International Journal of Current Engineering and Technology
©2021 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Support Vector Machines: Introduction and the Dual Formulation

Rashmi Pathak and Shyam Gupta

Department of Computer Engineering Siddhant College of Engineering, Sudumbre, Pune - 410501

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

Machine Learning means "learning through Machines". We make machine learn and predict the behavior in order to find a solution to a problem. Machine is set to make various predictions based on the learning mechanisms that have been incorporated in them. There are various techniques through which machines can learn. Learning can be supervised, Unsupervised or Semi-supervised. Under these learning schemes we have various classifications. Under Supervised learning we have Classification and Regression. Classification works on continuous values and Regression works on discrete values. Support Vector Machine is an efficient classifier which are mostly sort of linear and comes under supervised method of learning. SVM also find its application in real life for Face Detection, Bioinformatics, Handwriting recognition, image classification and many others. Though this paper we want to review the support vector machines concept and represent it in primal form. We have also stated dual formulation here as can be used to introduce the feature mapping that is mostly used in nonlinearly cases those are separable.

Keywords: *SVM, Classifier, Dual formation, Machine learning, Artificial Intelligence, Feature selection, Embedded methods, Support Vector machines, Mathematical programming.*

Introduction

Support Vector Machines is one of the most effective classifiers among those which are sort of linear. We are also able to handle particular cases of non-linearity using non-linear basis function. Support Vector Machines have been one of the most popular classifiers in recent times. This is because SVMs have clever way to prevent overfitting and we can work with large number of features at a time without getting involved in any complications. Before getting started on SVM, basic hands on in logistic regression is required. Below are the basics for the logistic regression before proceeding with SVMs.

Literature Survey

In this section, let us survey some major contributions towards SVM and its successful.

R. Burbidge et. al., [1] have shown that the support vector machine (SVM) classification algorithm, proves its potential for structure–activity relationship analysis. In a benchmark test, they compared SVM with various machine learning techniques currently used in this field. The classification task involves in predicting the inhibition of dihydrofolate reductase by pyrimidines, using the data obtained from the UCI machine learning repository. Among three tested artificial neural networks, they found that SVM is significantly better than all of these.

Giorgio Valentini [2] have proposed classification methods, based on non-linear SVM with polynomial and Gaussian kernels, and output coding (OC), ensembles of learning machines to separate normal from malignant tissues, to classify different types of lymphoma and to analyse the role of sets of coordinately expressed genes in carcinogenic processes of lymphoid tissues. By using gene expression data from "Lymphochip", he has shown that SVM can correctly separate the tumoural tissues, and OC ensembles can be successfully used to classify different types of lymphoma.

Shutao Li et. al., [3] have applied SVMs by taking DWFT as input for classifying texture, using translation-invariant texture features. They used a fusion scheme based on simple voting among multiple SVMs, each with a different setting of the kernel parameter, to alleviate the problem of selecting a proper value for the kernel parameter in SVM training and performed the experiments on a subset of natural textures from the Brodatz album. They claim that, as compared to the traditional Bayes classier and LVQ, SVMs, in general, produced more accurate classification results.

A training method to increase the efficiency of SVM has been presented by **Yiqiang Zhan** [4] for fast classification without system degradation. Experimental results on real prostate ultrasound images show good performance of their training method in discriminating the prostate tissues from other tissues and they claim that their proposed training method is able to generate more efficient SVMs with better classification abilities. **Yuchun Tang et. al.**, [5] have developed an innovative learning model called granular support vector machines for data classification problems by building just two information granules in the top-down way. The experiment results on three medical binary classification problems show that granular support vector machines proposed in their work provides an interesting new mechanism to address complex classification problems, which are common in medical or biological information processing applications.

Bo-Suk Yang et. al., [6] have presented a novel scheme to detect faulty products at semi-product stage in an automatic mass product line of reciprocating compressors for small-type refrigerators used in family electrical appliances. They presented the classification accuracy using the ANNs, SVM, LVQ, SOFM and SOFM with LVQ (SOFM-LVQ) and found SOFM-LVQ gives high accuracy and are the best techniques for classifying healthy and faulty conditions of small reciprocating compressors. The result shows SOFM with LVQ can improve the classification performance of SOFM but cannot eliminate the classification error, indicated in the concluding remarks.

Rung-Ching Chen [7] has proposed a web page classification method for extraction of feature vectors from both the LSA and WPFS methods by using a SVM based on a weighted voting schema. The LSA classifies semantically related web pages, offering users more complete information. The experimental results show that the anova kernel function yields the best result of these four kernel functions. The LSA-SVM, BPN and WVSVM were then compared and demonstrated that the WVSVM yields better accuracy even with a small data set.

Shu-Xin Du et. al., [8] have developed a Weighted support vector machines for classification where penalty of misclassification for each training sample is different. Two weighted support vector machines, namely weighted C-SVM and VSVM, have been developed for experimenting on breast cancer diagnosis which shows the effectiveness of the proposed methods. They have indicated that, the improvement obtained at the cost of the possible decrease of classification accuracy for the class with large training size and the possible decrease of the total classification accuracy.

Chih-Fong Tsai [9] has presented a two-level stacked generalization scheme composed of three generalizers having color texture of support vector machines (SVMs) for image classification. He has mainly investigated two training strategies based on two-fold cross-validation and non-cross-validation for the proposed classification scheme by evaluating their classification performances, margin of the hyperplane and numbers of support vectors of SVMs. The results show that the noncross-validation training method performs better, having higher correct classification rates, larger margin of the hyperplane and smaller numbers of support vectors.

Chin-Teng Lin et. al., [10] have proposed a support-vectorbased fuzzy neural network (SVFNN) to minimize the training and testing error for better performance. They have developed a learning algorithm consisting of three learning phases is to construct the SVFNN in which the fuzzy rules and membership functions are automatically determined by the clustering principle. To investigate the effectiveness of the proposed SVFNN classification, they applied the corresponding model to various datasets from the UCI Repository and Statlog collection. Experimental results show that the proposed SVFNN for pattern classification can achieve good classification performance with drastically reduced number of fuzzy kernel functions.

Kemal Polat [11] has developed a medical decision making system based on Least Square Support Vector Machine (LSSVM) which was applied on the task of diagnosing breast cancer and the most accurate learning methods was evaluated. He conducted the experiment on the WBCD dataset to diagnose breast cancer in a fully automatic manner using LSSVM. The results strongly suggest that LSSVM can aid in the diagnosis of breast cancer. In his conclusion he has claimed that on the exploration of large data sets the accuracy level may increase.

Sandeep Chaplot et. al., [12] have proposed and implemented a novel approach for classification of MR brain images using wavelet as an input to self-organizing maps and support vector machine .They have noticed classification percentage of more than 94% in case of self organizing maps and 98% in case of support vector machine. They have applied the method only to axial T2-weighted images at a particular depth inside the brain. The same method can be employed for T1- weighted, T2weighted, proton density and other types of MR images. Also they claim that with the help of above approaches, one can develop software for a diagnostic system for the detection of brain disorders like Alzheimer's, Huntington's, Parkinson's diseases etc.

Jin-Hyuk Hong et. al., [13] proposed a novel fingerprint classification method which effectively integrates NBs and OVA SVMs, which produces better accuracy than previously reported in the literature contained in the NIST-4 database. In their proposed method, several popular fingerprint features such as singularities, pseudo codes and the Finger Code were used, and the combination of methods described in the experimental analysis produced better results (90.8% for the five-class classification problem and 94.9% for the four-class classification problem with 1.8% rejection during the feature extraction phase of the Finger Code) than any of the component classifiers.

Fabien Lauer et. al., [14] have proposed different formulations of the optimization problem along with support vector machines (SVMs) for classification task. They have exposed the utility of concerns on the incorporation of prior knowledge into SVMs in their review of the literature. The methods are classified with respect to the categorization into three categories depending on the implementation approach via samples, in the kernel or in the problem formulation. They considered two main types of prior knowledge that can be included by these methods like class invariance and knowledge on the data.

M. Arun Kumar et. al., [15] have enhanced TSVM to least squares TSVM (LSTSVM), which is an immensely simple algorithm for generating linear/nonlinear binary classifiers using two non-parallel hyper planes/ hyper surfaces. In LSTSVM, they have solved the two primal problems of TSVM using proximal SVM (PSVM) idea instead of two dual problems usually solved in TSVM. They have further investigated the application of linear LSTSVM to text categorization using three benchmark text categorization datasets: reuters-21578, ohsumed and 20 Newsgroups (20NG) and based on the Comparison of experimental results, against linear PSVM shows that linear LSTSVM has better generalization on all the three text corpuses considered. Thus they claim that the performance of LSTSVM and PSVM on text categorization can greatly be improved by using it. **Wen Zhang et. al.,** [16] have implemented the multi-word extraction based on the syntactical structure of the noun multiword phrases. In order to use the multi-words for representation, they have developed two strategies based on the different semantic level of the multi-words: the first is the decomposition strategy using general concepts for representation and the second is combination strategy using subtopics of the general concepts for representation. IG method was employed as a scale to remove the multi-word from the feature set to study the robustness of the classification performance. Finally, a series of text classification tasks were carried out with SVM in linear and non-linear kernels, respectively, to analyze the effect of different kernel functions on classification performance.

Urmil B. Parikh et. al., [17] have proposed a new SVM based fault classification algorithm for a series compensated transmission line , which uses samples of three phase currents as well as the zero sequence current as input features to the SVMs for identification of the faulted phase(s). They have tested feasibility of the developed technique on an extensive data set of 25,200 test cases covering a wide range of operating conditions and they claim that accuracy of the proposed classification technique has been found to be at least 98%.

Alice Este et. al., [18] have introduced a new classification technique based on Support Vector Machines which is based on a flow representation that expresses the statistical properties of an application protocol. The classification mechanism presents a

relatively high complexity during the training phase, especially due to the tuning process of the involved configuration parameters. They have applied the proposed technique to three different data sets and almost in all cases, they found the accuracy of the classifier is very good with classification results (True Positives) going over the 90% mark and in general low False Positive rates. **Chih-Hung Wu et al.,** [19] have proposed HGASVM which can help innovators or firms in identifying (classifying) and searching critical documents that can assist their strategic decision making process. The contribution of their study is that the proposed algorithm is an effective patent classification system that can ensure the continuous and systematic use of patent information in a company's decision making processes. By using the HGASVM optimization approach, they have significantly improved the necessary steps, computation time and the times of trial anderror for building an effective SVM classification system.

Takuya Kitamura et. al., [20] have proposed two types of subspace based SVMs (SS_SVMs): subspace-based least squares SVMs (SSLS_SVMs) and subspace-based linear programming SVMs (SSLP_SVMs) where the similarity measure for each class is assumed as the separating hyperplane that separates the associated class with the remaining classes. Also the margin between classes is maximized under the constraints that the similarity measure associated with the class to which a data sample belongs is the largest among all the similarity measures which leads to a linear all-at-once SVM.

Arindam Chaudhuri et. al., [21] have implemented a novel Soft Computing tool viz., FSVM to study the problem of bankruptcy prediction in corporate organizations. The performance of FSVM is illustrated by experimental results which show that they are better capable of extracting useful information from the corporate data than traditional bankruptcy prediction methods. The procedure is easy to implement and is suitable for estimating unique default probability for each firm. The rating estimation done by FSVM is transparent and does not depend on heuristics or expert judgments which imply objectivity and high degree of robustness against user changes.

Jian Qu et. al., [22] have proposed an algorithm which adopts a developed data cleaning algorithm which is based on random subsampling validation and support vector classification to discover potential outliers and determines final outliers based on the measure of misclassification rate and uses the well-known sequential backward selection method to identify irrelevant features. The proposed data processing algorithm is applied to the slurry pump system in which the wear degrees of pump impellers are classified. The results indicate that the proposed data processing algorithm is able to achieve effective classifications and suggested that, it is promising to conduct data cleaning before classifications for a better result.

Nahla Barakat et. al., [23] have reviewed on a historical perspective for the SVM area of research and conceptually groups and analyzes the various techniques. In particular, they have proposed two alternative groupings; the first is based on the SVM (model) components utilized for rule extraction, while the second is based on the rule extraction approach. The analysis is followed by a comparative evaluation of the algorithms' salient features and relative performance as measured by a number of metrics. The paper concludes by highlighting potential research directions such as the need for rule extraction methods in the case of SVM incremental and active learning and other application domains, where special types of SVMs are utilized.

Saibal Dutta et. al., [24] have made an attempt to develop a robust heart beat recognition algorithm that can automatically classify normal/ PVC/other heart beats. The work proposes crosscorrelation as a formidable feature extraction tool, which when coupled with the LS-SVM classifiers, can be efficiently employed as an automated ECG beat classifier. The performance of the proposed scheme has been evaluated by considering several benchmark signals available in MIT/BIH arrhythmia database and the overall performance was found to be as encouraging as very close to 96%.

Hakan Cevikalp [25] have proposed two new clustering algorithms for the partition of data samples for SVM based BHDTs. The proposed methods have two major advantages over the traditional clustering algorithms like they are suitable when SVMs are used as the base classifier and the most commonly employed kmeans clustering algorithm may not be compatible with the SVM classifier as demonstrated in the synthetic database experiments which shows the proposed class based NCuts clustering method seemed more efficient than the proposed sample based clustering method.

Daoliang Li et. al., [26] have developed MSVMs which can classify the foreign fibers in cotton lint in an accurate and quick fashion. Three types of MSVMs, i.e., OAA-DTB MSVM, OAOVB MSVM and OAO-DAG MSVM were tested with the extracted feature vectors using leave-one-out cross validation. The results indicate that, the OAA-DTB MSVM cannot fulfill the requirement of accuracy of online measurement of content of foreign fiber in cotton lint. However, both of the two one-against-one MSVMs can satisfy the classification accuracy requirement and the OAO-DAG MSVM is the fastest.

John Shawe-Taylor et. al., [27] have presented a review of optimization techniques used for training SVMs. They have shown how to instantiate the KKT conditions for SVMs. Along with the introduction of the SVM algorithms, the characterization of effective kernels has also been presented, which is helpful to understand the SVMs with nonlinear classifiers. For the optimization methodologies applied to SVMs, they have reviewed interior point algorithms, chunking and SMO, coordinate descent, active set methods and Newton's method for solving the primal, stochastic sub gradient with projection and cutting plane algorithms. They believe that the optimization techniques introduced in this paper can be applied to other SVM-related research as well.

Yuan-Hai Shao et. al., [28] have proposed a fast TWSVM-type algorithm-the coordinate descent margin-based twin support vector machine (CDMTSVM) to binary classification. At the same time, they have also proposed a novel coordinate descent method to solve the dual problems. Compared to the original TWSVM, their CDMTSVM is not only faster, but also needs less memory storage. They claim that the obtained computational results on UCI and NDC datasets demonstrate that CDMTSVM obtains classification accuracy better than TWSVM with reduced computational effort for both linear and nonlinear kernels. **Ahmad Kazem et. al.,** [29] have developed a novel hybrid model based on a chaotic firefly algorithm and support vector regression for stock market price forecasting. Their contribution of the proposed algorithm is mainly the integration of chaotic motion with a firefly algorithm as a simple and novel optimization method. Compared with genetic algorithm-based SVR (SVR-GA), chaotic genetic algorithm-based SVR (SVR-CGA), firefly-based SVR (SVR-FA), artificial neural networks (ANNs) and adaptive neuro-fuzzy inference systems (ANFIS), the proposed model performs best based on two error measures, namely mean squared error (MSE) and mean absolute percent error (MAPE).

E.A. Zanaty [30] have constructed SVMs and computed its accuracy in data classification. They held a comparison between the SVMs and MLP classifier by considering different sizes of data sets with different attributes. Then, they have compared the results of the SVM algorithm to MLP classifier. The proposed GRPF kernel has achieved the best accuracy, especially with the datasets with many attributes. They believe that it greatly reduces the number of operations in the learning mode. It is well seen for large data sets, where SVM algorithm is usually much quicker. **Chin Heng Wan et. al.,** [31] have implemented a new text document classifier by integrating the K-nearest neighbour (KNN) classification approach with the support vector machine (SVM) training algorithm. The proposed Nearest Neighbour-Support Vector Machine hybrid classification approach is coined as SVMNN which avoids a major problem of the KNN in determining the appropriate value for parameter K in order to guarantee high classification effectiveness. By considering several benchmark text datasets for their experiments, it is shown that the classification accuracy of the SVM-NN approach has low impact on the value of parameter, as compared to the conventional KNN classification model.

Chien-Shun Lo et. al., [32] have proposed a SVM-based classifiers for MR image classification by presenting two sets of experiments: one set consists of computergenerated phantom images and the other set uses real MR images. From the experimental results, they found the correct rate of SVM classification is significantly better than CM in the case of SNR = 5 dB.

Accordingly, they have shown that the SVM has the capability for multi-spectral MR image segmentation and robustness against noise.

Yingjie Tian et. al., [33] have proposed a novel least squares support vector machine, named ε-least squares support vector machine (ε-LSSVM), for binary classification. They claim for improved advantages compared with the plain LSSVM by introducing the ε-insensitive loss function instead of the quadratic loss function into LSSVM. Experimental results on several benchmark datasets show the effectiveness of our method in sparseness, balance performance and classification accuracy. **Zhenning Wu et. al.,** [34] have proposed a PIM-clustering-based FSVM algorithm for classification problems with outliers or noises. The experiments have been conducted on five benchmark datasets to test the generalization performance of the PIM- FSVM algorithm. Their results have shown that the PIM-FSVM algorithm presents more reasonable memberships and is more robust than other methods used in their paper for classification problems with outliers or noises. Second, the computational complexity of the PIM-FSVM algorithm is presented, which is not more complex or even less complex than other methods. **Zhiquan Qi et. al.,** [35] have proposed a new Structural Twin Support Vector Machine (called S-TWSVM), which is sensitive to the structure of the data distribution. They firstly pointed out the shortcomings of the existing algorithms based on structural information and designed a new S-TWSVM algorithm and analysis with its advantages and relationships with other algorithms. Theoretical analysis and all experimental results shown that, the S-TWSVM can more fully exploit this prior structural information to improve the classification accuracy. **Himanshu Rai et. al.,** [36] have introduced a novel and efficient approach for iris feature extraction and recognition. They compared the recognition accuracy with the previous reported approaches for finding better recognition rate than using SVM or Hamming distance alone. They claim for the increase of efficiency, when they used separate feature extraction techniques for SVM and Hamming distance based classifier and proven that the accuracy of the proposed method is excellent for the CASIA as well as for the Chek image database in term of FAR and FRR. **Adil Baykasog lu et. al.,** [37] have presented a comparative analysis of the performances of GA, DE and FA on both static and dynamic multidimensional knapsack problems. One of the important contributions of their study is the development of FA2, which is designed for a more realistic reflection of the behaviors of fireflies and it requires less computational time. Particularly on stationary environments, FA2 obtains near optimal results with a significantly faster convergence capability. Thus, they assumed that FA2 was found more effective compared to GA, DE and FA for the problems studied. **Jui-Sheng Chou et. al.,** [38] have proposed several classifiers that can be applied when using CART, QUEST, C5.0, CHAID, and GASVM (a hybrid approach) to predict dispute propensity. In terms of accuracy, GASVM (89.30%) and C5.0 (83.25%) are the two best classification and regression-based models in predicting project disputes. Among all the models, GASVM provides the highest overall performance measurement score (0.871) considering accuracy, precision, sensitivity, and AUC. Notably, with the exception of GASVM, which was developed by the authors and implemented within a mathematical tool, all models are easily executed via open-source or commercial software. Compared to the baseline models (i.e., C5.0, CHAID, CART, and QUEST) and previous work, GASVM provides 5.89– 12.95% higher classification accuracy.

Zuriani Mustaffa et. al., [39], have reported empirical results that examine the feasibility of eABC-LSSVM in predicting prices of the time series of interest. The performance of their proposed prediction model was evaluated using four statistical metric, namely MAPE, PA, SMAPE and RMSPE and experimented using three different set of data arrangement, in order to choose the best data arrangement for generalization purposes. In addition, the proposed technique also has proven its capability in avoiding premature convergence that finally leads to a good generalization performance.

Youngdae Kim et. al., [40] have proposed an exact indexing solution for the SVM function queries, which is to find top-k results without evaluating the entire database. They first proposed key geometric properties of the kernel space – ranking instability and ordering stability – which is crucial for building indices in the kernel space. Based on them, they developed an index structure iKernel and processing algorithms and then presented clustering techniques in the kernel space to enhance the pruning effectiveness of the index. According to their experiments, iKernel is highly effective overall producing 1–5% of evaluation ratio on large data sets.

Zhen Yang et. al., [41] have proposed a DE-SVC for pattern recognition. Their research results shown that support vector machines and differential evolution can be used as effective and efficient data processing methods for two-stage HCNs synthesized by fine core/shell particles of PMMA/PAN. Compared with genetic algorithm, average iteration steps and training time of the prediction model based on improved DE-SVC have significantly shortened. They have claimed that SVC has more adaptive learning ability and higher prediction accuracy.

A.D. Dileep et. al., [42] have proposed two novel methods to build a better discriminatory IMK-based SVM classifier by considering a set of virtual feature vectors specific to each class depending on the approaches to multiclass classification using SVMs. They proposed a class-wise IMK based SVM for every class by using components of GMM built for a class and a pair wise IMK based SVM for every pair of classes by using components of GMM built for a pair of classes as the set of virtual feature vectors for that pair of classes in the one-againstone approach to multiclass classification. The performance of the SVM-based classifiers using the proposed class-specific IMKs is studied for speech emotion recognition and speaker identification tasks and compared with that of the SVMbased classifiers using the state-ofthe-art dynamic kernels. **Liu Hui** [43] have proposed a DFSVM algorithm for classification and adopted for detection cirrhosis from normal hepatic tissue MR imaging. They have extracted Six GLDM based texture features from medical MRI. The Experimental results shown that DFSVM can select important features and strengthen the specific feature by duplication caused by sampling with replacement in iteration. Their proposed DFSVM is compared with typical feature reduction approaches such as PCA, LDA and Weight-Inform Grain, and also compared with typical classifier ANN. The experiment result shown that DFSVM gets both high sensitivity and high specificity.

Yashar Maali et. al., [44] have proposed a self-advising SVM method for the improvement of the SVM performance by transferring more information from the training phase to the testing phase. This information is generated by using misclassified data in the training phase. Experimental results in their study shown improvement in accuracy, and the F-score and statistical tests reveal the significance of these improvements. They claimed that, by using the misclassified data in the training phase, overtraining can be avoided in their proposed method. SERSC

Shifei Ding et. al., [45] have enhanced LSPTSVM to nonlinear LSPTSVM(NLSPTSVM)for solving nonlinear classification problems efficiently. Similar to LSPTSVM, NLSPTSVM requires just the solution of two systems of linear equations in contrast to PTSVM which requires solving two QPPs. In order to boost performance of NLSPTSVM, they proposed a novel nonlinear recursive algorithm to improve its performance. Experimental results on synthetic two-moon dataset, several UCI datasets and NDC datasets shown that NLSPTSVM has good nonlinear classification capability.

Zhong Yin [46] have proposed two psycho physiological-datadriven classification frameworks for operator functional states (OFS) assessment in safety-critical human machine systems with stable generalization ability. They combined the recursive feature elimination (RFE) and least square support vector machine (LSSVM) and used for binary and multiclass feature selection. Feature selection results have revealed that different dimensions of OFS can be characterized by specific set of psycho physiological features. Performance comparison studies shown that reasonable high and stable classification accuracy of both classification frameworks can be achieved if the RFE procedure is properly implemented and utilized.

PROPOSED METHODOLOGY

An embedded method for feature selection using SVMs is proposed in this section. The reasoning behind this approach is that we can improve classification performance by eliminating the features that effect on the generalization of the classifier by optimizing the kernel function. The main idea is to penalize the use of features in the dual formulation of SVMs using a S. Maldonado et al. / Information Sciences 181 (2011) 115–128 117 gradient descent approximation for kernel optimization and feature elimination. The proposed method attempts to find the best suitable RBF-type kernel function for each problem with a minimal dimension by combining the parameters of generalization (using the 2-norm), goodness of fit and feature selection (using a 0-“norm” approximation).

A. Notation and preliminaries

For this approach we use the anisotropic Gaussian kernel:

$$k(x_{!} \quad x_s) = exp\left(-\sum_{j=1}^{n} \frac{(x_{ij}-x_{sj})}{2}\right)_{\&'''} \qquad (1)$$

in which the kernel shape is given by σ = [σ_{+}, $\sigma_{\&}$, $\sigma_{,,}$ … $\sigma_{(}$] n being the number of variables. Considering different widths in different dimensions, the importance of feature j is determined by ($\sigma_{)}$). For example, if $\sigma_{)}$ is very large, the particular variable j loses its importance since its contribution to the kernel function’s exponent will be close to zero. On the other hand, if $\sigma_{)}$ is very small then the contribution of the variable j to the exponent will be large thus increasing its importance. We propose the following change of variables f, in order to convert the feature selection process into a minimization problem: v = 2 ⁺ , ⁺ , ⁺ … , ⁺ 3 , which leads to:

’% ’\$ ’&   ”

\$

$K(X!, X", v) = exp$ (− -________|/∗1!%&/∗1#|- +
(2)

Where ∗ denotes the componentwise vector product operator, which is defined as $a * b$ = ($a_{+}b_{+}$ … . , $a_{(}b_{(}$)

B. KP-SVM Algorithms

The proposed approach (kernel-penalized SVM) incorporates feature selection in the dual formulation of SVMs. The formulation includes a penalization function $f(v)$ based on the 0-“norm” approximation and modifying the Gaussian kernel using a (anisotropic) width vector m as a decision variable. The feature penalization should be negative since the dual SVM is a maximization problem. The following embedded formulation of SVMs for feature selection is initially proposed:

$$\max_{\propto,v} \sum_{i=1}^{m} \propto_i - \frac{1}{2}\sum_{i,s=1}^{m} \propto_i \propto_s y_i y_s K(X_i S_s, V) - C_2 f(v), \quad (3)$$

subject to

$\sum^{4}{}_{!*+} \propto_{!} y_{!} = 0$ (4)

$0 \leq \propto_{!} \leq C, i = 1, \dots . m,$

$v_{)} \geq 0, j = 1, \dots . n.$

Notice that the values of v are always considered to be positive, in contrast to the weight vector w in formulation, since it is desirable that the kernel widths be positive values. Considering the 0-
"norm" approximation described in
$!|\nu|! \approx e^{"}(e - \exp(-\beta|\nu|))$ (5)
!
and since L$\nu_{)}$L = $\nu_{)}\forall_{)}$|, it is not necessary to use the 1-norm in the approximation. Along the lines of formula, the following feature penalization function is proposed, where the approximation parameter β is also considered. We also try different values for this parameter to study the influence of β in the final solution

$f(v) = e^{5}(e - \exp(\beta v)) = \sum^{(}{}_{)^{*}+}[1 - \exp\mathrm{R}{-\beta v_{)}}\mathrm{S}]$ (6)

Since the formulation is non-convex, we develop an iterative algorithm as an approximation for this formulation. We propose a 2-step methodology: first the traditional dual formulation of SVM for a fixed (isotropic) kernel width m is solved:
$\max\limits_{\propto} \sum_{i=1}^{m} \propto_i - \frac{1}{2}\sum_{i,,s=1}^{m} \propto_i \propto_S \, y_i y_S K(x_{!}, x_{"}, v)$, (7)
subject to
$\sum^{4}{}_{!^{*}+} \propto_{!} y_{!} = 0,$ (8)

$0 \leq \propto_{!} \leq C, i = 1, \dots . m,$

In the second step the algorithm solves, for a given solution a, the following non-linear formulation:

min/ $F(v) = \sum^{4}{}_{!,"^{*}+} \propto_{!}\propto_{6} \, y_{!}y_{6}K(x_{!}, x_{"}, v) + C_{\&}f(v)$, (9)
subject to
$v_{)} \geq 0, j = 1, \dots . n.$
The goal of formulation is to find a sparse solution, making zero as many components of ν as possible. We propose an iterative algorithm that updates the anisotropic kernel variable ν, using the gradient of the objective function, and eliminates the features that are close to zero (below a given threshold ϵ). The algorithm kernel width updating and feature elimination follows:

```
1.  Start with $v = v_{!}e$;
2.  cont = true; t=0;
3.  While(cont==true) do
4.  Train SVM( Step 1) for a given $v$;
5.  $v^{t+1} = v^{t} - \gamma\Delta F(v^{t})$;
6.  for all ( $v_j^{t+1} < \varepsilon$ ) do
7.  $V_j^{t+1} = 0$;
8.  end for
9.  If ( $v^{t+1} == v^{t}$) then
10. cont = false;
11. end if
12. t=t+1 ;
13. end while;
```

In the fourth line the algorithm adjusts the kernel variables by using the gradient descent procedure, incorporating a parameter γ, which has to be sufficiently small to avoid negative widths, especially at the first iterations. In this step the algorithm computes the gradient of the objective function in formulation for a given solution of SVMs $\propto$, obtained by training an SVM classifier using formulation.

$\Delta)F(v) = \sum 4!,"^{*}+ \, v)\mathrm{R}x!) - x")\mathrm{S}\& \propto!\propto 6 \, y!y6K(x!, x", v) + C_{\&}\beta exp(-\beta v)$ (10)

The lines 6, 7, and 8 of the algorithm represent the feature elimination step. When a kernel variable ν_7 in iteration t + 1 is below a threshold ϵ, we consider this feature irrelevant and we eliminate this feature by setting $\nu_7 = 0$. This variable will not be included in further iterations of the algorithm. The threshold ϵ has to be sufficiently small to avoid the elimination of relevant variables in the first iterations of the algorithm.
The lines 9, 10, and 11 of the algorithm represent the stopping criterion, which is reached when $||v^{89+} - v^{8}||_{+}$. It is also possible to monitor the convergence by considering the measure kmt+1 mt k1, which represents the variation of the kernel width between two consecutive iterations t and t + 1, Notice that the 1-norm penalty (LASSO penalty) can also be used instead of the 0-norm approximation., the 1-norm by itself can lead to good feature selection and classification results, without considering the 2-norm for robustness.

Result and Discussions

In this paper we present a novel embedded method for feature selection using SVMs. A comparison with other feature selection techniques shows the advantages of our approach:

- Empirically, KP-SVM outperforms other filter and wrapper techniques, based on its ability to adjust better to the data by optimizing the kernel function and simultaneously selecting an optimal feature subset for classification.
- Unlike most feature selection methods, it is not necessary to set the feature number to be selected a priori: KP-SVM deter- mines the optimal number of features according to the regularization parameter, C2.
- Any suitable kernel function can be used instead of the Gaussian.
- It can easily be generalized to variations of SVM, such as SV Regression and Multi-class SVM.

Even if several parameters should be tuned to obtain the final solution, the computational effort can be reduced since the feature subset is obtained automatically, reducing computational time by avoiding a further validation step in order to find the adequate number of ranked features. The proposed model selection methodology also reduces computational effort for finding the parameters.

KP-SVM attempts to find an optimal subset of features for classification. If, however, the goal of feature selection is to find a subset of a fixed size r among all n features, KP-SVM can be modified to accomplish this goal as well. The main idea is to con- struct a feature ranking with the removed features. Earlier removed variables have a lower rank than later removed variables. For features that have been eliminated as a batch in the same iteration, the ones with higher last value of mj have a better rank.

Our algorithm relies on a non-linear optimization problem, which is computationally treatable but expensive if the num- ber of input features is large. We could improve its performance by applying filter methods for feature selection before run- ning KP-SVM [47,48] or by developing hybrid models [49]. This way we can identify and remove irrelevant features at low cost. In several Credit Scoring projects we have performed for

Chilean financial institutions we used univariate analysis (ChiSquare Test for categorical features and the Kolmogorov– Smirnov Test for continuous ones) as a first filter for features selection with excellent results [50].

Future work can be done in several directions. First, it would be interesting to use the proposed method in combination with variations of SVM, such as Regression [51] or Multi-class. Also interesting would be the application of this approach with other kernel functions like polynomial kernel or with weighted support vector machines to compensate for the unde- sirable effects caused by unbalanced data sets in model construction; an issue which occurs for example in the domains of credit scoring and fraud detection.

There are many terms that are dig out when we talk about SVMs such as kernel functions, non-liner SVMs. Naïve Bayes is also used for similar sets of applications but once you are confident on your training data Support

References

[1]. R. Burbidge, M. Trotter, B. Buxton and S. Holden, "Drug design by machine learning: support vector machines for pharmaceutical data analysis", Computers and Chemistry, vol. 26, (2001), pp. 5-14.

[2]. G. Valentini, "Gene expression data analysis of human lymphoma using support vector machines and output coding ensembles", Artificial Intelligence in Medicine, vol. 26, (2002), pp. 281–304.

[3]. L. Shutao, J. T. Kwok, H. Zhua and Y. Wang, "Texture classication using the support vector machines", Pattern Recognition, vol. 36, (2003), pp. 2883-2893.

[4]. Y. Zhan and D. Shen, "Design efficient support vector machine for fast classification", Pattern Recognition, vol. 38, (2005), pp. 157-161.

[5]. Y. Tang, B. Jin, Y. Sun and Y. Zhang, "Granular support vector machines for medical binary classification problems", Proceedings of 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB '04, (2004), pp. 73-78.

[6]. B. Yang, W. Hwang, D. Kim and A. Tan, "Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines", Mechanical Systems and Signal Processing, vol. 19, (2005), pp. 371–390.

[7]. R. Chen and C. Hsieh, "Web page classification based on a support vector machine using a weighted vote schema", Expert Systems with Applications, vol. 31, (2006), pp. 427-435.

[8]. S. Du and S. Chen, "Weighted support vector machine for classification", IEEE International Conference on Systems, Man and Cybernetics, vol. 4, (2005), pp. 3866-3871.

[9]. C. Tsai, "Training support vector machines based on stacked generalization for image classification", Neurocomputing, vol. 64, (2005), pp. 497–503.

[10]. C. Lin, C. Yeh, S. Liang, J. Chung and N. Kumar, "Support-vectorbased fuzzy neural network for pattern classification", IEEE Transactions on Fuzzy Systems, vol. 14, no. 1, (2006), pp. 31-41.

[11]. K. Polat and S. Gune, "Breast cancer diagnosis using least square support vector machine", Digital Signal Processing, vol. 17, (2007), pp. 694–701.

[12]. S. Chaplot, L. M. Patnaik and N. R. Jagannathan, "Classification of magnetic resonance brain images using wavelets as input to support vector machine and neural network", Biomedical Signal Processing and Control, vol. 1, (2006), pp. 86–92.

[13]. J. Hong, J. Min, U. K. Cho and S. B. Cho, "Fingerprint classification using one-vs-all support vector machines dynamically ordered with naïve Bayes classifiers", Pattern Recognition, vol. 41, (2008), pp. 662671.

[14]. F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector machines for classification: A review", Neurocomputing, vol. 71, (2008), pp. 1578–1594.

[15]. A. M. Kumar and M. Gopal, "Least squares twin support vector machines for pattern classification", Expert Systems with Applications, vol. 36, (2009), pp. 7535–7543.

[16]. W. Zhang, T. Yoshida and X. Tang, "Text classification based on multi-word with support vector machine", Knowledge-Based Systems, vol. 21, (2008), pp. 879–886.

[17]. U. B. Parikh, B. Das and R. Maheshwari, "Fault classification technique for series compensated transmission line using support vector machine", Electrical Power and Energy Systems, vol. 32, (2010), pp. 629–636.

[18]. A. Este, F. Gringoli and L. Salgarelli, "Support Vector Machines for TCP traffic classification", Computer Networks, vol. 53, (2009), pp. 2476–2490.

[19]. C. Wu, Y. Ken and T. Huang, "Patent classification system using a new hybrid genetic algorithm support vector machine", Applied Soft Computing, vol. 10, (2010), pp. 1164–1177.

[20]. T. Kitamura, S. Takeuchi, S. Abe and K. Fukui, "Subspace-based support vector machines for pattern classification", Neural Networks, vol. 22, (2009), pp. 558-567.

[21]. A. Chaudhuri and K. De, "Fuzzy Support Vector Machine for bankruptcy prediction", Applied Soft Computing, vol. 11, (2011), pp. 2472–2486.

[22]. J. Qu and M. J. Zuo, "Support vector machine based data processing algorithm for wear degree classification of slurry pump systems", Measurement, vol. 43, (2010), pp. 781–791.

[23]. N. Barakat and A. P. Bradley, "Rule extraction from support vector machines: A review", Neurocomputing, vol. 74, (2010), pp. 178–190.

[24]. S. Dutta, A. Chatterjee and S. Munshi, "Correlation technique and least square support vector machine combine for frequency domain based ECG beat classification", Medical Engineering & Physics, vol. 32, (2010), pp. 1161–1169.

[25]. H. Cevikalp, "New clustering algorithms for the support vector machine based hierarchical classification", Pattern Recognition Letters, vol. 31, (2010), pp. 1285–1291.

[26]. D. Li, W. Yang and S. Wang, "Classification of foreign fibers in cotton lint using machine vision and multi-class support vector machine", Computers and Electronics in Agriculture, vol. 74, (2010), pp. 274–279.

[27]. J. Taylor and S. Sun, "A review of optimization methodologies in support vector machines", Neurocomputing, vol. 74, (2011), pp. 36093618.

[28]. Y. H. Shao and Y. N. Deng, "A coordinate descent margin based-twin support vector machine for classification", Neural Networks, vol. 25, (2012), pp. 114–121.

[29]. A. Kazem, E. Sharifi, F. K. Hussain, M. Saberi and O. K. Hussain, "Support vector regression with chaos-based firefly algorithm for stock market price forecasting", Applied Soft Computing, vol. 13, (2013), pp. 947–958.

[30]. E. A. Zanaty, "Support Vector Machines (SVMs) versus Multilayer Perception (MLP) in data classification", Egyptian Informatics Journal, vol. 13, (2012), pp. 177–183.

[31]. C. H. Wan, L. H. Lee, R. Rajkumar and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine", Expert Systems with Applications, vol. 39, (2012), pp. 11880–11888.

[32]. C. S. Lo and C. M. Wang, "Support vector machine for breast MR image classification", Computers and Mathematics with Applications, vol. 64, (2012), pp. 1153–1162.

[33]. Y. Tian, X. Ju, Z. Qi and Y. Shi, "Efficient sparse least squares support vector machines for pattern classification", Computers and Mathematics with Applications, vol. 66, (2013), pp. 1935-1947.

[34]. Z. Wu, H. Zhang and J. Liu, "A fuzzy support vector machine algorithm for classification based on a novel PIM fuzzy clustering method", Neurocomputing, vol. 125, (2014), pp. 119–124.

[35]. Z. Qi, Y. Tian and Y. Shi, "Structural twin support vector machine for classification", Knowledge-Based Systems, vol. 43, (2013), pp. 74–81.

[36]. H. Rai and A. Yadav, "Iris recognition using combined support vector machine and Hamming distance approach", Expert Systems with Applications, vol. 41, (2014), pp. 588–593.

[37]. A. Baykasoglu and F. B. Ozsoydan, "An improved firefly algorithm for solving dynamic multidimensional knapsack problems", Expert Systems with Applications, vol. 41, (2014), pp. 3712– 3725.

[38]. J. S. Chou, M. Y. Cheng, Y. W. Wu and A. D. Pham, "Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification", Expert Systems with Applications, vol. 41, (2014), pp. 3955–3964.

[39]. Z. Mustaffa, Y. Yusof and S. S. Kamaruddin, "Enhanced artificial bee colony for training least squares supportvector machines in commodity price forecasting", Journal of Computational Science, vol. 5, no. 2, (2014) March, pp. 196–205.

[40]. Y. Kim, I. Ko, W. S. Han and H. Yu, "iKernel: Exact indexing for support vector machines", Information Sciences, vol. 257, (2014), pp. 32–53.

[41]. Z. Yang, Q. Yu, W. Dong, X. Gu, W. Qiao and X. Liang, "Structure control classification and optimization model of hollow carbon nanosphere core polymer particle based on improved differential evolution support vector machine", Applied Mathematical Modelling, vol. 37, (2013), pp. 7442-7451.

[42]. A. D. Dileep and S. C. Chandra, "Class-specific GMM based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines", Speech Communication, vol. 57, (2014), pp. 126–143.

[43]. L. Hui, G. D. Mei and L. Xiang, "Cirrhosis classification based on MRI with duplicative-feature support vector machine (DFSVM)",

[44]. Biomedical Signal Processing and Control, vol. 8, (2013), pp. 346– 353.

[45]. Y. Maali and A. A. Jumaily, "Self-advising support vector machine", Knowledge-Based Systems, vol. 52, (2013), pp. 214–222.

[46]. S. Ding and X. Hua, "Recursive least squares projection twin support vector machines for nonlinear classification", Neurocomputing, (2014).

[47]. Z. Yin and J. Zhang, "Operator functional state classification using least-square support vector machine basedrecursive feature elimination technique", Computer methods and programs in biomedicine, vol. 113, (2014), pp. 101-115.

[48]. Y. Liu, Y.F. Zheng, FS-SFS: A novel feature selection method for support vector machines, Pattern Recognition 39 (2006) 1333–1345.

[49]. Ö. Uncu, I.B. Türksen, A novel feature selection approach: combining feature wrappers and filters, Information Sciences 177 (2007) 449– 466.

[50]. A. Unler, A. Murat, R.B. Chinnam, mr2PSO: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machineclassification, Information Sciences, in press, doi:10.1016/j.ins.2010.05.037.

[51]. S. Maldonado, R. Weber, A wrapper method for feature selection using support vectormachines, Information Sciences 179 (13) (2009) 2208–2217.

[52]. S. Maldonado, R. Weber, Feature selection for support vector regression via kernel penalization, in: Proceedings of the 2010 International Joint Conference on Neural Networks, Barcelona, Spain, 2010, pp. 1973–1979.