Time Series Analysis and Machine Learning Techniques on Various Datasets

Regie Felix (CSUSB) and Sophie D'Arcy (Dos Pueblos/Smith College)
Mentor: Nazli Dereli Principal Investigator: Dr. Ambuj Singh
Center of Bio-Image Informatics, ECE, UCSB

Introduction

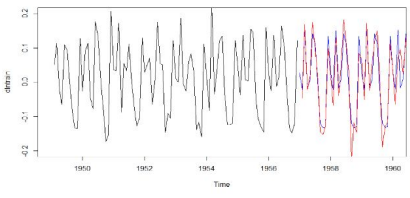
This summer, we were introduced to time series analysis and classification. Using the data-mining software R, we covered topics such as decomposition, classification, transformations, model-fitting, forecasting, as well as various machine learning techniques such as decision trees, ANN, and SVM. We applied these techniques to a various data sets to determine significant trends and predict future observations.

Time Series Analysis

A time series is a sequence of data taken at consistent intervals in time. We explored various time series analysis techniques such as decomposition (breaking up data into seasonal, trend, observed, and residual components), fitting ARIMA (autoregressive integrated moving average) models, and transformations (i.e. learning the differences between log, wavelet, and Fourier transforms and applying them correctly).

Application of Time Series Analysis

Our Application of the Box-Jenkins Approach
Data: AirPassengers (found in R), monthly totals of air passengers from 1949 to 1960.
Step 1: Separate data into training (first two-thirds) and testing (last third).
Step 2: Log transformation to remove nonlinearity in data.
Step 3: Regular and seasonal differencing of data to achieve stationarity and remove seasonality in data.
Step 4: Graph ACF and PACF
Step 5: Use graphs to identify the order (p,d,q) of ARIMA model.
Step 6: Fit ARIMA model.



Step 7: Compare model predictions to log-transformed test set. This is a basic graphical method for checking the accuracy of a model.

Time Series Classification

Classification is a function that categorizes each attribute to a particular class label. A useful tool in classifying is machine learning (algorithms that analyzes data and improves itself for future uses).

Steps in Classifying :
i) Preprocessing the data
ii) Feature Selection and Classification
iii) Evaluation and Improvement

Common Ways to Classify :
Decision Trees - uses nodes and connections to categorize data based on their class label
Artificial Neural Network (ANN) - inputs are brought to hidden units that calculate an output
Support Vector Analysis (SVM) - classifies data by obtaining the best split between different classes

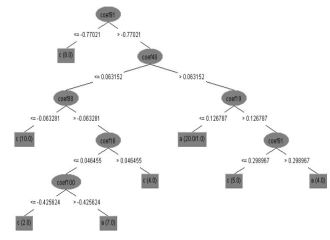
Application of Time Series Classification

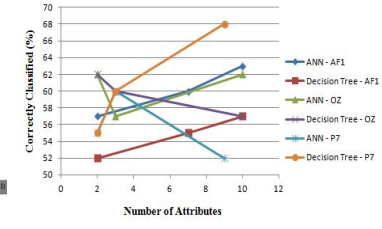
UCI KDD Dataset :
10 Alcoholics and 10 Non-Alcoholics had an EEG (electroencephalography) while being shown one picture, two pictures that match or do not match. The goal of this project was to classify the data and determine whether or not the patient is alcoholic or not based on their EEG results.

Preprocess :
- create directories, remove comments in files
- make data a time series, apply discrete wavelet transformation (DWT)
-create Attribute-Relation File Format (ARFF) file

Analysis :
- select the best attributes
- utilize those features to train the classifier

Evaluation and Improvement:
Both decision trees and ANN had a classification accuracy between 50-70% even after increasing the sample size and adding different sensors. This data set may be classified better in an SVM instead.





Conclusion

Throughout this internship, we have gained a better understanding of data analysis via machine learning and have familiarized ourselves with new programs such as R, Weka, and Java. Time series analysis is one way to examine data recorded in time intervals; it involves determining the trends of previous events in order to predict the future. We did this by processing our data, fitting it into an ARIMA model, and made predictions. Another way to analyze data is time series classification. This is when you create a model that categorizes data based on qualified attributes, which enables you to classify any unknown data. With this experience, we have been inspired to learn even more in the areas of bioinformatics and applied mathematics.

References

- An Introduction to Wavelets by : Amara Graps <http://www.amara.com/ftpstuff/IEEEwavelet.pdf>
- Introduction to Data Mining (2005) by : Pang-Ning Tan, et. al.
- Machine Learning (1997) by : Tom M. Mitchell
- Time Series Analysis and Forecasting by Example (2011) by : Bisgaard, Soren and Murat, Kulahci <http://onlinelibrary.wiley.com/book/10.1002/9781118056943>
- UCI KDD Archive (2005) <http://kdd.ics.uci.edu/>