CHAPTER 8

ACCELERATED STABLE SVM-RFE FOR FEATURE SELECTION

8.1 INTRODUCTION

The feature selection for classification is a dynamic research area in data mining and optimization. Among the existing methods the effectiveness of the SVM-RFE algorithm is generally considered remarkable. But the prime difficulty in using is its computational complexity. The problem becomes worse when bootstraps are required to achieve high feature selection constancy. To reduce the computational complexity and to improve the stability of the SVM-RFE two important processes are introduced in the proposed method.

First process is using only the support vectors for training process not all the training samples. This reduces the computational time for building a single SVM classifier drastically. Second process is a new feature ranking criteria, which takes into account both the discriminant ability of individual features and the feature’s ability to maintain the SVM’s margin as large as possible.

Instability in the feature selection process can be overcome by the Cross Validation technique or bootstrapping of the samples which must be performed before the selection process. The process mentioned above will definitely increase the computational time.

Here the reasons for the increase in the computational complexity of SVM-RFE, the feature selection algorithm used in this thesis is discussed. The fundamental factors affecting the computational complexity of SVM-RFE are its backward elimination nature and its greediness in selection. Theoretically searching backwards from the full set of features available may capture interesting features more easily than the forward selection method. But the backward elimination method is computationally expensive as it retrains SVM after removing each feature. It will be worse when the features present in the data sets are in thousands. Likewise, the greediness which is the other factor affecting the
computational complexity of SVM-RFE is achieved by not introducing a feature once again after its removal. There is a chance that the removed feature might influence the selection process in the later iterations and may affect stability of SVM-RFE method. As mentioned before the instability in the selection process can be overcome by bootstrapping or K-fold cross validation techniques which increase the computational complexity. It is thus necessary to find out ways to accelerate the SVM-RFE.

Here a more efficient version of SVM-RFE “Accelerated Stable SVM-RFE” is presented to reduce the computational time of feature selection and improve the generalization ability of the SVM classifier built from the feature subset selected by the proposed method.

8.2 HYPOTHESIS USED IN THE PROPOSED METHOD

By studying the characteristics of SVMs, it can be seen that the optimal hyper plane and the support vectors which determines the decision boundary of SVMs can be used to select the most informative features.

Hypothesis 1: Concentrate more on the samples (Support Vectors) that are close to the hyper plane which are likely to be misclassified after dimensionality reduction.

The support vectors available in the training data set alone decide the decision boundary not all the samples. Hence it is enough to train the SVM using these relatively small numbers of samples. On this basis we propose the modified version of SVM-RFE algorithm which makes SVM discriminative function to learn using the following procedure. During the first iteration all the training samples are used for training and the support vectors are identified. In the subsequent iterations only an active subset (support vectors) rather than the entire training set are dynamically maintained and used. This hypothesis will bring down boundary construction time of the SVM.

Hypothesis 2: Try to eliminate more number of irrelevant features in each iteration before retraining the SVM. Use the feature’s class discriminative ability along with the ability of maintaining the SVM’s margin as large as possible as the new ranking criteria.
The original SVM-RFE uses the squared coefficients $w^2$ of the weight vector $w$ obtained from Eqn. (4.8) as feature ranking criteria. The features having largest weight are assumed to be informative. Thus in each iteration of SVM-RFE the weight vector of all the features is computed by training the SVM classifier and eliminate the feature having the least weight. This process is continued until a feature subset of desirable size is generated. However, this way of feature ranking is a greedy method which may reduce the stability of the feature selection process. So some other factor has to be identified to still strengthen the feature ranking process. From the observations in (Schapire et al. 1998) it is decided that the ability of the features to maintain the SVM’s margin as large as possible can also be used to rank the features. The meaning is that the features which are not reducing the size of the margin are claimed as uninformative ones. It is noticed that while constructing SVM classifier a fat margin around the decision boundary is formed. However the dimensionality reduction may try to shrink the margin after each iteration of the SVM-RFE. The objective is to make the shrink in margin as small as possible. For this in the feature selection process, we should prefer the features which make more contribution to maintaining the margin large are preferred. To comprehend this idea, we bring in a coefficient $CLM_k$ (Coefficient of Large margin)

$$CLM_k = \frac{1}{S^+_n} \sum_{i \in S_+} x_{i,k} - \frac{1}{S^-_n} \sum_{j \in S_-} x_{j,k} \quad \ldots \ (8.1)$$

Where $S_+$ denotes the Support Vectors belonging to positive samples, $S_-$ denotes the Support Vectors belonging to negative samples, $S^+_n$ denotes the number of $S_+$ and $S^-_n$ denotes the number of $S_-$, and $x_{i,k}$ denotes the $k^{th}$ feature of Support Vector $i$ in the input data set. The larger $CLM_k$ indicates that the $k^{th}$ feature can make more contribution in maintaining the margin large. Therefore, $CLM_k$ can assist $|w_k|$ for feature ranking. The new criterion for ranking the features is $CLM_k \cdot |w_k|$.

8.3 PROPOSED ACCELERATED STABLE SVM-RFE USING HYPOTHESIS 1 AND 2

N training samples are considered with $k$ features each and their corresponding class labels. Initially train the SVM using all the features of all the training samples like the conventional SVM-RFE. After training SVM with these N sample and label pairs, the
discriminative function value \( f(X_i) \) for the Support Vectors can be found and optimal weight vector \( W \) of each feature as per the hypothesis 1. Now as per the hypothesis 2, to improve the generalization ability of the classifier, the parameter \( CLM_K \) is computed for all the features of support vectors. Since \( y_i f(X_i) \) of the support vectors control the Maximum Margin Distribution of the SVM calculate \( y_i f(X_i) \) for all the Support Vectors and store it in a linear Vector called Maximum Margin Distribution Vector (MMDV). In the next step remove maximum number of features before retraining the SVM which are not drifting the margin than that of a predefined threshold limit as stated in the hypothesis 2. The method also taking care of not losing important features by trying to remove more number of features in the initial iterations and gradually reducing the features in the later iterations. The working of the proposed method is depicted by the flow chart figure 8.1. The outline of the algorithm is given below.

(i) Initialization.

An empty ranked feature list \( RANK_{list} \).
The entire feature set to be ranked \( F = \{1,2,\ldots,K\} \)

(ii) Set \( l = K/2 \)

(iii) Train the SVM with \( K \) features of the \( N \) training sample set and obtain its optimal weight vector

\[
W_{[K]} = [w_1, w_2, w_3, \ldots, w_K], CLM_{[K]} = [CLM_1, CLM_2, \ldots, CLM_K] \text{ and bias } b_k.
\]

(iv) Rank the \( K \) features in the descending order of \( W(i) \cdot CLM(i), i = 1,2,\ldots,K \).

(v) Compute discriminative function for the Support Vectors using

\[
f(X_{i^{[k]}}) = \langle W_{[K]}^i, X_i \rangle + b_k, i = 1,2,\ldots,N \text{ Where } \alpha_i > 0
\]

The Margin Distribution of the SVM\( MD_{k}^n \) using

\[
y_i f(X_{i^{[k]}}), i = 1,2,\ldots,N \text{ Where } \alpha_i > 0
\]

(vi) Set \( W_{[K-j]} = [w_{i+1}, w_{i+2}, w_{i+3}, \ldots, w_K] \), \( CLM_{[K-j]} = [CLM_{i+1}, CLM_{i+2}, \ldots, CLM_K] \)

and \( X_{i^{[k-j]}} = [x_{i}^{j+1}, x_{i}^{j+2}, \ldots, x_{i}^{K}], i = 1,2,\ldots,n \) where \( n \) is the number of support vectors

(vii) Compute \( MD_{k-j}^n \) using \( y_i f(X_{i^{[k-j]}}), i = 1,2,\ldots,N \text{ Where } \alpha_i > 0 \)
(viii) Calculate $\Delta MD_k = \| MD_k^n - MD_{k-1} \|

(ix) If $\Delta MD_k > TH_{AMMD}$, let $l = l/2$ go to step (v)

(x) Else remove $l$ least significant features and place them on the top of $RANK_{list}$

(xi) Set $K = K - l$

(xii) If $K > 1$ go to step (iii) else display $RANK_{list}$ and Stop.
Start

Initialize RANK\textsubscript{list}
K --- No. of features to be Ranked

Train the SVM with N samples of K features

Obtain the Margin Distribution Vector \( MDF_K \)

Set \( l = k/2 \)

Obtain the Margin Distribution Vector \( MDF_{K-l} \)

\[ \| MDF_K - MDF_{K-l} \| > TH_{\text{MMD}} \]

Yes

Store the \( l \) features to the ranked list
Set \( K = K - l \)

No

Yes

\( K > 1 \)

No

Display the final ranked list RANK\textsubscript{list}

Stop

Figure 8.1 Flow chart of the accelerated stable SVM-RFE
8.4 DESCRIPTION OF THE DATA SET USED

The proposed method is tested for gene selection and classification of DNA microarray data. DNA microarray is a technology that measures expression levels of thousands of genes simultaneously. As microarray data contains thousands of features of which many are irrelevant or redundant it was found to be a good candidate to test the proposed method. The details of the four data sets used to test the performance of the proposed method are given in Table 8.1. The description of the breast cancer data set is available in section 7.3. The microarray for Acute Myeloid Leukemia (AML) and Acute Lymphoblast Leukemia (ALL) data set have the expression level of 7129 gens of 72 samples out of which 47 are ALL samples and 25 are AML samples (Golub et al. 1999). The colon cancer dataset has the expression levels of 2000 genes from 62 samples (Alon et al. 1999), likewise the Prostate cancer dataset has the expression levels of 12600 genes from 102 samples (Singh et al. 2002). To reduce the complexity of computation, an initial filtering was carried out using BRB Array Tool on all the four datasets to eliminate the irrelevant noisy genes by setting $p<0.001$ and 3 fold change.

**Table 8.1 Details of the data sets used**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>No. of Samples</th>
<th>Class 1</th>
<th>Class 2</th>
<th>No. of genes</th>
<th>Training Set</th>
<th>Testing Set</th>
<th>No. of genes after initial filtering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer dataset</td>
<td>86</td>
<td>43</td>
<td>43</td>
<td>22285</td>
<td>27</td>
<td>21</td>
<td>16 22</td>
</tr>
<tr>
<td>AML-ALL dataset</td>
<td>72</td>
<td>47</td>
<td>25</td>
<td>7129</td>
<td>25</td>
<td>13</td>
<td>22 12</td>
</tr>
<tr>
<td>Colon cancer dataset</td>
<td>62</td>
<td>22</td>
<td>40</td>
<td>2000</td>
<td>11</td>
<td>20</td>
<td>11 20</td>
</tr>
<tr>
<td>Prostate Cancer Dataset</td>
<td>102</td>
<td>52</td>
<td>50</td>
<td>12600</td>
<td>32</td>
<td>30</td>
<td>20 20</td>
</tr>
</tbody>
</table>

Note: AML- ALL data set Class1-ALL Class2-AML and other two data sets class 1- Normal class 2-cancerous

94
8.5 RESULTS AND DISCUSSION

Description of the data sets used in the experiments is given in Table 8.1. All the four data sets were randomly divided into training sets and testing sets. The training sets were used to identify the class discriminative genes and to train the SVM. The testing sets were employed on predicting the generalization ability of the trained SVM.

8.5.1 Comparison of Accelerated Stable SVM- RFE with other variants of SVM-RFE

The figure 8.2 gives the average classification error on the 43 testing samples. The classification error was computed for various training set sizes for the breast cancer data set. The performance of the proposed method along with the conventional SVM-RFE and E-RFE and RFE-Anealing (other variants of SVM-RFE) are also presented for comparison. It is very clear that the proposed method outperformed the other methods especially when the training set is small. As more and more number of training samples is used SVM-RFE also performed well like the proposed method. But comparing the other two methods E-RFE and RFE-Anealing which also removed more number of features in each iteration of feature selection, the proposed method gives appreciable generalization ability. Even on the other three data sets similar results are observed. It is a good indication that if the proposed method is adopted for cancer classification definitely the life of many cancer patients will be saved.

Figure 8.2 Comparison of feature selection methods on breast cancer data set
8.5.2 Comparison of the generalization ability of Accelerated Stable SVM-RFE with SVM-RFE

The Figures 8.3(a), 8.3(b) and 8.3(c) show the performance comparison of the proposed accelerated stable SVM-RFE and the original SVM-RFE on the breast cancer data set, AML-ALL dataset and colon cancer data set respectively. The experiments were conducted on both the training dataset and also the testing data set. For example in the case of the breast cancer data set the proposed algorithm was first trained using the 43 training samples. Here Leave One Out Cross Validation (LOOCV) is used to train the SVMs. At each of the 43 iterations 42 samples were used for selecting the top 10 genes and to train the SVM classifier, and the remaining 1 sample was used to test generalization ability of the classifier built. The classification error is the average of the classification errors on the test sample (1 sample) in the 43 iterations. In the testing phase top 10 genes were selected from the 43 training samples and the generalization ability of the classifier built was measured on the 43 testing samples.

Throughout the experiments it is noticed that the Accelerated Stable SVM-RFE has shown better performance compared to the original SVM-RFE on training as well as the testing data set. It was also observed that initially when the number of features was less the classification error was more and as the number of features increased it gradually got reduced and it was steady when the number of features was 5 to 10. The inference is that the top 10 genes identified can improve the classification accuracy of the classifier. This is observed in case of both the training and the testing data set. When we compare on the whole, the performance of both the algorithms the training set has given better performance than the testing set. The reason is that for the trained classifier the test samples are unknown samples. On the other hand the test sample used during the training phase would have been a training sample of some other iteration. The inference is that some more additional measures should be considered to improve the classification accuracy further.
Figure 8.3 (a) The comparison of the generalization ability of Accelerated stable SVM-RFE and SVM-RFE on Breast Cancer data set.

Figure 8.3 (b) The comparison of the generalization ability of Accelerated stable SVM-RFE and SVM-RFE on AML-ALL Data set.
Table 8.2 Performance Comparison between Accelerated Stable SVM- RFE and SVM-RFE

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Classification accuracy (%)</th>
<th>Run Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>Accelerated Stable SVM- RFE</td>
<td>SVM-RFE</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>98.8 97.4</td>
<td>92.9 92.4</td>
</tr>
<tr>
<td>AML-ALL dataset</td>
<td>97.4 96.4</td>
<td>96.3 95.4</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>95.3 94.1</td>
<td>84.8 84.8</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>96.1 94.8</td>
<td>94.2 94.0</td>
</tr>
</tbody>
</table>

The Table 8.2 list out the detailed result of the experiments conducted. The average computational time taken by the Accelerated Stable SVM-RFE and the original SVM-RFE in a single run is tabulated under Run time. It is in Seconds. The proposed method took
much shorter time than original method. The reasons are that the original method spent much of the time in training the classifiers with all the training set where as the proposed method carried out the training process only on a small dataset (support Vectors). Likewise the proposed method eliminated more number of irrelevant features compared to one feature by the original method per iteration. The last column shows the speedup gained by the proposed method. It is noted that the speed up factor is not less than 54. The variation in the speed up factor is because of the number of Support Vectors involved in the training in different data sets.

It is also noticed that the accelerated stable SVM-RFE has shown a performance improvement of 1% compared to the original SVM-RFE. This is seen in during the training as well as the testing phase. The tabulated results evidentially show that the proposed method has improved the efficiency of SVM-RFE without sacrificing the accuracy of either feature selection or the classification method.

8.5.3 Verification of the Reduction of Computational Complexity of Accelerated Stable SVM- RFE

As discussed before in the hypothesis 1 and 2, the highlighting features of the Accelerated Stable SVM-RFE are the usage of only the Support Vectors (SV) for training and not all training samples and elimination of more number of features per iteration. These two factors play a crucial role in the reduction of computational complexity of the proposed method compared to the original SVM-RFE. This can be easily understood from the computational complexity of the original SVM-RFE. It is stated as $O(\max(n,m)m^2)$, where $n$ is the number of iterations and $m$ is the number of features. It is apparent that as the number of features and number of samples decreases the computational complexity also decreases. In order to prove that the reduction of training set helps in decreasing the computational complexity plotted the change in the training dataset Reduction Factor (RF) against iterations has been plotted in figure 8.4. The RF is calculated using Eqn. 8.2 where $NSV_i$ is the Number of Support Vectors involved in $i^{th}$ iteration of training phase and $TNS$ is the Total Number of Samples in the training set.

$$RF(i) = \frac{NSV_i}{TNS} \quad (8.2)$$
Figure 8.4(a) Change of Reduction Factor during different iterations of the Accelerated Stable SVM-RFE for breast cancer dataset

Figure 8.4(b) Change of Reduction Factor during different iterations of the Accelerated Stable SVM-RFE for AML-ALL dataset
Figure 8.4(c) Change of Reduction Factor during different iterations of the Accelerated Stable SVM-RFE for Prostate cancer dataset

Figure 8.4(d) Change of Reduction Factor during different iterations of the Accelerated Stable SVM-RFE for colon cancer dataset
The change of reduction factor during different iterations of the Accelerates Stable SVM-RFE for feature selection from the breast cancer dataset, AML-ALL dataset, prostate cancer dataset and the colon cancer dataset are given in figures 8.4(a), 8.4(b), 8.4(c) and 8.4(d) respectively. From the graphs it is noticed that the value of RF varies from data set to data set. This is the reason for the variation in the speedup factor of the different data sets as given in table 8.2. It is also observed that only during the initial iteration the value of RF is one because in the initial iteration all the training samples are involved in training, following this in the subsequent iterations the value is very less and the change is also very less. The inference is that from the second iteration onwards the number of samples (SV) participating in the training process is very less compared to the original training set. This is the first evidence for the reduction of computation complexity of the proposed Accelerated Stable SVM-RFE. In the proposed algorithm it is seen that care is taken to reduce more number of irrelevant features during the initial iterations which is the second evidence for the reduction of computational complexity of the proposed method.

8.5.4 Verification of the stability of Accelerated Stable SVM- RFE

The proposed Accelerated Stable SVM-RFE has shown stability in the feature selection process. To verify that, an experiment is conducted on the breast cancer dataset. Figure 8.5 shows the total number of times each feature has been selected over 10 trials. The figure is plotted against Gene symbol versus number of times a gene being selected. From the figure it is seen that the genes with symbols KRT19, CD24, KRT7, GATA3, AGR2, TACSTD2 are selected all the 10 times where as ANK3 and KRT14 are selected one time and C8ORF4 and MFAP5 are not at all selected in the 10 trials. This clearly indicates that the proposed method gives very good stability in terms of feature selection also.
The stable feature selection framework proposed in the chapter 7 has improved the stability of the feature selection process. On the other hand the computational complexity has increased. To address this issue an Accelerated Stable SVM-RFE, a modified version of SVM-RFE is proposed and tested on four microarray gene expression data sets. This method has taken into account only the support vectors which are the reduced subset of the training data set for training SVM. The result shows that the speedup gained by the proposed method is not less than 54% compared to the original SVM-RFE. Similarly to enhance the stability of the feature selection process and also to improve the predictive ability of the classifier built a new feature ranking criteria is introduced. The results of the experiments show that the proposed method has improved the stability and in case of the classification accuracy the proposed method has shown an improvement of 1% over the original SVM-RFE.

The summary is that the proposed accelerated stable SVM-RFE method has remarkably speeded up the feature selection process and its stability mean while improved the generalization ability of the classifier built.