Article can be accessed online at http://www.publishingindia.com

PERFORMANCE EVALUATION OF FEATURE SELECTION MEASURES

Smita Chormunge*, Sudarson Jena**

**Department of Computer Science and Engineering, GITAM University, Hyderabad, Telangana, India.*
Email: smita2728@rediffmail.com
***Department of Information Technology, GITAM University, Hyderabad, Telangana, India.*
Email: sudarsonjena@gitam.edu

Abstract *Feature Selection is one of the approach for solving dimensionality problem. Numerous feature selection filter evaluation measures are used to produce good feature subset. This paper presents the comparison of Information Gain, Correlation and Gain Ratio filter measures to verify the performance of different filter evaluation measures on high dimensional datasets. Computational time required to evaluate dataset with respect to Naïve Bayes classifier is calculated by using filter measures. Experimental results on different datasets demonstrate that Correlation measure is favourable in terms of computational time than other measures.*

Keywords: *Feature selection, Correlation, Information Gain, Filter Measures, Gain Ratio*

Introduction

Technology development in different fields such as information retrieval, image analysis and text categorization not only contains complex and large amount of data but also these data have more number of attributes (or features). To extract such high dimensional data is challenging issue in data mining task. To solve this dimensionality problem feature selection approach are useful. Feature selection selects optimal feature subset by eliminating irrelevant and redundant data from original datasets [1].

To reduce dimensionality problem most of paper used evaluation measures. Feature subset is produce by using evaluation measures. Evaluation measures are classified into three kinds such as uncertainty measures, distance measures, and dependence measures. The data intrinsic category includes distance, entropy, and dependence measures. Recent research say that evaluation function is classified into five categories distance, information, dependence, consistency, and classier error rate [2]. There are many potential benefits of feature selection such as reducing computational time, facilitating data visualization and data understanding, minimize storage space, defying the curse of dimensionality to improve prediction performance.

In this paper different evaluation measures are compared with respect to Instance based classifier (IBK). We have calculated the computational time required to build the datasets by using Correlation, Information gain and Gain Ratio evaluation measure on UCI datasets. Experimental work performed on four different datasets which are from Text and Microarray domain.

Rest of paper is organized as follows. Detail description of Filter Evaluation measures are discussed in Section 2. Section 3 of this paper presents the analysis and results. Section 4 presents the concluding remark.

Evaluation Measures

In this section detail description of evaluation measures are discussed.

Correlation Filter Measure

Correlation is a bivariate analysis that measures the strengths of association between two attributes and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1 [4]. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two

variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is + indicates the positive relationship between the attributes and – indicates a negative relationship between the attributes.

$$r = \frac{N\sum xy - \sum(x)(y)}{\sqrt{N\sum x^2 - \sum(x^2)][N\sum y^2 - \sum(y^2)]}} \quad (1)$$

Where r is correlation coefficient, N is number of value in each data set, ∑ xy is sum of the products of paired scores, ∑x is sum of x scores, ∑y is a sum of y scores, ∑x2 is sum of squared x scores and ∑y2 is a sum of squared y scores .Pearson r correlation is the most widely used correlation statistic to measure the degree of the relationship between linearly related variables.

Information Gain Filter Measure

Information gain is the amount of information that's gained by knowing the value of the attribute, which is the entropy of the distribution before the split minus the entropy of the distribution after it. The largest information gain is equivalent to the smallest entropy. Information gain is a simplest attribute ranking measure which gained the amount of information by knowing the value of the attribute [3].

This measure is widely used in different applications such as text categorization, DNA analysis and image data analysis. Eq. (2) and (3) shows the information gain calculation. Consider A is an attribute and C is the class, then information gain is defines as [9],

Information gain= (Entropy of distribution before the split) - (Entropy of distribution after it)

$$IG_i = H(C) - H(C|A_i) \quad (2)$$

$$H(C) = -\sum_{c\in C} P(c)\log 2P(c) \quad (3)$$

$$H(C|A) = -\sum_{Q\in A} P(Q)\sum_{c\in C} p(c\,|\,Q)\log 2P(c\,|\,Q)$$

Gain Ratio Filter Measure

C4.5 a successor of ID3 uses an extension to information gain known as gain ratio [7]. It overcomes the bias of Information gain. It applies a kind of normalization to information gain using a split information value. The split information value represents the potential information generated by splitting the training data set D into v partitions, corresponding to v outcomes on attribute A

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} \times \log_2 \frac{|Dj|}{D} \quad (4)$$

High split Info when partitions have more or less the same size, Low split Info when few partitions hold most of the tuples.

Gain Ratio is defined as

$$Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Info(A)} \quad (5)$$

The attribute with the highest gain ratio is selected as the splitting attribute [8].

Results and Analysis

In this section datasets and results are discussed. For experimental work four well-known Text and Microarray datasets used. SRBCT, Lymphoma, Tr12 and Oh5, these four datasets are used for study [5]. Summary of datasets are presented in table 1. Weka tool is used for analysing the data [6].

Table 1: Summary of Datasets

Datasets	Features	Instances	Domain
SRBCT	2308	83	Microarray
Oh5	3013	918	Text
Lymphoma	4026	62	Microarray
Tr12	5803	313	Text

SRBCT is a Gene's data which contains 2308 features and 83 samples. It is taken from the microarray experiments of Small Round Blue Cell Tumors (SRBCT). Out of 83 samples 63 is training samples and 25 test samples. Lymphoma is a broad term encompassing a variety of cancers of the lymphatic system. It contains total 4026 genes and the samples are 62. There are all together three types of lymphomas. The first category, Chronic Lymphocytic Lymphoma, the second type Follicular Lymphoma and the third type Diffuse Large B-cell [10].

Text dataset tr12 is multi-class (1-of-n) attributes contains 5803 features and 313 records and Oh5 is multiclass text datasets have 3013 features and 918 samples [10].

Table 2: Comparison of Evaluation Measures

Datasets	Computational Time in seconds		
	Information Gain	Correlation	Gain Ratio
SRBCT	0.24	0.09	0.19
Oh5	3.4	1.91	3.88
Lymphoma	0.28	0.13	0.21
Tr12	2.77	1.57	3.8

High dimensional data is big challenge in data mining task. We evaluated the computational time required to execute data by using Information Gain, Correlation and Gain Ratio filter measures on different datasets. Computational time is calculated in seconds with respect to Naïve Bayes Classifier. To execute microarray datasets information gain measure takes more time than other two measures. Whereas executing text datasets Gain Ratio measure takes more time.

Correlation measures takes less time for evaluating text and microarray datasets as compared to other measures. Graphical representation of comparison results are shown in fig.1.

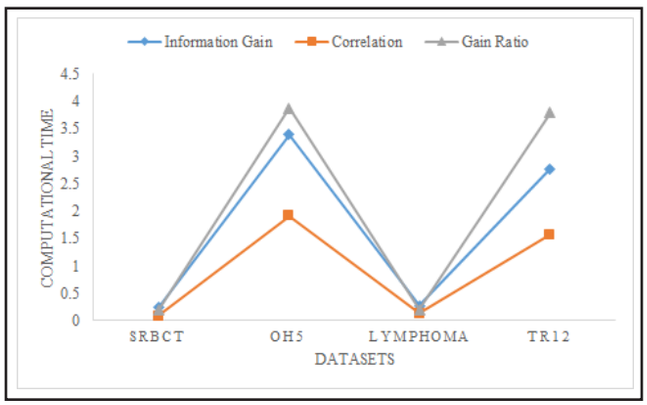

Fig. 1: Comparison of Filter Measures with Respect to Naïve Bayes Classifier

Conclusion

In this paper we have evaluated the computational time required to evaluate datasets by using Correlation, Information Gain and Gain Ratio filter evaluation measures. Microarray and Text datasets are used for empirical study. Computational time is calculated with respect to Naïve Bayes classifier. Based on results we found that Correlation measure is efficient for evaluating Microarray and Text datasets as compared to other evaluation measures. Further plan is to evaluate different evaluation measures on high dimensional datasets.

References

Song, Q., Ni, J., & Wang, G. (2013). Fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Transaction on Knowledge and Data Engineering*, January, 25(1), 1-14.

Guyon, J. I., & Elisseeff, A. (2013). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(1), 1157-1182.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(3), 1-16.

Biesiada, J., & Duch, W. (2008). Feature selection for high-dimensional data Pearson redundancy based filter. *Advances in Soft Computing*, 45, 242-249.

Chormunge, S., & Jena, S. (2015). Efficiency and effectiveness of clustering algorithms for high dimensional data. *International Journal of Computer Applications*, September, 125(11), 35-40.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2013). *WEKA Manual for Version 3-7-10*.

J.R. Quinlan, (1986). Induction of Decision Trees, Machine Learning 1: pp.81-106, Kluwer Academic Publishers, Boston.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques* (3rd ed.). San Francisco, Morgan KauffmANN Publishers.

Chormunge, S., & Jena, S. (2016). Efficient feature subset selection algorithm for high dimensional data. *International Journal of Electrical and Computer Engineering*, 6(4), 1880-1888.

Machine Learning & Data Mining Algorithms. Retrieved from http://tunedit.org/repo/Data/Text-wc