# A Novel Feature Selection Method Based on an Integrated Data Envelopment Analysis and Entropy Mode

Seyed Mojtaba Hosseini Bamakan, Peyman Gholami

RESEARCH CENTRE OF FICTITIOUS ECONOMY & DATA SCIENCE

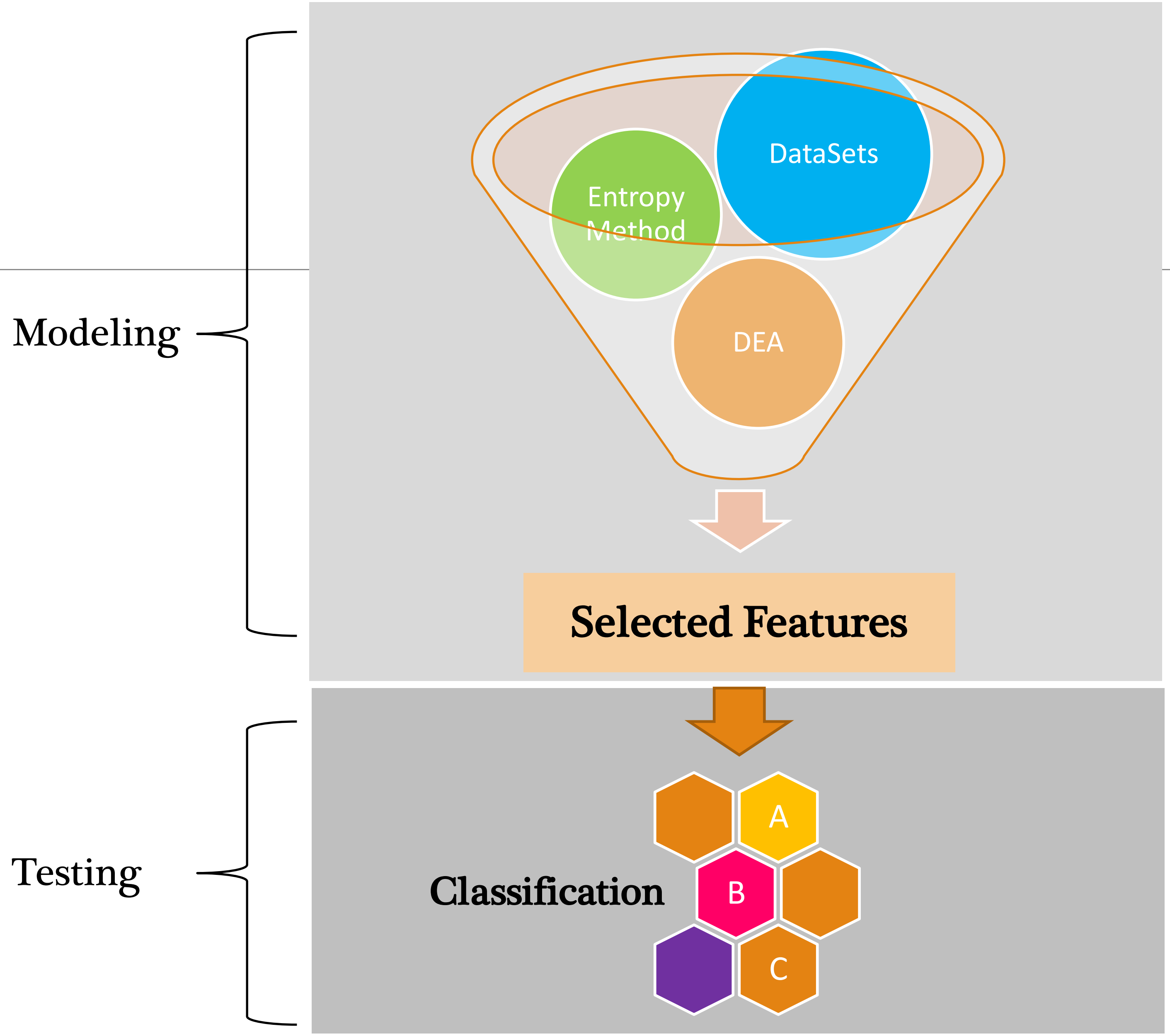
UNIVERSITY OF CHINESE ACADEMY OF SCIENCES

2014.06.03

# Abstract

Data mining is a one of the growing sciences in the world that can play a competitive advantages rule in many firms. Data mining algorithms based on their functions can be divided in four categories;

- Classification
- Feature selection
- Assassination rules
- Clustering


Modeling
Testing
DataSets
Entropy Method
DEA
Selected Features
Classification
A
B
C

# Abstract

- **Feature selection algorithms** mostly used for obtaining more precise and strong machine learning algorithms along with reducing the computation time.
- **Data Envelopment Analysis** which is a useful technique for determining the efficiency of decision-making units.
- **Entropy method** which its function is weighting the criteria to selecting the appropriate features.

# Problem Definition

Classification methods are widespread and strong tools to dealing with a real problems such as firm bankruptcy prediction, credit card assessment, intrusion detection, fraud detection and ets.

Totally, a classification machine contains the following four fundamental components:

(1) a set of attribute or characteristic values

(2) a sample training data set

(3) an acceptance domain

(4) a classification function.

The **accuracy of classification** and **predictive power** are two main issues related to classification methods.

# Problem Definition

Selecting an appropriate set of features to represent the main information of original datasets is an important factor that influences the accuracy of classification methods.

**The goal of this paper is to providing a novel feature selection method based on Data envelopment analysis and Entropy to gain more classification accuracy.**

# Feature selection

- Enhancing the classification accuracy and predictability ability
- Increasing the training process speed
- Decreasing the storage demands
- Better understanding and interpretability of a domain.

# Feature selection

Different kind of methods have been proposed feature reduction. Totally they can be divided in two main groups:

- feature extraction
- feature selection.

Although a number of comprehensive studies have been done on feature selection and classification methods to select the best subset of features to improve the accuracy of classification methods, this study focus on applying new model based on MCDM methods.

# Shannon’s entropy

Shannon’s entropy is a well-known method for calculating the weights for multiple criteria decision making problem.

Step 1: Normalize the decision matrix.

$$P_{ij} = \frac{x_{ij}}{\sum_{j=1}^{m} x_{ij}}, \qquad j = 1, \ldots, m \quad , \quad i = 1, \ldots, n$$

By normalizing the decision matrix we make a free unit matrix.

Step 2: By using formula 2 calculating the entropy:

$$E_j = -k \sum_{i=1}^{m} \left[ p_{ij} \, ln_{ij} \left( p_{ij} \right) \right] \rightarrow \left\{ \begin{array}{l} \forall_j = 1, 2, \ldots, n \\ k = \dfrac{1}{\ln(m)} \end{array} \right\}$$

N : Number of attribute

M : Number of Samples

# *Shannon’s entropy*

Step 3: Calculate the degree of deviation of each criteria from its entropy’s value:

$$d_j = 1 - E_j$$

Step 3: Calculate the degree of importance or weight of each criteria:

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \rightarrow \left(\forall_j = 1, 2, \ldots, n\right)_j \quad , \sum_{j=1}^{n} w_j = 1 \rightarrow \left(\forall_j = 1, 2, \ldots, n\right)_j$$

The entropy method is based on the variance of values in each criterion, so we can conclude if criteria have more deviation, then the value of its entropy would be increased and it shows that this criterion is more important for classification.

# Data Envelopment Analysis

DEA use to calculate the efficiency of Decision-making units (DMUs). This method is a non-parametric based on linear programming and was first proposed by Charnes, Cooper & Rhodes (1978).

The basic DEA model known as CCR model:

$$E_P = Max \frac{\sum_{r=1}^{s} u_r y_{rp}}{\sum_{i=1}^{m} v_i x_{ip}}$$

$$st: \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1$$

$$i = 1, 2, \dots, m \qquad j = 1, 2, \dots, n \qquad r = 1, 2, \dots, s$$

$$u_r, v_i \geq o$$

# Data Envelopment Analysis

There are two different methods to solve this problem, one can be output maximization or input minimization. Here, we choose the first method by placing denominator equal to 1, so in the following an output maximization CCR model presented:

$$E_P = Max \sum_{r=1}^{s} u_r y_{rp}$$

$$st: \sum_{i=1}^{m} v_i x_{ip} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq o$$

$$i = 1, 2, \ldots, m \qquad j = 1, 2, \ldots, n \qquad r = 1, 2, \ldots, s$$

$$u_r, v_i \geq o$$

# Experimental Evaluation

Table 1: Characteristics of selected datasets

| Row | Name | Number of attributes | Numbers of Instances | Numbers of Classes | Attribute Characteristics |
|---|---|---|---|---|---|
| 1 | Breast Cancer Wisconsin (Diagnostic) | 32 | 569 | 2 | Integer |
| 2 | Statlog (Landsat Satellite) Data Set | 36 | 6435 | 6 | Integer |
| 3 | Statlog (Vehicle Silhouettes) Data Set | 18 | 946 | 4 | Integer |

# Our proposed model

In the following steps we show how our model works:

Step1: compute the entropy value of each attribute in different classes by (a) at first separating the datasets according to their classes' type, (b) then calculating the Entropy of each attribute.

Step 2: considering each attribute as a Decision-Making Units (DMUs)

Step 3: placing the input of DMUs equal to 1.

Step 4: placing the output of DMUs equal to entropy value gain form step 1.

Step 5: compute the efficiency of each attribute.

Step 6: selecting the efficient attribute.

Step 7: applying other feature selection algorithms on the same datasets for selecting the features.

Step 8: comparing the result of our models with the result of step 7.

# The selected features

Table 2: The selected features from 1st dataset by different features selection algorithms and our proposed model

| Feature Selection Method | Selected Features | Number of selected features |
|---|---|---|
| CfS Subseteval | 2,7,8,14,19,21,23,24,25,27,28 | 11 |
| Consistency subset eval | 2,11,13,21,22,27,28,29 | 8 |
| Filtered Subset eval | 2,7,8,14,21,23,24,27,28 | 9 |
| Our Proposed Model | 7,8,11,13,14,16,17,20,26, 27 | 10 |

Table 3: The selected features from 2nd dataset by different features selection algorithms and our proposed model

| Feature Selection Method | Selected Features | Number of selected eatures |
|---|---|---|
| CfS Subseteval | 1,4,5,6,9,10,12,13,14,16,17,18,20,21,22,24,25,26,28,29,30,33,36 | 23 |
| Consistency subset eval | 1,2,7,10,11,17,18,24,28,29,31,33 | 12 |
| Filtered Subset eval | 1,4,5,6,9,10,12,13,14,16,17,18,20,21,22,24,25,26,28,29,30,33,36 | 23 |
| Our Proposed Model | 2,3,4,6,7,8,10,11,12,14,16,18,22,24,26,28,30,32,34,35,36 | 21 |

Table 4: The selected features from 3th dataset by different features selection algorithms and our proposed model

| Feature Selection Method | Selected Features | Number of selected features |
|---|---|---|
| CfS Subseteval | 4,5,6,7,8,9,11,12,14,15,16 | 11 |
| Consistency subset eval | 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18 | 18 |
| Filtered Subset eval | 4,5,6,7,8,9,11,12,14,15,16 | 11 |
| Our Proposed Model | 3,4,6,7,8,11,12,13,15,16 | 10 |

# Experimental Evaluation

Furthermore, we made the new datasets based on the selected features and then try to classify these new datasets by three classifications algorithms in SPSS clementine software and compare their accuracy. We used the 75% of each dataset as training dataset and the rest as testing dataset. The result showed in table 5 to 7.

Table 5: The accuracy of classification algorithms based on the selected feature for Breast Cancer Wisconsin data set

| | The accuracy of classification algorithms | | | |
|---|---|---|---|---|
| Feature Selection Method | SVM | C5.0 | Logistic Regression | average |
| CfS Subseteval | 87.84 | 92.57 | 95.95 | 92.12 |
| Consistency subset eval | 93.24 | 92.57 | 98.65 | 94.82 |
| Filtered Subset eval | 87.84 | 91.22 | 96.62 | 91.89 |
| Our Proposed Model | 89.86 | 93.92 | 95.95 | 93.24 |

# Experimental Evaluation

Table 6: The accuracy of classification algorithms based on the selected feature for Landsat Satellite data set

| | The accuracy of classification algorithms | | | |
|---|---|---|---|---|
| Feature Selection Method | SVM | C5.0 | Logistic Regression | Average |
| CfS Subseteval | 86.84 | 85.42 | 84.98 | 85.74 |
| Consistency subset eval | 87.10 | 85.42 | 84.89 | 85.80 |
| Filtered Subset eval | 86.84 | 85.42 | 84.98 | 85.74 |
| Our Proposed Model | 88.96 | 86.04 | 85.42 | 86.80 |

Table 7: The accuracy of classification algorithms based on the selected feature for Vehicle Silhouettes Data Set

| | The accuracy of classification algorithms | | | |
|---|---|---|---|---|
| Feature Selection Method | SVM | C5.0 | Logistic Regression | Average |
| CfS Subseteval | 58.74 | 66.50 | 67.96 | 64.40 |
| Consistency subset eval | 69.42 | 68.45 | 74.27 | 70.71 |
| Filtered Subset eval | 58.74 | 66.50 | 67.96 | 64.40 |
| Our Proposed Model | 61.17 | 73.3 | 64.56 | 66.34 |

# Conclusion and future work

As shown in the last sector by applying Data Envelopment Analysis and Entropy method for selecting the features, the result is comparable with the other method and in some of the cases it has a better result.

According to the acquired result we suggest other researchers to use different MCDM methods such as TOPSIS and SAW integrated with other methods of weighting such as Expected Value method for selecting the features.

Furthermore, our proposed model can be used for ranking the features instead of selecting them.

# References


1. Shi Y, Peng Y, Kou G, et al. Introduction to Data Mining Techniques via Multiple Criteria Optimization Approaches and Applications. Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications 2008;3.

2. Yan H, Wei Q. Data envelopment analysis classification machine. Information Sciences 2011;181:5029-5041.

3. Zhang D, Shi Y, Tian Y, et al. A class of classification and regression methods by multiobjective programming. Frontiers of Computer Science in China 2009;3:192-204.

4. Allwein EL, Schapire RE, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research 2001;1:113-141.

5. Qi Z, Tian Y, Shi Y. A nonparallel support vector machine for a classification problem with universum learning. Journal of Computational and Applied Mathematics 2014;263:288-298.

6. Chen Z-Y, Fan Z-P, Sun M. A hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioral data. European Journal of Operational Research 2012.

7. Anbazhagan S, Kumarappan N. A neural network approach to day-ahead deregulated electricity market prices classification. Electric Power Systems Research 2012;86:140-150.

8. Rennie JD, Rifkin R. Improving multiclass text classification with the support vector machine. 2001.

9. Smadja D, Touboul D, Cohen A, et al. Detection of Subclinical Keratoconus Using an Automated Decision Tree Classification. American journal of ophthalmology 2013.

10. Zhang Z, Shi Y, Zhang P, et al. A rough set-based multiple criteria linear programming approach for classification. Computational Science–ICCS 2008: Springer, 2008;476-485.

# References


11. Peng Y, Kou G, Chen Z, et al. Cross-validation and ensemble analyses on multiple-criteria linear programming classification for credit cardholder behavior. Computational Science-ICCS 2004: Springer, 2004;931-939.

12. Kwak W, Shi Y, Cheh JJ, et al. Multiple criteria linear programming data mining approach: An application for bankruptcy prediction. Data Mining and Knowledge Management: Springer, 2005;164-173.

13. De Stefano C, Fontanella F, Marrocco C, et al. A GA-based feature selection approach with an application to handwritten character recognition. Pattern Recognition Letters 2013.

14. Yao M, Qi M, Li J, et al. A novel classification method based on the ensemble learning and feature selection for aluminophosphate structural prediction. Microporous and Mesoporous Materials 2014;186:201-206.

15. Sakar CO, Kursun O, Gurgen F. A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy–Maximum Relevance filter method. Expert Systems with Applications 2012;39

16. Kohavi R, John GH. Wrappers for feature subset selection. Artificial intelligence 1997;97:273-324.

17. Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research 2003;3:1157-1182.

18. Sikora R, Piramuthu S. Framework for efficient feature selection in genetic algorithm based data mining. European Journal of Operational Research 2007;180:723-737.

19. Shirouyehzad H, Lotfi FH, Dabestani R. Aggregating the results of ranking models in data envelopment analysis by Shannon's entropy: a case study in hotel industry. International Journal of Modelling in Operations Management 2013;3:149-163.

20. Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. European journal of operational research 1978;2:429-444.

21. Azadeh A, Saberi M, Moghaddam RT, et al. An integrated Data Envelopment Analysis–Artificial Neural Network–Rough Set Algorithm for assessment of personnel efficiency. Expert Systems with Applications 2011;38:1364-1373.

22. Amado CA, Santos SP, Sequeira JF. Using Data Envelopment Analysis to support the design of process improvement interventions in electricity distribution. European Journal of Operational Research 2013.

23. Andersen P, Petersen NC. A procedure for ranking efficient units in data envelopment analysis. Management science 1993;39:1261-1264.

# Thank you