Stable Feature Extraction with the Help of Stochastic Information Measure

Alexander Lepskiy

Higher School of Economics, Moscow, Russia

Abstract. This article discusses the problem of extraction of such set of pattern features that is informative and stable with respect of stochastic noise. This is done through the stochastic information measure.

1 Introduction

In describing of non deterministic system we must take into account the nature of its uncertainty. In pattern recognition uncertainty could have both probabilistic nature, that defined by some additive measure, and more "imprecise" nature, which can be described, for example, by using monotonous nonadditive measures. The feature extraction is the important task of pattern recognition in general and image analyses in particular. This task consists to extraction of some minimal set of features or combinations of them from a set of features with given informativeness, which would have sufficient integral informativeness in solution of given problem of recognition. The are traditional approaches to solving this problem, such as correlation analysis (principal components analysis, etc.), discriminant analysis [2].

In this article we concretized the general problem of informative features extraction as follows. Suppose that pattern X is defined by ordered set $X = \{x_0, ..., x_{n-1}\}$. We call every subset $A \in 2^X$ a representation of pattern X. The problem is to find such representation A that it will be minimal on the one hand and it will be near to X on the other hand. Elements in the X may have different priorities, in other words, they may have different informativeness. Therefore, we will establish correspondence between the element $x \in X$ and the nonnegative number – a feature $\omega(x)$, that characterize the importance of the element x for the representation of pattern X. We will determine the degree of representativeness of the set $A \in 2^X$ for pattern X using the set function $\mu(A)$ and require from the set function μ to satisfy all the axioms of monotonous measure: 1) $\mu(\emptyset) = 0$, $\mu(X) = 1$ (normalization); 2) $\mu(A) \leq \mu(B)$ if $A \subseteq B$, $A, B \in 2^X$ (monotonicity). In addition the measure μ meets the certain additional conditions related to the specific task of pattern recognition. We will call such measures monotonous information measures. In [1] information measures were used to set and solve the task of finding the most minimal informative polygonal representation of the curve.

In certain tasks of pattern recognition, in particular, in image processing, image analysis and image recognition random nature of image features can be

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 54–59, 2011.

caused by some noisy effects. For example, if the pattern is a discrete plane curve that extracted on the image and features are some characteristics of curve points (eg, feature is a estimation of curvature in given point of discrete curve [3]), then a random character of features (eg, curvature) is caused by the noise of image. In this case the expectation $\mathbf{E}\left[\mathrm{M}(A)\right]$ characterize the level of informativeness of represenation $A \in 2^X$ and the variance $\sigma^2\left[\mathrm{M}(A)\right]$ characterize the level of stability of representation to noise pattern. Then there is the problem of finding the most stable and informative representation $A \in 2^X$ of the pattern X. The complexity of solutions of this problem will be determined by the degree of dependence of random features of each other. Stochastic measure of informativeness M will be additive measure if the features are independent random variables. Otherwise, the measure would be nonadditive measure. In this article mentioned task will be discussed and resolved in the case of probabilistic independence of random features. The case of random dependency of features is investigated in [4].

2 The Average Set Function of Information of Pattern

In general case the feature may be depend from all or some elements from the representation A on which it calculated i.e. $\omega(x) = \omega(x, A)$. For example, let $X = \Gamma$ be a discrete plane close curve: $\Gamma = (\mathbf{g}_k)_{k=0}^{n-1}$, $\mathbf{g}_k = x_k\mathbf{i} + y_k\mathbf{j}$, and $\omega(\mathbf{g}, A) = \|\mathbf{g} - \mathbf{g}_+(A)\|$, where $\mathbf{g}_+(A)$ is a point that follows from the point $\mathbf{g}$ in ordered representation A or $\omega(\mathbf{g}, A) = k_\varepsilon\left[A\right](\mathbf{g})$ is a some estimation of curvature of plane discrete curve that is calculated in $\varepsilon-$ neighbourhood of point $\mathbf{g}$ [3].

The degree of completeness of representation $A \in 2^X$ in describing the pattern X with respect to feature $\omega(x) = \omega(x, A)$ may be assigned with the follow set function

$$\mu(A) = \sum\nolimits_{x\in A} \omega(x, A) \Big/ \sum\nolimits_{x\in X} \omega(x, X), A \in 2^X, \tag{1}$$

which we call averaged set function of information of pattern X (with respect to feature ω). We put $\mu(A) = 0$ if the value $\omega(x, A)$ is not defined for set A (in particular $\mu(\emptyset) = 0$).

Example 1. Let $X = \Gamma$ be a discrete plane close curve: $\Gamma = (\mathbf{g}_k)_{k=0}^{n-1}$, $\mathbf{g}_k = x_k\mathbf{i}+y_k\mathbf{j}$ and $B = \{\mathbf{g}_{i_1}, ..., \mathbf{g}_{i_l}\} \subseteq \Gamma$. Also we introduce the set functions $\mu_L(B) = L(B)/L(\Gamma)$ and $\mu_S(B) = S(B)/S(\Gamma)$, where $L(B)$ and $S(B)$ are perimeter and square correspondingly of figure, that is limited by polygon with vertex in points of ordered set B. Then μ_L and μ_S are averaged set functions of information of curve Γ, where $\omega(\mathbf{g}, A) = \|\mathbf{g} - \mathbf{g}_+(A)\|$ for set function μ_L and $\omega(\mathbf{g}, A) = 0.5\left|\rho_O(\mathbf{g}) \times \rho_O(\mathbf{g}_+(A))\right|$ for set function μ_S, $\rho_O(\mathbf{g})$ is a radius vector of point $\mathbf{g} \in A$ with respect to arbitrary point O. Note that μ_L is a monotonous measure, and μ_S is a monotonous measure too if the domain bounded by polygons with vertexes in the points of discrete curve Γ is convex.

There is a question when the averaged set functions of information will be monotonous measure? It can be easily shown that follow proposition is true.

Proposition 1. *Averaged set functions of information μ on 2^X of form (1) are a monotonous measure iff μ obeys the following condition*

$$\sum_{x\in A}\left(\omega(x,A)-\omega(x,A\cup\{y\})\right)\le\omega(y,A\cup\{y\})$$

for any set $A\in 2^X$ and $y\in X\backslash A$.

Corollary 1. *If $\omega(x,A)=\omega(x,X)$ for all $x\in A$ and $A\in 2^X$ then set function μ of form (1) is an additive measure on 2^X.*

Let $A=\{x_{i_1},...,x_{i_s},x,\ x_{i_{s+1}},...,x_{i_l}\}\subseteq X$ be a some representation of pattern X. We denote $A_{k,m}(x)=\left\{x_{i_{s-k+1}},...,x_{i_s},x,\ \ x_{i_{s+1}},...,x_{i_{s+m}}\right\}\subseteq X$, $k+m\le l$.

Corollary 2. *Let $\omega(x,A)=\omega(x,A_{k,m}(x))$ for all $x\in A$ and $A\in 2^X$, $|A|>k+m$, then set function μ of form (1) is a monotonous measure iff*

$$\sum_{r=0}^{k-1}\left(\omega(x_{i_{s-r}},A)-\omega(x_{i_{s-r}},A\cup\{y\})\right)+\sum_{r=1}^{m}\left(\omega(x_{i_{s+r}},A)-\omega(x_{i_{s+r}},A\cup\{y\})\right)$$
$$\le\omega(y,A\cup\{y\})$$

for all $A\in 2^X$ and $y\in X\backslash A$. In particular, let $\omega(x,A)=\omega(x,A_{0,1}(x))$ for all $x\in A$ and $A\in 2^X$, $|A|>1$, then set function μ of form (1) is a monotonous measure iff $\omega(x_+(A),A)\le\omega(x_+(A),A\cup\{y\})+\omega(y,A\cup\{y\})$ for all $A\in 2^X$ and $y\in X\backslash A$, where $x_+(A)$ is a point that follows from the point x in ordered representation A.

3 The Stochastic Additive Average Information Measure

Let μ be a averaged information measure on 2^X of form (1) and $\omega(x,A)=\omega(x,X)=\omega(x)$ for all $x\in A$ and $A\in 2^X$. In other words the feature value at point x does not depend from considered representation. The estimation of curvature $\omega(\mathbf{g},A)=k_\varepsilon\left[\Gamma\right](\mathbf{g})$ in ε- neighbourhood of point $\mathbf{g}$ for discrete plane curve $\Gamma=(\mathbf{g}_k)_{k=0}^{n-1}$ is an example of this feature. Then (see corollary 1) the measure $\mu(A)=\sum_{x\in A}\omega(x)\big/\sum_{x\in X}\omega(x)$, $A\in 2^X$, is an additive measure. We suppose that feature characteristics will be independence random variables $\Omega(x)$, $x\in X$. In this case the value of information measure $\mathrm{M}(A)=\sum_{x\in A}\Omega(x)\big/\sum_{x\in X}\Omega(x)$ for fixed set $A\in 2^X$ will be random variable too. The set function $\mathrm{M}(A)$ will be an additive measure for any random event. We considered the example of this situation. Suppose that discrete plane curve $\Gamma=(\mathbf{g}_k)_{k=0}^{n-1}$, $\mathbf{g}_k=x_k\mathbf{i}+y_k\mathbf{j}$ was subjected to additive stochastic noncorrelated noise. In result we get a random curve $\tilde{\Gamma}=(\mathbf{G}_k)_{k=0}^{p-1}$, $\mathbf{G}_k=X_k\mathbf{i}+Y_k\mathbf{j}$, where $X_k=x_k+\eta_k$, $Y_k=y_k+\xi_k$, η_k, ξ_k are random noncorrelated variables and $\mathbf{E}[\eta_k]=\mathbf{E}[\xi_k]=0$, $\sigma^2[\eta_k]=\sigma^2_{x,k}$,

$\sigma^2[\xi_k] = \sigma^2_{y,k}$. Suppose that there is such basic set $\mathrm{B} \subseteq \Gamma$ for which random variables $\Omega(\mathbf{g})$, $\mathbf{g} \in \mathrm{B}$, are independent. In this case the feature characteristics $\omega(\mathbf{G}) = \Omega(\mathbf{g})$ will be random variables just as the value of measure $\mathrm{M}(A)$, $A \in 2^{\mathrm{B}}$, is.

Let $\mathbf{E}[\Omega(x)] = m_x$, $\sigma^2[\Omega(x)] = \sigma^2_x$. We will investigate numerical characteristics of random additive measure $\mathrm{M}(A)$, $A \in 2^X$.

3.1 The Numerical Characteristics of Stochastic Additive Average Information Measure

Let us find the mathematical expectation of random variable $\mathrm{M}(A)$ with a fixed $A \in 2^X$. Random variable $\mathrm{M}(A)$ is equal to a quotient two random variables $\xi = \sum_{x \in A} \Omega(x)$ and $\eta = \sum_{x \in X} \Omega(x)$.

Lemma 1. *Let ξ and η be random variables that taking values in the intervals l_ξ, l_η respectively on positive semiaxis and $l_\eta \subseteq ((1-\delta)\mathbf{E}[\eta], (1+\delta)\mathbf{E}[\eta])$, $l_\xi \subseteq (\mathbf{E}[\xi] - \delta\mathbf{E}[\eta], \mathbf{E}[\xi] + \delta\mathbf{E}[\eta])$. Then it is valid the following formulas for mean and variance of distribution of $\frac{\xi}{\eta}$ respectively*

$$\mathbf{E}\left[\frac{\xi}{\eta}\right] = \frac{\mathbf{E}[\xi]}{\mathbf{E}[\eta]} + \frac{\mathbf{E}[\xi]}{\mathbf{E}^3[\eta]}\sigma^2[\eta] - \frac{1}{\mathbf{E}^2[\eta]}\mathbf{K}[\xi,\eta] + r_1, \tag{2}$$

$$\sigma^2\left[\frac{\xi}{\eta}\right] = \frac{1}{\mathbf{E}^2[\eta]}\sigma^2[\xi] + \frac{\mathbf{E}^2[\xi]}{\mathbf{E}^4[\eta]}\sigma^2[\eta] - \frac{2\mathbf{E}[\xi]}{\mathbf{E}^3[\eta]}\mathbf{K}[\xi,\eta] + r_2, \tag{3}$$

where $\mathbf{K}[\xi,\eta]$ is a covariation of random variables ξ and η, could meet the certain additional conditions re i.e. $\mathbf{K}[\xi,\eta] = \mathbf{E}[(\xi - \mathbf{E}[\xi])(\eta - \mathbf{E}[\eta])]$; r_1, r_2 are the residuals that depend on numerical characteristics of ξ and η and $|r_1| \le \frac{\delta}{1-\delta} \cdot \frac{\mathbf{E}[\xi]+\mathbf{E}[\eta]}{\mathbf{E}^3[\eta]}\sigma^2[\eta] \le \frac{\mathbf{E}[\xi]+\mathbf{E}[\eta]}{(1-\delta)\mathbf{E}[\eta]}\delta^3$, $|r_2| \le C\delta^3$.

Let $\xi = \sum_{x \in A} \Omega(x)$ and $\eta = \sum_{x \in X} \Omega(x)$. Then $\mathbf{K}[\xi,\eta] = \mathbf{K}\left[\sum_{x \in A} \Omega(x), \sum_{y \in X} \Omega(y)\right] = \sum_{x \in A} \mathbf{K}[\Omega(x), \Omega(x)] + \sum_{x \in A, y \in X | x \neq y} \mathbf{K}[\Omega(x), \Omega(y)]$. Because random variables $\Omega(x)$, $x \in X$, are independent, $\mathbf{K}[\Omega(x), \Omega(y)] = 0$ for $x \neq y$, and $\mathbf{K}[\Omega(x), \Omega(x)] = \sigma^2_x$ by definition. Hence, $\mathbf{K}[\xi,\eta] = \sum_{x \in A} \sigma^2_x$. Using the notation $S(A) = \sum_{x \in A} m_x$, $D(A) = \sum_{x \in A} \sigma^2_x$, we can rewrite formulas (2), (3) for our case as follows

$$\mathbf{E}[\mathrm{M}(A)] = \frac{S(A)}{S(X)} + \frac{S(A)}{S^3(X)}D(X) - \frac{1}{S^2(X)}D(A) + r_1, \tag{4}$$

$$\sigma^2[\mathrm{M}(A)] = \frac{S(X)-2S(A)}{S^3(X)}D(A) + \frac{S^2(A)}{S^4(X)}D(X) + r_2. \tag{5}$$

Notice that, $\sigma^2[\mathrm{M}(X \backslash A)] = \sigma^2[1 - \mathrm{M}(A)] = \sigma^2[\mathrm{M}(A)]$ for $A \in 2^X$. We will use formulas (4) and (5) without their residuals. Respective values $\tilde{\mathbf{E}}[\mathrm{M}(A)] = \mathbf{E}[\mathrm{M}(A)] - r_1$, $\tilde{\sigma}^2[\mathrm{M}(A)] = \sigma^2[\mathrm{M}(A)] - r_2$ we will call estimations of numerical characteristics. The class of random variables $\{\mathrm{M}(A) : A \in 2^X\}$ satisfies all the requirements for a finitely additive stochastic measure [5]: 1) $\mathbf{E}[\mathrm{M}^2(A)] < \infty$ for all $A \in 2^X$; 2) $\mathrm{M}(A)$ is an almost probably finitely additive measure.

Note that mathematical expectation $\mathbf{E}[\mathrm{M}(A)]$ define the set function on 2^X and $\mathbf{E}[\mathrm{M}(\emptyset)] = 0$, $\mathbf{E}[\mathrm{M}(X)] = 1$. Since $S(A)$ and $D(A)$ are additive set function then measure $\tilde{\mathbf{E}}[\mathrm{M}(A)]$ will be additive too.

3.2 Finding of Optimal Stable Pattern Representation

We set a problem of finding such representation B of pattern X, which cardinality is less or equal to the given number $k \geq 3$ for which the summarized variance $\sum_{A\subseteq B} \tilde{\sigma}^2[\mathrm{M}(A)]$ will be minimal but the sum of squares of mathematical expectations of all representations $\sum_{A\subseteq B} \tilde{\mathbf{E}}^2[\mathrm{M}(A)]$ will be maximal. The value of summarized variance characterizes the stability of representation and all its subset to noise level of pattern and it depends also from number of elements in representation. The more elements in the presentation we have, then the summarized value of variance is greater. We will use mathematical expectations of nonnormalized information measures $S(A)$, $A \subseteq X$, instead of $\mathbf{E}[\mathrm{M}(A)]$, $A \subseteq X$, for simplification of computations. We introduce follow criteria: $f(X) = \sum_{A\subseteq X} \tilde{\sigma}^2[\mathrm{M}(A)] \Big/ \sum_{A\subseteq X} S^2(A)$, $|X| \leq k$.

Then it is necessary to find such set B, $3 \leq |B| \leq k$, for which $f(B) \to \min$. We simplify function $f(B)$. Let $S_2(B) = \sum_{x\in B} m_x^2$, $SD(B) = \sum_{x\in B} \sigma_x^2 m_x$.

Proposition 2. *If feature characteristics of pattern X are independent random variables then following equality is true for every $B \in 2^X$:*

$$f(B) = \frac{1}{S^4(B)} \left\{ D(B) - \frac{2S(B)}{S_2(B)+S^2(B)} SD(B) \right\}.$$

Corollary 3. *If $\sigma_x^2 = \sigma^2 = const$ for all $x \in B$ then*

$$f(B) = \frac{1}{S^4(B)} \left\{ |B| - \frac{2S^2(B)}{S_2(B)+S^2(B)} \right\} \sigma^2.$$

As $S^2(B)/|B| \leq S_2(B) \leq S^2(B)$ then following corollary is true.

Corollary 4. *If $\sigma_x^2 = \sigma^2 = const$ for all $x \in B$ then*

$$\frac{|B|^2-|B|}{(|B|+1)S^4(B)}\sigma^2 \leq f(B) \leq \frac{|B|-1}{S^4(B)}\sigma^2.$$

We will use the "inclusion-exclusion" procedure for finding of optimal representation which minimize the function $f(B)$. We estimate the variation of function $f(B)$ to the exclusion of element x from B and inclusion of element $y \in X\backslash B$.

Theorem 1. *If feature characteristics of pattern X are independent random variables then following asymptotic equality*

$$f((B\backslash\{x\}) \cup \{y\}) - f(B) = \frac{1}{S^4(B)} \left(Q_1(B)(m_y - m_x) + \sigma_y^2 - \sigma_x^2 \right) + o(\tau),$$

is true for every $x \in B$ and $y \in X\backslash B$, where $Q_1(B) = \frac{2\left(3S_2(B)+5S^2(B)\right)}{\left(S_2(B)+S^2(B)\right)^2} SD(B) - \frac{4}{S(B)} D(B)$, $\tau = \sqrt{\frac{1}{S^2(B)}\left(m_y^2 + m_y^2\right) + \frac{1}{D^2(B)}\left(\sigma_x^4 + \sigma_y^4\right)}$.

Corollary 5. *If* $\sigma_x^2 = \sigma^2 = const$ *for all* $x \in X$, *then for any* $x \in B$ *and* $y \in X \backslash B$ *we have* $f((B\backslash\{x\}) \cup \{y\}) - f(B) = Q_2(B)\sigma^2 (m_y - m_x) + o(\tau)$, *where* $Q_2(B) = \frac{2}{S^3(B)} \left(\frac{3S_2(B)+5S^2(B)}{(S_2(B)+S^2(B))^2} - \frac{2|B|}{S^2(B)} \right)$, $Q_2(B) < 0$ *for all* $B \subseteq X : \ |B| \geq 3$ *and* $\tau = \sqrt{\frac{1}{S^2(B)} \left(m_y^2 + m_y^2\right) + \frac{1}{D^2(B)} \left(\sigma_x^4 + \sigma_y^4\right)}$.

Formulas from theorem 1 and corollary 5 may be use for construction of algorithmic procedures for finding of representation B, which minimize the function f. Let $K_B(x, y) = Q_1(B)(m_y - m_x) + \sigma_y^2 - \sigma_x^2$. The algorithm for finding of representation B of cardinality k which minimize the function f (if random variables $\Omega(x)$, $x \in X$, are independent), consist of following steps:

1) we select the set B_0 consisting of k elements of pattern X with maximal values of informativeness $\mathbf{E}[\Omega(x)] = m_x$, $x \in X$, as a initial representation;

2) we compute the value $Q_1(B_0)$ from Theorem 1 and we find $(\tilde{x}, \tilde{y}) = \arg\min \{K_{B_0}(x, y) : \ x \in B, \ y \in X\backslash B, \ K_{B_0}(x, y) < 0\}$, if this pair is exist. Then $B_1 = (B_0\backslash\{\tilde{x}\}) \cup \{\tilde{y}\}$ is a new representation for which $f(B_1) \leq f(B_0)$ accurate within to small values of second order. This step will be repeating so long as will be pairs (x, y): $K_B(x, y) < 0$.

If $\sigma_x^2 = \sigma^2 = const$ for all $x \in X$, then following conclusion follows from the corollary 5, if we disregard small values: the optimal representation B with cardinality is less or equal k, which minimize the function f will be consist of elements with the greatest values $\mathbf{E}[\Omega(x)] = m_x$ (on the assumption of random variables $\Omega(x)$, $x \in X$, are independent).

Acknowledgement. This work was supported by the grants 10-07-00135, 10-07-00478, 11-07-00591 of RFBR (Russian Foundation for Basic Research).

References

[1] Bronevich, A., Lepskiy, A.: Geometrical Fuzzy Measures in Image Processing and Pattern Recognition. In: Proc. of the 10th IFSA World Congress, Istanbul, Turkey, pp. 151–154 (2003)

[2] Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification and Scene Analysis: Part I Pattern Classification. John Wiley and Sons, Chichester (1998)

[3] Lepskii, A.E.: On Stability of the Center of Masses of the Vector Representation in One Probabilistic Model of Noiseness of an Image Contour. Automation and Remote Control 68, 75–84 (2007)

[4] Lepskiy, A.E.: Application of Stochastic Information Measure in Problem of Finding of Optimal Polygonal Curve Representation. In: Proc. of Intern. Conf. Pattern Recognition and Image Analysis, Nizhni Novgorod, vol. 1, pp. 397–400 (2008)

[5] Shiryaev, A.N.: Probability (Graduate Texts in Mathematics). Springer, Heidelberg (1995)