IMPROVING SUPPORT VECTOR MACHINE CLASSIFICATION ACCURACY BASED ON KERNEL PARAMETERS OPTIMIZATION

Lubna B. Mohammed
Electrical and Computer Engineering,
Ryerson University,
Toronto, Canada
lubnabadri@ryerson.ca

Kaamran Raahemifar
Electrical and Computer Engineering,
Ryerson University,
Toronto, Canada
kraahemi@ee.ryerson.ca

ABSTRACT

Support Vector Machine (SVM) learning algorithm is considered as the most popular classification algorithm. It is a supervised learning technique that is mainly based on the conception of decision planes. These decision planes define decision boundaries which are used to separate a set of objects. It is important to extract the main features of the training datasets. These features can be used to define the separation boundaries. The separation boundaries can also be improved by tuning the parameters of the separation hyperplane. In literature, there are different techniques for feature selection and SVM parameters optimization that can be used to improve classification accuracy. There are a wide variety of applications that use SVM classification algorithm, such as text classification, disease diagnosis, gene analysis, and many others.

The aim of this paper is to investigate the techniques that can be used to improve the classification accuracy of SVM based on kernel parameters optimization. The datasets are collected from different applications; having different number of classes and different number of features. The analysis and comparison among different kernel parameters were implemented on different datasets to study the effect of the number of features, the number of classes, and kernel parameters on the performance of the classification process.

Keywords: Support Vector Machines (SVMs); Classification, Kernel functions; Kernel parameter.

1 INTRODUCTION

The universe of machine learning algorithms is divided into linear and nonlinear techniques. In linear techniques, the aim is to find the best fit hyperplane in the space of the input variables that fit to the distribution of the output variable as closely as possible. While in nonlinear techniques, the aim is to find local linear models by dividing the input space into segmented regions and search for local hyperplanes in boxes of carved-out input space (Coggeshall, Wu, and Coggeshall 2012). There are a number of nonlinear modeling techniques used for classification such as neural networks (Vachkov, Stoyanov, and Christova 2015) and (Charleonnan and Jaiyen 2015), decision trees (Lacey, Deng, and Xie 2014) and (Stankevich, Levashenko, and Zaitseva 2013), additive models (Prince and Aghajanian 2009), support vector machine (SVM) (Dubey and Saxena 2016), (Si Chen et al. 2012), and (Hsieh and Jiang 2011), fuzzy logic systems ("ANFIS: General Description for Modeling Dynamic Objects - IEEE Xplore Document" 2016), and others (Charleonnan and Jaiyen 2015) and (M. Wang et al. 2014).

SVM approach is a straightforward classifier with capabilities for classification in simple linear and non-linear models. In the binary classification problem, SVM identifies a hyperplane that best separates two types of data points in the multidimensional input space. The aim of the hyperplane is to provide the largest error tolerance on both sides of this separation plane. When unseen data is applied to the trained SVM, it should be able to assign the new data to one of the hyperplane sides. In real world applications, the input space cannot be linearly separated. The solution to nonlinearity in the datasets is using kernel functions during the training of SVM algorithm. These kernel functions are useful to convert the input space into higher dimensional feature space. Thus, the importance of SVM that differentiates it from other classification techniques is that it can extend the input space to a higher dimension (Coggeshall, Wu, and Coggeshall 2012).

In multiclass classification, N-class SVMs can be constructed from N SVMs which are combined in a coding scheme to generate a model for multiclass problems (Xu et al. 2010). Commonly, multiclass classification systems are evaluated by computing the accuracy of producing a correct coding using the testing datasets. Successful prediction of the classes is counted as a percentage of the whole datasets.

The performance of the classification process depends on many aspects such as the training algorithm, the feature extraction technique, the kernel function used, and the selected kernel's parameters (Sherin B. M. and Supriya M. H. 2015). The goal of this paper is to study and to analyze the strategies that can be used to improve the performance of SVM classification for different datasets. We study the influence of the selection of SVM parameters such as kernel functions and kernel parameters on the development of SVM classification.

The paper is organized as follows. Section II reviews the methods previously used for improving the classification accuracy based on different kernel parameter optimization techniques. In section III, the concept of SVM algorithm for two-class and multiclass classification is introduced. Also, in this section the kernel functions used in SVM and their parameter optimization are discussed. Experimental set up and results are discussed in section IV. Conclusions and future work are presented in section V.

2 RELATED WORK

There are a number of studies for development of SVM classification accuracy based on kernel parameters optimization. All these studies have aimed to enhance the performance of SVMs by choosing the suitable kernel functions or the suitable parameters of a kernel function applied to different datasets depending on the specific application. In (Li et al. 2010), an automatic parameter selection method has been designed which can be used for choosing the suitable parameter of the kernel function. The experiments involved comparing the proposed method with the traditional k-fold cross-validation technique. The datasets used in the training and testing of multiclass SVM classification are hyperspectral images taken from two sources. The first dataset is the Indian Pine, which has been taken in Indiana and includes forest and agricultural sites. The second dataset has been taken in Washington, DC mall as an urban area, and consists of hyper spectral digital images. Two kernel functions are used in their experiments; they are the polynomial kernel and the radial basis function (RBF) kernel. The aim of applying the classification technique is to generate a model capable of identifying the seven classes in the hyper spectral images. The results show that auto parameter selection method obtains a suitable parameter that provides the best classification accuracy and Kappa accuracy when compared with 5-fold technique. Also, the suggested method is less time consuming than the 5-fold cross-validation applied on the same datasets. The effect of dataset size on the accuracy of selecting the optimal parameter is also discussed. The result shows that as the sample size is increased, the accuracy of selecting the optimal parameter is decreased.

Another approach has been implemented in (L. Wang et al. 2011) based on genetic algorithm (GA). In this research, the classification is used to translate the brain signals of patients with certain medical conditions that damage parts of the brain, such as stroke, into control commands. In such conditions, although body movements might be compromised, and facial muscles might be paralyzed, the patient is usually conscious and the ability to perform eye movements is preserved. The training datasets are collected by instructing the patients to imagine a movement of the left hand, right hand or the foot (three classes).

Electroencephalogram (EEG) signals are recorded for each class using Ag/AgCl electrodes. The modeling of the three classes is performed using SVM classifier with RBF kernel function. It has been proven that optimization of the parameters of kernel function based on GA affects the performance of classifier and improves the accuracy results.

In (T. Wang and Xu 2012), a multiclass kernel polarization (MKP) is proposed which can be used for improvement of multiclass classification by enhancing the feature selection technique. The optimization problem is solved using the gradient-based search method. The training datasets are selected from two sources, Breast Cancer Wisconsin and Iris, during the experiments. The Breast Cancer is a two-class dataset and the Iris is a multiclass dataset. The experiment compares information gain (IG) algorithm with MKP algorithm for feature extraction. Both are suitable for the multiclass scenario. The prediction error of multiclass SVM is used to evaluate the feature subsets selection of IG and MKP. MKP method gives lower test error than traditional IG method. This means that improving the selection of kernel parameters affects the testing error of the selected features that directly enhance the classification accuracy.

Researchers in (Gaspar, Carbonell, and Oliveira 2012) studied the use of searching algorithm to optimize the parameters of SVM kernels. They used simulated annealing optimization technique to analyze and compare the results of classification for different kernels' parameters. The training datasets include nine heterogeneous dataset forms. They are from different application domains that consist of two classes of datasets. This research shows that the best improvements in classification accuracy were found when optimizing the parameters of the RBF, Polynomial and Linear kernels. The results showed poor classification performance when using other kernels.

The research work in (Sherin B. M. and Supriya M. H. 2015) used an algorithm which is inspired from studying the echolocation behavior of bats. The algorithm, named BAT, detects the target and avoids obstacles by using sonar which emits ultrasonic impulses. Bats visualize their surroundings by sensing the variation of the echo and the time variance between their ears. In (Sherin B. M. and Supriya M. H. 2015) the researchers implemented BAT algorithm optimization technique to select SVM kernel and kernel parameters. The aim was to identify and classify noise sources underwater in the ocean. The training datasets have the characteristic acoustic signatures captured by hydrophone. The modeling includes four classes. The performance was computed and compared with particle swarm optimization (PSO) based parameter optimization (Yassi and Moattar 2014). The updating of velocities and positions of the bats are similar to controlling the pace and range of swarming particles. However, BAT algorithm uses intensive local search strategy that does not exist in PSO. The results indicate higher classification accuracy can be obtained using BAT algorithm in the classification of underwater acoustic signatures.

Another optimization algorithm has been applied for SVM kernel parameter selection in the research work in (Zhang et al. 2016). The researchers proposed an improved fruit fly optimization algorithm (IFOA). While the fruit fly optimization algorithm (FOA) is trapped into the local optimum, the IFOA strengthens the search ability. The global and local search is stabilized by separating the fruit fly group into advanced and drawback subgroups. Experimental results show the comparison between the IFOA and other classic optimization algorithms. The datasets used in classification are three benchmarks taken from (UCI); they are the Glass, Segment, and German datasets. They have shown that IFOA is effective and computationally efficient in finding optimal parameters for SVM. The study requires further investigation in problems with larger sets of features. Table 1 compares the six methods used in optimization of SVM kernel parameters based classification problems.

3 SUPPORT VECTOR MACHINES

The classification goal of SVM is to generate a model based on the input datasets which can estimate the output class depending on the input features. The first stage in this process is the separation of the datasets into training set (L) and testing set (S). Each input in the training set includes number of attributes (observed features) and one output Yi called target value or label of the class. The output label belongs to one of two classes of either 1 or -1. The training dataset can be formulated as follows (Hsu, Chang, and Lin 2016):

$$\{X_i, Y_i\} \text{ where } i=1\ldots L,\ Y_i \in \{-1,1\},\ X \in R. \qquad (1)$$

Table 1: Comparing the six methods used in optimization of SVM kernels' parameters.

Algorithm	**Kernel Function**	**Dataset Type**	**Number of Classes**	**Pros**	**Cons**
Automatic Parameter Selection (Li et al. 2010)	Normalized kernel RBF function and Polynomial kernel	Hyperspectral image	Seven Classes	Less cost time and better classification accuracy	Classification accuracy decreases with larger dataset size.
Genetic Algorithm (L. Wang et al. 2011)	RBF kernel function	EEG data	Three classes	improves the classifier accuracy	Tested on EEG only.
Multiclass kernel polarization (MKP) (T. Wang and Xu 2012)	Gaussian ARD kernel	Breast Cancer Wisconsin and Iris	Binary and three classes	Enhance the classification accuracy by improving the selection of features	Tested on Medical data only
Simulated Annealing Optimization Technique (Gaspar, Carbonell, and Oliveira 2012)	Six different kernel functions	Nine datasets from different domains	Two classes	Improvements in classification accuracy when optimizing the parameters of the kernels: RBF, Polynomial and Linear.	The results showed poor classification performance when using other kernels.
The BAT algorithm (Sherin B. M. and Supriya M. H. 2015)	Linear, Polynomial, and Gaussian RBF kernel functions	Acoustic signatures	Four classes	High classification accuracy compared to PSO	Tested on acoustic signatures datasets only
Improved fruit fly optimization algorithm (IFOA) (Zhang et al. 2016)	Polynomial, RBF, and Sigmoid kernel functions	Glass, Segment and German datasets	Binary and multi-class	Effective and computational efficient in searching for the optimum parameters	Requires further in depth researches for problems with larger set of feature

During the training phase, the testing set should not appear to the learning algorithm. The classification technique is performed by finding the maximal margin from both classes. This margin separates the training datasets into two classes and produces a hyperplane at the center of the margin. When applying new datasets, the predicted output will be at one of the two sides of the hyperplane. Figure 1 shows the hyperplane H that separates two classes of data points such that:

$$X_i.w+b \geq +1 \text{ when } Y_i = +1$$

$$X_i.w+b \leq -1 \text{ when } Y_i = -1 \quad (2)$$

where w is the normal vector to the hyperplane and the optimal margin of a separating hyperplane is the shortest distance (d+ + d-).

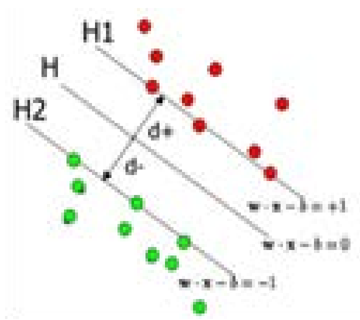

Figure 1. Classification problem where the optimal hyperplane separates two classes.

To compute the maximal margin, the distance between the two planes H1 and H2 is computed as 2/||w||. The points on the planes H1 and H2 are the support vectors that solve the classification problem in SVM algorithm. By introducing slack variable and the tunable parameter C, optimization problem using SVM can be written as the following equation (Gaspar, Carbonell, and Oliveira 2012).

$$\min_{w,\xi} \{ \frac{1}{2} ||w||^2 + C \sum_i \xi_i$$

$$\text{subject to: } y_i (w^\top x_i + b) \geq 1 - \xi_i \;\; ; i=1\ldots N \quad (3)$$

However, in real datasets, the feature space contains no clear separation hyperplane between the two classes. Therefore, we need to find a nonlinear mapping function that can transform the datasets in the feature space to a higher dimension feature space where a separating hyperplane can be found. These non-linear mapping functions are called kernel functions. Depending on the classification problem and the datasets, the optimal hyperplane that can separate the data with the larger margin can be chosen. Maximal margin classifier with good generalization characteristics requires implementing an optimization technique to select the optimal parameters of the kernel functions (Gaspar, Carbonell, and Oliveira 2012). There are many types of kernel functions; however, only the most commonly used ones are listed in Table 2. These kernels (except the linear kernel) include parameters that can be tuned to improve the classification performance. Reaching efficient classification results depends not only on the optimization of the kernel parameters, but also on other factors such as selecting the proper kernel for a specific application domain. Also, selecting the effective features in the training datasets may have significant impact on the result. Finally, the main optimization problem in SVM includes the misclassification penalty that can be adjusted to balance between misclassification error and model generalization (Gaspar, Carbonell, and Oliveira 2012).

Table 2: Most popular kernel functions

Kernel	**Function expression k(x, y)**
Linear	$x^T . y + c$
Polynomial	$(x^T . y + 1).d$
Radial Basis Function (RBF)	$\exp(-\frac{\|\|x-y\|\|^2}{2\sigma^2})$
Sigmoid (MLP)	$\tanh(\alpha x^T y + c)$
Cauchy	$(1+\frac{\|\|x-y\|\|^2}{\sigma})$-1
Log	$-\log(\|\|x-y\|\|^d + c)$

4 EXPERIMENT SETUP

To evaluate the performance of SVM based on the optimization of kernel parameters, a set of experiments was selected from UCI, StatLib, Statlog, and other sources ("LIBSVM Data: Classification, Regression, and Multi-Label," n.d.). All carried out using different types of datasets. These datasets are selected from UCI, StatLib, Statlog, and other sources ("LIBSVM Data: Classification, Regression, and Multi-Label," n.d.). All these data are stored in LIBSVM format. Table 3 shows the datasets used in the experiments. The table includes information about the number of classes, the size of the training and testing datasets, and the number of features in each dataset.

They are selected from different domains with different characteristics to study and analyze the peculiarity of each dataset. SVM classifier was built using Scikit (SciPy Toolkits) that is used to implement machine learning based on Python ("Scikit-Learn: Machine Learning in Python — Scikit-Learn 0.18.1 Documentation" 2016).

The datasets were divided between training and testing phases such that during the evaluation phase, the classifier can predict the class of any unknown input data (datasets that the model have not seen before). On each dataset, SVM classifier was evaluated by measuring the estimator performance and studying the peculiarity of each dataset while changing the kernel and tuning the parameters independently and keeping the penalty weight (C) equal to its default value 1. To evaluate the quality of predictions of a model, a cross-validation tool ("3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.18.1 Documentation" 2016) for evaluating estimator performance is used. Table 4 compares the prediction results of SVM for different types of datasets targeted for different applications. The variant skew of classes is calculated for each dataset. From the results, we can see that depending on the application, the change of kernel may increase or decrease the prediction accuracy. More research is required to examine the peculiarity of the datasets and its effect on the classification of SVM algorithm.

5 CONCLUSIONS AND FUTURE WORK

The approaches that can be used to develop SVM classification performance have been discussed in this paper. The experiment studies the influence of optimization of SVM parameters such as kernel functions and kernel parameters on the development of SVM classification. The results show that the change of kernel function may increase or decrease the performance of SVM classification depending on the type of dataset. The peculiarity of the dataset and its effect on the results of SVM classifications need to be further studied using machine learning techniques to find the optimal kernel and optimal kernels' parameters for a specific domain.

Table 3: Datasets used in building SVM classifier

Data Sets	Skew	Quality of Predictions (Linear Kernel)	Quality of Predictions (RBF Kernel)
w1a	5.606	0.977	0.970
w2a	5.428	0.981	0.970
W3a	5.602	0.983	0.970
W4a	5.579	0.984	0.970
W5a	5.676	0.985	0.971
W6a	5.456	0.986	0.972
W7a	5.513	0.987	0.973
W8a	5.538	0.987	0.974
pendigits	0.026	0.951	0.137
Australian	0.222	0.555	0.555
Diabetes	-0.634	0.651	0.651
Duke breast-cancer	-0.319	0.75	0.75
svmguide1	-0.617	0.957	0.669
svmguide3	1.229	0.170	0.024

REFERENCES

"3.1. Cross-Validation: Evaluating Estimator Performance — Scikit-Learn 0.18.1 Documentation." 2016. Accessed December 23. http://scikit-learn.org/stable/modules/cross_validation.html#cross-validation.

"ANFIS: General Description for Modeling Dynamic Objects - IEEE Xplore Document." 2016. Accessed December 19. doi:10.1109/AICCSA.2015.7507102.

Charleonnan, Anusorn, and Saichon Jaiyen. 2015. "A New Ensemble Model Based on Linear Mapping, Nonlinear Mapping, and Probability Theory for Classification Problems." In *2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 88–92. IEEE. doi:10.1109/JCSSE.2015.7219776.

Coggeshall, Stephen, James Wu, and Stephen Coggeshall. 2012. *Foundations of Predictive Analytics*. *Foundations of Predictive Analytics*.

Dubey, Vimal Kumar, and Amit Kumar Saxena. 2016. "Hybrid Classification Model of Correlation-Based Feature Selection and Support Vector Machine." In *2016 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*, 1–6. IEEE. doi:10.1109/ICCTAC.2016.7567338.

Gaspar, Paulo, Jaime Carbonell, and José Luís Oliveira. 2012. "On the Parameter Optimization of Support

Vector Machines for Binary Classification." *Journal of Integrative Bioinformatics* 9 (3): 201. doi:10.2390/biecoll-jib-2012-201.

Hsieh, Chen-Chiung, and Meng-Kai Jiang. 2011. "A Facial Expression Classification System Based on Active Shape Model and Support Vector Machine." In *2011 International Symposium on Computer Science and Society*, 311–14. IEEE. doi:10.1109/ISCCS.2011.102.

Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. 2016. "A Practical Guide to Support Vector Classification." Accessed December 19. http://www.csie.ntu.edu.tw/.

Lacey, Arron, Jingjing Deng, and Xianghua Xie. 2014. "Protein Classification Using Hidden Markov Models and Randomised Decision Trees." In *2014 7th International Conference on Biomedical Engineering and Informatics*, 659–64. IEEE. doi:10.1109/BMEI.2014.7002856.

Li, Cheng-Hsuan, Chin-Teng Lin, Bor-Chen Kuo, and Hsin-Hua Ho. 2010. "An Automatic Method for Selecting the Parameter of the Normalized Kernel Function to Support Vector Machines." In *2010 International Conference on Technologies and Applications of Artificial Intelligence*, 226–32. IEEE. doi:10.1109/TAAI.2010.46.

"LIBSVM Data: Classification, Regression, and Multi-Label." n.d. https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Prince, Simon J.D., and Jania Aghajanian. 2009. "Gender Classification in Uncontrolled Settings Using Additive Logistic Models." In *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2557–60. IEEE. doi:10.1109/ICIP.2009.5414004.

"Scikit-Learn: Machine Learning in Python — Scikit-Learn 0.18.1 Documentation." 2016. Accessed December 22. http://scikit-learn.org/stable/index.html#.

Sherin B. M., and Supriya M. H. 2015. "Selection and Parameter Optimization of SVM Kernel Function for Underwater Target Classification." In *2015 IEEE Underwater Technology (UT)*, 1–5. IEEE. doi:10.1109/UT.2015.7108260.

Si Chen, Ling Li-na, Yuan Rong-chang, and Sun Long-qing. 2012. "Classification Model of Seed Cotton Grade Based on Least Square Support Vector Machine Regression Method." In *2012 IEEE 6th International Conference on Information and Automation for Sustainability*, 198–202. IEEE. doi:10.1109/ICIAFS.2012.6419904.

Stankevich, Sergey, Vitaly Levashenko, and Elena Zaitseva. 2013. "Fuzzy Decision Tree Model Adaptation to Multi- and Hyperspectral Imagery Supervised Classification." In *The International Conference on Digital Technologies 2013*, 198–202. IEEE. doi:10.1109/DT.2013.6566311.

Vachkov, Gancho, Valentin Stoyanov, and Nikolinka Christova. 2015. "Incremental RBF Network Models for Nonlinear Approximation and Classification." In *2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. IEEE. doi:10.1109/FUZZ-IEEE.2015.7338093.

Wang, Lei, Guizhi Xu, Jiang Wang, Shuo Yang, Lei Guo, and Weili Yan. 2011. "GA-SVM Based Feature Selection and Parameters Optimization for BCI Research." In *2011 Seventh International Conference on Natural Computation*, 580–83. IEEE. doi:10.1109/ICNC.2011.6022083.

Wang, Mei, Yanan Guo, Xiaowei Li, Wei Mo, and Liang Wang. 2014. "Nonlinear Inertia Convergence Classification Model of Online Power." In *2014 International Symposium on Computer, Consumer and Control*, 630–33. IEEE. doi:10.1109/IS3C.2014.170.

Wang, Tinghua, and Shaoyuan Xu. 2012. "Multiclass Feature Selection via Kernel Parameter Optimization." *Proceedings - 2012 5th International Conference on Intelligent Computation Technology and Automation, ICICTA 2012*, no. 1: 213–16. doi:10.1109/ICICTA.2012.60.

Xu, Xiaogang, Songling Wang, Fei Li, Zhengren Wu, and Wei Sun. 2010. “Research on Fault Diagnosis Based on Wavelet Packet Multi-Class Classification SVM.” In *2010 International Conference on Manufacturing Automation*, 164–67. IEEE. doi:10.1109/ICMA.2010.41.

Yassi, Maryam, and Mohammad Hossein Moattar. 2014. “Optimal SVM Parameters Estimation Using Chaotic Accelerated Particle Swarm Optimization for Genetic Data Classification.” In *2014 International Congress on Technology, Communication and Knowledge (ICTCK)*, 1–9. IEEE. doi:10.1109/ICTCK.2014.7033505.

Zhang, Qiantu, Liqing Fang, Leilei Ma, and Yulong Zhao. 2016. “Research on Parameters Optimization of SVM Based on Improved Fruit Fly Optimization Algorithm” 8 (6). doi:10.7763/IJCTE.2016.V8.1096.