Significance of Reduced Features for Subcellular Bioimage Classification

Babu Kaji Baniya, Carol Lushbough, and Etienne Z. Gnimpieba

Department of Biomedical Engineering
University of South Dakota
GEAR Center, 4800 N. Career Ave., Sioux Falls, SD 57107
{babu.baniya,carol.lushbough,etienne.gnimpieba}@usd.edu

Abstract. High-throughput screening (HTS) system has the capability to produce thousands of images containing the millions of cells. An expert could categorize each cell's phenotype using visual inspection under a microscope. In fact, this manual approach is inefficient because image acquisition systems can produce massive amounts of cell image data per hour. Therefore, we propose an automated and efficient machine-learning model for phenotype detection from HTS system. Our goal is to find the most distinctive features (using feature selection and reduction), which will provide the best phenotype classification both in terms of accuracy and validation time from the feature pool. First, we used minimum redundancy and maximum relevance (MRMR) to select the most discriminant features and evaluate their corresponding impact on the model performance with a support vector machine (SVM) classifier. Second, we used principal component analysis (PCA) to reduce our feature to the most relevant feature list. The main difference is that MRMR does not transform the original features, unlike PCA. Later, we calculated an overall classification accuracy of original features (i.e., 1025 features) and compared with feature selection and reduction accuracies ($\sim$30 features). The feature selection method gives the highest accuracy than reduction and original features. We validated and evaluated our model against well-known benchmark problem (i.e. Hela dataset) with a classification accuracy of 92.70% and validation time in 0.41 seconds.

Keywords: features reduction, selection, phenotype, validation, MRMR

1 Introduction

Challenges in bioimages: nowadays, cell imaging is an emerging field focusing on the analysis of massive amount of cell images data produced by HTS systems. From these images, biologists can analyze the morphology of these cells and extract corresponding phenotypes such as subcellular compartments using visual inspection on a microscope. This approach is only effective under the small number of cell images. However, in the presence of millions of cells, visual classification becomes exhaustive and time-consuming. Even though automatic cell classification research has received tremendous attention in recent years, it still faces many challenges such as precise characterization of bioimages features selection and the classification of a subcellular compartment in their respective class or phenotypes [1].

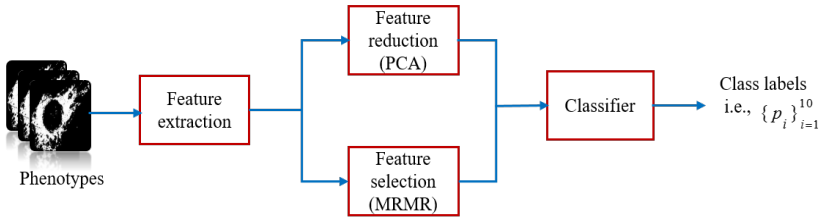

Fig. 1: Workflow of the proposed model

Contribution: in this paper, we developed a new model to identify the most discriminative features (through selection and reduction) from the (feature) pool using MRMR and PCA. We found that only minimum number of feature ($\sim$30 out of 1025 dimensions of the feature vectors) are sufficient to classify images into their respective phenotypes or classes with a better accuracy and computing time. Unlike other classifiers, we adopted the SVM because it can efficiently perform a non-linear classification using kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. We measured the performance of the system without feature reduction and got the accuracy 88.03%. Later, feature reduction and selection (using PCA and MRMR) techniques give an accuracy of 84.57% and 92.70% in 0.54 and 0.41 seconds respectively (Fig. 2).

2 Proposed Cell Classification System

The proposed model is shown in Fig.1. This includes bioimage dataset, feature extraction, and feature analysis before the classification. Feature extraction is a primary paradigm in classifying the phenotypically distinct cell within a species. Bioimage features can be highly diverse; therefore, the challenges lie in finding which cell feature is the most discriminative among the feature pool. There are different types of feature reduction and selection methodologies. We used PCA and MRMR for feature reduction and selection, where PCA is an orthogonal linear transformation that transforms the input feature to a new (feature) space. MRMR calculates the score of feature and rank them based on score level for selection [2]. These two are used to remove the irrelevant features before classification. The decision of classifying the features belongs to one of the classes i.e., $\{p_i\}_{i=1}^{10}$ (p represents phenotype) at each output.

Feature extraction: concise and relevant feature extraction and selection is a core issue in pattern recognition [3]. We implemented a set of general purpose features that were commonly used in machine learning and pattern recognition. Several features, previously used for distinguishing phenotype pattern classification obtained promising results [4]. For our application, we extracted different features such as morphological features, Haralick textures [5], and Zernike moment features [6]. This integration allows us to cover the maximum knowledge from an input image.

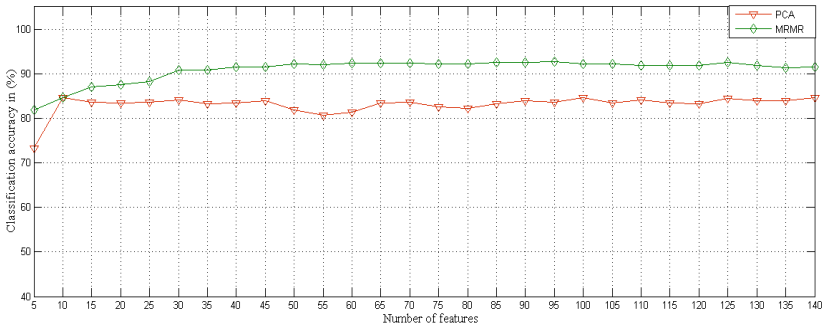

Fig. 2: Comparison of accuracies of original features with reduced (PCA) and selected (MRMR) features using SVM

3 Experiments

Dataset: for our classification scheme, we used a collection of 2D images from the HeLa dataset and publicly available [7]. The images include 10 organelles, which are Actin, DNA, Endosome, Endoplasmic Reticulum, Golgia, Golgpp, Lysosome, Microtubules, Mitochondria, and Nucleolus respectively.

Results and discussion: the proposed machine learning model lead us to several results (under the 10-fold cross-validations). First, we obtained the highest overall classification accuracy in the combination of all phenotypes (92.70%). Second, this accuracy was achieved by considering less than 30 bioimage features out of 1025 features. Third, we also selected the most discriminative bioimage feature from feature pool that led the new direction of phenotype classification and its interpretation. At the beginning, we performed an experiment by considering all bioimage features. This experiment aims to explore the classification accuracy with the features growth in an interval of 25. The experiment result showed that the classification accuracy increased up to 85.71% in 700 feature dimensions. After 700 features, no significant improvement in accuracy was observed. In the next stage, we performed 10-fold cross-validations using feature selection and reduction in a small interval (i.e., 5, 10, 15, 20, 25 etc.). We found a significant classification accuracy improvement within first 10 to 15 feature dimensions (out of 140 the most discriminative features). With MRMR, classification accuracy gradually increased up to 90% by considering 30 features, thereafter we observed that the accuracies remained stable up to the maximum feature number (Fig. 2). Furthermore, MRMR not only gave the higher accuracy but also selected the particular bioimage feature that contributed for the better classification accuracy. In case of PCA, original features transformed in to new feature-space. The transformed feature took for validation and accuracy increased up to 10-feature dimension i.e., shown in Fig. 2. The classification accuracy remained stable up to 140 feature vectors. This means that a minimum number of feature dimensions are sufficient to differentiate the phenotypes from each

other. Our automated classification system took the minimum validation time and higher classification accuracy as compared to Abbas at el. model [8] (in 10-phenotypes case).

4 Conclusion and Further Work

The objective of our proposed cell classification model is to improve the performance and minimize the computation time of existing systems by considering the minimum number of image features from pool. Beside these, we also found the particular bioimage feature for phenotype classification. We have achieved this by implementing the MRMR feature selection technique with SVM classifier. The redundant features were controlled by using the feature reduction and selection techniques i.e., PCA and MRMR. These reduced and selected features contribute to the accuracy and minimize validation time. We validated our method and evaluated it with the well-known benchmark problem (i.e., HeLa dataset). We achieved a classification accuracy of up to 92.7% at 0.41 seconds on average.

Our immediate plan is to extend the feature selection and reduction techniques (such as linear discriminant analysis, locality preserving projecting, factor analysis) for a particular phenotype to optimize the classification accuracy. This will be a genuine contribution across the bio-science plate-forms and multi-omics levels.

Acknowledgment: This work has been partially supported by the National Science Foundation/EPSCoR Award No. IIA-1355423 and by the state of South Dakota, through BioSNTR. And the Institutional Development Award (IDeA) from the National Institutes of Health for SD BRIN grant number P20GM103443.

References

1. I. Chebira and S. Streichan, "Tissue cartography: compressing bio-image data by dimensional reduction," *Nat Meth*, vol. 12, Dec 2015.
2. H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug 2005.
3. B. K. Baniya, J. Lee, and Z. N. Li, "Audio feature reduction and analysis for automatic music genre classification," pp. 457–462, Oct 2014.
4. M. V. Boland and M. K. Markey, "Automated recognition of patterns characteristic of subcellular structures in fluorescence microscopy images," *Cytometry*, vol. 33, no. 3, pp. 594–597 vol. 2, 1998.
5. R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.
6. M. R. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America*, vol. 70, no. 8, pp. 920–930, 1980.
7. "The Murphy Lab at Carnegie Mellon University," *http://murphylab.web.cmu.edu*.
8. S. S. Abbas, T. H. Dijkstra, and T. Hekes, "A comparative study of cell classifiers for image-based high-throughput screening," *Bioinformatics*, vol. 15, no. 342, Oct 2014.