Madhu Sujan Paudel

Prof. MD Ali, PhD

CS 420 Artificial Intelligence

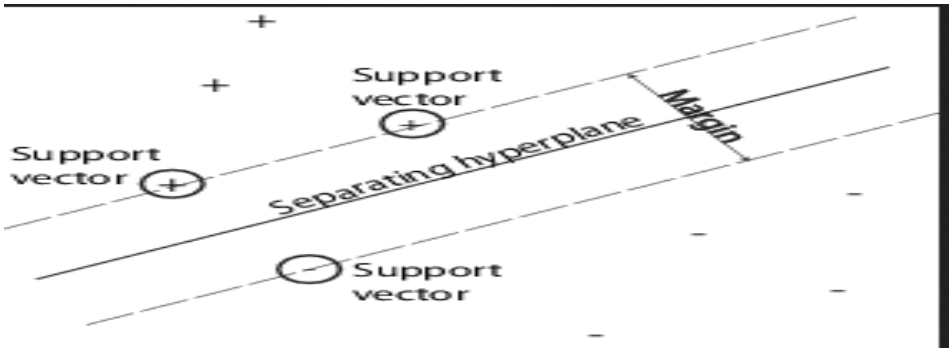
25th March, 2018

**Classification of dataset using SVM and research on effect of data type on accuracy of SVM**

## Introduction

Data mining is the extraction of implicit, previously unknown and potentially useful information from data. The main purpose of data mining is to classify the raw data and uncover relationship in data to predict an outcome. Classification of data is the process of organizing data into categories for its most effective and efficient use. In data classification a given set of data is given, called a training set where each record consists of several fields or attributes.

One way of classifying a dataset using machine learning model is use of SVM (Support Vector Machine). It is capable of performing linear or nonlinear classification, regression, and even outlier detection. And there has been successful use of SVM for classification, facial recognition, bioinformatics, and text categorization. As a classification tools, SVMs using the idea of kernel substitution has become more popular. SVM cannot easily deal with very large data set. So, for our project model composed of small data sets, SVM can be best machine learning model to use.



**Literature Review**

For our purposes, we use SVM for linear classification, face detection, text and hypertext categorization, classification of images, and bioinformatics. Most of my citied work on SVM are based on the research paper by leading computer scientist and data mining, *Chih-Jen Lin*, best known for open source library LIBSVM, an implementation of support vector machines. To understand the working of SVM model, training model and testing sets, I will take reference his scientific papers *LIBSVM: A Library for Support Vector Machines* written by *Chih-Chung Chang and Chih-Jen Lin* thoroughly explains various working models of SVM and its working algorithm. I will be using *LIBSVM*, SVM library, to obtain a model and use the model to predict information of the data set. Also, the book on *Hands-On Machine Learning with Scikit-Learn and Tensor Flow* by *Aurélien Géron* and paper on "High Dimensional Data Classification and Feature Selection Using Support Vector Machines" by Ghaddar et.al. provide a brief insight about SVM and is helpful for training the model.

**Methodology**

The classification of the data involves into separating data into training and testing sites, and each training sets contains one target value and attributes. Using SVM the main goal is to produce model which predicts the target values of the test data using the attributes. There are four common kernels particularly used in SVM model. For our weather dataset, we will be using linear kernel model of SVM.

Given a training set of instance-label pairs $(x_i , y_i )$, i = 1,…..,l where $x_i \in R^n$ and $y \in \{ 1, -1\}^l$ , the support vector machines (SVM)require the solution of the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned}$$

- Transform data to the format of an SVM package
- Randomly try a few kernels and parameters
- Test

We propose that beginners try the following procedure first:

- Transform data to the format of an SVM package
- Conduct simple scaling on the data
- Consider the RBF kernel
- $K(x,y) = e^{-\gamma\|x-y\|2}$
- Use cross-validation to find the best parameter
- C and $\gamma$
- Use the best parameter C and $\gamma$ to train the whole training set.
- Test

**Preliminary Data**

I will be using weather.csv dataset for this particular model. This data is about a person that will be playing or not in certain environment conditions. These conditions are of 5 types: outlook, temperature, humidity, windy and play. All the data set is nominal i.e. symbolic and use 14 instances. The first data set i.e. weather.csv is small and easy to use for this particular model. If possible as part of my research on SVM, I also might try to implement another data classifying heart disease in this program which has 12 instances. The latter data set composed of different fields like heart rate, cholesterol level, blood sugar etc. can be used to classify is a person is in risk of heart disease.

**Statement of Limitation**

SVM is ideal to use for small set of data and other advantages also include:

- Effective in high dimensional spaces and cases where number of dimensions is greater than the number of samples.
- It is also memory efficient due to the use a subset of training points in the decision function (support vectors)
- Different Kernel functions can be specified for the decision function, thus making it versatile to use.

Though SVM is effective to use in dataset classification particularly small dataset, classifying a large dataset can be time consuming for this type of model. And other disadvantages include:

- If the number of features is much greater than the number of samples, it might take a long time in training dataset.
- SVMs do not directly provide probability estimates which are calculated using an expensive five-fold cross-validation.

**Conclusion**

In this paper, we proposed a method to classify the weather.csv data for prediction if player plays a game or not. From a preliminary overview, it looks like using various SVM model has its own advantages and disadvantages. This paper will be focused on extensive research on finding the prediction of event from data set. Also, additional research will be done about classification of heart disease based on patient data field. I believe I will be able to get the desired outcome and find out what type of data is effective to use for SVM model.

References:

Géron, Aurélien. *Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. OReilly, 2017.

Ghaddar, Bissan, and Joe Naoum-Sawaya. "High Dimensional Data Classification and Feature Selection Using Support Vector Machines." *European Journal of Operational Research*, vol. 265, no. 3, Sept. 2017, pp. 993–1004., doi:10.1016/j.ejor.2017.08.040.

Hsu, Chih-Wei. *A Practical Guide to Support Vector Classification.* A Practical Guide to Support Vector Classification, 15 Apr. 2015, www.researchgate.net/publication/272039161_Evaluating_unsupervised_and_supervised_image_classification_methods_for_mapping_cotton_root_rot.

Karol Draszawka, Julian Szymański, "Thresholding strategies for large scale multi-label text classifier", *Human System Interaction (HSI) 2013 The 6th International Conference on*, pp. 350-355, 2013, ISSN 2158-2246.