Feature Selection after Feature Grouping

Jie Yao JIEYAO@CS.CMU.EDU
Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA 15213 USA

Abstract

Feature selection is an effective method to reduce the computational expense. However, the results of traditional feature selection algorithms have weak generalization ability when the training data is small. This kind of "brittleness" is due to the greedy search scheme and the overlapping features. This paper presents a feature grouping technique to find a good starting feature set for the feature selection algorithm. By this way, a controlled redundancy is introduced to the final selected feature set to improve the robustness of feature selection. The experiment results suggest that the new algorithm remarkablely improve the prediction accuracy without a significant decrease in the computational efficiency.

1. Introduction

Feature selection is an important aspect of many areas of supervised machine learning. Given a large candidate feature set, a small feature subset must be chosen to increase the prediction accuracy and the computational efficiency. The optimal feature set is the one, which give the most accurate predictions for future data and at the same time reduce the computational cost to the largest extent.

It is very clear that problem with more features is more complex than the problem with fewer features. Therefore, feature selection can also be viewed as an automated method to simplify complex problems. Feature selection technique can be used to simplify the problems, which tend to be out of control of human experts. The goal of this simplification may be a better and more efficient solution and even a better understanding for the problem. However, in practice feature selection algorithms are often challenged by some very hard problems.

In this paper, we use the microscope image classification problem as our example problem. The microscope image dataset is composed of images automatically extracted from on-line biological journal articles. We are going to classify the images into six image categories. This classification problem is difficult because there is no simple visual clue. In other words, there is little knowledge about which feature is the best. In practice, there are many features, which all are assumed relevant to some extent, borrowed from other research domain like image retrieval. Thus, a feature selection process becomes necessary to reduce the computational cost and even increase the prediction accuracy.

In our preliminary feature selection experiments, we use *Forward Sequential Selection* (FSS) algorithm to select features from a candidate feature set. The evaluation function is the leave-one-out cross validation accuracy of k Nearest Neighbor (k-NN) classifier. The output of the FSS algorithm is pretty small (typically 6 or 7 out of 139 features). However, the generalization ability of the selected feature set is poor, especially when the training dataset is small.

To increase the generalization ability of feature selection algorithm with small training data, we introduce a feature grouping process to choose starting feature set. First, features are grouped according to their feature value distributions on dataset. Then, for each feature groups, typical features are pre-selected to form a starting feature set, which assumed to cover a wide range of feature groups. Finally, conventional feature selection program is run on the basis of the pre-defined starting feature set.

Experiments results show that the FSS algorithm with this feature grouping technique can avoid the "over-fitting" effectively. The prediction accuracy performance increases without great change of the size of the selected feature set.

Section 2 reviews feature selection techniques and then analyzes the reason of the "brittleness". Section 3 presents feature grouping technique in detail, including the definition of the features' behavior for distance-based classifiers. Section 4 shows the experiment results of applying the FSS algorithm and feature grouping mechanism on the microscope image classification problem. Section 5 summarizes of the whole paper.

2. Feature Selection

2.1 Background

Detailed survey of feature selection can be found in [Langley 1994], [Blum & Langley 1997] and [Dash & Liu 1997]. The other useful publications on feature selection algorithms are [Aha & Banker1995], [Caruana & Freitag 1994], [John, Kohavi & Pfleger1994], [Kohavi & Sommerfield 1995]. Here we make a rough overview and focus on the forward sequential selection (FSS) algorithm which is studied intensively in this paper.

Generally speaking, most feature selection algorithms are composed of the following two components:

Search Algorithm

Feature selection can be viewed as a search problem in the feature space guided by the evaluation function. There are three categories of search algorithms: exponential, sequential and randomized. *Exponential* algorithms do a complete search for the optimal subset. *Sequential* or *heuristic* algorithms, which generate the subset incrementally (either increasing or decreasing), often have polynomial complexity. *Randomized* algorithms are non-determinative algorithms, including genetic algorithms (GA) and simulated annealing algorithms.

Exponential search algorithms are rarely used in piratical problems due to the high computational cost. Sequential search algorithms are most popular algorithms because they are simple and very easy to implement. Randomized search algorithms can avoid the local-minima problem by introducing randomness. However, they are complex and slow.

Feature Evaluation Function

Feature evaluation function assigns a score value to the feature sets to guide the search in the feature space. There are five popular evaluation functions.

Distance measure, also known as *separability* or *discrimination* measure, examines the difference between the conditional probabilities.

Information measure uses the information gain of selecting a feature.

Dependence measure or *correlation* measure uses the conditional probability of one variable given the value of other variables.

Classifier accuracy rate measure, is also called "*wrapper methods*" [Kohavi & Sommerfield 1995] uses the prediction accuracy on a separated testing dataset.

Most feature selection algorithms can be viewed as a combination of the different search algorithms and evaluation functions.

2.2 Forward Sequential Selection Algorithm

In this paper, we use forward sequential selection (FSS) algorithm as the basic feature selection algorithm, and the k-NN classifier leave-one-out cross validation accuracy of as the evaluation function.

FSS is a kind of greedy algorithm. It starts with an empty feature set and select features incrementally. In each cycle, all the possible feature sets are evaluated by prediction accuracies, and then the feature with the best performance is selected. This cycle repeats until some stopping conditions are satisfied. The trade-off of accuracy and the computational efficiency can be realized by choosing different stopping points. The most popular "no improvement" stopping condition has a strong bias on finding the smallest feature set.

FSS algorithm is very popular because it is simple, fast and of high accuracy. Though there are many advantages for FSS algorithm, we found the generalization performances of the FSS outputs vary greatly from good to very bad. The following section will discuss this "brittleness" problem.

2.3 Brittleness and Over-fitting

There are two aims of feature selection: finding the feature set with the best prediction accuracy and reducing the size of the feature set. Usually these two aims cannot be achieved at the same time. How to trade-off the accuracy and the feature size is always the key problem of feature selection.

To apply the conventional FSS algorithm to our image classification example, we found that the quality of the results depend heavily on the training dataset. To follow the terminology from [Vafaie & DeJong 1993], "brittleness" is used to refer this kind of weak generalization ability. Because the ability of generalization becomes worse when the training dataset is small, the brittleness can also viewed as a kind of "over-fitting".

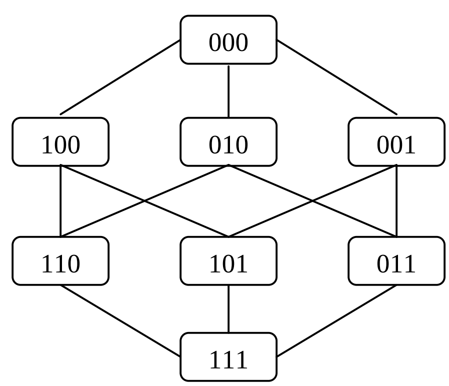

Figure 1. Sequential feature selection lattice.

One reason of the "brittleness" is the greedy search scheme. Figure 1 shows a sequential feature selection lattice for three features. A FSS feature selection process corresponds to a path from the node (000) to the node (111). It can be seen that in this sequential search scheme, not all points in the feature space is reachable. For example, if the feature selection program select the second feature first, that feature can not be discarded again. On the graph, the node (010) can not reach the node (101). This search scheme also accounts fr the possible local minima.

This reason is related closely to the FSS algorithm itself. We can compensate this part of "brittleness" by changing FSS algorithm. For example, allowing "backtrack" will enhance the feature space search ability and increase robustness. However, that kind of change will bring a remarkable decrease in the computational efficiency. We want to seek another possible route to increase robustness.

To describe the features' functions in the classification process, we introduce a concept of "feature behavior". For the distance-based classifiers, like k-NN, the feature behavior is determined by the feature value distribution over the data points (as shown in Figure 2.)

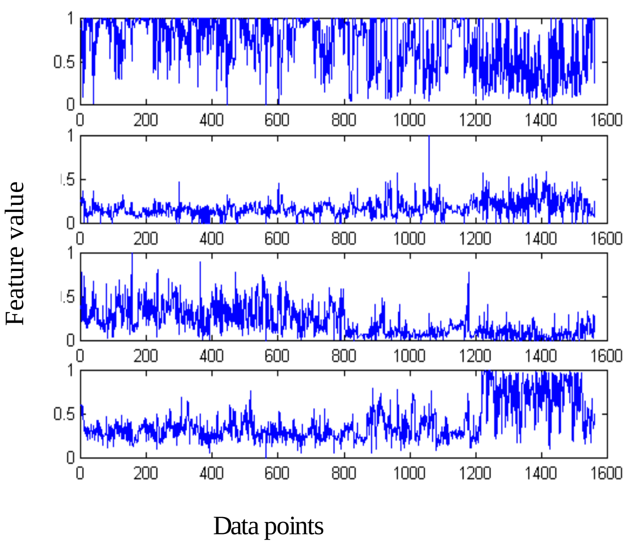

Figure 2. Feature value distribution over whole data points.

Features can be grouped by their feature value distributions. Each feature group has its specific focus and emphasis. In the case of small training data, the selected feature set may just include members of some feature groups but not include any features in the other feature groups. That is the other reason of over-fitting and weak generalization ability.

There are two methods to improve the robustness of FSS algorithm, one is to loose the stopping criterion to include more features after the feature selection process; the other is to use a non-empty starting feature set to include more features before the feature selection process.

In this paper, we adopt the second method by introducing a "feature grouping" process before the feature selection and form the starting feature set by selecting the typical features in each major feature groups.

3. Feature Grouping

This section describes the feature grouping technique. At first, a definition of feature behavior measure based on the feature value distribution is introduced. Then a k-Means clustering algorithm is used to group the features. Other related issues are also discussed.

For the distance-based classifiers, like k-NN, the feature behaviors are determined by their feature value distributions over dataset.

Let $A=(a_{i,j}), i=1,2,\cdots,m, j=1,2,\cdots,n$ represents a "feature-value" data table, where *m* is the number of data points and *n* is the size of features. $a_{i,j}$ is the *i* th data's value on the *j* th feature .

f(i), i=1,2,…n represents the features and $x(i)=[a_{i1},a_{i2},\cdots,a_{in}]$, $i=1,2,\cdots,m$ represents the *i* th data points.

The distance between data points $x(i)$ and $x(j)$ is the Euclidean distance

$$d(x(i),x(j))=\left(\sum_{k=1}^{n}(a_{i,k}-a_{j,k})^2\right)^{1/2} \qquad (1)$$

Viewing features as the data points in feature value space, the distance between any two features can be defined as the Euclidean distance between the feature value vectors. The distance between features *f(i)* and *f(j)* is defined as:

$$d(i,j)=\sum_{k=1}^{m}[a(k,i)-a(k,j)]^2 \qquad (2)$$

With the distance definition, we can group all the candidate features using k-means clustering algorithm. For each feature groups, one typical feature is chosen to form the pre-selected feature set, which is the starting feature set for FSS algorithm.

4. Results

4.1 Experiment Settings

4.1.1 Dataset

We run all experiments on the image set extracted from on-line *Journal of Cell Biology, vol.136, 1997*. There are six categories of images: Fluorescence Microscope Images, Transmitted Microscope Images, Electronic Microscope Images, *Gel* images, Charts/Graphs and Other images. The distribution is shown by figure 3.

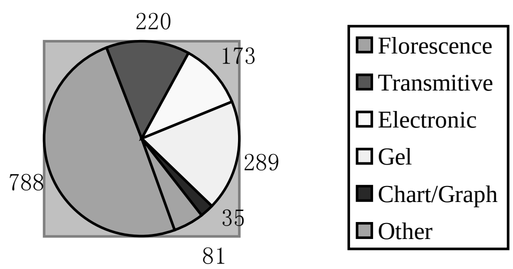

Figure 3. Distribution of Six Category of Images Extracted from On-line Journal (total 1586 images).

4.1.2 CANDIDATE FEATURE

Our candidate feature set comprised of 64 histogram feature set and 75 SLF(Subcelluar Localization Feature) set which are borrowed from two related research domain, *Content-Based Image Retrieval (CBIR)* [Pentland, Picard & sclaroff 1994] and the research of Protein Localization Pattern Analysis from Fluorescence Microscope Images [Markey, Boland & Murphy 1999], [Murphy, Boland & Vellliste 2000].

Histogram features count the numbers of pixels whose values are within the evenly cut intervals. They are called low-level features and its feature calculation process is very fast. Histogram features have been proven good at capturing the overall visual appearance of the images by CBIR research [Pentland, Picard & sclaroff 1994].

SLF feature set consists of texture features, moment features and morphological features. Most of them have high computational cost and are called high-level features. SLF set have been successfully used to describe the subcellular localization pattern of the florescence microscope images [Murphy, Boland & Vellliste 2000].

4.1.3 K-NN CLASSIFIER

We use k-NN classifier as our classification algorithm and the leave-one-out cross validation accuracy of k-NN classifier as the evaluation function. The parameter k is chosen based on the size of training data set manually.

The k-NN classifier has no separate training procedure which make the cross validation very fast [Aha 1998]. Another reason is that kNN is a distance-based classifier, which make it a natural example of those also distance-based image retrieval algorithms. Therefore, the results in this paper on the k-NN algorithm will have more impact on other image classification and retrieval problems.

4.2 Performance of Standard FSS Algorithm

The prediction accuracy of FSS selected feature set with different training datasets is shown in figure 4. For all size training dataset, the accuracy increases steadily with the number of features used. When the training data is small, the performance is bad.

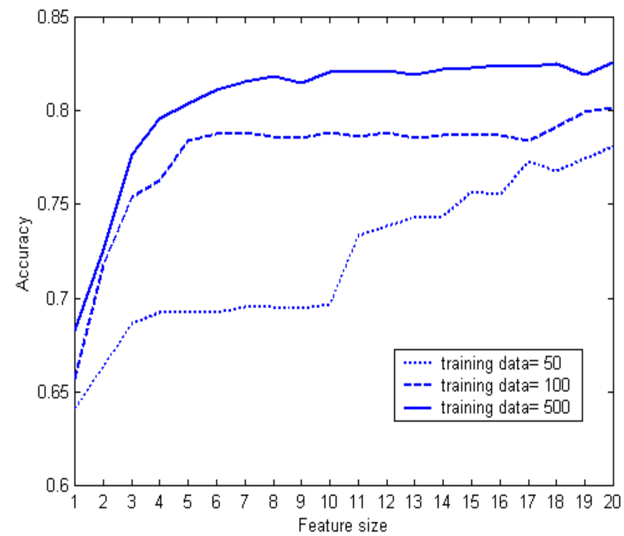

Figure 4. Performance of FSS algorithm with different training data set.

4.3 Feature Grouping Result

We use a simple k-means clustering algorithms [Jain, Murty & Flynn 1999] to group the features. The number of the clusters is decided manually. The k-means output of k=3 is shown in Table 1. The sample feature value distributions of each group are shown in figure 5, 6 and 7.

Table 1. K-means feature grouping output (k=3).

Group	1	2	3
Feature ID	9 11 60 62 63 64 65 66 68 70 72 74	6 10 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 45 46 47 48 49 50 51 52 53 54 55 56 59 61 67 71 75 79 80	1 2 3 4 5 7 8 44 57 58 69 73 76 77 78 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139

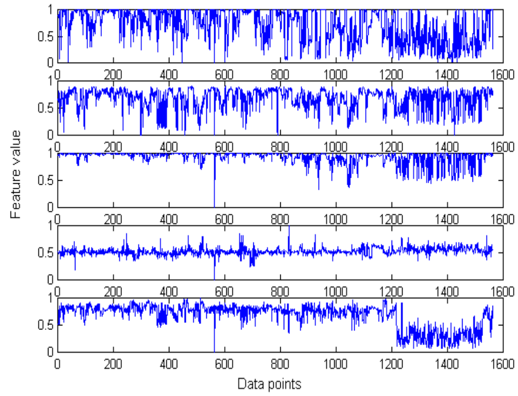

Figure 5. Sample feature value distributions of group1 (they have high feature value for all data).

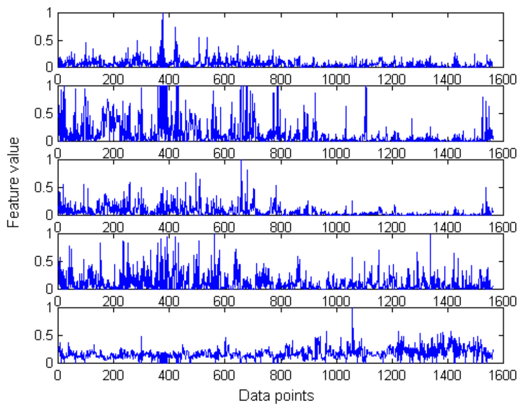

Figure 6. Sample feature value distributions of group2 (they have high feature value for some data and low value for the other data)

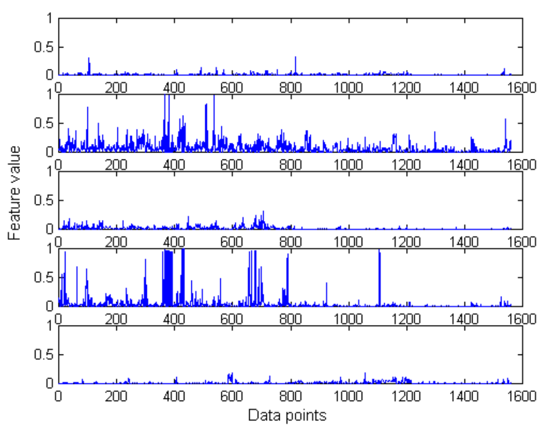

Figure 7. Sample feature value distributions of group 3 (they have low feature value for all data)

4.4 Performance of FSS after feature grouping

The prediction accuracy of revised FSS algorithm with training dataset size = 100 is shown in Figure 8. The increase of robustness is obvious.

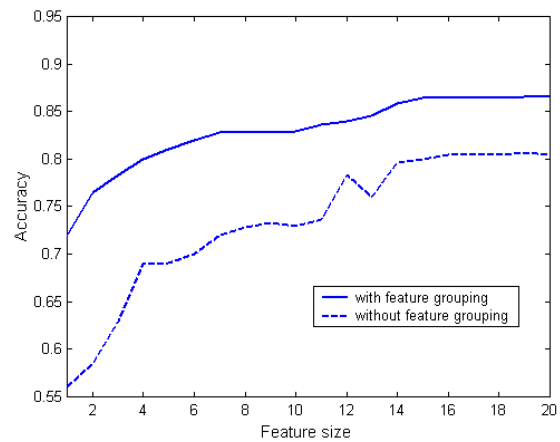

Figure 8. Performance of grouping FSS algorithm with datasize = 100, number of feature groups = 3.

5. Related Work

The brittleness of the result of feature selection has been noticed by [Vafaie & DeJong 1993]. It was also shown that the brittleness was related to the tendency to be trapped at local minima caused by the interdependency among the features. Instead of modifying FSS algorithm, they employed a GA algorithm to improve the robustness. They used AQ15 learning algorithm as their evaluation classifier. In their experiments on the Brodatz album of textures data, the GA based feature selection algorithm have little improvement on the actually recognition rate. In a smaller dataset of breast cancer data, the GA algorithm made an improvement from 40% to 50%. They did not give the performance on the size of the feature set or computational cost. Although their experiment result was not good so far, they stated clearly the brittleness problem of feature selection.

The feature grouping idea partly came from the research of [Valiaya, Figueiredo, Jain & Zhang 2001]. They utilized sequential floating forward selection (SFSS) and feature clustering (FC) techniques in their research of scenery (indoor /outdoor, city /landscape etc.) image classification. Their SFSS selected feature set (52 out of 600) had a little decrease in the accuracy. In addition, they reported a very long feature selection computing time. More importantly, they introduced a k-Means Feature Clustering (linear combination) technique as a simple alterative for SFSS algorithm. Instead of choose a different starting feature set as done in this paper, they used feature clusters to form a new feature by linear combination. Though the aim of feature clustering was to reduce redundancy, the idea of feature clustering gives birth to this paper. Moreover, comprised with the scenery images, our microscope images are more challenging. The images extracted from on-line journal are more diverse and noisy.

[Swets & Weng 1995] also proposed a feature selection method for the content-based image retrieval system. They used KL projection to generate most expressive features and used discriminant analysis projection to generate most discriminating features. Their experiment result on the face image retrieval problem was very good. There are two shortcomings with their projection feature selection method. The first, the computational cost is huge since the feature values must be calculated before generating new features. Also the generated features (non-linear combination of given features) were not explicit meaningful.

[Jaimes & Chang 2000] proposed a dynamic approach to select feature and classifier for the segmentation of baseball video images. They also used cross validation as their evaluation function. The classification of the image region was supposed to be easier than the classification of whole image because of the homogeneity of region.

The closest research work done on the similar image dataset was carried by R. F. Murphy, M. V. Boland and M. Velliste [Murphy, Boland & Velliste2000]. They selected 37 features from a pool of 84 features including texture features, moment features, and biological morphological features by "*stepwise discriminant analysis*" method to describe protein localization patterns. The difference of these two study was that our images extracted on-line are much more diverse.

6. Summary

The goal of the research presented here was to introduce a new feature selection algorithm to increase the generalization ability of the traditional forward sequential feature selection algorithm. We presented a feature grouping stage to choose the starting feature set on the basis of the analysis of the weakness of FSS algorithm. Experiments on the microscope image classification problem showed that this method was superior to the traditional FSS algorithm in prediction accuracy for new data. The feature grouping was based on the feature value distributions. Result shown that the feature grouping technique enhance the robustness of FSS algorithm. Non-distance based classifiers may use other definition of feature behaviors. Besides the FSS algorithm, the feature grouping technique presented here can also be used to other greedy feature selection algorithms.

There are still many limitations for the feature grouping technique:

Our discussion in section 3 only considers feature overlapping. Actually, the features are both overlapping and complimentary. That makes the feature grouping itself a difficult problem.

How to choose the pre-selected features is still a problem even the feature groups is known. In the future research, we want to consider different methods, such as "group center" and the most powerful features.

Acknowledgements

Thanks to Robert Murphy and Sebastian Thrun for the help in developing some of these ideas and for assistance in the running the experiments, to Meel Velliste for the help in feature calculation, and the anonymous reviewers for so many useful comments and suggestions.

References

Aha, D. W., & Banker, R. L. (1995). A comparative evaluation of sequential feature selection algorithms. *Proceeding of the Fifth International Workshop on Artificial Intelligence and Statistics*, 1-7.

Aha, D. W. (1998). Feature weighting for lazy learning algorithms. In Liu, H. & Motoda, H. (eds.) Feature Extraction, Construction and Selection: A Data Mining Perspective. Norwell, MA: Kluwer.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97, 245-271.

Boland, M. V. (1999). *Quantitative description and automated classification of cellular protein localization patterns in fluorescence microscope images of mammalian cell.* Ph.D. dissertation, Department of Biomedical Engineering, Carnegie Mellon University.

Caruana, R., & Freitag, D. (1994). Greedy attribute selection, *Proceedings of the Eleventh International Conference on Machine Learning,* 28-36, Morgan Kaufmann.

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis,* vol.1, no. 3, 131-156.

Jaimes, A., & Chang, S. F. (2000). Automatic selection of visual features and classifiers. *IS&T/SPIE Storage and*

Retrieval for Image and Video Databases VIII. 3972, San Jose.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data Clustering: A Review. *ACM Computing Surveys*, 31, 264-323.

John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning*, 121-129, Morgan Kaufmann.

Kohavi, R. & Sommerfield, D. (1995). Feature subset selection using the wrapper method: Overfitting and dynamic search space topology. *Proceedings of First International Conference on Knowledge Discovery and Data Mining*, 192-197.

Langley, P. (1994). Selection of relevant features in machine learning. *Proceedings of the AAAI Fall Symposium on Relevance*, 1-5.

Markey, M. K., Boland, M. V., & Murphy, R. F. (1999). Towards objective selection of representative microscope images. *Biophysical Journal*, 76, 2230-2237.

Murphy, R. F., Boland, M. V., & Velliste, M. (2000). Towards a systematic for protein subcellular location: Quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 251-259.

Pentland, A., Picard, R.W. & Sclaroff, S. (1994). Photobook: Content-based manipulation of image databases. *Proceeding of Storage Retrieval Image Video Databases II*, 34-47.

Swets, D. L., & Weng, J. J. (1995). Efficient content-based image retrieval using automatic feature selection. *International Conference on Computer Vision.*

Vafaie, H., & DeJong, K. (1993). Robust feature selection algorithms. *Proceedings of the Fifth Conference on Tools for Artificial Intelligence*, 356-363. Boston, IEEE Computer Society.

Vailaya, A., Figueiredo, A., Jain, A. K., & Zhang, H. J. (2001). Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10,117-129.