Why Relevant Features May Be Unuseful in Statistical Recognition of Two Classes

Leonid S. Fainzilberg

V.M.Glushkov Institute of Cybernetics
40, Prospect Akademika Glushkova,
252022, Kiev, UKRAINE
Fax + 380 44 2661570

Abstract

The interconnection between known definitions of irrelevant and unuseful features in statistical recognition of two classes is considered. It is demonstrated that any relevant feature in the sense of a posteriori probabilities change may be unuseful regarding average error probabilities change. The necessary and sufficient conditions of usefulness of binary feature according to error probabilities decrease are established.

1. Introduction

One of the important pattern recognition problems is the problem of feature selection, which has been studied in many papers, in particular in [1-11]. This problem may be considered as a two-stage one. At the first stage the initial set of features $x_1, ..., x_D$ is constructed, but this procedure is not formal : the constructor tries to include any measurement in initial set when the assumption of it's usefulness to recognized classes $V_1 , ... , V_M$ may be done. These assumptions often are based only on intuition and experience of the constructor.

According to Pudil, Ferri, Novovicova and Kittler [1] the main goal of the feature selection problem is to select a subset of d features from the initial set ($d < D$) without a significant degrading. It is assumed that at the second stage a suitable *formal criteria* has been chosen to evaluate the effectiveness of any feature. Just at this stage the most general statement of the feature selection problem can be formulated : we need to remove from initial set any unuseful measure and detect an optimal feature subset.

That is why the question comes into being: how to make judgment about usefulness of any feature when concrete pattern recognition problem is solved.

Adopting to statistical pattern recognition in Ben-Bassat [2] the definition of irrelevant feature was introduced and was shown that a feature is irrelevant *if and only if* it is identically distributed in classes. In several papers, in particular in Lewis [6], in Tou and Heydorn [7] the information criteria based on change of Shannon entropy was proposed. It is easily verified that irrelevant in the sense Ben-Bassat's feature necessarily does not change the Shannon entropy and vice versa. In other words the above mentioned definitions are equivalent.

At the same time in [9, 11] we have proved that change of Shannon entropy can be regarded only as necessary but not sufficient condition that feature is useful in the sense of error probability change. That is why known information criteria is not suitable for feature selection when statistical pattern recognition problem is solved.

In this paper the interrelation between above said criteria we consider and show when and why relevant and informative feature may be unuseful. We also establish the new necessary and sufficient conditions of feature usefulness in case when feature is binary (has only two possible values) and the problem of statistical recognition of two classes is solved. It is well known that such binary features named symptoms are widely used in pattern recognition practice, for example, in medical and technical diagnostic systems.

2. The concept of irrelevant and unuseful features

Let $V = \{V_1 , ..., V_M\}$ be set of M classes, $x \in X$ is a feature. Adopting to the Bayesian approach, let further $P(V_k)$ denote *a priori* probability for class V_k, $\sum_{k=1}^{M} P(V_k) = 1$ and let $p(x \mid V_k)$ be the conditional distribution of probability of feature values in class V_k .

It is known that for a zero-one cost matrix for correct and incorrect classification, the optimal decision rule assigns the pattern to the class with the highest *a posteriori* probability which may be obtained by Bayes formula

$$P(V_k \mid x) = P(V_k)\, p(x \mid V_k)\, [\sum_{i=1}^{M} P(V_i)\, p(x \mid V_i)]^{-1} \qquad (1)$$

In this case the Bayes risk is reduces to the average probability of error

$$P(e) = \sum_{x \in X} p(x)[1 - \max_i (V_i \mid x)]. \qquad (2)$$

Hence just $P(e)$ may be used as suitable criterion to evaluate the effectiveness of any feature.

In Ben-Bassat [2] the following definition was introduced.

Definition 1. A feature x is said to be irrelevant if the *a posteriori* probability of k-th class $P(V_k \mid x)$ remains unchanged regardless of the value observed for x. Namely

$$P(V_k \mid x) = P(V_k) \qquad (3)$$

for every possible value of $x \in X$, $k = 1, \ldots, M$ and $V_k \in [0, 1]$.

Not striving for the greatest generality, we shall confine ourselves to the study of two-class case where $M=2$ and $V = \{V_1, V_2\}$. In accordance with Ben-Bassat [1] in this case the feature x is irrelevant in sense (3) for $k=1,2$ if and only if this feature is identically distributed in classes:

$$p(x \mid V_1) = p(x \mid V_2). \qquad (4)$$

In Fainzilberg [9] we have introduced another definition in two-class case.

Definition 2 : A feature x is said to be unuseful (unuseful as such!) if the average probability of error decisions

$$P(\mathrm{e}) = \sum_{x \in X} p(x) \min \{P(V_1 \mid x), P(V_2 \mid x)\}, \qquad (5)$$

based on testing of this feature , is equal to the initial probability of error

$$P_0(\mathrm{e}) = \min \{P(V_1), P(V_2)\},$$

based only on *a priori* probabilities , i. e. if following equation is valid

$$P(\mathrm{e}) = P_0(\mathrm{e}). \qquad (6)$$

It is easily verified that when x is an irrelevant feature in the sense of (3) then following (5) this feature is definitely unuseful.

At the same time the opposite direction does not necessarily hold: we shall show that any feature x may be unuseful in sense (6) even when x is relevant and differently distributed in classes, i.e. $p(x \mid V_1) \neq p(x \mid V_2)$.

To avoid awkward notation let us prove the following two lemmas which would be utilized later.

Lemma 1. Let $P(V_1) \neq P(V_2)$. Then any feature x is unuseful *if and only if*

$$[P(V_1) - P(V_2)]\, [P(V_1 \mid x) - P(V_2 \mid x)] > 0 \quad x \in X, \qquad (7)$$

i.e. for every possible value x the sign of the difference of the *a posteriori* probabilities

$$\delta(x) = P(V_1 \mid x) - P(V_2 \mid x) \qquad (8)$$

is the same as the one of the *a priori* probabilities

$$\delta_0 = P(V_1) - P(V_2) \qquad . (9)$$

Proof : Suppose that the inequality (7) is valid, for example, that $P(V_1) > P(V_2)$ and $P(V_1 \mid x) > P(V_2 \mid x)$ for every possible $x \in X$. This means that $\min\{P(V_1), P(V_2)\} = P(V_2)$ and $\min\{P(V_1 \mid x), P(V_2 \mid x)\} = P(V_2 \mid x)$ for every possible $x \in X$. Hence if (7) is valid, then according to (5), $P(\mathrm{e}) = \sum_{x \in X} p(x)\, P(V_2 \mid x) =$ $P(V_2) = \min\{P(V_1), P(V_2)\} = P_0(\mathrm{e})$.

Now suppose that the inequality (7) is not valid, for example, that $P(V_1) > P(V_2)$, but there exists subset $X^* \subset X$ such that $P(V_1 \mid x) < P(V_2 \mid x)$ for $x \in X^*$. In this case in accordance with (5)

$$P(\mathrm{e}) = \sum_{x \in X \setminus X^*} p(x)\, P(V_2 \mid x) + \sum_{x \in X^*} p(x)\, P(V_1 \mid x). \qquad (10)$$

To carry out the change of $P(V_2 \mid x)$ for $P(V_1 \mid x)$ on the right-hand side of (10) and taking into account that $P(V_1 \mid x) < P(V_2 \mid x)$ for $x \in X^*$ we finally obtain that if $X^* \neq \varnothing$ then

$$P(\mathrm{e}) < \sum_{x \in X} p(x)\, P(V_2 \mid x), \text{ i.e. } P(\mathrm{e}) < P_0(\mathrm{e}).$$

Q.e.d.

It is important to note that the condition (7) of Lemma 1 does not require the satisfaction of the equation (3). In fact the condition (7) may be valid not only in case when feature x is identically distributed in classes , but in general if the *a posteriori* probabilities $P(V_1 \mid x)$ and $P(V_2 \mid x)$ are changed under x takes on different possible values, but sign of difference (8) remains unchanged.

Let us show that this situation may be possible only when the *a priori* probabilities of classes are not equal.

We shall confine ourselves to the simplest case when x is a binary feature taking only two possible values $x = x^+$ and $x = x^-$ with conditional probabilities $p(x^+ \mid V_k)$, $p(x^- \mid V_k)$, $k = 1,2$. It is clear that $p(x^- \mid V_k) = 1 - p(x^+ \mid V_k)$.

Lemma 2. If the *a priori* probabilities of classes are not equal, i.e. $P(V_1) \neq P(V_2)$ then for any given conditional probability $p(x^+ \mid V_1)$ may be point such conditional probability $p(x^+ \mid V_2)$ that

$$p(x^+ \mid V_1) \neq p(x^+ \mid V_2) \qquad (11)$$

and equation (6) is valid, i.e. x is relevant but unuseful feature.

Proof : First of all let us note that without loss of generality we may assume that $P(V_2) < P(V_1)$. Let us

introduce the set Ω of possible values $p(x^+ \mid V_1)$ and $p(x^+ \mid V_2)$, which satisfies the following inequalities

$$p(x^+ \mid V_1) > \lambda_0 \; p(x^+ \mid V_2) \tag{12}$$

$$p(x^+ \mid V_1) < \lambda_0 \; p(x^+ \mid V_2) + (1- \lambda_0) \tag{13}$$

where $\lambda_0 = P(V_2)/P(V_1)$ denotes the ratio of the *a priori* probabilities.

In the view of (12), (13) it is clear, that if $P(V_2) < P(V_1)$ than $\Omega \neq \varnothing$ because in this case $\lambda_0 < 1$. Hence for given $p(x^+ \mid V_1)$ one may assign the corresponding $p(x^+ \mid V_2)$ which satisfies the (11) - (13) simultaneously. Then taking into account that

$$p(x) = P(V_1)\, p(x \mid V_1) + P(V_2)\, p(x \mid V_2)$$

when $x = x^+$ and $x = x^-$ and

$$p(x^- \mid V_k) = 1 - p(x^+ \mid V_k)$$

for k=1,2 from (12) and (13) by Bayes formula we have $p(V_1 \mid x^+) > p(V_2 \mid x^+)$ and $p(V_1 \mid x^-) > p(V_2 \mid x^-)$. This implies that condition (7) is valid and in accordance with Lemma 1 the equation (6) is valid certainly.

The proof of Lemma 2 in case when $P(V_2) > P(V_1)$ may be similar, but in this case instead of (12), (13) the following inequalities have to be considered :

$$p(x^+ \mid V_1) < \lambda_0 \; p(x^+ \mid V_2), \tag{14}$$

$$p(x^+ \mid V_1) > \lambda_0 \; p(x^+ \mid V_2) + (1- \lambda_0) \tag{15}$$

which determine the set $\Omega \neq \varnothing$ when $\lambda_0 > 1$.

Consequence 1: If the *a priori* probabilities of classes are equal, i.e. $P(V_1) = P(V_2) = 0.5$ and x is differently distributed in classes, i.e. $p(x^+ \mid V_1) \neq p(x^+ \mid V_2)$ then x is a useful feature certainly. This result follows immediately from (12), (13) or (14), (15) because $\Omega = \varnothing$ when $\lambda_0 = 1$.

Fig.1 and fig.2 show the domain Ω under different values of λ_0.

It should be observed here that domain Ω is reduced when λ_0 is approximating to 1 and $\Omega = \varnothing$ when $\lambda_0 = 1$.

Example : Let $P(V_1) = 0.8$; $P(V_2) = 0.2$. Then $\lambda_0 = 0.25$ and domain Ω is bounded by lines which are shown on Fig.1.

Let $(x^+ \mid V_1) = 0.7$. In this case we may assume the corresponding value $p(x^+ \mid V_2) = 0.2$, because the point $P = (0.2, 0.7)$ belongs to domain Ω.

It is worth noticing that feature x is differently distributed in classes : $p(x^+ \mid V_1) \neq p(x^+ \mid V_2)$ and therefore x is not irrelevant in accordance with Ben-Bassat.

At the same time by Bayes formula we have $p(V_1 \mid x^+) = 0.933$; $p(V_2 \mid x^+) = 0.067$; $p(V_1 \mid x^-) = 0.6$; $p(V_1 \mid x^-) = 0.4$. Hence $p(V_1 \mid x^+) > p(V_2 \mid x^+)$ and $p(V_1 \mid x^-) > p(V_2 \mid x^-)$ and therefore the decision about classes does not change after testing any possible value of feature x. Consequently $P(e) = P_0(e) = 0.2$ and relevant in Ben-Bassat's feature x is not useful.

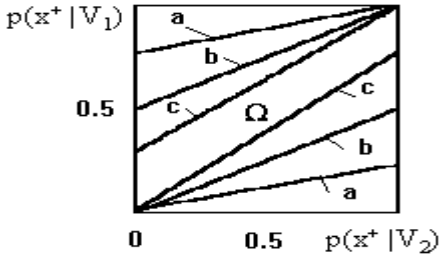

Figure 1: Domain Ω of unuseful feature distribution
(a) - $\lambda_0 = 0.25$; (b) - $\lambda_0 = 0.5$; (c)- $\lambda_0 = 0.75$;

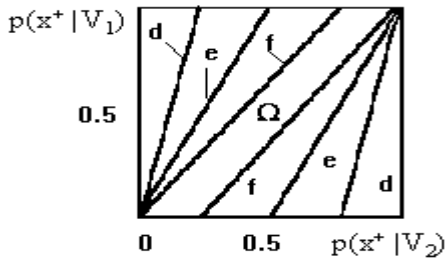

Figure 2 : Domain Ω of unuseful feature distribution
(d) - $\lambda_0 = 4.0$; (e) - $\lambda_0 = 2.0$; (f) - $\lambda_0 = 1.33$.

Using above mentioned lemmas and consequence one can easily formulate the following theorem which defines necessary and sufficient conditions that binary feature is useful.

Theorem 1: Any binary feature x is useful as such in statistical recognition of two classes *if end only if* one of the following conditions is valid

a) $p(x^+ \mid V_1) \neq p(x^+ \mid V_2)$ when $\lambda_0 = 1$;

b) $p(x^+ \mid V_1) < \lambda_0 \; p(x^+ \mid V_2)$ or
$p(x^+ \mid V_1) > \lambda_0 \; p(x^+ \mid V_2) + (1- \lambda_0)$ when $\lambda_0 < 1$;

c) $p(x^+ \mid V_1) > \lambda_0 \; p(x^+ \mid V_2)$ or
$p(x^+ \mid V_1) < \lambda_0 \; p(x^+ \mid V_2) + (1- \lambda_0)$ when $\lambda_0 > 1$,

where $\lambda_0 = P(V_2)/P(V_1)$

3. Informative Condition of Feature Usefulness

In several papers, in particular in [5,6,10], the change of Shannon's entropy is used as a means of feature selection. Let us establish the interconnection between this informative criterion and above mentioned definitions.

It is easily verified that in showing two-class case for any given $x \in X$, the conditional probability of error

$$P(e \mid x) = \min \{P(V_1 \mid x), P(V_2 \mid x)\}$$

uniquely determines the particular conditional Shannon's entropy

$$h(V \mid x) = - P(\mathrm{e} \mid x)\log P(\mathrm{e} \mid x) - [1- P(\mathrm{e} \mid x)] \log[1-P(\mathrm{e} \mid x)].$$

At the same time the analogous interconnection between average probability of error

$$P(\mathrm{e}) = \sum_{x \in X} p(x)\, P(\mathrm{e} \mid x)$$

and average conditional Shannon's entropy

$$H(V \mid X) = \sum_{x \in X} p(x)\, h(V \mid x)$$

does not exist. It may be pointed the least upper boundary and the greatest lower boundary of $H(V \mid X)$ for given $P(\mathrm{e})$ only, which according to [10,11] in two-class case may be written as follows

$$\sup H(V \mid X) = - P(\mathrm{e}) \log P(\mathrm{e}) -[1- P(\mathrm{e})] \log [1- P(\mathrm{e})], \quad (16)$$

$$\inf H(V \mid X) = 2\, P(\mathrm{e}). \quad (17)$$

The question comes into being : whether it is possible to make judgment about usefulness of any feature in sense that average error probability change in reference to change of average Shannon's entropy? The answer to this question gives following

Theorem 2 : Any feature x is useful, i.e. $P(\mathrm{e}) < P_0(\mathrm{e})$ when average Shannon's entropy change so that

$$H(V) - H(V \mid X) > I_0 \quad (18)$$

where $H(V) = - P(V_1) \log P(V_1) - P(V_2) \log P(V_2)$ is the initial (unconditional) Shannon's entropy of set $V = \{V_1, V_2\}$, $H(V \mid X)$ is the average conditional Shannon's entropy, which estimates the uncertainty of the set V after testing any possible value of this feature and

$$I_0 = \log(1+ \lambda_0) - \lambda_0 [1+ \lambda_0]^{-1} \log \lambda_0 -$$

$$- 2 \min\{[1+ \lambda_0]^{-1}, \lambda_0 [1+ \lambda_0]^{-1}\} \quad (19)$$

is the so-called threshold of information depending only on the ratio of the *a priori* probabilities of classes $\lambda_0 = \mathrm{P}(\mathrm{V}_2)/\mathrm{P}(\mathrm{V}_1)$.

Proof : Let the condition (18) be valid. Then taking into account that $P_0(\mathrm{e}) = \min(P(V_1),P(V_2))$ with reference to (19) this condition can be finally regarded as

$$H(V \mid X) < 2\, P_0(\mathrm{e}). \quad (20)$$

We will show that in this case $P(\mathrm{e}) < P_0(\mathrm{e})$. To this end let us assume otherwise that $P(\mathrm{e}) = P_0(\mathrm{e})$ (the case $P(\mathrm{e}) > P_0(\mathrm{e})$ is impossible). In this assumption in accordance with (17) $\inf H(V \mid X) = 2P_0$ (e), but this contradicts to the strict inequality (20). Theorem 2 is proved.

Fig. 3 shows the relationship between threshold I_0 and ratio λ_0 when $0 \le \lambda_0 \le 1$. It is possible to use this relationship in case $\lambda_0 > 1$, because in this case according to (19) $I_0(\lambda_0) = I_0(1/\lambda_0)$ and $0 \le [\lambda_0]^{-1} \le 1$.

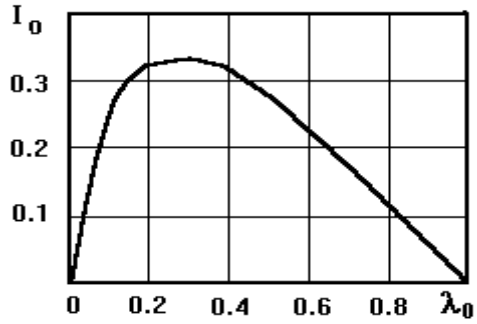

Figure 3 : Relationship between information threshold I_0 and ratio λ_0.

It should be noted here that the condition of Theorem 2 is sufficient but not necessary.

Consequence 2 : If the *a priori* probabilities of classes are equal, i.e. $P(V_1) = P(V_2) = 0.5$ then any feature x, which is differently distributed in classes is useful in the sense that $P(\mathrm{e}) < P_0(\mathrm{e})$. This fact follows immediately from (18) with reference to (19) so long as $I_0 = 0$ when $\lambda_0 = 1$ and $H(V \mid X) < H(V)$ when $p(x \mid V_1) \neq p(x \mid V_2)$.

It is interesting that the same result that Definition 1 and Definition 2 are equivalent *only when* $P(V_1) = P(V_2)$ we have proved in a different way.

4. Conclusion

In this paper it is showed that, in general, the condition $p(x \mid V_1) \neq p(x \mid V_2)$, which means that x is relevant in the sense Ben-Bassat [2] and informative in sense Lewis [6], can be regarded only as *necessary but not sufficient* condition that this feature x is useful in the sense of average error probability $P(\mathrm{e})$ change. Therefore, in general, when $P(V_1) \neq P(V_2)$ any irrelevant feature is unuseful without fail, but any relevant feature may be as useful as unuseful (see fig.4).

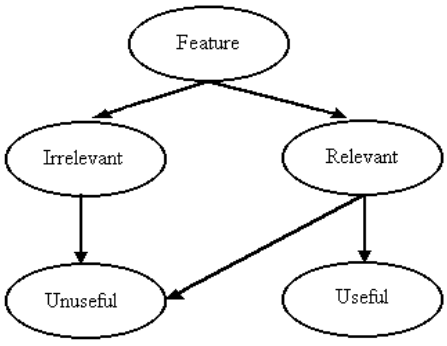

Figure 4: The interconnection between relevant and useful feature when $P(V_1) \neq P(V_2)$.

At the same time if $P(V_1) = P(V_2) = 0.5$ then, in accordance with Consequence 1 and Consequence 2, $P(e) < \min\{P(V_1), P(V_2)\}$ certainly when $p(x \mid V_1) \neq p(x \mid V_2)$. Hence, in this case any relevant feature is useful and any irrelevant one is unuseful (see fig. 5).

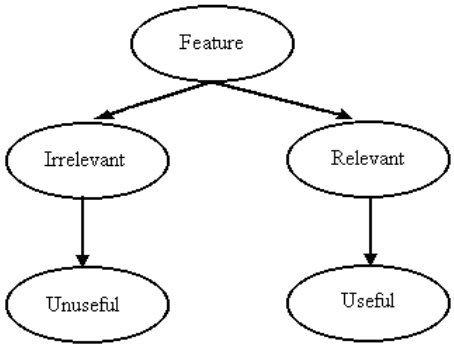

Figure 5: The interconnection between relevant and useful feature when $P(V_1) = P(V_2) = 0.5$.

We also have established the necessary and sufficient conditions which guarantee that binary feature is useful in the statistical recognition of two classes.

The conditions we have considered in this paper were applied in solving the problems of feature selection for statistical recognition of true and false thermal effects of phase transformation in cooling steel (see Fainzilberg [12]) and computer analysis of phase space electro-cardio graphic image (see Fainzilberg and Potapova [13]).

References

[1] P.Pudil, F.J.Ferri, J.Novovicova, J.Kittler," Floating Search Methods for Feature selection with Nonmonotonic Criterion Function". In Proc. of the 12-th Int. Conf. on Pattern Recognition, vol.2, pp.279-283, Jerusalem, 1994.

[2] M.Ben-Bassat, "Irrelevant Features in Pattern Recognition", IEEE Trans. Comput., vol.C-27, N8, pp. 749-766, August 1978.

[3] J.Kittler, "Feature Selection and Extraction", Handbook of Pattern Recognition and Image Processing, 1986.

[4] P.M.Narendra, K.Fukunaga "A Branch and Bound Algorithm for Feature Subset Selection", IEEE Trans. Comput., vol. C-26, pp. 917-922, Sept. 1977.

[5] W.Siedlecki, J.Sklansky, "On Automatic Feature Selection", International Journal of Pattern Recognition and Artificial Intelligence, vol. 2, pp. 197-220, 1988.

[6] P.M.Lewis, "The Characteristic Selection Problem in Recognition System", IRE Trans. Inform. Theory, vol.8, N 2, pp. 171-178, 1962.

[7] J.T.Tou, R.P. Heydorn, "Some Approaches to Optimum Feature Selection", In Computer and Information Sciences Academic Press, N 4, vol. 11,pp. 57-89, 1967.

[8] S.D.Stearns, "On Selecting Features for Pattern Classifiers", In Third Int. Conf. on Pattern recognition, pp. 71-75, Coronado, CA, 1976.

[9] L.S.Fainzilberg, "Interconnection Between Feature Properties and Probability of Error in Statistical Recognition of Two Classes", In Proc. of the 12-th Int. Conf. on Pattern Recognition, vol. 2, pp. 544-546, Jerusalem, 1994.

[10] V.A.Kovalevsky, "Optimal Decision Making in Pattern Recognition", (in Russian), 1976.

[11] L.S.Zhitetsky, L.S.Fainzilberg, "On Informative Approach to Estimation of Feature Usefulness in Statistical Pattern Recognition", Izwestija Akademii Nauk SSSR. Technicheskaja Kibernetika (in Russian), N 4, pp. 120-126, 1983.

[12] 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.

[13] L.S.Fainzilberg, T.P.Potapova, "Computer Analysis and Recognition of Cognitive Phase Space Electro-Cardio Graphics Image", In Proc. of the 6-th Int. Conf. on Computer Analysis of Images and Patterns, pp. 668-673, Prague, 1995.