Joint Conference of 1996 International Computer Symposium
December 19~21, Kaohsiung, Taiwan, R.O.C.

Recognition of Handwritten Chinese Postal Addresses Using A Dual-Expert Classification Scheme

Yih-Ming Su[1,2] and Jhing-Fa Wang[2]

[1]Department of Electronic Engineering, Kaohsiung Polytechnic Institute,
Kaohsiung County, Taiwan

[2]Institute of Information Engineering National Cheng Kung University,
Tainan, Taiwan

Abstract

A mail sorting system based on the development of OCR system is proposed in the paper for the automatic mail processing. The system to recognize the handwritten Chinese postal addresses on standard envelopes is constructed by using dual-expert classification scheme. Each architecture of dual-expert classification scheme includes feature extraction stage, recognition stage, and semantic processing stage. The black local density feature (BLDF) and multi-layer percetron (MLP) are used in the first expert classification for handwritten Chinese characters (HCCs). The white local density feature (WLDF) and Bayesian network are used in the other expert classification for HCCs. Since the names of city on addresses contain contextual information, the performance of recognition can be increased by the semantic processing. Finally, the comparison of results from the two expert classifications is used to reduce error rate of address recognition. By using dual-expert classification, we could improve the performance of the system with 70.2% correction rate and 1.3% error rate in our experimental results.

1. Introduction

A new mail sorting system with speedy, robust, and automatic characteristics, as shown in Fig. 1, is constructed as a testbed for this study. The system includes a sorting machine and an address recognition strategy. The sorting machine contains four mechanisms, namely, envelope extraction, envelope delivery, envelope assignment, and envelope box. The mail sorting system takes a handwritten postal address image from a standard envelope, as shown in Fig. 2, and determines a unique mail delivery address. The system requires the development of an OCR system for handwritten/machine-printed digital ZIP codes and word addresses. Moreover, the recognition of word addresses is more important than that of digital ZIP codes because the digital ZIP codes are usually neglected by personal custom. In this system, the recognition of word addresses is described in the following section.

The studies of the automatic mail sorting system in Japan[1], US[2], and UK[3] have been developed during the past decades. Although the study of the automatic mail sorting system has just been developed in Taiwan, the OCR system was developed with skillful and fruitful researches. The work on the optical recognition of handwritten Chinese characters that has been reported since 1980 was reviewed by Hilderbrandt[4]. Cheng[5] also described typical research on Chinese optical character recognition in Taiwan. The architecture of the dual-expert classification as shown in Fig. 3, includes character segmentation, two kinds of feature extraction, two classifiers, semantic processing and verification. The paper is organized as follows. Section 2 describes the segmentation and feature extraction of each HCC. Section 3 gives the structural description of the four-layer percetron and the Bayesian network. Section 4 includes the semantic and verification processing. The experimental results are discussed in Section 5. Section 6 gives some concluding remarks and discussion.

2. Segmentation and feature extraction

An envelope image is grabbed by a camera with 512*480 resolution and interlace scan. The preprocessing for shift correction, skew correction, and thresholding is needs in our methods. The shift and skew of the image are produced from the mechanism problems. For vertical writing form in a standard envelope, as shown in Fig. 2, a specific extent of the original image is processed and is partitioned into one or two character lines by performing the vertical projection. It is able to find obviously that there is a peak value in a specific location of the histogram and this characteristic represents a vertical line of square boundary line on a standard envelope. The line is referred to segment a right neighboring character line. The vertical character line is partitioned from the original image by scanning the vertical and horizontal projections. The respective characters are segmented from the character line by projection processing. Although these characters are segmented from a character address, these characters are coarse because the strokes of the HCC have many redundant and extended parts. The coarse characters have to be again processed by eliminating extended strokes according to judgment of the width of the strokes.

For a selected feature, we hope it can be a unique representation in each HCC. However, each HCC exists a wide variation for different writers. It is hard to find an optimal feature that only represents the variation among the different categories. Fortunately, in our special application, the processing of the feature extraction is simple to speed up the HCC recognition because the number of categories is not many. In addition, the processing for size normalization and thinning of all HCCs is not needed in our method before feature extraction is performed. To reserve the structural information between the strokes within an HCC, we first partition the HCC into 64 grid frames (i.e., 8*8 blocks) by nonlinear division as shown in Fig. 4. The method of the nonlinear division is used by the mentioned projection to build a horizontal histogram. Then, the distribution of the histogram is partitioned into 8 different intervals such that each interval consists of an equal number of black pixels. The 7 horizontal lines that cut the horizontal histogram constitute the horizontal boundaries for the grid frame. In similar way, the vertical direction of character image is also partitioned into 8 different intervals by 7 vertical boundary lines. The density of black pixels of each grid frame is regarded as an element of feature vector. The black local density feature (BLDF) vector is defined to be

$$BLDF(i) = BD(i)/A(i), \qquad for\ i=1,2,...,64,$$

where *BLDF* denotes the first feature vector, *BD(i)* is the number of black pixels in the *i*th grid frame, and *A(i)* is the area in the *i*th grid frame of the HCC. The other feature vector that is the white local density feature (WLDF) vector is defined to be

$$WLDF(i) = WD(i)/A(i), \qquad for\ i=1,2,...,64,$$

where *WLDF* denotes the second feature vector, *WD(i)* is the number of white pixels in the *i*th grid frame, and *A(i)* is the area in the *i*th grid frame of the HCC.

3. Description of the two classifiers

In past few years, the multi-layer percetron (MLP) has been successfully and widely applied in the field of OCR as a pattern classifier. The objective of training the network is to adjust the weights so that input feature vector can produce the desired output through the network. Therefore, in this experiment, the four-layer perceptron is selected as the first classifier. The network is fully connected between adjacent layers. The number of nodes of input layer corresponds directly to the dimensions of feature vector. The number of nodes of output layer corresponds directly to the number of categories of HCCs. The back-propagation training algorithm[6] is briefly described as follows.

(1) Initialization: Set all weights and node biases to small random numbers.

(2) Forward computation: Let $X = \{x_1, x_2, ..., x_n\}$ be the n dimensional feature vector. Then the net internal activity level $v_j^l(t)$ for node j in layer l is

$$v_j^l(t) = \sum_{i=0}^{n} w_{ji}^l(t) a_i^{l-1} - \theta_j^l$$

where $a_i^{l-1}(t)$ is the function signal of node i in the previous layer l-1 at iteration t, $w_{ji}^l(t)$ is the weight of node j in layer l that is fed from node i in layer l-1, and θ_j^l is the bias of node j in layer l. Assuming the use of a logistic function for sigmoidal nonlinearity, the output signal of node j in layer l is

$$a_j^l(t) = \frac{1}{1 + \exp(-v_j^l(t))}.$$

If node j is in the first hidden layer (i.e., l=1), set $a_j^0(t) = x_j(t)$ where x_j is the jth element of feature vector X. If node j is in the output layer (i.e., l=L), set $o_j(t) = a_j^L(t)$. Then compute the error signal

$$e_j(t) = d_j(t) - o_j(t), \quad \text{for node } j \text{ in output layer,}$$

where $d_j(t)$ is the jth element of the desired output.
(3) Backward computation: The purpose of the processing is used to adjust synaptic weights between the adjacent layer according to the following equation:

$$w_{ji}^l(t+1) = w_{ji}^l(t) + \alpha[w_{ji}^l - w_{ji}^l(t-1)] + \eta\delta_j^l(t)a_i^{l-1}(t),$$

where η is a learning rate, α is a momentum constant, and δ_j is a local gradient of the network. The δ_j can be described as follows:

$$\delta_j^L(t) = e_j^L(t)o_j(t)[1 - o_j(t)] \quad \text{for node } j \text{ in output layer,}$$

$$\delta_j^l(t) = a_j^l(t)[1 - a_j^l(t)]\sum_k \delta_k^{l+1}(t)w_{kj}^{l+1}(t) \quad \text{for node } j \text{ in hidden layer } l.$$

(4) Iteration: The forward and backward is recurrently performed until the error computed over the entire training set is at a minimum.

The other classifier being a Bayesian network is described as following section. Let $X = (x_1, ..., x_n)$ be the input feature vector of an HCC, which belongs to one of M categories $W_i, i = 1, ..., M$. Consider the decision problem consisting of determining whether X belongs to W_i. Let $P(X|W_i)$ be the conditional probability of X given category W_i. Let $d_i(X)$ be called Bayes decision rule[7] and be defined by

$$d_i(X) = P(X|W_i) * P(W_i), \quad i=1,...,M,$$

where an unknown pattern X is assigned to category W_i if for that pattern $d_i(X) > d_j(X)$ for all $j \neq i$. The structure of Bayesian network, as shown in Fig. 5, has three layers: input layer, Gaussian layer, mixture layer. The function and connection of nodes are described: (1) input layer: the input vector of the layer is a feature vector. Therefore, the number of nodes in the layer is equal to the number of the dimension of the feature vector; (2) Gaussian layer: the links between input layer and this layer are full connections. Let weights of all links be equal to 1. The number of nodes in the layer is dependent on the number of subclasses for each category. The output of each node is performed by the processing of mapping distance by means of the mean, variance, and probability parameters of each subclass. The three parameters are created by training phase; (3) mixture layer: the number of nodes in the layer is just only one and the output of the node is performed by selecting a minimal mapping distance among the outputs of nodes in the Gaussian layer. In training phase, the mean, variance, and probability parameters of each subclass in all categories were obtained by the processing of maxmin-distance and K-means algorithm [7]. The maxmin-distance algorithm is used to determine the number of subclasses for each category and the number of subclasses is used as an initial parameter for performing K-means algorithm. The K-means algorithm is used to determine the mean, variance, and probability parameters of each subclass. In recognition phase, let $X = (x_1, ..., x_n)$ be an n-dimensional unknown feature vector. The normal density function of subclass W_i is defined to be

$$P(X|W_i) = \frac{1}{(2\pi)^{k/2}(\Pi_{k=1}^i \sigma_k^2)^{0.5}} * \exp[\frac{-1}{2}\sum_{k=1}^{i}\frac{(X-\mu_k)^2}{\sigma_k^2}],$$

where each density is completed specified by its mean μ_k and variance σ_k of each subclass. The decision function for subclass W_i may be chosen as $d_i(X) = P(X|W_i) * P(W_i)$. Because of the exponential form of the normal density function, however, it is more convenient to work with the natural logarithm of the decision function. The mapping distance defined to be $D_i(X) = \ln P(X|W_i) + \ln P(W_i)$ represents the different degree among subclasses corresponding to the input feature vector X . Finally, a Bayesian network is constructed to correspond to a category. The input feature vector is send to the input layer of all Bayesian network and the outputs of all Bayesian networks are regarded as different degrees corresponding to input feature vector. The input feature vector belongs to which categories are according to minimum of these different degrees from all nets.

4. Use of semantic information and verification

The above methods can be extended again to incorporate a semantic information. Once all relevant HCCs on addresses have passed through the recognition stage, semantic information is used to further improve the performance of address recognition .

Each classifier performs to recognize the relevant HCCs and finds 5 possible candidate characters for each HCC by selecting the first 5 higher outputs of recognition. The processing of matching pair is used to find whether the relevant characters have contextual relation through the searching of a look-up table. The table was built by the address information in advance. If the two or more pairs are found in the processing of matching pair, the best one of all pairs is determined by selecting minimum summation. The summation is defined to be

$$Summation(j) = sum(\frac{output_j^1}{sumoutput^1}, \frac{output_j^2}{sumoutput^2}), \quad for\ j = 1, ..., M,$$

where *Summation(j)* is the sum of normal outputs of relevant HCCs in the *j*th pair, $output_j^1$ is the output of the first character in the *j*th pair, $sumoutput^1$ is the sum of the outputs of all candidate characters in the first character, and the $output_j^2$ and $sumoutput^2$ denoted the second character are similar to the description of $output_j^1$ and $sumoutput^1$. In order to reduce error rate of address recognition, the two results from performing character recognition and semantic processing after are verified by comparison operation. If the two results are not equal to each other, the address recognition is regarded as rejection of erroneous recognition. The processing can promote the accuracy of the mail sorting system to avoid delivering error addresses.

5. Experimental results

In order to investigate the performance of the mail sorting system based on scheme of the dual-expert classification, a database set of handwritten Chinese characters has been used as a source of training and test pattern. The database set includes 36 categories each of which has 80 patterns of HCCs being 50 training data and 30 test data. Due to 36 categories consisting of all characters from the names of city, the 36 categories can be partitioned into two groups and each group is composed of characters according to present sequence of the names of city. Thus, the segmented HCCs on addresses pass through the recognition stage in turn. The architecture of four-layer percetron is constructed with 64 nodes in input layer, 100 nodes in the first hidden layer, 20 nodes in the second hidden layer, and 18 nodes in the output layer. The parameters of the net include a learning-rate parameter $\eta = 0.01$, a momentum parameter $\alpha = 0.9$, and rejection parameter about determining the performance of recognition in the classifier. The architecture of the other classifier, Bayesian network, is constructed with 64 nodes in input layer, the number of nodes in Gaussian layer being dependent on the number of subclasses in each category, and one node in mixture layer. The parameters of the net conclude separation parameter about maximum distance among subclasses and rejection parameter about determining the performance of recognition in the classifier. Recognition performances of respective classification in addresses are shown in Table 1, 2 according to different rejection parameter. The experimental results give the demonstration that the rejection parameter may affect the error rate and correction rate. The higher rejection is, the lower error rate is. The performance of address recognition using dual-expert classification scheme is shown in the table 3. A dual-expert classification has prominent effects in reducing error rate of recognition and keeping correction rate of recognition.

6. Conclusions and discussions

The search described in this paper has been aimed at developing dual-expert classification for automatic mail processing. The scheme of dual-expert classification is focused on reducing the error rate of address recognition. Besides, the appropriate combination of two classifiers may produce highly reliable performance. In addition, these simple and efficient approaches in character segmentation and feature extraction are used to reduce computing time.

Our future research efforts will emphasize on promoting the correction rate of address recognition and further recognizing the other portion of handwritten addresses such as street and section names. In addition, the issues of the character segmentation and feature extraction have to been enhanced such that the mail sorting system has better performance.

References

[1] Y. Tokunaga, "History and current state of postal mechanization in Japan," Pattern Recognition Letters, vol. 14, no. 4, pp. 277-280, April 1993.

[2] Edward Cohen, Jonathan J. Hull and Sargur N. SriHari, "Understanding handwritten text in a structured environment: determining ZIP codes from addresses," Int. J. of Pattern Recognition and

Artificial Intelligence, vol. 5, no. 1&2, pp. 221-264, 1991.

[3] A. C. Downton, R.W.S. Tregidgo and C. G. Leedham, "Recognition of handwritten British postal addresses," Proc. Int. Workshop on Frontiers in Handwriting Recognition, France, Sept. 1991.

[4] Thmomas H. Hildebrandt and Wenta Liu, "Optical recognition of handwritten Chines characters advance since 1980," Pattern Recognition, vol. 26, no. 2, pp. 205-225, 1993.

[5] F. H. Cheng and W. H. Hsu, "Research on Chinese OCR in Taiwan",Int. J. of Pattern Recognition and Artificial Intelligence, vol. 5, no. 1&2, pp. 139-164, 1991.

[6] Simon Haykin, "Neural networks- A Comprehensive Foundation", Macmillan College Publishing Company, 1994.

[7] J. T. Ton and R. C. Gonzalez, " Pattern Recognition Principles," Addison-Wesley Publishing Co.

Table 1. The performance of Bayesian network

Rejection Parameter	Correction Rate	Error Rate	Rejection Rate
0.3	75.9%	15.2%	8.9%
0.4	70.1%	10.2%	19.7%
0.5	76.9%	8.3%	22.8%

Table 2. The performance of four-layer percetron

Rejection Parameter	Correction Rate	Error Rate	Rejection Rate
0.6	78.8%	10.1%	11.1%
0.8	77.38%	8.51%	14.11%
1	67.1%	6.76%	26.14%

Table 3. The performance of the dual-expert classification

Correction Rate	Error Rate	Rejection Rate
70.2%	1.3%	28.5%

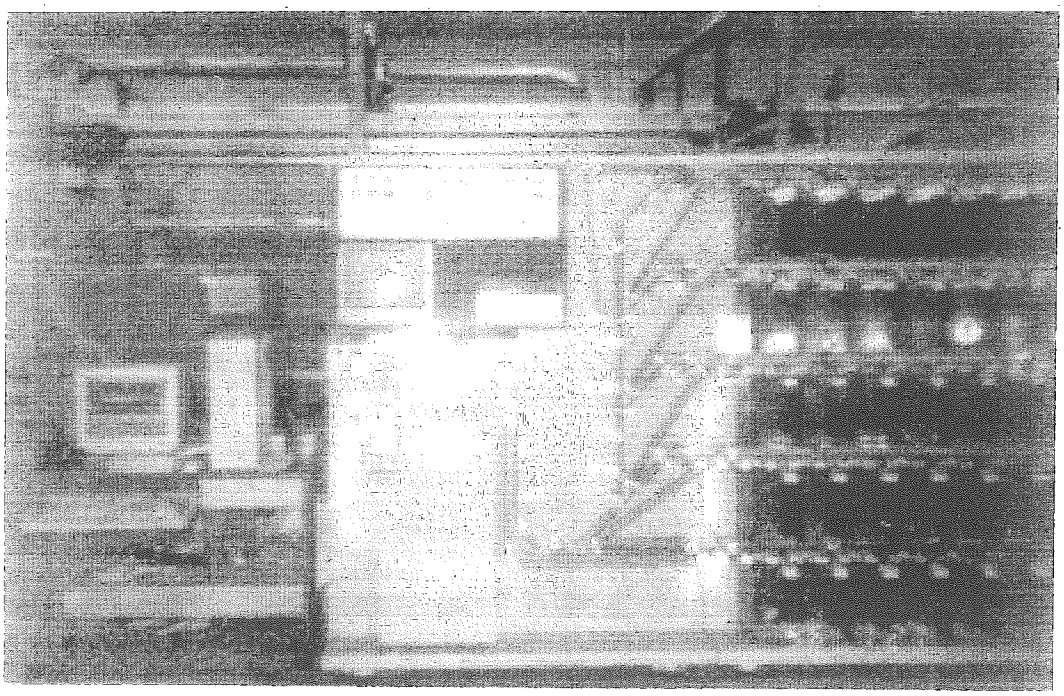

Fig. 1. The new mail sorting system used in this study.

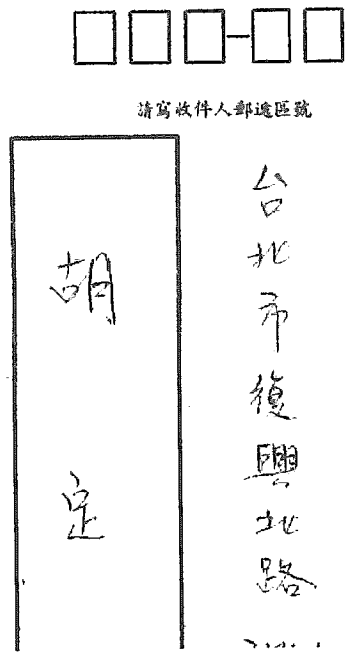

Fig. 2. An example of a standard envelope.

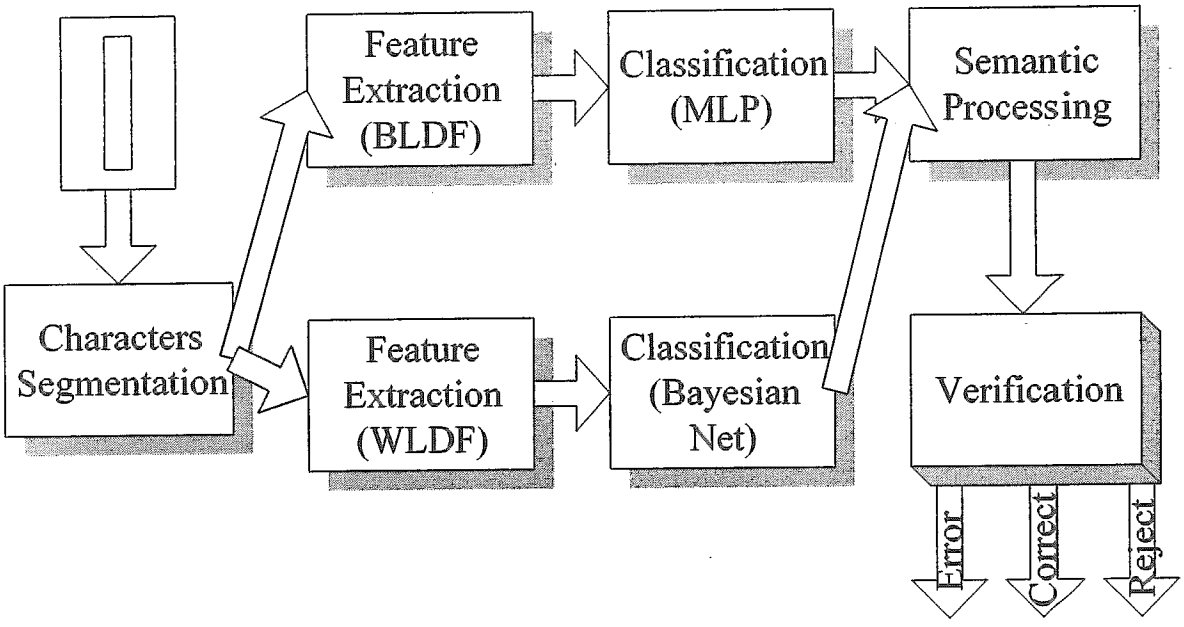

Fig. 3. The architecture of the dual-expert classification.

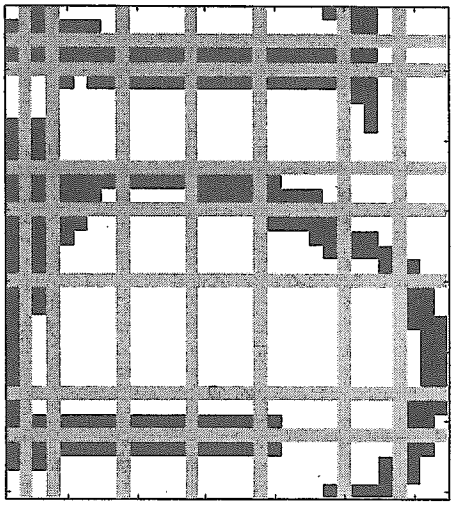

Fig. 4. An 8*8 nonlinear grid frame

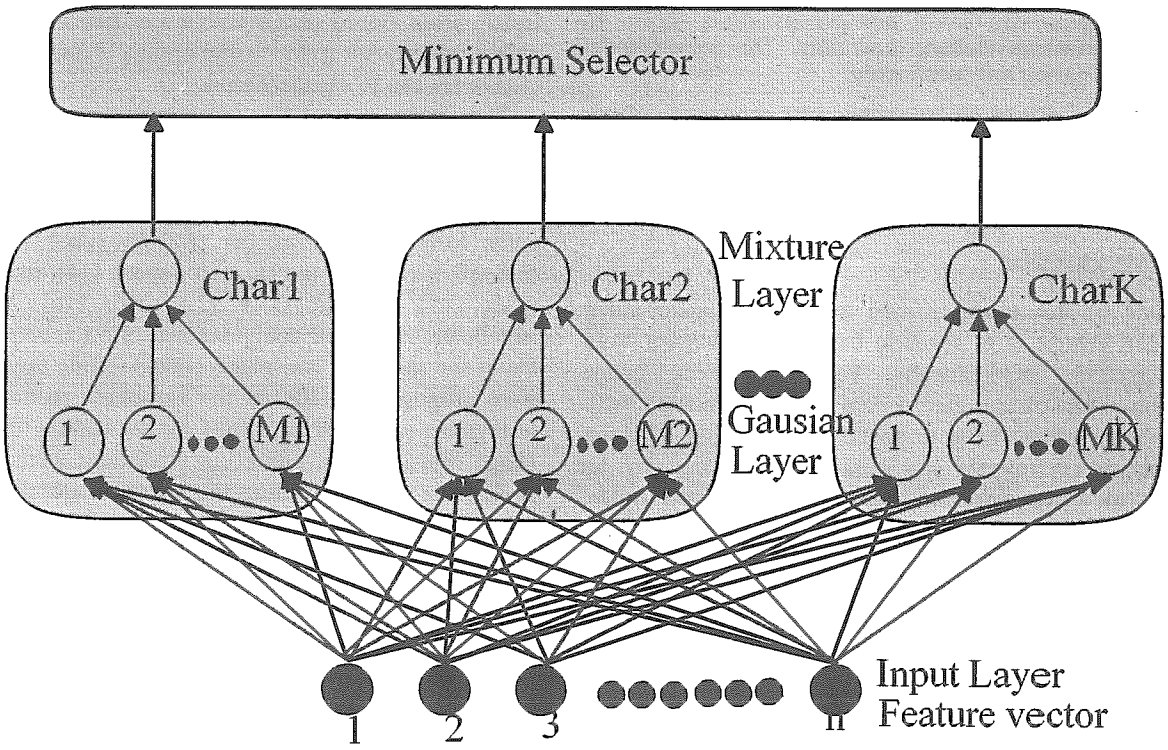

Fig. 5. The structure of Bayesian network.