Stable Feature Selection with Support Vector Machines

Iman Kamkar(✉), Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh

Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, Australia
ikamkar@deakin.edu.au

Abstract. The support vector machine (SVM) is a popular method for classification, well known for finding the maximum-margin hyperplane. Combining SVM with l_1-norm penalty further enables it to simultaneously perform feature selection and margin maximization within a single framework. However, l_1-norm SVM shows instability in selecting features in presence of correlated features. We propose a new method to increase the stability of l_1-norm SVM by encouraging similarities between feature weights based on feature correlations, which is captured via a feature covariance matrix. Our proposed method can capture both positive and negative correlations between features. We formulate the model as a convex optimization problem and propose a solution based on alternating minimization. Using both synthetic and real-world datasets, we show that our model achieves better stability and classification accuracy compared to several state-of-the-art regularized classification methods.

1 Introduction

High dimensional datasets have become increasingly popular in many real-world applications. However, it is generally believed that in these datasets often only a small number of features are informative and the remaining features are either noisy or contain irrelevant information. Hence, selecting truly informative features is essential for many real applications [8] and improves the prediction accuracy of the model.

One of the important attributes of feature selection methods is their "stability" in selecting informative features. The feature stability is defined as the variation in obtained feature sets due to small changes in dataset [18] and is crucial in applications where selected features are used for knowledge discovery and decision makings [12]. For example, in clinical domain, explaining the risk factors in prognosis is as important as the prognosis itself. Consequently, stable features in spite of data resampling, are critical for clinical adoption [7].

A widely used strategy for feature selection that imposes sparsity on regression or classification coefficients is l_1-norm regularization. Perhaps the most well-known example is Lasso that minimizes the sum of squared errors while penalizing the l_1-norm of the regression coefficients [13]. The idea of using l_1-norm penalty to automatically select features has also been extended to classification problems.

B. Pfahringer and J. Renz (Eds.): AI 2015, LNAI 9457, pp. 298–308, 2015.
DOI: 10.1007/978-3-319-26350-2_26

Zhu et al. in [21] proposed l_1-norm support vector machine that can perform feature selection and binary classification, simultaneously. Although using l_1-norm regularization has shown success in many applications and has been generalized for different settings [21], it shows instability in presence of correlated features. The reason for such instability is that it tends to assign a nonzero weight to only a single feature among a group of correlated features [19,20].

Different methods have been proposed to address the instability of l_1-norm methods. Many of these methods try to find the groups of correlated features because these groups are consistent to the variation of training data. In presence of feature grouping information, groups Lasso [19] can be used and if features have an intrinsic hierarchical structure, tree-Lasso can be considered as a solution for stabilizing Lasso [10]. When there are ordering between features which imposes correlation among them, fused-Lasso can be used as a remedy to increase the stability of Lasso by selecting neighboring features [14]. Use of these methods requires that we know the structure of the data. However, such a structure is not available in many applications, which renders these methods inapplicable. There are limited works that try to solve the instability of l_1-norm methods in general context by incorporating feature similarities. Elastic net [22] is one of these methods that assigns comparable weights to similar features by using a combination of l_1 and l_2 penalty. However, it results in a longer lists of features compared to Lasso. Another method is Oscar that performs feature grouping and feature selection, simultaneously [1]. This method uses a combination of l_1 and pairwise l_∞ norm penalties to impose sparsity and equal feature weights for highly correlated features. The features with equal weights automatically form a group. Although Oscar tends to increase the stability of Lasso by grouping correlated features, assigning equal weights to features that are partially correlated may degrade the performance of the method [3].

All the methods discussed above, are proposed to increase the feature stability of Lasso, where its loss function is residual sum of squares or the logit function. As mentioned before, l_1-norm penalty terms are also combined with support vector machines to encourage sparsity. However, limited research has been done to address the instability of l_1-norm in these methods. To the best of our knowledge the only work done to address the instability in l_1-norm support vector machines is combining SVM with elastic net penalty [16,17]. However, this method does not properly exploit the feature correlations.

To address the instability in l_1-norm SVM, we propose a regularization formulation that encourages the similarities between features based on their relatedness. In our formulation, the relatedness between features is captured through a feature covariance matrix. Our method can perform feature selection and capture both positive and negative correlations between features through a convex objective function. In summary, our contributions are as follows:

– Proposal of a new model aimed to improve the stability of l_1-norm support vector machines by capturing the similarities between features based on their relatedness via a feature covariance matrix.

- Proposal of a convex optimization formulation for the model and a solution based on an alternating optimization.
- Demonstration of improved feature stability in terms of two stability measures, Jaccard similarity measure and Spearman's rank correlation coefficient in comparison with several baseline methods namely, Lasso, l_1-SVM and Elastic net SVM.
- Demonstration of improved classification accuracy of the model in comparison with the above baseline methods.

2 Framework

We propose a new model to address the instability of l_1-SVM in selecting informative features. We consider a binary classification problem with training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the feature vector and $y_i \in \{-1, 1\}$ is the class label. In general, we make two assumptions: (1) We are dealing with high dimensional but sparse setting. By sparsity we mean that the majority of the features are not predictive of the outcome. (2) Among the features, there are sets of features with high levels of correlations. In this context, l_1-SVM shows instability in selecting informative features because it randomly assigns a nonzero weight to a single feature among a group of correlated features and so with small changes in dataset, another feature maybe selected from the correlated group. To overcome this problem, the similarities between the features can be encouraged based on their relatedness. To this end, we use a feature covariance matrix to capture relationships between features. Our proposed model, is the solution to the following optimization problem:

$$\begin{aligned} \arg \min_{\beta_0, \beta, \Omega} \frac{1}{n}(1 - y_i(\beta_0 + \mathbf{x}_i^T \beta))_+ + \lambda \|\beta\|_1 + \frac{\eta}{2} \beta^T \Omega^{-1} \beta \\ \text{s.t.} \qquad \Omega \succeq 0, \mathrm{tr}(\Omega) = 1, \end{aligned} \tag{1}$$

where β is the vector of feature weights and β_0 is the intercept. Also, Ω is the covariance matrix that models the relationships between features, λ and η are the tuning parameters and $(1 - T)_+ = \max(T, 0)$ is the hinge loss. The term $\beta^T \Omega^{-1} \beta$ ensures that feature weights follow the feature correlations, i.e. if two features are highly correlated their feature weights would become very high. We refer to the above model as **Covariance SVM (C-SVM)**.

2.1 Algorithm for Covariance-SVM

Although the objective function in (1) is convex with respect to all variables, it is not straight forward due to the non-smooth convexity. To solve this problem, we introduce an iterative algorithm that alternatively updates β and Ω as follows:

Optimizing w.r.t. β ***when*** Ω ***is Fixed:*** In this situation, the objective function can be stated as:

$$\arg\min_{\beta_0,\beta} \frac{1}{n}(1 - y_i(\beta_0 + \mathbf{x}_i^T\beta))_+ + \lambda\|\beta\|_1 + \frac{\eta}{2}\beta^T\Omega^{-1}\beta. \tag{2}$$

This problem can be solved using the alternate direction method of multipliers (ADMM), which has recently become a method of choice for solving many large-scale problems [2]. Because of the nondifferentiability of the hinge loss and l_1 norm term in (2), we introduce some auxiliary variables to handle these two nondifferentiable terms. Suppose $X = (x_{ij})_{i=1,j=1}^{n,p}$ and Y be a diagonal matrix, where its diagonal elements are the vector $y = (y_1 \ldots, y_n)^T$. So the problem in (2), can be reformulated as:

$$\begin{aligned} \arg\min_{\beta_0,\beta} \frac{1}{n}\sum_{i=1}^{n}(a_i)_+ + \lambda\|z\|_1 + \frac{\eta}{2}\beta^T\Omega^{-1}\beta \\ \text{s.t.} \quad \mathbf{a} = \mathbf{1} - Y(X\beta + \beta_0\mathbf{1}),\ z = \beta, \end{aligned} \tag{3}$$

where $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{1}$ is a column vector of 1's with length n. The augmented Lagrangian function of (3) is

$$\begin{aligned} L(\beta_0, \beta, a, z, u, v) = \frac{1}{n}\sum_{i=1}^{n}(a_i)_+ + \lambda\|z\|_1 + \frac{\eta}{2}\beta^T\Omega^{-1}\beta \\ + \langle u, \mathbf{1} - Y(X\beta + \beta_0\mathbf{1}) - a\rangle + \langle v, \beta - z\rangle, \end{aligned} \tag{4}$$

where $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ are dual variables corresponding to the first and the second constraints in Eq. (3), respectively. $\langle ., . \rangle$ is the inner product in the Euclidean space and μ_1 and μ_2 control the convergence behavior and are usually set to 1. By solving the above equation w.r.t u,v, (β_0, β), a and z we have:

$$\begin{cases} (\beta_0^{k+1}, \beta^{k+1}) = \arg\min_{\beta_0,\beta} L(\beta_0, \beta, a^k, z^k, u^k, v^k), \\ a^{k+1} = \arg\min_a L(\beta_0^{k+1}, \beta^{k+1}, a, z^k, u^k, v^k), \\ z^{k+1} = \arg\min_c L(\beta_0^{k+1}, \beta^{k+1}, a^{k+1}, z, u^k, v^k), \\ u^{k+1} = u^k + \mu_1(\mathbf{1} - Y(X\beta^{k+1} + \beta_0^{k+1}\mathbf{1}) - a^{k+1}), \\ v^{k+1} = v^k + \mu_2(\beta^{k+1} - z^{k+1}). \end{cases} \tag{5}$$

The first term in (5) is a quadratic and differentiable objective function, so its solution can be found by solving a set of linear equations:

$$\begin{aligned} \begin{pmatrix} \lambda_2\Omega^{-1} + \mu_2\mathbf{I} + \mu_1 X^T X & \mu_1 X^T\mathbf{I} \\ \mu_1\mathbf{1}^T X & \mu_1 n \end{pmatrix} \begin{pmatrix} \beta^{k+1} \\ \beta_0^{k+1} \end{pmatrix} \\ = \begin{pmatrix} X^T Y u^k - \mu_1 X^T Y(a^k - \mathbf{1}) - v^k + \mu_2 z^k \\ \mathbf{1}^T Y u^k - \mu_1\mathbf{1}^T Y(a^k - \mathbf{1}) \end{pmatrix}. \end{aligned} \tag{6}$$

The second term in (5) can be solved by using Proposition 1.

Proposition 1. *Let* $h_\lambda(w) = \arg\min_x \lambda x_+ + \frac{1}{2}\|x - w\|_2^2$. *Then* $h_\lambda(w) = w - \lambda$ *for* $w > \lambda$, $h_\lambda(w) = 0$ *for* $0 \leq w \leq \lambda$ *and* $h_\lambda(w) = w$ *for* $w < 0$.

So the second term in (5), can be written as

$$\frac{\|u\|_2^2}{2\mu_1} + \frac{\mu_1}{2}\|\mathbf{1} - Y(X\beta^{k+1} + \beta_0^{k+1}\mathbf{1}) - a\|_2^2 + \left\langle u^k, \mathbf{1} - Y(X\beta^{k+1} + \beta_0^{k+1}\mathbf{1}) - a\right\rangle$$
$$= \frac{\mu_1}{2}\|a - (\mathbf{1} + \frac{u}{\mu_1} - Y(X\beta^{k+1} + \beta_0^{k+1}\mathbf{1}))\|_2^2.$$

From above equation and Proposition 1, we can update a^{k+1} as follows:

$$a^{k+1} = H_{\frac{1}{n\mu_1}}(\mathbf{1} + \frac{u^k}{\mu_1} - Y(X\beta^{k+1} + \beta_0^{k+1}\mathbf{1})), \tag{7}$$

where $H_\lambda(w) = (h_\lambda(w_1), h_\lambda(w_2), \ldots, h_\lambda(w_n))^T$.

The third equation in (5) can be solved using soft thresholding. So we have

$$z^{k+1} = S_{\frac{\lambda}{\mu_2}}\left(\frac{v^k}{\mu_2} + \beta^{k+1}\right), \tag{8}$$

where S_λ is the soft threshold operator defined on vector space and $S_\lambda(w) = (s_\lambda(w_1), \ldots, s_\lambda(w_p))$, where $s_\lambda(w_i) = sgn(w_i)\max\{0, |w_i| - \lambda\}$.

By combining (5)–(8), we obtain the ADMM algorithm for solving the objective function (1) with respect to β when Ω is fixed.

Optimizing w.r.t Ω ***when*** β ***is fixed:*** In this situation, the optimization problem for finding Ω becomes

$$\min_{\Omega} \beta^T \Omega^{-1} \beta \text{ such that } \Omega \succeq 0, \text{tr}(\Omega) = 1$$

Let $B = \beta\beta^T$, as $\beta^T\Omega^{-1}\beta = \text{tr}(\beta^T\Omega^{-1}\beta) = \text{tr}(\Omega^{-1}\beta\beta^T)$ and $\text{tr}(\Omega) = 1$, so

$$\text{tr}(\Omega^{-1}B) = \text{tr}(\Omega^{-1}B)\text{tr}(\Omega) = \text{tr}((\Omega^{-\frac{1}{2}}B^{\frac{1}{2}})(B^{\frac{1}{2}}\Omega^{-\frac{1}{2}}))\text{tr}(\Omega^{\frac{1}{2}}\Omega^{\frac{1}{2}})$$
$$\geq (\text{tr}(\Omega^{-\frac{1}{2}}B^{\frac{1}{2}}\Omega^{\frac{1}{2}}))^2 = (\text{tr}(B^{\frac{1}{2}}))^2.$$

The inequality holds because of Cauchy-Schwarz inequality for the Frobenius norm. From this inequality, we can say that $\text{tr}(\Omega^{-1}B)$ achieves its minimum value $(\text{tr}(B^{\frac{1}{2}}))^2$ if and only if $\Omega^{-\frac{1}{2}}B^{\frac{1}{2}} = \zeta\Omega^{\frac{1}{2}}$ for some constant ζ and $\text{tr}(\Omega) = 1$. So Ω can be obtained from $\Omega = \frac{(\beta\beta^T)^{\frac{1}{2}}}{\text{tr}((\beta\beta^T)^{\frac{1}{2}})}$.

3 Experiments

In this section, we perform experiments using both synthetic and real datasets and compare the classification accuracy and feature stability of the C-SVM with several baselines that deemed to be closest to our work, namely Lasso [13], l_1-norm SVM [21] and Elastic net SVM (ENSVM) [16].

3.1 Tuning Parameter Selection

In case of synthetic data set, we use a validation set to select the tuning parameters λ and η. We train each model on the training set and use the validation set to select the best tuning parameter for the final model. The performance of each model is evaluated using the test set. In case of real data sets, we use the 5-fold cross validation to select the best tuning parameters.

3.2 Performance Metrics

Feature Stability Measures. To compare the feature stability of C-SVM with other methods we use two similarity measures, Jaccard similarity measure (JSM), which considers the indices of the selected features in its evaluation process and Spearman's rank correlation coefficient (SRCC), which considers the rank of the selected features for evaluating stability. Jaccard measures the similarities between any two sets of selected features S_q and $S_{q'}$ as $JSM(S_q, S_{q'}) = \frac{|S_q \cap S_{q'}|}{|S_q \cup S_{q'}|}$, where JSM$\in [0, 1]$ and 0 means there are no similarities between the two sets and 1 means the two sets are identical. SRCC measures similarity between two rankings r and r' as $SRCC(r, r') = 1 - 6\sum_j \frac{(r_j - r'_j)}{p(p^2-1)}$, where SRCC$\in [-1, 1]$ and 1 shows the two rankings are identical, 0 shows there is no correlation between two rankings and -1 shows that rankings are in inverse order. In our experiments we generate M sub-samples of the training set and apply each algorithm to each sub-sample to obtain its selected feature set. We use JSM and SRCC to evaluate the similarity between each pair of selected features and finally, average similarities over all pairs to obtain the stability of each algorithm.

Classification Accuracy. To compare the classification accuracy of C-SVM with other baselines, F-measure and AUC score are used [9].

3.3 Simulation Results

We consider a binary classification problem in a p dimensional space where only the first 50 features are relevant for classification and the remaining features are noise. To this end, we generate n instances where half of them belong to $+1$ class and the other half belong to -1 class. Instances in positive class are i.i.d drawn from a normal distribution with mean $\mu_+ = (\underbrace{1, \ldots, 1}_{50}, \underbrace{0, \ldots, 0}_{p-50})^T$ and covariance

$$\Sigma = \begin{pmatrix} \Sigma^*_{50\times 50} & \mathbf{0}_{50\times(p-50)} \\ \mathbf{0}_{(p-50)\times 50} & \mathbf{I}_{(p-50)\times(p-50)} \end{pmatrix},$$

where in Σ^* the diagonal elements are 1 and others are all equal to ρ. The mean for negative class is $\mu_- = (\underbrace{-1, \ldots, -1}_{50}, \underbrace{0, \ldots, 0}_{p-50})$. In this situation, the Bayes optimal classification rule depends on $x_1, \ldots, x_{50}$, which are highly correlated if ρ is large.

Table 1. Stability performance of C-SVM compared to other baselines for Synthetic dataset. Means and standard error over 50 iterations are reported.

Synthetic data		Lasso	l_1-SVM	ENSVM	C-SVM
$\rho = 0$	JSM	0.652 (0.015)	0.649 (0.019)	0.655 (0.026)	**0.662** (0.031)
	SRCC	0.452 (0.031)	0.448 (0.037)	0.463 (0.027)	**0.487** (0.025)
$\rho = 0.8$	JSM	0.447 (0.027)	0.510 (0.021)	0.571 (0.034)	**0.603** (0.027)
	SRCC	0.305 (0.032)	0.319 (0.018)	0.368 (0.032)	**0.407** (0.034)

We explore two values of ρ, 0 and 0.8, where $\rho = 0$ simulates the situation that informative features are uncorrelated to each other and $\rho = 0.8$, simulates the situation that those features are highly correlated. The stability performance of each method, measured in terms of JSM and SRCC, is shown in Table 1. The high value of SRCC implies that ranks of features do not vary a lot for different training sets and high value of JSM means that the selected features do not change significantly when there is a slight change in the training set. As the table implies, when there is no correlation among variables ($\rho = 0$), the stability of C-SVM is comparable to other baselines. However, when the correlation among features is high ($\rho = 0.8$) Lasso and l_1-SVM show low stability performance in terms of both JSM and SRCC. However, ENSVM that incorporates l_2-norm penalty shows better stability performance compared to Lasso and l_1-SVM. For C-SVM, we can see that as this model encourages similarities between features based on their relatedness, it shows the best stability compared to the baselines in terms of both JSM and SRCC.

Figure 1 shows the classification performance of C-SVM in terms of two classification measures, F-measure, and AUC and compares them with other baselines. As shown, the classification performance of C-SVM outperforms other baselines in terms of the both classification measures.

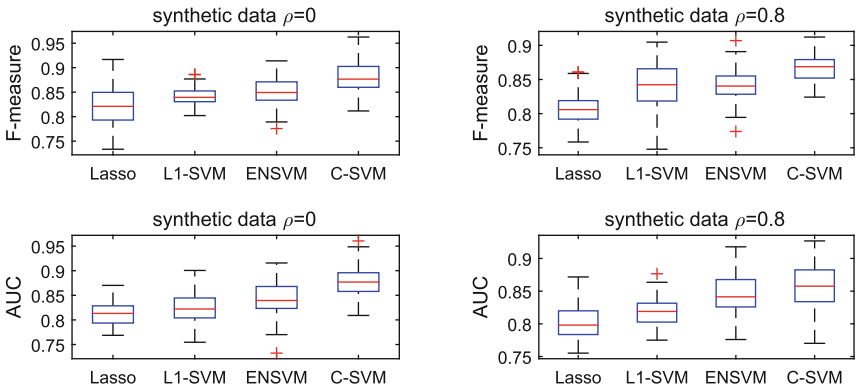

Fig. 1. Classification performance of C-SVM and baseline methods in terms of F-measure and AUC for Synthetic dataset.

3.4 Application on Real Datasets

In this section, we evaluate the performance of C-SVM on real datasets and compare it with other baselines. The datasets used are as follows:

Breast Cancer Dataset: This dataset is compiled by [15] and consists of gene expression data for 8141 genes in 295 breast cancer tumors (87 metastatic and 217 non-metastatic). As the dataset is very imbalanced, we balance it by using 3 replicates of each positive (metastasis) sample while keeping all replicates in the same fold during cross validation.

Cancer Dataset: This dataset is obtained from a large regional hospital in Australia. There are eleven different cancer types in this data recorded from patients visiting the hospital during 2010–2012. Patient data is acquired from Electronic Medical Records (EMR). The dataset consists of 4293 patients with 3867 variables including International Classification of Disease 10 (ICD-10), procedure and diagnosis related Group (DRG) codes of each patient as well as demographic data (age, gender and postcode). Using this dataset, our goal is to predict 1 year mortality of patients while ensuring the stable feature sets.

AMI Dataset: This dataset is also obtained from the same hospital in Australia. It involves patients admitted with AMI conditions and discharged later between 2007–2011. The task is to predict if a patient will be re-admitted to the hospital within 30 days after discharge. The dataset consists of 2941 patients with 2504 variables include International Classification of Disease 10 (ICD-10), procedure and diagnosis-related Group (DRG) codes of each admission; details of procedures; and departments involved in the patient's care.

Experimental Results

Stability Performance. The comparison between stability performance of C-SVM and other baselines in terms of JSM and SRCC for real datasets are presented in Table 2. For Breast cancer dataset, C-SVM shows the best stability performance in terms of JSM (0.620). However, in terms of SRCC, ENSVM represents the best stability (0.512), which is closely followed by C-SVM (0.509). For Cancer dataset, C-SVM shows the best stability performances with JSM = 0.631 and SRCC = 0.518. In terms of both JSM and SRCC, C-SVM is followed by ENSVM with JSM = 0.568 and SRCC = 0.427. For AMI dataset, again C-SVM shows the best stability in terms of both JSM (0.572) and SRCC (0.509), which is followed by ENSVM with JSM = 0.516 and SRCC = 0.457. As seen, stability performances of Lasso and l_1-SVM are close to each other and these methods show the least stability, the reason for which is that these methods use only l_1 regularization term which is unstable in selecting correlated features.

Classification Performance. Figure 2 shows the classification performance of C-SVM and other baselines in terms of F-measure and AUC for real-world datasets. For Breast cancer dataset, we can see that the classification performance of C-SVM outperforms other methods in terms of both F-measure and

Table 2. Stability performance of C-SVM compared to other baselines for real data sets. Means and standard error over 50 iterations are reported.

Real data		Lasso	l_1-SVM	ENSVM	C-SVM
Breast cancer	JSM	0.352 (0.028)	0.348 (0.025)	0.551 (0.021)	**0.620** (0.031)
	SRCC	0.237 (0.028)	0.235 (0.019)	**0.512** (0.026)	0.509 (0.023)
Cancer	JSM	0.420 (0.019)	0.423 (0.025)	0.568 (0.018)	**0.631** (0.026)
	SRCC	0.273 (0.021)	0.276 (0.030)	0.427 (0.022)	**0.518** (0.015)
AMI	JSM	0.372 (0.026)	0.368 (0.033)	0.516 (0.036)	**0.572** (0.036)
	SRCC	0.268 (0.031)	0.270 (0.024)	0.457 (0.028)	**0.509** (0.021)

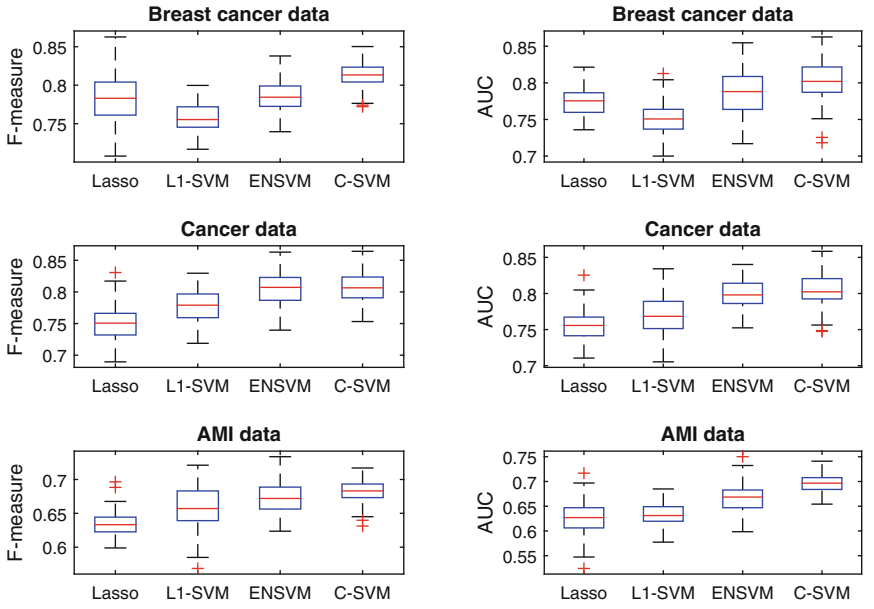

Fig. 2. Classification performance of C-SVM and other baselines in terms of accuracy, F-measure and AUC for Cancer dataset.

AUC. For Cancer data, we can see that C-SVM along with ENSVM show the best classification performance in terms of both F-measure and AUC. Turning to the AMI data, again C-SVM shows the best accuracy among other baselines in terms of the two classification measures.

Estimated Covariance Matrix. As feature names for AMI and Cancer datasets are available, we show the estimated covariance matrix for these datasets in Fig. 3 and we further discuss about some of the correlated features estimated in their Ω matrix. For better representation, we show the correlation matrix computed from Ω matrix by standardizing its values as $\Omega_{st}(i,j) = \frac{\Omega(i,j)}{\sqrt{\Omega(i,j)\Omega(i,j)}}$. In Ω matrix of

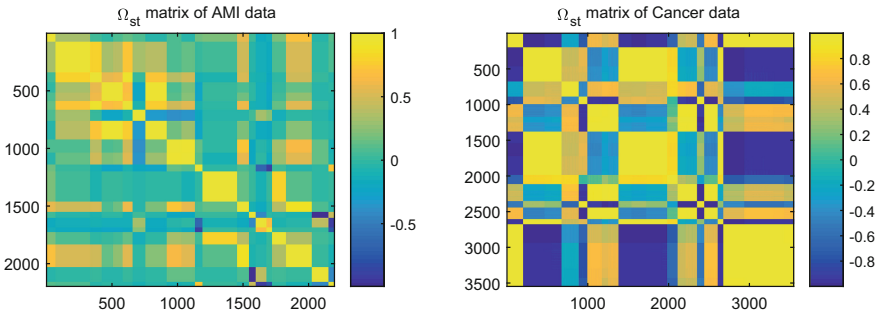

Fig. 3. The pictorial representation of estimated covariance matrix for real data sets. For better representation, we show the correlation matrix computed from Ω matrix by standardizing its values.

AMI dataset, the first group are the features related to cardiac troponin and the last group are features related to discharge sodium values. Both of these features are reported as important risk factors for Mayocardial infarction [6,11]. In Ω matrix obtained for Cancer dataset, the first group are the features related to diabetes mellitus and the last group are the features related to anemia, where both of these features are important risk factors for cancer survival prediction [4,5].

4 Conclusion

In this paper, we propose a method that can increase feature stability of l_1-norm SVM in presence of highly correlated features. The method can capture both the positive and negative relations between features using a covariance matrix, therefore the highly correlated features could be selected or rejected together by the model. We propose a convex formulation for the model that can be solved using an alternating optimization algorithm. We show the proposed method is more stable and more accurate than many existing methods.

References

1. Bondell, H.D., Reich, B.J.: Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics **64**(1), 115–123 (2008)
2. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
3. Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.H.: Correlated variables in regression: clustering and sparse estimation. J. Stat. Planning Infer. **143**(11), 1835–1858 (2013)
4. Caro, J.J., Salas, M., Ward, A., Goss, G.: Anemia as an independent prognostic factor for survival in patients with cancer. Cancer **91**(12), 2214–2221 (2001)

5. Coughlin, S.S., Calle, E.E., Teras, L.R., Petrelli, J., Thun, M.J.: Diabetes mellitus as a predictor of cancer mortality in a large cohort of us adults. Am. J. Epidemiol. **159**(12), 1160–1167 (2004)
6. Eapen, Z.J., Liang, L., Fonarow, G.C., Heidenreich, P.A., Curtis, L.H., Peterson, E.D., Hernandez, A.F.: Validated, electronic health record deployable prediction models for assessing patient risk of 30-day rehospitalization and mortality in older heart failure patients. JACC Heart Fail. **1**(3), 245–251 (2013)
7. Ein-Dor, L., Kela, I., Getz, G., Givol, D., Domany, E.: Outcome signature genes in breast cancer: is there a unique set? Bioinformatics **21**(2), 171–178 (2005)
8. Fan, J., Li, R.: Statistical challenges with high dimensionality: feature selection in knowledge discovery (2006). arXiv preprint math/0602133
9. Han, J., Kamber, M., Pei, J.: Data Mining: Concepts and Techniques. Elsevier, Massachussets (2011)
10. Kamkar, I., Gupta, S.K., Phung, D., Venkatesh, S.: Stable feature selection for clinical prediction: exploiting ICD tree structure using tree-lasso. J. Biomed. Inf. **53**, 277–290 (2015)
11. Mair, J., Artner-Dworzak, E., Lechleitner, P., Smidt, J., Wagner, I., Dienstl, F., Puschendorf, B.: Cardiac troponin T in diagnosis of acute myocardial infarction. Clin. Chem. **37**(6), 845–852 (1991)
12. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part II. LNCS (LNAI), vol. 5212, pp. 313–325. Springer, Heidelberg (2008)
13. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. Ser. B (Methodological) **58**(1), 267–288 (1996)
14. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K.: Sparsity and smoothness via the fused lasso. J. Roy. Stat. Soc. Ser. B (Statist. Method.) **67**(1), 91–108 (2005)
15. Van De Vijver, M.J., He, Y.D., van't Veer, L.J., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Marton, M.J., et al.: A gene-expression signature as a predictor of survival in breast cancer. N. Engl. J. Med. **347**(25), 1999–2009 (2002)
16. Wang, L., Zhu, J., Zou, H.: The doubly regularized support vector machine. Stat. Sinica **16**(2), 589 (2006)
17. Ye, G.B., Chen, Y., Xie, X.: Efficient variable selection in support vector machines via the alternating direction method of multipliers. In: International Conference on Artificial Intelligence and Statistics, pp. 832–840 (2011)
18. Yu, L., Ding, C., Loscalzo, S.: Stable feature selection via dense feature groups. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 803–811. ACM (2008)
19. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. Roy. Stat. Soc. Ser. B (Stat. Method.) **68**(1), 49–67 (2006)
20. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)
21. Zhu, J., Rosset, S., Hastie, T., Tibshirani, R.: 1-norm support vector machines. Adv. Neural Inf. Process. Syst. **16**(1), 49–56 (2004)
22. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. Ser. B (Stat. Method.) **67**(2), 301–320 (2005)