FEATURE SELECTION BASED ON SURVIVAL CAUCHY-SCHWARTZ MUTUAL INFORMATION

Badong Chen[1], Xiaohan Yang[1], Hua Qu[1], Jihong Zhao[1], Nanning Zheng[1], Jose C. Principe[2]

1. School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
2. Department of Electrical and Computer Engineering, University of Florida, Gainesville, USA
(chenbd@mail.xjtu.edu.cn, principe@cnel.ufl.edu)

ABSTRACT

Feature selection techniques play a crucial role in machine learning tasks such as regression and classification. Many filter methods of feature selection are based on the mutual information (e.g. MIFS, MIFS-U, NMIFS, and mRMR methods). In this work, a new mutual information is defined based on the cross survival information potential (CSIP) and Cauchy-Schwartz divergence (CSD), called the survival Cauchy-Schwartz mutual information (SCS-MI). We apply this new mutual information to select an informative subset of features for a SVM classifier. Experimental results illustrate the desirable performance of the new method.

Index Terms— Feature selection, survival information potential, Cauchy-Schwartz divergence, classification

1. INTRODUCTION

Selecting an informative subset of candidate features is very important in machine learning since it has a crucial impact on the computational cost and generalization performance of the learning algorithms. The feature selection techniques in classification can be, in general, divided into approaches that are classifier-dependent ("wrapper" or "embedded" methods) and classifier-independent ("filter" methods). The filter methods define a heuristic *scoring criterion* to evaluate the *relevance* of the data independently of any particular classifier. Many feature selection criteria in the literature are designed based on the fundamental concept of mutual information (MI) [1-6]. Several typical examples are listed in Table 1, where C denotes the class label, S denotes the set of currently selected features, f denotes a candidate feature that is not selected so far, β is a control parameter, H and I denote, respectively, Shannon's entropy and mutual information [7]:

$$H(X) = -\int p_X(x) \log p_X(x) dx \qquad (1)$$

Table 1 Several mutual information based feature selection criteria

Criterion	Criterion function $J(f)$
MIFS [2]	$I(C;f) - \beta \sum_{s \in S} \left(I(s;f) \right)$
mRMR [4]	$I(C;f) - \frac{1}{\lvert S \rvert} \sum_{s \in S} \left(I(s;f) \right)$
MIFS-U [3]	$I(C;f) - \beta \sum_{s \in S} \left(\frac{I(f;s)}{H(s)} I(C;s) \right)$
mMIFS-U [5]	$I(C;f) - \max_{s} \left(\frac{I(f;s)}{H(s)} I(C;s) \right)$
NMIFS [6]	$I(C;f) - \sum_{s \in S} \left(\frac{I(f;s)}{\min\{H(f), H(s)\} \lvert S \rvert} \right)$

$$I(X;Y) = \int p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} dxdy \qquad (2)$$

where p_X , p_Y , and p_{XY} denote the corresponding marginal and joint probability densities (or the probability masses for discrete variables).

In recent years, some new definitions of entropy and mutual information are proposed based on cumulative distribution functions or survival functions of the random variables, such as the cumulative residual entropy (CRE) [8,9], cross cumulative residual entropy (CCRE) [8,9], survival exponential entropy [10], and survival information potential (SIP) [11]. Compared with the traditional definitions, these new definitions have some merits such as the validity in a wide range of distributions, robustness, and the simplicity in computation. The CCRE has been successfully used in image registration [8,9], and the SIP finds applications in adaptive systems training [11].

In this paper, we define a new mutual information, called the survival Cauchy-Schwartz mutual information (SCS-MI), based on the cross SIP (CSIP) and Cauchy-Schwartz divergence (CSD) [12]. The proposed mutual information can be easily estimated from samples (just by comparing the data values and carrying out a multiplication), without the choice of any free parameters. This new mutual information is then applied to select an informative subset of features for a SVM classifier, and the experimental results confirm its good performance in feature selection .

This work was supported by National Natural Science Foundation of China (No. 61372152).

2. SURVIVAL CAUCHY-SCHWARTZ MI

2.1. Definitions

Before presenting the definition of SCS-MI, we give the definitions of the cross survival information potential (CSIP) and the survival Cauchy-Schwartz divergence (SCSD).

Let X and Y be two non-negative random variables with the same dimension, $X, Y \in \mathbb{R}_+^m$. Denote $\bar{F}(.)$ and $\bar{G}(.)$ respectively, the survival functions of X and Y,

$$\bar{F}(x) = P(X > x) = P(X_1 > x_1, \cdots, X_m > x_m) \tag{3}$$

where $X = (X_1, X_2, \cdots, X_m)$, and $x = (x_1, \cdots, x_m)$. Then the CSIP between X and Y (or $\bar{F}$ and $\bar{G}$) is defined by

$$S_c(X,Y) = S_c(\bar{F}, \bar{G}) = \int_{\mathbb{R}_+^m} \bar{F}(x)\bar{G}(x)dx \tag{4}$$

If the distributions of X and Y are identical, the CSIP equals the quadratic SIP (QSIP) [11]:

$$S_c(X,Y) = S_c(X,X) = \int_{\mathbb{R}_+^m} \bar{F}^2(x)dx \tag{5}$$

In addition, the following equality holds:

$$S_c(X,Y) = E\left(\prod_{i=1}^m \min(X_i, Y_i)\right) \tag{6}$$

where E denotes the expectation operator. The above equality can be easily proved as follows:

$$\begin{aligned}
&\bar{F}(x)\bar{G}(x) \\
&= P(X_1 > x_1, \cdots, X_m > x_m) \times P(Y_1 > x_1, \cdots, Y_m > x_m) \\
&= P(\min(X_1, Y_1) > x_1, \cdots, \min(X_m, Y_m) > x_m)
\end{aligned}$$

And hence

$$\begin{aligned}
S_c(X,Y) &= \int_{\mathbb{R}_+^m} \bar{F}(x)\bar{G}(x)dx \\
&= \int_{\mathbb{R}_+^m} P(\min(X_1, Y_1) > x_1, \cdots, \min(X_m, Y_m) > x_m)dx \\
&= \int_{\mathbb{R}_+^m} E\left[\prod_{i=1}^m \mathbb{I}(\min(X_i, Y_i) > x_i)\right]dx \\
&= E\left[\int_{\mathbb{R}_+^m} \left(\prod_{i=1}^m \mathbb{I}(\min(X_i, Y_i) > x_i)\right)dx\right] \\
&= E\left(\prod_{i=1}^m \min(X_i, Y_i)\right)
\end{aligned} \tag{7}$$

where $\mathbb{I}(.)$ denotes the indicator function.

Based on the CSIP, we define the survival Cauchy-Schwartz divergence (SCSD) between X and Y as

$$\begin{aligned}
D_{SCS}(X,Y) &= -\log\left(\frac{S_c(X,Y)}{\sqrt{S_c(X,X)S_c(Y,Y)}}\right) \\
&= -\log\left(\frac{\int_{\mathbb{R}_+^m} \bar{F}(x)\bar{G}(x)dx}{\sqrt{\int_{\mathbb{R}_+^m} \bar{F}^2(x)dx \int_{\mathbb{R}_+^m} \bar{G}^2(x)dx}}\right)
\end{aligned} \tag{8}$$

which is in form identical to the Cauchy-Schwartz divergence (CSD) defined in [12]. By Cauchy-Schwartz inequality, we have

$$D_{SCS}(X,Y) \geq 0 \tag{9}$$

where equality holds if and only if $\bar{F}(x) = \gamma\bar{G}(x)$ for a constant scalar γ. As $\bar{F}(0) = \bar{G}(0) = 1$, we conclude that $D_{SCS}(X,Y) = 0$ if and only if $\bar{F}(x) = \bar{G}(x)$.

Suppose now $X \in \mathbb{R}_+^{m_1}$, $Y \in \mathbb{R}_+^{m_2}$. Based on the SCSD, we define the survival Cauchy-Schwartz mutual information (SCS-MI) between X and Y as

$$\begin{aligned}
I_{SCS}(X,Y) &= D_{SCS}(\bar{H}, \bar{F}\bar{G}) \\
&= -\log\left(\frac{S_c(\bar{H}, \bar{F}\bar{G})}{\sqrt{S_c(\bar{H}, \bar{H})S_c(\bar{F}\bar{G}, \bar{F}\bar{G})}}\right) \\
&= -\log\left(\frac{\int_{\mathbb{R}_+^m} \bar{H}(z)\bar{F}(x)\bar{G}(y)dz}{\sqrt{\int_{\mathbb{R}_+^m} \bar{H}^2(z)dz \int_{\mathbb{R}_+^m} \bar{F}^2(x)\bar{G}^2(y)dz}}\right)
\end{aligned} \tag{10}$$

where $Z = (X,Y) \in \mathbb{R}_+^m$, $m = m_1 + m_2$, and $\bar{H}(.)$ is the survival function of Z. Clearly, we have $I_{SCS}(X,Y) \geq 0$, with equality if and only if X and Y are independent (i.e. $\bar{H} = \bar{F}\bar{G}$).

2.2. Estimators

We can easily estimate the CSIP, SCSD, and SCS-MI from the sample data. Given N sample data of $X \in \mathbb{R}_+^m$, $\{x(1), x(2), \cdots, x(N)\}$, the survival function of X can be estimated as

$$\begin{aligned}
\hat{\bar{F}}(x) &= \int_{\bar{\Omega}(x)} \left(\frac{1}{N}\sum_{i=1}^N \delta(\tau - x(i))\right)d\tau \\
&= \frac{1}{N}\sum_{i=1}^N (\bar{u}(x - x(i)))
\end{aligned} \tag{11}$$

where $\bar{\Omega}(x) = \{\xi \in R_+^m : \xi_1 > x_1, \cdots, \xi_m > x_m\}$, $\delta(.)$ is the multivariate Dirac δ function, and

$$\bar{u}(x - x(i)) = \int_{\bar{\Omega}(x)} \delta(\tau - x(i))d\tau \tag{12}$$

$$\hat{I}_{SCS}(X,Y) = -\log\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{l=1}^{N}\left(\prod_{k=1}^{m_1}\min\left(x_k(i),x_k(j)\right)\prod_{k=1}^{m_2}\min\left(y_k(i),y_k(l)\right)\right)}{\sqrt{\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m_1}\min\left(x_k(i),x_k(j)\right)\prod_{k=1}^{m_2}\min\left(y_k(i),y_k(j)\right)\right)\right\}\times\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m_1}\min\left(x_k(i),x_k(j)\right)\right)\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m_2}\min\left(y_k(i),y_k(j)\right)\right)\right\}}} \right) \quad (17)$$

Substituting (11) into $S_c(X,X)$, we obtain

$$\begin{aligned}\hat{S}_c(X,X) &= \int_{\mathbb{R}_+^m} \hat{\bar{F}}^2(x)\,dx \\ &= \frac{1}{N^2}\int_{\mathbb{R}_+^m}\left(\sum_{i=1}^{N}\bar{u}\left(x-x(i)\right)\right)^2 dx \\ &= \frac{1}{N^2}\int_{\mathbb{R}_+^m}\left(\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\bar{u}\left(x-x(i)\right)\bar{u}\left(x-x(j)\right)\right)\right)dx \\ &\overset{(a)}{=} \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m}\min\left(x_k(i),x_k(j)\right)\right)\end{aligned} \quad (13)$$

where (a) follows from

$$\int_{\mathbb{R}_+^m}\left(\bar{u}\left(x-x(i)\right)\bar{u}\left(x-x(j)\right)\right)dx = \prod_{k=1}^{m}\min\left(x_k(i),x_k(j)\right) \quad (14)$$

By similar derivation, we obtain $\hat{S}_c(Y,Y)$ and $\hat{S}_c(X,Y)$, and hence, the SCSD can be estimated as

$$\begin{aligned}\hat{D}_{SCS}(X,Y) &= -\log\left(\frac{\hat{S}_c(X,Y)}{\sqrt{\hat{S}_c(X,X)\hat{S}_c(Y,Y)}}\right) \\ &= -\log\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m}\min\left(x_k(i),y_k(j)\right)\right)}{\sqrt{\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m}\min\left(x_k(i),x_k(j)\right)\right)\right\}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\prod_{k=1}^{m}\min\left(y_k(i),y_k(j)\right)\right)\right\}}}\right)\end{aligned} \quad (15)$$

Then, the SCS-MI between $X\in\mathbb{R}_+^{m_1}$ and $Y\in\mathbb{R}_+^{m_2}$ can be simply estimated as

$$\hat{I}_{SCS}(X,Y) = \hat{D}_{SCS}\left(\bar{H},\bar{F}\bar{G}\right) \quad (16)$$

The detailed expression is shown in (17) at the top of this page.

Remark: The SCS-MI has some merits: 1) it has consistent definition in the continuous and discrete domains; 2) it can be computed from sample data without density estimation and the choice of free parameters; 3) it is a more robust measure since the survival function is more regular than the density function (note that the density is computed as the derivative of the distribution).

The SCS-MI is defined only for non-negative random variables, but this will not prohibit its practical applicability, since in most practical situations, the sample data are always bounded, and one can easily obtain positive data by simple translation.

3. APPLICATION TO FEATURE SELECTION

The SCS-MI has many potential applications in areas where traditional mutual information is applied. In this work, we focus only on the feature selection problem. Specifically, one can design some new feature selection criterion based on the SCS-MI. For example, a selection criterion similar to the MIFS-U can be designed as

$$J(f) = I_{SCS}(C;f) - \beta\sum_{s\in S}\left(\frac{I_{SCS}(f;s)}{S_c(s,s)}I_{SCS}(C;s)\right) \quad (18)$$

which we refer to as the "SCS- MIFS-U" criterion.

In the following, we evaluate the performance of the selected features induced by different criteria through conducting experiments on two data sets: Pima Indians Diabetes Data Set [13], and Heart Disease Data Set [14], which are set in the UC-Irvine repository. Table 2 lists the brief information of the two data sets. The performance of the SCS- MIFS-U criterion is compared with the results of MIFS, MIFS-U, mRMR, mMIFS-U, and NMIFS (see Table 1 for details). In all cases, the control parameter β for MIFS , MIFS-U and SCS-MIFS-U was experimentally set at 0.8.

Table 2 Brief Information of the Data Sets Used

Dataset	Feature num	Sample num	Classes
Pima	8	768	2
Heart	13	303	2

We consider the Support Vector Machine (SVM) as the classifier to evaluate the selected feature subsets and show the effectiveness of the new criterion. In the experiments we use the LIBSVM package, which supports both 2-class and multiclass classification. Both data sets used were split into two disjoint sets: training (70%), and testing(30%).

1) Pima Indians Diabetes Data Set

First of all, we normalized every input feature of this data set to have the values in [0, 1]. Table 3 shows the rates of correct classification obtained by SVM. It compares the performance of SCS- MIFS-U for the entire range of feature selection with the performance of other five criteria. The bold numbers in the table indicate that this criterion performs better than the rest of the criteria. As one can see clearly, the SCS-MIFS-U gets slightly higher classification accuracy than other criteria except only when the number of the selected features is 6.

Table 3 Correct Classification Rates for Pima Indians Diabetes Data Set (%)

Num of Selected Feature	MIFS	MIFS-U	mRMR	NMIFS	mMIFS_U	SCS-MIFS-U
1	64.94	64.94	64.94	64.94	64.94	**65.80**
2	65.37	65.80	65.37	64.94	65.80	**74.46**
3	67.10	68.40	67.10	**77.06**	76.19	**77.06**
4	68.40	78.79	68.40	77.49	**78.79**	**78.79**
5	68.83	78.36	68.83	**78.79**	**78.79**	78.36
6	70.56	77.49	70.56	**80.52**	**80.52**	**80.52**
7	78.36	78.36	78.36	79.22	**79.65**	**79.65**
ALL(8)	79.65					

Table 4 Correct Classification Rates for Heart Disease Data Set (%)

Num of Selected Feature	MIFS	MIFS-U	mRMR	NMIFS	mMIFS_U	SCS-MIFS-U
1	54.35	54.35	54.35	54.35	54.35	**73.91**
2	71.74	**73.91**	71.74	71.74	**73.91**	**73.91**
3	71.74	76.09	71.74	**81.52**	76.09	79.35
4	71.74	75.00	70.65	76.09	75.00	**78.26**
5	69.57	78.26	73.91	78.26	78.26	**79.35**
6	70.65	**78.26**	76.09	**78.26**	**78.26**	**78.26**
7	**78.26**	**78.26**	76.09	**78.26**	**78.26**	**78.26**
8	72.83	76.09	77.17	76.09	77.17	**78.26**
9	76.09	76.09	76.09	76.09	78.26	**80.43**
10	**78.26**	**78.26**	**78.26**	**78.26**	**78.26**	**78.26**
11	77.17	77.17	77.17	75.00	**80.43**	77.17
12	75.00	75.00	75.00	75.00	**78.26**	76.09
ALL(13)	76.09					

2) Heart Disease Data Set

Table 4 reports the correct classification rates for different numbers of the selected features on Heart Disease data set. It shows that SCS- MIFS-U performs better than other criteria in most cases.

4. CONCLUSION

Mutual information as a heuristic measure of relevance has been broadly used in feature selection. In this paper, a new mutual information, called the survival Cauchy-Schwartz mutual information (SCS-MI), is define based on the cross survival information potential (CSIP) and Cauchy-Schwartz divergence (CSD). This mutual information can be directly estimated from sample data without density estimation and choice of free parameters. Experimental results on two data sets with SVM classifier suggest that the feature selection criteria based on SCS-MI may perform very well in classification tasks.

REFERENCES

[1] G. Brown, A. Pocock, M. J. Zhao, and M. Luján, "Conditional likelihood maximization: A unifying framework for information theoretic feature selection, " *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.

[2] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.

[3] N. Kwak and C.-H. Choi, "Input feature selection for classification problems," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 143–159, Jan. 2002.

[4] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[5] J. Novovicova, P. Somol, M. Haindl, P. Pudil, "Conditional Mutual Information Based Feature Selection for Classification Task," *Progress in Pattern Recognition, Image Analysis and Applications, Lecture Notes in Computer Science*, vol. 4756, pp. 417-426, 2007.

[6] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized Mutual Information Feature Selection, " *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, 2009.
[7] Cover, T. M., Thomas, J. A., *Element of Information Theory*, Chichester, U.K.: Wiley, 1991.
[8] M. Rao, Y. Chen, B. C. Vemuri, and F. Wang, "Cumulative residual entropy: A new measure of information," *IEEE Trans. Inf. Theory*, vol.50, no. 6, pp. 1220–1228, 2004.
[9] F. Wang, B. C. Vemuri, "Non-rigid multi-modal image registration using cross-cumulative residual entropy, " *International Journal of Computer Vision*, vol. 74, no.2, pp. 201–215, 2007.
[10] K. Zografos and S. Nadarajah, "Survival exponential entropies," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 1239–1246, 2005.
[11] B. Chen, P. Zhu, and J. C. Principe, "Survival Information Potential: A New Criterion for Adaptive System Training, " *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1184–1194, 2012.
[12] Principe, J. C., *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*, Springer, New York, 2010.
[13] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
[14] http://archive.ics.uci.edu/ml/datasets/Heart+Disease