Decision tree and bagging algorithm for the automatic identification of epithelial cell of wound

Hai Ling LIU, Xian ZANG
School of Electronic and Information Engineering, Jingling Institute of Technology, Nanjing, Jiangsu 210069, China
hailingliu@jit.edu.cn

Abstract. In this paper, a pattern recognition algorithm of automatic identification of epithelial cell was researched. Decision tree and bagging algorithm were used to find the most important features for the automatic identification of epithelial cell of wound. Different features were used to evaluate the identification performance. Three features are regarded as the most important ones. The method and algorithm can also be used to other applications with respect to automatic cell recognition.

1. Introduction
The motion of epithelial cell is very important to the healing of wound. The research on the analysis and mathematical model of the microscopic image provide an effective tool for this aim. Through analyzing the parameters of motion in specific environment for a type of epithelial cell and setting up the motion model, we can understand the mechanics of wound healing quantitatively. The premise of all these is to find the most distinguishing characteristics and identify the epithelial cell from its background.

1.1. Cell culture and drug treatment
HCT116 human epithelial cell line obtained from American Type Culture Collection (Manassas, USA) was grown in RPMI 1640 medium (Invitrogen, USA) supplemented with 10% (v/v) fetal bovine serum (FBS; Invitrogen) and 1% penicillin-streptomycin (Welgene, Korea) in a 37°C humidified incubator in an atmosphere of 5% CO2. Drug treatment of cells was performed by adding 50 M oxaliplatin (L-OHP; Boryung Pharmaceutical, Korea) or vehicle (for control sample) to the culture medium and incubating for 48 h.

1.2. Hoechst 33342 (HO) staining
As nuclear condensation and fragmentation is best well known features of dead epithelial cells and this is also very simple and easy way to detect them. Therefore, it is a well-established method widely used in the detection of cell that the nuclei are stained to observe their alteration. In our experiments, the cells were incubated with 1 µg/ml HO for the final 10 min of drug treatment in the 37°C incubator and then, both floating and attached cells were collected by centrifugation. The pooled cell pellets were washed with ice-cold phosphate-buffered saline (PBS), fixed in 3.7% formaldehyde on ice, washed again with PBS, and a fraction of the suspension was centrifuged in a cytopspiner (Thermo Shandon, Pittsburgh, PA).
1.3. Experimental images
The original image size captured is 2560 × 1920. Fig.1 shows the light image and the Hoechst image in our experiment. In Fig.1 (b), it is easy to see that: (1) live epithelial cells are of low brightness, big size and circular shape and dead ones due to wound are of high brightness, tiny size and near-circular shape.

1.4. Feature extraction and analysis
In the normal framework of image-based cell recognition system, the image preprocessing and image segmentation are two initial steps. However, they will not be elaborated in this paper. We only focus on the features. A good feature should remain unchanged if variations take place within a class, and it should reveal important differences when discriminating between patterns of different classes. In other words, patterns are described with as little loss as possible of pertinent information.

In this experiments, for each case, 10 features were estimated automatically from illumination intensity, morphological and textural nuclear features [1]. Information about nuclear size and shape was captured by morphological features, which constituted measurements of nuclear area, roundness and concavity [1]. The concavity attempts to measure the severity of concavities, or the indentations of a nucleus [1]. The remaining two textural features that encoded chromatin distribution of the cell nucleus. These features were estimated by means of nuclear histograms and the co-occurrence matrix [2].

Table.1 lists some typical images of normal epithelial cells, dead cells and their respective feature values. Ten features were used in our experiments and the description for each feature is listed in Table.2.

1.5. Pattern recognition research
In this paper, the importance of all features and the classification accuracy were evaluated. The classification algorithm used here is decision tree algorithm.

1.5.1. Decision tree algorithm
Decision tree algorithm has been used broadly for several years. It is an approximate discrete function method and can yield many useful expressions. It is one of the most important methods for classification. This algorithm’s terms follow the “tree ”metaphor.It has a root, which is the first split point of the data attribute for building a decision tree. It also has leaves, so that every path from root to leaf will form a rule that is easily understood.
Since the decision tree is built by given data, the data value and character will be more important. For example, the amount of data will affect the result of the tree building procedure. The type of attribute value will also affect the tree model. Decision trees need two kinds of data: training and testing. Training data, which are usually the bigger part of data, are used for constructing trees. The more training data collected, the higher the accuracy of the results. The other group of data, testing, is used to get the accuracy rate and misclassification rate of the decision tree. Many decision tree algorithms have been developed. One of the most famous one is ID3 [3,4], whose choice of split attribute is based on information entropy. C4.5 is an extension of ID3 [5], which improves computing efficiency, deals with continuous values, handles attributes with missing values, avoids over-fitting, and performs other functions.

1.5.2. Bagging algorithm
In data mining, an approach to make decisions more reliable is to combine the outputs of different models. Several machine learning techniques do this by learning an ensemble of models and using them in combination: prominent among these is a scheme called “bagging” [6].

Bagging predictor is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and performs a plurality vote when predicting a class. The multiple versions are created by making bootstrap replicate of the learning set and using these as new learning sets. Test on real and simulated data sets using classification and regression trees and subset selection in linear regression have shown that bagging can provide substantial gains in accuracy [7].

Bagging attempts to neutralize the instability of learning methods by simulating the process using a given training set. Instead of sampling a fresh, independent training dataset each time, the original training data is altered by deleting some instances and replicating others. For the replacement, instances are randomly sampled from the original dataset to create a new one of the same size. This sampling procedure inevitably replicates some of the instances and deletes the others.

2. Experiment and discussion

2.1. Data description
The dataset includes 10 categories of features extracted from 25000 cells in 25 microscopic images. The total number of data is $10 \times 25000$. 
Table 1 Typical Features Comparison Between Normal Epithelial Cell And Dead Epithelial Cell

| No. | Image | MeanInt | MedInt | StdInt | MajAxl | MinAxl | Peri | Area   | Eccen | Contr | Entr |
|-----|-------|----------|--------|--------|--------|--------|------|--------|-------|-------|------|
| 1   |       | 0.298    | 0.315  | 0.086  | 309.54 | 296.14 | 1028.12 | 71906  | 0.29  | 0.17  | 2.39 |
| 2   |       | 0.319    | 0.333  | 0.07   | 254.59 | 242.17 | 866.33  | 48240  | 0.31  | 0.24  | 2.42 |
| 3   |       | 0.376    | 0.392  | 0.044  | 295.55 | 208.87 | 895.51  | 47877  | 0.71  | 0.23  | 1.75 |
| 4   |       | 0.318    | 0.329  | 0.027  | 268.75 | 251.15 | 891.15  | 53179  | 0.36  | 0.24  | 2.47 |
| 5   |       | 0.285    | 0.293  | 0.052  | 294.53 | 286.89 | 994.41  | 66592  | 0.23  | 0.21  | 2.29 |
| 6   |       | 0.342    | 0.371  | 0.066  | 268.23 | 253.94 | 897.08  | 53418  | 0.32  | 0.26  | 2.23 |
| 7   |       | 0.468    | 0.447  | 0.097  | 146.19 | 133.73 | 496.21  | 15153  | 0.41  | 0.28  | 2.54 |
| 8   |       | 0.466    | 0.454  | 0.121  | 183.65 | 159.58 | 584.65  | 23069  | 0.49  | 0.33  | 2.53 |
| 9   |       | 0.489    | 0.481  | 0.135  | 145.96 | 133.81 | 476.75  | 15331  | 0.41  | 0.34  | 2.65 |
| 10  |       | 0.472    | 0.452  | 0.103  | 129.31 | 111.13 | 410.02  | 11308  | 0.51  | 0.39  | 2.73 |
| 11  |       | 0.523    | 0.511  | 0.129  | 146.52 | 140.36 | 487.16  | 16172  | 0.29  | 0.32  | 2.51 |
| 12  |       | 0.503    | 0.497  | 0.112  | 132.01 | 106.45 | 405.78  | 11057  | 0.59  | 0.41  | 2.63 |

Table 2 Feature Description

| No. | Name     | Description                                                                 |
|-----|----------|-----------------------------------------------------------------------------|
| 1   | MeanInt  | Mean of pixel intensity values                                              |
| 2   | MedInt   | Median of pixel intensity values                                            |
| 3   | StdInten | Standard deviation of pixel intensity values                                |
| 4   | MajAxl   | The length (in pixel) of the major axis of the ellipse shape cell           |
| 5   | MinAxl   | The length (in pixel) of the major axis of the ellipse shape cell            |
| 6   | Peri     | The total number of pixels around the boundary of each cell                 |
| 7   | Area     | The actual number of pixels inside the cell region                          |
| 8   | Eccen    | The eccentricity of the ellipse shape cell (Value between 0 and 1. 0 is actually a circle and 1 is line segment) |
| 9   | Contr    | Co-occurrence texture feature contrast                                      |
| 10  | Entr     | Co-occurrence texture feature entropy                                       |

2.2. Experiments
In order to compare the experimental results of Bagging decision tree algorithm, another two classification methods were also applied here, decision tree algorithm and Bagging Naïve Bayes, All
experiments were implemented in Matlab 2011B platform and using Matlab statistical toolbox. The first step to construct the classification ensemble will be to find a good leaf size for the individual trees. Here three different sizes of 4, 6 and 10 were tried. 30 trees were used. For reproducibility and fair comparisons, the random number generator were reinitialized, which was used to sample with replacement from the data in each time we build a classifier. The error are comparable for the three leaf-size options. A leaf size of 10 will be used because it results in leaner trees and more efficient computations. Here we did not split the data into training and test subsets. This is done internally, it is implicit in the sampling procedure that underlines the method. At each bootstrap iteration, the bootstrap replica is the training set, and any customers left out (“out-of-bag”) are used as test points to estimate the out-of-bag classification error reported above.

Next, whether all the features are important for the accuracy of our classifier was researched, which was done by turning on the feature importance measure and plot the results to visually find the most important features. It is easy to see that feature 3 (MedianIntensity), 7 (MinorAxisLength) and 8 (Area) stand out of the rest while feature 4 (Eccentricity), 9 (Contrast) and 10 (Entropy) has the least ability to separate dead epithelial cells from normal live ones. Generally speaking, illumination features and size features has most separation ability while texture features has less separation ability, which exactly demonstrate the intuitive feeling that dead and live epithelial cell give to us.

The following figure shows comparison of the classification of using all features and only 3 most contributing features. It is evident that when the number of trees is not very big, the classification results using three features are comparable to the results using all ten features.

![Feature importance graph](image1)
![Classification error using most important features](image2)

### 3. Performance measurement

Many different metrics are used in machine learning and data mining to build and evaluate models. I employed four performance measures: precision, recall and F-measure.

Less formally, precision measures the percentage of the actual live cell among the cells that were classified as live cell. Recall measures the percentage of the actual live cells that were discovered. F-measure balances between precision and recall. According to the feature importance results, it is known that MedianIntensity, MinorAxisLength and Area are three features with best classification ability. I experimented on classification results using these three features and the remaining features respectively. The training data is 320 and test data is 80. Table 3 shows the performance results of three methods.
Table 3 The classification performance using seven secondary features

| Seven features | Decision tree | Bagging decision tree | Bagging Naïve Bayes |
|---------------|--------------|----------------------|---------------------|
| Precision     | 89.60%       | 90.89%               | 91.51%              |
| Recall        | 86.59%       | 87.24%               | 88.34%              |
| F_measure     | 84.96%       | 85.12%               | 86.27%              |

Table 4 The classification performance results using three primary features

| Three features | Decision tree | Bagging decision tree | Bagging Naïve Bayes |
|---------------|--------------|----------------------|---------------------|
| Precision     | 97.34%       | 97.25%               | 98.31%              |
| Recall        | 95.08%       | 96.11%               | 97.21%              |
| F_measure     | 96.43%       | 97.18%               | 98.35%              |

4. Conclusions
In this paper, a decision tree and bagging algorithm was used to find the most important features for the automatic identification of live and dead epithelial cells. The classification performance was also compared using different feature groups. It is easy to see that MedianIntensity, MinorAxisLength and Area are three most important features for the automatic identification with best identification performance. We think that the method and algorithm can also be used in other similar applications.

Acknowledgments
The work in this paper is supported by “The research initiation funds for the returned students from abroad of ministry of education”and “Jinling Institute of Technology (JIT) under grant 40610060”.

References
[1] Spyridonos P, Ravazoula P, Cavoures D, Berberidis K, Nikiforidis G. Computer based grading of hematoxylin-eosin stained tissue sections of urinary bladder carcinomas. Med Inform Internet Med 2001;26(3):179—90.
[2] Harralick R, Shanmugam K. Textural features for image classification. IEEE Trans Syst Man Cybernet 1973;3(6): 610—21.
[3] Quinlan, J.R., “Learning effective classification procedures and their application to chess ending games.” Machine Learning: An Artificial Intelligence Approach. Ed. M. Kaufmann. Vol. III.1983.463-482
[4] Quinlan, J.R., “Induction of Decision Trees,” Machine Learning. Vol.1. 1986.81-106
[5] Prather, J.C., Lobach.D.F., “Medical data ming knowledge discovery in a clinical data warehouse.” Proc AMIA Annual Fall Symposium. 1997:p.101-105
[6] H.W. Ian, E.F., “Data mining: Practical machine learning tools and techniques.” 2005: Morgan Kaufmann
[7] B. Leo, “Bagging Predictors, in Machine Learning.” 1996. p.123-14