Interconnection Between Features Properties and Probability of Error in Statistical Recognition of Two Classes

Leonid S. Fainzilberg

V.M.Glushkow Institute of Cybernetics
40, Prospect Akademika Glushkowa, MSP 252650 KIEV 207 Ukraine

Abstract

The concept of informative and useful features in statistical recognition of two classes is introduced. It is demonstrated that any feature can be informative but not useful in combination with other features. The sufficient conditions which assume the usefulness of feature according to error probability are consider.

1. Introduction

When the pattern recognition problem is solved the situation in which it is necessary to estimate the usefulness of the features can often be met. Adopting to statistical pattern recognition one of the approaches to solution the problem of feature selection which has been proposed in several papers, in particular in [1, 2], allows to use the information criteria such as Shannon entropy. At the same time as is well known the selection of features by probability of error is more suitable [3, 4].

In this paper we consider conditions which ensure the usefulness of any feature in combination with other features in sense of probability of errors under incomplete a priori information about feature distributions. As far as we know this approach is considered for the first time.

2. Formal description of problem.

Let $V = \{V_1,...,V_M\}$ be set of M classes, $x = (x_1,...,x_N)$ is initial description vector, which includes N features $x_1,...,x_N$. Adopting the Bayesian approach, let further $P(V_k)$ denotes a prior probability for class V_k, $\sum_{k-1}^{M} P(V_k) = 1$ and let $p(x|V_k)$ is the conditional distribution of probability of feature values in class V_k. In this paper we are limited by case of two classes $(M = 2)$.

Definition 1: Any feature x_n $(1 \le n \le N)$ is said to be *informative itself* if

$$H(V) - H(V|X_n) > 0, \qquad (1)$$

where $H(V) = -\sum_{k-1}^{2} p(V_k) \log P(V_k)$ is the initial entropy (by Shannon) of set $V = \{V_1, V_2\}$ and

$$H(V|X_n) = -\sum_{x_n \in X_n}^{2} p(x_n) \sum_{k-1} P(V_k|x_n) \log P(V_k|x_n)$$

is the middle conditional entropy, which estimates the uncertainty of decisions, which are making only with reference to a posterior probabilities $P(V_1|x_n)$ and $P(V_2|x_n)$.

Definition 2: A feature x_n $(1 \le n \le N)$ is said to be *informative in combination* with other $N-1$ features if

$$H(V|X^{(N-1)}) - H(V|X^{(N)}) > 0, \qquad (2)$$

where

$$H(V|X^{(N)}) = -\sum_{x^{(N)} \in X^{(N)}} p(x^{(N)}) \sum_{k-1}^{2} P(V_k|x^{(N)}) \log P(V_k|x^{(N)}),$$

$$H(V|X^{(N-1)}) = -\sum_{x^{(N-1)} \in X^{(N-1)}} p(x^{(N-1)}) \sum_{k-1}^{2} P(V_k|x^{(N-1)}) \log P(V_k|x^{(N-1)}),$$

i.e. the middle conditional entropy is changing (is increasing) as a result of transformation of full vector $x^{(N)} = (x_1,...,x_N)$ to the reduced vector $x^{(N-1)}$, which doesn't contain this feature x_n.

Definition 3: Any feature x_n $(1 \le n \le N)$ is said to be *useful itself* if

$$P_1(e) < P_0(e), \qquad (3)$$

where $P_0(e) = \min\{P(V_1), P(V_2)\}$ is a priori probability of error and

$$P_1(e) = \sum_{x_n \in X_n} p(x_n) \min \{P(V_1|x_n), P(V_2|x_n)\}$$

is the middle probability of error decisions, which are making only with reference to a posterior probabilities $P(V_1|x_n)$ and $P(V_2|x_n)$.

Definition 4: A feature x_n $(1 \le n \le N)$ is said to be *useful in combination with* other $N-1$ features if

$$P_N(e) < P_{N-1}(e), \qquad (4)$$

where

$$P_N(e) = \sum_{x^{(N)} \in X^{(N)}} p(x^{(N)}) \min \{P(V_1|x^{(N)}), P(V_2|x^{(N)})\}$$

and

$$P_{N-1}(e) = \sum_{x^{(N-1)} \in X^{(N-1)}} p(x^{(N-1)}) \min \{P(V_1|x^{(N-1)}), P(V_2|x^{(N-1)})\}$$

i.e. the middle probability of error is changing (is increas-

ing) as a result of transformation of full vector $x^{(N)} = (x_1,...,x_N)$ to the reduced vector $x^{(N-1)}$ which doesn't contain this feature x_n.

In fact the verification of conditions (1) - (4) is not very difficult problem when we know a priori probabilities $P(V_k)$ and the conditional distribution $p(x^{(N)}|V_k)$. However, this case is unusual. Usually one must solve such problem when $P(V_k)$ and $p(x^{(N)}|V_k)$ are unknown.

In this paper we consider the sufficient conditions which ensure that any feature is useful in the sense of (4) when incomplete a priori information takes place.

3. Main results

At first let us consider some interest properties of features which we have discovered in case when problem of statistical recognition of two classes was solved.

Let x_n and x_m $(1 \le n \le N,\ 1 \le m \le N,\ n \ne m)$ are any of features in vector $x = (x_1,...,x_N)$. We shall call these features as *unconditionally independent*, when

$$p(x_n, x_m) \equiv p(x_n)p(x_m), \tag{5}$$

and *conditionally independent* for class V_k, when

$$p(x_n,x_m|V_k) \equiv p(x_n|V_k)\ p(x_m|V_k). \tag{6}$$

It is turned out [6], that following properties are valid.

Property 1: Let x_n and x_m are unconditionally independent according to (5) and both of these features are differently distributed in classes, i.e.

$$p(x_n|V_1) \ne p(x_n|V_2), \tag{7}$$

$$p(x_m|V_1)\ ) \ne p(x_m|V_2). \tag{8}$$

Then these features are conditionally dependent at least in one of classes, i.e.

$$p(x_n,x_m|V_k)\ ) \ne p(x_n|V_k)\ p(x_m|V_k),\ k=1 \text{ or/and } k=2. \tag{9}$$

Property 2: Let x_n and x_m are unconditionally independent according to (5) and at least one of these features is identically distributed in classes, i.e.

$$p(x_n|V_1) \equiv p(x_n|V_2), \tag{10}$$

or/and

$$p(x_m|V_1) \equiv p(x_m|V_2). \tag{11}$$

Under these assumptions if x_n and x_m are conditionally independent in one class then x_n and x_m certainly are conditionally independent in another and vice versa, if x_n and x_m are conditionally dependent in one class then x_n and x_m certainly are conditionally dependent in another.

Property 3: Let x_n and x_m are unconditionally independent according to (5) and both are identically distributed in classes, i.e. (10) and (11) are valid.

Under these assumptions if x_n and x_m are conditionally dependent then this dependence is different in classes:

$$p(x_m|x_n,V_1)\ ) \ne p(x_m|x_n,V_2). \tag{12}$$

It is clear that the above properties can be easily generalized on case when instead of two features x_n and x_m may be considered any subsets of features, particular x_n and $x^{(N-1)}$. This particular case will be utilized later.

It should be observed here that the strictly inequality (4) can be valid only under the restrictions

$$P_{N-1}(e) \ne 0 \tag{13}$$

and

$$P_N(e) < P_0(e). \tag{14}$$

The verification of restriction (13) may be possible on basis of following theorems [6].

Theorem 1: Assume that each of $N-1$ features have a crossing distributions in classes, i.e.

$$X_{1n} \cap X_{2n} \ne \varnothing \quad \text{for each } n = 1,2,...,N-1, \tag{15}$$

where $X_{kn} = \{x_n : p(x_n|V_k) \ne 0\}$. Then the recognition of classes V_1 and V_2 without errors $(P_{N-1}(e) = 0)$ is possible only when these features are conditionally dependent at least in one of classes, i.e.

$$p(x^{(N-1)}|V_k) \ne \prod_{n-1}^{N-1} p(x_n|V_k),\ k=1 \text{ or/and } k=2 \tag{16}$$

Theorem 2: Assume that each of $N-1$ features are identically distributed in classes, i.e.

$$p(x_n|V_1) \equiv p(x_n|V_2) \text{ for each } n=1,2,...,N-1. \tag{17}$$

Then the recognition of classes V_1 and V_2 without errors ($P_{N-1}(e) = 0$) is possible only when these features are conditionally dependent in both classes, i.e.

$$p(x^{(N-1)}|V_k)\ ) \ne \prod_{n-1}^{N-1} p(x_n|V_k),\ k=1 \text{ and } k=2. \tag{18}$$

Recall that according to generalization of Properties 1-3 the conditions of above Theorems may be satisfied when features are unconditionally independent, i.e.

$$p(x^{(N-1)}) \equiv \prod_{n-1}^{N-1} p(x_n).$$

It is easy to make sure that feature x_n does not satisfy the condition (1) if and only if

$$p(x_n|x^{(N-1)},V_1) \equiv p(x_N|x^{(n-1)},V_2). \tag{19}$$

It follows immediately from (19) that the feature x_n cannot be informative itself in sense of (1) in two cases:

a) when x_n is conditionally independent in both classes and is identically distributed in these classes;

b) when conditional dependence between x_n and other $N-1$ features is identical in classes.

It is interesting to observe that in case b) the equality $H(V|X^{(N-1)}) = H(V|X^{(N)})$ is valid even when inequality (3) takes place, i.e. even when x_n is useful feature itself.

Obviously that $P_N(e) = P_{N-1}(e)$ if (19) is valid. At the same time it can be shown that equality $P_N(e) = P_{N-1}(e)$ may be possible not only in trivial case when $P(V_k|x^{(N)}) \equiv P(V_k|x^{(N-1)})$ but in general when transformation of $x^{(N)}$ to $x^{(N-1)}$ leads to changing a posterior probability $P(V_k|)$.

It follows from this that inequality (2) can be regarded *only as necessary but not sufficient condition* of usefulness of x_n in combination with other *N-1* features in the sense of Definition 4.

To demonstrate this fact we shall consider following

Example: Let $P(V_1) = 0.4$, $P(V_2) = 0.6$, $N = 2$. Assume that both features have two values $(x_1 = x_1^1, x_1^2; x_2 = x_2^1, x_2^2)$ being $p(x_1^1, x_2^1|V_1) = 0.56$; $p(x_1^1, x_2^2|V_1) = 0.14$; $p(x_1^2, x_2^1|V_1) = 0.24$; $p(x_1^2, x_2^2|V_1) = 0.06$; $p(x_1^1, x_2^1|V_2) = 0.02$; $p(x_1^1, x_2^2|V_1) = 0.08$; $p(x_1^2, x_2^1|V_2) = 0.18$; $p(x_1^2, x_2^2|V_2) = 0.72$. As is easily seen x_1 and x_2 are conditionally independent in both classes.

According to data of this example we have $H(V|X_1, X_2) = 0.512$; $H(V|X_1) = 0.678$. Consequently $H(V|X_1) - H(V|X_1, X_2) = 0.161 > 0$, but $P_2(e) = P_1(e) = 0.18$, i.e. x_2 is informative but not useful in combination with x_1.

The question comes into being: whether it is possible to make judgment about usefulness of features in sense of error probability with reference to changing of Shannon entropy? The answer on this question gives following

Theorem 3: [7] If under transformation of $x^{(N)}$ to $x^{(N-1)}$ the middle conditional entropy $H(V|)$ is changing so that

$$H(V|X^{(N-1)}) - H(V|X^{(N)}) > I^*, \qquad (20)$$

where

$$I^* = - H(V|X^{(N)}) [1 + 0.5 \log 0.5 H(V|X^{(N)})] - \\ - [1 - 0.5 H(V|X^{(N)})] \log [1 - 0.5 H(V|X^{(N)})], \qquad (21)$$

then x_n is useful in combination with other *N-1* features in sense of Definition 4.

As is seen on the right-hand side of inequality (20) in contrast to (2) the "threshold" $I^* > 0$ is present. This threshold according to (21) is dependent on $H(V|X^{(N)})$ exclusively. The maximum $I^*_{max} = \log_2 1.25$ is attained when $H(V|X^{(N)}) = 0.4$.

Theorem 4: [7] Feature x_n is useful itself in sense of Definition 3 when

$$H(V) - H(V|X_n) > I_0^*, \qquad (22)$$

where

$$I_0^* = \log (1 + \lambda_0) - \lambda_0[1 + \lambda_0]^{-1} \log \lambda_0 - \\ - 2\min\{[1 + \lambda_0]^{-1}, \lambda_0[1 + \lambda_0]^{-1}\}. \qquad (23)$$

Here $\lambda_0 = P(V_2)[P(V_1)]^{-1}$.

Consequence 1: Feature x_n which is informative itself is useful itself only when a priori probabilities of two classes is equal, i.e. $P(V_1) = P(V_2)$. This fact results from (22) with reference (23) so long as $I_0^* = 0$ when $\lambda_0 = 1$.

It is worth noticing that the generalizations of Theorem 4 on case $N > 1$ gives possibility to check also the restriction (14).

Theorem 5: [5] Under the restrictions (13) and (14) the feature x_n $(1 \le n \le N)$ is useful in combination with other *N-1* features according to (4) if following conditions are fulfilled:

1^0. x_n is differently distributed in classes, i.e.

$$p(x_n|V_1)) \neq p(x_n|V_2); \qquad (24)$$

2^0. x_n and other *N-1* features are conditionally independent in both classes, i.e.

$$p(x_n|x^{(N-1)}, V_1) \equiv p(x_n|x^{(N-1)}, V_2), \quad k=1,2; \qquad (25)$$

3^0. The conditional distributions $p(x^{(N-1)}|V_k)$ are continuous functions of $x^{(N-1)}$ and sets

$$X_k^{(N-1)} = \{x^{(N-1)} : p(x^{(N-1)}|V_k) \neq 0\}, \; k=1,2$$

are connected domains.

Theorem 6: [5] Under the restrictions (13) and (14) the feature x_n $(1 \le n \le N)$ is useful in combination with other *N-1* features according to (4) if following conditions are fulfilled:

1^0. x_n and other *N-1* features are conditionally dependent at least in one of classes, i.e.

$$p(x_n|x^{(N-1)}, V_k)) \neq p(x_n|V_k), \quad k=1 \text{ or/and } k=2 \qquad (26)$$

2^0. this dependence is different in classes:

$$p(x_n|x^{(N-1)}, V_1)) \neq p(x_n|x^{(N-1)}, V_2), \quad \forall x^{(N-1)} \in X^{(N-1)} \qquad (27)$$

and condition 3^0 of Theorem 5 is also fulfilled.

It is interesting to note that if feature x_n satisfies only the condition (24) of Theorem 5 this fact itself doesn't denote that x_n is useful in combination with other *N-1* features. Moreover, in accordance with Consequence 1 this feature may be useless itself when $P(V_1) \neq P(V_2)$.

In other hand feature x_n, which is identically distributed in classes, i.e. $p(x_n|V_1) \equiv p(x_n|V_2)$ undoubtedly is useful in sense of Definition 4 when all conditions of Theorem 6 are fulfilled.

It is interesting also, that in accordance with (23) the quantitative estimation of feature usefulness may be derived for following special case.

Special case. Suppose that all conditions of the Theorem 5 are satisfied. Assume further that x_n is not useful itself, i.e. $P_1(e) = P_0(e)$. Then following estimation of interconnection between error probabilities $P_{N-1}(e)$ and $P_N(e)$ are valid:

$$P_{N-1}(e) \le - 0.5 P_N(e) \log P_N(e) - 0.5[1 - P_N(e)] * \\ * \log [1 - P_N(e)] + 0.5 \log(1 + \lambda_0) - 0.5\lambda_0[1 + \lambda_0]^{-1} * \\ * \log \lambda_0 - \min\{[1 + \lambda_0]^{-1}, \lambda_0[1 + \lambda_0]^{-1}\}. \qquad (28)$$

The results which we have considered in this paper were successfully applied when solving the problem of selection the useful features for statistical recognition of true and false thermal effects of phase transformation in cooling steel (see Fainzilberg [8]).

References

[1] P.M.Lewis, "The characteristic selection problem in recognition system", IRE Trans. Inform. Theory, 1962, vol.8, N2, pp.171-178.

[2] J.T.Tou , R.P.Heydorn, "Some approaches to optimum feature selection", In Computer and Information Sciences. Academic Press, New York, 1967, vol. 11, pp.57-89.

[3] M.Ben-Bassat, S. Gal, "Properties and convergence of a posteriori probabilities in classification problems", Pattern Recognition, 1977, vol. 9, pp.99-107.

[4] M.Ben-Bassat, "Irrelevant features in pattern recognition",

IEEE Transactions on computers, August 1978, vol.C-27, N8, pp.749-766.

[5] L.S.Fainzilberg, "The application of statistical pattern recognition methods to thermal analysis of metal composition", Kibernetika, 1978, N6, pp.159-162 (on Russian).

[6] L.S.Fainzilberg, "On question to recognition of two classes without error by combination of crossing features", Kibernetika, 1982, N4, pp.104-109 (on Russian).

[7] L.S.Zhitetsky , L.S.Fainzilberg, "On informative approach to estimation of features usefulness in statistical pattern recognition", Izwestija Akademii Nauk SSSR. Technicheskaja Kibernetika, 1983, N4, pp.120 -126 (on Russian).

[8] L.S.Fainzilberg, "Method and device for discriminating thermal effect of phase transformation of metals and alloys in the process of their cooling", USA Patent N 4198679, 1980.