MACHINE LEARNING – A SURVEY ON SUPPORT VECTOR MACHINE

M.SARASWATHI

M.Sc. [Computer Science] II year

Kanchi Mamunivar Government Institute of Post Graduate Studies and Research

(Autonomous) Puducherry

Abstract

Support vector machine are a specific type of Machine Learning algorithm that are among the most widely used for many statistical learning problems such as spam filtering ,text classification ,handwriting analysis ,face recognition etc .,Support vector machine a new classification method for both linear and non-linear data and support vector machine are a set of related supervised learning methods used for classification and regression .In this paper an analyses is carried out on support vector machine and how it performs on linear data and high dimensional data .It scales relatively well to high dimensional data.

Keywords: SVM, Linear, Classification, Regression, High dimensional data.

1. INTRODUCTION

Machine learning is a branch of Artificial Intelligence (AI) that allows computer systems to learn directly from examples, data and experience. Machinelearning exploresalgorithms. Machine learning is building model for real/high dimensional data/big data**.** The purpose of machine learning is to learn from the data. Machine learning model used for prediction,decision making or solving tasks and solving real world problems.The fundamental principle of Machine learning is to develop cost effective approximate solutions to complex problemsby exploiting the tolerance for imprecision. Machine learning is not a single methodology. Rather, it is a coalition (or) consortium of distinct methodologies.

Machine Learning involves various types of learning. (a)Supervised learning (b) Unsupervised learning (c) Reinforcement learning.A machine can be modeled based on the algorithms. There are different ways an algorithm can model a problem based on its interaction with the experience or past data. Machine Learning algorithms such as Artificial Neural

Network, Naive Bayes, KNN, Decisiontrees, Support Vector Machine, Random Forest etc., and,its mainly concentrate on Support Vector Machine technique.

2.SUPPORT VECTOR MACHINE

A brief history of Support Vector Machine(SVM)

The SV algorithm is a nonlinear generalization of the Generalized Portrait algorithm developed in Russia in the sixties (Vapnik and Lerner 1963, Vapnik and Chervonenkis 1964).As such, it is firmly grounded in the framework of statistical learning theory, or VC theory, which has been developed over the last three decades by Vapnik and Chervonenkis (1974) and Vapnik (1982, 1995)[1].Lately, Support vector machine (SVM), proposed by V. Vapnik in mid-1990, is probably the most popular machine learning algorithm [2]. Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. Classification is used to predict the class labels, so it can be used to categorize various datasets. It is based on the model of applying mapping function on the dependent variable that can be used to predict the independent variable [3].On the other hand, regression can be applied to continuous data instead of discrete data as in classification. Further, it can be classified as linear regression based on a single independent variable whereas polynomial regression is based on multiple hands to achieve a system to produce the program based on the input, statistical analysis and the predicted outcomes[4].

SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables. Several recent studies have reported that the SVM (support vector machines) generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms.

SVM has been employed a wide range of real-world problems such as text categorization, hand-written digit recognition, image recognition, image classification and object detection etc.,SVM is a classifier that uses a hyper plane to separate the data points into classes or groups(Fig 2.1) .The data points are actually vectors in an n- dimensional space .SVM does not use all the data points to (vectors) to make this boundary .It only uses a few of these vectors that support it to separate the points .These points(vectors) that support SVM are called Support Vector.

SVM tries to make a decision boundary in such a way that the separation between the two classes. Hyper plane are decision boundary that helps to classify the data points. It is basically used for two class classification problems. But it can be used for multi-class problems by one-against-rest approach [5].SVM are also works well on linear data and nonlinear data(fig 2.2).When the data has linearly not separable,the SVM uses the Kernel function.If data is linear, a separating hyper plane may be used to divide the data. However, it is often the case that the data is far from linear and the datasets are inseparable. To allow for this kernel are used to non-linearly map the input data to a high-dimensional space [6]. There are different kernel functions are used for non – linear data,

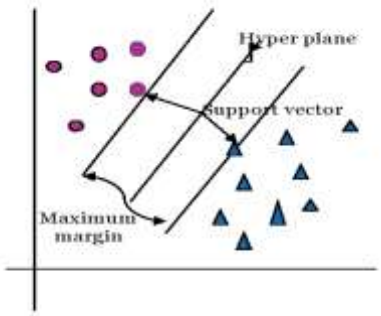

Fig (2.1) SVM

1. Polynomial kernel
2. Gaussian kernel
3. Gaussian Radial Basis Function(RBF)
4. Sigmoid Kernel etc.

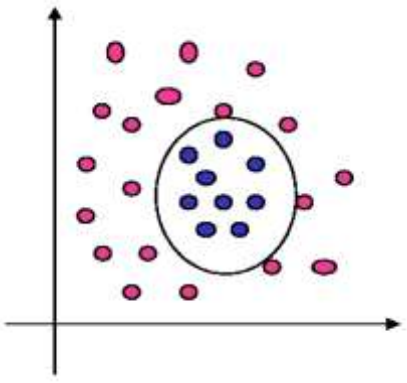

Fig (2.2) nonlinear data

In this paper, an analysis the SVM performance of using the handwritten digit and wine recognition dataset will be implemented on python and Scikit toolkit.

3. SOFTWARE TO BE USED

3.1Anaconda3

Anaconda is an open source distribution for python. It is used for machine learning, datascience, deeplearning. It have more than 300 library for data science .Anaconda comes with a wide variety of tools to easily collect data from various sources using various machine learning and AI algorithm.

In anaconda software,using Jupiter Notebook file. It is feel free to use and accessing. The extension of file was saved .ipynb.

3.2 Python

Python was created in the early 1990‘s by Guido van Rossum.It is a high level programming language. And also it is anopen source programming language.Python is free to download and use it.Python is very easy to learn and understand.

3.3 Python in Machine learning

Using python in machine learning application is working very efficient manner .There are many machine learning applications written in Python. Machine learning is a way to write a logic .so that a machine can learn and solve a particular problem on its own.

3.4 Scikit – Module

Scikit-learn (are also known as sklearn) was initially developed by David Cournapeau as Google summer of code project in 2007.Scikit Open-source Machine learninglibrary for Python.Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. It is on Numpy, Scipy and matplotlib,this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering .Simple and efficient tools for predictive data analysis .A major benefit of this library is the BSD license it’s distributed under. This license allows you to decide whether to upstream your changes without any restriction on commercial use .And, download all of the source code, documentationfrom http://scikit-learn.sourceforge.net.

Installing the scikit - learn library usingpip install –U scikit-learn orconda install scikit-learn.

3.5 Comparing with other tools

Various Machine Learning tools to be used,such that

3.5.1 Weka Tool

Weka is a collection of machine learning algorithms for data mining tasks. These algorithms can either be applied directly to a data set or can be called from your own Java code [7]. Weka provides a simple Command-line explorer which is a simple interface for typing commands .But ,the connection of excel was poor compared to others tools ,it can only handled a small datasets .It relatively slow (java) .Weka does not have the facility to save parameters for scaling to apply to future datasets. Loaded the handwritten digit dataset in weka tool, the error was encountered .Because it cannot handle the large dataset .So, the scikit was better than weka tool.

3.5.1 Orange Tool

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting[8] .The main disadvantages of orange has list of machine learning algorithms is limited .And also it is weak in classical statistics .The installation process is very long .It has limited report capability.

4. DATASETS

The SVM model will be build by using the handwritten digit dataset and wine recognition dataset.

4.1. Optical RecognitionHandwrittenDigits dataDescription

Number of Instances: 5620

Number of Attributes: 64

Attribute Information: 8x8 image of integer pixels in the range 0.16.Downloadingthis data from UCI ML repository,http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

The data set contains images of hand-written digits: 10 classes where each class refers to a digit.

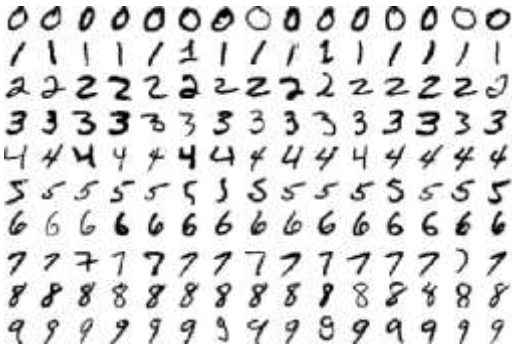

Fig 4.1 Handwritten digit data[9]

The hand written digits(fig 4.1)contains major problem , every person have own style of hand writing .No two person doesn't written the same style of hand writing.It differsbetween size ,width ,orientation . So the problem can be classified the digits due to the major similarity between digits such as 1 and 7 ,5 and 6 ,3 and 8 ,2 and 5 ,2 and 7 etc.,

To develop a model using Support Vector Machine which should correctly classify the handwritten digit from 0 to 9 based on pixel values given as features. Thus,this is a basic classification problem of 10 class.

4.2**Wine Recognition Dataset Description**

Number of Instances: 178 (50 in each of three classes)

Number of Attributes: 13 numeric, predictive attributes and the class.

Attribute Information

- Alcohol
- Malic acid
- Ash
- Alcalinity of ash
- Magnesium
- Total phenols
- Flavonoid's
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
-OD280/OD315ofdiluted wines
- Praline

Class

- class_0
- class_1
- class_2

Downloading this dataset from UCI ML Repository,

https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

Nowadays the sales of wine increase in online compared to outlets. To better understand the buying behavior a stimulus-response model was built to anticipate the sales rate [10].The brands of wine has tremendous improvement and the market is highly competitive. Structural Equation modeling was developed to assess the models of wine brand and it was stated that wine experience is related its brand [11].

These two datasets are used to analyses the performance.

So,this classification can be computed by using SVM algorithm .It builds on SVM model of analyses the performance.

5. Implementation Results

5.1 Handwritten digit Recognition

The main application of machine learning methods over the last decade has determined efficacious in conforming decisive systems which are competing to human performance and which accomplish far improved than manually written classical artificial intelligence systems used in the beginnings of digit recognition technology [12].

Using Scikit library for further implementation .First ,split the data into train and test data of 70-30%.The training data would correctly classified(fig 5.1).

Fig 5.1 Training data

Performing the cross validation of various parameters the given training data is converted into training and testing data.

Calculate the accuracy of linear kernel model ,the accuracy has been predicted in training and testing data is 100.0% and 96.16%.Second implementation for the sigmoid kernel ,the accuracy has been predicted in training and testing data is 10.79% and 10.29%.Third implementation for the Polynomial kernel function ,the accuracy has been predicted in training and testing 100% and 97.21%.Fourth implementation for the RBF kernel function , the accuracy has been predicted in training and testing 100% and 48.46%. Thus the results produced, the SVM technique has given the greater performance in handwritten digit dataset.

5.2Wine Recognition Data

This data will use handle the classification problem .in this problem going to predict the accuracy of wine(three classes). To classify an unlabeled wine according to its characteristic features such as alcohol content, flavor, hue etc.,Among these several features, commonly takenflavonoids for distributions of these classes(Fig5.1 (a)). Before going to implement,split the data into training and testing data.

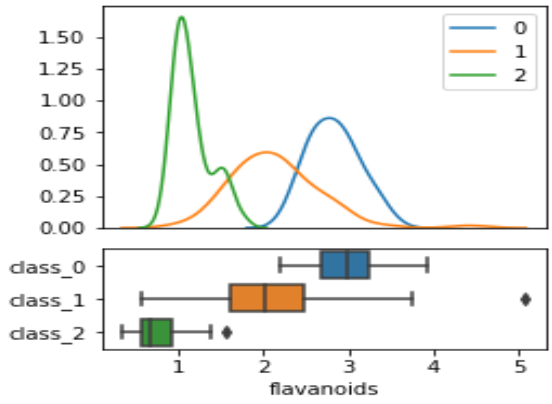

Fig 5.2(a)

The Scikit module has been imported for split the dataset using function train_test_split() of successfully split the data 80-20%.And perform the cross validation of training data .Then ,to find the classification of wine data and see the result of classify data in kernel functions are linear kernel, polynomial and RBF kernel (refer fig 5.2(b)).

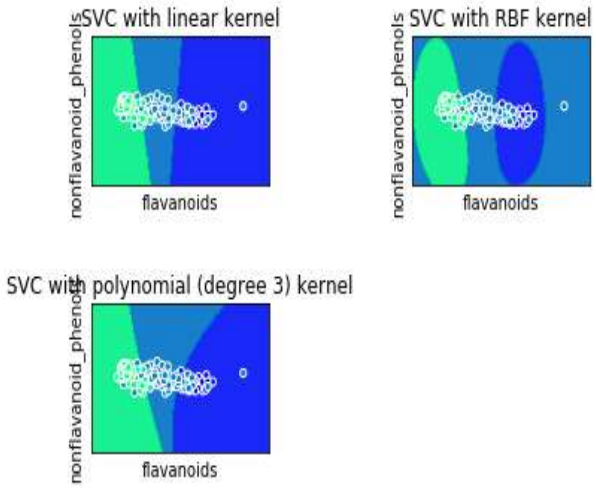

Fig 5.2(b)

From the above performance have been seen thus the SVM technique will give higher performance, correctly classified and predicted the data .It is well suited for wine recognition data.

6. Conclusion and Future work

This paper generalizes the result of Support Vector Machine performance .Mainly ,the Support Vector Machine has been used for classification problems .The main aim of this paper analyses the Support Vector Machine algorithm using with two different datasets, Handwritten digit recognition and Wine recognition dataset ,gives the reliable performance .Some drawback were seen when developing in Support Vector Machine model,in case of using large set of dataset.The vector cannot be classified as well. For future,if the no. of data was large you to choose Neural Network.Various different dataset are chosen for research and applied on Support Vector Machine to see the different type of classifications.

REFERENCES

[1].A. Smola and B. Scholkopf. A tutorial on support vector regression. Statistics and Computing, 14:199–222, 2004.

[2] A Tutorial for Support Vector Machine; Wei-Lun Chao weilunchao760414@gmail.com Graduate Institute of Communication Engineering, National Taiwan University Draft version: Dec. 30, 2011

[3]J. Alcala-Fdez et al. “KEEL: a software tool to assess evolutionary algorithms for data mining problems,” in Soft Computing, 2009, vol. 13, pp. 307-318

[4]R. Mikut, and R. Markus, “Data mining tools," in Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 2011, vol. 1, pp. 431-443.

[5]Arora, S. Bhattacharjee, D. Nasipuri, M. Malik, L. Kundu, M and Basu, D. K.2010. “Performance Comparison of SVM and ANN for Handwritten Devnagari Character Recognition” IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, pp-18-26.

[6] Comparative study of data mining tools-semantic scholar by K Rangra-2014 –cited by 82-Related articles Volume 4, Issue 6, and June 2014.

[7] Comparative study of data mining tools-semantic scholar by K Rangra-2014 –cited by 82-Related articles Volume 4, Issue 6, and June 2014.

[8] .Constanza Bianchi, “Consumer Brand Loyalty in the Chilean Wine Industry”, Journal of Food Products Marketing., vol. 21, no. 4, 2015, pp. 442-460.

[9] D. Veena Parboteeah, D. Christopher Taylor, and A. Nelson Barber, “Exploring impulse purchasing of wine in the online environment”, Journal of Wine Research., vol. 27, no. 4, 2016, pp. 322-339

[12]Seewald, A. K. (2011).On the brittleness of handwritten digit recognition models .ISRN Machine Vision, 2012.