The penalty term pen denotes the intercept of the hyperplane. The class assignment for a test dataset \( x \) can be achieved by solving 'penalizedSVM' implemented penalization functions \( L \) feature selection methods have been proposed. In the R package gene selection. To overcome this restriction, a number of penalized high-dimension data. Most powerful supervised classification techniques, especially for classification methods play an important role in data analysis in a wide range of scientific applications. Classification methods open to the wide R community. The command help (package=penalizedSVM) points to the available selection methods for SVM classification using penalty functions. We implemented a new quick version of \( L_1 \) penalty (LASSO). The second implemented method, Smoothly Clipped Absolute Deviation (SCAD) was up to now not available in R. Thus, this package makes feature selection SVM methods open to the wide R community. The command help (package=penalizedSVM) points to the available help pages.

2 METHODS

LASSO (\( L_1 \)): the use of a \( L_1 \) penalty function is a common approach in regression originally proposed by Tibshirani (1996). Tibshirani described a technique, called the LASSO for ‘least absolute shrinkage and selection operator’, for parameter estimation. Bradley and Mangasarian (1998) adapted the \( L_1 \)-regularization to SVM. The penalization term has the form

\[
pen_1(w) = \sum_{j=1}^{d} |w_j| = \sum_{j=1}^{d} |w_j|,
\]

As a result of singularity of the \( L_1 \) penalty function, the \( L_1 \) SVM can automatically select genes by shrinking the small coefficients of the hyperplane to exactly zero. Thus, the \( L_1 \) SVM is an effective feature selection tool. Recently, Fung and Mangasarian (2004) have published a fast \( L_2 \) SVM modification using a Newton Linear Programming SVM (NLPSVM). We implemented this modification in the package.

SCAD: it is a non-convex penalty function first proposed by Fan (1997) and discussed in Fan and Li (2001). Zhang et al. (2006) combined SVM with SCAD for feature selection. The penalization term for SCAD SVM has the form

\[
pen(w) = \sum_{j=1}^{d} \psi(w_j),
\]

Often we do not only require low prediction error but also we need to identify covariates playing an important role in discrimination between the classes and to assess their contribution to the classifier. To do so, a number of feature selection methods have been proposed for SVMs.

Feature selection methods can be subdivided into two classes: filter and wrapper methods (Blum and Langley, 1997; Weston et al., 2000). Filter methods, e.g. Recursive Feature Elimination (RFE) (Guyon et al., 2002) drop irrelevant features before the learning algorithm constructs the prediction rule. Wrapper methods provide the selection within the optimization procedure, which increases the prediction power (Markowetz and Spang, 2005).

The R package ‘penalizedSVM’ provides two wrapper feature selection methods for SVM classification using penalty functions.

\[ f(x) = \sum_{j=1}^{d} w_j x_j + b, \]

where \( w = (w_1, \ldots, w_d) \) are the coefficients of the hyperplane and \( b \) denotes the intercept of the hyperplane. The class assignment for a test dataset \( x_{\text{test}} \) is given as \( y_{\text{test}} = \text{sign}(f(x_{\text{test}})) \).

The problem of finding the optimal hyperplane with maximal margin is solved by convex optimization. Maximizing the margin can be achieved by solving

\[
\min_b \sum_{i=1}^{n} (1 - y_i f(x_i)) + \text{pen}(w).
\]

The penalty term \( \text{pen}(w) = \lambda \|w\|^2 \) for SVM has the form of \( L_2 \) norm (‘ridge penalty’), which only shrinks the coefficients but does not set them exact to zero.

*To whom correspondence should be addressed.

© The Author 2009. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
where the SCAD penalty function for each coefficient $w_j$ is defined as

$$ p_S(w_j) = \begin{cases} \lambda_1 |w_j| & \text{if } |w_j| \leq \lambda, \\ \frac{\lambda_1}{\lambda} w_j - \frac{\lambda_1}{2\lambda^2} |w_j|^2 & \text{if } \lambda < |w_j| \leq \lambda_0, \\ \frac{\lambda_1}{2\lambda} w_j^2 & \text{if } |w_j| > \lambda_0, \end{cases} $$

with tuning parameters $a > 2$ and $\lambda > 0$. From Fan and Li (2001), suggested value $a = 3.7$ is used in the package. $p_S(w_j)$ corresponds to a quadratic spline function with knots at $\lambda$ and $\lambda_0$. For small coefficients, the SCAD has the same behavior as the $L_1$. For large coefficients, however, the SCAD applies a constant penalty, in contrast to the $L_1$ penalty, which increases linearly as the coefficient increases. This absolute maximum of the SCAD penalty, which is independent from the input data, decreases the possible bias for estimating large coefficients. The $L_1$ penalty, which is applied in standard SVM without feature selection, is plotted for comparison only (Supplementary Fig. 1).

It has been shown that the SCAD penalty has better theoretical properties as the $L_1$ function (Fan and Li, 2001).

3 AN EXAMPLE SESSION

To illustrate the use of ‘penalized SVM’, an example session with simulated data is shown. First, we simulated training data with $n = 200$ samples, each with $ng = 1000$ features, where only $nsg = 10$ out of 1000 features have an impact on the class labels.

```r
> train<-sim.data(n = 200, ng = 1000, nsg = 10, seed=123)
```

Since SVM algorithms are very sensitive to the choice of the tuning parameters, the user should predefine a possible set of tuning parameters, $\Lambda = \{\lambda_k, k = 1, \ldots, K\}$.

After defining a set of candidate tuning parameters

```r
> Lambda.scad <- seq(0.01 , 0.05 , 0.01)
```

we trained the classifier using the SCAD penalty (for $L_1$ see the help page ?svm.fan)

```r
> fit.scad<- svm.fs(x=train$x), y=train$y, fs.method="scad", lambda1.set=Lambda.scad)
```

The final tuning parameter was selected using generalized cross-validation. Optionally, this can be combined with an outer $k$-fold cross-validation. After training the model, the separating hyperplane was described by only a subset of features. Note that the final predictor could also be the null model, i.e. containing no feature at all. To validate the prediction, we used test data with the same number of features.

```r
> test<- simulate.data(n = 200, ng = 1000, nsg = 10, seed=124)
```

The predicted classes for the test data result from

```r
> predict(fit.scad,new.data=test$x), new.data.labels=test$y)
```

The function `predict(fit)` predicts class labels for a new set of samples. In case the true class labels for test samples are available, the prediction accuracy, misclassification table, sensitivity and specificity are also calculated. In this example, the misclassification error was 21% with sensitivity 74% and specificity 83%.

Depending on sample size, number of tested tuning parameters and eventually the use of outer cross-validation, the calculation time can increase from seconds to several minutes. In the example above, the computation time was 80 s using a desktop computer with 2 GB RAM and 3 GHz CPU speed.

ACKNOWLEDGEMENTS

The development of the SCAD and its original implementation in MATLAB is the work of H. H. Zhang. The MATLAB implementation of the $L_1$ norm SVM was kindly provided by Glenn M. Fung and O. L. Mangasarian.

We are grateful to the editor and the referees for their helpful comments.

Conflict of Interest: none declared.

REFERENCES


