STUDY ON RELAVANCE FEATURE SELECTION METHODS

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 - *Feature selection are available in the field were thousands of variables with high dimension. Feature selection methods which provide reducing computation time. Several methods that provide improve the efficiency of the predictors and better understandings over the data that generated. So that feature construction, feature ranking process needs the efficient search methods are provided. Feature selection algorithm is used to reduce the high dimensionality. Feature selection is the process which reduces the inputs for processing and analysis or to finding the most meaningful inputs. Extracted the data from existing data are called feature extraction process.*

Key Words: **Feature selection, Feature extraction, Selection algorithms, Feature ranking, Feature construction.**

1. INTRODUCTION

Feature selection is a process which reduces the input for processing. Always the data contains more information for the builded model or the wrong information. Feature selection not only improves the quality of the model it also processing the model more efficient. In order to reduce the dimensionality of the data feature selection technique are used. Ant colony selection optimization algorithm techniques are used for feature selection process. The ants are creating a short path for finding their food. Feature extraction, classification are done using this technique. The increasing of datasets the classification is very difficult. So that new quadratic programming feature selection process is arise. The weight of the feature text which is higher weight is used for classifier. So the redundancy of the each feature is determined. Improve the efficiency of the predictors the variable selection is done using the variable ranking process. Mainly filter method predicts the pre-processing step of the variable ranking. Wrapper, filter, embedded methods are used for variable ranking process. The feature subset selection is done arising fast correlation based filter ad forward selection process. The sequential process is does not interact with some data in the Features. A greedy algorithm is used to help fast and scalable optimization of the feature. There is a big issue that discovered the relevant features from large no of documents patterns by the user preference of the predictor feature. So that the overall document is classified into D- and+ patterns. Apriori property is applied for find out the sequential patterns of the documents. So that the several algorithms are used for find out the features from the documents. The SPADE algorithm is avoiding the complexity problem of the features .In the supervised techniques the labels are known, unsupervised learning the datasets are unknown. Semi supervised learning small subsets of labels. The related works shows the different types of feature selection methods.

2. LITERATURE SURVEY

In classification system feature selection and feature extraction are the main steps. In order to reduce the dimensionality of the documents feature selection is basically used. The feature selection in the text categorization is discussed in [1] a feature selection algorithm is used to reduce the high dimensionality. The algorithm is performed basically on the ant colony optimization which is inspired by the ants are created a shortest path for finding their food. The computational complexity of the algorithm is very less so it is very easily to implement. There are several steps for the algorithm initialization, solution generation and evaluation of ants, evaluation of the selected subsets, check the stop criterion, generation of new ants.

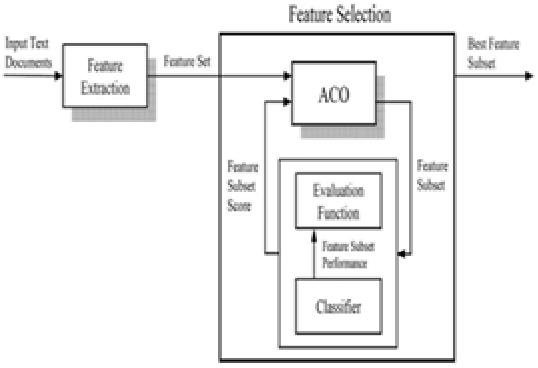

Fig 1.1 ACO layout

Due to the increasing of data sets in the real world the classification accuracy of the subset of feature is very difficult [2]. To reduce the quadratic optimization problem new selection method is provided quadratic programming feature selection. This is used for reduce the computational complexity by using Nystrom method. To minimize the multivariate quadratic function object classifier used N samples and M variables. Then find out the weights of text in the feature selection process if the feature which has higher weight is used for classifier. So the redundancy of the each feature can be easily determined. There is high time and space complexity for solving the quadratic programming problems. So that Nystrom method used matrix diagonalization for reformulate the optimization problem. Nystrom helps to sample the variables without remove the information. So that features selection of the huge documents by using the quadratic programming functions.

In the case of large data sets there is a big issue [3] to find out the variable for appropriate user needs. So that the variable selection techniques needs where data set have thousands of documents. Main goal is to improve the efficiency of the predictors and better understanding over the data that generated. So that feature construction, feature ranking like process needs the efficient search methods are provided. Variables are selected by using variable ranking process that is a filter method predictor use the preprocessing step for to determine the variable rank. Variable subset selection is done using by their process wrapper, filters and embedded method. Wrapper method is that according the power of predictive scores the black box subset by using machine learning. Filter method is provided predictor independently chooses the subset variables from the preprocessing step. Embedded method is that perform the selection of variables at the time of training the documents which directly given to the machine learning process.

Feature subset selection is used in the classification process [4]. To ranked the features in a document and compare the list with the mostly used algorithms. Basically used two techniques one is Fast Correlation Based Filter (FCBF) and second one is Sequential Forward Selection (SFS).This methods are applied on to the any high dimensional domains. Feature is ranked by using the evaluation measure, list the feature is compare with the ranked list. Evaluation done by any wrapper, filter or embedded method. Repeating the process up to the algorithm returns the subsets. This does not have any irrelevant data or redundancy of the feature from the whole document. The main disadvantage to the sequential forward generation is that it is does not interact with some features. So that the feature from two process are compare joined together to form a candidate generation.

Sub modular selection for the feature in the large documents [5] is used in the several fields. Dimensions of the high dimensional space, subset of data for training, also sub modular optimization. An accelerated greedy algorithm that used to help fast and scalable optimization of the features. The method is still have some outperforms but it is computationally efficient. It perform mutual information between the features and variables .Any pattern recognition task can involves in the high dimensionality spaces for the feature selection of the overall documents that the. Technique high sub modular selection for the feature selection is helping for the total extraction of the documents.

A novel approach that is [6] used for the helping for users to select correct relevant features. There is a big issue that discovered the relevant feature form large no of documents, patterns by the user preference of the particular features. So that the overall document is divided into two groups positive patterns and negative patterns. The classification basically on the specificity of the higher level features over than the lower level ones. The training set is D and classified the positive patterns D+ and negative patterns D-.In this technique the overall document is split into a paragraph for easily understanding the relevant features. The deployment is that the find out the support of the each patterns by the presence of the patterns in the documents. Pattern Taxonomy Mining algorithm is basically used in the paper finds out weight of each feature in the document by evaluation of support. Applying apriori property for find out the sequential patterns of the features in the documents.

In the case of information filtering, monitoring, categorizing, personalizing or searched application [7] were they performed by using the text categorization process. In text categorization process the feature selection is an effective and efficient process. Feature selection process is mainly depends on the term frequency or the document frequency of a document. Document frequency is depends on the discriminative power of the document and Gini index which is examined with the term frequency. These two metrics are comparing with the datasets. Term frequency is mainly observed the total scattering of the features in the overall document class and also the rate at which the data was covered. So that the text categorization is comparing the both term frequency and the document frequency. Wrapper, filter and embedded methods are used for the feature selection of the large no. of documents.

Now a days the availability of documents increased more [8] .The documents contains relevant and irrelevant features the extraction of relevant features is very difficult. Some automatic feature selection is available in the internet. Novel approach for the feature selection is that the association rule based feature selection process. This is the unsupervised selection technique. This stated that the text documents are associated to each other. These associated documents are extracted by using association rule .The measure of the association rule of the documents are find out by relative confidence. If a data set R then X->Y is called the association rule. Were X and Y are the terms in the data set. Confidence is the term where transaction that contains Y according to the transaction Y. Support is that the transaction containing X and y according to the number of transactions.

Supp. (R) = P(X^Y); Conf. (R) = P(X^Y)/P(X) .

The class of unknown data can be discovered by using the classification process [9].Several classification techniques are used .Mainly Bayesian, Decision Tree etc. The relevant data are mined before removed the irrelevant data from the document. Removed the data by applying wrapper, filter and embedded method are used. Improve the efficiency of the mining data, optimized the parameter of the models are by applying feature selection techniques. Mainly classification

Process is applied supervised, unsupervised or semi supervised. In the supervised technique the labels are known, unsupervised learning were the dataset are unknown, semi supervised learning small subset of labels are available. Classification is applied in the training set of the each document and testing the each term. Rule based classifier is mainly used for. To

Mining the sequential patterns or the fast discovery of the sequential patterns are determined by a algorithm that is called SPADE algorithm [10]. In the existing systems are provided multiple scanning of the database so that needs more computation time and also provide complex hash structure. So that avoids this complex problem by using the SPADE algorithm. The algorithm is used that the overall document is divided into sub problems .Then applying the joining process for mixed up the sub problems of the documents. This algorithm can be used in the real world problems for easy searching the data. Determine the frequent patterns by applying two different searching strategies. Searching features are placed in a database by applying Id to the terms. That can be reducing the main memory problems. First scanning for generating the first frequent generation second scan for generating second sequence.

Clustering is a technique [11] that used in the data mining process, so the feature selection from the clustered document is very difficult. Some method is now a day's produced new model is applied to the wrapper method, so that the model is superior than the other models. The model is alternative in supervised learning problems. EM algorithm is used for mainly in the wrapper model; two steps are mainly used expectation step and the maximization step. E step used the current parameters and M step used in the estimate parameters. The selected terms can be easily recognized by using these clustering problems, so that the wrapping model is helpful for the clustering technique.

The automatic partitioning were potentially uses [12] but we focus that one web content into Several subgroups .So that feature selection process is done very easily. Partition is done by using the segmentation algorithms. After the segmentation process the overall in documents are clustered together. That have the minimum of cost function .That is used by which the segmentation of the process easy for those calculated the relevance feature from the huge amount of the document.

Pattern Discovery For Text Mining

Many techniques are proposed for mining the useful patterns from documents. There is a still issue that the how to effectively use and update the discovered patterns. Term based approach are [13] suffer the problem which polysemy and synonymy many hypothesis are proved that the pattern based approach is better than the term based approach but there is no all the experiments are support this. In order to improve the efficiency of the discovered patterns, innovative and effective pattern discovery process is also used. The given patterns are split into 2 documents negative and positive, and then the patterns are structured in to taxonomy by using a relation. So that it recover problem that raised in the text mining low frequency of the data and misinterpretation.

Information Filtering [14] is basically used in the case of large volume of documents are used in the datasets. User modeling and the filtering are the two main components that are used in the information filtering process. Filtering of a document is very useful in the user preference search that depends upon the user profile and also the model of the user system. In the information filtering is used the various approaches for modeling the user profile and also the documents of the particular user. These approaches are including in the term, pattern and phrase based modeling. Knowledge based and statistical based concepts mainly used in the information filtering process. Term weighting approach is used in the case of term based approach is used. To carry the semantic structure of the document the phrase based approach is used, but the phrases are low frequency. Topic based model is overcome the overall disadvantages by combining the data mining techniques.

High dimensional data are analyzed by [15] using logistic regression which has high computational and statistical challenge. In order to avoid over fitting of the documents are by proposed a approach is called simple Bayesian logistic regression which used a Laplace. The model is also a sparse predictive model. Classification of the documents is used by support vector machines that classify and derived compact predictive models. The statistical classification of the text representation by applying text weighting and text processing. The text process are diagnosed the character string and term weighting find out the no. of terms that appeared in a document. Logistic regression classifier produces a more compact classification of the documents. Effective process is logistic regression process. Bayesian also help to classify the documents.

There are several techniques are available for [16] find out the relevant features in a document, but most of them are not guaranteed ones because there is some problems are occurred. The new technique that applied to produce the update, recognize and categorized text features. Some experiments are done to the different documents .Some students have some doubt that can be solved by classifying the student learning materials. So the student can be improved the efficiency of studying things .Most probably used term based models for extract the irrelevant documents from the data sets. Then a SVM classifier that classify the text documents. So that there a model that found for solving the students doubts that is a recommendation model called material recommendation model. Online questions are provided for each staff they send to the student for recover the solution.

There are a group of documents are available [17] in the real world which contains relevant and irrelevant terms. To reduce the terms from the documents by applying feature selection methods. Sometimes selection of feature is not sufficient for the documents so there is a redundancy of the terms are also checked. So the paper shows that the redundancy analysis of each terms in a document. So a new frame work is produced by the combination of relevance and redundancy of the documents. The analysis of the both process are done by using correlation based method. Feature relevance is classified in to basically three group's irrelevant, strongly and weekly relevant features. Strong relevant are seed for the optimal subset, weekly relevant is not necessary for the classification but sometimes it need, irrelevance is not necessary. If the two terms are completely crenelated by each other then they are redundant feature. Feature redundancy is calculated by applying the feature correlation process.

3. CONCLUSIONS

Feature selection technique is used in several fields. There are some algorithms are also used for extracting the relevant features from the documents. Variable selection techniques are reducing the high dimensionality of the feature text. So that the feature selection process is helped for extracting the relevant feature from the data. Relevance feature selection process is used in the case of large amount of data in database. So that the relevance feature selection process reducing the number of dimensions of a dataset. These features are improved the accuracy of the classifier.

ACKNOWLEDGEMENT

We are grateful to our project guide and HOD of the department Prof. Anil A.R for her support, remarks and suggestions for preparing the paper. We are also thankful to the staff members of the computer science department. We are also thankful to all the persons who support for preparing the paper directly and indirectly.

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

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

[4] 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

[5] 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.

[6] 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.

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

[8] Tien Dung Do, Siu Cheung Hui and Alvis C.M. Fong ,"Associative Feature Selection for Text Mining ", , International Journal of Information Technology, Vol. 12 No.4 2006.

[9] Suita Beniwal*, Jitender Arora, "Classification and Feature Selection Techniques in Data Mining ", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 6, and August – 2012, ISSN: 2278-018.

[10] Mohammaed J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences", , Machine Learning, 42, 31–60, 2001.

[11] Luis Talavera ,"An evaluation of filter and wrapper methods for feature selection in categorical clustering.

[12] Richard Romero and Adam Berger ,"Automatic Partitioning of Web Pages Using Clustering.

[13] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu ,"Effective Pattern Discovery for Text Mining", , IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 1, JANUARY 2012.

[14] Sreerekha R S, Smith E S ,"A Review on Modeling Methods for Information Filtering "Mtech Student, JETIR (ISSN-2349-5162), June 2016, Volume 3, Issue 6.

[15] Alexander GENKIN DIMACS, "Large-Scale Bayesian Logistic Regression for Text Categorization", Rutgers University Piscataway, TECHNOMETRICS, AUGUST 2007, VOL. 49, NO.

[16] Ms. Raja Saranya Kumari ,Ms. R. Divya Lakshmi, ".Efficient Classification of Text and Improving Learning Experience, IJSTE - International Journal of Science Technology & Engineering | Volume 3 | Issue 01 | July 2016.

[17] Lei Yu, Huan Liu "Efficient Feature Selection via Analysis of Relevance and Redundancy", , Journal of Machine Learning Research 5 (2004) 1205–1224 Submitted 1/04; Revised 5/04; Published 10/04.