A Novel Entropy-Based Approach to Feature Selection

Chia-Hao Tu and Chunshien Li(✉)

Laboratory of Intelligent Systems and Applications, Department of Information Management, National Central University, Taoyuan City, Taiwan
jamesli@mgt.ncu.edu.tw

Abstract. The amount of features in datasets has increased significantly in the age of big data. Processing such datasets requires an enormous amount of computing power, which exceeds the capability of traditional machines. Based on mutual information and selection gain, the novel feature selection approach is proposed. With Mackey-Glass, S&P 500, and TAIEX time series datasets, we investigated how good the proposed approach could perform feature selection for a compact subset of feature variables optimal or near optimal, through comparing the results by the proposed approach to those by the brute force method. With these results, we determine the proposed approach can establish a subset solution optimal or near optimal to the problem of feature selection with very fast calculation.

Keywords: Feature selection · Probability density estimation · Information entropy · Time series dataset

1 Introduction

In the field of machine learning, feature selection is an issue that has always received much attention, especially in the past few years, due to the amount of features in datasets increased significantly. A good feature selection method not only can select relatively important feature subsets but also can reduce the amount of calculation required for the model while maintaining forecast/estimation accuracy. Feature selection refers to using certain designated methods or rules to select an important subset of feature variables from a dataset, of which the corresponding data are to be used as training and testing data for establishing machine learning models, in the hope of producing a model with good capabilities in estimation or prediction. With rapid advances in computer and database technologies, datasets with thousands of features are now ubiquitous in pattern recognition, data mining, and machine learning [1, 5, 10, 12]. Processing such datasets requires an enormous amount of processing power, which exceeds the capability of traditional machines. The use of feature selection eliminates irrelevant, redundant, and noisy data and allows machine learning models to operate normally.

In this study, we propose a novel feature selection algorithm based on mutual information and selection gain, which uses Claude Shannon's information theory as its foundation. With selection gain, the proposed method can select features from a large amount of data to form a compact feature subset to the target variable. The proposed method is intuitive, easy to manage and explain, and is expected to be fast with good performance.

© Springer International Publishing AG 2017
N.T. Nguyen et al. (Eds.): ACIIDS 2017, Part I, LNAI 10191, pp. 445–454, 2017.
DOI: 10.1007/978-3-319-54472-4_42

Hence, we designed experiments to test the algorithm for feature selection on time series datasets, and compared the results to those by the brute force method to investigate how good the proposed algorithm could quickly achieve near optimal solution.

The rest of this paper is organized as follows. Section 2 introduces previous studies in literature on feature selection and probability distribution. Section 3 describes a novel approach to feature selection based on mutual information and selection gain. The experimental results are presented in Sect. 4. Finally, Sect. 5 concludes the paper.

2 Literature Review

Using a correlation between feature selection and machine learning procedures, John et al. [11] first categorized feature selection methods into two types: "filter" or "wrapper". Later researchers also adopted these classifications. In addition to filter and wrapper, Guyon and Elisseeff [9] proposed a new type known as "embedded". The following is a brief overview of the filter, wrapper, and embedded types:

- For filter methods, feature selection is completely an independent procedure from machine learning for computational models. The feature subsets generated through feature selection are assessed using certain metrics.
- In wrapper methods, various subsets of feature variables generated by the feature selection are inputted into machine learning models for assessment, and the best subset is kept. Loughrey [13] mentioned that wrappers are much more computationally expensive and have a risk of overfitting to the model considered.
- Embedded methods take the feature selection procedure and integrate it with the machine learning procedure. This type is only used in a few specific models.

Within filter methods, certain metrics need to be defined as evaluation functions and used to assess the feature subsets selected to decide whether to keep or eliminate them. In different applications, the selection metrics will also be different. For example, in the field of text classification applications, Forman [7] listed 12 metrics that had been used in previous studies and compared them. These metrics include: Chi-Squared, Information Gain (IG), Odds Ratio, Bi-Normal Separation (BNS), etc. Dash and Liu [4] categorized evaluation functions into four types according to metrics evaluation content: Distance Measures, Information Measures, Dependence Measures and Consistency Measures.

Shannon proposed the concept of entropy and mutual information in his 1948 paper on information theory. *Entropy* is used to measure the total amount of disordered information of a single random variable, while *mutual information* is used to measure mutual dependence between two random variables. These types of intuitive concepts have been widely used in applications across different fields. Naghibi et al. [15], Peng et al. [17], and Torkkola [20] all had used this method in their feature selection studies, obtaining good results.

A random variable is a variable whose value is subject to variations due to chance. It can take on a set of possible different values, each with an associated probability, in contrast to other mathematical variables. A random variable can be either discrete or

continuous, depending on its legal values. A probability distribution, usually defined by a mathematical function, describes all possible events within a given range and their likelihoods, with the constraint that the integration (or summation) of likelihoods of all possible events must be equal to one. If a random variable is continuous, the distribution function is called a probability density function. Otherwise, it is called a probability mass function (or simply probability function). In the real world, probability distribution can be estimated using observations.

The methods of density estimation can be classified into parametric and non-parametric estimation methods. In parametric estimation methods, the probability density function is assumed to be given, and the parameter values are unknown. Hence, the parameter values are estimated. When the assumptions are correct, parametric methods will produce more accurate and precise estimates than non-parametric methods. However, if the assumptions are wrong, a larger error will result. Parzen [16] and Rosenblatt [18] proposed the kernel density estimation method, also known as the Parzen window estimation, a non-parametric estimation method that is commonly used. In the non-parametric estimation methods, only the assumption that is *similar inputs have similar outputs* was made, where selecting the probability density function is not required. This can avoid false assumption when used in probability density function, required by parametric estimation methods.

Although selecting the kernel function is required in the kernel density estimation, the final curve shape is not closely correlated with the selected kernel function [6], which is due to the fact that the neighboring wave crests generated by the kernel density estimation may be synthesized. Considering the ease of use of a function of wave synthesis, generally the Gauss function is used as the kernel function. The kernel density estimation method is very widely used. In the studies of feature selection, Alibeigi et al. [2], Azmandian et al. [3], Geng and Hu [8], Supriyanto et al. [19] and Zhang and Wang [21] had developed some feature selection methods based on the concept of kernel density estimation.

3 Proposed Approach

3.1 Entropy-Based Information

Assume that there are n discrete events as possible results for a random variable X, denoted as $\{x_i, i = 1, 2, \ldots, n\}$. The corresponding probability of each event is denoted as $\{p(x_i), i = 1, 2, \ldots, n\}$. In 1948, Shannon defined the concept of *entropy* to measure the amount of uncertainty. If additional information to X is added, the corresponding entropy will be decreased, because uncertainty of X is subsided. Thus, the change of entropy of X can be viewed as additional information to X, and vice versus. For a discrete random variable X, the entropy is denoted by $H(X)$, defined below.

$$H(X) = -\sum\nolimits_{x_i \in U_X} p(x_i) \log p(x_i) \tag{1}$$

If there are two discrete random variables X and Y, the joint entropy can be defined below for the total amount of uncertainty between the two random variables, denoted as $H(X, Y)$.

$$H(X,Y) = -\sum_{x_i \in U_X} \sum_{y_j \in U_Y} p(x_i, y_j) \log p(x_i, y_j) \tag{2}$$

The joint entropy must be not greater than the sum of the individual entropies of the two variables, as shown below.

$$H(X,Y) \leq H(X) + H(Y) \tag{3}$$

If X is already known, then conditional entropy can be used to calculate the amount of uncertainty for Y, denoted as $H(Y|X)$.

$$H(Y|X) = -\sum_{x_i \in U_X} \sum_{y_j \in U_Y} p(x_i, y_j) \log p(y_j|x_i) \tag{4}$$

The relationship for joint entropy, and conditional entropy can be written below.

$$\begin{aligned} H(Y|X) &= H(X,Y) - H(X) \\ \text{or } H(X,Y) &= H(X) + H(Y|X) \end{aligned} \tag{5}$$

From Eqs. (3) and (5), the following result can be established.

$$H(Y) \geq H(Y|X) \tag{6}$$

From formula (6), if X is known, the amount of uncertainty in Y can be reduced. The *mutual information* is denoted as $I(X, Y)$, defined as follows.

$$I(X,Y) = H(Y) - H(Y|X) \tag{7}$$

Equations (1) to (7) are used for discrete cases. If continuous random variables considered, entropy equations for $H(Y)$, and $H(Y|X)$ are given, respectively, as follows.

$$H(Y) = -\int_{y \in U_Y} p(y) \log p(y) dy \tag{8}$$

$$H(Y|X) = -\int_{x \in U_X} \int_{y \in U_Y} p(y|x) p(x) \log p(y|x) dx\, dy \tag{9}$$

The purpose of entropy-based feature selection is to find a feature subset, denoted as FS, with maximum dependency and minimum redundancy. Previous studies just applied the mutual information value between X and Y, denoted as Eq. (7), for this purpose [17]. In this study, we take a new approach, using negative and positive values of X and Y, to break mutual information into four types: $I(X_+, Y_+), I(X_+, Y_-), I(X_-, Y_+)$ and $I(X_-, Y_-)$. Then the values of the four types of

mutual information are averaged to become an element of influence information matrix (IIM), as shown in the algorithm given in Sect. 3.2.

3.2 Proposed Method

For a dataset, the data can be re-organized into a matrix, each column of which is viewed as a variable. These variables are then expressed in the paired form of $(\vec{X}, Y)$, where $\vec{X}$ is an n-dimensional vector of feature variables, $\vec{X} = [X_1, X_2, \ldots, X_n]$, and Y is the corresponding target variable. The objective of feature selection is to select a compact subset of feature variables from $\vec{X}$ so that the target variable can have as effective information supplied by the compact subset as possible. All or part of the variables of the subset can be used as input (feature) variables, when utilized afterwards in machine learning applications. The subset is denoted by $\mathrm{FS} = \{f_k, k = 1, 2, \ldots, m\}$, where m, $m \leq n$, refers to the amount of feature variables selected into FS and f_k is the feature variable at the kth selection. Note that the parameters m and n are usually pre-given, depending on applications.

The procedure of feature selection is given as follows.

Step 1. Calculate the influence information matrix (IIM) for all feature and target variables, using the dataset corresponding to these variables. Each component of IIM is calculated using mutual information given in Eq. (7). Note that IIM can be an asymmetry matrix because the influence from one variable to another can be different and vice versus.

Step 2. Select the feature variable, which is with the best influence information to the target variable in IIM, to the empty FS initially.

Step 3. Calculate selection gains for feature variables $\{X_i\}$ that are not in FS, respectively. And, select the feature variable that is with the best selection gain to FS. The function of selection gain is composed of two parts. The 1st part is the influence information from X_i to the target variable. The 2nd one is the average influence information between X_i and each variable in FS. The selection gain function for the feature variable X_i is given below.

$$\mathrm{gain}(X_i) = \mathrm{IIM}(X_i, Y) - \frac{1}{2|\mathrm{FS}|} \sum_{k=1}^{|\mathrm{FS}|} \left(\mathrm{IIM}(X_i, f_k) + \mathrm{IIM}(f_k, X_i)\right) \tag{10}$$

where $\mathrm{IIM}(X_i, Y)$ represents the influence information from X_i to the target variable Y; f_k is the feature variable at the kth selection to FS; $|\mathrm{FS}|$ indicates the size of FS so far. Repeat Step 3 until $|\mathrm{FS}| = m$ or $\mathrm{gain}(X_i) < 0$.

4 Experiments

For experiments, we collected three datasets of time series: Mackey-Glass, S&P 500, and TAIEX. S&P 500 and TAIEX daily exchange data were collected from Google Finance, and in order to simplify the experiment, only the closing indices were used, and all feature and target variables were generated according to close stock index.

In 1977, Mackey and Glass used first-order differential-delay equations to describe physiological control systems with the Mackey-Glass differential equation [14].

$$\frac{dx(t)}{dt} = \beta \frac{x(t-\tau)}{1+\{x(t-\tau)\}^n} - \gamma x(t) \tag{11}$$

where $\{\beta, \gamma, \tau, n\}$ are positive real-valued parameters, and $x(t-\tau)$ represents the value of the variable x at time $(t-\tau)$. At different τ the equation will exhibit different physical behaviors. If τ is small, a periodic phenomenon appears. If τ is no less than 17, a chaotic phenomenon appears. Chaotic Mackey-Glass time series has been extensively applied to testing criteria for the degree of accuracy of various models using neural network, fuzzy logic, and others. The parameter settings for the Mackey-Glass equation in (11) were given as $\beta = 0.2$, $\gamma = 0.1$, $\tau = 17$, and $n = 10$. The sampling time was 1 s. For x(t <= tau), values were set to be random in [0,1]. For the Mackey-Glass dataset in the study, 1000 data points for $t = 1001$ to 2000 were used.

Standard & Poor's 500 (S&P 500) is an American stock market index based on the market capitalizations of 500 major companies having common stock listed on the NYSE or NASDAQ. The constituents of the S&P 500 index are selected by a committee and are reviewed over time. Because of the strict selection process for the S&P 500 index, many consider it one of the best representations of the U.S. stock market as well as a bellwether for the U.S. economy. For the S&P 500 dataset in the study, 283 close stock index data from the 18th June 2015 to the 1st August 2016 were collected.

Taiwan Stock Exchange Capitalization Weighted Stock Index (TAIEX) uses the Passche Formula and uses the market capitalization of listed stocks for weighting to calculate the stock price index. This is similar to the US S&P 500, and it is regarded as an index that reflects the market value of the overall stock exchanges. For the TAIEX dataset in the study, 274 close stock index data from the 17th June 2015 to the 1st August 2016 were collected.

For the proposed approach, all the original datasets must be pre-processed and converted to datasets with 30 features to comply with the experimental setup. For each dataset, we calculated its difference series in time order, denoted as $\{\Delta y(k), k = 1, 2, \ldots, n_D\}$, where $\Delta y(k) = y(k+1) - y(k)$; $y(k)$ the datum at the kth time order; n_D the size of the dataset transformed. The dataset was then reorganized as $\left\{\left(\vec{x}(i), d(i)\right), i = 1, 2, \ldots\right\}$, where $\vec{x}(i) = \left[\Delta y(i+30-j), j = 30, 29, \ldots, 1\right]^{\mathrm{T}}$ and $d(i) = \Delta y(i+30)$. After pre-processing, the Mackey-Glass dataset was re-organized to be with 969 data pairs; the S&P 500 dataset with 252 data pairs; the TAIEX dataset with 243 data pairs.

The proposed approach was compared to the brute force method for performance comparison in terms of computing time and contents of compact subsets for feature

selection. The brute force method could exhaustively form all possible subsets of features, for each of which it then computed to see its information gain contribution to the target variable. In this way, this compared method surely could find out the optimal subset of feature variables. However, it is computationally intensive as its name shows. By comparing the results by both the proposed approach and the brute force method, we investigated how good the proposed approach could establish the optimal or sub-optimal subset of feature variables. Note that to consider the computing time consumption by the brute force method we used a maximum of eight feature variables only in the three experiments. The device used for the experiments was a desktop with Intel Core i7-6500U processor, 16 GB of DDR3 memory and 256 GB SATA SSD.

For the Mackey-Glass, S&P 500, and TAIEX data preprocessed, each was used to test both the proposed algorithm (denoted as EBAFS for short henceforth) and the brute force method. The results are shown below. The feature subsets by both methods are shown in Tables 1 to 3 for Mackey-Glass, S&P 500, and TAIEX, respectively. The curve of feature selection for S&P 500 by the proposed approach is shown in Fig. 1.

Table 1. Feature selection by the EBAFS and the brute force method (Mackey-Glass)

EBAFS (proposed)		Brute force method	
Feature ID	Selection gain	Feature ID	Selection gain
1	0.8276	1	0.8276
23	0.0998	23	0.0998
20	0.0691	20	0.0691
14	0.0698	14	0.0698
16	0.0488	16	0.0488
12	0.0546	12	0.0546
18	0.0390	18	0.0390
13	0.0333	13	0.0333

Table 2. Feature selection by the EBAFS and the brute force method (S&P 500)

EBAFS (proposed)		Brute force method	
Feature ID	Selection gain	Feature ID	Selection gain
9	0.4932	9	0.4932
3	0.0328	3	0.0328
20	0.0196	7	0.0167
16	0.0206	22	0.0226
22	0.0175	16	0.0175
18	0.0156	1	0.0181
1	0.0134	18	0.0139
7	0.0132	13	0.0124

Table 3. Feature selection by the EBAFS and the brute force method (TAIEX)

EBAFS (proposed)		Brute force method	
Feature ID	Selection gain	Feature ID	Selection gain
22	0.5875	22	0.5875
30	0.0674	30	0.0674
28	0.0530	28	0.0530
5	0.0438	5	0.0438
3	0.0381	3	0.0381
21	0.0326	21	0.0326
1	0.0320	1	0.0320
20	0.0286	20	0.0286

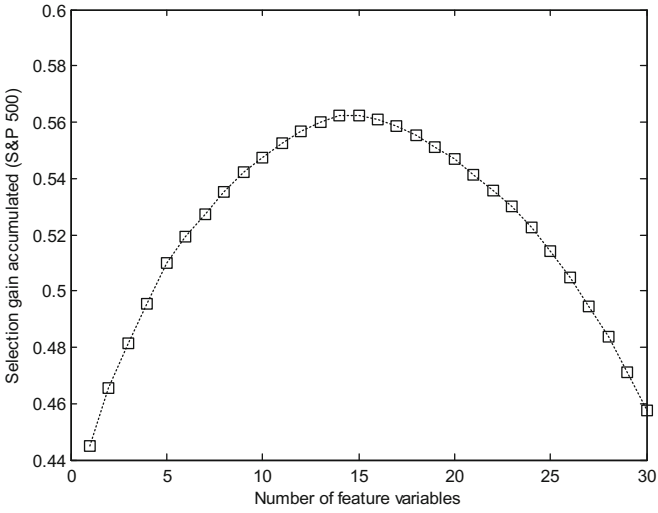

Fig. 1. Curve of feature selection (S&P 500)

As shown in Tables 1 and 3, for Mackey-Glass and TAIEX datasets, the subsets of feature variables by the proposed approach are completely consistent with the results found by the brute force method. As shown in Table 2, for S&P 500 dataset, the result by the EBAFS (the proposed approach) shows petite nonconformity from that by the brute force method (viewed as optimal solution). By comparing to the brute force method, we can see that the results of the three experiments by the proposed EBAFS are to be optimal or near optimal for feature selection.

The computing time spent by the proposed and compared methods is shown in Table 4, showing that the proposed method is much faster. Through the experimental results, the proposed approach can indeed achieve excellent performance in terms of the contents of compact feature subsets optimal or near optimal and faster computing time spent when compared to the brute force method.

Table 4. Computing time spent by the proposed method and the compared brute force method

Method	Mackey-Glass	S&P 500	TAIEX
Brute force method	3,543 s	3,234 s	3,197 s
EBAFS (proposed)	0.018 s	0.021 s	0.016 s
Times faster by EBAFS	196,833 times	154,000 times	199,813 times

5 Conclusion and Future Work

Feature selection has become more and more important for the problems faced in big data. We have proposed a novel entropy-based approach in the paper to feature selection, by which some influentially important feature variables to target variable can be effectively selected to form a compact subset. Such a subset can be utilized in the future for machine learning of intelligent computing models, such as fuzzy systems, neural networks, and others. This proposed approach is with the merits in terms of much less computing time spent when compared to the brute force method and capability of finding optimal or near optimal solution to feature selection.

For verifying the research idea of the proposed approach, we designed and performed experiments on chaotic Mackey-Glass, S&P 500, and TAIEX time series datasets. The results by the proposed method were compared to those by the brute force method for performance comparison, giving promising consequences, as shown in Tables 1 to 4. We have determined that the proposed approach can establish the subset solution optimal or near optimal in feature selection with very fast calculation. Furthermore, the method is intuitive and easy to understand.

Research topics to extend for the study in the future include finding a more efficient way to reduce further computational resources when establishing an influence information matrix for feature selection and applying on real-world problems using the proposed approach.

Acknowledgments. This study was supported by the research project with funding no. MOST 104-2221-E-008-116, Ministry of Science & Technology, Taiwan.

References

1. Aksakalli, V., Malekipirbazari, M.: Feature selection via binary simultaneous perturbation stochastic approximation. Pattern Recogn. Lett. **75**, 41–47 (2016)
2. Alibeigi, M., Hashemi, S., Hamzeh, A.: Unsupervised feature selection using feature density functions. Int. J. Electr. Electron. Eng. **3**(7), 394–399 (2009)
3. Azmandian, F., Dy, J.G., Aslam, J.A., Kaeli, D.R.: Local kernel density ratio-based feature selection for outlier detection. In ACML, pp. 49–64, November 2012
4. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**(3), 131–156 (1997)
5. Dash, M., Liu, H.: Consistency-based search in feature selection. Artif. Intell. **151**(1), 155–176 (2003)

6. De Smith, M.J.: STATSREF: Statistical Analysis Handbook - a web-based statistics resource. The Winchelsea Press, Winchelsea (2015)
7. Forman, G.: An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. **3**, 1289–1305 (2003)
8. Geng, X., Hu, G.: Unsupervised feature selection by kernel density estimation in wavelet-based spike sorting. Biomed. Sig. Process. Control **7**(2), 112–117 (2012)
9. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
10. Jain, A., Zongker, D.: Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. **19**(2), 153–158 (1997)
11. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In Proceedings of the Eleventh International Conference on Machine Learning, pp. 121–129 (1994)
12. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1), 273–324 (1997)
13. Loughrey, J., Cunningham, P.: Overfitting in wrapper-based feature subset selection: the harder you try the worse it gets. In: Bramer, M., Coenen, F., Allen, T. (eds.) Research and Development in Intelligent Systems XXI, pp. 33–43. Springer, London (2005)
14. Mackey, M., Glass, L.: Oscillation and chaos in physiological control systems. Science **197**(4300), 287–289 (1977)
15. Naghibi, T., Hoffmann, S., Pfister, B.: A semidefinite programming based search strategy for feature selection with mutual information measure. IEEE Trans. Pattern Anal. Mach. Intell. **37**(8), 1529–1541 (2015)
16. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat. **33**(3), 1065–1076 (1962)
17. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)
18. Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. Ann. Math. Stat. **27**(3), 832–837 (1956)
19. Supriyanto, C., Yusof, N., Nurhadiono, B.: Two-level feature selection for Naive Bayes with kernel density estimation in question classification based on Bloom's cognitive levels. In: 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), pp. 237–241, October 2013
20. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. J. Mach. Learn. Res. **3**, 1415–1438 (2003)
21. Zhang, J., Wang, S.: A novel single-feature and synergetic-features selection method by using ISE-based KDE and random permutation. Chin. J. Electron. **25**(1), 114–120 (2016)