Confidence Estimation in Stem Cell Classification

Zahra Rajabi, Jana Kosecka
Dept. of Computer Science
George Mason University
Fairfax, USA
zrajabi, kosecka @gmu.edu

Peter Bajcsy
National Institute of Standard and Technology
Information Technology Laboratory
peter.bajcsy@nist.gov

***Abstract* - We study the problem of supervised classification of stem cell colonies and confidence estimation of the attained classification labels. The problem is investigated in the application context of heterogeneity labels of stem cell colonies observed by using fluorescent microscopy imaging. Given the features of colonies using numerous image statistics, we report the classification results using adaptive *k*-Nearest Neighbor (NN) algorithm. This algorithm minimizes typical *k*-NN classification bias by giving more weight to more informative features in predicting class posterior probabilities. We then estimate the confidence of each prediction for unlabeled data using transductive p-value and strangeness metrics. We show that such an introspection can gradually increase the accuracy of learned model, quantify false positives, and guide the resource-limited manual colony annotation process to provide training labels for the less confident unlabeled samples.**

***Index Terms*— k-Nearest Neighbor, Confidence Estimation**

I. INTRODUCTION

Recent advances in automated microscopy imaging techniques enables scientists to capture large amount of imagery to classify cell colony heterogeneity and monitor the heterogeneity changes over time. Providing biologists with automated classification capabilities is central to gaining further insights into human diseases, drug discovery, and stem cell treatments. The terabyte-sized image representations in stem cell biology necessitate establishment of not only an automated qualitative and quantitative characterization of cell colony heterogeneity under various biological conditions but also the confidence estimation of those automated analyses. In this context, automated classification analyses must be robust and have an introspection capability of quantifying uncertainty for the predicted class labels. In this paper, we use an adaptive *k-Nearest Neighbor (*NN) classification of heterogeneity of stem cell colonies, and describe a confidence metric to predict the uncertainty of classification.

II. RELATED WORK

One area of research in computational cell biology is automated characterization of stem cells in stem cell therapies. The characteristics include, for instance, colony growth rate and function (i.e., pluripotency). The function of a colony can be measured via heterogeneity of cells forming the colony and their protein expression. This colony heterogeneity is perceived as a textured image region that can be labeled manually by experts. A common approach to classification in cell microscopy images is to use some manually labeled colony instances in a supervised setting to generate models. The models can be used for future classifications/predictions and ideally have good generalization capability. In [7], a supervised segmentation technique based on Gaussian Mixture Model (GMM) Bayes classifier assigns pixels in sub-cellular images into biologically meaningful regions. The biases from such model assumptions using GMM are analyzed in [5]. The work refutes the conditional independence of cellular visual features (intensity, texture, shape). Similarly, the k-Nearest Neighbor (NN) classifiers can also suffer from a bias due to the assumption of locally constant class conditional probabilities. To address this bias one can use an adaptive k-NN algorithm proposed in [1]. The adaptive k-NN method considers feature relevance while identifying neighbors.

In addition to the classification bias, we address the confidence estimation of obtained classification results which requires some knowledge of ground truth labels. It is shown in [5] that classification accuracy can be estimated as a function of GMM components. However, we need to provide a confidence measure based on the uncertainty of a classifier. The confidence measure assigns a probability that unlabeled test data is misclassified.

Another restricting issue with classical supervised learning methods is the limited number of labeled instances. An insufficient number of training samples (labeled instances) often prevents us from learning a classification model with good generalization capacity. Furthermore, obtaining annotations for training samples, specifically in highly specialized areas such as cell biology, requires exact expertise and is a time consuming effort. Thus, new paradigms for collecting training samples have been proposed to reduce human supervision with the goal of increasing learned model performance. Many research teams look at using typically abundant unlabeled data. For example, active learning methods start with a small amount of labeled data and identify the most informative unlabeled instances to be labeled by a human expert iteratively. In the active learning frameworks, the capability of assessing the uncertainty of current classifier's label is essential. A commonly used measure is classification entropy [8]. However, entropy is often unreliable due to unimportant classes being included in the estimation of label likelihood. To address the confidence estimation and incorporate the concepts from active learning, we estimate the confidence of a classification label using strangeness [3] in the context of transductive learning. Transductive learning estimates the properties of unlabeled (test) instances directly from specific labeled (training) instances.

III. Proposed Approach

In order to understand heterogeneity of cell colonies from microscopy images, we explore characterization of stem cell colonies represented by groups of image features related to textural appearance, shape signatures, and intensity. Using such image features, we classify segmented colonies into three classes of {*homogeneous, heterogeneous and dark*}. Next, we evaluate the discriminative power of different features to predict class labels. This is achieved by adaptively emphasizing more relevant features and by applying the weighted *k*-NN algorithm as explored in [4] in order to improve classification accuracy.

To compute the confidence of each classification label, we use the strangeness measure to associate a confidence value with the output of our weighted k-NN non-parametric approach. This confidence value can be used for future solicitation of new annotations to improve current performance and also as an introspective tool to identify incorrectly classified samples. We show that p-value confidence estimates can be used for assessing which test data instances may have been classified incorrectly.

Dataset: In our experiments, we consider the annotated experimental dataset of stem cell colonies accessible from the NIST web interface at https://isg.nist.gov/deepzoomweb/. The test data consist of three replicas of stem cell colonies growing in a 10 cm dish over a period of five days. The stem cell colonies are imaged by using phase contrast and green fluorescent protein (GFP) modalities with Oct4 stain used as a GFP marker. A total of 396 (18x22) or 320 (16x20) fields of view were stitched together to form a large composite image (i.e., a mosaic) consisting of hundreds of colonies at each time point. Each composite image consists of about 22,912 x 20,775 pixels with 16 bits per pixel (bpp). The images are segmented, and colonies are tracked over time. The example images of the colonies are shown in Fig. 1. Each colony represents a classification example $\boldsymbol{x}$ that is characterized by 74 dimensional feature vector extracted from each imaging modality. We analyze the colony examples formed from the GFP imaging modality and characterized by intensity, shape and texture statistics. The set of manually annotated colonies includes 68 homogeneous, 47 heterogeneous and 24 dark samples. Thus, each colony example $\boldsymbol{x}$ is associated with its label y in a training data set.

IV. K-NN

K-Nearest Neighbors algorithm (*k*-NN) is a non-parametric method used for supervised classification. An unlabeled sample is classified by assigning the label which is the most frequent among the k training examples nearest to the unlabeled sample based on a chosen distance function. The class label of a query example is assigned to be the majority class label among the retrieved test examples. In the case of a tie, the class label with the smallest distance is taken.

We first normalize all features to have a zero mean and unit variance. This is achieved by subtracting mean from each sample and normalizing it by the variance. After normalization, k-NN and adaptive k-NN classification algorithms are applied to the normalized feature dataset. The classification model is built by a percentage split of the normalized data into training and test subsets.

We performed the *k*-NN classification with all combinations of the following training percentages {50%, 60%, 70%, 80%} and k values {1, 3, 5, 7, 9}. In our baseline experiments we used the *k*-NN classifier with Euclidean distance metric. Average accuracies of the *k*-NN classification using different k values and split percentages are shown in Table 1.

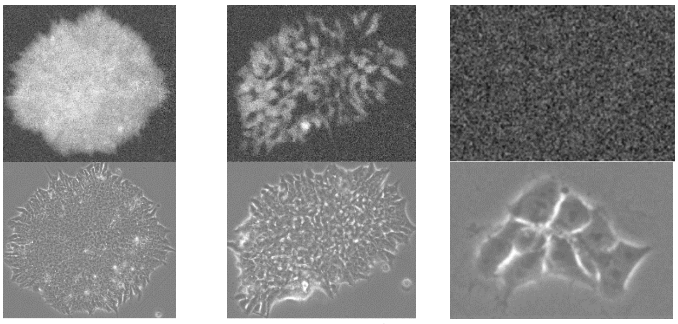

Figure 1) Top: Corrected GFP images of homogeneous, heterogeneous and dark colonies. Bottom: The corresponding phase contrast images of the colonies in the top row.

TABLE I. Accuracy of *k*-NN classifier metric for a range of k's and percentage splits of training data

Split	k=1	k=3	k=5	k=7	k=9
50%	82	77	81	76	79
60%	80	80	82	83	84
70%	81	90	87	89	89
80%	86	92	92	92	92

V. Adaptive *k*-NN

A k-NN classification algorithm assumes a locally constant class posterior probability and uses the Euclidean distance to compute the nearest neighbors around a query sample. However, this assumption is prone to a bias especially around a class boundary. Therefore, we examined a locally adaptive metric for nearest neighbor classification proposed in [1] in which posterior probabilities are adaptive to query locations. In this approach, the goal is to estimate the relevance of a feature group (or channel) i by computing its ability to predict the class posterior probabilities locally at the query sample. To do so, let us consider a query sample with the feature vector $\boldsymbol{x_0}$, and $\boldsymbol{x}$ be the nearest neighbor of $\boldsymbol{x_0}$ computed according to the Euclidean distance. After splitting the input feature vector into intensity, shape and texture groups/channels, we compute the total weighted distance between the query sample during the test phase and training sample according to [2] as follows:

$$d_w^2 = w_1 d_I^2 + w_2 d_T^2 + w_3 d_S^2 \qquad (1)$$

where $[d_I^2, d_T^2, d_S^2]^t$ are the three squared Euclidean distances between the intensity, texture and shape channels of feature vectors computed at every query sample $\boldsymbol{x_0}$ with respect to training data $\boldsymbol{x}$, and $\boldsymbol{W} = [w_1, w_2, w_3]^t \in \Re^3$ defines the weights for the distances of the three feature channels. To compute weights $\boldsymbol{W}$, first we compute the class conditional expectation of posterior $P(j|\boldsymbol{x})$ denoted by $\bar{P}(j|x_i = z_i)$, for

each class label $j = \{1,2,3\}$ corresponding to homogeneous, heterogeneous and dark class labels, given that x_i represents the i^{th} channel of $\boldsymbol{x}$ and assumes value z_i. The ability of i-th feature channel to predict $P(j|\boldsymbol{z})$ is defined as follows:

$$r_i(\boldsymbol{z}) = \sum_{j=1}^{L} \frac{(P(j|\boldsymbol{z}) - \bar{P}(j|x_i = z_i))^2}{\bar{P}(j|x_i = z_i)} \qquad (2)$$

where L is the number of classes. The smaller the difference of these two probabilities, $P(j|\boldsymbol{z})$ and $\bar{P}(j|x_i = z_i)$, the more information feature channel i carries in predicting the posterior probability $P(j|\boldsymbol{z})$ locally at $\boldsymbol{z}$. The details of estimation of $P(j|\boldsymbol{z})$ using local neighbourhood of $\boldsymbol{z}$ can be found in [1]. Given the summation over all class labels for such r_i's at the query point and averaging over all such summations in the neighborhood of the query sample, we can compute the feature relevance factor $\bar{r}_i$. The relevance factor measures how well on average the class posterior probability is approximated by a channel i in the vicinity of the query point $\boldsymbol{x_0}$

$$\bar{r}_i(\boldsymbol{x}_0) = \frac{1}{k} \sum_{z \in N_k(x_0)} r_i(\boldsymbol{z}) \qquad (3)$$

where $N_k(\boldsymbol{x_0})$ is the neighborhood of point $\boldsymbol{x_0}$. The weight calculation is repeated iteratively according to the equation below

$$w_i(\boldsymbol{x}_0) = \frac{\exp(c\, R_i(x_0))}{\sum_{l=1}^{q} \exp(c\, R_l(x_0))} \qquad (4)$$

where $R_i(\boldsymbol{x}_0) = \max_{j=1,2,3} \bar{r}_j(\boldsymbol{x}_0) - \bar{r}_i(\boldsymbol{x}_0)$ is the relevance of i-th feature channel (i.e. subset of features) with maximal relevance, $q=3$ is the number of features channels, and here $c=5$ is a parameter that can be chosen to affect the influence of $\bar{r}_i(x_0)$ on w_i . The number of all iterations is set to 5. More details about the weight computation can be found in [1]. Once the weights are computed, the nearest neighbors are retrieved using the weighted distance $\boldsymbol{d}_w$ in Eq. (1). In our case, the weights are computed for two different percentage splits of 70% and 80% of training data. The weight distributions across colonies of each feature channels, intensity, shape, and texture, for the test data are plotted in Fig. 2. We can see that the weights are smaller for the shape feature channel (around *0.2)* which implies less capacity of the feature channel to predict class probability. The texture channel has the largest weights between 0.4 and 0.6. The results of adaptive k-NN can be found on Table II for all feature channels. While the weight computation correctly corroborates our intuition about relevance of different feature channels, the absolute values of classification accuracy are comparable with Table I.

VI. Confidence Estimation using Strangeness

In this section, the goal is to quantify the confidence of the prediction using a *strangeness* measure. This measure characterizes the uncertainty of a sample (instance) with respect to its own label and provides the k-NN classifier with an introspection ability. For each example $\boldsymbol{x}_i$ in this dataset, a strangeness α_i^y with respect to a class y is computed as:

$$\alpha_i^y = \frac{\sum_{r=1}^{k} d_{ir}^{y}}{\sum_{r=1}^{k} d_{ir}^{\bar{y}}} \qquad (5)$$

where y is the predicted class label y_i for instance $\boldsymbol{x}_i$, d_{ir}^{y} is the r-th shortest distance between a point i and another point with class label y, $d_{ir}^{\bar{y}}$ is the r-th shortest distance between point i and another point with the class label other than class y, and k is the number of nearest neighbors considered. The strangeness measure is a ratio of the sum of k nearest distances from the same class to the sum of the k nearest distances from all other classes. It measures how "strange" an instance in question is with respect to its semantic category. An example closer to other class instances in comparison to its own class instances has higher strangeness and vice versa.

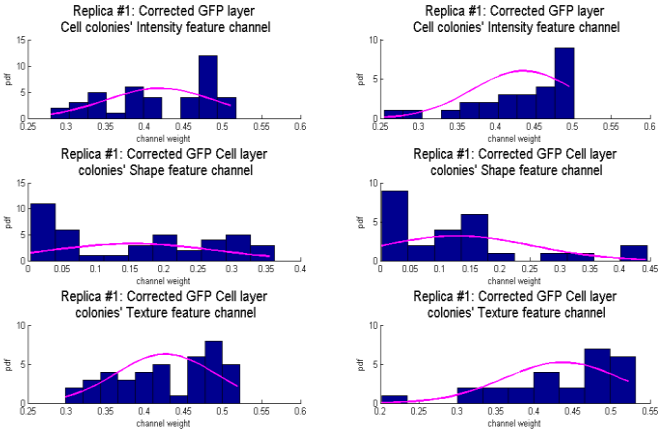

Figure 2) Distribution of learned weights that reflect the relevance of the individual feature channels of intensity (top row), shape (middle row), and texture (bottom row) for predicting class labels using weighted k-nearest neighbor algorithm trained on 70% (left) and 80% (right) of the training samples.

After computing the strangeness, we compute transductive **p-value** statistics according to [3]. The p-value is a measure of the probability of obtaining a result equal to or more extreme than what was actually observed under the null hypothesis (i.e., class label assignment). The p-value quantifies how well the data supports our classification hypothesis. The smaller the p-value the greater the evidence against the hypothesis (a sample does not belong to a class) and vice versa. Therefore, by using the strangeness values obtained for test data, we can compute a p-value measure t_r^y for all test samples r with respect to each y-labeled class as *homogeneous*, *heterogeneous* or *dark* colony according to the equation below:

$$t_r^y = \sum_{\substack{i=1 \\ y_i = y_r = y}}^{n} 1\{\alpha_{iy_i} > \alpha_{ry_r}\}/n \qquad (6)$$

where n is the number of instances in the entire training set with the label y and $1\{.\}$ is the indicator function, which is 1 when the i-th example from the training set of the same class has strangeness value greater than the one of the test point denoted by r. The p-value t_r^y can be viewed as a measure of the probability of having instances in the class with strangeness greater than or equal to that of test point r. Using Eq. 5 and Eq. 6, the strangeness and p-values are computed for all test points.

TABLE II. Average accuracies of adaptive k-NN classifier for a range of Ks and percentage splits of training data.

Split	K=1	K=3	K=5	K=7	K=9
50%	83	80	82	74	78
60%	81	79	81	82	83
70%	79	90	90	87	87
80%	86	90	96	92	92

To estimate the confidence in a label from the adaptive *k*-NN classification, we consider a split of the data into 80% training and 20% testing samples and use adaptive k-NN described in previous section, with *k* is set to 5. We compute p-values for all 26 examples in the test set with respect to *homogeneous*, *heterogeneous* and *dark* classes (20% rounded down of 68 homogeneous, 47 heterogeneous and 24 dark). We visualize colonies with high p-values which are labeled correctly. Visualization of the p-values for different examples and with respect to *homogeneous* and *heterogeneous* classes can be seen in Fig. 3. Note that in all the test examples True Positives have a transductive p-value greater than 0.5. This indicates that the p-value can be effectively used for quantifying which examples are true positives. The color coding shows that examples with high p-values are often true positives (TP), while examples with low p-values with respect to the class are mostly true negatives (TN) and a few are false positives (FP) and false negatives (FN). The colonies with p-values less than 0.2 were misclassified by adaptive *k*-NN classifier for k=5 and split 80%. Their images are shown in Fig. 4. The vertical lines in Fig. 3 separate indices of homogenous, heterogeneous and dark test examples. By association of p-value estimates with classification output, we can estimate which test samples might have been classified incorrectly.

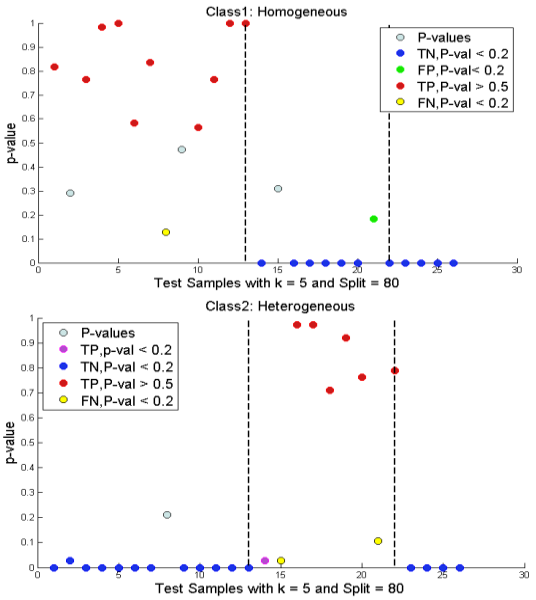

Figure 3) P-values for the subset of samples (replica #1) in the test dataset with respect to homogeneous (top) and heterogeneous (bottom) computed for k = 5 and split 80%. The color schema in the legend corresponds to combinations of p-value ranges and classification evaluations (TP=true positive, TN=true negative, FP=false positives, and FN=false negatives).

VII. Conclusion

We have demonstrated the effectiveness of k-NN and adaptive k-NN classifiers together with their per sample confidence estimates. Observing the p-value confidence estimates enabled us to assess which examples may have been classified incorrectly. Fig. 4 illustrates examples of the incorrect classification of a homogenous colony due to the inappropriate choice of the parameter *k* for the k-NN classifier. Using *k=5 f*or these examples the three most distant neighbors were from incorrect classes, while with *k=2* the examples were correctly classified. This suggests that choosing the parameter *k* adaptively per example may be more appropriate.

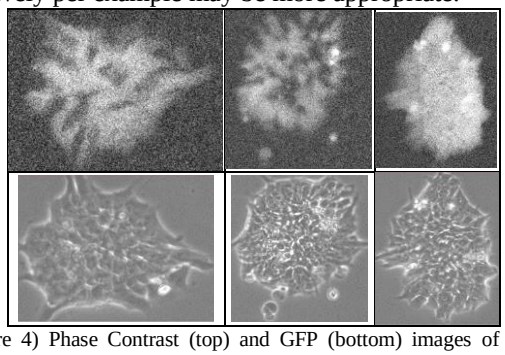

Figure 4) Phase Contrast (top) and GFP (bottom) images of three test colonies which have lower p-values than 0.2 within classes and are misclassified. Left: Colony index 15 in heterogeneous p-value plot is misclassified to homogeneous (FN with p-value < 0.2). Middle: Colony index 21 is a heterogeneous colony that is misclassified to homogeneous (FP with p-value < 0.2). Right: Colony index 8 in the homogeneous p-value plot is misclassified to heterogeneous (FN for homogenous class).

VIII. Disclaimer

Commercial products are identified in this document in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

References

[1] C. Domeniconi, J. Peng, and D. Gunopulos. Locally adaptive metric nearest-neighbor classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(9):1281-1285, 2002.

[2] G. Singh and J. Kosecka, “Nonparametric Scene Parsing with Adaptive Feature Relevance and Semantic Context”, Proceedings of CVPR, 3151-3157, 2013.

[3] K. Proedrou, I. Nouretdinov, V. Vovk, and A. Gammerman. “Transductive Confidence machines for pattern recognition”. In ECML, pages 381-390, 2002.

[4] G. Singh and J. Kosecka, “Introspective Semantic Segmentation”, 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), 714-720.

[5] A. Cardone, J. Amelot, Y. -Shian Li-Baboud, M. Brady and P. Bajcsy, “Biases from model assumptions in texture sub-cellular image segmentation”, SPIE Newsroom, 13 November, 2012.

[6] M. Halter, Y.-S. Li-Baboud, A. Peskin, P. Bajcsy, D. Hoeppner, and A. L. Plant, "Addressing Uncertainty in the Automated Evaluation of Stem Cell Colony Quality," in CYTO 2013 May 19-22, 2013, San Diego, CA..

[7] J. Gu, J. Chen, Q. Zhou, H. Zhang, Gaussian mixture model of texture for extracting residential area from high-resolution remotely sensed imagery. Updating Geo-spatial Databases Imagery/5th ISPRS Workshop. DMGISs, p. 157-162, 2004.

[8] A. Holub, P. Perona, and M. Burl. Entropy-based active learning for object recognition. In CVPR Workshop, pages 1-8, 2008.