

Feature Selection for On-line Image Classification

Jie Yao

CALD, SCS, CMU

Abstract

We propose research on the application of feature selection technique to the problem of on-line image classification. That is, for the specific problem of classifying the images extracted from on-line biological journals, we try to find out a subset of “power features” which best describe those images and have highest classification accuracy. The image classification problem proposed here is to classify the fluorescence microscope images from all kinds of on-line images (including graphs, *gel* images and other types of microscope images). The fluorescence microscope images will contribute to the analysis of protein subcellular localization patterns. The project focuses on most studied Sequential Selection Algorithms, which is accurate and efficient. Since the features we used to analysis come from different levels, including low-level features such as histogram and moments features and high-level features such as shape and object features, the project also study the hierarchy structure of features and the “feature clustering” technique. Our prior work demonstrated the potential of these techniques on increasing classification accuracy. We also try to get a better insight into the underlying concept of this real-world family of images.

1. Introduction

1.1 Problem

In the classification problems, more input features often do not induce better accuracy. This bad effect of irrelevant/redundant features to the classifier is called as “*curse of dimensionality*”[Duda&Hart1973][Jain,Duin&Mao2000][Ripley1996]. In the proposal, we deal with this problem in the context of on-line images classification, which is a challenge due to the complexity of the two-dimensional image data, and the classification result was mostly determined by the quality of feature set. By measuring the relevance or contribution of features to the aim of classifying images, we choose the most powerful and useful set of features and eliminate the irrelevant features. This “feature filtering” procedure will reduce the computational cost and avoid overfitting.

In this project, the images were extracted from on-line *Journal of Cell Biology*, vol.136, 1997. In this data set, there are fluorescence microscope images, electronic microscope images, transmitted microscope images, *gel* images, graph images and other images. In this project, we consider two classification problems. Two-class classification problem require a separation of fluorescence microscope images and all other. Six-class classification problem require classifying all images categories. We are going to select features from a pool of histogram features, texture features, moment features and morphological features, which perform very well in the problem of protein localization patterns analysis from fluorescence microscope images [Murphy,Boland&Velliste2000][Boland1999].

1.2 Goal

The goal of this research is finding out the most “powerful” image features, which achieve best classification performance. And also the selected features will help us getting a better insight into the underlying concept of this real-world image classification problem.

1.3 Impact

This research will have a great impact on image classification and retrieval applications, especially the application of images from an open source, like WWW. Moreover, the result may give us some knowledge of the structure of visual information.

2. Related Work

Feature selection has been studied by statistics and machine learning communities for several decades. Several publications have reported performance improvements for such measures when feature selection algorithms are used. Detailed survey of feature selection method can be found in [Langley1994][Dash&Liu1997]. Recently, more attention has been received by feature selection because of enthusiastic research in Data Mining. The specific survey of forward and backward sequential feature selection algorithms and their variants can be found in [Aha&Banker1995].

The image classification problem as a part of “Content-Based Image Retrieval (CBIR)” research can be found in [Valiaya, Figueiredo,Jain&Zhang2001]. They utilized sequential floating forward selection method and feature clustering method in the research of vacation image retrieval. A hierarchical classification was used to enhance the performance of content-based retrieval systems by filtering out irrelevant classes. These give us the intuition to study the hierarchical structure of features corresponding to the hierarchical of image classes.

The closest research to this project is carried by R. F. Murphy, M. V. Boland and M. Velliste [Murphy,Boland&Velliste2000]. They selected 37 good features from a pool of 84 features including Haralick texture features [Harlick1979], Zernike moment features, and biological morphological features by “*stepwise discriminant analysis*” [Klecka1980] method to describe protein localization patterns. The difference of these two study is that our images extracted on-line are much more diverse and the huge mount of data also require more sufficient algorithms. Our project can be viewed as a “preprocess” procedure.

3. Approach

3.1 Background of feature selection algorithms

Most feature selection algorithms are typically composed of the following two components:

1. **Search algorithm:** search the space of feature subsets, which has size 2^d where d is the number of features. There are three categories of search algorithm: exponential, sequential and randomized. Exponential algorithms (branch and bound, exhaustive) do a complete search for the optimal subset according to the evaluation function used. The optimality of the feature subset is guaranteed because the procedure can backtrack. Sequential or heuristic algorithms, which basically generate the subset incrementally (either increasing or decreasing), often have polynomial complexity. Randomized algorithms include genetic and simulated annealing search methods. These algorithms attain high accuracies but they require biases to yield small subsets. How this is best done remains an open issue.

2. **Evaluation function:** assign a score to a feature subset. There are five kinds of evaluation function: distance measure, information measure, dependence measure and classifier error rate measure. Distance measure, also known as separability and discrimination measure, examines the difference between the conditional probabilities. For a two-class problem, X feature is preferred if X induces a greater difference between the two-class conditional probabilities. Information measure determines the information gain from a feature. Dependence measure or correlation measure qualify the ability to predict the value of one variable from the value of another. The coefficient is a classical dependence measure and can be used to find the correlation between a feature and a class. Classifier error rate measure is also called “wrapper methods” [Kohavi& Sommerfield 1995]. As the features are selected using the classifier that later on uses these selected features in predicting the labels of unseen instances.

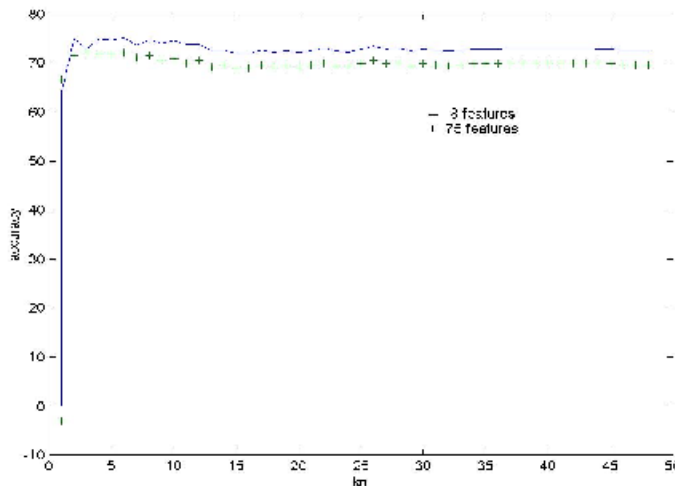
Most feature selection algorithms can be viewed as the combinations of different search algorithms and evaluation functions. The tradeoffs between high accuracy and small model size and computational cost always need consideration. In this project, we use sequential search algorithm and k-NN classifier “leave-one-out” error rate as the evaluation function.

3.2 Prior Work

In our prior research, we utilizing the "stepwise discriminant analysis"[Klecka1980]

method provided by SAS software to select the most discriminant features. It sequentially identifies those features that maximize a criterion which describes their ability to separate classes from one another while at the same time keeping the individual classes as tightly clustered as possible.

The k-NN (k nearest neighbor) classification result show that the selected 8 features perform better than all 75 features.



3.3 Sequential feature selection algorithms

Sequential features selection algorithms are the most used feature selection methods due to the simplicity and efficiency. The most common sequential feature selection algorithms are forward sequential selection (FSS) and backward sequential selection (BSS). FSS begin with zero features, evaluate all feature subsets with exactly one feature, and selects the one with the best performance. It then adds to this subset the feature that yields the best performance for subsets of the next larger size. This cycle repeats until no improvement is obtained from extending the current subset. BSS instead begins with all features and repeatedly removes a feature whose removal yields the maximal performance improvement.

In this project, we use k-NN algorithm as the evaluation classifier and “leave-one-out” cross validation error as the measure of accuracy.

It is well known that nearest neighbor algorithms perform not well under some situations,

however, such non-parameter classifiers are excellent choices for evaluation functions. Moreover k-NN algorithm expresses clearly the property of “data localization”: data that belong to the same class should be close to each other and those in different classes should be relatively farther away, according to some distance metric.

After the feature selection, we will validate the result on other classification algorithms, such as BP neural network and SVM. With the selected features, these algorithms are expected to do much better than k-NN classifier.

3.4 Hierarchical feature structure and “feature clustering”

Most feature selection algorithms, such as sequential selection algorithms, assume that features have little or no relation between each other. But in a contrary, features are closely related. Some features are extracted for the same aim and capture the same information. Some features can be viewed as one multidimensional feature. For example, histogram features (a vector of 64 or 256 single features) are more related each other than the morphological features.

It is more complex to consider the interaction of features. For the simplest situation, we study the “feature clustering” method for the similar features. One simple heuristic procedure was introduced by [Valiaya,Figueiredo,Jain&Zhang2001]. Every cluster is averaged to form a new feature. Thus the number of clusters determines the final number of features. Although this method does not guarantee an optimal solution, it does attempt to eliminate highly correlated features in high-dimensional feature spaces.

4. Work Plan

Month1. We will develop and implement the basic forward sequential selection algorithm and the backward sequential selection algorithm with k-NN evaluation function.

Month2. We will compare the performance of the FSS and BSS algorithms and make further revision to get better performance.

Month3. We will develop “feature clustering” algorithm.

Month4. We will investigate the result of the selection procedure and try to draw a conclusion about choosing features for this on-line image classification problem.

Reference:

[Aha&Banker1995] D. W. Aha and R. L. Banker, “A Comparative Evaluation of Sequential Feature Selection Algorithms”, Proceeding of the Fifth International Workshop on Artificial Intelligence and Statistics, pp. 1-7, 1995

[Boland1999] M. V. Boland, “Quantitative Description and Automated Classification of Cellular Protein Localization Patterns in Fluorescence Microscope Images of Mammalian Cells”, Ph.D. Dissertation, Department of Biomedical Engineering, CMU, 1999

[Caruana&Freitag1994] R. Caruana and D. Freitag, “Greedy Attribute Selection”, Proceedings of Eleventh International Conference on Machine Learning, pp. 28-36, 1994

[Craven1999] M. Craven, J. Kumlien, “Constructing Biological Knowledge Bases by Extracting Information from Text Sources”, Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pp. 77, Aug. 1999

[Dash&Liu1997] M. Dash and H. Liu, “Feature Selection for Classification”, Intelligent Data Analysis, vol.1, no. 3, pp. 131-156, 1997

[Djreaba2000] C. Djreaba, “When Image Indexing Meets Knowledge Discovery,” Proceedings of the International Workshop on Multimedia Data Mining, Aug. 2000

[Duda&Hart1973] R. Duda and P. Hart, "Pattern Classification and Scene Analysis", New York: Wiley, 1973

[Faloutsos etc.1994] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic and W. Equitz, "Efficient and effective querying by image content", Journal of Intelligent Information System, vol. 3, pp.231-263, 1994

[Favela&Meza1999] J. Favela, V. Meza, "Image-Retrieval Agent: Integrating Image Content and Text", IEEE Intelligent Systems, pp. 36-39, Sept./Oct., 1999

[Harlick1979] R. M. Haralick, "Statistical and Structural approaches to Texture," Proceedings of IEEE, vol. 67, pp. 786-804, 1979

[Jain,Duin&Mao2000] A.K. Jain, R. Duin and J. Mao, "Statistical Pattern Recognition: A Review", IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.22, pp.4-38, 2000

[John,Kohavi&Pfleger1994] G. John, R. Kohavi, K. Pfleger, "Irrelevant Features and the subset selection problem", Proceedings of the Eleventh International Conference on Machine Learning, pp. 121-129, 1994

[Klecka1980] W. Klecka, "Discriminant Analysis, Quantitative Applications in the Social Sciences" Sage University Paper, vol. 19, Beverly Hills and London, 1980

[Kohavi&Sommerfield1995] R. Kohavi and D. Sommerfield, "Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology", Proceedings of First International Conference on Knowledge Discovery and Data Mining, pp. 192-197, 1995

[Langley1994] P. Langley, "Selection of Relevant Features in Machine Learning", Proceedings of the AAAI Fall Symposium on Relevanxe, pp. 1-5, 1994

[Markey,Boland&Murphy1999] M. K. Markey, M. V. Boland, and R. F. Murphy, "Towards Objective Selection of Representative Microscope Images", Biophysical Journal, vol. 76, pp. 2230-2237, 1999

[Moore&Lee1994] A. W. Moore and M. S. Lee, "Efficient Algorithms for Minimizing Cross Validation Error", Proceedings of Eleventh International Conference on Machine Learning, pp. 190-198, 1994

[Murphy,Boland&Velliste2000] R. F. Murphy, M. V. Boland, M. Velliste, "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), pp. 251-259, Aug. 2000

[Pentland,Picard&Sclaroff1994] A. Pentland, R.W. Picard and S. Sclaroff, “Photobook: Content-based Manipulation of image databases”, Proceeding of Storage Retrieval Image Video Databases II, pp. 34-47, 1994

[Ripley1996] B. Ripley, “Pattern Recognition and Neural Networks”, Cambridge, U.K.: Cambridge University Press, 1996

[Valiaya,Figueiredo,Jain&Zhang2001] A. Vailaya, A. Figueiredo, A. K. Jain and H. J. Zhang, “Image Classification for Content-Based Indexing”, IEEE Transactions on Image Processing, vol. 10, no. 1, 2001

[Valiaya,Zhong&Jain1996] A. Valiaya, Y. Zhong and A. K. Jain, “A Hierarchical System for Efficient Image Retrieval”, Proceeding for International Conference on Pattern Recognition, Aug. 1996

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.