2nd International Conference on Information Technology and Quantitative Management, ITQM 2014

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

Seyed Mojtaba Hosseini Bamakan[a]*, Peyman Gholami[b]

aResearch Center of Fictitious Economy & Data Science, University of Chinese Academy of Sciences, Beijing 100090, China;
bTABM (Dept Collection Company Affiliated to Mellat Bank of Iran)

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 and Clustering. One of the most important of these functions is feature selection which has been increasingly developed and many researchers provide variety of algorithms to deal with this function in recent years. Feature selection algorithms mostly used for obtaining more precise and strong machine learning algorithms along with reducing the computation time. Another growing science is Multiple Criteria Decision Making techniques witch it also has variety of methods. In this paper, we use both Data Envelopment Analysis which is a useful technique for determining the efficiency of decision-making units and Entropy method which its function is weighting the criteria to selecting the appropriate features. Hence, our novel integrated method has been analyzed by implementing in a testing environment and we apply it on three datasets of UCI's datasets, so the result showed our innovated approach has comparable accuracy with the other feature selections algorithms.

*Keywords:*Data mining, Feature Selection Algorithm, Entropy, Data Envelopment Analysis, Classification

1. Introduction

Nowadays, all the business looks at the data mining as a powerful and prevalent tool to gain competitive advantages in the information technology era. Generating data has been growing with unbelievable speed by the different parts of businesses, so the functions of data mining such as data pre-processing, classification, clustering,

* Corresponding author. Tel.: +86-13261345552, +98-913-359-1263
  E-mail address: s_mojtabahossini@yahoo.com

association, forecasting, estimation, assessment and features selection getting a significant role in this time. We can consider data mining methods based on the variety of disciplines, for instance statistics, computer science, machine learning and operations research [1].

Classification methods as a widespread and potent tools to dealing with a real problems such as firm bankruptcy prediction, credit card assessment, intrusion detection, fraud detection and others has been frequently used. By considering training datasets, classification method decided whether a data belong to a specific class or not. According to Yan and Wei "in general, a classification machine contains the following four fundamental components: (1) a set of attribute or characteristic values according with some assumptions or a postulate system, (2) a sample training data set, (3) an acceptance domain and (4) a classification filter function. The classification process is to first construct the acceptance domain from the sample training data set, according to the given attribute values, and then to judge if a new data is in the acceptance or be rejected" [2, 1].

A classification method can be considered as robust one if the similarity of members in a class is high and the similarity between the different classes is low. The accuracy of classification and predictive power are two main issues related to classification methods. Support vector machine (SVM), neural network (NN), decision tree (C.50), fuzzy set, rough set, linear programming (LP) are some examples of classifications methods [3].

Allwein et al [4] used binary learning algorithm based on margin and presented a framework for multiclass classification. A nonparallel support vector machine has been used for universum learning problem by [5]. [6] provide a hierarchical multiple kernel support vector machine for customer churn prediction using longitudinal behavioural data. Anbazhagan and Kumarappan [7] provide a neural network approach to day-ahead deregulated electricity market prices classification. Rennie and Rifkin [8] compared Naïve Bayes and support vector machine with each other for text classification. Smadja et al [9] used decision tree for Detection of Subclinical Keratoconus. Shi et al used multiple criteria linear programming (MCLP) – based data mining method to apply in practical classification problems such as, a rough set-based MCLP approach [10], using MCLP for credit cardholder behaviour [11] and application of MCLP for bankruptcy prediction [12].

As shown in literature review, classification algorithms can be applied in different situations to solve real problems. But it should be noted that reducing the attribute space of a dataset is a significant part 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. Our proposed model has been examined by three real databases. This paper is organized as follow. In section 2, we describe the ordinary feature selection methods. Section 3 gives an overview of data envelopment and Shannon Entropy and their results are shown in section 4. Finally, we conclude and providing feature work in Section 5.

2. Feature selection

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 [13]. Enhancing the classification accuracy and predictability ability, increasing the training process speed and decreasing the storage demands are some of the potential advantages of feature selections algorithms. Hence, by reducing the feature set domain, better understanding and interpretability of a domain can be gain [14,15].

To make a smaller feature set based on the initial feature space to obtaining more classification accuracy and precision, different kind of methods have been proposed. These methods, mostly named as feature reduction can be divided in two main groups, called feature extraction and feature selection [13,14,16].

When the original feature sets transform to a new smaller feature space, it called feature extraction. On the other hand, when the original features do not face by any transformation function just selecting some of them it called feature selection [13].

Feature selection algorithms can be categorized as two main group, filter and wrapper. Filter functions is based on the statistical properties of feature subset, while wrapper function is based on performance of specific classifier training. Filters methods is much faster than wrapper because of its non-iterative computation on the dataset [13,17]. Although filter methods are faster, wrapper methods can obtain more recognition rates, because of their interactions between the classifier and the dataset. By the way, because of their iterative computation aspects, we can consider these methods computationally expensive [14, 13,18].

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. According to the result the accuracy of our proposed model is acceptable compare with other three algorithms of feature selection.

3. Describe the models

3.1. Shannon's entropy

There are different methods for calculating the weights for multiple criteria decision making problem and Shannon's entropy is a well-known method among them. Consider P_{ij} in decision matrix for alternatives' evaluation [19]. The original procedure of Shannon's entropy can be expressed in a series of steps:

Step 1: Normalize the decision matrix

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

By normalizing the decision matrix we make a free unit matrix. This process transforms different scales and units among various criteria into common measurable units to allow for comparisons of different criteria.

Step 2: By using formula 2 calculating the entropy:

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

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

$$d_j = 1 - E_j \tag{3}$$

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

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \rightarrow \left(\forall_j = 1, 2, \dots, n \right)_j \quad , \sum_{j=1}^{n} w_j = 1 \rightarrow \left(\forall_j = 1, 2, \dots, n \right)_j \tag{4}$$

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.

3.2. Data Envelopment Analysis

One of the well-known methods to calculate the efficiency of Decision-making units (DMUs) is data envelopment analysis (DEA). This method is a non-parametric based on linear programming and was first proposed by Charnes et al [20,21]. The assumption is that each DMU is a transformation process, by consuming some resources as inputs try to produce some outputs [22]. According to the Azadeh et al, "DEA models can be input or output

oriented and can be specified as constant returns to scale (CRS) or variable returns to scale (VRS)" [21]. Bellow we present the basic DEA model known as CCR model [20]:

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

$$st: \ \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1 \tag{5}$$

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

$$u_r, v_i \geq o$$

We can consider x as an inputs and y as an outputs. There are two different methods to solve this problem, one can be output maximization or input minimization. Input-oriented CCR model, it means compared with other peers, a DMU is efficiency if it consumes fewer inputs to produce greater, at least the same, outputs. Here, we choose the first method by placing denominator equal to 1, so in the following an output maximization CCR model presented:

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

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

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \leq o \tag{6}$$

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

$$u_r, v_i \geq o$$

Where, objective function changes between 0-1 and if the efficiency of examined unit becomes 1, this unit is efficient unless it considers as an inefficient unit. It should be noted that sometimes more than one unit ranked as an efficient unit so we used Andersen-Peterson model for ranking DMUs by omitting the constraint of that DMU. By using this model the value of efficiency can be greater than 1 [23].

4. Experimental Evaluation

In this section, we introduce the three datasets has been used for testing our model. We choose these datasets from UCI datasets, the details information of them illustrated in table 1.

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 (top 60% efficient DMUs has been selected).

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.

By going through the aforementioned steps we can select the features from the defined datasets. The results sorted in tables 2 to 4. For apply the first three method we used the WEKA software and for applying our model we used MATLAB software.

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

In the reset of this section, 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

Feature Selection Method	The accuracy of classification algorithms			
	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

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

Feature Selection Method	The accuracy of classification algorithms			
	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

Feature Selection Method	The accuracy of classification algorithms			
	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

5. Conclusion and future work

Multiple Criteria Decision Making mostly used for ranking the criteria in a real world problem. Our novel integrated model has been proposed in this paper is an example of applying MCDM method in other scientific domain. As shown in sector 4 by applying Data Envelopment and Entropy method for selecting the features, the result is comparable with the other method and in most of cases it has a better result but it can only use for integer attribute. 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.
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.