in each of the response classes, and use Bayes’ theorem to flip these around into estimates for the probability of the response category given the value of X.

Linear discriminant analysis (LDA) assumes that observations within each class are drawn from a covariance matrix that is common to all classes. However, unlike **LDA**, **Quadratic discriminant analysis (QDA)** assumes that each class has its own covariance matrix. That is, it assumes that observation from the kth class is of the form $X \sim N(\mu_k, \Sigma_k)$, where $\sum k$ is a covariance matrix for the kth class. So, before processing the discriminantanalysis, we have to check whether the covariance matrix from each group is equal or not. We will use linear discriminantanalysis if the covariance matrix is equal to each other, or we prefer to use Quadratic discriminantanalysis. Since **LDA** is a much less flexible classifier than **QDA**, and so has substantially lower variance. Although it can potentially lead to enhance the prediction performance, we can’t violate the assumption for each method.



The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
21554.595563	1770	<.0001

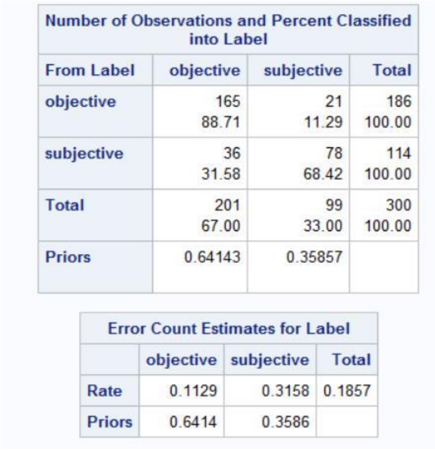
(Table 4 homogeneity of covariance matrices)

The p-value is smaller than 0.0001, which indicates the covariance matrix are significantly different between the two groups. According to the output, we will use QDA instead of LDA in our study. We use function ranuni(value) to assign a random from 0 to 1 for each sample and proceed with the if function to split it. After splitting the dataset, **Proc discrim** will apply for the dataset.

```
proc discrim data = train testdata = test method = normal pool = no
     distance list testout = result ;
class label;
priors prop;
var […variable insert…];
run;
```

Our first step is to split the dataset into train and test. Generally, 70% of the available data is allocated for training, and the remaining 30% are referred to test data, so we can assess how well our model performs on an out of sample dataset. The idea is that more training data would be better because it makes the classifier better whilst more test data makes the error prediction more accurate. Of course, the 70:30 proportion is not mandatory and could be varied by different data sets.

For **Proc Discrim** Function, "no" is set for pool option, since we already get there exists significant difference covariance matrices between the subjective and objective group. Besides, we set "prop" for priors, since the default setting assumes the proportion between subjective or objective paper stands the same.



Number of Observations and Percent Classified into Label

From Label	objective	subjective	Total
objective	165 88.71	21 11.29	186 100.00
subjective	36 31.58	78 68.42	114 100.00
Total	201 67.00	99 33.00	300 100.00
Priors	0.64143	0.35857	

Error Count Estimates for Label

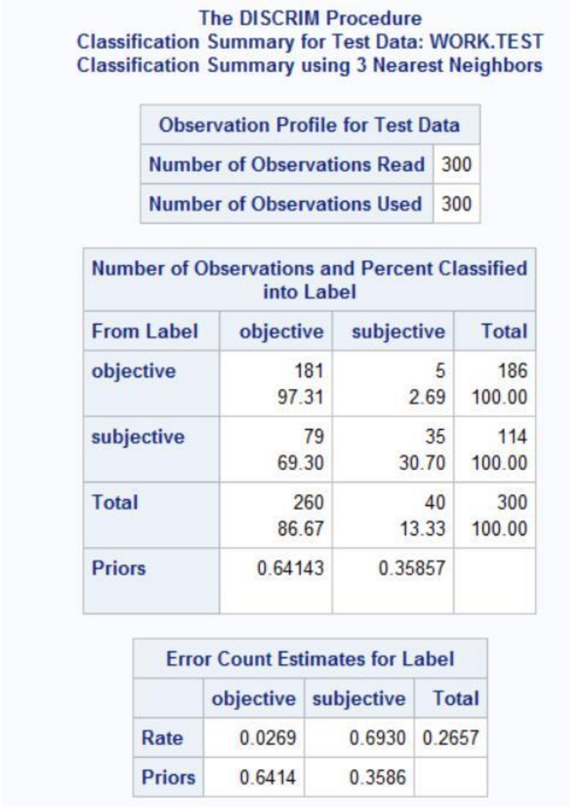
	objective	subjective	Total
Rate	0.1129	0.3158	0.1857
Priors	0.6414	0.3586	

(Table 5 QDA OUTPUT)

The final output is the classification result for the "test" dataset. From the table above, "test" data includes 300 observations, and the error rate for misclassification is 18.67%.

Next, we proceed with **K- Nearest Neighbors** algorithm, which is also a type of machine learning method. In order to make a prediction for an observation X=x, the K training observations that are closest to x are identified. So, **KNN** is a completely non-parametric method for classifier: no assumptions are made about the shape of the decision boundary. **QDA** may give better results, however, in many cases, **KNN** as a non-parametric method can be superior.

SAS code almost performs the same with the **QDA** process. SAS acknowledges "NPAR" as KNN in method option. Besides, it would be difficult for us to decide the value of K, since the best choice of K depends upon the data. Larger values of k reduce the effect of the noise on the classification but make boundaries between classes less distinct. The proper K value varies from 3 to 10, we select K= 3 that makes the error rate reaches the minimum level. The value of K was chosen using the cross-validation approach.



The DISCRIM Procedure
Classification Summary for Test Data: WORK.TEST
Classification Summary using 3 Nearest Neighbors

Observation Profile for Test Data	
Number of Observations Read	300
Number of Observations Used	300

Number of Observations and Percent Classified into Label

From Label	objective	subjective	Total
objective	181 97.31	5 2.69	186 100.00
subjective	79 69.30	35 30.70	114 100.00
Total	260 86.67	40 13.33	300 100.00
Priors	0.64143	0.35857	

Error Count Estimates for Label

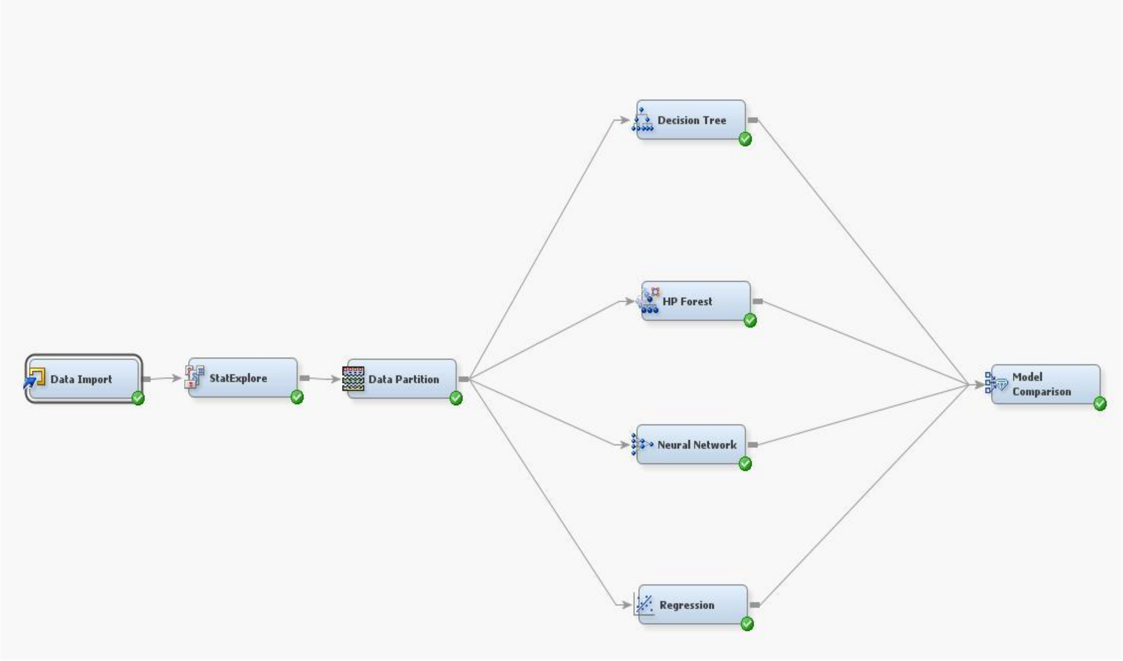
	objective	subjective	Total
Rate	0.0269	0.6930	0.2657
Priors	0.6414	0.3586	

(Table 6 KNN output)

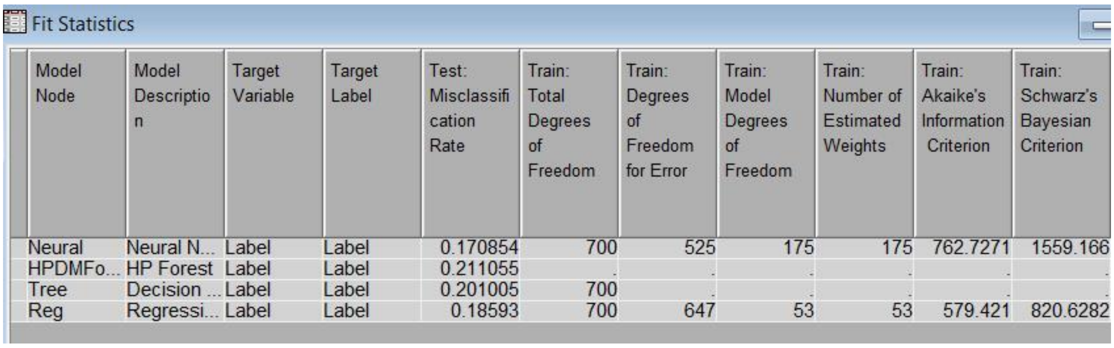
Through the table above, we get the **KNN** algorithm result, the misclassification rate for test data is 26.57%, which is much worse than **QDA**.

Other machine learning method:

We will use **SAS Enterprise Miner** platform instead of **SAS 9.4** to deal with the sports article dataset. SAS Enterprise Miner offers many features and functionalities for the business analysis to model the data. The Logistic method, Decision Tree, Random Forest will be applied for the statistic analysis. In **SAS Enterprise Miner**, the data mining process is driven by a process flow diagram. Besides, **SAS Enterprise Miner** allows researchers to easily proceed with the machine learning method on the SAS platform.



According to our design, we draw the flow diagram. After confirming the target variable which is "label", we set 70% training / 30% test in the data partition node. The proportion is the same as we proceed with **KNN** and **QDA**. Since we need to make sure that the researcher could compare these two methods under the same level.



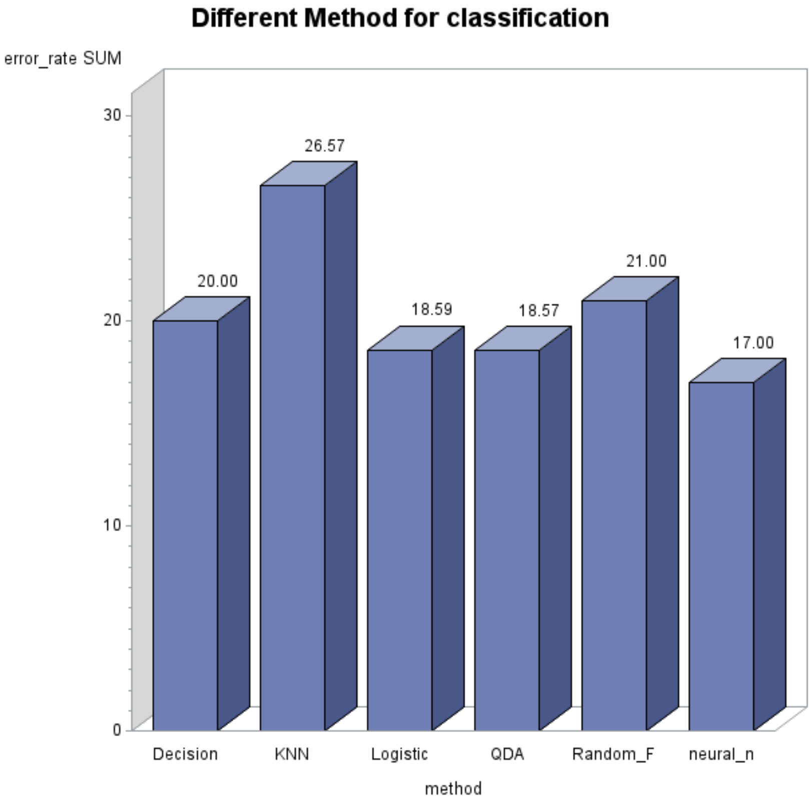
Fit Statistics

Model Node	Model Description	Target Variable	Target Label	Test: Misclassification Rate	Train: Total Degrees of Freedom	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom	Train: Number of Estimated Weights	Train: Akaike's Information Criterion	Train: Schwarz's Bayesian Criterion
Neural	Neural N...	Label	Label	0.170854	700	525	175	175	762.7271	1559.166
HPDMFo...	HP Forest	Label	Label	0.211055	.	.	.	.	.	.
Tree	Decision ...	Label	Label	0.201005	700	.	.	.	.	.
Reg	Regressi...	Label	Label	0.18593	700	647	53	53	579.421	820.6282

From the output, the misclassification for Neural Network finishes the minimum misclassification rate –17 %, the misclassification rate for Random Forest, Decision Tree, and Logistic regression respectively are 21 %, 20%, and 18.6%.

Conclusion:

Based on SAS /STAT and SAS Enterprise Miner, the cluster is used to preprocess the data, after that, we use a lot of classification methods to proceed with the classification procedure, which includes Quadratic Discriminant Analysis, K- Nearest Neighbor, logistic prediction, decision tree, random forest, and neural network. We got the error rate for each method. We compare the error rate for each method, and our goal is to get a higher identification rate as possible as we can.



It won't be surprised for us that the fancy method neural network performs best. KNN's error rate is worse than any method we use in this paper, which indicates the decision boundary is not highly non-linear. That's also explaining why QDA performs relatively better among so many methods. Though not flexible as KNN, QDA can perform better in the presence of a limited number of training observations.

It is always good to compare the results of different analytic techniques; this can either help to confirm results or highlight how different modeling assumptions and assumptions and characteristics uncover insights. We operate SAS to proceed with different methods to classify binary variables and hope it helpful for any researchers to deal with the categorical groups.

REFERENCE:

1. Nadine Hajj, Yara Rizk, and Mariette Awad. 'A Subjectivity Classification Framework for Sports Articles using Cortical Algorithms for Feature Selection,' Springer Neural Computing and Applications, 2018.

2. Yara Rizk, and Mariette Awad 'Syntactic Genetic Algorithm for a Subjectivity Analysis of Sports Articles,' International Conference on Cybernetic Intelligent Systems, Limerick, Ireland, 2012.

3. Parra-Frutos, I. Comput Stat (2013) Testing homogeneity of variances with unequal sample sizes. 28: 1269.doi:10.1007/s00180-012-0353-x.

4. Getting Started with **SAS ® Enterprise Miner 7.1**

5. **"An introduction to statistical learning**", Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani.

6. **"Pattern Classification" ,** Richard O. Duda

Contact Information:

Your comments and questions are valued and encouraged. Contact the author at:

YIDA BAO, Math PHD Candidate, SAS certified advanced programmer

Department of Mathematics and Statistics, Auburn University

E-mail: yzb0010@auburn.edu

Philippe Gaillard, Associate Professor

Department of Mathematics and Statistics, Auburn University

Director of Statistical Consulting Center

E-mail: prg0007@auburn.edu