The Empirical Study of the Naive Bayes Classifier in the Case of Markov Chain Recognition Task

Andrzej Zolnierek and Bartlomiej Rubacha

Chair of Systems and Computer Networks, Wroclaw University of Technology,
Wybrzeze Wyspianskiego 27, 50-370 Wroclaw, Poland
andrzej.zolnierek@pwr.wroc.pl

Summary. In this paper the problems of sequential pattern recognition are considered. As a statistical model of dependence, in the sequences of patterns, the first-order Markov chain is assumed. Additionally, the assumption about independence between the attributes in the feature vector is made. The pattern recognition algorithms with such assumption are called in the literature "naive Bayes algorithm". In this paper such approach is made to the pattern recognition algorithm for first-order Markov chain and some results of numerical investigation are presented. The main goal of these investigations was to verify if it is reasonable to make such assumption in the real recognition tasks.

1 Introduction

In many pattern recognition problems there are exists dependencies among the patterns to be recognized. Additionally there are exists dependencies in the feature vectors of recognized patterns. For instance, this situation is typical for recognition of state in technological processes or in the computer-aided medical diagnosis [10], [11] to name only a few. In such cases, we should recognize the sequence of patterns, for example the sequence of states of technological processes or the sequence of patient's states in medical diagnosis. In this paper, similar as in the [6], [9], [12] the assumption of Markov dependence in the sequence of classes of recognized patterns is made. Furthermore, it is assumed that the features vectors are conditional independent. Based on this model and using Bayes' approach in the case of complete probabilistic information the pattern recognition algorithms for first-order Markov chains can be presented in recurrent form [5], [9].

In the practice the pattern recognition algorithms in the case of complete statistical information are the base on which we can propose the algorithms with learning, because usually instead of knowing needed probabilities and class density functions we have so called set of learning sequences. One conceptually simple and effective method of building the algorithms with learning is to replace these probabilities by their estimators and to calculate the needed values of density functions using nonparametric estimators like Parzen estimator for example[4]. In such cases, for calculation reasons and to simplify the learning procedures, the assumption about the conditional independency among the attributes in the feature vector is usually made. Such classifiers are known in the literature as naive Bayes classifiers [1], [2], [8]. There are plenty of papers, in which the naive Bayes classifier is examined. Some interesting results, concerning text recognition, can be found in [7].

However, in this paper some results of simulation investigations of naive Bayes classifier in the case of Markov chain recognition task are presented. In these investigations, the sequences of patterns with dependency in two dimensional features vector, which form the first-order Markov chains, were generated by simulation and next in the case of complete probabilistic information the pattern recognition algorithm for Markov chain was tested. In this paper we want to investigate if taking into account the dependencies in the feature vectors can really improve the quality of recognition in comparison with the attempt in which assumption of class conditional independence in the feature vector is made. Additionally, results of Markov chain algorithms were compared with results, which were obtained using very well known Bayes' decision rule for independent patterns.

2 Statement of the problem

Let us consider the classical problem of pattern recognition that is concerned with the assignment of a given pattern to one of m possible classes from the set of classes $M = \{1, \ldots, m\}$. Let x_n, which takes values in $r-$dimensional Euclidean space $\mathbb{E}^r$, $x_n = \left[x_n^1, x_n^2, \ldots, x_n^r\right]^T$ denote the vector of measured features of the n-th pattern to be recognized and let j_n denote the label of the class to which the pattern in question belongs. Thus: $\bar{x}_n = \{x_1, x_2, ..., x_n\}$,$\bar{j}_n = \{j_1, j_2, ..., j_n\}$ denote the sequence of feature vectors and true identities, respectively. In this paper it is assumed that x_n, j_n are observed values of random variables X_n, J_n for $n = 1, 2, \ldots$, while random variablesX_n are multidimensional. It is also assumed that the sequence of classifications forms first-order Markov chain:

$$P\left(J_n = j_n \left| J_{n-1} = j_{n-1}, J_{n-2} = j_{n-2}, \ldots, J_1 = j_1\right.\right) = \tag{1}$$

$$= P\left(J_n = j_n \left| J_{n-1} = j_{n-1}\right.\right).$$

Note that the first-order Markov chain is described by the initial probabilities:

$$P\left(J_1 = j_1\right) = p_{j1}\ , \qquad j_1 \in M, \tag{2}$$

and the set of transition probabilities:

$$p^n_{j_n, j_{n-1}} = P\left(J_n = j_n \mid J_{n-1} = j_{n-1}\right) , \; j_n , j_{n-1} \in M , \; n = 2, 3, \ldots . \quad (3)$$

Let $f(x_n | j)$ be the conditional density function of random variable X_n given that $J_n = j$, $j \in M$, identically for all natural n. For simplicity we shall assume conditional independence among variables X_n, n = 1, 2, ..., which implies that

$$\bar{f}_n(\bar{x}_n \mid \bar{j}_n) = \prod_{\alpha=1}^{n} f(x_\alpha \mid j_\alpha) , \quad n = 1, 2, \ldots . \quad (4)$$

This assumption states that, given the true identity of a pattern, the distribution of measurement vector is independent of the features and true identities of previous and future patterns, but it is dependent only on the true identity of the pattern in question.

In the naive Bayes pattern recognition algorithms we also assume that the attributes in the features vectors x_n , $n = 1, 2, \ldots$given the class $j \in M$are conditional independent :

$$f(x_n \mid j) = f\left(x_n^1, x_n^2, \ldots, x_n^r \mid j\right) = \prod_{\beta=1}^{r} f_\beta\left(x_n^\beta \mid j\right) \quad (5)$$

This assumption greatly simplify the learning procedures, but we lose the part of complete probabilistic information.

3 Pattern recognition algorithm

In this part, under the assumption of complete statistical information and using Bayes' approach, the decision rules of pattern recognition algorithm for second-order Markov chains are presented. In the papers [5], [6] it is shown, that for the special case of a 0-1 loss function, i.e.:

$$L(i_n, j_n) = 0 \; if \; i_n = j_n \; or \; 1 \; otherwise \; for \; n = 1, 2, \ldots, \quad (6)$$

where $L(i_n, j_n)$ is the loss incurred by the classifier if a pattern from the class j_n is assigned to the class i_n at the moment n, the Bayes' decision rule assigns the n-th recognized pattern to the class i_n with the highest a posterior probability after observing $\bar{x}_n$ for all natural n :

$$i_n = \Psi_n^*(\bar{x}_n) , \quad n = 1, 2, \ldots, \quad (7)$$

if for every $s \in M$, $s \neq i_n$,

$$p_{1,n}(i_n \mid \bar{x}_n) > p_{1,n}(s \mid \bar{x}_n) \quad (8)$$

where: $p_{1,n}\left(j_n \,|\, \bar{x}_n\right) = P\left(J_n = j_n \,\middle|\, \bar{X}_n = \bar{x}_n\right)$, and $f_{1,n}\left(\bar{x}_n\right)$ denotes the joint conditional density function of the sequence of random variables $\bar{X}_n$. Instead of calculate the probabilities (8) it is sufficient to maximize only the following discriminate functions:

$$d\left(j_n, \bar{x}_n\right) = f_{2,n}\left(\bar{x}_n \,|\, j_n\right) \cdot p_{2,n}\left(j_n\right) = p_{1,n}\left(j_n \,|\, \bar{x}_n\right) \cdot f_{1,n}\left(\bar{x}_n\right) \qquad (9)$$

$$n = 1, 2, \ldots \qquad j_n \in M$$

where $p_{2,n}\left(j_n\right) = P\left(J_n = j_n\right)$ and $f_{2,n}\left(\bar{x}_n \,|\, j_n\right)$ denotes the joint conditional density function of the sequence of random variables $\bar{X}_n$ given that $J_n = j_n$. In the papers [5], [6] it is shown, that the discriminate functions can be calculated recursively as follows:

$$d\left(j_n, \bar{x}_n\right) = f\left(x_n \,|\, j_n\right) \cdot \sum_{j_{n-1}=1}^{m} p^{n}_{j_n, j_{n-1}} \cdot d\left(j_{n-1}, \bar{x}_{n-1}\right) \qquad (10)$$

for all natural $n \geq 2$ and for every $j_n, j_{n-1} \in M$. At the first step of classification the discriminate functions can be obtained immediately:

$$d\left(j_1, x_1\right) = f\left(x_1 \,|\, j_1\right) p_{j_1} \quad n = 1, j_1 \in M \qquad (11)$$

In order to classify the n-th pattern we take into account in our recognition task the whole sequence of patterns to this moment ("context") calculating the discriminant functions (9) according to the recursive formulas (10), (11). Then the pattern recognition algorithm for first-order Markov chain classifies the n-th recognized pattern to this class for which the discriminant function is maximal. The difference between the naive Bayes classifier and complete algorithm in the case of Markov chain recognition consists in the way in which the needed values of density functions in the discriminant functions are obtained. In the first case we simplify calculation using assumption (5) and in the second case we take complex class density functions.

4 Simulation investigations

During the simulation investigations we considered the set of three classes $M = \{1, 2, 3\}$. The parameters of Markov chain were constant i.e. $p_{j1} = [0.25 \;\; 0.5 \;\; 0.25]^T$ and the transition probabilities (3) formed constant transition matrix:

$$P = \begin{bmatrix} 0.6 & 0.6 & 0.1 \\ 0.1 & 0.1 & 0.8 \\ 0.3 & 0.3 & 0.1 \end{bmatrix}$$

where for example $P(J_n = 2 \mid J_{n-1} = 2) = 0.8$ for every n. In each class two-dimensional Gaussian density functions were assumed. This density function was described by the vector of mean values $m_j = [m_{j\,1}, m_{j\,2}]^{T}$ and the covariance matrix:

$$\Sigma_j = \begin{bmatrix} \sigma_{j\,1}^2 & \rho\,\sigma_{j\,1}\sigma_{j\,2} \\ \rho\,\sigma_{j\,1}\sigma_{j\,2} & \sigma_{j\,2}^2 \end{bmatrix},$$

with the same parameter of correlation ρ for j=1,2,3.

In every experiment two hundred testing sequences of length equal to ten were generated using program Matlab 6.5. The vectors of mean values were constant and equal:

$$m_1 = [70,\ 110]^T,\ m_2 = [80,\ 190]^T,\ m_3 = [90,\ 270]^T.$$

The variances of each attribute in every class were the same i.e. $\sigma_{j\,1}^2 = \sigma_{j\,2}^2 = 900$ for j = 1, 2, 3. In this way according to the changes of correlation parameter the covariance matrix for every class was changing in the same way. In these simulation investigations three different coefficients of correlation were taken into account : $\rho = 0.3$, $\rho = 0.5$, $\rho = 0.7$ so the covariance matrix for each class were:

$$\Sigma_j = \begin{bmatrix} 900 & 270 \\ 270 & 900 \end{bmatrix}, \Sigma_j = \begin{bmatrix} 900 & 450 \\ 450 & 900 \end{bmatrix}, \Sigma_j = \begin{bmatrix} 900 & 630 \\ 630 & 900 \end{bmatrix}$$

respectively. Of course in the case of Gaussian two dimensional density function the correlation parameter ρ is the measure of dependence between attributes in the feature vectors. Let us notice, that treating our attributes as independent the covariance matrix become diagonal and the value of joint density function in needed point x_n can be calculated for every class as follows:

$$f(x_n \mid j_n) = f(x_n^1, x_n^2 \mid j_n) = f_1(x_n^1 \mid j_n) \cdot f_2(x_n^2 \mid j_n) \qquad (12)$$
$$n = 1, 2, \ldots, \quad j_n \in M$$

where:

$$f_1(x_n^1 \mid 1) = N(70,\ 30),\ f_2(x_n^2 \mid 1) = N(110,\ 30)$$
$$f_1(x_n^1 \mid 2) = N(80,\ 30),\ f_2(x_n^2 \mid 2) = N(190,\ 30)$$
$$f_1(x_n^1 \mid 3) = N(90,\ 30),\ f_2(x_n^2 \mid 3) = N(270,\ 30)$$

Then calculating the values of density functions either according to the joint distribution $f(x_n^1, x_n^2 \mid j_n)$ or to the formula $f_1(x_n^1 \mid j_n) \cdot f_2(x_n^2 \mid j_n)$ we can investigate, in the case of complete probabilistic information, how important in the recognition task is taking into account the dependencies among attributes in the feature vectors. The results of pattern recognition algorithms for first-order Markov chain in these two above mentioned methods of calculating the values of density functions were compared with the results obtained using algorithm which does not take into account any dependencies in the sequences of patterns. The decision functions of this algorithm are as follows:

$$d\,(j_n, x_n) = f\,(x_n\,|\,j_n)\;p_{j_n} \quad , n = 1, 2, \ldots , j_n \in M \tag{13}$$

The probabilities:

$$P\,(J_n = j_n) = p_{j_n} \,, \quad j_n \in M,\; n = 1, 2, \ldots \tag{14}$$

we can assume as constant and equal to the initial probabilities of Markov chain $P\,(\,J_1 = j_1\,) = p_{j_1}\,,\quad j_1 \in M$, or we can calculate them according to the step of classification and to the recursive formula:

$$p_{j_n} = \sum_{j_{n-1}}^{m} p^{n}_{j_n\,,\,j_{n-1}} p_{j_{n-1}} \,, \quad j_n\,, j_{n-1} \in M\,, \quad n = 2, 3, \ldots \tag{15}$$

with the initial condition for n =1:

$$P\,(\,J_1 = j_1\,) = p_{j_1}\,, \quad j_1 \in M. \tag{16}$$

In this way during simulation investigations the following pattern recognition algorithms were compared by calculating the percentage of correctly classified patterns:

A - algorithm for Markov chain with dependent attributes,
B - algorithm for Markov chain with independent attributes,
C - modified Bayes algorithm with dependent attributes,
D - modified Bayes algorithm with independent attributes,
E - classical Bayes algorithm with dependent attributes,
F - classical Bayes algorithm with independent attributes.

The results of numerical example, in which the correlation parameter ρ was changed, are presented in the table 1:

Table 1. The results of simulation

Algorithm	A	B	C	D	E	F
$\rho = 0.3$	92,4	91,1	89,5	88,4	87,4	86,7
$\rho = 0.5$	93,7	90,4	90,4	87,1	89,6	85,7
$\rho = 0.7$	95,8	89,6	94,6	86,7	93,8	83,6

Looking at these results we can see the improvement of the quality of classification, taking into account the dependencies among the attributes in the feature vector. In addition, as it can be expected, the improvement is bigger if the correlation parameter is growing up. In this example of Markov chain we can notice, that the difference between the algorithm A and B for every investigated degree of correlation is less significant that the difference between algorithms C and D, or between E and F. This difference for Markov chain recognition is less than 6 % in the case of complete statistical information. In the real problems we have to solve the problem of learning, which can be very complicated if we want to use nonparametric technique (like Parzen method for example) and if we want to take into account the dependency in the feature vector. Additionally,in the case of discrete attribute the number of needed samples in the learning sequences must be greater because we have to estimate all needed joint probabilities.

5 Conclusions

In this paper some results of numerical investigation for pattern recognition algorithm for first-order Markov chain with the "naive" assumption concerning the feature vector are presented. This assumption state, that in every class the attributes in the feature vector are conditionally independent. From theoretical point of view, i.e. in the case of complete probabilistic information in which we know the class conditional joint distribution of feature vector or the class conditional joint density function, there is no problem in calculating the values of decision functions. However the problem remains if have to consider the algorithms with learning. In such case the simplifying assumption about the independencies among the attributes in the feature vector is very useful. First assuming the independency in the feature vector, the calculation of estimates of needed joint probabilities requires less training samples. Second, using nonparametric estimators of values of density function we have the problem of finding the proper kernel function for the kind of dependence in the feature vector. That's way such assumption is usual made. Concluding, in the case of Markov chain recognition, if we can accept less quality of pattern recognition algorithm (about 6 %) the assumption of independence of attribute in the feature vector can be made. Similar investigations should be done for another cases of Markov chain recognition tasks, i.e. for higher-order or controlled Markov chains.

References

1. Duda R, Hart P (1973) Pattern classification and scene analysis. John Wiley, New York
2. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classfiers. Machine learning 29:131-163
3. Fu K (1974) Syntactic methods in pattern recognition. New York Academic Press
4. Greblicki W (1978) Pattern recognition procedures with nonparametric density estimates. In: IEEE Trans. on SMC 8:809–812
5. Kurzynski M (1997) Pattern recognition-statistical approach. Publishers of Wroclaw University of Technology.

6. Kurzynski M, Zolnierek A (1980) A recursive classifying decision rule for second-order Markov chain. Control and Cybernetics 9:141–147
7. McCallum, Nigam K (1998) A comparison of event models for naive Bayes text classification. In AAAI-98 Workshop on Learning and Text Categorization, Madison, WI, USA:41–48
8. Mitchel T (1997) Machine learning. McGraw Hill, New York
9. Raviv J (1967) Decision making in Markov chain applied to the problem of pattern recognition. In: IEEE Trans. on IT 21:536–551
10. Zolnierek A (1982) Computer-aided recognition of the human acid-base state. In Proc. of 6-th Int. Conf. on Pattern Recognition:1219
11. Zolnierek A (1983) Pattern recognition algorithms for controlled Markov chains and their application to medical diagnosis. Pattern Recognition Letters 1:299–303
12. Zolnierek A (2003) The simulation investigations of pattern recognition algorithm for second-order Markov chains. In: Proc. of the 37-th conference, Brno, Czech Republic, Acta MOSIS 92:29–35