# Probabilistic Classifier with Supervised Learning

Dong-Chul Park

***Abstract*—** A novel classifier based on stochastic characteristics of training data is proposed in this paper for a supervised learning algorithm of machine learning schemes. The proposed classifier, called Probabilistic Classifier with Supervised Learning (PCSL), utilizes stochastic characteristics of training data in its training stage and produces a class probability for each category for the classifier. For a given data for classification to the PCSL, the proposed classifier produces a probability to be classified for each class. The proposed PCSL, then, integrates the probabilities for all classes with the previously obtained classifier performance table and produces the class output for the given data with the most probable class decision. The proposed PCSL is applied to a set of satellite image data for its performance evaluation. The results show that the proposed PCSL is fairly compared with conventional classifiers in terms of classification accuracy.

***Keywords*—** data classification, probability, supervised learning, SOM, classifier, feature.

## I. Introduction

RECENT increase in the amount of data for big data analysis requires a very fast analysis and classification method that provides automated content-based categorization and retrieval of data[1][2]. Several search techniques based on annotations attach different annotations to each datum. An identifier for each datum is achieved by using the annotation for individual datum. However, the annotation process is not free of charge and this process based search method is sometimes inefficient for several reasons including the mismatch between those who apply annotations to data and those who use the annotations when the size of the database is large in big data case. However, search techniques based on features can offer a search alternative with efficiency and objectiveness to search techniques based on annotations using numerical values as features of each datum. Search techniques based on features require no preprocessing to the data, but require proper feature extraction methods that can represent the data properly.

Automatic content-based retrieval of multimedia data from a database needs two important tasks: proper feature extraction process and utilization of efficient classification algorithms. Proper feature extraction should be able to describe objects more precisely. At the same time, the extracted features should be able to discriminate different classes of data sets.

In order to design a classifier for a higher classification accuracy, the concept of probability is introduced in classifier design[3]. This concept of probability is applied to the training data and classification procedure. The supervised learning algorithm allows that the target value of each data can be used to formulate the classifier. The proposed classifier in this paper utilizes the training data in formulating the classifier with minimal computational burden for training procedure.

The rest of this paper is organized as follows: Section II and Section III provide the brief summaries of Naïve Bayed Classifier and Centroid Neural Network, respectively. An efficient method for classifier design with supervised learning is proposed with how to formulate the expertise table for the proposed Probabilistic Classifier with Supervised Learning in Section IV. Experiments on a set of satellite image data classification is performed and its evaluation results are reported in Section V. Section VI concludes this paper.

Dong-Chul Park is with Department of Electronics Engineering, Myong Ji University, Yong In, Gyeonggi-do, 449-728, South Korea (e-mail: parkd@mju.ac.kr).

## II. Naïve Bayes Classifier

Mean Shift The Naive Bayes Classifier is based on the *Bayes'* theorem of probability [4][5]. In *Bayes'* theorem, the conditional probability that an event *x* belongs to a class *k* can be calculated from the conditional probabilities of finding particular events on each class and the unconditional probability of event in each class. That is, for given data, $\mathbf{x} \in \mathrm{X}$, and C classes where X is a random variable, the conditional probability that an event $\mathbf{x}$ belongs to a class k can be calculated by the following equation:

$$\mathrm{P}(\mathrm{c}_k|\mathbf{x}) = \frac{\mathrm{P}(\mathrm{c_k})P(\mathbf{x}|c_k)}{P(\mathbf{x})} \tag{1}$$

This equation shows that calculating $\mathrm{P}(\mathrm{c_k}|\mathbf{x})$ is the pattern classification problem itself since it finds the probability that the given data $\mathbf{x}$ belongs to class k and we can decide the optimum class by choosing the class with highest probability among all possible classes, C, which can minimizes the classification error. In doing so, we need to estimate $\mathrm{P}(\mathbf{x}|\mathrm{c_k})$ and it requires an assumption that any particular value of vector $\mathbf{x}$ conditional on $\mathrm{c_k}$ is statistically independent on each dimension and can be written as follows:

$$\mathrm{P}(\mathbf{x}|\mathrm{c}_k) = \prod_{i=1}^{n} P(\mathrm{x}_i|c_k) \tag{2}$$

where $\mathbf{x}$ is a n-dimensional vector data $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n})$.

The Naive Bayes classifier is based on Eq. (2) and assumes that each feature be statistically independent [13]. This assumption results in simpler calculation cost and efficient data processing. By combining Eq.(1) and Eq. (2), the Naive Bayes classifier can be summarized as the following equation:

$$\mathrm{k} = \mathrm{argmax}_k\, P(c_k) \prod_{i=0}^{n} P(\mathrm{x}_i|c_k) \tag{3}$$

where the denominator $\mathrm{P}(\mathbf{x})$ is omitted since the value is the same for all class.

The Naive Bayes classifier is often referred as the MAP(maximum a posteriori) decision rule. Note that the assumption of statistically independence in each feature sometimes does not hold in some cases and causes problems in some practical cases. However, various applications and experimental studies show that training schemes based on MAP decision rule with the Naive Bayes assumptions yields an optimal classifier even when the assumption does not hold.

## III. Centroid Neural Network

The CNN algorithm is an unsupervised competitive learning algorithm based on the classical k-means clustering algorithm [6]. It finds the centroids of clusters at each presentation of the data vector. The CNN first introduces definitions of the winner neuron and the loser neuron. When a data $x_i$ is given to the network at the epoch (k), the winner neuron at the epoch (k) is the neuron with the minimum distance to $x_i$. The loser neuron at the epoch (k) to $x_i$ is the neuron that was the winner of $x_i$ at the epoch (k-1) but is not the winner of $x_i$ at the epoch (k). The CNN updates its weights only when the status of the output neuron for the presenting data has changed when compared to the status from the previous epoch.

When an input vector $\mathbf{x}$ is presented to the network at epoch n, the weight update equations for winner neuron j and loser neuron i in CNN can be summarized as follows:

$$\mathrm{w}_j(n+1) = w_j(n) + \frac{1}{N_j+1}[x(n) - w_j(n)]$$

$$\mathrm{w}_i(n+1) = w_i(n) - \frac{1}{N_i-1}[x(n) - w_i(n)] \qquad (4)$$

where $\mathrm{w}_j(\mathrm{n})$ and $\mathrm{w}_i(\mathrm{n})$ represent the weight vectors of the winner neuron and the loser neuron, iteration, respectively. The CNN has several advantages over conventional algorithms such as SOM or k-means algorithm when used for clustering and unsupervised competitive learning. The CNN requires neither a predetermined schedule for learning gain nor the total number of iterations for clustering. It always converges to sub-optimal solutions while conventional algorithms such as SOM may give unstable results depending on the initial learning gains and the total number of iterations. More detailed description on the CNN can be found in [6][7].

## IV. Probabilistic Classifier With Supervised Learning

When a set of total N training data with M classes are given and each datum has its own label of class, the classifier is supposed to produce the class of each datum as accurate as possible. In designing a classifier, the complexity of classifier is another consideration point in addition to the classification accuracy. Assume that a data $x_{ij}$ represent the i-th data and its class j and

$$\mathrm{X} = \{\mathrm{X}_i | 1 < \mathrm{i} < \mathrm{N}_j\} = \{\mathrm{x}_{ij} | 1 < i < N_j, 1 < j < M\} \qquad (4)$$

$$\mathrm{M} = \sum_{k=1}^{N} M_k$$

In designing PCSL, we assume that each training datum $x_{ij}$ be noisy and can be considered as a part of a probability density function with a certain set of parameters. For the training data of each class, $\mathrm{X}_i = \{\mathrm{x}_{ij} \mid 1 \le \mathrm{j} \le \mathrm{M}\}$, find the probability density function based an Gaussian probability, $\mathrm{P}_i(\mathbf{x})$.

When the probability density functions for all classes are obtained, that is,

$$\text{Classifier } \mathrm{C} = \{\mathrm{P}_j(\mathbf{x}) | 1 < \mathrm{j} < \mathrm{M}\} \qquad (5)$$

and one data, $\mathbf{y}$, is given to the classifier to classify its class.

Then, the classifier C calculates the probability that the data $\mathbf{y}$ belongs to each class as follows:

$$\mathrm{P}(\mathbf{y}) = \{\mathrm{P}_j(\mathbf{y}) | 1 \le \mathrm{j} \le \mathrm{M}\} \qquad (6)$$

In order to satisfy the axiom of probability, the normalized version of probability, $\mathrm{NP}(\mathbf{y})$, is calculated as follows:

$$\mathrm{NP}(\mathbf{y}) = \left\{\frac{\mathrm{P}_j(\mathbf{y})}{\sum_{i=1}^{M} P_j(\mathbf{y})} | 1 \le \mathrm{j} \le \mathrm{M}\right\} \qquad (7)$$

In order to utilize the classification tendency of the classifier with previous classification history, it is natural to construct the classification tendency by the formation of conventional expertise table. The expertise table consists of the probabilities of classification accuracies for all classes for a given data as shown in the following:

$$Q_k = \begin{bmatrix} q_{11} & q_{12} & \cdots & q_{1N} \\ q_{21} & q_{22} & \cdots & q_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ q_{N1} & q_{N2} & \cdots & q_{NN} \end{bmatrix} \qquad (8)$$

$$\sum_{j=1}^{N} P_{ij} = 1$$

where $q_{ij}$ represents the probability that the classifier classifies the data of Class i with Class j. $q_{ij}$ can be obtained through the training stage with the given training data and can be updated with each classification result.

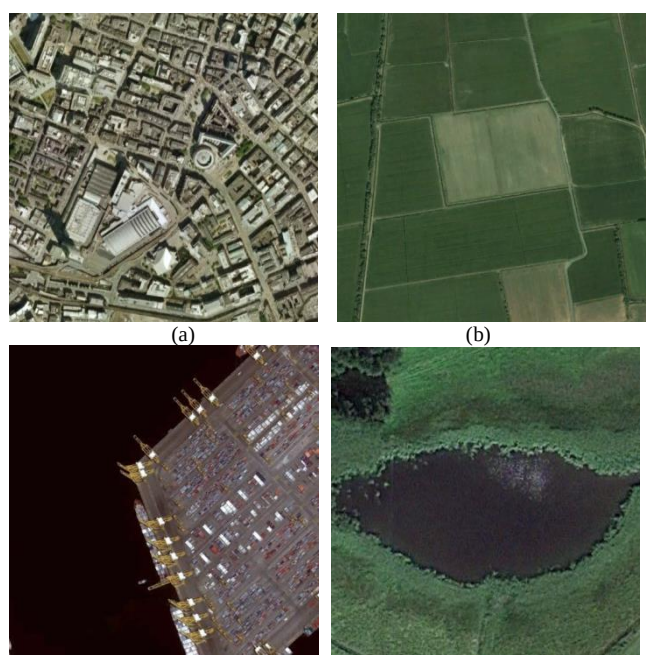

Fig. 1 Example of satellite image data set: (a) urban, (b) farm area, (c) port area, and (d) pond area.

With the probability that the data $\mathbf{y}$ belongs to each class, $\mathrm{NP}(\mathbf{y})$, is obtained, the final decision can be obtained with the expertise table. For example, with $\mathrm{NP}(\mathbf{y})$ given, the possibility, $\mathrm{P}(\mathrm{i}|\mathbf{y})$, that the datum $\mathbf{y}$ belongs to the class i can be calculated as follows:

$$\mathrm{P}(\mathrm{i}|\mathbf{y}) = \sum_{i=1}^{M} P_j(\boldsymbol{y}) q_{ij} \tag{9}$$

Therefore, the final decision of the class, $\mathrm{Class}(\mathbf{y})$, to where the datum $\mathbf{y}$ belongs can be obtained by the following calculation:

$$\mathrm{Class}(\mathbf{y}) = max_{i=1}^{M} P(i|\boldsymbol{y} = max_{i=1}^{M} \sum_{i=1}^{M} P_j(\boldsymbol{y}) q_{ij}) \tag{10}$$

Note that the above calculation can be performed without excessive computational efforts unlike other conventional classifiers . Similar advantages can be expected through Naïve Bayes classifier.

TABLE I
CLASSIFICATION ACCURACIES FOR DIFFERENT CASES

| Data | HSV feature | | | WPT feature | | |
|---|---|---|---|---|---|---|
| | MLPNN | CNN | PCSL | MLPNN | CNN | PCSL |
| *Industry* | 0.827 | 0.816 | 0.862 | 0.754 | 0.742 | 0.784 |
| *Mountain* | 0.775 | 0.782 | 0.820 | 0.688 | 0.716 | 0.775 |
| *Port* | 0.708 | 0.716 | 0.764 | 0.646 | 0.668 | 0.684 |
| *Pond* | 0.714 | 0.695 | 0.692 | 0.718 | 0.702 | 0.720 |
| Average | 0.756 | 0.752 | 0.784 | 0.702 | 0.707 | 0.741 |

## V. EXPERIMENTS AND RESULTS

For the evaluation of the proposed Probabilistic Classifier with Supervised Learning (PCSL), a satellite image data classification problem [8] was used for experiments. The satellite image data set consisted of different image classes in which each class contained different areas. Fig. 1 shows examples of an industrial area, a mountain area, a port area and a pond area. Each class consists of 100 images with different views resulting in a total of 400 images in the data set. From this data set, 80 images in each class were randomly chosen for training classifiers while the remaining images were used for evaluating classifiers. In order to obtain the expertise table, the training data is again divided by 2 parts: 90% of pure training and 10% of evaluating and obtaining the expertise table. (This expertise table is obtained by experimenting 10 times with randomly selected 90% pure training data and the 10% evaluation data.) This training data and test data combination was collected 10 times. That is, there were 10 different combinations of randomly chosen training data and test data sets. The accuracy of different algorithms used in experiments is reported by mean and variance from these 10 data sets.

Through experiments, the proposed Probabilistic Classifier with Supervised Learning was compared with Centroid Neural Network based classifier in terms of classification accuracy. In order to describe the texture information of images, the following image representation methods were used:

1) Hue-Saturation-Value (HSV): The HSV feature is a simple transformation of the RGB (Red-Green-Blue) feature and HSV color histogram is has shown good results in practice especially for image indexing and retrieval tasks, where feature extraction has to be as simple and as fast as possible. [9][10].

2) Wavelet Packet Transform (WPT): The wavelet transform provides a precise and unifying framework for the analysis and characterization of a signal at different scales [11]. In experiments, 6 step 2-D wavelet packet transform is used. The WPT produced 68 dimensional feature vector for each image.

The dimensions of each feature vector obtained from HSV and WPT are 200 and 68, respectively. When training a basic supervised classifier such as multi-layer perceptron type neural networks, the above high dimensional data have problems in its convergence. In order to solve this problem, SOFM (Self-Organizing Feature Map)[12] is adopted for reducing data dimension to 3 dimension for each of two feature vectors.

TABLE II
EXAMPLE OF EXPERTISE TABLE FOR HSV

| | *Urban* | *Farm* | *Port* | *Pond* |
|---|---|---|---|---|
| *Urban* | 0.823 | 0.022 | 0.130 | 0.025 |
| *Farm* | 0.048 | 0.790 | 0.013 | 0.155 |
| *Port* | 0.166 | 0.012 | 0.724 | 0.098 |
| *Pond* | 0.120 | 0.164 | 0.018 | 0.698 |
| Overall | 0.759 | | | |

TABLE III
EXAMPLE OF EXPERTISE TABLE FOR WPT

| | *Urban* | *Farm* | *Port* | *Pond* |
|---|---|---|---|---|
| *Urban* | 0.761 | 0.055 | 0.152 | 0.032 |
| *Farm* | 0.068 | 0.733 | 0.123 | 0.076 |
| *Port* | 0.286 | 0.036 | 0.660 | 0.018 |
| *Pond* | 0.036 | 0.162 | 0.148 | 0.654 |
| Overall | 0.702 | | | |

Table I summarizes the average classification accuracies on different classifiers that utilize each feature vector, HSV and WPT. The results shown in Table I imply that the proposed PCSL scheme is fairly compared with the classifiers based on conventional artificial neural networks with a supervised architecture (multi-layered perceptron type neural network (MLPNN) trained with Error-Back-Propagation algorithm) and Centroid Neural Network as a unsupervised architecture. The CNN algorithm is used with 8 code vectors. The proposed PCSL outperforms MLPNN and CNN by 3.6% and 4.3% in terms of classification accuracy for HSV feature case, respectively. Similar performance is shown fro WPT feature case. Since the proposed PCSL utilizes the classifier's history of classification tendency in terms of expertise table, it is somewhat natural to show higher classification accuracies over the other two classifier schemes and these results confirm the effectiveness of the proposed Probabilistic Classifier with Supervised Learning for satellite image classification tasks.

Table II and Table III show the examples of expertise tables for the proposed Probabilistic Classifier with Supervised Learning obtained in these satellite image classification experiments for HSV feature case and WPT feature case, respectively.

## VI. CONCLUSIONS

In this paper, a novel classifier based on stochastic characteristics of training data is proposed for reducing the computational loads involved in classifier design with a supervised learning algorithm of machine learning schemes. The proposed Probabilistic Classifier with Supervised Learning

utilizes stochastic characteristics of training data in its training stage and produces a class probability for each category for the classifier. For a given data for classification to the proposed Probabilistic Classifier with Supervised Learning (PCSL), the proposed classifier produces a probability to be classified for each class. The proposed PCSL, then, integrates the probabilities for all classes with the previously obtained classifier performance table and produces the class output for the given data with the most probable class decision. The proposed PCSL is applied to a set of satellite image data for its performance evaluation. The results show that the proposed PCSL is fairly compared with conventional classifiers in terms of classification accuracy. When applied to sets of satellite image data classification problem, the proposed method shows promising results in terms of classification accuracy with reduced training time. With further experiments on larger training data sets, the training speed can be measured and its advantage will be witnessed further

## ACKNOWLEDGMENT

This work was supported by the IT R&D program of The MKE/KEIT (10040191, The development of Automotive Synchronous Ethernet combined IVN/OVN and Safety control system for 1Gbps class).

## REFERENCES

[1] M. Lew et al., “Content-based Multimedia Information Retrieval: State of the Art and Challenge”, ACM Transactions on Multimedia Computing, Communications, and Applications, 2006, pp. 1–19.
[2] D. Cardoso et al., “Iterative Technique for Content-Based Image Retrieval using Multiple SVM Ensembles”, J. Clerk Maxwell, A Treatise on Electricity and Magnetism,, vol.2, pp.68-73.
[3] J. Tou and R. Gonzalez, Pattern Recognition Principles, Addison-Wesley Pub. Co., 1974
[4] I. Rish, “An empirical study of the naive Bayes classifier”, in IJCAI Workshop on Empirical Methods in AI, 2011.
[5] H. Zhang, “The Optimality of Naive Bayes”, in Proc. FLAIRS, 2004.
[6] D.-C. Park, “Centroid Neural Network for Unsupervised Competitive Learning”, IEEE Trans on Neural Network, vol.11, no.2, 2000, pp. 520-548.
[7] D.-C.Park, O.H. Kwon, and J. Chung, “ Centroid neural network with a divergence measure for GPDF data clustering”, IEEE Trans on Neural Network, vol. 19, no. 6, 2008, pp.948-957.
[8] T.-H. Kim, D.-C. Park, and D.-M. Woo, “Satellite Image Classification Using a Fuzzy c-Means Algorithm with Mercer Kernel”, in Procedia Information Technology and Computer Science, 2012
[9] C.L. Novak and S.A. Shafer, “Anatomy of a color histogram”, in Proc. IEEE Computer Society Conf, 1992, pp.299-605.
[10] J. Ma, “Content-Based Image Retrieval with HSV Color Space and Texture Features”, in Proc. Int. Conf. on Web Information Systems and Mining, 2009, pp.61-63.
[11] Z.Z. Wang and J.H. Yong, “Texture analysis and classification with linear regression model based on wavelet transform”, IEEE Trans. on Image Processing, vol.17, no.8, pp.1421-1430.
[12] T. Kohonen, “The self-organizing map”, Proceedings of IEEE, vol.78, no.9, 1990, pp.1464-1480.

**Dong-Chul Park** (M’90-SM’99) received the B.S. degree in electronics engineering from Sogang University, Seoul, Korea, in 1980, the M.S. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, Seoul, Korea, in 1982, and the Ph.D. degree in electrical engineering, with a dissertation on system identifications using artificial neural networks, from the University of Washington (UW), Seattle, in 1990. From 1990 to 1994, he was with the Department of Electrical and Computer Engineering, Florida International University, The State University of Florida, Miami. Since 1994, he has been with the Department of Electronics Engineering, MyongJi University, Korea, where he is a Professor. From 2000 to 2001, he was a Visiting Professor at UW. He is a pioneer in the area of electrical load forecasting using artificial neural networks. He has published more than 130 papers, including 40 archival journals in the area of neural network algorithms and their applications to various engineering problems including financial engineering, image compression, speech recognition, time-series prediction, and pattern recognition. Dr. Park was a member of the Editorial Board for the IEEE TRANSACTIONS ON NEURAL NETWORKS from 2000 to 2002.