한국통신학회 1987년도 추계학술발표회 논문집 (1987. 11. 14)

상호정보를 이용한 특성 추출에 대하여

최 형 일 ° 박 상 규 *
° 숭실대학교 전산학과 * 한양대 전통과

On the Feature Selection using Information Measure

H.I. Choi° and S.K. Park*
° Soongsil Univ. * Han-Yang Univ

1. Introduction.

Probably the most important aspect of pattern recognition is that of feature selection. With proper and efficient feature selection, both simple and sophisticated classification algorithms can be more easily implemented owing to the large dimensionality reduction provided by the feature selection process. In addition, the selected features may show better class separability and possibly may allow simpler decision surfaces.

To select the best set of m features from a collection of N features, one needs a measure which indicates the effectiveness of the set of features for classification purposes. As such a measure the *mutual information* has been widely used [1,2].

However, to obtain the mutual information between a source of classes and a set of m features, one has to estimate underlying m-th order joint distributions which requires a lot of memory storage and computation time.

In this paper, we suggest a new approach which approximates m-th order mutual information as the sum of second order interactions. This approach utilizes *divergence* and *expected divergence* measure [3] with the aid of maximum weight dependence tree searching method [4].

The organization of this paper is as follows. In section 2, we interpret the mutual information as the measure evaluating feature effectiveness. In section 3, we propose an approach to approximate m-th order mutual information as the sum of second order interactions, and compare the accuracy of the proposed method with those of other approximating methods appearing in the literature.

2. The mutual information.

Suppose there is a source of classes $\mathbf{S} = \{C_1,\ldots,C_K\}$ such that the source provides features from one of K classes sequentially. Assuming a zero memory source, let $P(S_i)$ describe the distribution of the i-th class of the source. Then the average uncertainty of the source is known as the *source entropy*, and is defined as

$$H(\mathbf{S}) = -\sum_{i=1}^{K} P(S_i)\cdot \log P(S_i) \qquad (1)$$

The value of this entropy is greater than or equal to zero, and is maximum when $P(S_i) = 1/K$ for all i.

Let us next consider what can be learned about the source of classes by observing a set of m features $\mathbf{F} = \{F_1, \ldots, F_m\}$; each F_i is quantized to a total of L_i quantum levels and Ω represents an sample space of the set of random variables $\mathbf{F}$. Then the *conditional entropy* of the source given an observation of m features $\mathbf{F}$ becomes

$$H(S|F) = -\sum_{i=1}^{K} \sum_{\Omega} P(S_i, F) \cdot \log P(S_i|F) \quad (2)$$

Then the mutual information between the source **S** and the features **F** is defined as

$$I(S:F) = H(S) - H(S|F) = H(F) - H(F|S) \quad (3)$$

That is, $I(S:F)$ is the amount of information gained by observing the given set of features **F**. In other words, the uncertainty of the source has been reduced by the amount of the conditional entropy.

This mutual information measure has many interesting properties. One of them is that it provides tight upper and lower bounds to the probability of misclassification [1,5]. Thus one can roughly say that for given two sets of features **F** and **F'**, if $I(S:F) > I(S:F')$ then feature set **F** is more effective than feature set **F'** for the classification purpose.

3. Approximation of the mutual information.

We suggest an approach for approximating m-th order mutual information as the sum of lower order interactions, using divergence and expected divergence measure with the aid of maximum weight dependence tree searching method.

Lewis [6] suggested using the divergence as a measure of closeness of approximation. One of the important properties of this measure is [3]

$$D(P : P_a) \geq \sum_{\Omega} P(F) - \sum_{\Omega} P_a(F) \quad (4)$$

where P_a is a approximated version of P, and equality holds when $P_a(F) = P(F)$ for all **F**. Furthermore, if $\sum P_a(F) = 1$, then $D(P : P_a) \geq 0$.

According to the chain rule, a joint probability distribution can be expressed as a chain product of conditional probabilities.

$$P(F) = P(F_1)P(F_2|F_1)\cdots P(F_m|F_1 \cdots F_{m-1}) \quad (5)$$

Chow [4] suggested approximating this probability by imposing a limit on the maximum number of variables upon which each variable may be conditioned. That is, in the 2-nd order approximation,

$$P_a(F) \equiv \prod_{i=1}^{m} P(F_{n(i)}|F_{n(j(i))}), \quad 0 \leq j(i) < i \quad (6)$$

where $(n(1), \cdots, n(m))$ is an unknown permutation of integers $1, \cdots, m$, and $P(F_i|F_0)$ is by definition equal to $P(F_i)$. The accuracy of $P_a(F)$ depends on how to choose m-1 pairs of $(n(i),n(j(i)))$s in (5). For notational convenience, $F_{n(i)}$ and $F_{n(j(i))}$ will be denoted by F_i and $F_{j(i)}$, respectively.

We now express $D(P : P_a)$ as follows,

$$D(P : P_a) = -H(F) + \sum_{i=1}^{m} H(F_i) - \sum_{i=1}^{m-1} I(F_i : F_{j(i)}) \quad (7)$$

where $I(F_i : F_{j(i)})$ is by definition equal to zero.

Since $H(F)$ and $\sum H(F_i)$ do not depend on a specific pairs of $(i,j(i))$s, minimizing $D(P : P_a)$ is equivalent to maximizing $\sum I(F_i : F_{j(i)})$.

To find m-1 pairs of $(i,j(i))$s which maximize $\sum I(F_i, F_{j(i)})$, Chow [4] suggested following algorithm where $I(F_i : F_{j(i)})$ is considered as a weight of $(i,j(i))$.

1) Order all the mutual information pairs in the order of decreasing weight.
2) Select the first and 2-nd pairs.
3) Add next pair if it does not form a *cycle* with previously selected ones, and reject it otherwise.
4) Stop processing when m-1 pairs are selected.

Once we find such m-1 pairs, Equation (4) and (7) yields approximated $H(\mathbf{F})$ as follows;

$$H(\mathbf{F}) \approx H_a(\mathbf{F}) = \sum_{i=1}^{m} H(F_i) - \sum_{i=1}^{m-1} I(F_i : F_{j(i)}) \quad (8)$$

We next approximate the conditional entropy in a similar way using expected divergence measure. We first define $P(\mathbf{F}|S_i)$ and $P_a(\mathbf{F}|S_i)$ as in (6)

$$Q_i \equiv P(\mathbf{F}|S_i) = P(F_1 \cdots F_m|S_i) \quad (9)$$

$$Q_{ai} \equiv P_a(\mathbf{F}|S_i) = \prod_{i=1}^{m} P(F_{I(i)}|F_{I(j(i))}) \quad (10)$$

For notational convenience, $F_{I(i)}$ and $F_{I(j(i))}$ will be denoted by F_i and $F_{j(i)}$, respectively. Then the expected divergence between Q and Q_a becomes

$$E[\mathbf{D}(Q : Q_a)] \geq 0 \quad (11)$$

with equality when

$$P(\mathbf{F}|S_i) = P_a(\mathbf{F}|S_i) \quad \text{for all } \mathbf{F} \text{ and } i$$

On the other hand, one can also show that

$$E[\mathbf{D}(Q : Q_a)] \quad (12)$$
$$= -H(\mathbf{F}|\mathbf{S}) + \sum_{i=1}^{m} H(F_i|\mathbf{S}) - \sum_{i=1}^{m-1} I(F_i : F_{j(i)}|\mathbf{S})$$

where $I(F_1 : F_0|\mathbf{S})$ is by definition equal to zero.

Since $H(\mathbf{F}|\mathbf{S})$ and $\sum H(F_i|\mathbf{S})$ do not depend on a specific pairs of $(i,j(i))$s, minimizing $\mathbf{D}(Q : Q_a)$ is equivalent to maximizing $\sum I(F_i : F_{j(i)}|\mathbf{S})$. To find m-1 pairs of $(i,j(i))$s which maximize $\sum I(F_i, F_{j(i)}|\mathbf{S})$, we can utilize Chow's algoritnm in a similiar way. Once we find such m-1 pairs, Equation (11) and (12) yields approximated $H(\mathbf{F}|\mathbf{S})$ as follows;

$$H(\mathbf{F}|\mathbf{S}) \approx H_a(\mathbf{F}|\mathbf{S}) \quad (13)$$
$$= \sum_{i=1}^{m} \left(H(F_i) - H(F_i|\mathbf{S}) \right)$$
$$- \sum_{i=1}^{m-1} \left(I(F_i : F_{j(i)}) - I(F_i : F_{j(i)}|\mathbf{S}) \right)$$

Finally, by combining Equation (8) and (13), we can obtain the approximated mutual information.

$$I_a(\mathbf{F} : \mathbf{S}) = H_a(\mathbf{F}) - H_a(\mathbf{F}|\mathbf{S}) \quad (14)$$

We next evaluate the proposed approximating method using synthetically generated binary distributions, and compare the accuracy of this method with those of Michael and Lin's method [7] and Chen's method [6]. The following table lists conditional binary distributions $\mathbf{F} = (F_1 \cdots F_4)$, given class information $\mathbf{S} = (S_1 S_2)$ with equiprobable distribution. $I(\mathbf{S}:\mathbf{F})$ represents the true mutual information, $I_{a1}(\mathbf{S}:\mathbf{F})$ approximated one using the proposed method, $I_{a2}(\mathbf{S}:\mathbf{F})$ using Michael and Lin's method; $I_{a3}(\mathbf{S}:\mathbf{F})$ using Chen's method.

This experiment clearly shows that the proposed method yields more accurate results than the others. We expect that the superiority of the method would become more significant as the number of variables m increases.

reference

[1] C.H. Chen, "Theoretical Comparison of a class of Feature Selection of in Pattern Recognition," IRE T. on inform. Theory, pp171-178, Feb. 1962.

[2] M. Mata, "On the Use of I-Divergence for Generating Distribution Approximation," IEEE T. on PAMI pp661-664, Nov. 1983.

[3] R. Ash, *Information Theory*, New York: Interscience, 1965.

[4] C.K. Chow and C. Lin, "Approximating Discrete Proabability Distributions with Dependence Tree," IEEE T. on IT, pp462-467, May, 1968

[5] M. Hellman, "Probability of Error, Equivocation, and the Chernoff Bound," IEEE T. on IT, pp368-372, July 1970

[6] P.M. Lewis, "Approximating Probability Distributions to Reduce Storage Requirements," Information and Control, pp214-225, Sep. 1959

[7] C.H. Chen, "On a Class of Computationally Efficient Feature Selection Criterion," Pattern Recognition, vol 7, pp87-94, 1975

[8] M. Hellman and J. Ravis, "Probability of Error, Equivocation, and Chernoff Bound," IEEE T. on Inform. Theory, pp368-372, July 1970

$F_1F_2F_3F_4$	$P(F\|S_1)$	$P(F\|S_2)$
0 0 0 0	0.30	0.01
0 0 0 1	0.15	0.02
0 0 1 0	0.07	0.03
0 0 1 1	0.08	0.01
0 1 0 0	0.05	0.05
0 1 0 1	0.01	0.10
0 1 1 0	0.03	0.05
0 1 1 1	0.04	0.07
1 0 0 0	0.06	0.02
1 0 0 1	0.04	0.02
1 0 1 0	0.10	0.01
1 0 1 1	0.01	0.12
1 1 0 0	0.02	0.04
1 1 0 1	0.02	0.14
1 1 1 0	0.01	0.01
1 1 1 1	0.01	0.30

$I(S:F) = 0.4713$

$I(S:F) - I_{a1}(S:F) = 0.0177$

$I(S:F) - I_{a2}(S:F) = 0.0778$

$I(S:F) - I_{a3}(S:F) = -0.0197$