TUTA/IOE/PCU

Journal of the Institute of Engineering
*July 2019, Vol. **15 (No. 2):** 202-209*

On the Image Pixels Classification Methods

Santosh Ghimire
Department of Applied Sciences, Institute of Engineering,
Pulchowk Campus, Tribhuvan University, Nepal
Corresponding author: *santoshghimire@ioe.edu.np*

Received: *Feb 7, 2019* **Revised:** *March 15, 2019* **Accepted:** *March 18, 2019*

Abstract: In this article, we first discuss about the images and image pixels classifications. Then we briefly discuss the importance of classification of images and finally focus on various methods of classification which can be implemented to classify image pixels.

Key words: Image, parametric test, LDA, QDA, polyclass, SVM

1. Introduction

We can define an image as a two dimensional function, say $f(x, y)$. Here x and y represent plane coordinates and the amplitude of f at (x, y) is called grey level or intensity of image a that point. When the values given by x, y and the amplitude of f are all finite, discrete quantities, the resulting image is called a digital image. More precisely, a rectangular black-and white image is a martrix of real numbers in which all the entries represent the level of grey at that point. The level of grey ranges from 0 to 255 in which 0 represents the darkest spot and 255 represents the brightest spot. The elements of digital image are usually called pixels, short for picture elements. Moreover, a color image is generated by taking the tensor product of three matrices which are the decompositions of the original image into blue, green, and red components. The pixels representing a particular feature or a color in an image show more homogenity in terms of distribution followed by the data set of pixels. Hence by comparing image pixels with each other, and to pixels of known identity, we can form different groups of similar image pixels. The image pixels groups so formed are called image classes. Image pixels classification is a process of assigning pixels to different classes of interest in the imgae. Broadly speaking, classification is a multivariate analysis task and as the name suggests, it basically deals with classifiying a new observation into one of the classes of interest.

We use random sample of locations from each of class of interest to build the recognition system. Then the random sample obtained from a class is called the training data of that class. All the classification methods assume that the image in the context depicts one or more image features on it and that each of the features is from one of exclusive and distinct classes. In general, there are two different approaches of image classification: supervised and unsupervised. The

supervised classification is based on the idea that a user can select sample pixels in an image that work as representative of classes of interest in the image and then direct the image processing software to use these choices as references for the classification of all other pixels in the image. In the unsupervised classification, as the name suggests, groupings of pixels with common characteristics are based on the software analysis of an image without user providing sample classes for the classification. Image pixels classification is very important in today's digital age as it has numerous applications. It is broadly used in remote sensing. Main areas of application of classification in remote sensing are mineral exploration, determination of surface composition and land-use analysis. Another area where the classification is widely used is medical field. Digital image classification is used in chromosome karyotyping, study of anatomical surgery, computer integrated surgery and in many other cases in medical treatment. Digital techniques are widely used in astronomical applications, face recognition, traffic control systems, agricultural imaging, computer vision etc. The pixels classifications are employed to perform segmentation of color images.

Commonly used statistical methods that can be implemented for image pixels classification are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), classification tree, polyclass method, maximum likelihood and Bayes classifier. Commonly used computer-based classifiers include nearest-neighbor classier, K-nearest-neighbor, neural networks, and support vector machine. A classification method based on hypothesis testings was developed by Liao and Akritas. This is powerful nonparametric classification method which can allow the variation within a class be taken into account through the test statistics without making distributional assumptions. However, the implementation of their method in the context of images reveals that the method can fail to correctly classify many image pixels in the given image due to small p-values. In 2012, Ghimire and Wang modified this method by introducing minimum distance into test-based classification and came up with a new classifier for image pixels. This method is called Ghimire and Wang classifier.

2. Various Methods of Classifications

In this section, we discuss various methods of classifications which can be implemented to classify image pixels. We discuss Bayesian Classification Method [3], Linear discriminant analysis [3], Quadratic discriminant analysis [3], Classification tree [1], Test based classifications [4], Support vector machine [6], Polyclass [5], Ghimire and Wang's classifier [2]. We begin with Bayesian Classification.

2.1 Bayesian Classification

Bayesian classification is a statistical method for classification which assumes an underlying probabilistic model, the Bayes theorem. Bayesian classification is named after Thomas Bayes, who proposed the Bayes theorem. Now, we describe the classification of a pattern vector by the Bayes classifier. Suppose that there are k classes of interest, given by $\omega_j, j = 1,2,3, \ldots k$ and x is a n-dimensional pattern vector. The probability that a pattern vector x belongs to a class ω_j is

given by $P(\omega_j|x)$. Using the Bayes Theorem we have $P(\omega_j \cap x) = P(x|\omega_j)\,P(\omega_j)$ where $P(x|\omega_j)$ is the probability density function of the pattern vector x in the class ω_j and $P(\omega_j)$ is the probability of occurrence of the class ω_j. The decision function for the Bayesian classification is

$$d_j(x) = P(\omega_j|x) \propto P(x|\omega_j)\,P(\omega_j)$$

Thus a pattern vector x belongs to class ω_j if $d_j(x) > d_i(x)\ for\ i = 1,2,3,\ldots.k, i \neq j$. It is often assumed that the data from a class of interest have Gaussian distribution i.e.

$$P(x|\omega_j) = \frac{1}{(2\pi)^{\frac{n}{2}}|C_j|^{\frac{1}{2}}}\ exp\left(-\frac{1}{2}\,[\,(x-m_j)^T\ C_j^{-1}(x-m_j)]\right)$$

where C_j and m_j are the covariance matrix and mean vector of class ω_j and $|C_j|$ is the determinant of C_j. As ln is monotonic function, decision function remains invariant under the ln transformation. Then the decision function becomes

$$d_j(x) \propto -\frac{1}{2}\ln|C_j| - \frac{1}{2}\left[\,(x-m_j)^T\ C_j^{-1}(x-m_j)\right] + \ln P(\omega_j) - \frac{n}{2}\ln(2\pi).$$

We note that the term $-\frac{n}{2}\ln(2\pi)$ is independent of number of classes, the decision function for the Bayesian classification is given by

$$d_j(x) = P(\omega_j|x) \propto -\frac{1}{2}\ln|C_j| - \frac{1}{2}\left[\,(x-m_j)^T\ C_j^{-1}(x-m_j)\right] + \ln P(\omega_j).$$

2.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a method in multivariate analysis and gives us the separation of different classes of objects. It follows the principle of total probability of misclassification and assume the normality distribution for data in each class. We now give a brief overview of binary classification using LDA. Let p_1 and p_2 be prior probabilities of two classes, say π_1 and π_2. We would like to assign an object Y to one the two classes. Let Y be characterized by some vector $X = [x_1,\ x_2, \ldots x_p]^T$. Now by using the Bayes's rule, the conditional probability of each class is given by:

$$P(\pi_i|X) = \frac{P(X|\pi_i)p_i}{\sum_{j=1}^{2} P(X|\pi_j)p_j}$$

where $P(\pi_i|X)$ is the posterior probability and $P(X|\pi_i)$ is called the likelihood function of π_i. The prior probabilities are assumed to be given. If they are not known, then the uniform distribution is used so that $p_1 = p_2$. We assume that the conditional distributions are multivariate normal i.e.

$$P(X|\pi_i) = \frac{1}{|\sum_i|^{\frac{1}{2}}(2\pi)^{\frac{p}{2}}}\ exp\left(-\frac{1}{2}[(X-\mu_i)^T\ \sum_i^{-1}(X-\mu_i)]\right)$$

where μ_i, $\sum_i$ are mean and covariance matrices. In this method, it is assumed that the classes have common covariance matrix i.e. $\sum_1 = \sum_2$. Thus we have after simplifications:

$$\log\left[\frac{P(\pi_1|X=x)}{P(\pi_2|X=x)}\right] = \log\left(\frac{p_1}{p_2}\right) - \frac{1}{2}\,(\mu_1+\mu_2)^T \sum^{-1}(\mu_1-\mu_2) + x^T \sum^{-1}(\mu_1-\mu_2).$$

Hence by minimizing posterior probability of misclassification, a new observation x_0 belongs to class 1 if

$$x_0{}^T \sum^{-1}(\mu_1-\mu_2) - \frac{1}{2}\,(\mu_1+\mu_2)^T \sum^{-1}(\mu_1-\mu_2) > log\left(\frac{p_1}{p_2}\right).$$

The decision boundary between classes π_1 and π_2 i.e. the set where $P(\pi_1|X=x) = P(\pi_2|X=x)$ is linear in x and is a hyperplane in p-dimension with p>1. In practice, the mean and covariance matrix of classes are known and are estimated by the training data.

2.3 Quadratic Discriminat Analysis

Quadratic discriminant analysis follows similar principle as LDA and also assumes that the distributions are normal. This method is different from the linear discriminant analysis in the sense that it allows the classes to have different covariance matrices. Because of this the decision boundary between the classes is quadratic. Then using the discussion in the LDA, we have after simplification, x_0 belongs class π_1 if

$$-\frac{1}{2}\,x_0{}^T(\sum_1^{-1}-\sum_2^{-1})x_0 + (\mu_1{}^T\sum_1^{-1}-\mu_2{}^T\sum_2^{-1})x_0 - K > log\left(\frac{p_2}{p_1}\right)$$

where

$$K = \frac{1}{2}\log\left[\frac{|\sum_1|}{|\sum_2|}\right] + \frac{1}{2}\,(\mu_1{}^T\sum_1^{-1}\mu_1 - \mu_2{}^T\sum_1^{-1}\mu_2)\,.$$

Here the surface that separates the classes is quadratic. Hence, we use the term quadratic in QDA. We estimate the class parameters mean and covariance by using the training data.

2.4 Classification Tree

Classification tree method is also known as decision tree method. This method is an observational method which is used in the classification of explanatory variable. In this method there are no prior assumptions about the data to be classified. Therefore, it is a non-parametric technique. It is simply based on the idea of partition testing. By the means of this method, the input domain of a test object is regarded under various aspects. Then for each such aspect, we form disjoint and complete classification. The stepwise partition of the input domain is represented graphically in the form of a tree. For this reason, it is called classification tree method. Tree structured classifiers are constructed by repeated splits of subsets of the feature space into two descendent subsets beginning with the feature space itself. More precisely, the decision tree is constructed by recursively partitioning the data set into purer, more homogeneous subsets depending on a set of tests applied to one or more attribute values at each node in the tree. All the algorithms developed to split the training data at each internal node of a decision tree into regions that contain examples from just one class, either minimize the impurity of the training data or maximize the goodness of split. The goodness of split is measured by an impurity function defined for each node. The possible impurity functions include entropy, the

misclassification rate, and the Gini index. The procedure of creating a tree classifier involves the following three steps: The selection of splits, the decisions when to declare a node terminal or to continue splitting it and the assignment of each terminal node to a class. The class labels are assigned to terminal nodes based on a majority vote or a weighted vote when it is assumed that certain classes are more likely than others. A tree is composed of a root node which contains all the data, a set of internal nodes (splits), and a set of terminal nodes which are called leaves. Each node in a decision tree has only one parent node and two or more descendent nodes. The data is classified by moving down the tree and sequentially subdividing it according to the decision framework defined by the tree until a leaf is reached. Decision tree classifiers divide the data into subsets, which contain only a single class.

2.5 Test-Based Classification

The test-based classification was introduced by Liao and Akritas. This test-based classification does not need any assumptions on the form of the distribution of classes. We now discuss the main idea behind this test-based classification. Liao and Akritas employ hypothesis testing in their classification method. The p values of the hypothesis tests which are essentially the values that provide evidence to reject or fail to reject the null hypothesis is the main idea behind Liao & Akritas's classification method. Suppose that there are two classes, say class 1 and class 2. Let x_0 be a test point. Suppose that class means from two classes are μ_1 and μ_2. Let us consider training vectors with observations $(x_{11}, x_{12}, \ldots . x_{1n_1})$ and $(x_{21}, x_{22}, \ldots . x_{2n_2})$ from class 1 and class 2 respectively. For the classification of test point x_0, the following two tests are conducted:

- Test 1: Place x_0 with the observations from class 1 and use $(x_{11}, x_{12}, \ldots . x_{1n_1})$ and $(x_{21}, x_{22}, \ldots . x_{2n_2})$ to test the null hypothesis $H_0 : \mu_1 = \mu_2$.
- Test 2: Place x_0 with the observations from class 2 and use $(x_{11}, x_{12}, \ldots . x_{1n_1})$ and $(x_{21}, x_{22}, \ldots . x_{2n_2})$ to test the null hypothesis $H_0 : \mu_1 = \mu_2$.

Then, the decision rule for the classification is that x_0 belongs to class 1 if PV_1 is less than PV_2 . Similarly x_0 belongs to class 2 if 1 if PV_1 is greater than PV_2. This binary classification is then extended to more than two classes cases.

2.6 Support Vector Machine

Support vector machines are simply a set of related supervised learning methods which analyze data and recognize patterns. The SVM's perform pattern recognition between two point classes with the help of a surface obtained by using certain points of training data and these points are called support vectors. The SVM's is a non-probabilistic binary linear classifier which constructs a hyperplane or a set of hyperplane for the classification. We consider both linearly separable and non-separable data. The basic idea behind the SVM classification in the linearly separable data is to choose a hyperplane which gives us the maximum separation of two groups of data. In other words, we choose the hyperplane which has the largest margin where margin is the summation of shortest distance from the separating hyperplane to the nearest data of both classes. Such a

hyperplane is called maximum-margin hyperplane. In order to address the non-linearly separable data, SVM does a mapping from the input space to a higher dimensional space where the data is linearly separable and a maximal separating hyperplane is constructed there. Now we give the basic theory of SVM, mostly taken from Vapnik (1982). Suppose that we are given a set S of points $x_i, x_i \in R^n, i = 1,2, \dots N$ and each x_i belongs to either of the two classes. We assign a label $y_i \in \{1, -1\}$.

We need to find equation of hyperplane which divides S with all the points of one class in same side and maximizing the minimum distance between either of the two classes and the hyperplane. A hyperplane can be represented by $W.X - b = 0$ where represents dot product, W is normal vector and b is the distance from the origin. When the data are linearly separable, W and b are chosen to maximize the distance between parallel hyperplane which separates the data. These hyperplanes are given by $W.X - b = 1, W.X - b = -1$ But the distance between these hyperplanes is $\frac{2}{||W||}$ where $||W|$ is norm of W. So we minimize ‖W‖. For this we need, $W.X - b \geq 1,$ for x_i to be in first class and $W.X - b \leq 1$ for $x_i$to be in second class. Thus we need to minimize ‖W‖ subjected to the condition $y_i(W.X - b) \geq 1, i = 1,2, \dots N.$ After the construction of the hyperplane, it separates the data into two distinct classes.

2.7 Polyclass

Polyclass model fits a polychotomus logistic regression model using linear splines and their tensor product. It provides estimates for conditional class probabilities which can be estimated to predict class labels. We now give an overview of Polyclass model. Suppose that Y is a qualitative random variable that takes on a finite number K + 1 of values that we refer to as classes. Depending on a vector of predictors X ∈ R. We would like to predict Y. As stated earlier, Polyclass uses piecewise linear splines and selected tensor products to model the conditional class probabilities. Precisely suppose $P(Y = k|X = x) > 0$ for $k \in K = \{1, 2, \dots K + 1\}$ and $x \in X$, where X is a subset of R^M over which X ranges. We set

$$\theta(K|x) = \log \frac{P(Y = k|X = x)}{P(Y = K + 1|X = x)}, x \in X \; and \; k \in K.$$

Then $\theta(K + 1|x) = 0$ for $x \in X$ and

$$P(Y = k|X = x) = \frac{\exp \theta(k|x)}{\exp \theta(1|x) + \cdots . + \exp \theta(K + 1|x)}, x \in X \; and \; k \in K.$$

This is referred as the polychotomous regression model. When K=1 it is known as the logistic regression model. Let J be a positive integer and G be a J dimensional linear space of functions on X with basis $B_1, B_2, \dots B_J$ Let us consider the model $\theta(k|x) = \theta(k|x; \beta k) = \sum_{j=1}^{J} \beta_{jk} \; B_j(x), x \in X \; and \; k \in K;$

where β is the JK-dimensional column vector consisting of the entries $\beta_1, \beta_2, \dots \beta_k$. Then we set

$$P(Y = k|X = x; \beta) = \frac{\exp \theta(k|x;\beta)}{\exp \theta(1|x;\beta) + \cdots . + \exp \theta(K+1|x;\beta)} \text{ for } \beta \in R^{JK}, x \in X \; and \; k \in K.$$

The maximum likelihood estimate of $\theta(k|x)$ is given by $\hat{\theta}(k|x) = \theta(k|x, \hat{\beta})$ where $\hat{\beta}$ is the maximum likelihood estimate given by $l(\hat{\beta}) = \max_{\beta} l(\beta)$. Then the Polyclass rule of classification is to assign a case with X=x to a class k having the maximum value of $\hat{\theta}(k|x)$. In Polyclass, there are K parameters for each basis function which increases the amount of computation needed for large data sets.

2.8 Ghimire and Wang's Method

We begin with the discussion of binary classification. As the name suggests, we consider two classes in the given image. Let us consider two image pixels with their means μ_1 and μ_2 and x_0 be a randomely selected test point in the image. Let the training vectors are the observations $(x_{11}, x_{12}, x_{13}, \dots, x_{1n_1})$ and $(x_{21}, x_{22}, x_{23}, \dots, x_{2n_2})$ respectively from class 1 and class 2. Then we perform following two tests where two statistical tests, namely Wilcoxon rank sum test and t-test depeding upon the distribution of image classes are used.

- **Test 1:** Place x_0 with the observations from class 1 and use $(x_0, x_{11}, x_{12}, x_{13}, \dots, x_{1n_1})$ and $(x_{21}, x_{22}, x_{23}, \dots, x_{2n_2})$ to test the null hypothesis H_0. The H_0 for the Wilcoxon rank sum test is that class 1 and class 2 have identical distribution and the H_0 for the t-test is $\mu_1 = \mu_2$.
- **Test 2:** Place x_0 with the observations from class 2 and use $(x_{11}, x_{12}, x_{13}, \dots, x_{1n_1})$ and $(x_0, x_{21}, x_{22}, x_{23}, \dots, x_{2n_2})$ to test the null hypothesis H_0 .The H_0 for the Wilcoxon rank sum test is that class 1 and class 2 have identical distribution and the H_0 for the t-test is $\mu_1 = \mu_2$.

Let us denote the p-values from the test 1 and test 2 by $PV_1(x_0)$ and $PV_2(x_0)$ respectively whereas p_1 and p_2 will be reserved to denote the prior probabilitites of classes. We note that a small by $PV_1(x_0)$ and a large $PV_2(x_0)$ suggests that putting this obervation in class 1 will maintain the difference of the classes whereas putting this observation in class 2 will blur the boundary between the two classes. Depending upon the two different scenarious of p-values, we present the detailed classification for binary classification as follows:

- If max (PV1, PV2)≥0.001 (threshold), i.e. at least one of the test p-value is larger than the thershold value, then a test point x_0 belongs to class 1 or class 2 depending on PV1(1-prior of class 1) is smaller or greater than PV2(1-prior of class 2).
- If max (PV1, PV2)<0.001 (threshold), i.e. both test p-values are smaller than the thershold value, then test point x_0 belongs to class 1 if distance of x_0 to class 1 is less than distnace of x_0 to class 2. We classify x_0 as coming from class 2 if distance of x_0 to class 2 is less than distance of x_0 to class 1.

The distance of a point x_0 to a class can take one of the traditional forms such as complete linkage, single linkage, average linkage etc. or simply, the distance between x_0 and the central tendency of class pixel values. In our experiments, we employ the distance of x_0 to the mean

pixel values of each class. Now we discuss the multiclass classification. Here we consider more than two classes in the image. We extend the ideas of binary classification discussed above to multiclass classification. Assume that there are k pixel classes in the image with means μ_1, μ_2, … , μ_k and prior probabilities p_1, p_2, …, p_k respectively. Let x_0 be a test point which we would like to classify. Here we perform the hypothesis testings as many times as the number of classes by placing the test observation in one of the classes every time. We do a series of hypothesis testing in which we test to see the sample evidence that x_0 belongs to each of the classes based on the training data. We choose Kruskall-Wallis and ANOVA as our statistical tests depending on the distribution of classes. Let $PV_1(x_0)$, $PV_2(x_0)$, … , $PV_k(x_0)$ denote the p-values of Test 1, Test 2 ,… and Test k respectively. When all the test p-values are larger than the threshold, then x_0 is classified to the class obtained by eliminating classes, one at a time and comparing $(1 - p_i) \times PV_i(x_0)$. There are other various methods of classification available in literature. Moreover, the comparison between the abovementioned methods to classify image pixels can be found in [2].

3. Conclusion

We studied the image and its classification and discussed in detail about the various classification methods which can used to classify image pixels.

References

[1] Breiman L, Friedman J, Olshen AR and Stone JC (1998), *Classification and Regression Tree*, Chapman and Hall, CRC.

[2] Ghimire S and Wang H (2012), Classification of image pixels based on minimum distance and hypothesis testing, *Computational Statistics and Data Analysis*, **56:** 2273-2283.

[3] Hastie T, Tibshirani R and Friedman J (2001), *The Elements of Statistical Learning*, Second Edition, Spring-Verlag, New York.

[4] Liao SM and Akritas M (2007), Test-Based Classification: A Linkage Between Classification and Statistical Testing, *Statistics and Probability Letters*, **77:** 1269-1281.

[5] Stone JC, Hansen HM, Kooperberg C and Truong KY (1997), Polynomial Splines and their Tensor Products in extended Linear Modeling, *Annals of Statistics*, **25:** 1371-1470.

[6] Vapnik V (1982), *The Nature of Statistical Learning Theory*, Spring-Verlag, New York.