Stochastic Systems Project 2

Topic: Support Vector Machines

a) B.E. Boser, I.M Guyon, V.Vapnik "A Training Algorithm for Optimal Margin Classifiers"

b) C.Cortes and V.Vapnik "Support-vector Networks" Machine Learning vol.20 pp273-297, 1995

Critique Project by

Tharun Battula - 824000197

Texas A&M University

Faculty: Dr. P. R. Kumar

Report compiled in partial fulfillment of the requirements for the

ECEN 755 Stochastic Systems, spring 2016

1 Abstract

The support vector machine is a classic training algorithm to make decision boundary with maximizing margin between two classes. This linear training model developed can be used by wide range nonlinear classification either by increasing dimensionality or by using kernel machines. The solution is expressed in a class separating decision boundary with support hyperplanes which are linear combination of supporting vectors that are closest to support boundary.

The original SVM algorithm was invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis in 1963. Later in 1992, Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik suggested a way to create nonlinear classifiers by applying the kernel trick to maximum-margin hyperplanes through first paper. The other paper proposes standard soft margin was proposed by Corinna Cortes and Vapnik in 1993 and published in 1995. These works are breakthrough in classification and learning problems developing into kernel methods. This training algorithm is carefully developed by the authors of and strongly supported from the experimental results. First paper introduces the support vector machines fundamental formulation in direct and dual space, explaining properties of solution and results. The second paper explains model as support vector networks combining with basic neural network. It adds new concepts of soft margin hyperplane for treatment of for non-separable training data. More mathematical foundation for establishing the kernel machines or convolution of dot product is provided in this paper. The theory, analysis, computation, and experiments on optical character recognition benchmark data are discussed. This report is my understanding of concepts, analysis and observations and critique as a reader.

2 Introduction

2.1.1 Paper1

Author introduces paper by comparing previous work classifiers and requirement for optimal capacity classifiers as large capacity classifiers with lot of parameters yield good performance yet fail to generalize. Paper id described using maximum margin literature and his prior work on support vector patterns [Vapnik

82]. The work proposed here is better than mean square error model, as this classifier's margin offers robustness to unlearned data. Advantages of maximum margin classifier in terms of limited computational accuracy and evaluation is discussed. The core decision function developed for linear model can be applied on different type of nonlinear inputs to form complex model classifier as long as parameters are linear. The dual space interpretation explained in the paper has efficiency improvements for complex models such as kernel models, perceptron, polynomial models for very large training sets Introduction.

2.1.2 Paper 2

In Paper 2 author introduces as Fisher's work on the models of Gaussian distribution function as the first algorithm in pattern recognition history. Explains the quadratic decision function and inference on linear function and the relevance in generalization and lower complexity models. Paper discusses about perceptron's model from Rossenbelts's work along with illustration. It's modeling as piecewise linear models and developments in neural networks is discussed. Author creates the context for his article and explains problems in dealing with higher dimension data through an example. Author's previous work on finding hyperplanes for class separable data is briefly introduced. The paper1's work is explained and its restrictions on handling data with errors is mentioned. Authors arguments are strongly supported logical explanations e.g., formula for expectation of probability of error calculated is mentioned. The second paper approaches multiple classifiers for a network i.e. forming a vector and matrix notation for calculations.

3 Optimal Margin Hyperplane

For given vectors x_i and class labels $y_i = \pm 1$, for robustness on testing data, task is to find a linear parametric classifier with largest margin. This problem can be framed as

$y = \mathrm{sign}(\mathbf{w} \cdot \mathbf{X} + b)$ Or $y_i(\mathbf{w} \cdot \mathbf{X}_i + b) \geq 1$ minimizing the margin can be stated as $\min\limits_{\mathbf{w},b} : \|\mathbf{w}\|$ Explanation on maximizing margin equivalent to minimizing ||w|| is provided.

For this inequality constrained optimization, author uses Lagrange multipliers optimization techniqiues on parameters α_i *and* b . To show that the class boundary is an optimal hyperplane which can be shown linear combination of training patterns with below equations.

The final decision equation can be summarized as

$$\mathbf{w} = \sum_{i \in S} \alpha_i y_i \mathbf{X}_i \qquad y = \operatorname{sign}(\sum_{i \in S} \alpha_i y_i \mathbf{X}_i \cdot \mathbf{X} + b)$$

$$\mathcal{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right)$$

$$\mathbf{w} = \sum_{i \in S} \alpha_i y_i \mathbf{X}_i \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

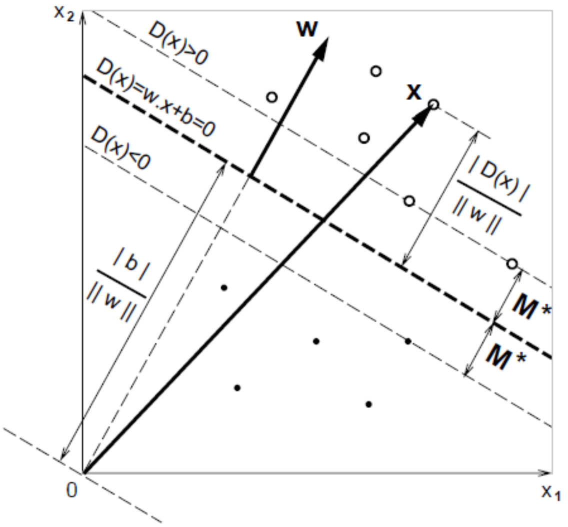


Figure 1 Optimal Hyper Plane disntnguishing classes, support vectors and marginc can be noted

The parameter of α_i , can be solved using Lagrangian dual problem originating in the constrained optimization problem.

$$\text{max.} \;\; W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

The similar explanation is provided in matrix notation specified in the appendix of paper2. Author uses terming Xi as data pattern to indicate kernel function for nonlinear classifiers, reader to be careful at notation understanding in linear domain but expanding to nonlinear problems. Author explains that he applied this method iteratively on chunks of training data on improving the margin gradually. Since this method is for separable data classes, to make it generalized, author mentions method of automatic or manual removal of outliers in the algorithm pertaining to bigger alpha values. This is the restriction of paper -1 but still the more control on outliers keeps this algorithm in a better place than least square estimation.

4 Soft Margins Hyperplane

As the theory is in paper 1 is limited restricted case of separable data case. Author in the Paper 2 develops mathematical formulation on margins for training data with errors. This is mainly done considering the slack for the data patterns where it could deviate but in optimization the constraint for minimization of this deviation is added. The equation becomes

$$\begin{cases} \mathbf{w}^T\mathbf{x}_i + b \geq 1 - \xi_i & y_i = 1 \\ \mathbf{w}^T\mathbf{x}_i + b \leq -1 + \xi_i & y_i = -1 \\ \xi_i \geq 0 & \forall i \end{cases}$$

Minimize $\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i^k$

subject to $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$

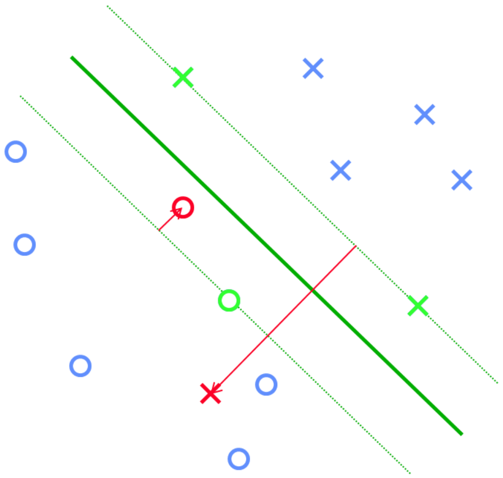


Figure 2 Soft margin Hyper Plane distinguishing classes, errors misclassified information can be noted

The support vectors equation becomes with inequality, for taking care of margin and error vectors. The same Lagrange optimizations equations can be calculated. It converts to dual problem where we need to calculate

$$\text{max. } W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1,j=1}^{n} \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

With the algebraic manipulation paper shows the, the lagrangian dual problem is not a direct quadratic problem as there is a non-polynomial maximum alpha term gets added in the soft margin classifier with more. This is solved in quadratic programming in added extra dimension or convex programming methods. Author presents the deviation equation thoughtfully with **slack variable exponent k**. This gives extra control for modelling the error tolerance for the user. However by changing the exponent of slack variable problem becomes complex in solving equation. For example k is made infinity virtually no error is allowed in minimization problem and for linearly inseparable data a solution may not exist. For simplicity papers considers square of variable as it blends in quadratic equation solution. Paper is thorough in this mathematical formulation.

4.1 Non Linear Boundary

The developed hyperplane algorithm can be extended to bigger set of classes where we could make the data pattern as a functions of data variable allowing the classifier to learn complex models.

4.2 Direct Space to Dual space

Equations for dual space are analyzed for the case of nonlinear or higher dimensional space, with simple replacement of functions as below

$$\mathbf{X}_i = \Phi(\mathbf{x}_i) \qquad \mathbf{X}_i \cdot \mathbf{X} = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) \qquad y = \operatorname{sign}(\sum_{i \in S} \alpha_i y_i \mathbf{X}_i \cdot \mathbf{X} + b) \qquad y = \operatorname{sign}(\sum_{i \in S} \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b)$$

$$\mathbf{X} = \Phi(\mathbf{x})$$

$$w_i = \sum_{k=1}^{p} \alpha_k \varphi_i(\mathbf{x}_k).$$

Based on Mercer's theorem this inner product of $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) = K(\mathbf{x}_i, \mathbf{x})$. The paper provides details on this conversion citing other research and general kernel machine which are in state of the art. Explains dual space conversion here. There is **confusion in notation** here, at equation 6 where they establish correspondence between linear and nonlinear decision functions, paper 1 claims the α_k as dual space parameters. In the next discussions paper uses same α_k as Lagrange multipliers in constructing hyperplanes. For my understanding these are different, if they are related by extending the algebra, they could have been clearly specified. Similarly the paper's usage of terms of "dual space" in nonlinear to linear correspondence and usage in "dual problem" for Lagrange may cause a reader to get confused. If they are related, relation can be established in paper.

The paper extends the discussion to nonlinear boundary case and how the optimizations are done with respect to α_k and b. In this choice of bias b is discussed in two conditions and it is best in terms and when decision boundary is between halfway between two classes. [Vap 82][VC 74]. Some of the valid kernels are Polynomial Splines Gaussian Radial Basis Function Sigmoid (Two layer perceptron)

4.2.1 Convolution of Dot product in Feature space

The second paper thoroughly introduces the previous concept with mathematical notation and theoretical treatment. Feature space is defined and dot product in Hilbertspace is cited with the corresponding contributions and Mercer's theorem are discussed in brief. This overview of the dot product in feature space lays strong foundation for creating nonlinear classifiers and model formation.

For a given kernel function K(xi, xj), the transformation f(.) is given by its Eigen functions from functional analysis, but Eigen functions can be difficult to construct explicitly. People specify the kernel function without worrying about the exact transformation. A kernel function can be thought as simple inner product, as a cosine measure or length similarity, is really a similarity measure between the objects in vector space. With this intuition in nonlinear model design, we can look for a kernel function which measures similarity between data in best way.

4.3 Properties and Features

Paper 1 and Paper 2 discuss the features of proposes solutions. Paper 1 claims how the large margin helps the data insensitivity for small variation of parameters w and α **.** The robustness of the solution is discussed , omitting support patterns leading some other optimum margin is proposed , removing feature with big α , useful in data cleaning applications. The solution is very good as it is not iterative, no complex parameters, does not depend on initial conditions. Leave on out method to estimate upper bound of generalization error. The theoretical understanding on the model for the performance is discussed. With leave one out cross validation method, the worst case linearly independent vector and frequency analysis reported. The analysis in terms of authors famous work from Vapnik-Chervonenkis(VC) dimension is mentioned in the context of frequency of errors and relationships.

4.3.1 Support Vector Network Features

Constructing the decision rules by support vector networks is efficient as the quadratic programming uses the dot product directly in the solution. The curse of dimensionality is reduced by making simplified combined function in evaluating quadratic programming models. The optimal decision function is unique.

The support vector network is a universal machine it offers flexibility in modeling various different types of nonlinear surfaces by observing data and expressing in dot product format according to mercers theorem.

The ability to generalization for support vector networks is discussed. Author presents good logical explanations and formal approach for generalization ability control of generalization ability of learning machine. With respect to below equation the smaller the VC-dimension of set of function in the learning machine, the smaller the confidence interval, but the larger the value of error frequency.

Pr(test error) ≤ Frequency(training error) + confidence Interval

Author gives more detail in VC-dimension analysis and concludes that even for data without any training errors if we control the confidence term (C) in the formation and solving the method to generalize the model we will get better results as robust generalized model, and the support vector network offers such flexibility.

Author offers suggestions to the readers about the **computational considerations** and how model could be optimized. The dot product calculations to only in the dual problem solving. Since the support vector gets constructed only on some data pattern an approximate can be made to omit evaluations which do not contribute. Quadratic optimizations are solved through ellipsoid method in polynomial time, iterative algorithms will be efficient but may not guarantee convergence.

4.4 Experiments and Results

The method is developed for two classes, but extended it ten hypersurfaces one for each class separating from others is formed. Though this idea is generally used and known for its good results, I believe in some case of algorithm with joint decision for all classes together might work better. For this classifier this idea might not be useful. The results are mainly cover for the image data base collection and object recognition challenge, some other data with application could also have been used for showing the power of SVM and its broadness.

4.4.1 Paper 1 – [Vapnik 92]

Author provides strong results and detailed analysis on effects of the proposed methods on optical character recognition data (0-9 digits). Two data base of images is considered, DB1 (1200) is cleaned data. Half is used for training other half is used for evaluation. DB2 (7000) contains images from postal mails without any cleaning.

The best performance is observed for DB2 for fourth order polynomial model compared to results of neural network. The experiments are done for different order of polynomial, RBF kernel, power series with varying parameters, and perceptron model. There is drastic improvement from linear to polynomial order 2, this is explained by VC dimensionality reasoning explained in paper and similar evaluation is supported by power series results comparison. Results are compared in terms of error and average number of supporting vector patterns for hyperplane. DB2 pose challenge in terms of complexity, with orientation of data, scaling. Author cites the corresponding research work for preprocessing. An evaluation of Gaussian smoothing of data in preprocessing and corresponding results is provided marking the importance of preprocessing or cleaning. The results in DB2 still pose problems in terms of invariance over the large data set. Other similarity functions might work better drawing inference as RBF is found as better margin classifier for the illustration with image. In summary, author evaluates most of the multiple combinations model and parameters and supports his arguments well. However only one type of data set is considered, or some other parameters or some nonlinear classifiers are skipped for evaluation.

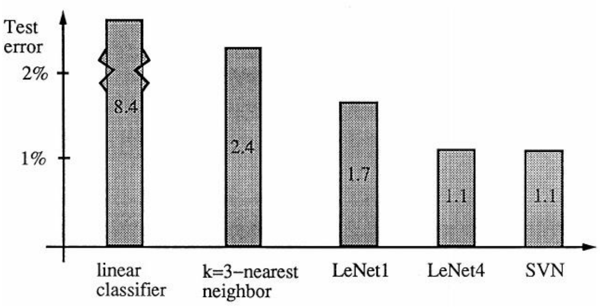
4.4.2 Paper 2 – [Vapnik 95]

Experiments are conducted, in the plane, for digit recognition and over MNIST data base, covering more in different domain for covering wideness of use case of method proposed. This paper differentiates soft margin improvements over the previous work. Experiments in plane 2 cover degree two polynomial showcasing soft margin yield good results. It might have supported if it covered more models.

Models in plane are covered for illustration and visualization purposes along with evaluation and improvement algorithm and tuning during the algorithm creation. Data set of different resolution of images from USPS collection is used – It is Small data base of 7300 for training and 2000 for testing data set. There is another data set of NIST data, covering larger base of 60,000 training and 10,000 test cases are considered. Experimental research done for different polynomial degrees. The current research has evolved and accordingly benchmark comparison is done, such as decision trees and neural networks of 2 and 5 layers are compared. There is lot of learning from the first paper (92), as they continued work, it is reflected in choices made for evaluation such as polynomial order, preprocessing of data choices (centering, deslanting, smoothing). For smoothing Gaussian kernels are used.

The results look very promising in polynomial degree, degree 1 performs the least as a linear classifier, however after degree 2 the performance does not improve much and author mentions this as no overfitting problem. The dimensionality of space increases significantly as feature space increases exponentially. This is good takeaway that increasing polynomial order does not necessarily over fit for SVM. Support vectors count for increasing degree do not change, and the main boundary and core time complexity inside will not change much. The beautiful point to note here is the training time for worst case degree (10^{16} dimensionality) is less than best performing neural network assuring the quality as well.

Due to limited time author could evaluate for MNIST data for only 4th order polynomial. Training error and misclassification rate is reported with some examples of failure case. Author collectively uses analysis and algorithm from other authors to compare results from neural networks (Lenet2 Lenet4) and K nearest neighbor algorithms for all 10 digit classes. The error rate is 1.1% remarkable result compared to other classifiers with prior knowledge on geometry of problem.



In conclusion author opens up for future work suggestions other dot product creation using mercer's theorem and additive improvement of results.

4.5 Summary

The idea of support vector machine and its remarkable contribution is truly inspiring. The two papers are well thought and well written. Arguments are well supported with mathematical ideas and evaluated well. The terminology used is clear and conventional. It is interesting to note that there is no occurrence of word “support vector machine” in the first paper. The paper uses the term of support patterns, whereas second paper develops the concept into support vector networks.

The paper1 talks mainly about minimizing maximum loss and mathematical formulation for forming high margin hyperplanes. The problem of finding linear classifier with (l+1) parameters (dimension of features) is converted into dimension of data size by using LaGrange’s dual equivalent. Though this sounds very complex and counter intuitive to go in higher dimension, it is efficient because of the solution being only dependent on the support vectors which is of very low dimension. Dual problem requires quadratic optimization whereas primal problem has convex optimization. This simplicity in exploiting dimensionality and usage of support patterns makes the support vector machines into elegant classifier.

The paper 2 mainly discusses work is using the first papers work and ideas and makes contribution towards generalizing the solution to data which is non-separable and also generalizing nonlinear classifiers by finding the kernel as convolution of dot products. (Famously known as kernel trick). The soft margin hyperplane make it very powerful in terms of tolerance to error opens up solution for many applications. The second papers evaluation and results are remarkable and insightful.

The key characteristics such as capacity control and ease of changing the implemented decision surface render the support vector network an extremely powerful and universal machine.

4.6 Other References

[1] V. Vapnik, *The Nature of Statistical Learning Theory,* 2nd Ed., Springer, 2000.

[2] L. Bottou et al. Comparison of classifier methods: a case study in handwritten digit recognition. Proceedings of the 12th IAPR International Conference on Pattern Recognition, vol. 2, pp. 77-82.

[3] Summary of SVM - https://en.wikipedia.org/wiki/Support_vector_machine

[4] http://www.cs.cmu.edu/~aarti/Class/10701/slides/Lecture12.pdf

[5] http://www.isn.ucsd.edu/courses/776/slides/kernel-learning.pdf