A SURVEY ON RELEVANCE FEATURE SELECTION DISCOVERY

Revathy M[1], [2]Minu Lalitha Madhavu

[1]PG Scholar, Department of Computer Science and Engineering, Sree Buddha College of Engineering, Alappuzha, India

[2]Assistant Professor, Department of Computer Science and Engineering, Sree Buddha College of Engineering, Alappuzha, India

---***---

Abstract - *Relevance feature discovery is a big challenging issue in determine the quality of the user searched documents. New user is wanted a most relevant feature for its appropriate searching for text, document, images, etc. In early days term based and pattern based technique are used for find out the most relevant feature for documents. Now days feature selection methods and different types of clustering (Partition based clustering, density based clustering, hierarchical clustering etc.)are used. We are conducting a survey based on different techniques in feature selection and relevance feature discovery.*

Keywords: Educational data mining, Feature method, Feature selection, Feature extraction

1. INTRODUCTION

Different types of search engines collect millions of queries for several items, the engines provides the result for every query. The users is not interested to spend more time to reading all the search results they want the relevant data. Many tools are available checking the repeated terms of the search result and remove that redundant word. If a query is" apple" some user interested searching about the fruit apple and some others searching about the phone apple so there is a confusion and lots of data are arising. Information filtering is a technique for find out the most relevant data from the searching results so that the user can be spends lesser time. Filtering is checking the users feedback documents about the search and remove the unwanted data this type of filtering is called the adaptive filtering.

There is a challenging issue to find out useful phrases for pattern text mining. Sequential pattern is that the some words are repeatedly appearing in a sentence or paragraphs. Many researchers find out the pattern based mining for sequential data. There is two types of patterns are used inter pattern and intra pattern, where a inter pattern means the repeated terms in a sentences and intra pattern means the repeat terms are appeared in a paragraph.so the increased efficiency of the pattern extraction feature selection is used, which reduce the computation time, prediction performance, better understanding of a data.

2. LITERATURE SURVEY

M. Aghdam [1] presented in classification systems feature extraction and feature selection is mostly used. Feature selection reduce the size of the datasets which is the feature is cannot process further. Text categorization is the major problem in feature selection so improve the quality of feature selection an algorithm is used which is based on ant colony optimization which is similar to real ant that found a short path for food. In the algorithm first treat the training set and extracted the most relevant feature then get a group of feature set. Then apply the feature selection technique where ACO is used, there is a classifier is used for evaluate the feature and select the best subset. Finally the evaluation function produced the best subset.

Yuefeng [2] presented a paper that most commonly used feature selection is using term based approach but recently pattern based technique is also used. In pattern based technique there is patterns are grouped into both positive and negative patterns. According to their specificity the low level terms can be easily removed and find out the high level term feature .Revising positive feature algorithm is used for find out the high level terms, it also determine negative terms which is closed to the feature in the positive terms.

Nouman Azma[3] stated that when a repeated terms are appeared in sentence or paragraph can be make noise in the document so filter technique used for remove the noise.so the text categorization is the basic problem of the filtration in order to avoid the problem feature selection is used. The main metrics that used in the feature selection is the term frequency or the term that repeatedly appeared in the document .Mainly Gini index is used for measure the frequency of the term, frequency based metrics can be determined by how much the data is scattered in the document. Filter method, wrapper method and hybrid methods are mainly provided feature selection technique

,mostly used feature selection is filter technique which is extracted feature independently.in this paper provided the term frequency which is calculated by using Gini index and discriminative power measure. DPM is useful to reduce the feature set in both negative and positive patterns ,the Gini index is used to provide the splitting the terms.

Girish Chandrasekhar [4] paper tells that feature selection is useful reduce the computation time, easy understanding of the data .It focus on variable remove using filter, wrapper and embedded method. In filter method evaluating the general data independently which is differ from mining algorithm, wrapper method is to require predetermined algorithm of mining for the evaluation technique, hybrid model which is the combination of filter and wrapper model.

Irene Rodriguez-Lujan [5] proposed a new technique for improve the quality of the feature selection Quadratic Programming Feature Selection(QPFS).It is helpful to calculate the optimization problem for reducing the computational time of large data sets. It is applied on both small and large scale data sets where compute the high computational efficiency of the large data sets. QPFS is combined with the other method is called the Nystrom method which is used correlation coefficient because it is efficient than other previous method for calculating the feature selection .All the experiments are shows that the correlation coefficient is much better than other selection procedure.

R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruiz [6] are stated feature subset selection for the classification. There is mainly two algorithms are used for subset selection fast correlation based filter (FCBF) and sequential forward selection(SFS) hybrid feature selection is used better than the wrapper method. It shows that the subset feature selection is showed error prone of the all classification. Mainly the feature selection is classified into feature rank and feature subset selection.FR select the relevance feature of the each class and then ranked them foe the top most relevant feature. So that these two subset selection algorithms are helpful for classifying them.

Isabelle Guyon [7] prepared paper based on the variable and feature selection process. Variable feature selection is mainly used in the data sets which is tens or hundreds .In internet documents the text processing ,gene expression analysis is the some examples of having large data sets and providing variable feature selection. It is mainly focused on improving the prediction performance, faster and cost effective predictors ,better understanding of the data .It providing a better feature ranking, feature selection, efficient search method, feature validity assessments.

Yuzong Liu, Kai Wei[8] they applied for selecting subsets from high dimensional scores, sub sets data from training, sub modular functions etc. sub set selection have two applications selection for phone training , phone segment classification. Presented a new method feature selection provide optimal guarantee. It helpful for calculating the large feature selection in variety of application. This method is applicable in the feature selection problem of the pattern recognition high dimensional spaces in feature.

Jović[9] they provide a review feature selection methods and application for data preprocessing for data reduction. This is used to find the accurate data models. Mainly feature selection application classification, clustering and regression task. Mainly used feature selection filter, wrapper, embedded and recent new hybrid techniques. Applying new algorithms for the hybrid method in computation heuristic algorithm and genetic algorithms .It provide application in bioinformatics, image processing, text mining.

Mathew Shardlow[10] conduct a analysis of a feature selection technique. There are several technique are used. Filter, wrapper and hybrid method mostly a hybrid technique is used that is called Ranked Forward Search. When small subset of feature is selected then the accuracy of the selection is increased by the forward search. This technique is reduce the no of dimensionality of the data sets. So that the feature selection can be easily done.

3. CONCLUSIONS

Feature selection most important process in the classification and mining process. Find out the most relevant feature is helpful to reduce the computation time, easily understanding of the data so there are several technique are used for retrieve the relevant feature. This survey shows that the several method that improve the efficiency of the feature selection. Hybrid method is have new technique for feature selection process.

REFERENCES

[1] M. Aghdam, N. Ghasem-Aghaee, and M. Basiri, “Text feature selection using ant colony optimization,” in Expert Syst. Appl., vol. 36, pp. 6843–6853, 2009.

[2] Yuefeng Li “Positive and Negative Patterns for Relevance Feature Discovery”,Discipline

of Computer Science Queensland University of Technology Brisbane, QLD 4001, Australia y2.li@qut.edu.au optimization," in Expert Syst. Appl., vol. 36, pp. 6843–6853, 2009.

[3] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," Expert Syst. Appl., vol. 39, no. 5, pp. 4760–4768, 2012.

[4] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," in Comput. Electr. Eng., vol. 40, pp. 16–28, 2014.

[5] Irene Rodriguez-Lujan, Quadratic Programming Feature Selection" Departamento de Ingenier´ıa Inform´atica and IIC Universidad Aut´onoma de Madrid 28049 Madrid,

[6] R. Ruiz, J.C. Riquelme, J.S. Aguilar-Ruiz, M. García-Torres" Fast feature selection aimed at high-dimensional data via hybrid-sequential-ranked searches".

[7] Isabelle Guyon "An Introduction to Variable and Feature Selection "Clopinet 955 Creston Road Berkeley, CA 94708-1501, USA

[8] Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, Jeff Bilmes," SUBMODULAR FEATURE SELECTION FOR HIGH-DIMENSIONAL ACOUSTIC SCORE SPACES" Department of Electrical Engineering, University of Washington Seattle

[9] A. Jović K. Brkić and N. Bogunović "A review of feature selection methods with applications "Faculty of Electrical Engineering and Computing, University of Zagreb / Department of Electronics, Microelectronics,

[10] Matthew Shardlow "An Analysis of Feature Selection Techniques"