Jul 1st, 12:00 AM

Renyi's-entropy-based Approach for Selecting the Significant Input Variables for the Ecological data

Can-Tao Liu

Bao-Gang Hu

International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

antao

Renyi's-entropy-based Approach for Selecting the Significant Input Variables for the Ecological data

Can-Tao Liu[a,b] **Bao-Gang Hu**[a,b]
a. NLPR /LIAMA, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, P. R. China
b. Graduate University of Chinese Academy of Sciences, Beijing 100049, P. R. China
(newcantao@yahoo.com.cn, hubg@nlpr.ia.ac.cn)

Abstract: Recently, data-driven approaches including machine-learning (ML) techniques have played a key role in the research on ecological data and models. One of the most important steps in the application of a ML technique is the selection of significant model input variables. Among ML methods, artificial neural networks and genetic algorithm are widely used for the sake of the above aim; however entropy-based learning methods have not been well adopted in the field of selecting the significant input variables for ecological model. In this paper, we utilize Renyi's entropy to estimate mutual information, and then compute maximum relevance and minimum redundancy of the input variables by the mutual information for selecting a compact input subset. This work is a case on forest cover type dataset obtained from US Forest Service Region 2 Resource Information System. A detailed analysis of the whole discrete variables of the dataset for their much redundancy was made. First we fully understand the amount of information of these features and their relevance and redundancy. Then we study which features are more important for forest cover type with feature selection method based on mutual information. The results show the discrete attributes of the dataset contain little effective information, with much redundancy. Only 17 variables of 44 attributes are kept for discrete values. The method proposed in the paper is helpful to make the good decision and measuring due to increased data transparency in ecological informatics. In all, by utilizing information theory as the mathematical infrastructure, the new view to study ecological data can be acquired.

1. INTRODUCTION

Large scale and high complexity have become available in ecological data and biological traits. Hence, data-driven approaches including machine-learning (ML) techniques have played a key role in prediction/forecasting model of ecosystems [D'heygere et al., 2003]. However, for any forecasting model, the selection of appropriate model inputs is extremely important for the prediction accuracy due to noise data [Muttil et al., 2007] and model computational performance for "the curse of dimensionality" [Bowden et al., 2005]. The aim of selecting input variables is to determine a compact input subset from a superset of potentially useful inputs, which will lead to a superior model as measured by some optimality criterion.

Recently, many researchers have studied input variable selection for the application of data-driven models in ecological modeling. Among data-driven approaches artificial neural networks (ANN) and genetic algorithm (GA) are widely used for the sake of selection of significant model input variables. Lee et al. [2003] applied ANN for modeling of coastal

algal blooms. They noted that the use of all possible input variables may present the model with noise, rather than useful information, because the effects of some of the input variables may be duplicated. D'heygere et al., [2003] used GA to reduce the number of input variables to a half or less without affecting the predictive power of benthic macro invertebrates obviously. Watts et al. [2008] utilized ANN to determine the relative contribution of abiotic factors that influenced the establishment of insect pest species. The contribution of each input variable was analyzed according to the results trained and evaluated on the relevant data by the multi-layer perceptron.

On the other hand, mutual information (MI) and entropy have a small amount of application in ecological informatics. They have been applied to clustering and ordering data sets of benthic macroinvertebrates [Walley et al., 2001] based on information theories. Rimet et al. [2005] used MI and regression maximization to explore the complexity of diatom assemblages. However, feature selection algorithms based on mutual information have not been studied almost at all in the field of selecting the significant input variables for ecological models.

In the present paper, we employ MI and Renyi's entropy to select the significant variables of ecological data, which is case on forest cover type dataset [Blackard, 1998; UCI databases]. First we estimate MI with Renyi's entropy instead of Shannon's entropy, which reduces the computational complexity. Then we evaluate minimal redundancy and maximal relevance to get the significant input variables with MI. Finally we verify our conclusions according to classification accuracy of SVM classifier.

2. METHODS

2.1 Renyi's Entropy

The entropy measure provided by Shannon is not common use for the difficulty of estimating [Battiti, 1994]. An alternative measure is Renyi's entropy [Renyi, 1970]. Principe [2000] shows that Renyi's quadratic entropy of continuous variables can be estimated by one non-parametric method, which is the Parzen window method with a kernel function on top of each sample. It turns out that Renyi's quadratic measure, when combined with the Parzen density estimation method using Gaussian kernels, provides significant computational savings.

For the continuous variable X with the probability density function $f_X(x)$, Renyi's quadratic differential entropy is defined as

$$H_{R_2} = -\log\left(\int f_X^2(x)dx\right), \tag{1}$$

Assumed that a data set $\chi = \{x_i\}_{i=1}^N$ is independently and identically drawn from $f(x)$, Gaussian kernel density estimation [Renyi, 1970, Principe 2000] is

$$\hat{f}_X(x) \propto \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{x}-\mathbf{x_i}, \sigma\mathbf{I}), \tag{2}$$

where $G(\mathbf{x}, \sigma\mathbf{I})$ is a Gaussian kernel evaluated at $\mathbf{x}$, having a diagonal, isotropic covariance matrix. Substituting Eq. (2) to Eq. (1), we arrive at the following nonparametric estimator

$$\hat{H}_{R_2}(X) = -\int \log f_X^2(x)dx = -\log\sum_{k=1}^{N}\sum_{j=1}^{N} G(\mathbf{x_k}-\mathbf{x_j}, \sqrt{2}\sigma\mathbf{I}) + \text{const} \cdot \tag{3}$$

Therefore, Renyi's quadratic entropy can be estimated as a sum of local interactions, as defined by the kernel, over all pairs of samples. Because of symmetry, only half of these need to be evaluated in practice.

2.2 Mutual Information

Entropy is the uncertainty measures of a single random variable. MI is a quantity that measures the mutual dependence of two variables. Given two random variables X and Y, their MI is defined in terms of their probabilistic density functions *f(x)*, *f(y)* and *f(x,y)* given by Eq. (4).

$$I(X;Y) = \iint f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy \cdot \tag{4}$$

The entropy and the MI have the following relationships:

$$\begin{aligned} &I(X;Y) = I(Y;X) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}, \tag{5}$$

where *H(.)* is the entropy of the respective variables, univariate or multivariate variables, and *H(.|.)* is the conditional entropy.

2.3 Renyi's-entropy-based approach for selecting the significant input variables

For selecting the significant input variables there are many entropy-based feature selection techniques like MIFS (Mutual Information Feature Selection) [Battiti, 1994], MIFS-U [Kwak and Choi, 2002], mRMR [Minimal-Redundancy-Maximal-Relevance, Peng et al., 2006], NMIFS (Normalized MIFS) [Estévez et al., 2009] methods in the area of machine learning. These methods utilize Eq. (5) to estimate MI, and adopt Shannon's entropy. As mentioned in section 2.1, Shannon's entropy is computationally expensive. However, by replacing Shannon entropy by Renyi's entropy, the computational complexity can be reduced from $O(N^2)$ to $O(N)$ [Renyi, 1970, Principe 2000].

In this paper, we focus on the mRMR [Peng et al., 2006] method, and estimate MI with Renyi's entropy instead of Shannon's entropy. The principles and the algorithm flow chart introduced as follows.

Given the input data χ tabled as N samples and M input variables $F = \{\mathbf{f}_i, i = 1, \cdots, M\}$, and the target classification variable c, the objective of the problem selecting significant input variable problem is to find a subset S with m input variables $\{\mathbf{f}_i\}$, which "optimally" characterizes c. The mRMR's formulation for the selection of the kth significant input variable is followed as

$$\underset{x_j \in F - S_{k-1}}{Max} \left[I(\mathbf{f_j}, \mathbf{c}) - \frac{1}{k-1} \sum_{x_i \in S_{m-1}} I(\mathbf{f_j}, \mathbf{f_i}) \right] \cdot \tag{6}$$

We can thus utilize Eq. (3) to estimate Renyi's entropy, and then get estimation of MI with Eq. (5), and finally select the significant input variable from Eq. (6).

The complete algorithm is as follows.

(1) (Initialization) Set $\mathbf{F}$ = 'whole input variable set', $\mathbf{S}$ = 'empty set'.

(2) $\forall \mathbf{f_i} \in \mathbf{F}$, compute $I(\mathbf{C}; \mathbf{f_i})$.

(3) Find the input variable $\hat{\mathbf{f}}_\mathbf{i}$ that maximizes $I(\mathbf{C}; \mathbf{f_i})$, set $\mathbf{F} \leftarrow \mathbf{F} \setminus \{\hat{\mathbf{f}}_\mathbf{i}\}$, $\mathbf{S} \leftarrow \hat{\mathbf{f}}_\mathbf{i}$.

(4) (Greedy selection) repeat step a and b until desired number m of input variables are selected.

(a) (Computation of the MI between variables) for all pairs of variables $(\mathbf{f_i}, \mathbf{f_s})$ with $\mathbf{f_i} \in \mathbf{F}$, $\mathbf{f_s} \in \mathbf{S}$, compute $I(\mathbf{f_i}; \mathbf{f_s})$ if not yet available.

(b) (Selection of the next significant input variable) choose the input variable $\mathbf{f_i} \in \mathbf{F}$ that maximizes Eq.6. Set $\mathbf{F} \leftarrow \mathbf{F} \setminus \{\mathbf{f_i}\}$, $\mathbf{S} \leftarrow \mathbf{f_i}$.

(5) Output the subset $\mathbf{S}$ containing m selected significant input variables.

3. STUDY AREA

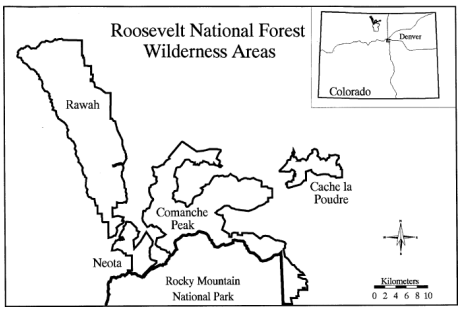

Figure 1 Study area location map [Blackard, 1998]

The forest cover type data [Blackard, 1998; UCI databases] obtained from US Forest Service (USFS) Region 2 Resource Information System (RIS) consisted of four study areas, as shown in Fig. 1. There are four wilderness areas, Rawah (29 628 hectares or 73 213 acres, area 1), Neota (3904 hectares or 9647 acres, area 2), Comanche Peak (27 389 hectares or 67 680 acres, area 3), and Cache la Poudre (3817 hectares or 9433 acres, area 4) in the Roosevelt National Forest in northern Colorado. These have experienced relatively little direct human management disturbances. Therefore forest cover type data is a result of natural ecological processes rather than the product of active forest management.

Seven forest cover type classes defined in these four areas were spruce/fir (type 1), lodgepole pine (type 2), Ponderosa pine (type 3), cottonwood/willow (type 4), aspen (type 5), Douglas-fir (type 6), and krummholz (type 7).

The actual forest cover type for a given observation (30 x 30 meter cell) consists of 12 independent variables, 54 attributes, and 1 class label. The total number of 54 attributes of cover type data availably includes the following 12 measures defined with 10 independent quantitative variables, four binary wilderness areas and forty binary soil type variables. The following table describes these variables. The order of this listing corresponds to the order of numerals along the rows of the database.

Table 1 Measures of forest cover type data set [Blackard, 1998]

Name	Data type	Measurement	Description
Elevation	quantitative	meters	Elevation in meters
Aspect	quantitative	azimuth	Aspect in degrees azimuth
Slope	quantitative	degrees	Slope in degrees
Horizontal_Distance _To_Hydrology	quantitative	meters	Horz Dist to nearest surface water features
Vertical_Distance_To _Hydrology	quantitative	meters	Vert Dist to nearest surface water features
Horizontal_Distance _To_Roadways	quantitative	meters	Horz Dist to nearest roadway
Hillshade_9am	quantitative	0 to 255 index	Hillshade index at 9am, summer solstice
Hillshade_Noon	quantitative	0 to 225 index	Hillshade index at noon, summer soltice
Hillshade_3pm	quantitative	0 to 225 index	Hillshade index at 3pm, summer solstice
Horizontal_Distance _To_Fire_Points	quantitative	meters	Horz Dist to nearest wildfire ignition points
Wilderness_Area (4 binary columns)	qualitative	0 (absence) or 1 (presence)	Wilderness area designation
Soil_Type (40 binary columns)	qualitative	0 (absence) or 1 (presence)	Soil Type designation
Cover_Type (7 types)	integer	1 to 7	Forest Cover Type designation

4. EXPERIMENTS AND RESULTS

4.1 Experiments for selecting the significant input variables

4.1.1 Experiments description

There are 581,012 observations with 54 attributes including 10 continuous variables and 44 discrete variables, as listed in Table 1. Intuitively, continuous variables are more important than discrete ones, as two measures are defined in 44 discrete attributes. The experiments are all focused on discrete input variables by reason of their much redundancy.

The steps of experiments are as follows. We first compute the MI of discrete input variables and class label, and then compute the MI between discrete input variables, furthermore we select the significant discrete input variables according to Eq.(6) and the previous results. Finally, we verify our conclusions in the light of classification accuracy. The significant input variables selected by this scheme have good performance on various types of classifiers. Herein, we only considered Support Vector Machine (SVM) as classifier. The LIBSVM package [Hsu et al., 2002] is selected, which is an open source software, provides the interface of programming language such as java, C#, and MATLAB, supporting both 2-class and multiclass classification problem.

4.1.2 Results and Discussion

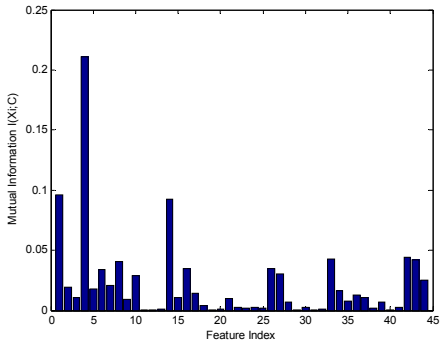

Figure 2 MI of 44 discrete input variables and class label

The results are shown in Figure 2. For explicit illustration, 1-44 is instead of the variable index 11-54 of discrete variables, similar to the following figures. The MI maximal value of discrete input variables and class label, which reflects the relevance between them, is 0.211; the other values are less than 0.1. The value less than 0.01 are the half. Furthermore, 7 values are close to zero (less than 0.001). The minimal is associated with the 19th feature index, which is 2.61e-005. It revealed that the soil type has no effect on forest cover type. From Figure 3, the 4th discrete input variable or wildness area 4 is the most important or relevance to the class label, the second is the 1st or wildness area 1, and the third is the14th or soil type 10.

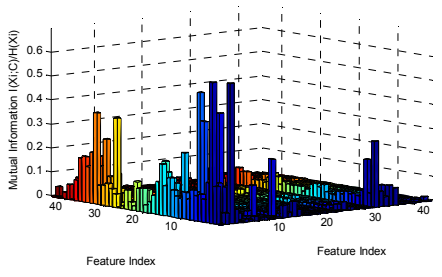

Figure 3 The ratio between MI of 44 discrete input variables and their entropy

The MI between 44 discrete input variables is small (to save space, we do not show the result); however, the ratio between MI of 44 discrete input variables and their entropy cannot be ignored, especially the ratio value is close to 0.6 for the first 4 discrete variables, as shown according to Figure 3. From information theory, it means much redundancy between discrete input variables. In Figure 3, the diagonal elements, which represent the ratio between self-information and its entropy, should be 1. However, for the sake of clarity the diagonal elements are forced to 0.

According to Eq. (6), the index of input variables sequentially selected is {4, 42, 1, 14, 43,

8, 26, 6, 27, 44, 16, 10, 7, 5, 2, 33, 17, 34, 21, 9 …}. The corresponding classification accuracy is listed in Table 2. As the classification accuracy of the 17th significant input variable selected is the maximal, the significant input variable selected subset is $\Omega = \{4, 42, 1, 14, 43, 8, 26, 6, 27, 44, 16, 10, 7, 5, 2, 33, 17\}$, the number being the variable index.

Table 2 The significant variables and their corresponding classification accuracy

NO.	1	2	3	4	5
Index of Variables	4	42	1	14	43
Classification accuracy	0.642	0.646	0.650	0.655	0.661
NO.	6	7	8	9	10
Index of Variables	8	26	6	27	44
Classification accuracy	0.665	0.669	0.676	0.680	0.684
NO.	11	12	13	14	15
Index of Variables	16	10	7	5	2
Classification accuracy	0.689	0.691	0.693	0.693	0.695
NO.	16	17	18	19	20
Index of Variables	33	17	34	21	9
Classification accuracy	0.692	0.696	0.694	0.696	0.695

Classification accuracy, as listed in Table 2, is from 0.642 to 0.696. It reveals that 44 discrete variables have slight influence on the forest cover type data. To verify it, we compute the MI of the first continuous variable and class label, which is 8.436; however, the MI maximal value of discrete input variables and class label is 0. 211.

4.2 Experiments for comparing between entropy-based approaches

As mentioned in section 2.3, there are many entropy-based feature selection approaches. In this paper, mRMR [Peng et al., 2006], NMIFS [Estevez et al., 2009] and our proposed methods RMIFS (Renyi's-entropy-based approach of Mutual Information Feature Selection) are compared in the computational speed and classification accuracy because these three methods are based on mRMR. MI is estimated by Renyi's entropy in RMIFS, however, by Shannon's entropy in the others.

4.2.1 Computational complexity

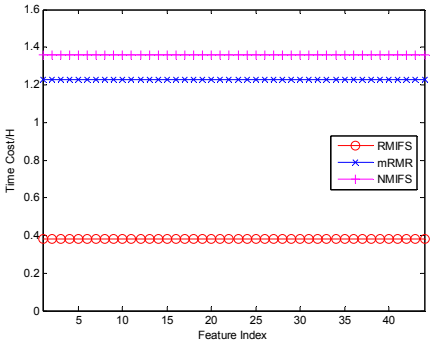

Figure 4 Time cost (hours) for selecting input variables

We compared the average computational time cost to select the top features for these methods. All experiments were run on an ordinary PC with Intel Centrino 2 P8600 CPU and Scilab 5.2. The results in Figure 4 demonstrate that three methods are almost constant of the number of variables, which is because of their computational complexity relying on the number of samples, not the number of attributes selected. However, time cost is different, RMIFS (the proposed method) is the least, mRMR is the second, and NMIFS is the most, which shows that RMIFS is computationally more efficient than other three methods due to estimating MI with Renyi's entropy.

4.2.2 Classification accuracy

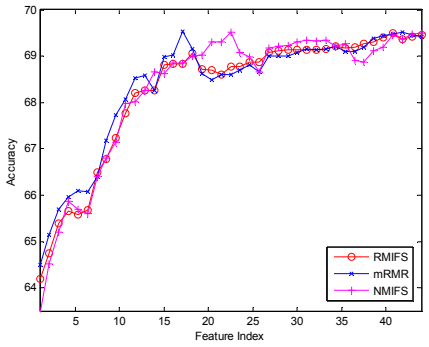

Figure 5 Comparison of classification accuracy

Like the section 4.1, we employed LIB-SVM as classifier for its convenience. The results of these three methods are shown in Figure 5, which represents the relation between accuracy and selected variables. Figure 5 shows mRMR get slightly higher classification accuracy than RMIFS and NMIFS when the number of the selected variables is within the range between 1 and 18. If the number of selected variables is within the range between 18 and 25, NMIFS lead to higher classification accuracy. And the accuracy value is close when the number is more than 25. However, the accuracy is close, and the largest disparity is 0.5%. So the computational time is critical for these methods.

5. Conclusion

In this paper, it is demonstrated to select the significant input variables for the ecological data with Renyi's-entropy-based approach, based cases on forest cover type data. The results revealed the discrete attributes of the dataset contain little effective information, and much redundancy, and only 17 variables of 44 attributes are kept. The method proposed in the paper increases ecological data transparency and is helpful to make the good decision for us to planning and measuring the ecological informatics.

For proposed method (RMIFS) reduce the computational complexity, comparisons of three methods based on mRMR about time cost and accuracy are made. The results show these methods classification accuracy is close, and slightly disparity. However, RMIFS is computationally more efficient than other two methods.

In all, by utilizing information theory as the mathematical infrastructure, the new view to study ecological data can be acquired.

ACKNOWLEDGEMENTS
This work has been partly supported by MOST of China (No. 2007DFC10740).

REFERENCES

Blackard J., Comparison of Neural Networks and Discriminant Analysis in Predicting Forest Cover Types, Ph.D. dissertation, Colorado State University, Fort Collins, 1998.

Battiti R., Using MI for selecting features in supervised neural net learning, *IEEE Trans. on Neural Networks*, 5(4), 537-550, 1994.

Bowden G.J., Dandy G.C., Maier H.R., Input determination for neural network models in water resources applications. Part 1-background and methodology. *Journal of Hydrology*, 301, 75-92, 2005.

Cover T.M. and Thomas J.A., Elements of Information Theory, 2nd edition. NewYork: John Wiley, 2005.

D'heygere T., Goethals P. L.M., Pauw N. D., Use of genetic algorithms to select input

variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecological Modelling*, 160, 291-300, 2003

Estévez P. A., Tesmer M., Perez C. A., and Zurada J. M., Normalized MI Feature Selection, *IEEE Trans. Neural Netw.*, 20(2), 189-201, 2009

Hsu C.W.and Lin C.J., A Comparison of Methods for Multi-Class Support Vector Machines, *IEEE Trans. Neural Networks,* v13, pp. 415-425, 2002.

Kwak N. and Choi C.-H., Input feature selection for classification problems, *IEEE Trans. Neural Netw.*, 3(1), pp. 143–159, 2002.

Lee J.H.W., Huang Y., Dickman M., Jayawardena A.W., Neural Network Modelling of Coastal Algal Blooms. *Ecological Modelling* , 159, 179 – 201, 2003.

Muttil N., Chau K. W., Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of Artificial Intelligence*, 20, 735–744, 2007.

Peng H., Long, F. and Ding C., "Feature selection based on MI: Criteria of max-dependency, max-relevance and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(8), 1226–1238, 2005.

Pricipe J.C., Xu D., and Fisher J.W.. Information theoretic learning. In Simon Haykin, editor, Unsupervised Adaptive Filtering. Wiley, New York, 2000.

Rimet, F., Cauchie, H.M., Tudesque, L., Ector, L., Use of artificial intelligence (MIR-max) and chemical index to define type diatom assemblages in Rhône basin and Mediterranean region. In: Lek, S., Scardi, M., Verdonschot, P., Descy, J., Park, Y.-S. (Eds.), Modelling Community Structure in Freshwater Ecosystems. Springer, Berlin, pp. 288–303, 2005..

Renyi A., *Probability Theory*, North-Holland Publishing Company, Amsterdam, Nether-lands, 1970.

UCI Repository of Machine Learning Databases. Dep. Information and Computer Sciences, Univ. California, Irvine. [Online] http://archive.ics.uci.edu/ml/datasets/Covertype

Walley, W.J, O'Connor, M.A., Unsupervised pattern recognition for the interpretation of ecological data. *Ecological Modelling*, 146, 219–230, 2001..

Watts M. J. and Worner S.P., Using artificial neural networks to determine the relative contribution of abiotic factors influencing the establishment of insect pest species. *Ecological Informatics*, 3, 64-74, 2008.