10-610 KDD LAB Proposal Revised

Feature Selection for Fluorescence Image Classification

Jie Yao

CALD, SCS, CMU

Abstract

We propose research on the application of feature selection technique to the problem of Fluorescence image classification. That is, for the problem of classifying the fluorescence microscope images extracted from on-line biological journals, we try to find out the feature subset achieving the highest classification accuracy. Because of the diversity and noise in the images data, it is very difficult to extract the proper features manually. We are going to apply machine learning techniques to find the feature set automatically. In this project, we consider the techniques of ***Feature Selection and Feature Grouping***. By feature selection, we want to reduce the redundancy of the features set by eliminating the irrelevant features; by feature grouping, we want to preserve the feature robustness. Feature grouping can be viewed as an extension to the standard feature selection approach.

Key words:

Feature selection, Image classification, Feature grouping, Fluorescence

1. Introduction

1.1 Curse of dimensionality

The course of dimensionality [Duda&Hart1973][Jain, Duin&Mao2000] is well known for the machine learning community. It refers to the phenomenon of more input features decrease the classification accuracy . This bad effect of more input features is due to the irrelevant features to the specific classification problem. The existence of irrelevant and redundant features can overwhelm the relevant features and lead the classifier to a wrong way. In fact, the quality of features is critical for nearly all machine learning problem. For some problems, finding the proper feature set is more important than algorithms choosing. The fluorescence image classification problem investigated in this project is such a problem.

1.2 Fluorescence image classification problem

Fluorescence image classification problem classify the fluorescence microscope images from all other images extracted from on-line journal. In this project, we use the 1861 images collected from 110 articles of *Journal of Cell Biology, vol.136, 1997*. Due to the broad subject coverage of the journal, the images cover a range of image types, species type, and microscope types. The distribution of different image types shown in Table 1.

Image Type	Fluorescence Microscope Image	Transmitted Microscope Image	Electron Microscope Image	Gel	Chart& Graph	Other
Number	788	220	173	289	35	81

Table 1

We are more interested into the Fluorescence microscope images, since they are very important in the determination of the Protein Sub-cellular Location Pattern [Boland1999] [Murphy,Boland&Velliste2000].

The problem is not easy because that

1) The fluorescence microscope images are much diverse due to the difference of conditions including the experiment material, research goal and microscopes used.

2) The noise that is introduced during the process of publishing and extraction makes the classification problem even harder.

For this problem, it is very difficult to choose an optimal and robust feature set manually. With optimal, we mean the size of the feature set. With robust, we mean the capability of describing the diverse and noisy image data. We plan to employ feature selectionJ and feature grouping techniques to choose the feature set automatically.

1.3 Feature selection and grouping

The problem with the features is not the lack of features, but on the contrary, we have too many features. A blind choice of features unavoidable include many irrelevant features that will do harm to the classification performance. So, the application of feature selection and grouping techniques are a must for our fluorescence image classification problem.

With feature selection, we want to reduce the size of the feature set and deleting the irrelevant features. The small size and more relevant feature set will increase the classification accuracy and reduce the computation time. The outputs of feature selection algorithms are optimal, but such optimal feature set may perform badly in some practical problems. The reason is that the minimal feature sets depend too heavily on the specific classification problem and the known data set. By grouping features first according to their behaviors, we can pre-select one or two typical features for all the groups to make improvement of robustness.

Our candidate feature sets are the histogram feature set and Sub-cellular Location Feature (SLF) set [Boland1999]. The histogram features have been used in the research of Content-Based Image Retrieval (CBIR) and been proven good at capturing the overall visual information of the images. [Faloutsos etc.1994] [Pentland,Picard&Sclaroff1994] The SLF set, which includes texture features, moment features and morphological features, have been used to describe complex sub-cellular patterns in the images [Murphy, Boland&Velliste2000] A detailed description of the SLF features can be found in the thesis of M. Boland [Boland1999].

1.3 Goal and Impact

The goal of this project is finding an optimal and robust feature set for the fluorescence image classification problem. This research will have a great impact on the image classification and retrieval applications, especially the open source image classification problems.

2. Related Work

Feature selection has been studied by statistics and machine learning researchers for many years. Detailed survey of feature selection can be found in [Langley1994] [Blum&Langley1997] [Dash&Liu1997]. The other useful publications on feature selection are [Aha&Banker1995][Caruana&Freitag1994][John,Kohavi&Pfleger1994] [Ng1998] [Kohavi& Sommerfield 1995].

In the traditional content-base image retrieval research, human experts select features. Decisions are usually based either on domain specific knowledge or on past performance of features. In the case of lots of possible features but little domain knowledge, the above method will fail. So recently, more and more research has been done on the automatic feature selection. Our approach differ from other feature selection research in the following aspects:

1. Our image data are more challenging. The images extracted from on-line journal are more diverse and noisy than other image data.

2. We consider both the efficiency and robustness in a simple way by adding grouping before feature selection.

The closest research to this project is 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 [Harlick1979], 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 efficient algorithm.

[Swets&Weng1995] proposed a self-organized framework to organized the features used in the content-based image retrieval system. They use K-L projection to generate most expressive features and use discriminant analysis projection to generate most discriminating features. The experiment result of face image retrieval is rather high. But the selected features(non-linear combination of features) are not explicit meaningful .

[Jaimes&Chang2000] proposed a dynamic approach to feature and classifier selection for the segmentation task of baseball video image. They also use cross-validation as the evaluation of features. The classification of the image region is easier than the classification of whole image because the homogeneity of region.

[Valiaya, Figueiredo,Jain&Zhang2001] utilized sequential floating forward selection(SFFS) and feature clustering (FC) techniques in their research of scenery (indoor/outdoor, city/landscape etc.) image classification. Their SFFS selected feature set (52 from 600) has a little decrease (87% for subset feature and 88.2% for whole feature) in the accuracy. And they also report a very long computing time. The Feature Clustering (linear combination) technique has been used as a simple alterative for SFFS. The experiment result shows FC is better than SFFS.

[Vafaie&De Jong] also noticed the brittleness of the result of feature selection. It is also shown that the brittleness is the tendency to get trapped at local minima caused by interdependency among the features. They employ GA algorithm for feature selection for improving the robustness without sacrificing too much in speed and accuracy. Feature grouping can be viewed as a simple heuristic approach to avoid trapped into local minima.

3. Approach

3.1 Background of Feature Selection

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

1. **Search algorithm**

Feature selection can be viewed as a search problem in the space of feature subsets according to the evaluation function. There are three categories of search algorithm: exponential, sequential and randomized. Exponential algorithms do a complete search for the optimal subset. 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.

2. **Feature evaluation function**

Feature evaluation function assign a score to a feature subset. There are five kinds of evaluation function.

Distance measure, also known as *separability* and *discrimination* measure, examines the difference between the conditional probabilities.
Information measure uses the information gain from a feature.
Dependence measure or *correlation* measure uses the conditional probability of one

variable given the value of another variable.

Classifier error rate measure is also called *wrapper methods* [Kohavi& Sommerfield 1995] uses the error or accuracy of the classifier on a test dataset.

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 cross-validation accuracy as the evaluation function.

3.2 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. It is shown that FSS outperforms BSS [Aha&Banker1995].

3.3 Feature Grouping

For the distance-based classifiers, like k-NN, the contribution of a feature to classification is determined by its values over the data. So, the distance of features can be calculated from the feature vector over the data. The formal definition of the distance between features is:

Let $A_{m\times n}$ represent a $m \times n$ data table, where *m* is the number of instances and *n* is the number of features.

Let *f(i), i=1,2, n* represent the features, $a(i, j), i = 1,2,...m, j = 1,2...n$ represent the value of feature *j* at instance *i*.

The distance of feature *f(i)* and *f(j)* can be defined as: $\sum_{k=1}^{m}[a(k,i) - a(k, j)]^2$.

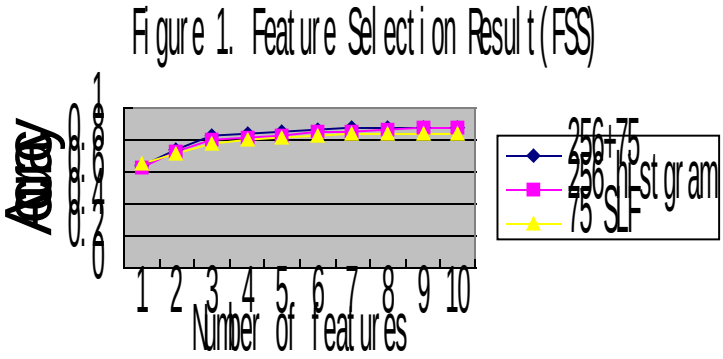
Under the above distance definition of the feature, we will use k-means clustering method to implement the feature grouping . The parameter k will be determined manually. From the grouped features, we can pre-select one or two feature as the starting feature set for FSS method. For BSS, the pre-selected features can be viewed as irremovable features.

3.4 Evaluation

For a feature selection problem, both the finally selected feature set and its classification accuracy are interested. We will use feature - accuracy curves as the basic evaluation of the feature selection. Like other machine learning problem, the data will be divided to training set and testing data randomly. The training date here will be used for feature selection; the testing data will be used for the evaluation of the accuracy of the finally selected feature set. The experiment will be repeated several times. The robustness can also be tested the same way.

3.5 Prior Results

Figure 1 shows the feature selection (FSS) results and its k-NN classifier accuracy. The three candidate feature sets are 75 SLF, 256 histogram feature set and the combination of them. It can be seen that the small number of selected features achieve satisfied classification accuracy.



4. Work Plan

Feb. 5-10 Image feature calculation.

Feb. 11- Mar. 10 Work on Feature selection, including the FSS and BSS algorithms.

Mar. 11- Mar. 30 Work on feature grouping.

April 1-10 Project write up

4. Summary

We propose an application of feature selection techniques to fluorescence image classification problem. Our project try to find the image features that meet the requirement of efficiency, accuracy and robustness.

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

[Blum&Langley1997] A.L. Blum and P. Langley, Selection of Relevant Features and Examples in Machine Learning , Artificial Intelligence, vol. 97, pp.245--271, 1997

[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

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

[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

[Harlick1979] R. M. Haralick, ¡° Statistical and Structural approaches to Texture¡± Proceedings o IEEE, vol. 67, pp. 786-804, 1979

[Jaimes&Chang2000], A. Jaimes and S. F. Chang, Automatic Selection of Visual Features and Classifiers , IS&T/SPIE Storage and Retrieval for Image and Video Databases VIII, vol. 3972, San Jose, Jan. 2000

[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

[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

[Ng1998] A.Y. Ng, On Feature Selection: Learning with Exponentially Irrelevant Features as Training Examples , International Conference on Machine Learning, 1998

[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

[Swets&Weng1995] D. L. Swets and J. J. Weng, Efficient Content-based Image Retrieval Using Automatic Feature Selection , International Conference on Computer Vision, Nov. 1995

[Vafaie&De Jong1993], H. Vafaie and K. De Jong, Robust Feature Selection Algorithms , Proceedings of the Fifth Conference on Tools for Artificial Intelligence. pp. 356 363, Boston, IEEE Computer Society, 1993

[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.