Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.862 – 867

RESEARCH ARTICLE

Correlation Based Feature Selection with Irrelevant Feature Removal

P.Velavan[1], S.Subashini[2]

[1]PG Student, Department of CSE & N.P.R College of Engg and Tech, Dindigul, Tamil Nadu, India
[2]Assistant Professor, Department of CSE & N.P.R College of Engg and Tech, Dindigul, Tamil Nadu, India
[1] vela.friends@yahoo.com

Abstract-For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. A correlation based feature selection algorithm is developed in traditional methods. A feature subset selection research has focused on searching single query feature subset selection. Also irrelevant and redundant features are removed by this correlation based feature selection algorithm to provide accuracy of targeted data. In proposed work, plan to explore different types of correlation measures, with correlation based feature selection of feature space. Using this correlation based feature selection; multiple correlation/query can be used to get target feature subset selection

Keywords: Feature Subset Selection, Classification, Wrapper Methods, Filter Methods

I. INTRODUCTION

Data mining is the search for valuable information in large volumes of data. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, "Knowledge Discovery in Databases", or KDD. Data mining is the process of finding patterns and relations in large databases. The primary purpose of the data mining is to extract information from huge amounts of raw data. Data mining using statistical methods have been quite successful. Experimental studies show that the learning performance of irrelevant features is greatly improved when these algorithms are used to preprocess the training data by eliminating the irrelevant features from feature's consideration[1].Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Dramatic advances in data capture, processing power, data transmission and storage capabilities are enabling organizations to integrate their various databases into *data warehouses*. Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data[2].

Information can be converted into *knowledge* about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior[3]. Thus a manufacturer or retailer could determine which items are most susceptible to promotional efforts. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[4].With the aim of choosing a subset of good features with respect to the target concepts, feature subset selection is an effective way for reducing dimensionality, removing irrelevant data, increasing learning accuracy, and improving result comprehensibility. Many feature subset selection methods have been proposed and studied for machine learning applications. Previously many feature selection algorithm can be used. But Their computational complexity is low, but the accuracy of the learning algorithms is not guaranteed. With respect to good feature selection, i have choosing "correlation based feature selection algorithm". This algorithm will be provide better result than previous feature selection algorithms.

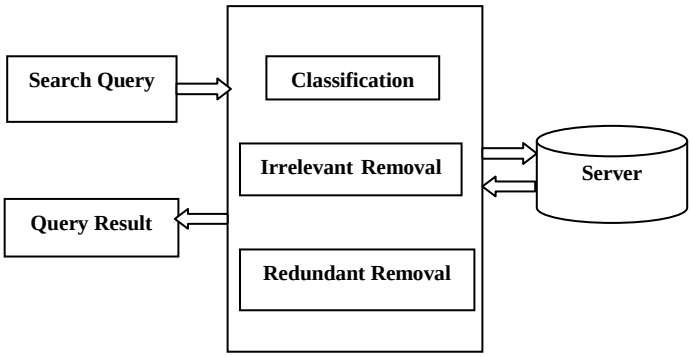


Fig1: Feature selection process

Feature subset selection can be viewed as the process of identifying and removing as many irrelevant and redundant features as possible. In existing work, a fast clustering-based feature selection algorithm is used. Features are divided into clusters by using graph-theoretic clustering methods. The most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent; the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features[5]. But it can't feature the different types of correlation measures.

Disadvantages

Previous feature selection algorithm can't explore correlation measures. Irrelevant and redundant data will be affect original data. So accuracy of the result of output can be affected.

Advantage:

The efficiently and effectively deal with both irrelevant and redundant features and also as different correlation measures, and obtain a good feature subset. In proposed work, traditionally, feature subset selection research has focused on searching single query feature subset selection. In our proposed work, i have plan to explore different types of correlation measures, using Correlation based feature selection algorithm. In this method, we can use multiple correlation/ query for feature subset selection.

II. RELATED WORK

Admin Database

The admin of particular database can process this, multiple files uploaded with the file or text form, also upload relevant image to text file. The files that were uploaded successfully are moved to a given directory. Users also can upload their files with required authentication. The class may reject files that exceed a given size limit. The descriptions are picked from the value of a form text field that is submitted with the file field data[1][6]. Various features are selected from database.

Correlation Based Feature Selection Algorithm

Inputs: D (F_1, F_2…, F_m, C) – the given data set

θ – the T-Relevance threshold

Output: S- Selected feature subset.

// Part 1: Classification//

D : Set of tuples

Each Tuple is an 'n' dimensional attribute vector

Let there be 'm' Classes : C1,C2,C3…Cm

Naïve Bayes classifier predicts D belongs to Class Ci

// Part 2: Irrelevant Feature Removal //

for i= 1 to m **do**

T- Relevance = SU (F_i, C)

If T-Relevance > θ **then**

S = SU{Fi};

//Part3: Representative Feature Selection //

Corr = S

for each Var in Corr do

If SU(Fi,Fj) ε Var then

Remove Corr[Fi,Fj]

Else

S = Var

Return S

Thus the above algorithm helps to differentiate that relevant and irrelevant features. Based on this process redundancy of relevant features can be identified, also that can be removed with selecting representative features at once.

Classification (Probability based naive bayes)

The classification refers to accuracy of target data. This classification can be applied with user's query. The Probability based naive bayes is one of the classifier which is used to classifying data based on the user's query. The result of this classification, helps to identifying with accuracy of original data from that various data which is stored in database[7].

Irrelevant Feature Removal (Correlation based feature selection algorithm)

The CFS algorithm obtains features relevant to the target concept by eliminating irrelevant from classifier results. The irrelevant feature removal is straightforward once the right relevance measure is defined or selected. The relevant features have strong correlation with target concept so are always necessary for a best subset. Feature subset selection can be the process that identifies and retains the strong irrelevant features and select relevant from feature classifiers[8]. This is a nonlinear estimation of correlation between feature values or feature values and target classes[Fig 1].

Redundant Feature Removal (Correlation based feature selection algorithm)

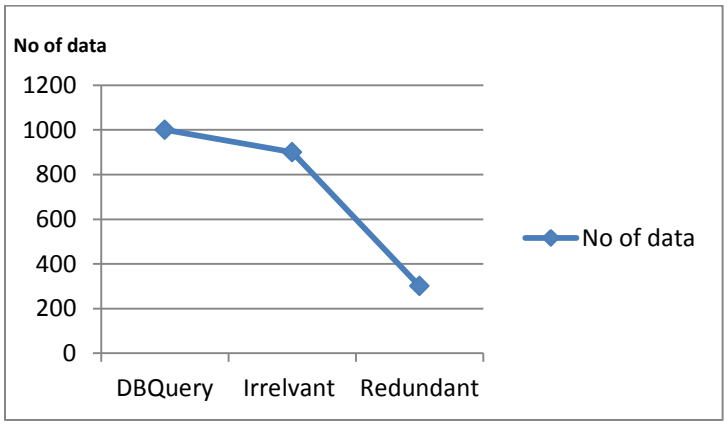
The latter removes redundant features from relevant ones via choosing representatives from classifier results and thus produces the final subset using CFS algorithm. The redundant feature elimination is a bit of sophisticated. Redundant features are assembled and representative feature can be taken out of the classifier results. Thus, notions of feature redundancy are normally in terms of feature correlation and feature-target concept correlation. As a result, only a very small number of discriminative features are selected[8][9].

Correlation Measures

Correlation Measures seek to quantify statistically how variables are closely related. Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases[10].

Results and Analysis

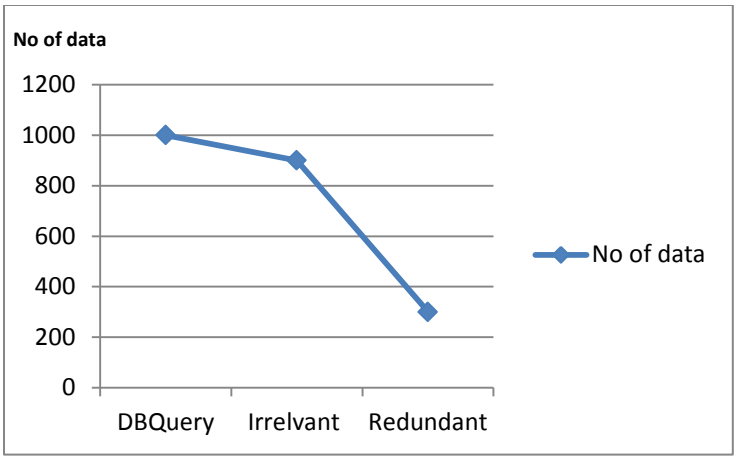
The previous feature selection algorithm has been provide data accuracy as well as good results as shown in following figure.



FAST Algorithm Result

This result has some percent of accuracy of data and limited process with irrelevant feature removal and redundant feature removal.

Thus the following figure will be consisting of that result of developed system. Based on these two results improvements of that feature selection concept can be proved.



Correlation Based Feature Selection Result

Comparison of Result

- The previous feature selection algorithm has limited points of result accuracy as 800 irrelevant data can be removed from that 1000 database query.
- Also 100 redundant data can b removed from that same 1000 db query.
- But, the developing present system provide result as 900 irrelevant data can be removed from that 1000 database query.
- Also 300 redundant data can be removed from that same 1000 db query.
- Result of that both previous and present feature selection algorithm can be determined as better data accuracy on result of user query.

III. CONCLUSION

We present that more concept on feature subset selection, different feature selection algorithm presenting that process of classifying features, removing irrelevant features. We have compared the various feature selection methods and algorithm to identifying features as well as Distributional Clustering of Words for Text Classification. We also found that Correlation based feature selection algorithm obtains the rank of text classification accuracy with different types of classifiers. Also different types of correlation measures helps to reducing that result page and targeted data result can be achieved. At the same time, FCBF,CFS and Relief-f is a good alternative for image and text data.

REFERENCES

[1] H. Almuallim and T.G. Dietterich, "Algorithms for Identifying Relevant Features," Proc. Ninth Canadian Conf. Artificial Intelligence, pp. 38-45, 1992.
[2] H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence, vol. 69, nos. 1/2, pp. 279-305, 1994.
[3] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "A Feature Set Measure Based on Relief," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.
[4] L.D. Baker and A.K. McCallum, "Distributional Clustering of Words for Text Classification," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
[5] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.
[6] D.A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.
[7] R. Battiti, "Irrelevant Features and the Subset Selection Problem," IEEE Trans. Neural Networks, vol. 5, no. 4, pp. 537-550, July 1994.
[8] A. Arauzo-Azofra, J.M. Benitez, and J.L. Castro, "An Extensive Empirical Study of Feature Selection Metrics for Text Classification," Proc. Fifth Int'l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004.

[9] L.D. Baker and A.K. McCallum, "An Introduction to Variable and Feature Selection," Proc. 21st Ann. Int'l ACM SIGIR Conf. Research and Development in information Retrieval, pp. 96-103, 1998.
[10] D.A. Bell and H. Wang, "Wrappers for Feature Subset Selection," Machine Learning, vol. 41, no. 2, pp. 175-195, 2000.

