High Dimensional Data Classification and Feature Selection using Support Vector Machines

Bissan Ghaddar*
Ivey Business School, 1255 Western Road, London, ON Canada, N6G 0N1, bghaddar@uwaterloo.ca
Joe Naoum-Sawaya
Ivey Business School, 1255 Western Road, London, ON Canada, N6G 0N1, jnaoumsa@uwaterloo.ca

In many big-data systems, large amounts of information are recorded and stored for analytics purposes. Often however, this vast amount of information does not offer additional benefits for optimal decision making, but may rather be complicating and too costly for collection, storage, and processing. For instance, tumor classification using high-throughput microarray data is challenging due to the presence of a large number of noisy features that do not contribute to the reduction of classification errors. For such problems, the general aim is to find a limited number of genes that highly differentiate among the classes. Thus in this paper, we address a specific class of machine learning, namely the problem of feature selection within support vector machine classification that deals with finding an accurate binary classifier that uses a minimal number of features. We introduce a new approach based on iteratively adjusting a bound on the l_1–$norm$ of the classifier vector in order to force the number of selected features to converge towards the desired maximum limit. We analyze two real-life classification problems with high dimensional features. The first case is the medical diagnosis of tumors based on microarray data where we present a generic approach for cancer classification based on gene expression. The second case deals with sentiment classification of on-line reviews from Amazon, Yelp, and IMDb. The results show that the proposed classification and feature selection approach is simple, computationally tractable, and achieves low error rates which are key for the construction of advanced decision-support systems.

Key words: Analytics, Classification, Support Vector Machines, Feature Selection, Machine Learning

1. Introduction

Data driven decision support systems are nowadays an integral part of many businesses. Particularly, with the recent widespread of cognitive computing, significant effort is being made to push the frontier by which computers can assist humans in deriving insights and in making decisions through the analysis of a multitude of increasingly complex data streams. For example, financial institutions heavily rely on decision support systems to decide on whether to grant a loan to a customer or not. Such a decision is basically a binary classification that uses data analysis to identify

* Corresponding Author

the features of safe and risky loans (Baesens et al. 2003). Support Vector Machines (SVMs) is a class of data driven machine learning approach that deals with predictive binary classification, i.e. the assignment of class labels to unlabeled data. Using a large set of observations with known labels (training set), SVM finds a maximum margin function that separates the observations into two classes where each observation is a point in a multidimensional space of feature measurements. New unlabeled data are then assigned a class based on their geometric position relative to the classifier function. Given the vast amount of complex features that modern systems use, finding the classifier function often requires the simplification of the features space by identifying the dimensions that have the most distinguishing power. It is therefore essential to jointly optimize the feature selection and the classification in order to ensure the best performance of the decision support system.

Feature selection and classification have several areas of applications such as medical diagnosis which is traditionally based on the diagnosis by a physician of a patient's signs and symptoms (Ustun and Rudin 2016). Recent technological innovations are aiming to assist the physician in making highly accurate diagnosis particularly for challenging cases by relying on methodologies that are capable to integrate and evaluate large amounts of data and information such as expression genes. One example is IBM's Watson computer system which can identify lung, prostate, and breast cancers by analyzing a vast array of data and medical information (Steadman 2013, Friedman 2015). Another area of application is sentiment analytics to derive insights by analyzing raw text mined from microblogging services such as twitter and on-line customer reviews. Such services help businesses in analyzing vast amounts of real time data to essentially identify the sentiment of the crowd and analyze responses, trends, and behaviors (Cui et al. 2006, Das and Chen 2007, Yegulalp 2015). In such applications, the data observations are formed by a large collection of features that often adversely impact classification in terms of error rates and computational requirements. The difficulty of accurate classification is potentially due to the noisy features that are not relevant for classification but rather lead to the accumulation of relatively large errors. Thus the importance of feature selection which reduces the classifier function to the dimensions that are deemed most relevant for accurate classification.

1.1. Literature Review

Even without feature selection, finding a good classifier function is computational demanding. To address such challenges, Suykens and Vandewalle (1999) proposed a least squares formulation of SVM where finding the classifier requires the solution of a set of linear equations as compared to the quadratic program of traditional SVM (Cortes and Vapnik 1995, Vapnik 1995, 1998). Bradley and Mangasarian (2000) presented a linear support vector machine based on approximating the

quadratic distance function leading to a linear program that is computationally tractable compared to the quadratic counterpart. Rinaldi and Sciandrone (2010) proposed an iterative approach where a quadratic program is solved to train the SVM classifier followed by a linear program that eliminates features while maintaining the classification of the data. Alternatively, Ferris and Munson (2002) presented a specialized interior point method for solving the quadratic programming formulation of SVM and presented results for a data set consisting of 60 million observations with 34 features each. Other variants of SVM have also been developed. Lanckriet et al. (2003) proposed a robust optimization approach to construct a classifier where the probability of a correct classification is maximized given a mean and covariance matrix for each class. Lanckriet et al.'s model however forces the worst case accuracies for the binary classes to be equal. This assumption is relaxed in Huang et al. (2004) who proposed an approach to minimize the error rate of future data classification.

One of the advantages of support vector machines is solution sparsity which refers to the ability to express the classification function in terms of a relatively small portion of the training data. Solution sparsity is mainly desired to decrease the computational requirements for classification in practice. Downs et al. (2002) proposed an approach that deletes linearly dependent support vectors however without any guarantees on the resulting classification quality. Keerthi et al. (2006) presented a greedy approach to select a subset of the training set to approximate the SVM's optimal solution and showed that good approximations can be achieved with relatively small subsets of the initial training data. Wu et al. (2006) added a constraint to the SVM optimization problem to control the sparsity of the resulting solution and proposed a specialized optimization algorithm for the modified problem. While sparsity in the support vectors that define the classification function can be obtained with the aforementioned approaches, the resulting classifier often spans the majority of the defining features (Guyon and Elisseeff 2003). Ideally, it may be assumed that all the features are needed for accurate classification, however as discussed in Chan et al. (2007), in several applications many of the features may be either noise or too noisy for classification and ignoring such features may actually lead to an improved generalized classification. Furthermore, several features might be redundant (convex combination of other features) and ignoring them would not affect the performance but would rather simplify the interpretation of the results and the classification of the unlabeled data. Another major reason for favoring a low dimensional space for classification is that feature collection in practice may be too costly or sometimes infeasible. For instance, feature collection may necessitate the deployment of expensive sensors and would also require data storage and transmission.

Several methods have been proposed for embedded SVM feature selection. Guyon et al. (2002) and Weston et al. (2003) propose sequential approaches that iterate between selecting a subset

of features and optimizing the SVM classifier. This approach is known as the Recursive Feature Elimination (RFE) SVM where features whose removal lead to the largest margin of class separation are removed using backward elimination. Maldonado and Weber (2009) presented an iterative approach where at each iteration the feature with the least contribution to the classifier accuracy is removed and SVM is re-trained. Alternatively, feature selection and the SVM classifier can be optimized simultaneously. The approaches of Bradley and Mangasarian (2000), Fung and Mangasarian (2004), and Zhu et al. (2004) replace the quadratic part of SVM, i.e. the l_2–$norm$, by a linear approximation, i.e. the l_1–$norm$, which results in an optimization problem that leads to sparse features. While replacing the l_2–$norm$ by the l_1–$norm$ has a significant impact on reducing the computational requirement for solving the optimization problem, the l_1–$norm$ leads to relatively reduced classification accuracy due to the inability to maximize the margin of separation between the two classes of data. To overcome the disadvantages that are due to the l_1–$norm$, Zou and Hastie (2005) and Wang et al. (2006) proposed the elastic-net regularization technique that penalizes both the l_1–$norm$ and l_2–$norm$ in the objective function. Penalizing the l_1–$norm$ enables feature selection while the penalty on the l_2–$norm$ favors the joint selection or removal of highly correlated features which improves the accuracy of the resulting classifier. Another important characteristic is that the elastic-net regularization allows the number of selected features to exceed the sample size, in contrast to the case where only the l_1–$norm$ is used which limits the number of features to the sample size. This important characteristic of the elastic-net regularization is particularly relevant for the cases where the number features are much larger than the sample size such as cancer classification and gene selection using microarray data. Dunbar et al. (2010) proposed a reformulation of the elastic-net as a simple convex quadratic minimization problem and presented an iterative approach for its solution. Neumann et al. (2005) and Maldonado et al. (2011) proposed the extension of the SVM optimization model by the addition of a penalty term on the number of features that are used in the classifier, i.e. the l_0–$norm$ or the cardinality of the classifier vector and thus sparsity is encouraged by increasing the penalty. Using a similar approach, several sparse SVM models are presented in Chan et al. (2007) based on the addition of a constraint that limits the cardinality of the classifier. Since the resulting optimization model is not convex and computationally challenging to solve, two convex relaxations are proposed. The first is a quadratically-constrained quadratic program and the second is a semidefinite program. The semidefinite program remains challenging to solve for a large set of features while the quadratically-constrained quadratic program is a weak relaxation of the original problem and often leads to lower classification accuracy. Recently, Maldonado et al. (2014a) used a similar constraint that limits the cardinality of the classifier however without considering the l_2–$norm$ of the classifier function thus obtaining a mixed integer linear program. Maldonado et al. (2014b) proposed a feature selection approach based on

recursive feature elimination that is tailored for the case of imbalanced instances when one of the two classes has significantly more samples than the other. Rather than setting a limit on the desired number of features, Aytug (2015) sets a penalty on the number of selected features and proposed a Benders Decomposition approach for the solution of the resulting problem. While solving the optimization problems exactly remains computationally expensive, Bertolazzi et al. (2016) uses randomized metaheuristics for selecting the features.

Inspired by the convex relaxation of the cardinality constraint that is proposed in Chan et al. (2007), this paper proposes a joint feature selection and support vector machine optimization approach that iteratively provides support vector machine classifiers that span a desired number of features. The proposed approach is based on iteratively adjusting the limit on the l_1–$norm$ of the classifier vector thus forcing the number of selected features to converge towards the desired maximum limit. By using the l_1–$norm$, the proposed support vector machine optimization has commonalities with the convex relaxation that is proposed in Chan et al. (2007) as well as the l_2-l_1-SVM that is proposed in Neumann et al. (2005). However, as detailed in Sections 3 and 4, several distinguishing differences exist. Compared to Chan et al. (2007), the proposed approach leads to relaxations that are iteratively improved in order to meet the desired limit on the number of selected features while the relaxation that is proposed in Chan et al. (2007) often leads to selecting features beyond the desired limit. Alternatively the l_2-l_1-SVM of Neumann et al. (2005) does not impose a limit on the number of features but rather penalizes the selection of features in the objective function. The selection of an appropriate penalty parameter may potentially be challenging and does not guarantee a maximum number of features as proposed in this paper. While imposing a hard limit on the number of selected features may not necessarily be an advanctage for all applications, it allows the evaluation of the quality of classification as a function of number of selected features. Thus the limit on the number of features is an additional degree of freedom to balance the quality of classification and the dimension of the data to classify. The proposed approach also has commonalities with the elastic-net doubly regularized SVM of Wang et al. (2006) which is formulated as a quadratic programming problem with a non-linear l_1–$norm$ constraint. In fact, the doubly regularized SVM can be seen as a relaxed version of the SVM that is proposed in this paper which imposes an additional limit on the number of selected features.

Using the proposed approach, we analyze two real-life classes of classification problems with high dimensional features and varying characteristics. The first case is the medical diagnosis of tumors based on microarray data and the second test case deals with sentiment classification of online reviews from Amazon, Yelp, and IMDb. The results show that the proposed feature selection and support vector machine classification is intuitive to implement and computationally tractable which are essential characteristics for practical cases that involve data sets with large number of

features. The resulting classifier also outperforms competing approaches in terms of the accuracy of the classifications. We note however that while in this paper the focus is on linear classifiers, for non-linear classifiers, the challenge remains in the ability to exploit the kernel functions to avoid solving problems over intractable high dimensional space.

Following this introductory section, the classic support vector machine problem formulations are presented in Section 2. The feature selection approach that is proposed in this paper is described in Section 3. The medical diagnosis case and the sentiment classification case are presented in Section 4. Finally, a conclusion is provided in Section 5 and future research directions are highlighted.

2. Classic SVM Problem Formulation

SVMs form a class of supervised machine learning algorithms which train the classifier function using pre-labeled data. Given the training data set $\{x_i\}_{i=1}^m$ where each observation x_i has n features such that $x_i \in \mathbb{R}^n$ and a corresponding label $y_i \in \{-1, 1\}$, the objective of the support vector machine problem is to identify a hyperplane $\{w \mid w^T x - b = 0\}$ that separates the two classes of points with a maximal separation margin as measured by the l_2–$norm$. Since perfect separation between the two classes is often infeasible, slack variables ξ_i are introduced to allow errors in the classification of the data that may not be linearly separable (Vapnik 1995). The standard SVM problem was thus formulated in Cortes and Vapnik (1995) as the following convex quadratic problem:

$$\textbf{(SVM-Q)} \quad \min_{w,b} \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{m} \xi_i \tag{1}$$

$$\text{s.t. } y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, m, \tag{2}$$

$$\xi_i \geq 0 \quad \forall i = 1, \cdots, m. \tag{3}$$

The objective function balances between maximizing the separation margin and minimizing the classification error given an error weight C. Problem (SVM-Q) is a convex quadratic program that can be solved efficiently for large scale instances using modern optimization solvers.

Given all the n features that define each observation in the training data, the feature selection problem selects a subset of at most $r > 0$ features while maximizing the separation margin between the two classes of data. As proposed in Chan et al. (2007), a cardinality limit on the vector w is imposed and the resulting problem is

$$\textbf{(SVM-C1)} \quad \min_{w,b,\xi} \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{m} \xi_i \tag{4}$$

$$\text{s.t. } y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, m, \tag{5}$$

$$\text{Cardinality}(w) \leq r, \tag{6}$$

$$\xi_i \geq 0 \quad \forall i = 1, \cdots, m, \tag{7}$$

where Cardinality(w) denotes the number of nonzero elements of w, i.e. the l_0–$norm$ of w. By introducing the binary variables

$$z_j = \begin{cases} 1 & \text{if } w_j \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

(SVM-C1) can be rewritten as the following mixed integer quadratic problem

$$\textbf{(SVM-C2)} \quad \min_{w,b,\xi,z} \frac{1}{2} \| w \|_2^2 + C \sum_{i=1}^{m} \xi_i \tag{8}$$

$$\text{s.t. } y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, m, \tag{9}$$

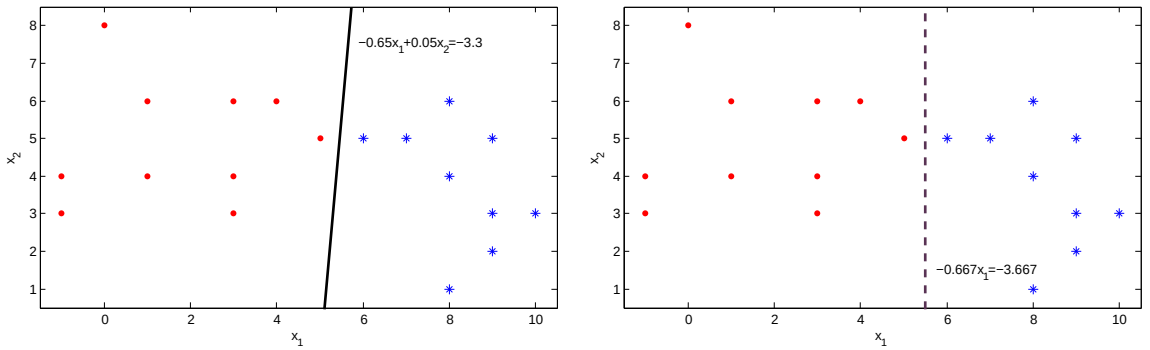
$$-Mz_j \leq w_j \leq Mz_j, \quad \forall j = 1, \cdots, n, \tag{10}$$

$$\xi_i \geq 0 \quad \forall i = 1, \cdots, m, \tag{11}$$

$$\sum_{j=1}^{n} z_j \leq r, \tag{12}$$

$$z_j \in \{0, 1\}, \quad \forall j = 1, \cdots, n. \tag{13}$$

Given a sufficiently large value for parameter M, Constraints (10) force z_j to take a value of 1 when w_j is not zero. While problem (SVM-C2) is a mixed integer quadratic program that can be solved by a state of the art optimization solver such as CPLEX, (SVM-C2) remains computationally expensive for large scale classification problems. Figure 1 provides an example of a small instance solved using (SVM-Q) and (SVM-C2) that illustrates the value of feature selection. The instance consists of two features, x_1 and x_2, and 19 training points. The solution of (SVM-Q) results in two features being selected while the relevant feature for the classifier is x_1 which is obtained by solving (SVM-C2) with $r = 1$.



(a) Solution of (SVM-Q). (b) Solution of (SVM-C2).

Figure 1 SVM example with and without feature selection.

Rather than using binary variables as in (SVM-C2), Chan et al. (2007) proposed a non-convex relaxation of the cardinality, i.e. Constraint (6), based on the Cauchy-Schwartz inequality such that

$$\text{Cardinality}(w) \leq r \Rightarrow \|w\|_1^2 \leq r\|w\|_2^2. \tag{14}$$

Constraint (14) is a non-convex constraint. Alternatively, given a decision variable t, the following weaker relaxation

$$\|w\|_2^2 \leq t, \ \|w\|_1^2 \leq rt, \tag{15}$$

is convex. As proposed by Chan et al. (2007), problem (SVM-C1) can then be relaxed to the following quadratically-constrained quadratic program

$$\textbf{(SVM-CR1)} \quad \min_{w,b,\xi} \ \frac{1}{2}t + C\sum_{i=1}^{m}\xi_i \tag{16}$$

$$\text{s.t. } y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, m, \tag{17}$$

$$\|w\|_2^2 \leq t, \tag{18}$$

$$\|w\|_1^2 \leq rt, \tag{19}$$

$$t \geq 0, \tag{20}$$

$$\xi_i \geq 0 \quad \forall i = 1, \cdots, m. \tag{21}$$

The solution of problem (SVM-CR1) is efficient in terms of computational time, however as shown in Section 4 the resulting optimal solutions are weak and violate the imposed desired features limit.

In order to solve the feature selection problem effectively for big data sets with large number of features, the following section proposes a new approach for the joint feature selection and support vector machine classification.

3. Proposed Approach

The weak bound that is provided by (SVM-CR1) is due to the loose limit that is imposed on $\|w\|_1^2$ by Constraint (19). Thus the approach that is proposed in this paper imposes an alternative limit using a parameter θ and Constraint (19) is replaced by $\|w\|_1^2 \leq \theta$. The value of θ is iteratively adjusted to obtain improving bounds by solving the following quadratically constrained quadratic problem

$$\textbf{(SVM-CR2)} \quad \min_{w,b,\xi} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i \tag{22}$$

$$\text{s.t. } y_i(w^T x_i - b) \geq 1 - \xi_i, \quad \forall i = 1, \cdots, m, \tag{23}$$

$$\|w\|_1^2 \leq \theta, \tag{24}$$

$$\xi_i \geq 0 \quad \forall i = 1, \cdots, m. \tag{25}$$

Hence the proposed optimal classifier can be obtained by finding a value for θ that is low enough so that the cardinality of w does not violate the maximum limit and is large enough to allow the lowest possible value for objective function (22). Note that as discussed earlier, (SVM-CR2) has commonalities with the elastic-net doubly regularized SVM of Wang et al. (2006). In fact, for fixed value of θ, there exists tuning parameters for the doubly regularized SVM of Wang et al. (2006) such that the resulting solution is identical to the solution of (SVM-CR2). The major difference with the doubly regularized SVM is that the proposed approach aims to solve problem (SVM-C2), i.e. to find a value for θ such that the cardinality of w does not violate the maximum limit r. The doubly regularized SVM can be seen as a relaxed version of (SVM-C2) as it relaxes the limit on the selected number of features r and rather considers r as a parameter that can be tuned.

To find the best value for θ, a bisection based algorithm based on iteratively maintaining the following two values of θ is proposed. A lower bound θ_l denotes a value of θ where solving problem (SVM-CR2) results in a classifier w such that Cardinality$(w) \leq r$. A starting value of θ_l may be the trivial $\theta_l = 0$ which leads to Cardinality$(w) = 0$. We note that with $\theta = \theta_l$, (SVM-CR2) results in a feasible solution to (SVM-C1) and hence a feasible classifier. The other value of θ that is maintained at each iteration is an upper bound θ_u such that solving problem (SVM-CR2) with $\theta = \theta_u$ results in a classifier w where Cardinality$(w) > r$. A good starting value for θ_u can be easily obtained by solving problem (SVM-Q) and setting $\theta_u = \|\overline{w}\|_1^2$, where $\overline{w}$ is the optimal classifier as obtained from (SVM-Q).

Given a starting $(\theta_l,\ \theta_u)$, at each iteration k a new value $\theta^k = \frac{\theta_l + \theta_u}{2}$ is computed and problem (SVM-CR2) is solved with $\theta = \theta^k$. If the resulting classifier $\tilde{w}^k$ has Cardinality$(w) \leq r$ then θ_l is set to $\theta_l = \theta^k$ otherwise θ_u is set to $\theta_u = \theta^k$. The iterative algorithm stops when the upper and lower bounds on θ are within a desired gap. The proposed joint feature selection and support vector machine (FS-SVM) classification algorithm is as follows

Initialization:
Set the iteration count $k = 0$;
Set $\theta_l = 0$;
Solve (SVM-Q) and set $\theta_u = \|\overline{w}\|_1^2$, where $\overline{w}$ is the optimal classifier given by (SVM-Q);
Set the desired gap ϵ;
while $\frac{\theta_u - \theta_l}{\theta_u} \leq \epsilon$ **do**
  $\theta^k = \frac{\theta_u + \theta_l}{2}$;
  Solve (SVM-CR2) with $\theta = \theta_k$ and obtain an optimal classifier $\tilde{w}^k$;
  if *Cardinality*$(w) \leq r$ **then**
    Set $\theta_l = \theta^k$;
  else
    Set $\theta_u = \theta^k$;
  end
  $k = k + 1$;
end

Algorithm 1: Joint feature selection and support vector machine (FS-SVM).

The classifier resulting from (FS-SVM) is the one that corresponds to $\theta = \theta_l$ thus guaranteeing Cardinality$(w) \leq r$. We note that in the computation of Cardinality(w), the value of w_i is considered to be zero indicating that the feature is not selected if $|w_i| \leq \gamma$ (in the computational testing, a value of $\gamma = 10^{-6}$ is considered). By including this threshold γ below which the value of w_i is forced to be exactly zero, then as observed in the computational testing section, for a sufficiently small ϵ (e.g. $\epsilon = 10^{-6}$) the proposed (FS-SVM) is capable of finding classifiers where Cardinality$(w) = r$.

The application of the proposed feature selection and support vector machine classification on two real cases is presented in the following section.

4. Test Cases and Results

This section presents an evaluation of the proposed support vector machine feature selection and classification. A description of the test cases is presented first followed by the implementation details and the computational results.

4.1. Data Sets

The experiments are performed on two classes of high dimensional real world datasets. The first class is sentiment classification of on-line reviews using data collected from Amazon, IMDb, and Yelp. The second class is cancer classification based on gene expressions for leukemia, prostate cancer, and lung cancer. The two datasets are described in details next.

4.1.1. Sentiment Classification With the rapid growth of on-line social networks, sensing social sentiment is becoming an integral part of marketing analytics (Netzer et al. 2012). A crucial element of on-line sentiment analytics is the classification of the general opinion of subject matter

experts and end users between positive or negative. Labeling such opinions provides firms with growing opportunities to gain insights about their customers, receive real time feedback of user opinion, and monitor the response of users to initiatives such as marketing campaigns and product launches. Equally important, such tools allow firms to quickly identify and respond to potentially unforeseen user responses which otherwise would lead to unintended costly consequences. However, processing the vast quantity of information that is posted on-line is challenging due to the qualitative nature of the data and the scale of the affecting variables. We thus evaluate the potential of using the proposed support vector machine feature selection and classification to develop a generic sentiment classification approach based on an optimized number of features. The focus is on a set of on-line reviews posted on Amazon, IMDb, and Yelp. The data set can be obtained from the UCI Machine Learning Repository (Lichman 2013) and each set of reviews contains 1000 raw text sentences (500 positive and 500 negative). For the evaluation, a hold-out validation strategy is used where the reviews are randomly split into a training set and a testing set each containing 250 positive and 250 negative reviews. For each data set, the features are extracted by enumerating all the words that appear in the reviews. We note that we have eliminated all the words that appear only once from the set of features as well as all special characters. The following are the details of each of the data sets.

Amazon: The Amazon set is formed of customer reviews of mobile phones and accessories that have been sold on amazon.com. The reviews that have a score of 4 and 5 are classified as positive while the reviews that have a score of 1 and 2 are classified as negative (the ones with a score of 3 are considered neutral and not included). The total number of features for the Amazon data set is 1874.

IMDb: The IMDb set is formed of reviews of movies on the IMDb website. Each review is labeled as either positive or negative. The total number of features in the IMDb data set is 3082.

Yelp: The Yelp set is formed of restaurant reviews from Yelp website. Similar to the Amazon data set, the reviews that have a score of 4 and 5 are classified as positive while the reviews that have a score of 1 and 2 are classified as negative (the ones with a score of 3 are considered neutral and not included). The total number of features in the Yelp data set is 2058.

4.1.2. Cancer Classification Cognitive computer systems are leading to new innovations in health analytics. Oncologists along with technology companies are integrating new supercomputing tools in medical decision making. For instance, IBM Watson is now being used to process the ever increasing health data and clinical information to develop advanced health decision support systems (Furlow 2016). Particularly, the classification of tumors based on gene expressions is of great interest due to the dimensions and the complexity of the underlying problem. The importance of

this problem is in the correct identification of the tumor based on the genetic fingerprint in order to allow for personalized cancer treatments. For instance, Steve Gold (Vice President at IBM Watson Group) highlighted that "Genomics is the secret to unlocking personalized medicine" (Friedman 2015). Tumor classification based on gene expression deals with using microarray data to identify cancer (Golub et al. 1999). Due to the large dimension of the microarray data, feature selection is key in identifying the genes that are most important to characterize tumor. Hence in this test case, we evaluate the use of the proposed feature selection and classification support vector machine in identifying three types of cancer. The three datasets that are used are leukemia, prostate cancer, and lung cancer and can be obtained from *http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi*. As detailed next, the datasets have different characteristics that can help in assessing the performance of the proposed support vector machine classifiers. Particularly, the prostate cancer dataset is formed of samples that contain a very large number of features. On the other hand, the lung cancer dataset includes two classes that are significantly different in terms of number of samples. Similar to the sentiment classification case, a hold-out validation strategy is used where the datasets are randomly split into a training set and a testing set. The following are the details of each of the datasets.

Leukemia: The leukemia dataset includes a total of 72 samples out of which 47 samples are acute lymphocytic leukemia and 25 samples are acute mylogenous leukemia. Each sample is formed of 7130 features. The dataset is divided into a training set with 24 lymphocytic leukemia cases and 13 samples with mylogenous leukemia. The test set is formed of the remaining samples which include 23 lymphocytic leukemia and 12 mylogenous leukemia cases.

Prostate Cancer: The prostate cancer dataset includes 50 normal samples and 52 tumor samples. This dataset is characterized by the presence of a large number of features where each sample is formed of 12601 features. The dataset is divided into a training set and a test set each containing 25 normal samples and 26 tumor samples.

Lung Cancer: The lung cancer dataset includes 17 normal samples and 139 samples with adenocarcinomas tumor. Each sample is formed of 1000 features. The dataset is thus divided into a training set with 9 normal samples and 70 cancerous samples and a test set with 8 normal samples and 69 cancerous samples. This test set thus evaluates the feature selection and classification in the presence of unbalanced classes where one class has many more samples than the other class.

4.2. Implementation and Evaluation Details

The proposed feature selection and classification support vector machine (FS-SVM) is compared to solving problems (SVM-C2) and (SVM-CR1) as well as to recursive feature elimination (RFE). All approaches are implemented using Julia and CPLEX 12.6 with default settings is used to solve

the underlying optimization models. The computational experiments are conducted on a Lenovo Thinkstation P300 with 32GB of RAM and the number of CPU cores that can be used is limited to 4. Two performance criteria are used to compare the methods. The first criterion is the ability of the method to find a classifier with the desired number features within a reasonable computational time. The second criterion is the accuracy of the classification for the presented datasets.

4.3. Results and Discussion

To assess the effect of feature selection, we evaluate the impact of varying the maximum number of allowable features r on the performance of the classifier in terms of error rate and computational time to compute the optimal classifier. A key difference between the sentiment classification datasets and the cancer classification datasets is the density of the samples. As one might expect, each of the reviews in the sentiment classification dataset contains a small subset of the words that are used in all the reviews and hence the sentiment classification dataset is characterized by a sparse features matrix. In that case, parameter r should remain reasonably large for accurate classification otherwise the selected features, i.e. keywords, will be too few and many sentences will not contain any of those keywords. Thus for the sentiment classification datasets, parameter r is varied between 10% and 50% of the number of features. On the otherhand, the samples in the cancer classification datasets are dense where all the features in a sample contain a value. Reducing the number of features then boils down to reducing the multicollinearity of the features that are retained. As the results show, parameter r can then have a small value while maintaining a small classification error. Thus for the cancer classification datasets, r is varied between 2% and 20% of the number of features. We note that besides parameter r, the error weight parameter C can be also used to tune (FS-SVM), (SVM-C2), and (SVM-CR1) however since the focus of this paper is on feature selection, a constant error weight $C = 10^6$ is considered for all the methods (We evaluated cases with $C = 10^{-6},\ 10^{-2},\ 10^2, 10^6$ and more accurate classifiers were obtained as C increased, thus we fixed $C = 10^6$).

4.3.1. Sentiment Classification The results for the sentiment classification are shown in Table 1. We first notice that although (SVM-CR1) is computationally efficient outperforming (FS-SVM) and (SVM-C2) for the majority of the tested instances, (SVM-CR1) is a weak relaxation and often selects features that exceed the maximum desired limit. In fact as shown in Table 1, (SVM-CR1) selects almost double the allowable maximum features when r is set between 10% and 40%. Problems with r between 10% and 30% are the most challenging to solve for (SVM-C2). With the exception of IMDb with $r = 30\%$, the computational time limit of 1 hour is reached before an optimal solution is found. For these challenging instances, (FS-SVM) is significantly faster and

	Joint Feature Selection SVM (FS-SVM)			(SVM-C2)			(SVM-CR1)			Recursive Feature Elimination (RFE)		
Features Limit	Selected Features	Classification Error*	Solution Time(s)	Selected Features	Classification Error*	Solution Time(s)/Gap	Selected Features	Classification Error*	Solution Time(s)	Selected Features	Classification Error*	Solution Time(s)
Amazon (Total Features: 1874)												
10%	187	**97/500** (19%)	64	187	102/500 (20%)	>1hour (58%)	371	74/500 (15%)	3	187	103/500 (21%)	587
20%	374	**73/500** (15%)	47	374	77/500 (15%)	>1hour (16%)	753	76/500 (15%)	2	374	101/500 (20%)	594
30%	562	77/500 (15%)	38	562	**75/500** (15%)	>1hour (3%)	907	76/500 (15%)	2	562	95/500 (19%)	577
40%	749	**74/500** (15%)	34	749	77/500 (15%)	2	907	76/500 (15%)	2	749	90/500 (18%)	542
50%	907	76/500 (15%)	26	907	76/500 (15%)	2	907	76/500 (15%)	2	916	97/500 (19%)	491
IMDb (Total Features: 3082)												
10%	279	135/500 (27%)	481	308	**131/500** (26%)	>1hour (40%)	619	107/500 (21%)	4	308	130/500 (26%)	1552
20%	616	**106/500** (21%)	388	616	113/500 (23%)	>1hour (9%)	1275	109/500 (22%)	4	616	123/500 (25%)	1469
30%	924	**105/500** (21%)	351	924	115/500 (23%)	616	1380	109/500 (22%)	4	924	122/500 (22%)	1456
40%	1223	**109/500** (22%)	297	1232	111/500 (22%)	65	1380	109/500 (22%)	4	1232	126/500 (25%)	1424
50%	1283	109/500 (22%)	32	1283	109/500 (22%)	2	1380	109/500 (22%)	4	1398	127/500 (25%)	1229
Yelp (Total Features: 2058)												
10%	195	**107/500** (21%)	127	205	118/500 (24%)	>1hour (66%)	413	108/500 (22%)	2	205	117/500 (23%)	664
20%	404	**103/500** (21%)	117	411	106/500 (21%)	>1hour (20%)	775	93/500 (19%)	2	411	117/500 (23%)	620
30%	617	**93/500** (19%)	113	617	98/500 (20%)	>1hour (5%)	1030	97/500 (19%)	2	617	119/500 (24%)	621
40%	823	**97/500** (19%)	47	823	99/500 (20%)	51	1030	97/500 (19%)	2	823	118/500 (24%)	590
50%	947	97/500 (19%)	22	947	97/500 (19%)	2	1030	97/500 (19%)	2	1029	114/500 (23%)	512

*: The classification error denotes the number of instances in the test set that were assigned an incorrect label based on the optimal classifier that is computed by each method.

Table 1 Computational Results: Sentiment Classification

	Top Selected Positive Words			Top Selected Negative Words		
Rank	Amazon	IMDb	Yelp	Amazon	IMDb	Yelp
1.	definitely	liked	fantastic	not	ugly	not
2.	love	masculine	delicious	contact	boring	english
3.	great	right	great	internet	girl	never
4.	good	brilliant	prompt	industrial	bad	batter
5.	best	actually	restaurant	try	empty	total
6.	nice	love	expect	slides	prejudice	though
7.	incredible	cool	amazing	fairly	excessively	different
8.	easy	nice	buffet	drawback	nothing	worst
9.	excellent	funny	quickly	wireless	skip	longer
10.	comfortable	enjoyed	seated	low	long	bland

Table 2 Words with the largest absolute weight in the classifier functions when $r = 10\%$.

often finds a solution with lower number of features and obtains a lower error rate than (SVM-C2). (RFE) results in higher classification error compared to the other methods.

The results that are shown in Table 1 also illustrate the importance of feature selection. For the three test sets, we notice that at most 30% of the features are needed to achieve the lowest classification error. In fact, even when the limit on the features exceeds 50%, the limit on the number of features becomes non binding and the models do not use all the allowable dimensions thus indicating that the additional features are rather insignificant and might hinder the quality of the classifier. Finally, the results also reveal that IMDb reviews are the hardest to correctly classify where a minimum classification error rate of 21% is obtained with (FS-SVM) ($r = 30\%$). A minimum classification error rate of 15% is obtained for Amazon with (FS-SVM) ($r = 20\%$) while the minimum error rate for Yelp is 19% as obtained by (FS-SVM) ($r = 30\%$).

The weights of the selected features in the classifier function also reveal valuable information about the reason of the positive or negative review. For instance as shown in Table 2, *comfortable* is among the top 10 features of a positive review regarding mobile phones on Amazon which may indicate that customers value the comfort of a mobile phone. On the other hand the top features of a negative review indicate complaints about *internet* and *wireless* which reveals that the customers are facing difficulties with the wireless capabilities of the phones. In fact, by checking the reviews,

several complaints about the internet can be found such as "*Internet is excruciatingly slow*". Other selected features such as *love*, *best*, and *good* often indicate a positive reviews while features such as *not*, *fairly*, and *drawback* indicate a negative review. The IMDb top selected features indicate that the presence of funny scenes often leads to a positive review such as "*The scenes are often funny and occasionally touching as the characters evaluate their lives and where they are going*". Similarly, the selected features of the Yelp reviews indicate that restaurants that offer buffets are well desired and receive positive reviews such as "*Today is the second time I've been to their lunch buffet and it was pretty good*". Thus, a thorough analysis of the selected features offers several key insights about the preferences of customer and highlights areas of improvement and the potential opportunities.

	Joint Feature Selection SVM (FS-SVM)			(SVM-C2)			(SVM-CR1)			Recursive Feature Elimination (RFE)		
Features Limit	Selected Features	Classification Error*	Solution Time(s)	Selected Features	Classification Error*	Solution Time(s)/Gap	Selected Features	Classification Error*	Solution Time(s)	Selected Features	Classification Error*	Solution Time(s)
Leukemia (Total Features: 7130)												
2%	142	**7/35** (20%)	1431	142	9/35 (26%)	>1hour (87%)	238	7/35 (20%)	35	142	7/35 (20%)	1948
4%	285	**7/35** (20%)	1105	285	9/35 (26%)	>1hour (76%)	410	5/35 (14%)	39	285	7/35 (20%)	1964
6%	427	**5/35** (14%)	968	427	9/35 (26%)	>1hour (79%)	559	5/35 (14%)	59	428	5/35 (14%)	1922
8%	570	**4/35** (11%)	791	570	8/35 (23%)	>1hour (79%)	658	4/35 (11%)	35	570	4/35 (11%)	1878
10%	713	**4/35** (11%)	688	713	6/35 (17%)	>1hour (75%)	753	4/35 (11%)	40	713	4/35 (11%)	1843
12%	855	**4/35** (11%)	70	855	6/35 (17%)	>1hour (75%)	824	4/35 (11%)	56	855	4/35 (11%)	1932
14%	991	4/35 (11%)	42	991	5/35 (14%)	>1hour (68%)	878	4/35 (11%)	51	999	**3/35** (9%)	1983
16%	991	**4/35** (11%)	42	907	5/35 (14%)	>1hour (68%)	915	4/35 (11%)	54	1141	4/35 (11%)	2010
18%	991	**4/35** (11%)	45	907	5/35 (14%)	>1hour (66%)	957	4/35 (11%)	53	1284	4/35 (11%)	1629
20%	991	**4/35** (11%)	44	907	5/35 (14%)	>1hour (65%)	991	4/35 (11%)	57	1426	4/35 (11%)	1713
Prostate Cancer (Total Features: 12601)												
2%	252	7/51 (14%)	2357	252	9/51 (18%)	>1hour (80%)	395	7/51 (14%)	119	252	**6/51** (12%)	3038
4%	504	6/51 (12%)	1889	504	8/51 (16%)	>1hour (70%)	699	6/51 (12%)	165	504	6/51 (12%)	3035
6%	756	6/51 (12%)	1486	756	8/51 (16%)	>1hour (63%)	925	6/51 (12%)	181	756	6/51 (12%)	3033
8%	1008	6/51 (12%)	1168	1008	8/51 (16%)	>1hour (70%)	1093	6/51 (12%)	151	1008	6/51 (12%)	3033
10%	1260	6/51 (12%)	100	1260	6/51 (12%)	>1hour (61%)	1245	6/51 (12%)	130	1260	6/51 (12%)	3025
12%	1512	6/51 (12%)	87	1512	6/51 (12%)	>1hour (56%)	1352	6/51 (12%)	128	1512	6/51 (12%)	3023
14%	1538	6/51 (12%)	83	1538	6/51 (12%)	>1hour (59%)	1414	6/51 (12%)	127	1765	6/51 (12%)	3022
16%	1538	6/51 (12%)	87	1538	6/51 (12%)	>1hour (59%)	1453	6/51 (12%)	126	2016	6/51 (12%)	3010
18%	1538	6/51 (12%)	82	1538	6/51 (12%)	>1hour (59%)	1502	6/51 (12%)	119	2268	6/51 (12%)	3007
20%	1538	6/51 (12%)	83	1538	6/51 (12%)	>1hour (59%)	1538	6/51 (12%)	107	2520	6/51 (12%)	2988
Lung Cancer (Total Features: 1000)												
2%	20	**2/77** (3%)	282	20	2/77 (3%)	>1hour (86%)	41	2/77 (3%)	1	20	2/77 (3%)	240
4%	40	**2/77** (3%)	191	40	2/77 (3%)	>1hour (75%)	69	2/77 (3%)	1	40	2/77 (3%)	237
6%	60	**2/77** (3%)	184	60	2/77 (3%)	>1hour (65%)	100	1/77 (1%)	1	60	2/77 (3%)	234
8%	80	**1/77** (1%)	149	80	1/77 (1%)	>1hour (59%)	134	1/77 (1%)	1	80	1/77 (1%)	233
10%	100	1/77 (1%)	118	100	1/77 (1%)	>1hour (60%)	156	1/77 (1%)	1	100	1/77 (1%)	231
12%	120	1/77 (1%)	93	120	1/77 (1%)	>1hour (46%)	195	1/77 (1%)	1	120	1/77 (1%)	231
14%	140	1/77 (1%)	83	140	1/77 (1%)	>1hour (41%)	228	1/77 (1%)	1	140	1/77 (1%)	226
16%	160	1/77 (1%)	79	160	1/77 (1%)	>1hour (37%)	252	1/77 (1%)	1	160	1/77 (1%)	211
18%	180	1/77 (1%)	76	180	1/77 (1%)	>1hour (33%)	276	1/77 (1%)	1	180	1/77 (1%)	203
20%	200	1/77 (1%)	75	200	1/77 (1%)	>1hour (29%)	310	1/77 (1%)	1	200	1/77 (1%)	198

*: The classification error denotes the number of instances in the test set that were assigned an incorrect label based on the optimal classifier that is computed by each method.

Table 3 Results: Gene Expression

4.3.2. Cancer Classification This test case evaluates the potential of using the proposed feature selection and classification support vector machine as an analytical approach for cancer identification based on gene microarrays. Classification functions using a varying number of features are obtained using the training sets and then evaluated on the test sets. The focus is on the more challenging cases where the limit on the number of features is tight (2%–20% of the total number of features). The results that are displayed in Table 3 show that classification with high accuracy can be obtained using a very limited number of features. The lung cancer cases are the most

	$\epsilon = 0.001$					$\epsilon = 0.00001$					$\epsilon = 0.0000001$				
Features	Class. Error	Iter.	Sol. Time	$\frac{1}{2}\|w\|_2^2$	$C\sum_{i=1}^m \xi_i$	Class. Error	Iter.	Sol. Time	$\frac{1}{2}\|w\|_2^2$	$C\sum_{i=1}^m \xi_i$	Class. Error	Iter.	Sol. Time	$\frac{1}{2}\|w\|_2^2$	$C\sum_{i=1}^m \xi_i$
Amazon (Total Features: 1874)															
10%	99/500	12	32	53.30	6.21e-2	97/500	19	53	53.30	6.21e-2	97/500	25	64	53.30	6.21e-2
20%	74/500	11	30	64.10	9.43e-3	73/500	18	40	63.85	6.44e-3	73/500	24	47	63.85	0.67e-3
30%	77/500	11	29	48.68	0.49e-3	77/500	18	36	48.67	1.79e-3	77/500	25	38	48.67	1.79e-3
40%	74/500	11	27	44.95	0.73e-3	74/500	18	30	44.94	8.03e-3	74/500	24	34	44.94	0.34e-3
IMDb (Total Features: 3082)															
10%	135/500	12	83	53.96	6.67e-2	135/500	19	125	53.94	6.67e-2	135/500	25	481	54.20	6.67e-2
20%	108/500	11	79	69.43	1.88e-3	106/500	18	114	69.34	4.26e-3	106/500	24	388	69.34	1.18e-3
30%	105/500	11	63	56.75	4.51e-3	105/500	18	98	56.73	3.15e-3	105/500	24	351	56.73	3.72e-3
40%	109/500	11	59	53.38	1.44e-3	109/500	18	95	53.38	0.36e-3	109/500	23	297	53.38	0.27e-3
Yelp (Total Features: 2058)															
10%	108/500	12	38	48.21	8.08e-2	107/500	19	57	48.19	8.08e-2	107/500	25	127	48.18	8.08e-2
20%	103/500	11	36	73.92	1.09e-2	103/500	17	54	73.95	1.06e-2	103/500	24	117	73.95	1.06e-2
30%	93/500	11	33	58.81	0.28e-3	93/500	18	52	58.76	0.06e-3	93/500	24	113	58.76	0.06e-3
40%	97/500	11	31	52.25	6.21e-3	97/500	17	41	52.24	3.77e-3	97/500	22	47	52.24	1.47e-3

Table 4 Computational Performance: Sentiment Classification

accurate to classify with only one case misclassified out of 77 using 8% of the features. Increasing the number of features however doesn't increase the accuracy thus highlighting the importance of feature selection. Reducing the number of features to 2% results in the misclassification of only two cases. Similar results are obtained for leukemia where four cases out of 35 are misclassified when 8% or more features are used. Reducing the number of features to 2% increases the misclassification to seven cases out of 35. For the prostate cancer, a misclassification rate as low as six cases out of 51 is obtained with 4% of the features. Reducing the number of features further to 2% leads to one additional misclassification. The results thus showcase the robustness of the proposed feature selection and classification in identifying the important features for accurate classification for cases with varying characteristics such as the case of prostate cancer where the dataset contains a very large number of features and the case of unbalanced classes such as the lung cancer dataset.

While the proposed (FS-SVM) leads to more accurate classification compared to solving (SVM-C2), the approach is also more computationally efficient than (SVM-C2) which reaches the time-limit of 1 hour of computational time. Similar to the sentiment analytics test case, (SVM-CR1) is computationally very efficient however leads to classifiers that exceed the imposed features limit. (RFE) was able to obtain less classification error for the Leukemia test set with $r = 14\%$ and the Prostate cancer test set with $r = 2\%$. Finally, Figures 2–4 list the 20 most informative genes that are selected for the classification of each of the cancer cases which are obtained by setting the maximum number of desired features to 20. Each row in the figures correspond to a gene and the columns correspond to the samples. The intensity of the color in each cell indicates the value of the normalized expression levels. The figures reveal clear differences between the normal samples and the cancer cases. The lung cancer case in particular (Figure 4) exhibits obvious distinctive gene expression levels between the normal samples and the samples with lung cancer. This obvious distinction illustrates the importance of feature selection in identifying the most informative expressions to achieve high classification accuracy as in the case of lung cancer.

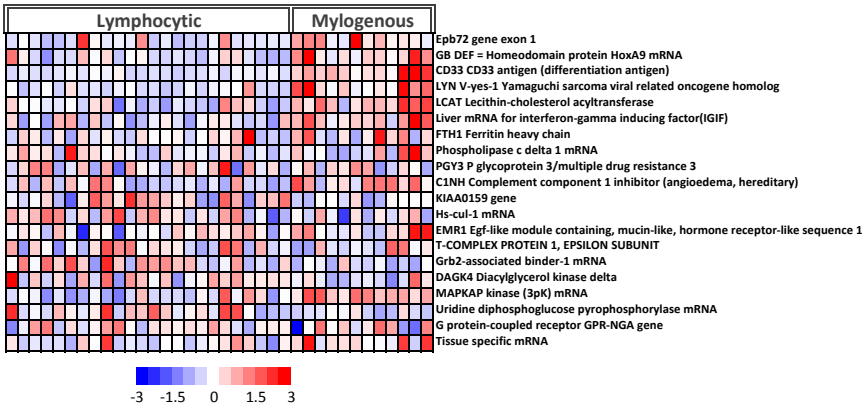


Figure 2 Leukemia: Top 20 selected genes and the normalized expression levels for each test set sample.

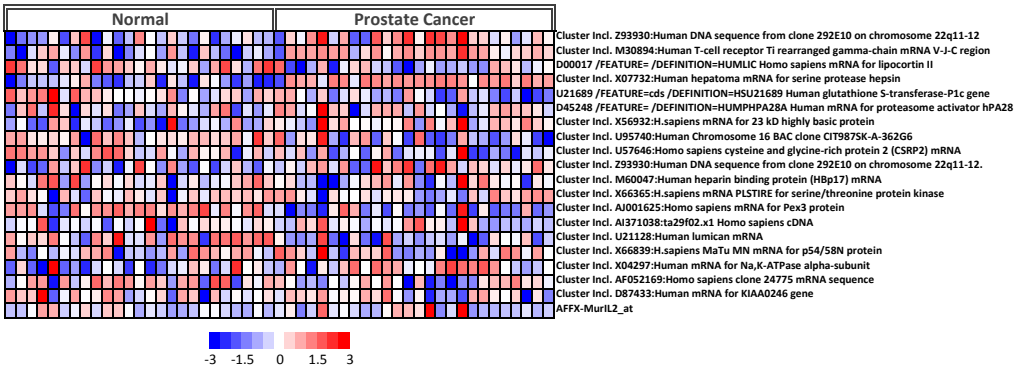


Figure 3 Prostate Cancer: Top 20 selected genes and the normalized expression levels for each test set sample.

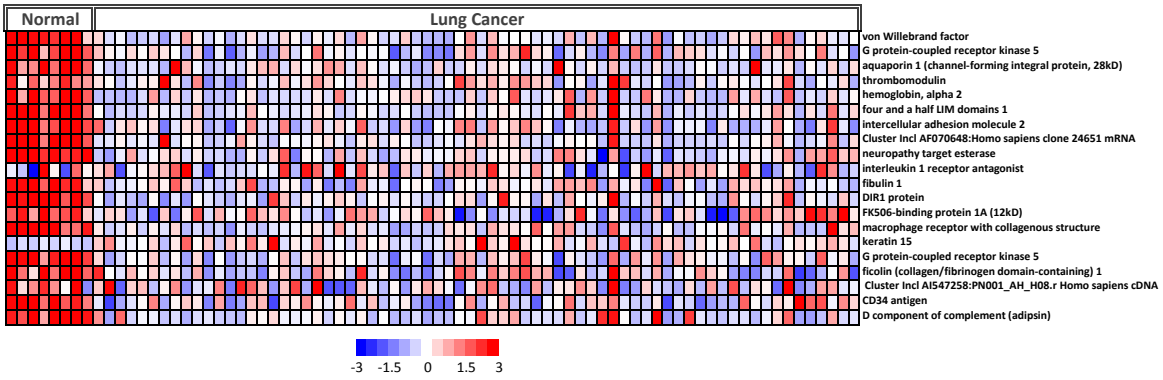


Figure 4 Lung Cancer: Top 20 selected genes and the normalized expression levels for each test set sample.

4.4. Computational Performance

The computational performance of (FS-SVM) is dependent on the gap parameter ϵ. In this section, the classification error, the number of iterations, the computational time, and the values of the two components of the objective function (22) are presented in Tables 4–5 for varying values of ϵ. The results show that the computational time and the number of iterations of Algorithm 1 decrease with an increase in the desired gap ϵ. Furthermore, for the majority of the tested instances, the

	$\epsilon = 0.001$					$\epsilon = 0.00001$					$\epsilon = 0.0000001$				
Features	Class. Error	Iter.	Sol. Time	$\frac{1}{2}\|w\|_2^2$	$C\sum_{i=1}^{m}\xi_i$	Class. Error	Iter.	Sol. Time	$\frac{1}{2}\|w\|_2^2$	$C\sum_{i=1}^{m}\xi_i$	Class. Error	Iter.	Sol. Time	$\frac{1}{2}\|w\|_2^2$	$C\sum_{i=1}^{m}\xi_i$
Leukemia (Total Features: 7130															
2%	7/35	20	529	0.0149	0.03e-6	7/35	27	744	0.0149	0.02e-6	7/35	32	1431	0.0149	0.02e-6
4%	7/35	19	453	0.0081	0.08e-6	7/35	27	701	0.0081	0.01e-6	7/35	32	1105	0.0081	0.05e-6
6%	5/35	19	434	0.0056	0.01e-6	5/35	27	674	0.0056	0.04e-6	5/35	31	968	0.0056	0.03e-6
8%	4/35	18	371	0.0043	0.02e-6	4/35	27	566	0.0043	0.01e-6	4/35	30	791	0.0043	0.02e-6
Prostate Cancer (Total Features: 12601)															
2%	7/51	26	1603	0.0184	0.20e-6	7/51	32	2294	0.0184	0.51e-6	7/51	38	2357	0.0184	0.01e-6
4%	6/51	26	1546	0.0103	0.13e-6	6/51	31	1808	0.0103	0.03e-6	6/51	34	1889	0.0103	0.03e-6
6%	6/51	26	1391	0.0074	0.06e-6	6/51	31	1454	0.0074	0.06e-6	6/51	35	1486	0.0074	0.06e-6
8%	6/51	25	956	0.0058	0.09e-6	6/51	30	1091	0.0058	0.01e-6	6/51	35	1168	0.0058	0.01e-6
Lung Cancer (Total Features: 1000)															
2%	2/77	12	26	0.0444	6.25e-6	2/77	17	127	0.0441	2.12e-6	2/77	24	282	0.0441	2.57e-6
4%	2/77	11	24	0.2720	0.36e-6	2/77	17	116	0.0271	0.35e-6	2/77	23	191	0.0271	0.30e-6
6%	2/77	10	22	0.0178	1.08e-6	2/77	17	76	0.0178	1.81e-6	2/77	23	184	0.0178	0.09e-6
8%	1/77	10	21	0.0129	1.39e-6	1/77	17	55	0.0129	1.39e-6	1/77	23	149	0.0129	1.39e-6

Table 5 Computational Performance: Gene Expression

classification error remains identical even when ϵ is increased from 10^{-6} to 10^{-3}. Particularly, for the sentiment classification test case shown in Table 4, the classification error increased only for $\epsilon = 10^{-3}$ and when a low number of features is set, i.e. $r = 10\%$ and $r = 20\%$ (IMDb only). For the remaining instances as well as for the gene expression test case shown in Table 5, the same classification error is achieved.

5. Conclusion and Future Work

This paper introduced a new joint support vector machine classification and feature selection based on iteratively adjusting a bound on the l_1–*norm* of the classifier function in order to enforce the desired sparsity level. The main characteristic of the proposed approach is its intuitive implementation and computational tractability for applications that contain high dimensional features where the direct application of standard feature selection models is computationally intractable. The proposed approach is demonstrated on two important classification problems. The first is sentiment classification of on-line reviews which is nowadays an essential practice for firms to quickly analyze and respond to market feedback. The second application is cancer classification based on gene expressions which aims to assist medical physicians in making highly accurate diagnosis by integrating the vast availability of medical data. The results demonstrate the capabilities of the proposed approach in dealing with large data sets that contain up to 12,600 features and in achieving low misclassification compared to other approaches. The analysis of the features also provides valuable insights that highlight the factors that are most important in decision making. Finally, while the support vector machine approach that is considered in this paper considers equal importance for the two data classes, we note that for cancer classification a false negative might have a greater consequence than a false positive and thus adapting the proposed approach to consider class importance may be desired for such use case.

Future work will also focus on addressing feature selection for non-linear classifiers. The challenge is in the ability to exploit the kernel functions to avoid solving problems over intractable

high dimensional space. Furthermore, exploring support vector machine classification and feature selection to gain insight in new niche applications is also of particular interest for future research.

Acknowledgments

We are very grateful to two anonymous referees for their valuable feedback and comments that helped improve the content of the paper. Bissan Ghaddar was supported by NSERC Discovery Grant RGPIN-2017-04185 and Joe Naoum-Sawaya was supported by NSERC Discovery Grant RGPIN-2017-03962.

References

Aytug H (2015) Feature selection for support vector machines using generalized benders decomposition. *European Journal of Operational Research* 244(1):210–218.

Baesens B, Setiono R, Mues C, Vanthienen J (2003) Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science* 49(3):312–329.

Bertolazzi P, Felici G, Festa P, Fiscon G, Weitschek E (2016) Integer programming models for feature selection: New extensions and a randomized solution algorithm. *European Journal of Operational Research* 250(2):389–399.

Bradley P, Mangasarian O (2000) Massive data discrimination via linear support vector machines. *Optimization Methods and Software* 13(1):1–10.

Chan AB, Vasconcelos N, Lanckriet GR (2007) Direct convex relaxations of sparse SVM. *Proceedings of the 24th international conference on Machine learning*, 145–153 (ACM).

Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20(3):273–297.

Cui G, Wong ML, Lui HK (2006) Machine learning for direct marketing response models: Bayesian networks with evolutionary programming. *Management Science* 52(4):597–612.

Das SR, Chen MY (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53(9):1375–1388.

Downs T, Gates KE, Masters A (2002) Exact simplification of support vector solutions. *The Journal of Machine Learning Research* 2:293–297.

Dunbar M, Murray JM, Cysique LA, Brew BJ, Jeyakumar V (2010) Simultaneous classification and feature selection via convex quadratic programming with application to HIV-associated neurocognitive disorder assessment. *European Journal of Operational Research* 206(2):470–478.

Ferris MC, Munson TS (2002) Interior-point methods for massive support vector machines. *SIAM Journal on Optimization* 13(3):783–804.

Friedman L (2015) IBM's Watson computer can now do in a matter of minutes what it takes cancer doctors weeks to perform. `http://www.businessinsider.com/r-ibms-watson-to-guide-cancer-therapies-at-14-centers-2015-5`.

Fung GM, Mangasarian OL (2004) A feature selection newton method for support vector machine classification. *Computational Optimization and Applications* 28(2):185–202.

Furlow B (2016) IBM Watson collaboration aims to improve oncology decision support tools. `http://www.cancernetwork.com/mbcc-2016/ibm-watson-collaboration-aims-improve-oncology-decision-support-tools`.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286(5439):531–537.

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *Journal of machine learning research* 3(Mar):1157–1182.

Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. *Machine Learning* 46(1-3):389–422.

Huang K, Yang H, King I, Lyu M, Chan L (2004) The minimum error minimax probability machine. *The Journal of Machine Learning Research* 5:1253–1286.

Keerthi SS, Chapelle O, DeCoste D (2006) Building support vector machines with reduced classifier complexity. *The Journal of Machine Learning Research* 7:1493–1515.

Lanckriet G, El-Ghaoui L, Bhattacharyya C, Jordan M (2003) A robust minimax approach to classification. *The Journal of Machine Learning Research* 3:555–582.

Lichman M (2013) UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

Maldonado S, Pérez J, Weber R, Labbé M (2014a) Feature selection for support vector machines via mixed integer linear programming. *Information sciences* 279:163–175.

Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. *Information Sciences* 179(13):2208–2217.

Maldonado S, Weber R, Basak J (2011) Simultaneous feature selection and classification using kernel-penalized support vector machines. *Information Sciences* 181(1):115–128.

Maldonado S, Weber R, Famili F (2014b) Feature selection for high-dimensional class-imbalanced data sets using support vector machines. *Information Sciences* 286:228–246.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text mining. *Marketing Science* 31(3):521–543.

Neumann J, Schnörr C, Steidl G (2005) Combined SVM-based feature selection and classification. *Machine Learning* 61(1-3):129–150.

Rinaldi F, Sciandrone M (2010) Feature selection combining linear support vector machines and concave optimization. *Optimization Methods & Software* 25(1):117–128.

Steadman I (2013) IBM's Watson is better at diagnosing cancer than human doctors. http://www.wired.co.uk/news/archive/2013-02/11/ibm-watson-medical-doctor.

Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural Processing Letters* 9(3):293–300.

Ustun B, Rudin C (2016) Supersparse linear integer models for optimized medical scoring systems. *Machine Learning* 102(3):349–391.

Vapnik V (1995) *The Nature of Statistical Learning Theory* (Springer-Verlag).

Vapnik V (1998) *Statistical learning theory*, volume 1 (Wiley).

Wang L, Zhu J, Zou H (2006) The doubly regularized support vector machine. *Statistica Sinica* 589–615.

Weston J, Elisseeff A, Schölkopf B, Tipping M (2003) Use of the zero norm with linear models and kernel methods. *The Journal of Machine Learning Research* 3:1439–1461.

Wu M, Schölkopf B, Bakır G (2006) A direct method for building sparse kernel learning algorithms. *The Journal of Machine Learning Research* 7:603–624.

Yegulalp S (2015) IBM's Watson mines twitter for sentiments. http://www.infoworld.com/article/2897602/big-data/ibms-watson-now-mines-twitter-for-sentiments-good-bad-and-ugly.html.

Zhu J, Rosset S, Hastie T, Tibshirani R (2004) 1-norm support vector machines. *Advances in Neural Information Processing Systems* 16(1):49–56.

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.