Chapter 6

Probabilistic Approach in Feature Selection

Contents

In this chapter a new algorithm of feature selection using probabilistic approach is developed. The main objective is to reduce number of features before applying the classification technique to avoid the risk of over fitting and hence to get better generalisation accuracy. The proposed algorithm named Filter technique and Partial Forward Search (F_PFS) algorithm is presenting a new weighted probabilistic approach for feature selection and decides the best models of support vector machines(SVMs) to diagnose various skin diseases.

In section 6.1, importance of feature selection techniques are discussed in brief. Sec-

tion 6.2 discusses various feature selection techniques and related work. In section 6.3 a novel method (Weighted Probabilistic Approach) for Feature Selection named F_PFS is developed and provided algorithm of the new approach [101]. Experimental setup, Experiments and results to assess the effectiveness of the new algorithm is showcased in section 6.4, which is followed by the summary of the chapter in section 6.5.

6.1 Introduction

Dimensionality reduction is an active field of research in statistical learning theory and in machine learning for classification problem. The objective of the dimensionality reduction is to remove noisy (irrelevant features) and redundant features. In many applications, datasets are having large pool of features. In these types of datasets, usually dimensionality of the set is very high but available samples for training the data for classification are small in numbers. When we use classifier such as SVM for such types of datasets, then classification problem suffers from over fitting. Feature extraction and feature reduction are dimensionality reduction techniques. Feature extraction techniques are considering combination of original features and produces a new space of features with lower dimensionality. Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Canonical Correlation Analysis (CCA) are some feature extraction techniques. But, there is no physical meaning of the new features and hence analysis becomes difficult [126]. Another technique of dimensionality reduction is feature selection techniques. In this technique a new space of features from the original feature set with a small number of discriminative features is obtained without any transformation. In this technique physical meaning of original features is maintained, so considered better than feature extraction techniques. Many times features in the data sets are highly correlated, which unnecessary increase the dimensionality. Also, some of the features in the data sets are not correlated to class (output), and so increases noise. Therefore, as a pre-step of classification, an appropriate model with proper feature selection is required to reduce over fitting problem, to reduce computational cost and to increase accuracy, especially when model construction uses classifier such as support vector machine.

6.2 Techniques of Feature Selection

There are three main techniques of feature selection:

1. **Filter Method**:
   The Filter method is one of the feature selection techniques, in which feature selection is carried out using various statistical measures, without using any machine learning algorithm or classifier. So, it finds the subset of feature set without involving any learning algorithm. In this method each individual feature is evaluated and the less interesting features are suppressed. Filter methods use some statistical measures to give ranks to features and use a threshold to obtain a subset of feature set [17]. Filter methods such as correlation based filter technique ([45],[141]) and mutual information technique ([9], [45], [70], [75], [141]) use measure of dependency between two attributes to determine rank. Shen *et. al.* have given ranks to all features using posterior probabilistic approach [122]. Many feature selection algorithms emphasis on finding correlation between features and labels and established optimal set of relevant features. But, relevance of feature does not mean that it is in optimal subset of features. Similar is also true for irrelevance of features. Many machine learning algorithms such as induction of decision tree algorithm, instance-based algorithm are facing the problem of irrelevant features. Their performance degrades if irrelevant features are added in the set of features. If more irreverent features are added to the feature set, the algorithm such as Naive-Based, accuracy does not change significantly but if correlated features are added then its performance affects. Relief is another algorithm which searches not only most relevance features but extracts both weak and strong relevant features [70]. Filter method index is calculated based on single feature without considering orthogonality between features, which is one of the weaknesses of the methods [68]. The model constructed from this techniques are faster and more general as it does not involve any classifier. But, sometimes it can not effectively make the feature data space smaller.

2. **Wrapper Method**:
   The Wrapper is another method of feature selection. It uses classifier performance as an objective function to evaluate feature. It conducts a search for best feature subset using induction algorithm. Sequential selection search al-

gorithms and heuristic search algorithms are two main techniques of wrapper methods. Sequential selection search algorithms include forward sequential search procedure and backward sequential search procedure. Forward sequential search procedure starts with empty set, add one feature every time and for each subset finds accuracy using some classifier. The subset of feature set which is giving the optimum value of the objective function under study is considered as the best subset of the feature set. Backward sequential search procedure starts with full set, removes one feature in every step and finds the best model. Another technique, which is heuristic search algorithm evaluates different subsets to find optimum value of the objective function. In this method for n features, the size of the search space is $O(2^n)$, which is a NP-hard problem [17]. Kohavi and John presented a more formal discussion of this kind of methodology by introducing variability using various classifiers and search strategies [70]. Filtered and Supported Sequential Forward Search (FS_SFS) algorithm takes into account both the discriminate ability of individual features and the correlation between them. It filter out nonessential features and reduces search space [79]. Combination of filter method and wrapper method i.e. Hybrid feature selection method is used by Xie *et al.* [136]. Wrapper methods are very slow therefore to reduce computational cost for larger data sets, less number of folds can be used [70]. Due to these limitations sometimes we may not find the best values of parameters.

3. **Embedded Method**:

   Third approach of feature selection is embedded methods, which induces combine advantage of both filter and wrapper techniques. It use prior knowledge and then use classifier to construct the model with less number of features. Selected features are sensitive to the structure of the classifier [68]. Classification trees, random forests, Least Angle Regression, Recursive feature elimination are some examples of embedded methods [68]. For training data set consisting of labeled as well as unlabelled data, Xu *et al.* have suggested semi supervised learning algorithm [139]. They used the embedded feature selection method to extract information about unlabelled data. Duval *et. al.* have used an embedded approach used for classification of microarray datasets [34]. The algorithm is a combination of problem specific cross over operator and dedicated local search procedure. Embedded methods are similar to Wrapper methods, but less computationally expensive and less prone to over fitting.

6.3 A Novel Approach for Feature Selection

We develope a new algorithm of feature selection, to be called F_PFS Method uses good features of both filter and wrapper techniques [101]. It uses weighted probability of each feature to assign rank. If a feature frequently occurs in the dataset, it is considered as a high probability feature and which indicates its importance in the prediction. In any diagnosis, the common symptoms (features) are focused first. The common features are the features having high probability. So, we have included all common features in our base model. This method works for both balanced as well as for imbalanced data sets. It can be applied to multiclass data classification also. The algorithm not only works for binary inputs, but also works on datasets in which inputs are given in scale according to the intensity of features.

The method is divided into three phases.

1. In the first phase Filter Technique of feature selection is used and determine weighted probability of each feature.
2. In the second phase the features are arranged in the descending order of weighted probability value and its average is determined, which is called threshold value. Finally obtain the base model which includes only those features whose weighted probabilities are more than the threshold.
3. In the third phase Support Vector Machine is used as classifier to find the best model. Wrapper method is started with the base model and use Forward Sequential Search algorithm.

Proposed Algorithm: Take a training set $\{\mathbf{x}_i, \mathbf{y}_i\}$, where each $\mathbf{x}_i$, $i = 1, 2, ..., m$ be the n dimensional vector indicating n features $f_1, f_2, ..., f_n$ in each sample, m be the total number of training samples and $\mathbf{y}_i$ be the corresponding class label taking the values $i = 1, 2, 3, ..., NC$ where NC indicate the number of classes. r_i, $i = 0, 1, 2, ..., l$ be the score given to each feature in the data set according to the intensity of the feature, where l is an integer indicating highest score (maximum

intensity) of the feature. Let d_k denote the number of training instances in the k^{th} class where, $k = 1, 2, ..., NC$.

1. For each class k, $k = 1, 2, ..., NC$ find total number of scores n_i, $i = 0, 1, 2, ..., l$ corresponding to r_i, $i = 0, 1, 2, ..., l$ respectively for the j^{th} feature, $j = 1, 2, ..., n$.
2. Find $R = \sum_{i=0}^{l} r_i$ which is total of scores given to each feature.
3. For each feature the probability of r_i, $i = 0, 1, 2, ..., l$ for the class k, $k = 1, 2, ..., NC$ be $p_{ri} = \frac{n_i}{d_k}$, $i = 0, 1, 2, ..., l$.
4. Then the probability of the j^{th} feature for the class k be $p_k = \sum_{i=0}^{l} \left(\left(\frac{r_i}{R} \right) p_{ri} \right)$, $k = 1, 2, ..., NC$ $\left(\sum_{i=0}^{l} \frac{r_i}{R} = 1 \right)$.
5. Calculate the weight of j^{th} feature for class k as $w_k = \frac{\frac{m}{d_k}}{\sum_{k=1}^{NC} \frac{m}{d_k}}$.
6. Weighted probability of the j^{th} feature is $p_j = \sum_{k=1}^{NC} w_k p_k$, $j = 1, 2, ..., n$.
7. Find the threshold which is the average of the weighted probability of n features for the entire training dataset. i.e. $T = \sum_{j=1}^{n} \frac{p_j}{n}$.
8. Arrange the features in the decreasing order of weighted probabilities.
9. Set the base model as the subset of feature set including only those features whose weighted probabilities are more than threshold value.
10. Apply Partial Forward Sequential Search Algorithm which starts finding accuracy of the base model using Support Vector Machine (SVM). Radial Bases Function (RBF) $exp\left(-\gamma \|x - y\|^2\right)$ is used as kernel function in SVM.
11. Add one feature at each step with weighted probability just lower than that of the feature added in the previous step. Each time find the accuracy of the model obtained by adding new feature using SVM learning algorithm.
12. Compare accuracy of all models and select that model as the best model which gives the highest classification accuracy.

6.4 Experiments and Results

For empirical verification, the proposed algorithm is applied to two skin datasets: Dataset-I and Dataset-II (Appendix-A). SVM is implemented using LIBSVM-3.18 [76]. The 10 folds cross validation (refer 2.4.8) criteria is used to set values of the parameter of RBF kernel and regularization parameter C of the SVM optimization problem (4.2.14) in each case. For both dataset experiments are carried out on 60-40%, 70-30% and 80-20% training-testing data partitions. Radial Basis Function (RBF) is used as kernel function which is defined as $exp\left(-\gamma \left\|x-y\right\|^2\right)$.
Table 6.1 and Table 6.2 show the results of our proposed method applied to the Dataset-I and Dataset-II respectively. The comparison of the F_PFS algorithm with IFSFS (Improved F-score and sequential forward search algorithm) discussed is given in Table 6.3 [136].

Improved F-score:

This method is discussed in [136]. We have make comparison of this technique with our feature selection algorithm called F_PFS. For m training samples and k(k $\geq$ 2) number of classes, where $n_j, j = 1, 2, ...k$ be the number of samples in j^{th} class, the improved F-score for the i^{th} feature is defined as:

$$F_i = \frac{\sum_{j=1}^{k}(\bar{x_i}^{(j)} - \bar{x_i})^2}{\sum_{j=1}^{k}\frac{1}{n_j-1}\sum_{l=1}^{n_j}(x_{l,i}^{(j)} - \bar{x_i}^{(j)})^2} \tag{6.4.1}$$

where, $\bar{x_i}^{(j)}$ and $\bar{x_i}$ are the averages of the ith feature of the whole dataset and j^{th} dataset respectively. $x_{l,i}^{(j)}$ is the i^{th} feature of the l^{th} sample in the j^{th} dataset. The algorithm is applied Dataset-I and the corresponding results are discussed in Table 6.3. In IFSFS algorithm filter method is used, in which ranking of features of Erythomoto Squamous skin disease is assigned using improved F-score. Then wrapper method is applied to find the best subset of feature set which gives the best accuracy.

Table 6.1: F_PFS Method applied on DataSet-I

Model #	Total Features	Selected Features	SVM Classification Accuracy		
			60-40%	70-30%	80-20%
1	19	3,20,4,31,30, 5,14,8,38,13, 9,11,17,7,1, 21,16,26,32	77.13%	80.14 %	77.66 %
2	20	Model #1 + feature # 24	74.47%	82.27%	74.47%
3	21	Model #2 + feature # 25	79.26%	82.98%	75.53%
4	22	Model #3 + feature #43	85.65%	84.40%	76.60%
5	23	Model #4 + feature # 23	82.45%	86.52%	77.66%
6	24	Model #5 + feature # 27	77.13%	85.82%	78.72%
7	25	Model #6 + feature # 47	82.45%	84.40%	80.85%
8	26	Model #7 + feature # 37	80.85%	84.40%	84.04%
9	27	Model #8 + feature # 6	84.57%	87.23%	81.91%
10	28	Model #9 + feature # 10	82.45%	86.52%	80.85%

11	29	Model #10 + feature # 40	82.98%	85.12%	84.04%
12	30	Model #11 + feature # 42	86.70%	84.40%	86.7%
13	31	Model #12 + feature # 33	85.64%	85.11%	86.70%
14	32	Model #13 + feature # 29	86.70%	82.98%	86.70%
15	33	Model #14 + feature # 28	81.91%	84.40%	87.23%
16	34	Model #15 + feature # 34	82.45%	84.40%	87.23%
17	35	Model #16 + feature # 2	**86.70%**	**89.36%**	**88.30%**
18	36	Model #17 + feature # 12	86.70%	86.52%	84.04%
19	37	Model #18 + feature # 35	85.64%	88.65%	84.04%
20	38	Model #19 + feature # 18	84.04%	89.36%	86.17%
21	39	Model #20 + feature # 36	84.04%	87.23%	84.04%
22	40	Model #21 + feature # 39	84.57%	89.36%	84.04%
23	41	Model #22 + feature # 41	84.57%	87.23%	85.11%

24	42	of Model #23 + feature # 22	84.57%	87.23%	84.04%
25	43	Model #24 + feature # 44	83.51%	87.23%	86.17%
26	44	Model #25 + feature # 15	82.98%	87.23%	86.17%
27	45	Model #26 + feature # 45	85.64%	89.36%	84.04%
28	46	Model #27 + feature # 46	84.57%	87.23%	81.92%
29	47	Model #28 + feature # 19	85.64%	87.23%	81.92%

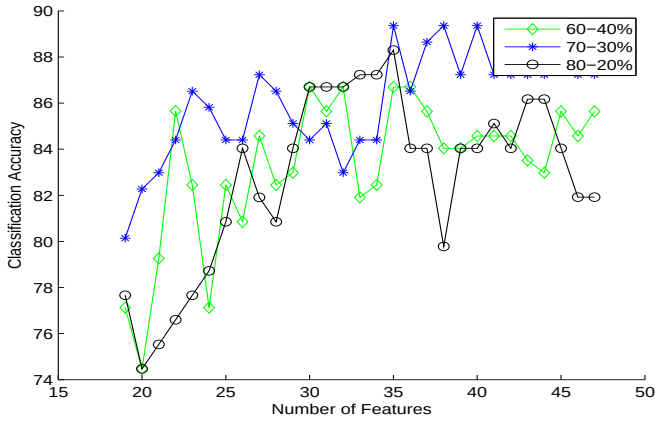

Figure 6.1: Graph of F_PFS Method applied to Dataset-I

Table 6.2: F_PFS Method applied on Dataset-II

Model #	Total Features	Selected Features	SVM Classification Accuracy		
			60-40%	70-30%	80-20%
1	13	1,17,32,2,16,28, 3,19,4,7,31,9,30	84.93%	79.09%	93.15%
2	14	Model #1 + feature # 18	86.30%	84.55%	90.41%
3	15	Model #2 + feature # 21	86.99%	88.18%	94.52%
4	16	Model #3 + feature #5	90.41%	92.73%	94.52%
5	17	Model #4 + feature # 15	93.15%	94.55%	94.52%
6	18	Model #5 + feature # 33	89.73%	95.45%	95.89%
7	19	Model #6 + feature # 10	92.46%	95.45%	95.89%
8	20	Model #7 + feature # 14	**93.15%**	**97.27%**	**95.89%**
9	21	Model #8 + feature # 24	93.15%	96.36%	95.89%
10	22	Model #9 + feature # 27	91.78%	92.73%	94.52%

11	23	Model #10 + feature # 6	91.78%	92.73%	94.52%
12	24	Model #11 + feature # 25	91.78%	92.73%	94.52%
13	25	Model #12 + feature # 11	91.10%	92.73%	94.52%
14	26	Model #13 + feature # 12	91.10%	92.73%	94.52%
15	27	Model #14 + feature # 8	89.04%	92.73%	94.52%
16	28	Model #15 + feature # 20	89.73%	90.91%	94.52%
17	29	Model #16 + feature # 26	92.47%	92.73%	93.15%
18	30	Model #17 + feature # 22	93.15%	90.91%	93.15%
19	31	Model #18 + feature # 23	89.73%	90.91%	94.52%
20	32	Model #19 + feature # 13	91.78%	92.73%	94.52%
21	33	Model #20 + feature # 24	91.78%	92.73%	94.52%

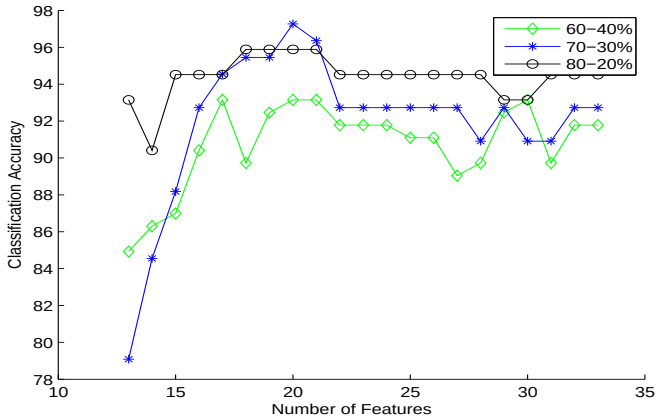

Figure 6.2: Graph of F_PFS Method applied to Dataset-II

Table 6.3: IFSFS applied on Dataset-I

Model #	Total Features	Selected Features	SVM Classification Accuracy		
			60-40%	70-30%	80-20%
1	16	32, 17, 21, 16, 14, 25,47, 38, 24, 27, 1, 26, 15, 20, 13, 28	80.32%	83.69%	85.11%
2	17	Model #1 + feature # 23	80.32%	83.69%	82.98%
3	18	Model #2 + feature # 5	82.45%	81.56%	86.17%
4	19	Model #3 + feature #30	86.70%	83.69%	80.85%
5	20	Model #4 + feature # 33	83.51%	83.69%	82.98%
6	21	Model #5 + feature # 45	82.98%	83.69%	80.85%

7	22	Model #6 + feature # 3	84.57%	82.98%	80.85%
8	23	Model #7 + feature # 22	83.51%	81.56%	80.85%
9	24	Model #8 + feature # 31	86.17%	82.98%	84.04%
10	25	Model #9 + feature # 4	84.04%	83.69%	84.04%
11	26	Model #10 + feature # 29	85.11%	87.23%	85.11%
12	27	Model #11 + feature # 35	85.11%	86.53%	84.04%
13	28	Model #12 + feature # 2	84.57%	86.53%	84.04%
14	29	Model #13 + feature # 41	85.11%	85.11%	82.98%
15	30	Model #14 + feature # 18	84.04%	86.53%	85.11%
16	31	Model #15 + feature # 44	85.11%	84.40%	85.11%
17	32	Model #16 + feature # 8	85.64%	86.53%	84.04%
18	33	Model #17 + feature # 12	85.64%	87.23%	84.04%
19	34	Model #18 + feature # 40	86.17%	85.82%	84.04%

20	35	Model #19 + feature # 10	86.17%	86.53%	82.98%
21	36	Model #20 + feature # 37	85.64%	86.53%	82.98%
22	37	Model #21 + feature # 42	85.11%	86.53%	85.11%
23	38	Model #22 + feature # 11	**85.11%**	**89.36%**	**85.11%**
24	39	Model #23 + feature # 43	86.17%	88.65%	85.11%
25	40	Model #24 + feature # 9	86.17%	87.94%	87.23%
26	41	Model #25 + feature # 39	85.64%	88.65%	87.23%
27	42	Model #26 + feature # 6	87.23%	87.23%	87.23%
28	43	Model #27 + feature # 19	87.23%	87.23%	87.23%
29	44	Model #28 + feature # 34	84.57%	87.23%	86.17%
30	45	Model #29 + feature # 46	85.11%	87.23%	86.17%
31	46	Model #30 + feature # 36	86.17%	87.23%	86.17%
32	47	Model #31 + feature # 7	85.11%	87.23%	86.17%

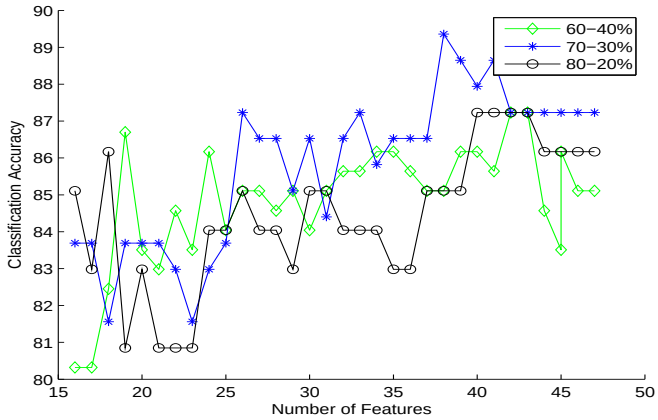

Figure 6.3: Graph of IFSFS method applied to Dataset-I

Result Analysis:

Table 6.1 and Table 6.2 display the results of our proposed method applied to the Dataset-I and Dataset-II respectively. For comparison purpose of proposed F_PFS algorithm with that of IFSFS algorithm, Improved F-score formula is applied on Dataset-I to assign rank and corresponding results are discussed in Table 6.3.
For Dataset-I, 19 features have weighted probability greater than threshold value (average weighted probability) and that of for Dataset-II, there are 13 features. i.e. base model for Dataset-I contains 19 features out of total 47 features and that of Dataset-II 13 features out of total 33 features.

From Table 6.1 it is observed that by applying Wrapper method to Dataset-I, the highest accuracy achieved is 89.36% for 35 features (model #17) for 70-30% data partitions. For the same model accuracies are achieved as 86.70% and 88.30% for 60-40% and 80-20% data partitions respectively.

When F_PFS algorithm is applied on Dataset-II (Table-6.2), highest accuracy achieved is 97.27% for 20 features out of 33 features (model #8) by taking 70-30% data partitions. When IFSFS algorithm is applied on the Dataset-II, it achieved highest accuracy of 94.44% for the 70-30% training-testing data partitions. This reveal the efficiency of newly developed algorithm F_PFS. When IFSFS method is applied to Dataset-I (Table-6.3), the same highest accuracy of 89.36% is obtained for 70-30% data partitions for subset of feature set consist of 38 features, while from F_PFS it is achieved for 35 features only.

To give more clarity and easy analysis of the work, graphical representation of results discussed in Tables 6.1, 6.2 and 6.3 are given in Fig 6.1, Fig 6.2 and Fig 6.3 respectively.

6.5 Summary

In this chapter a novel hybrid feature selection technique(algorithm) is presented. This technique exploits the advantage of the feature selection techniques, Filter and Wrapper methods. During Filter phase, the novel algorithm uses weighted probability approach to assign rank to each feature. Using Wrapper method, classifier is trained starting from the subset of the training dataset, which contains the features, whose ranks are more than average rank. Using forward sequential search algorithm, the classification accuracy of the entire dataset is obtained using SVM as classifier. The novel approach of feature selection (F_PFS) is applied on two different Skin Datasets described in the Appendix-A. The results show that new approach of feature selection(F_PFS) reduces 26% features from Dataset-I and 39% features from Dataset-II with good classification accuracies for both datasets and also reduces computational efforts due to the concept of base model.