J. Basic. Appl. Sci. Res., 2(2)1197-1203, 2012

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

A New Method for Filtered Attribute in Data Mining Using Entropy

Peyman Gholami [1,*] , Azade Bazle[2] , Meysam Eftekhary[3], Payam Gholami[4]

Young researchers Club, Arak Branch, Islamic Azad University, Arak, Iran
[2,3] Department of Industrial Engineering, Arak Branch, Islamic Azad University, Arak, Iran
[4] Department of Mechanic Engineering, Arak Branch, Islamic Azad University, Arak, Iran

ABSTRACT

Filtered Attribute (feature ranking) is one of the main tasks that data mining algorithms have been proposed and the feature ranking algorithms have been used to determine the importance and ranking features. The entropy is one of the multi-criteria decision making techniques when the decision maker cannot find the weights for criteria and the entropy is offered weighting the criteria. Entropy in this paper using a new method for ranking features in the two and multiple class datasets and proposed method was implemented on 7 dataset and finally, the accuracy of the results by the correlation coefficient between the proposed methods and other ranking algorithms has been examined in respect to the high correlation coefficient. We have concluded that the proposed method is an appropriate method for ranking.

KEY WORDS: data mining, correlation coefficient, feature ranking algorithms, multi- criteria decision making, entropy.

INTRODUCTION

Data mining and knowledge discovery (DMKD) has made predominant progress during the past decades [16]. It uses algorithms, and techniques from many disciplines, including statistics, databases, machine learning, pattern recognition, artificial intelligence, data visualization, and optimization [15].

One of the most important data mining tasks is to determine the importance (rating) of features. That much research has been done in this respect and different algorithms to rank the features have been offered that can be pointed to Fisher Score and CFS and Gain Ratio algorithms. But so far decision making algorithms in determining the importance of features are not used.

In MCDM problems, since that the evaluation of criteria leads to diverse opinions and meanings, each attribute should be imported with a specific importance weight[1], question rises up here and that is ‘‘how this importance weight should be calculated’’? In literature, most of the typical MCDM methods delegate this part to decision makers, while sometimes it would be useful to engage end-users into the decision making process.

To obtain a better weighting system, we may categorize weighting methods into two categories: subjective methods and objective methods [2]. While subjective methods determine weights solely based on the preference or judgments of decision makers, objective methods use mathematical models, such as entropy method or multiple objective programming, automatically without considering the decision makers’ preferences.

The approach with objective weighting is particularly applicable for occasions where reliable subjective weights cannot be obtained [3].

Later research has applied this measure to a wide range of applications including:

- Spectral analysis [18];
- Language modeling [6];
- Economics [7].

Materials and Methods:

In This section we investigate 3 common algorithms that have been used in feature ranking and Shannon entropy and our proposed method.

Gain ratio:

The gain ratio was introduced by Quinlan in [2]. A function of this metric is that it can efficiently assess the correlation of an attribute to the class conception. The larger the gain ratio is, the more connection the attribute has with the class conception. It is efficiently used to compute the correlation of attributes with respect to the class conception of an incomplete data set in [5].So, frequencies of missing values are distributed across other values in proportion to their frequencies. Here we accept the method in [5]

Theorem1. Given the contingency table M = (m_{ij}) k _ l of an attribute. A with respect to the class variable C, the gain ratio Gr(A,C) of A with respect to C can be computed:

***Corresponding Author:** Peyman Gholami, Young researchers Club, Arak Branch, Islamic Azad University, Arak, Iran. Email: peyman711@gmail.com

$$Gr(A,C) = \frac{\sum_{i=1}^{l} col_i \ln col_i - \sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij} \ln m_{ij}}{\sum_{i=1}^{k} raw_i \ln raw_i - s \ln s} + 1 \qquad (1)$$

Where:

$$raw_i = \sum_{j=1}^{l} m_{ij}\ , col_j = \sum_{i=1}^{k} m_{ij} \text{ and } s = \sum_{i=1}^{k} raw_i \qquad (2)$$

The process of computing the gain ratio Gr(A, C) of an attribute A with respect to the class variable C of an incomplete data set D can be described as follows.

1) Count the following frequencies:

$$n(A = a_i, C = c_j), n(A = ?, C = c_j), n(A = a_i, C = ?) \text{ and } n(A = ?, C = ?)$$

(2) Compute the following summations with frequencies That demanded in step(1)

$$sa_i = \sum_{j=1}^{l} n(A = a_i, C = c_j),$$

$$sa_j = \sum_{j=1}^{k} n(A = a_i, C = c_j),$$

$$sa_i = \sum_{j=1}^{l} sa_i \qquad (3)$$

where k is the number of all values A may take on and l is the number of all classes.

(3) Construct the contingency table M = (m_{ij}) of A with respect to C, where m_{ij} can be computed by distributing frequencies of missing values across other values in proportion to their frequencies

$$m_{ij} = n(A = a_i, C = c_j) + \frac{sa_i}{sum} \times n(A = ?, C = c_j) + \frac{sc_i}{sum} \times n(A = a_i, C = ?) + \frac{n(A = a_i, C = c_j)}{sum} \times n(A = ?, C = ?) \qquad (4)$$

Compute Gr(A,C) with formula (1) and (4) in Theorem 1.

In order to count the frequencies in step (1), the data set needs to be seen only once. So, the computation of gain ratio can be very efficient and measurable for very large data sets with many samples. Furthermore, because the frequencies of missing values are scattered across other values in proportion to their frequencies, the statistic information involved in these frequencies can be utilized completely. And so, we selected gain ratio to evaluate the correlation of attributes with respect to the class conception of an incomplete data set.

Fisher score

To evaluate the discrimination power of each feature, we have been using the statistical criteria of Fisher scores that are defined as follows:

$$Fr = \frac{\sum_{i=1}^{c} n_i(\mu_i - \mu)^2}{\sum_{i=1}^{c} n_i \sigma_i^2} \qquad (5)$$

Where n_i is the number of samples in ith class, μ_i is the mean values of a feature in ith class, σ_i is the variance values of a feature in ith class, μ is the mean values of a feature in total samples. Suppose x_{ij} is the values of jth feature in ith class, then μ, μ_i, σ_i are defined as following:

$$\mu = \frac{\sum_i \sum_j x_{ij}}{\sum_i n_i} \quad (6)$$

$$\mu_i = \frac{\sum_j x_{ij}}{n_i} \quad (7)$$

$$\sigma_i = \sqrt{\frac{\sum_j (x_{ij} - \mu_i)^2}{n_i - 1}} \quad (8)$$

When the difference between μ value and μ_i value is high or σ_i value is very small, the Fisher score would be great. If a feature has similar property values in the same class and has very different values in other classes, the Fisher score would be very large. In this case, the features for discriminating samples from different classes are very distinct and use the scores for weighing the features would be very useful. [9]

CFS (Correlation based Feature Selection):

CFs algorithm based on correlation coefficient has been established and are based on a subset of the set of k features are selected and the average correlation coefficient between the features are computed and put equal $\overline{r_{ff}}$. Then calculate average correlation coefficient between features and classes and put equal $\overline{r_{cf}}$ and then put in following equation and each subset that had the greatest amount, it's features will be selected[8].

$$M_S = \frac{K\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (9)$$

Shannon entropy and objective weights:

Shannon and Weaver (1947) proposed the entropy conception, which is a measure of uncertainty in information formulated in terms of probability theory [13]. Since the entropy concept is well suited for measuring the relative contrast intensities of criteria to represent the average intrinsic information transmitted to the decision maker (Zeleny, 1996), conveniently it would be a proper option for our purpose.[14]

Shannon developed measure H that satisfied the following properties for all pi within the estimated joint probability distribution P [10]:

H is a continuous positive function;

If all pi are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n; and,

3. For all, $n \geq 2$, $H(p_1, p_2, \ldots, p_n) = h(p_1 + p_2, p_3, \ldots, p_n) + (p_1 + p_2)H\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right)$ (10)

He showed that the only function that satisfied these properties is:

$$H_{Shannon} = -\sum_i p_i \log(p_i) \quad (11)$$

Shannon's conception is capable of being deployed as a weighting method [11, 12]. Through the following steps[17]:

Step 1: Normalize the evaluation index as:

$$P_{ij} = \frac{x_{ij}}{\sum_j x_{ij}} \quad (12)$$

Step 2: Calculate entropy measure of every index using the following equation:

$$e_j = -k \sum_{j=1}^{m} P_{ij} \ln(P_{ij}) \quad (13)$$

Where $k = (\ln(m))^{-1}$

Step 3: Define the divergence through:

$$div_j = 1 - e_j \quad (14)$$

the more the div_j is, the more important the criterion jth is

Step 4: Obtain the normalized weights of indexes as:

$$w_j = \frac{div_j}{\sum_j div_j} \quad (15)$$

Proposed method:

This paper has proposed a method for filtered Attributed hat has 4 stages, which is as follows:

1. Datasets based on the classes are separated.
2. Entropy value of each features in each class are calculated.
3. Entropy value obtained in each class for each of the features is pulsed together.
4. Entropy value obtained for each features are normalized by the linear normalization.

RESULTS

Shannon entropy is used for weighting the features. At first we selected the appropriate dataset.
The data used in this study are 7 public-domain data sets from 7 application domains that have been shown in table 1.

Table 1 (the attributes of 7 public- domain datasets used for this paper)

Data set	Features number	Number of samples	Number of Classes
Labor	14	57	**2**
Segment	19	1500	**7**
Soybean	35	683	**19**
Cardiotogoraphy	21	2126	**3**
Breast cancer	9	699	**2**
Hepatitis	19	155	**2**
Ionosphere	34	351	**2**

The next step is to weight each of the features in each dataset with our method and different methods. We selected seven feature ranking algorithms, which results in Tables 2 to 8 are:

Table 2(filtered attribute of the first dataset using chi-square algorithm and entropy)

Feature number	The importance based on the chi-square algorithm	The importance based on the entropy
1	40.7	**0.49**
2	0	**0**
3	0	**0**
4	0	**0**
5	0	**0**
6	0	**0**
7	0	**0**
8	20.7	**0.26**
9	0	**0**
10	0	**0.01**
11	0	**0**
12	0	**0.01**
13	0	**0**
14	27.2	**0.25**

Table 3 (filtered attribute of the second dataset using gain ratio algorithm and entropy)

Feature number	The importance based on the gain ratio algorithm	The importance based on the entropy
1	0.08	**0.02**
2	0.39	**0.07**
2	0	**0**
4	0	**0**
5	0.11	**0.03**
6	0.2	**0.03**
7	0.17	**0.02**
8	0.2	**0.02**
9	0.22	**0.03**
10	0.52	**0.08**
11	0.55	**0.09**
12	0.47	**0.08**
13	0.5	**0.09**
14	0.38	**0.08**
15	0.4	**0.06**
16	0.43	**0.06**
17	0.51	**0.08**
18	0.41	**0.09**
19	0.56	**0.1**

Table 4 (filtered attribute of the third dataset using the one feature evaluator algorithm and entropy)

Feature number	The importance based on the One feature evaluator algorithm	The importance based on the entropy
1	26.06	**0.03**
2	24.45	**0.03**
3	22.98	**0.03**
4	24.74	**0.03**
5	20.64	**0.02**
6	18.59	**0.02**
7	26.50	**0.03**
8	29.86	**0.03**
9	27.52	**0.03**
10	22.98	**0.03**
11	28.55	**0.04**
12	16.98	**0.01**
13	30.30	**0.04**
14	30.45	**0.04**
15	30.89	**0.02**
16	27.81	**0.02**
17	23.86	**0.02**
18	27.37	**0.03**
19	28.55	**0.03**
20	25.76	**0.03**
21	32.35	**0.04**
22	35.87	**0.04**
23	28.55	**0.03**
24	26.94	**0.03**
25	16.69	**0.01**
26	25.18	**0.02**
27	18.74	**0.01**
28	36.60	**0.04**
29	39.97	**0.05**
30	27.37	**0.04**
31	26.20	**0.03**
32	26.35	**0.02**
33	26.64	**0.02**
34	26.64	**0.03**
35	27.67	**0.03**

Table 5 (filtered attribute of the fourth dataset using the symmetrical uncert algorithm and entropy)

Feature number	The importance based on the symmetrical uncut algorithm	The importance based on the entropy
1	0.08	**0.04**
2	0.13	**0.06**
3	0.02	**0.01**
4	0.07	**0.03**
5	0.18	**0.1**
6	0.18	**0.1**
7	0.19	**0.1**
8	0.08	**0.03**
9	0.05	**0.02**
10	0.01	**0.01**
11	0.18	**0.08**
12	0.09	**0.05**
13	0.09	**0.05**
14	0.02	**0.01**
15	0.01	**0.01**
16	0	**0**
17	0.12	**0.06**
18	0.16	**0.07**
19	0.13	**0.06**
20	0.11	**0.06**
21	0.02	**0.01**

Table 6 (filtered attribute of the fifth dataset using the algorithm relief attribute eval and entropy)

Features number	Importance based on relief attribute veal algorithm	The importance based on the entropy
1	0.08	**0.04**
2	0.13	**0.06**
3	0.02	**0.01**
4	0.07	**0.03**
5	0.18	**0.1**
6	0.18	**0.1**
7	0.19	**0.1**
8	0.08	**0.03**
9	0.05	**0.02**

Table 7 (filtered attribute of the sixth dataset using filtered attribute (Fisher Score) algorithm and entropy)

Features number	Importance based on filtered attribute (Fisher Score) algorithm	The importance based on the entropy
1	0.08	**0.15**
2	0	**0**
3	0	**0**
4	0	**0.01**
5	0.03	**0.05**
6	0	**0**
7	0	**0**
8	0	**0**
9	0	**0**
10	0	**0**
11	0.03	**0.06**
12	0.09	**0.16**
13	0.09	**0.17**
14	0.1	**0.22**
15	0	**0**
16	0	**0**
17	0	**0**
18	0.08	**0.13**
19	0	**0.01**

Table 8 (filtered attribute of the seventh dataset using info gain and entropy)

Features number	The Importance based on the info gain algorithm	The importance based on the entropy
1	0.17	**0.02**
2	0	**0**
3	0.38	**0.04**
4	0.32	**0.03**
5	0.46	**0.05**
6	0.44	**0.04**
7	0.35	**0.03**
8	0.36	**0.03**
9	0.28	**0.03**
10	0.19	**0.02**
11	0.27	**0.03**
12	0.29	**0.03**
13	0.36	**0.03**
14	0.22	**0.02**
15	0.31	**0.03**
16	0.31	**0.03**
17	0.30	**0.02**
18	0.19	**0.01**
19	0.25	**0.03**
20	0.18	**0.04**
21	0.37	**0.02**
22	0.34	**0.03**
23	0.32	**0.03**
24	0.18	**0.02**
25	0.28	**0.02**
26	0.16	**0.01**
27	0.32	**0.02**
28	0.27	**0.03**
29	0.39	**0.03**
30	0.11	**0.01**
31	0.34	**0.04**
32	0.16	**0.01**
33	0.40	**0.03**
34	0.37	**0.04**

The proposed method is compared with other conventional methods by examining the correlation coefficient between these algorithms and our method that is shown in Table 9.

Table 9 (check the correlation coefficient between feature ranking algorithms and entropy)

Dataset	1	2	3	4	5	6	7	Average
Correlation coefficient	0.99	0.96	0.81	0.98	0.99	0.99	0.78	**0.93**

DISUSSION

The approach used in this paper a new approach that was not previously used. The results of correlation coefficient of the proposed algorithm with other algorithms indicates that the Filtered Attribute of our proposed algorithm has a high correlation with other algorithms so that proposed algorithm has the highest correlation coefficient (99%) with chi-square algorithms, relief attribute veal and filtered attribute. It also has a 98% correlation coefficient with the symmetrical uncut algorithm and a correlation coefficient of 96% with Gain Ratio algorithm too. Other researchers are advised to implement other used techniques in multi-criteria decision making heightening techniques and compare the results with results obtained in this paper. The researchers also recommended that the implementation of the approach to take on a greater number of datasets. The researchers also recommended to attention to sensitivity analysis of our proposed method and others algorithms.

REFERENCES

1. Chen, M. F. , Tzeng, G. H., & Ding, C. G, 2003. Fuzzy MCDM approach to select service provider. In IEEE international conference on fuzzy systems: 572–577.

2. Wang, T. C., & Lee, H. D, 2009. Developing a fuzzy TOPSIS approach based on subjective weights and objective weights. Expert Systems with Applications, 36: 8980–8985.

3. Deng, H., Yeh, C. H., & Willis, R. J, 2000. Inter-company comparison using modified TOPSIS with objective weights. Computers and Operations Research, 27: 963–973.

4. J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufman, San Francisco, CA, 1993.

5. I.H. Witten, E. Frank, 2005. Data Mining Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufman, San Francisco, CA.

6. Rosenfeld, R, 1994. Adaptive statistical language modeling: A maximum entropy approach. Ph.D. Thesis, School of Computer Science, Carnegie Mellon University .

7. Golan, A., Judge, G., & Miller, D, 1996. Maximum entropy econometrics: Robust estimation with limited data. New York: John, Wiley and Sons.

8. M A. Hall, 1999. Correlation-based Feature Selection for Machine Learning, Dr thesis, Hamilton, New Zealand.

9. E.Roghanian, A.Bazleh, P.Gholami, M.Ahmadi, 2011. A Novel Classification Method Aided SAW. International Journal of Advanced Research in Computer Science, 2(1): 410-415.

10. Zitnick, C. L., & Kanade T, 2004. Maximum entropy for collaborative filtering . In ACM proceedings of the 20th conference on uncertainty in artificial intelligence, pp: 636–643.

11. Lihong, M., Yanping, Z., & Zhiwei, Z, 2008. Improved VIKOR algorithm based on AHP and Shannon entropy in the selection of thermal power enterprise's coal suppliers. In International conference on information management, innovation management and industrial engineering, pp: 129–133.

12. Wang, T. C., & Lee, H. D, 2009. Developing a fuzzy TOPSIS approach based on subjective weights and objective weights. Expert Systems with Applications, 36: 8980–8985.

13. Shannon, C. E., & Weaver, W, 1947. The mathematical theory of communication. Urbana: The University of Illinois Press.

14. Zeleny, M, 1996. Multiple criteria decision making. New York: Springer.

15. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery: an overview, in: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996: 1–34.

16. Y. Peng, G. Kou, Y. Shi, Z. Chen, 2008. A descriptive framework for the field of data mining and knowledge discovery, International Journal of Information Technology and Decision Making 7 (4): 639–682.

17. P.Gholami, A.Alem Tabriz, A.Bazleh, 2011. A Novel Method for Classification Aided TOPSIS. 2nd International Conference on Contemporary Issues in Computer and Information Science, pp: 200-204.

18. Burg, J, 1967. Maximum entropy spectral analysis. In 37th meeting of the Society of Exploration Geophysicists, Oklahoma City.