University of Benghazi
Faculty of Education Al marj
Global Libyan Journal

بنغازي جامعة
كلية التربية المرج
المجلة الليبية العالمية

العدد الواحد والعشرون –02/ يونيو 2017

Data reduction based on information gain approach

*** Hend M. farkash, ** Mona A. El.zuway.**

Benghazi, Libya

ABSTRACT

Many processes of Machine Learning need to the real values and appropriate size of data, that we get through search and Data Mining, and extract the necessary knowledge to work with them. The Preprocessing techniques are basic processes that we do when using any set of data in the different areas of Data Classification or Pattern Recognition،Where the applied these processes in any machine learning field will be difficult without the basic procedure for the preprocessing of the data and reduce its size and get rid of any confusion or noise. So give us effort and time, and produce more results that are accurate. In this paper, we offer study and apply method for used to reduce the data and extract the appropriate features to work on, finally reached a satisfactory outcomes.

KEYWORDS: Feature Selection, Data Reduction, Information Gain, Machine Learning, Preprocessing.

INTRODUCTION

Machine learning is a type of Artificial Intelligence (AI) that provides computers with the ability to learn without being explicitly programmed.

Machine learning offers methods for automatic learning of patterns from data and making intelligent decisions based on learned behavior, This becomes often necessary in the area of medicine, where the large dimensionality of data and highly variable environments[11].

Data preprocessing techniques when applied to data first, they improve the quality of the results and the required time, and these techniques include data cleaning, data integration, data transformation and data reduction, Feature selection can be performed to reduce the dimensionality of the data as a preprocessing step prior to classification[8].

This paper presents use information gain as a filter method to reduce attributes in dataset, and apply it for arrhythmia data [3].

METHODOLOGY

In this work, we studied how to used Information Gain (IG) to reduce the set of data representing the cases of heart disease data and our approach is divided into two tracks:

First one: study of the apply Gain ratio for raw data after the initial preprocessing operations which is removing missing value and normalization.

Second one: study of the apply Gain ratio for data without initial preprocessing operations which is removing missing value and normalization.

- **Data collection**

The data used in this work represents ECG signals, these are UCI" Arrhythmia" data set. . ECG is well known in medical diagnosis processes [1] [2], but it has taken a long time to become one of the methods that are used to diagnosis heart diseases. In this work we dealing with method

for preprocessing and reduction Arrhythmia data to support diagnosis processes، UCI" Arrhythmia" dataset [3] content 452 pattern each pattern has 279 attribute .

- **Preprocessing**

The data are preprocessed by manipulating missing data and applying normalization algorithm to the data , then reduction size of data by gain ratio. This operation can be arranged as the following processes in following figure(1):

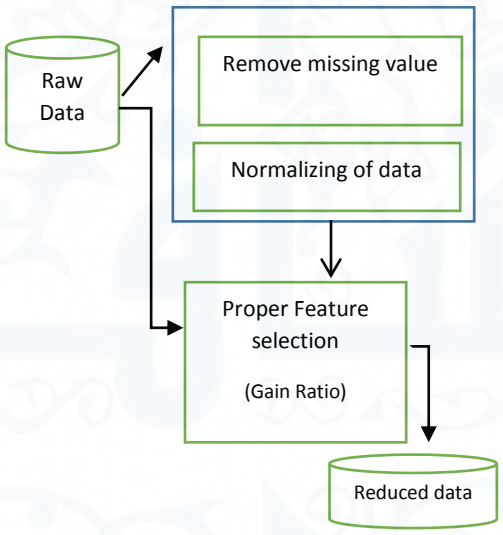

Figure (1): Preprocessing operations

To Remove missing value and normalizing raw data performed process is summarized in the following steps:

Remove personal data

The total numbers attributes (inputs) 279, some of them are personal information that is not related to classes this type of data can be removed from the input data array directly before any other preparation process. This will reduce the number of inputs used into 275 inputs.

Replace missing values with mean values

In the resulted input data array some of the attributed values are missing. They are marked by giving them a large value number, which is 999 this value needs to be replaced by average values, to perform this process required following steps:

1. Arrange the data in each class and specify the instant number in each class.

2. Replace the missing values with the mean value and calculate the maximum and minimum value.

Normalizing of data

To reduce affect for high frequency, data, must be defined in determined range of data between two boundaries. To perform this process, it is required to do following steps:

1. Determine maximum and minimum values of each attribute.
2. Determine the lower and upper limits of normalized range.
3. Remove attributes that has zero value for all patterns.

- **Information gain**

Feature selection can be performed to reduce the dimensionally of the data as a preprocessing step prior to work on [8]. The information gain measure is used to select the proper feature by compute importance for each attribute, which provided as input to any supervised learning techniques need to dealig with, such as data classification and Pattern Recognition [10].

We want to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.

The information gain is defined as[7]:

$$IG(A,S) = H(S) - \sum_{t \in T} P(t)H(t).....EQ1$$

Where,

H(S)-Entropy of set S

T-The subsets created from splitting set S by attribute A such that s =$U_{t \in T}$ t

P(t) –The proportion of the number of elements in t to number of elements in set S .

H(t) –Entropy of subset t.

$$H(S) = -\sum_{x \in X} P(x)\log_2 P(x)....EQ2$$

Where,

S-The current(data) set for which entropy is being calculated .

x – Set of classes in S.

P(x)-The proporation of the number elements in class x to the number of element in set S.

Based on this principles choose the highest Gain value and add to reduction set[9].

RESULTS

This step shows the results obtained in this research work. MATLAB was used to program the principles of this work; matlab is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation [4]. Matlab was used for computing and programming algorithms for initial preprocessing and gain ratio.

After initial preprocessing the size of raw data becomes 256 attribute for each pattern, in first track computing gain ratio for each attribute after initial preprocessing then we get 87 attribute. In second one computing gain ratio for each attribute then we get 107, all preprocessing cases shown in following table (1):

Process	Data Size
Remove personal information	275 attribute * 452 pattern
Removing missing value and normalizing	256 attribute *452 pattern
Apply Gain ratio after initial preprocessing …….(first track)	87 attribute * 452 pattren
Apply gain ratio in raw data directly…(Second track)	107 attribute * 452 pattren

Table (1): data size when apply gain ratio

CONCLUSION

This work shows that idea to data preprocessing before using it. In particular, the focus was on effect data reduction by information gain ratio in two cases, after conducting several operations on Arrhythmia data. According to table (1) We have reached satisfactory results. Information gain reduce size data and gives us Fewer feature to used it, generally in future can be used this data in classification with supervised learning using Back Propagation Neural Network (BPNN).

REFERENCES

[1] "ECG library," http://www.ecglibrary.com/ ecgurls.html, Dec. 2007.

[2] A.L. Goldberger, Clinical Electrocardiography: A Simplified Approach, Mosby, 1999.

[3] "Arrhythmia," http://archive.ics,uci.edu/ml/datasets/Arrhythmia, Jan. 2008.

[4] "Learning matlab version 6 (Release 12)," 2001.

[5] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems, Addison Wesley, 2002.

[6] "Id3_algorithm", http://informatic-ar.com/id3_algorithm/ NOV. 2016.

[7] "Measuring Entropy (data disorder) and Information Gain"http://mariuszprzydatek.com/ NOV. 2016.

[8] S. Doraisamy, S. Golzari, N. M. Norowi, MD. N. B. Sulaiman, N.I. Udzir, "A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music" ,2008 , PP 331-336.

[9] B. Azhagusundari, A. S. Thanamani. "Feature Selection based on Information Gain", iJITEE, ISSN: 2278-3075, V.2, Issue-2, January 2013.

[10] A. Karegowde, M. A. Jararam " Comparative study of attribute selection using gain ratio and correlation based feature selection", research gate ,January 2010.

[11] http://campar.in.tum.de/view/Chair/ResearchIssueMLmedical ,NOV 2016