JMLR: Workshop and Conference Proceedings 11 (2010) 142–151 Workshop on Applications of Pattern Analysis

A Novel Hybrid Feature Selection Method Based on IFSFFS and SVM for the Diagnosis of Erythemato-Squamous Diseases

Juanying Xie[1,2] XIEJUANY@SNNU.EDU.CN
Weixin Xie[1,3] WXXIE@SZU.EDU.CN
Chunxia Wang[2] JYXIE@163.COM
Xinbo Gao[1] XBGAO.XIDIAN@GMAIL.COM
1 School of Electronic Engineering, Xidian University, Xi'an 710071, P. R. China.
2 School of Computer Science, Shaanxi Normal University, Xi'an 710062, P. R. China.
3 School of Information Engineering, Shenzhen University, Shenzhen 518060, P. R. China.

Editors: Tom Diethe, Nello Cristianini, John Shawe-Taylor

Abstract

This paper developed a diagnosis model based on Support Vector Machines (SVM) with a novel hybrid feature selection method to diagnose erythemato-squamous diseases. Our hybrid feature selection method, named IFSFFS (Improved F-score and Sequential Forward Floating Search), combines the advantages of filters and wrappers to select the optimal feature subset from the original feature set. In our IFSFFS, we firstly generalized the original F-score to the improved F-score measuring the discrimination of more than two sets of real numbers. Then we proposed to combine Sequential Forward Floating Search (SFFS) and our improved F-score to accomplish the optimal feature subset selection. Where, our improved F-score is an evaluation criterion for filters, while SFFS and SVM compose an evaluation system of wrappers. The best parameters of kernel function of SVM are found out by grid search technique with ten-fold cross validation. Experiments have been conducted on five random training-test partitions of the erythemato-squamous diseases dataset from UCI machine learning database. The experimental results show that our SVM-based model with IFSFFS achieved the optimal classification accuracy with no more than 14 features as well.

Keywords: F-score, support vector machines, feature selection, sequential forward floating search, erythemato-squamous diseases

1. Introduction

The differential diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. Erythemato-squamous diseases all share the clinical features of erythema and scaling with very little differences. There are six groups in erythemato-squamous diseases. They are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, chronic der-matitis and pityriasis rubra pilaris. These diseases are frequently seen in the outpatient dermatology departments (Güvenir et al., 1998; Govenir and Emeksiz, 2000). Usually a biopsy is necessary for the diagnosis of these diseases, but unfortunately these diseases share many

histopathological features as well. Another difficulty for the differential diagnosis is that one disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation with 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope (Güvenir et al., 1998).

SVM, developed by Cortes, Vapnik and Cristianini respectively (Cortes and Vapnik, 1995; Critianini and Shawe-Taylor, 2005), is a relatively new machine learning technique and has been studied increasingly due to its many attractive characters and its excellent generalization performance. SVM has been widely in areas such as image recognition, text classification, speaker identification, cancer diagnosis and bioinformatics. SVM can handle a nonlinear classification efficiently by mapping samples from low dimensional input space into high dimensional feature space with a nonlinear kernel function. In feature space, SVM tries to maximize the generalization performance by solving a quadratic programming problem, and then finds the optimal separating hyperplane, which is the maximal margin hyperplane.

Although SVM has many outstanding advantages, its classification performance and generalization ability are often influenced by the high dimension or the number of feature variables of samples. In the diagnosis of erythemato-squamous diseases, there are 34 feature variables, which may lead to the risk of "over-fitting". Feature selection plays an important role in dimension reduction for classification (Liu and Zheng, 2006). In this study, we developed a diagnosis model based on SVM with a hybrid feature selection to diagnose the erythemato-squamous diseases. Our hybrid feature selection method, named IFSFFS, combines filter and wrapper methods to find out the optimal feature subset from the original feature set, where our improved F-score is an evaluation criterion of filters, and SFFS and SVM an evaluation of wrappers. In addition, in order to construct a sound diagnosis model, we randomly partitioned datasets into different training-test portions, and then employed the grid search technique with ten-fold cross validation on the training dataset of five different training-test partitions to find the optimal parameters for the kernel function of SVM.

This paper is organized as follows. Section 2 summarizes the available researches on the diagnosis of erythemato-squamous diseases. Section 3 describes the general feature selection methods and our improved F-score and our SFFS. Section 4 reviews the principles of SVM. Section 5 discusses our research design. Section 6 demonstrates and analyzes our experiment results in detail. Finally, in section 7, our conclusion arrives.

2. Related Work on Diagnosis of Erythemato-Squamous Diseases

There have been several studies reported focusing on the diagnosis of erythemato-squamous diseases using dermatology dataset, and all available methods achieved comparatively high classification accuracy. Übeyli and Güler proposed an approach based on adaptive neuro-fuzzy inference systems for detection of erythemato-squamous diseases (Übeyli and Güler, 2005), and got the classification accuracy of 95.5%. Luukka and Leppälampi obtained 97.02% using fuzzy similarity classifier for diagnosis of erythemato-squamous diseases (Luukka and Leppälampi, 2006). The methods based on fuzzy weighted pre-processing, K-NN

based weighted pre-processing, and decision tree classifier for the diagnosis of erythemato-squamous diseases were proposed by Polat and Günes (Polat and Günes, 2006), and the classification accuracy they reached are 88.18%, 97.57%, and 99.00%, respectively. Nanni obtained 97.22%, 97.22%, 97.5%, 98.1%, 97.22%, 97.5%, 97.8%, and 98.3% using LSVM, RS, B1_5, B1_10, B1_15, B2_5, B2_10, and B2_15 algorithms (Nanni, 2006). Luukka presented similarity classifier using similarity measure derived from Yu's norm in classification of medical data sets (Luukka, 2007), and the classification accuracy for the diagnosis of erythemato-squamous diseases was 97.8%. Übeyli obtained 98.32% classification accuracy on the differential diagnosis of erythemato-squamous diseases (Übeyli, 2008), using multiclass support vector machines with the error correcting output codes. Polat and Günes obtained 96.71% classification correct rate on diagnosis of erythemato-squamous diseases using a novel hybrid intelligence method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problem (Polat and Günes, 2009). Übeyli obtained about 97.77% classification accuracy using combined neural networks model to guide model selection for the diagnosis of erythemato-squamous diseases (Übeyli, 2009). Liu and Sun et al. obtained 96.72%, 92.18%, 95.08%, and 92.20% using feature selection algorithm with dynamic mutual information, which was estimated using four typical classifiers named Naive Bayes, 1-Nearest neighbor, C4.5 and RIPPER (Liu et al., 2009). Karabatak and Ince proposed a new feature selection method based on association rules and neural network for diagnosis of erythemato-squamous diseases, and their correct classification rate was 98.61%, and the dimension of feature space was reduced from 34 to 24 (Karabatak and Ince, 2009).

In this paper, we developed a diagnosis model based on Support Vector Machines (SVM) with a novel hybrid feature selection method to diagnose erythemato-squamous diseases. In our project, we firstly generalized the original F-score from computing the discrimination of two sets of real numbers to computing the one of more than two sets of real numbers, and proposed the hybrid feature selection method IFSFS (Improved F-score and Sequential Forward Search, IFSFS) to diagnose the erythemato-squamous diseases (Xie et al., 2009). SFFS as a floating search method introduced by Pudil et al. is characterized by changing number of features included at different stages of procedure (Pudil et al., 1994). SFFS can overcome the problem of SFS and provide the close to optimal solution at an affordable computational cost. Therefore, in this work, we combined our improved F-score and SFFS to propose a novel hybrid feature selection method named IFSFS.

3. Feature Selection

Feature selection plays an important role in building classification systems (Liu et al., 2002). It can not only reduce the dimension of data, but also lower the computation costs and gain a good classification performance. The general feature selection algorithms comprise two categories: the filter methods and wrapper methods (Blum and Langley, 1997; Talavera, 2005). The filter methods identify a feature subset from the original feature set with a given evaluation criterion which is independent of learning algorithms. The wrapper methods choose those features with high prediction performance estimated by a specified learning algorithm (Kohavi and John, 1997; Guyon and Elisseeff, 2003).

This paper presents a novel hybrid feature selection method IFSFFS, where the improved F-score plays the role of filters, and SFFS with SVM acts as wrappers. Using IFSFFS we eliminated all redundant features and remained all necessary ones, so that we can achieve good performance in diagnosing the erythemato-squamous diseases. In the following subsection, our improved F-score and our IFSFFS are introduced.

3.1. Our Improved F-score

The original F-score is used to measure the discrimination of two sets of real numbers (Akay, 2009). Now we generalize it to the improved F-score to measure the discrimination between more than two sets of real numbers. Here are the definitions of our improved F-score. Given training vectors $x_k,\ k = 1, 2, \cdots, n$, and the number of subsets of dataset l ($l \geq 2$), if the size of the jth subset is $n_j,\ j = 1, 2, \cdots, l$, then the F-score of the ith feature is defined as

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

where $\bar{x}_i$ and $\bar{x}_i^{(j)}$ are the average of the ith feature of the whole dataset and the jth subset respectively, and $x_{k,i}^{(j)}$ is the ith feature of the kth instance in the jth subset. The numerator of Equation (1) indicates the discrimination between each subset, and denominator the one within each subset. The larger the F-score is, the more likely this feature is discriminative.

3.2. Our Sequential Forward Floating Search (SFFS)

Our SFFS converted the original SFFS procedure as follows:
Step 1: Compute the F-score for each feature via our improved F-score as described in Equation (1) on training subset, and sort features in descending order according to their F-score values. Initialize the destination subset empty and the source subset with all features;
Step 2: Select the top one feature from the source subset and add it to the destination subset;
Step 3: Execute Section 4.2 to get the optimal parameter for the kernel function of SVM, and build the optimal predictor model to classify the training subset of samples according to the current destination subset, and get the classification accuracy of training subset;
Step 4: Select the next feature and add it to the destination subset;
Step 5: Execute Section 4.2 to get the optimal parameter of SVM, and use the corresponding SVM classifier with optimal parameters to classify the exemplars in training subset according to the current destination subset, and get the training classification accuracy;
Step 6: If the training classification accuracy is not improved then eliminate the feature that has just been added to the destination subset;
Step 7: Go to Step 4 until all features in source subset have been processed.

4. Our Research Design

This subsection introduced our research idea in detail, including how we randomly partitioned the dataset into five different training-test portions, and how we chose the optimal parameters for kernel functions of SVM, and how we built the diagnosis model for erythemato-squamous diseases.

4.1. Partition the Dataset Randomly

In order to get a reasonable result and make the results statistically meaningful, we randomly scattered the datasets, so that we can randomly partition it into training and testing portions. To fulfill this aim, we created a two-dimension array with 5000 rows and 2 columns randomly, and each element in the array is between 1 and 358, where 358 is the size of the dataset we used. We exchanged the two samples in the dataset according to the values of each row of the two-dimension array. For example, if the values of the ith row are p and q, then we exchange the pth sample with the qth sample of the dataset. After that, we partitioned the dataset into five sorts of training-test portions.

4.2. Choosing Model Parameters

We chose RBF kernel function for SVM ($K(x, x^{'}) = \exp(-\|x - x^{'}\|^2/\gamma^2)$), and the parameters for RBF kernel function are the penalty parameter C and the kernel function parameter γ. To get good generalization, we applied a grid search and 10-fold cross validation process to select the optimal parameters. We consider a grid search space of (C, γ), where $\log_2 C \in \{-5, \cdots, 15\}$, and $\log_2 \gamma \in \{-15, \cdots, 5\}$.

4.3. SVM-based Diagnosis Model with IFSFFS

The diagnosis model we proposed for erythemato-squamous diseases combined IFSFFS and SVM constructing a hybrid feature selection process provided with the advantages of filters and wrappers. In filter process, we initially calculate the improved F-score value for each feature, and sort features in descending order according to their F-score values. In wrapper procedure, a subset of the original training set is generated by including the features with top N F-scores, where $N = 1, 2, \cdots, m$, and m is the total number of features. Once the subset is generated, a grid search with 10-fold cross validation is carried out to find the optimal value of (C, γ). We choose the (C, γ) leading to the highest classification accuracy on training dataset to construct the SVM predictor model. This procedure via our SFFS is carried out until all features are tested. Finally, we constructed the diagnosis model which is the best classifier with the optimal parameters of SVM and with the least necessary features.

5. Experiment Results and Analysis

In this subsection, we described the erythemato-squamous diseases dataset which we used in our study, and demonstrated all experiment results on the dermatology dataset. Finally the results were analyzed in detail.

5.1. The Erythemato-Squamous Diseases Dataset

Table 3 in Appendix 1 shows the dataset of erythemato-squamous diseases we used in our work, which consists of 358 exemplars with 34 features of each example and 6 classes in total. One will notice that we deleted 8 samples which have missing values from the original datasets of erythemato-squamous diseases. In the dataset, the family history feature has the value 1 if any of these diseases has been observed in the family and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

5.2. Experiment Results and Analysis

Training-test partitions	Selected features	Training accuracy (%)	Testing accuracy (%)
50 - 50%	27,31,2,15,22,9,14,5,26	100.0000	97.2222
60 - 40%	33,29,20,31,6,7,15,22,21,14,26,2	98.5849	93.8356
70 - 30%	33,29,31,20,15,7,22,8,14,5,26	99.1935	98.1818
80 - 20%	33,29,31,6,20,7,15,22,16,14,5,26	98.9437	98.6486
90 - 10%	33,31,6,20,22,15,7,25,28,10,16,14,5,26	98.7500	100.0000

Table 1: Summary of Classification Accuracies on Five Different Training-Test Partitions

We calculated the importance of each feature via our improved F-score on five different training subsets corresponding to five different training-test partitions respectively. Then we use our SFFS procedure to construct the diagnoses model for erythemato-squamous diseases. We repeated the experiments five times with different train/test splits We summarized the optimal performance of this experiment on five random training-test partitions in Table 1. The figures in Table 1 show that the best performance is that the testing accuracy can achieve 100% with 14 features selected on 90-10% training-test portion. The average training accuracy is about 99.09%, and testing accuracy 97.58%. The average number of being selected features is nearly 12. Table 1 also tells us that the selected features according to five different partitions are not always the same. The common ones here are the features of 31, 26, 22, 15, and 14. We can conclude that these five features are the necessary ones which should be considered in diagnosing erythemato-squamous diseases. These results imply that our method is to be improved further, so that we can get more objective and stable results. This also is our doing work now. Furthermore, it must be pointed out that each run of our algorithm cannot always get the one and same result. But the best testing accuracy we can achieve is nearly always up to 100%, and the selected feature subset has a slight variation, but the size of it is around ten. The most common features we achieved are 31, 26, 22, and 15 feature subset, and the average testing accuracy is around 98%. The reason for this variation is that we preprocess the dataset before partitioning it into five different training-test portions to get the random partitions. Anyway, we can say that the

Author	Method	Classification accuracy (%)
Übeyli and Güler (2005)	ANFIS	95.50
Luukka and Leppälampi (2006)	Fuzzy similarity-based classification	97.02
Polat and Günes (2006)	Fuzzy weighted pre-processing	88.18
	K-NN based weighted pre-processing	97.57
	Decision tree	99.00
Nanni (2006)	LSVM	97.22
	RS	97.22
	B1_5	97.50
	B1_10	98.10
	B1_15	97.22
	B2_5	97.50
	B2_10	97.80
	B2_15	98.30
Luukka (2007)	Similarity measure	97.80
Übeyli (2008)	Multiclass SVM with the ECOC	98.32
Polat and Günes (2009)	C4.5 and one-against-all	96.71
Übeylii (2009)	CNN	97.77
Liu et al. (2009)	Naive Bayes	96.72
	1-NN	92.18
	C4.5	95.08
	PIPPER	92.20
Karabatak and Ince (2009)	AR and NN	98.61
This study	IFSFFS and SVM	100 (best) 97.58 (avg.)

Table 2: Classification Accuracies of Our Method and Other Classifiers from Literature

most common features are the necessary ones which must be considered in constructing the predictor of diagnosing erythemato-squamous diseases.

Table 2 summarizes the classification accuracies of all available methods. About our method we gave the best testing accuracy and the average one of this run. From Table 2 we can see that our method using SVM with IFSFFS can achieve the optimal classification accuracy among that of all available methods to diagnose erythemato-squamous diseases. Although the average one is not the best, the size of the selected featue subset is the minimum compared to the available ones.

6. Conclusions

In this study we generalized the original F-score to the improved F-score to measure the discrimination between more than two sets of real numbers. Then we proposed a SVM-

based diagnosis model with a novel hybrid feature selection method, named IFSFFS, for the diagnosis of erythemato-squamous diseases. This novel hybrid feature selection method combines the advantage of filters and wrappers, where our improved F-score is used as an evaluation criteria of filters, and SFFS with SVM as wrappers for feature selection procedure to find out the optimal feature subset. Experiment results show that our new hybrid feature selection method obtained the promising classification accuracies for diagnosis of erythemato-squamous diseases.

However, there are many aspects in this work to be improved. The accuracy was computed by the ratio of correctly classified exemplars to the exemplars under considering, which did not consider the classification accuracy of each class. We have started to redefine the classification accuracy taking classification accuracy of each class into account in our experiment to overcome the disadvantage, so that we can get more objective and stable results. In addition, 10-fold cross validation will not only limited to the procedure of selecting the optimal parameters for classifier model as that in this work, it will also be considered to be introduced into our research procedure to substitute the random five training-test partitions, so that the result will be more statistical. Furthermore, we will consider the relevance between different features during feature selection procedures.

Acknowledgments

The authors wish to thank the helpful comments and suggestions from anonymous reviewers. The authors are grateful to the authors who provide the useful benchmark data sets. This work was supported in part by the grant of the Fundamental Research Funds for the Central Universities of GK200901006, and was also supported by Natural Science Basis Research Plan in Shaanxi Province of P. R. China (Program No. 2010JM3004).

References

M. F. Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36:3240–3247, 2009.

A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 1-2:245–271, 1997.

C. Cortes and V. N. Vapnik. Support vector networks. *JMLR*, 3:273–297, 1995.

N. Critianini and J. Shawe-Taylor. *An introduction to Support Vector Machines and other kernel-based learning methods*. China Machine Press, Beijing, 2005.

H. A. Govenir and N. Emeksiz. An expert system for the differential diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 18:43–49, 2000.

H. A. Güvenir, G. Demiröz, and N. İlter. Learning differential diagnosis of erythemato-squamous diseases using voting feature intervals. *Artificial Intelligence in Medicine*, 13: 147–165, 1998.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *JMLR*, 3: 1157–1182, 2003.

M. Karabatak and M. C. Ince. A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 36: 12500–12505, 2009.

R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, 1-2:273–324, 1997.

H. Q. Liu, J. Y. Li, and L. Wong. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics*, 13: 51–60, 2002.

H. W. Liu, J.G. Sun, L. Liu, and H. J. Zhang. Feature selection with dynamic mutual information. *Pattern Recognition*, 42:1330–1339, 2009.

Y. Liu and Y. F. Zheng. Fs_sfs: A novel feature selection method for support vector machines. *Pattern recognition*, 39:1333–1345, 2006.

P. Luukka. Similarity classifier using similarity measure derived from yu's norms in classification of medical data sets. *Computers in Biology and Medicine*, 37:1133–1140, 2007.

P. Luukka and T. Leppälampi. Similarity classifier with generalized mean applied to medical data. *Computers in Biology and Medicine*, 36:1026–1040, 2006.

L. Nanni. An ensemble of classifiers for the diagnosis of erythemato-squamous diseases. *Neurocomputing*, 69:842–845, 2006.

K. Polat and S. Günes. The effect to diagnostic accuracy of decision tree classifier of fuzzy and k-nn based weighted pre-processing methods to diagnosis of erythemato-squamous diseases. *Digital Signal Processing*, 16:922–930, 2006.

K. Polat and S. Günes. A novel hybrid intelligent method based on c4.5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36:1587–1592, 2009.

P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 11:1119–1125, 1994.

L. Talavera. An evaluation of filter and wrapper methods for feature selection in categorical clustering. In *In: proceeding of 6th International Symposium on Intelligent Data Analysis*, pages 440–451, 2005. Madrid, Spain.

E. D. Übeyli. Multiclass support vector machines for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 35:1733–1740, 2008.

E. D. Übeyli. Combined neural networks for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 36:5107–5112, 2009.

E. D. Übeyli and İ. Güler. Automatic detection of erythemato-squamous diseases using adaptive neuro-fuzzy inference systems. *Computers in Biology and Medicine*, 35:421–433, 2005.

J. Y. Xie, W. X. Xie, C. X. Wang, and X. B. Gao. Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Systems with Applications*, 2009. Under review.

Appendix 1

Erythemato-Squamous Diseases	Features	
(number of patients)	Clinical	Histopathological
Psoriasis (111)	Feature 1: Erythema	Feature 12: Melanin incontinence
Seboreic dermatitis (60)	Feature 2: Scaling	Feature 13: Eosrinophils in the infiltrate
Lichen planus (71)	Feature 3: Definite borders	Feature 14: PNL infiltrate
Pityriasis rosea (48)	Feature 4: Itching	Feature 15: Fibrosis of the papillary dermis
Chronic dermatitis (48)	Feature 5: Koebner phenomenon	Feature 16: Exocytosis
Pityriasis rubra pilaris (20)	Feature 6: Polygonal papules	Feature 17: Acanthosis
	Feature 7: Follicular papules	Feature 18: Hyperkeratosis
	Feature 8: Oral mucosal involvement	Feature 19: Parakeratosis
	Feature 9: Kneeand elbow involvement	Feature 20: Clubbing of the rete ridges
	Feature 10: Scalp involvement	Feature 21: Elongation of the rete ridges
	Feature 11: Family history	Feature 22: Thinning of the suprapapillary epidermis
	Feature 34: Age	Feature 23: Pongiform pustule
		Feature 24: Munro microabcess
		Feature 25: Focal hypergranulosis
		Feature 26: Disappearance of the granular layer
		Feature 27: Vacuolization and damage of basal layer
		Feature 28: Spongiosis
		Feature 29: Saw-tooth appearance of retes
		Feature 30: Follicular horn plug
		Feature 31: Perifollicular parakeratosis
		Feature 32: Inflammatory mononuclear infiltrate
		Feature 33: Band-like infiltrate

Table 3: Eryrhenato-Squamous Diseases Dataset from UCI