Project Proposal(Modified)

Yubo Feng

April 3, 2015

Outline

In this project, we will explore learning of pollen image classification: given a set of pollen images and corresponding labels, we will build various models to learn image pattern and try to do multi-classification, then we compare different performances of models, and try to find a relative better model to be the classifier of this problem. To be more specific, images are different kinds of pollens and labels are type of the plants that this kind of pollen belongs to. The model we try to use including important methods within the class: based on naive bayes network as baseline, non-parameter classification method like KNN and parameter method like NN, KVM and so on.

Data Set

In this project, we choose a public dataset from Pollen Laboratory at the The University of Newcastle, Callaghan, NSW, Australia. The dataset contains about 50+ types different geometry characteristics pollen classes, 200 different pollen types, about 1000+ pollen images. However, in this project, in order to make things simpler, we only choose 10 types of pollen (# of class), each contains at least 4 images and at most 7 images, average 6; in another word, dataset will contain 60 sample images; the reason why we do in this way lies in that the dataset is no longer a simple txt file, but mixtures of pollen images and titles, which means it takes a lot of time to filter the image and label them; so we only choose some of them to do our project instead of do them all.

Importance

Over 20% of all the world?s plants are already at the edge of becoming extinct. Saving earth?s biodiversity for future generations is an important global task and as many methods as available must be combined to achieve this goal. This involves mapping plants distribution by collecting pollen and identifying them in a laboratory environment. However, pollen grain classification has been an expensive qualitative process, involving observation and discrimination of features by a highly qualified palynologist. It is still the most accurate and effective method. But it certainly limits research progress, taking considerable amounts of time and resources. Automatic recognition of pollen grains can overcome these problems, producing purely objective results faster. Therefore, as we could see, it is necessary to develop good classifier for this problem; recent year, machine learning methods were widely used, different models are built and various approaches were tried, it is still a interesting area.

Metholds

Recent works concentrate on two approaches:

- Traditional Shallow Architecture for classification: pre-processing Feature Extraction → trainable classifier → test classifier. For this method, the feature itself cannot be learned, classifier is naive, like linear classifier, kernel machine, nearest neighbor, also the most frequently used algorithm is Kernel machine.

- Deep learning Hierarchical Representations: learning a hierarchy of internal representations from low-level features to mid-level invariant representations, to object identities. Representations are increasingly invariant as we go up the layers

In this project we will choose traditional approach, but cover a lot of methods. The whole project process will be consisted by data pre-processing, feature reduction, multiple models training and test. More details as follows:
(1). for data process, since we get 60 sample images for 10 types, then there is a limitation for source data set, in this case, k-fold cross validation will be used to create sample points.
(2). a lot of work will be focused on feature extraction: according to previous works, for a single image, according to pollen's geometric characteristics, 48 different features could be conducted, by using feature reduction algorithms, like PCA, we could choose some significant feature to be our final datasets.
(3). For model training, we will choose naive bayes to be our baseline, then we will use Multi-Logistic Regression, NN, KVM, Decision Tree, KNN, also, to make things more interesting, mixture method will also be used, like bagging and boosting based on model mentioned above. Finally, several models will be tested and compared to show their ability of classify.
Besides above steps, we will explore the impact of feature selection on model selection, in order to find the best features for classifier to use.

Previous Work

Ma da lina Cosmina Popescu, Lucian Mircea Sasu, *Feature extraction, feature selection and machine learning for image classification: A case study*, OPTIM, 2014.
M. Langford, G. Taylor, and J. Flenley, *Computerized identification of pollen grains by texture analysis*, Review of Palaeobotany and Palynology, vol. 1, no. 4, pp. 197?203, Nov. 2010.
M. Rodrguez-Damin, E. Cernadas, A. Formella, and P. de S-Otero, *Pollen classification using brightness-based and shape-based descrip- tors*, ICPR (2), 2004, pp. 212?215.
J. Fan and J. Lv, *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica, vol. 20, no. 1, pp. 101? 148, 2010.
Marcos del Pozo-Baos, Jaime R. Ticay-Rivas, Jous Cabrera-Falcn, *Image Processing for Pollen Classification*, *Biodiversity Enrichment in a Diverse World*, Chapter 19, ISBN 978-953-51-0718-7, 2012.
P. LI, W. J. TRELOAR, J. R. FLENLEY* and L. EMPSON, *Towards automation of palynology 2: the use of texture measures and neural network analysis for automated identification of optical images of pollen grains*, JQS, 2004.

Test

In test part, since we use k-fold cross validation to separate and create dataset, then test data are easy to identify; we will use several methods to test our models: for examples, miss-classification error rate, confusion matrix. After these, we will compare the influence of different feature selection strategy by testing models on different sets of features, then test their accuracy.

Schedule

Learning methods: 1 weeks, including raw data processing and different feature set building
Coding and testing: 1 week, including programing code of different models and training model, test model, getting experiment results
Report composition: 1 week, summary, report building