International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-01

EFFECTIVE AND PRAGMATIC APPROACH TOWARDS CLUSTER ANALYSIS FOR MULTIDIMENSIONAL BIG DATA USING FEATURE SUBSET SELECTION

Lavisha
Research Scholar
Department of Computer Science & Engineering
Doon Valley Institute of Engineering & Technology
Karnal, Haryana, India

Sakshi Mehta
Assistant Professor
Department of Computer Science & Engineering
Doon Valley Institute of Engineering & Technology
Karnal, Haryana, India

ABSTRACT

Data mining comprise the huge set of data with enormous and high dimensional format because of the number of attributes can easily be in the unexpected and large volume. With this understanding and analysis with the nature of high-dimensional space, or hyperspace, is very important, especially when hyperspace is not behaving like more familiar geometry in multiple dimensions. Data is classically broken down for the analysis for a data warehouse into dimensions such as time period, product segment and the geographical location. Measurements are separated into classes. For time these could be months, quarters or years. Highlight subset choice is the procedure of selecting a subset of important element for the development of a model or better characterization and portrayal of the data. The center idea of the component subset choice strategy is that the crude data we are utilizing has numerous improper, excess and unessential elements. Repetitive elements are those components those give no data when contrasted with officially chose highlight and superfluous element can be considered as an element of no utilization in setting of the data. Highlight determination systems are the subset of the Feature extraction field. Highlight extraction makes new elements from the property set or the first components, while highlight determination gives back a subset of the elements. In this work, we have utilized Correlation based Feature determination for the diminishment of the measurements. Which makes us force random or some predefined limitations on the quantity of characteristics considered amid the development of a model. We have additionally taken qualities in thought as highlight as our decision or we can dispose of a characteristic taking into account the handiness of that particular trait for examination. In this research work, we come to know about the attributes we are selected by make use of the co-related algorithm in which subset co-relation is measured in a very effective manner.

KEYWORDS – Data Mining, Big Data, Multidimensional Data

INTRODUCTION

A lot of data are presently being utilized, handled and dissected as experimental, mechanical and numerous different structures because of a boom in the field of computerization and digitization procedures. This extensive measure of data goes about as gigantic and huge asset for the knowledge based revelation and choice emotionally supportive networks. For instance is the case of the case of shopping in a general store, when shopping is done in a grocery store, the trade administrator is locked in out the checking of standardized identification of things and store all the data about the thing shopped and exchange made into the database. The grocery store then utilizing this database can bring and extricate the knowledge based and imperative data for advertising by breaking down the business data in its shopping exchange database .

DATA MINING AND KNOWLEDGE DISCOVERY

In order to analyze, understand deeply, and to make use of the large amount of heterogeneous data, a multitier approach is to be used, this approach is known as Data Mining, It means that to mine some Knowledge from the raw data.

In simple cases and scenarios Data Mining refers to the unique process and paradigm for the identification of interesting and knowledge based patterns out of the abundant amount of data .

The Knowledge Discovery in databases process consists of the following steps leading from raw data to a form of new knowledge:

1. **Data cleaning:** it is sometimes also known as data refining. It is a step in which noisy or irrelevant data are removed from the collection of the database.
2. **Data integration:** at this step, multiple refined heterogeneous data sources may be combined together in a one form.
3. **Data selection:** at this step, a relevant data according to the knowledge we want to extract from given data for the analysis is selected and retrieved from the data collection.
4. **Data transformation:** it is the phase in which the selected data is transformed into a form that is appropriate for the mining process.

5. **Data mining:** it is the most important step in which some clever techniques are applied to extract the useful patterns or knowledge.
6. **Pattern evaluation:** in this step, strictly interesting patterns representing knowledge are identified based on given measures.
7. **Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

MULTI-DIMENSIONAL DATA

In data mining typically the data is very high dimensional or we can say that Big Data is sometimes difficult to understand, since the number of attributes can easily be in the hundreds or thousands. "Understanding the nature of high-dimensional space, or hyperspace, is very important, especially since hyperspace does not behave like the more familiar geometry in two or three dimensions.Data broken down in analysis for a data warehouse into dimensions such as time period, product segment and the geographical location. Dimensions are broken down into categories. For time these could be months, quarters or years.

They can represent multi dimensional data in the form of tabular form in one dimension, two dimension and d-dimension that is the authors can store the multidimensional data in multidimensional database in multi dimensions like for example in 3 dimension the authors can store the data in one of the form known as hypercubes. In the logical multidimensional model of the big data, a cube can represents the all of the measures with the same shape, that is, the exact same dimensions. In a cube shape, each edge represents a dimension. The dimension members are aligned on the edges and divide the cube shape into cells in which data values are stored.In an analytic workspace, the cube shape also represents the physical storage of multidimensional measures, in contrast with two-dimensional relational tables. An advantage of the cube shape is that it can be rotated: there is no one right way to manipulate or view the data. This is an important part of multidimensional data storage, calculation, and display, because different analysts need to view the data in different ways.

INTRODUCTION TO FEATURE SELECTION

Now we come to our main point that is on the basis of which we understand the basics of data mining so efficiently is Feature Selection. The amount of high-dimensional data that exists and is publically available on the internet has greatly increased in the past few years. Therefore, machine learning methods have difficulty in dealing with the large number of input features, which is posing an interesting challenge for researchers. In order to use machine learning methods effectively, pre-processing of the data is essential. Feature selection is one of the most frequent and important techniques in data pre-processing, and has become an essential component of the machine learning process. It is also known as variable selection, attribute selection, or variable subset selection in machine learning and statistics. It is the process of detecting relevant features and removing irrelevant, redundant, or noisy data. This process speeds up data mining algorithms, improves predictive accuracy, and increases comprehensibility. Irrelevant features are those that provide no useful information, and redundant features provide no more information than the currently selected features. In terms of supervised inductive learning, feature selection gives a set of candidate features using one of the three approaches:

1. The specified size of the subset of features that optimizes an evaluation measure
2. The smaller size of the subset that satisfies a certain restriction on evaluation measures
3. In general, the subset with the best commitment among size and evaluation measure

Therefore, the correct use of feature selection algorithms will leads us to the accurate results for selecting features improves inductive learning, either in term of generalization capacity, learning speed, or reducing the complexity of the induced model. In the process of feature selection, irrelevant and redundant features or noise in the data may be hinder in many situations, because they are not relevant and important with respect to the class concept such as microarray data analysis. When the number of samples is much less than the features, then machine learning gets particularly difficult, because the search space will be sparsely populated. Therefore, the model will not able to differentiate accurately between noise and relevant data. There are two major approaches to feature selection. The first is Individual Evaluation, and the second is Subset Evaluation. Ranking of the features is known as Individual Evaluation. In Individual Evaluation, the weight of an individual feature is assigned according to its degree of relevance. In Subset Evaluation, candidate feature subsets are constructed using search strategy.

Subset generation is relating to the search in which each of the state specifies a candidate subset for evaluation in the search space. Two basic issues determine the nature of the subset generation process. First, successor generation decides the search starting point, which influences the search direction. Second, search organization is responsible for the feature selection process with a specific strategy, such as sequential search, exponential search or random search. A newly generated subset must be evaluated by a certain evaluation criteria. Therefore, many evaluation criteria have been proposed in the literature to determine the goodness of the candidate subset of the features. Base on their dependency on mining algorithms, evaluation criteria can be categorized into groups: independent and dependent criteria. Independent criteria exploit the essential characteristics of the training data without involving any mining algorithms to evaluate the goodness of a feature set or feature. And dependent criteria involve predetermined mining algorithms for feature selection to select features based on the performance of the mining algorithm applied to the selected subset of features. Finally, to stop the selection process, stop criteria must be determined. Feature selection process stops at validation procedure. It is not the part of feature selection process, but feature selection method must be validate by carrying out different tests and comparisons with previously established results or comparison with the results of competing methods using artificial datasets, real world datasets, or both.

The relationship between the inductive learning method and feature selection algorithm infers a model. There are three general approaches for feature selection. First, the Filter Approach exploits the general characteristics of training data with independent of the mining algorithm. Second, the Wrapper Approach explores the relationship between relevance and optimal feature subset selection. It searches for an optimal feature subset adapted to the specific mining algorithm. And third, the Embedded Approach is done with a specific learning algorithm that performs feature selection in the process of training.

Feature selection methods

Many, feature selection methods have been proposed in the literature, that can be used and their comparative study is a very difficult task. Without knowing the relevant features in advance of the real data set, it is very difficult to find out the effectiveness of the feature selection methods, because data sets may include many challenges such as the huge number of irrelevant and redundant features, noisy data, and high dimensionality in term of features or samples. Therefore, the performance of the feature selection method relies on the performance of the learning method. There are many performance measures mentioned in the literature such as accuracy, computer resources, ratio of feature selection, etc. Most researchers agree that there is no so-called “best method”. Therefore; the new feature selection methods are constantly increasing to tackle the specific problem (as mentioned above) with different strategies

1. To ensure a better behavior of feature selection using an ensemble method
2. Combining with other techniques such as tree ensemble and feature extraction
3. Reinterpreting existing algorithms
4. Creating a new method to deal with still-unresolved problems
5. To combine several feature selection methods

Many good studies of existing feature selection methods have been done in the literature, for example, an experimental study of eight filter methods (using mutual information) is used in 33 datasets and for the text classification problem, 12 feature selection methods are compared. The capability of the survival ReliefF algorithm (sReliefF) and tuned sReliefF approach are evaluated. Seven filters, two embedded methods, and two wrappers are applied in 11 synthetic datasets (tested by four classifiers), which are used for comparative study of feature selection performances in the presence of irrelevant features, noise in the data, redundancy, and the small ratio between the number of attributes and samples. Related to the high-dimensional dataset (in both samples and attributes), the performance of feature selection methods are studied for the multiple-class problem.

In a theoretical perspective, guidelines by which we can select feature selection algorithms are presented, where algorithms are categorized based on three perspectives, namely search organization, evaluation criteria, and data mining tasks. In characterizations of feature selection algorithms are presented with their definitions of feature relevance. In the application perspective, many real-world applications like intrusion detection], text categorization DNA microarray analysis music information retrieval image retrieval information retrieval customer relationship management, Genomic analysis and remote sensing are considered. We can understand the daily related problems in an very effiecient manner.As some methods are given below to learn them in a good manner.

LITERATURE SURVEY

Shima Kashef, An Advanced ACO Algorithm for Feature subset Selection, 2015, "In this paper, the authors present a new feature selection technique based on Ant Colony Optimization (ACO) by combining two models of ACO. The proposed algorithm has a strong search capability in the problem space and can efficiently find a minimal feature subset. This algorithm is compared with some powerful algorithms, including IBGSA, CatfishBPSO, ACOFS, BACO, ACO with and without heuristic desirability, BGA and BPSO. In order to evaluate the performance of these approaches, experiments were performed using twelve datasets from the UCI machine learning repository. The experimental results confirm our algorithm and provide obvious evidences, allowing us to conclude that our method achieves a better feature set in terms of classification accuracy and number of selected features. Further investigation on the parameters values and testing the ABACO model with other heuristic functions are an area of future research.

Arnab Roy et al., 2014 This paperthey have introduced two crossover operators, MMX-BLX exploit andMMX-BLXexplore, for simultaneously solving multiple feature/subset selection problems where the features may have numeric attributes and the subset sizes are not predefined. These operators differ on the level of exploration and exploitation they perform; one is designed to produce convergence controlled mutation and the other exhibits a quasiconstant mutation rate. They illustrate the characteristic of these operators by evolving pattern detectors to distinguish alcoholics from controls using their visually evoked response potentials (VERPs).

Min Wei et al., 2014 In this research paperConventional mutual information (MI) based feature selection (FS) methods are unable to handle heterogeneous feature subset selection properly because of data format differences or estimation methods of MI between feature subset and class label. A way to solve this problem is feature transformation (FT). In this study, a novel unsupervised feature transformation (UFT) which can transform non-numerical features into numerical features is developed and tested. The UFT process is MI-based and independent of class label. MI-based FS algorithms, such as Parzen window feature selector (PWFS), minimum redundancy maximum relevance feature selection (mRMR), and normalized MI feature selection (NMIFS), can all adopt UFT for pre-processing of non-numerical features.

Guangtao Wang et al., 2014 In this research paperMany feature subset selection (FSS) algorithms have been proposed, but not all of them are appropriate for a given feature selection problem. At the same time, so far there is rarely a good way to choose appropriate FSS algorithms for the problem at hand. Thus, FSS algorithm automatic recommendation is very important and practically useful. In this paper, a meta learning based FSS algorithm automatic recommendation method is presented. The proposed method first identifies the data sets that are most similar to the one at hand by the k-nearest neighbor classification algorithm, and the distances among these data sets are calculated based on the commonly-used data set characteristics.

Dongsong Zheng et al., 2013 In this research paperFeature subset selection, as an important preprocessing step to knowledge discovery and machine learning, is effective in reducing irrelevant features, compressing repeated data, and improving classification accuracy. Rough set theory is important tool in selecting feature subset of large-scale data. In this work, fuzzy rough feature subset selection concept is introduced, and the efficient fuzzy rough measure of feature significance measure is designed, what's more, a quick filter feature selection approach which can efficiently identify relevant features as well as redundancy among relevant features with fuzzy rough approach, is presented.

M. Akhil Jabbar et al., 2013 In thispaperHeart disease is the leading cause of death in India and worldwide. India is in the middle of a major economic and industrial transition. The life style changes have led to rise in hypertension, obesity, smoking, diabetes and in turn heart disease. Disease diagnosis often done based on doctors experience and personal opinion rather than the data hidden in the medical data base, which leads to wrong diagnosis and increases diagnosis costs which in turn affects the quality of services provided by hospitals to the patients. Medical data mining is to search knowledgeable data for effective medical diagnosis.K nearest neighbor is one of the widely used data mining technique in classification. It is a straight forward classifier where samples are classified based on the class of their nearest neighbor .Medical data bases are high volume in nature.

A. Srikrishna et al., 2013 In this research paper Feature subset selection is a process of selecting a subset of minimal, relevant features and is a pre processing technique for a wide variety of applications. High dimensional data clustering is a challenging task in data mining. Reduced set of features helps to make the patterns easier to understand. Reduced set of features are more significant if they are application specific. Almost all existing feature subset selection algorithms are not automatic and are not application specific. This paper made an attempt to find the feature subset for optimal clusters while clustering.

Qianhong Wu et al., 2013 In thispaperFeature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. 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. Based on these criteria, a fast clustering-based feature selection algorithm, FAST, is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features.

Guorong Li et al., 2012 In this research paperthey propose a new feature subset evaluation method for feature selection in object tracking. According to the fact that a feature which is useless by itself could become a good one when it is used together with some other features, they propose to evaluate feature subsets as a whole for object tracking instead of scoring each feature individually and find out the most distinguishable subset for tracking. In the paper, they use a special tree to formalize the feature subset space. Then conditional entropy is used to evaluating feature subset and a simple but efficient greedy search algorithm is developed to search this tree to obtain the optimal k-feature subset quickly. Furthermore, their online k-feature subset selection method is integrated into particle filter for robust tracking. Extensive experiments demonstrate that kfeature subset selected by their method is more discriminative and thus can improve tracking performance considerably.

Tao Chen et al., 2012 In this research paper existing models for cluster analysis typically consist of a number of attributes that describe the objects to be partitioned and one single latent variable that represents the clusters to be identified. When one analyzes data using such a model, one is looking for one way to cluster data that is jointly defined by all the attributes. In other words, one performs one-dimensional clustering. This is not always appropriate. For complex data with many attributes, it is more reasonable to consider multidimensional clustering, i.e., to partition data along multiple dimensions. In this paper, they present a method for performing multidimensional clustering on categorical data and show its superiority over uni-dimensional clustering.

Quanquan Gu et al., 2012 In thispaperFisher score is one of the most widely used supervised feature selection methods. However, it selects each feature independently according to their scores under the Fisher criterion, which leads to a suboptimal subset of features. In this paper, they present a generalized Fisher score to jointly select features. It aims at finding a subset of features, which maximize the lower bound of traditional Fisher score. The resulting feature selection problem is a mixed integer programming, which can be reformulated as a quadratically constrained linear programming (QCLP). It is solved by cutting plane algorithm, in each iteration of which a multiple kernel learning problem is solved alternatively by multivariate ridge regression and projected gradient descent.

John Q.Gan et al., 2011 Thispaperin brain-computer interface (BCI) development, temporal/spectral/ spatial/statistical features can be extracted from multiple electroencephalography (EEG) signals and the number of features available could be up to thousands. Therefore, feature subset selection is an important and challenging problem in BCI design.

Sequential forward floating search (SFFS) has been well recognized as one of the best feature selection methods. This paper proposes a filter-dominating hybrid SFFS method, aiming at high efficiency and insignificant accuracy sacrifice for high-dimensional feature subset selection.

Deng Cai et al., 2010 in this research paperin many data analysis tasks, one is often confronted with very high dimensional data. Feature selection techniques are designed to find the relevant feature subset of the original features which can facilitate clustering, classification and retrieval. In this paper, they consider the feature selection problem in unsupervised learning scenario, which is particularly difficult due to the absence of class labels that would guide the search for relevant information. The feature selection problem is essentially a combinatorial optimization problem which is computationally expensive.

Zheng Zhaoet al., 2009 In this research paper the evolving and adapting capabilities of robust intelligence are best manifested in its ability to learn. Machine learning enables computer systems to learn, and improve performance. Feature selection facilitates machine learning (e.g., classification) by aiming to remove irrelevant features. Feature (attribute) interaction presents a challenge to feature subset selection for classification. This is because a feature by itself might have little correlation with the target concept, but when it is combined with some other features; they can be strongly correlated with the target concept. Thus, the unintentional removal of these features may result in poor classification performance.

Hong Zeng et al., 2008 This paper addresses the problem of feature selection for the high dimensional data clustering. This is a difficult problem because the ground truth class labels that can guide the selection are unavailable in clustering. Besides, the data may have a large number of features and the irrelevant ones can ruin the clustering. In this paper, they propose a novel feature weighting scheme for a kernel based clustering criterion, in which the weight for each feature is a measure of its contribution to the clustering task. Accordingly, they give a well-defined objective function, which can be explicitly solved in an iterative way.

Gert Van Dijcket al., 2006 In this paperhybrid filter/wrapper feature subset selection algorithm for regression is proposed. First, features are filtered by means of a relevance and redundancy filter using mutual information between regression and target variables. They introduce permutation tests to find statistically significant relevant and redundant features". Second, a wrapper searches for good candidate feature subsets by taking the regression model into account.

Problem Formulation

Feature selection refers to the involvement of identification of subset of assorted and key based features which produces the most compatible results as main and original entire set of features to be extracted. "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. Based on these criteria, a clustering-based feature selection algorithm is proposed and experimentally evaluated in their work."

The algorithm works in two steps.

1. "In the first step, features are divided into clusters by using graph-theoretic clustering methods.
2. In the second step, 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 the algorithm has a high probability of producing a subset of useful and independent features. To ensure the efficiency of the algorithm, the authors adopt the efficient minimum-spanning tree clustering method. The efficiency and effectiveness of the algorithm are evaluated through an empirical study.

"Extensive experiments will be carried out to compare the algorithm and several representative feature selection algorithms, including, Fast Correlation-Based Filter (FCBF), Relief, CFS. (Correlation based Feature Selection), with respect to various types of well-known classifiers, namely, the probability-based Naive Bayes, the tree-based, the instance-based, and the rule-based Classifiers before and after feature selection application.

To ensure the efficiency of the algorithm, the authors adopt the efficient minimum-spanning tree clustering method."

Research Objectives

"Feature selection or variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features.

Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. In this research work following objectives are selected,

Result And Analysis

The Supermarket dataset

The data is nominal and each instance represents a customer transaction at a supermarket, the products purchased and the departments involved. The attributes are aggregated to the department level in the supermarket, so and a value of t" indicates that the customer's shopping cart contained at least one product from that department. The data contains 4,627 instances and 217 attributes. The data is denormalized. Each attribute is binary and either has a value ("t" for true) or no value ("?" for missing). There is a nominal class attribute called "total" that indicates whether the transaction was less than $100 (low) or greater than $100 (high).

```
@relation supermarket
@attribute 'department1'{t}
@attribute 'department2'{t}
@attribute 'department3'{t}
@attribute 'department4'{t}
…
@attribute 'Baby needs' { t}
```

@attribute 'tea' { t}
…@attribute 'total'{low, high}% low<100
@data
f,f,f,f,f,f,f,f,f,f,f,t,t,t,f,t,f,t,f,f,t,f,f,f,t,t,t,t,f,t,f,t,t,f,f,f,f,f,
f,t,t,t,f,f,f,f,f,f,f,t,…… high

Screen Shots

The Correlation based Feature Subset algorithm Arguments, Max Number of Subsets taken is 5.

Possible Search Methods:

1. Best First
2. Greedy

Data File: Supermarket.arff.

In this, Screenshot we are describing the project properties of Supermarket Dataset that we are taken as the arff file extension format.We generally use this screenshot to represent the file format that we are taking to evaluate the file

CFS Evaluator for Supermarket Data Set

In these Screenshots,we are showing the running attributes of the supermarket dataset

Table 1 SuperMarket Dataset

Supermarket Dataset	
CFS Based Cluster	63.713%
Greedy Based Cluster	55.857%
Mean Absolute Error	0.362%
Time Elapsed(CFS)	243
Time Elapsed(Greedy)	1090

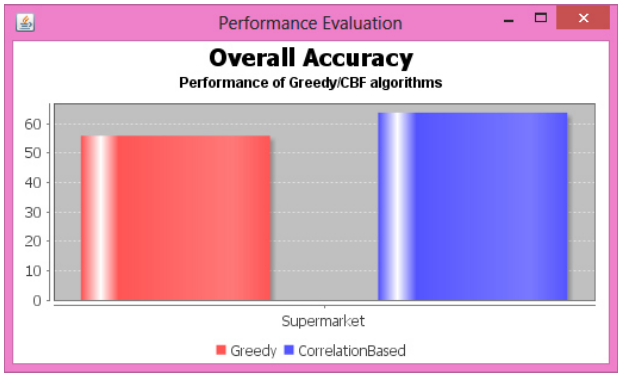

Figure 1 - Performance Greedy/CBF algorithm in SuperMarket Dataset

Data File:Onlineretail.arff

In this database generally ,we get the details of data in Boolean form that is either in True or false format with some list of attributes that will be used to find out the details of online shopping done by the customer.We compare these outputs on the basis of certain which attributes will selected from the given attributes.

This database values are used to evaluate the accuracy ,time taken & mean absolute error taken by the both of the algorithms shown in Table.

Retail Dataset

In these Screenshots,we are showing the running attributes of the online retail dataset,this datset includes the almost 20 attributes taken by the naïve baise classifiers & the CBF algorithm to evaluate performance by the different parameters like accuracy, time, mean abosulte error.These are actually the final running state of the feature subset selection.Here we are showing the table of Online

Retail Dataset to evaluate the accuracy,time taken by the two different algorithms.

Table 2 Online Retail dataset

Online Retail Dataset	
CFS Based Cluster	76.5%
Greedy Based Cluster	47.287%
Mean Absolute Error	0.1776%
Time Elapsed(CFS)	32
Time Elapsed(Greedy)	140

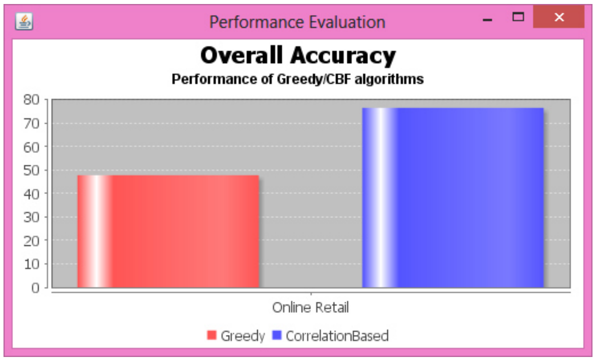

Figure 2 - Performance Greedy/CBF algorithm in Online Retail Dataset.

Conclusion and Future Work

Feature subset selection is basically the process for selecting a subset of useful(relevant) feature for the construction of a model or better classification and description of the data. The core concept of the feature subset selection technique is that the raw data we are using has many un appropriate, redundant and irrelevant features. Redundant features are those features those provide no information as compared to already selected feature and irrelevant feature can be considered as a feature of no use in context of the information. Feature selection techniques are the subset of the Feature extraction field. Feature extraction makes new features from the attribute set or the original features, whereas feature selection returns a subset of the features.In machine learning and statistics, feature selection plays a very important role as a process for the selection of the relevant features from the pool of features for use in model construction and for the better classification and understanding of the data.

In this work, we have employed Correlation based Feature selection for the reduction of the dimensions. Which makes us impose random or some predefined constraints on the number of attributes considered during the construction of a model. We can also take an attribute in consideration as feature as our choice or we can discard an attribute based on the usefulness of that specific attribute for analysis.

Extensive experiments were carried out to compare the algorithm and several representative feature selection algorithms, including well-known classifiers, namely, the Best First, the tree-based, the instance-based, and the Greedy Step before and after feature selection application. The research has found that CFS Subset algorithm is fast and can reduce dimensionality of underlying dataset very quickly. For instance the CFS algorithm reduced the time taken by the greedy algorithm & will produce more accuracy only single thread and a single thread pool.

Feature selection mainly involves the identification of a subset of the most useful features that produces almost similar results as the original entire set of features. The efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a Correlation based feature selection algorithm was proposed and experimentally evaluated in our work.

For the future work, we plan to explore different types of correlation measures, and study some formal properties of feature space. The work reported in this thesis assumes noise-free training data but in future we will try to evaluation real world datasets. A direct way to deal with classification noise is to modify the given algorithms by relaxing the requirement of explainingall the conflicts generated from the training data. Other future direction is to develop faster algorithms for tree analysis. Another second direction is to extend the work presented to cover continuous and mixed data.

References

[1] "Kashef, Shima, and Hossein Nezamabadi-pour. An advanced ACO algorithm for feature subset selection. Neurocomputing 147 (2015): 271-279

[2] Arnab Roy, J. David Schaffer, and Craig B. Laramee. New crossover operators for multiple subset selection tasks. arXiv preprint arXiv:1408.1297 (2014).

[3] Min Wei, Tommy WS Chow, and Rosa HM Chan. Mutual Information-Based Unsupervised Feature Transformation for Heterogeneous Feature Subset Selection. arXiv preprint arXiv:1411.6400 (2014).

[4] Guangtao Wang, Qinbao Song, Heli Sun, Xueying Zhang, Baowen Xu, and Yuming Zhou. A feature subset selection algorithm automatic recommendation method. arXiv preprint arXiv:1402.0570 (2014).

[5] Wenger, Etienne. Artificial intelligence and tutoring systems: computational and cognitive approaches to the communication of knowledge. Morgan Kaufmann, 2014.

[6] Kumar, Vipin, and Sonajharia Minz. Feature Selection. SmartCR 4, no. 3 (2014): 211-229

[7] Dongsong Zheng, and Changsheng Zhang. Selecting Feature Subset for Large-scale Data via Fuzzy Rough Approach. Journal of Convergence Information Technology 8, no. 9 (2013).

[8] M. Akhil Jabbar, B. L. Deekshatulu, and Priti Chandra. HEART DISEASE CLASSIFICATION USING NEAREST NEIGHBOR CLASSIFIER WITH FEATURE SUBSET SELECTION. Anale. Seria Informatica 11 (2013).

[9] A. Srikrishna, B. Eswara Reddy, and V. Sesha Srinivas. Automatic Feature Subset Selection using Genetic Algorithm for Clustering. International Journal on Recent Trends in Engineering & Technology 9, no. 1 (2013).

[10] Qinbao Song, Jingjie Ni, and Guangtao Wang. A fast clustering-based feature subset selection algorithm for high-dimensional data. Knowledge and Data Engineering, IEEE Transactions on 25, no. 1 (2013): 1-14.

[11] Baskar, S. S., L. Arockiam, and S. Charles. A Systematic Approach on Data Pre-processing In Data Mining. Compusoft 2, no. 11 (2013): 33

[12] Guorong Li, Qingming Huang, Junbiao Pang, Shuqiang Jiang, and Lei Qin. Online selection of the best k-feature subset for object tracking. Journal of Visual Communication and Image Representation 23, no. 2 (2012): 254-263.

[13] Tao Chen, Nevin L. Zhang, Tengfei Liu, Kin Man Poon, and Yi Wang. Model-based multidimensional clustering of categorical data. Artificial Intelligence 176, no. 1 (2012): 2246-2269.

[14] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725 (2012).

[15] John Q.Gan,Bashar Awwad Shiekh Hasan, and Chun Sing Louis Tsui. A hybrid approach to feature subset selection for brain-computer interface design. In Intelligent Data Engineering and Automated Learning-IDEAL 2011, pp. 279-286. Springer Berlin Heidelberg, 2011.

[16] LaValle, Steve, Eric Lesser, Rebecca Shockley, Michael S. Hopkins, and Nina Kruschwitz. Big data, analytics and the path from insights to value. MIT sloan management review 52, no. 2 (2011): 21.

[17] Deng Cai, Chiyuan Zhang, and Xiaofei He. Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 333-342. ACM, 2010.

[18] Zheng Zhao and Huan Liu. Searching for interacting features in subset selection. Intelligent Data Analysis 13, no. 2 (2009): 207-228.

[19] Katakis, Ioannis, Grigorios Tsoumakas, Evangelos Banos, Nick Bassiliades, and Ioannis Vlahavas. An adaptive personalized news dissemination system. Journal of Intelligent Information Systems 32, no. 2 (2009): 191-212

[20] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18

[21] Hong Zeng and Yiu-ming Cheung. Feature selection for clustering on high dimensional data. In PRICAI 2008: Trends in Artificial Intelligence, pp. 913-922. Springer Berlin Heidelberg, 2008.

[22] Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan et al. Top 10 algorithms in data mining. Knowledge and information systems 14, no. 1 (2008): 1-37.

[23] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. Supervised machine learning: A review of classification techniques. (2007): 3-24.

[24] Gert Van Dijck and Marc M. Van Hulle. Speeding up the wrapper feature subset selection in regression by mutual information relevance and redundancy analysis. In Artificial Neural Networks–ICANN 2006, pp. 31-40. Springer Berlin Heidelberg, 2006

[25] Peng, Hanchuan, Fuhui Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27, no. 8 (2005): 1226-1238.

[26] Witten, Ian H., and Eibe Frank. Data Mining: Practical machine learning tools andtechniques. Morgan Kaufmann, 2005

[27] Inmon, William H. Building the data warehouse. John wiley & sons, 2005.

[28] Yu, Lei, and Huan Liu. Efficient feature selection via analysis of relevance and redundancy. The Journal of Machine Learning Research 5 (2004): 1205-1224.

[29] Guyon, Isabelle, and André Elisseeff. An introduction to variable and feature selection. The Journal of Machine Learning Research 3 (2003): 1157-1182.

[30] Abelló, Alberto, José Samos, and Félix Saltor. A framework for the classification and description of multidimensional data models. In Database and Expert Systems Applications, pp. 668-677. Springer Berlin Heidelberg, 2001.

[31] Stefanovic, Nebojsa, Jiawei Han, and Krzysztof Koperski. Object-based selective materialization for efficient implementation of spatial data cubes. Knowledge and Data Engineering, IEEE Transactions on 12, no. 6 (2000): 938-958.

[32] Zhang, Guoqiang Peter. Neural networks for classification: a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 30, no. 4 (2000): 451-462.

[33] Hall, Mark A. Correlation-based feature selection for machine learning. PhD diss., The University of Waikato, 1999.

[34] Friedman, Nir, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. Machine learning 29, no. 2-3 (1997): 131-163.

[35] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. AI magazine 17, no. 3 (1996): 37.

[36] Chen, Ming-Syan, Jiawei Han, and Philip S. Yu. Data mining: an overview from a database perspective. Knowledge and data Engineering, IEEE Transactions on 8, no. 6 (1996): 866-883.

[37] Benjamin, Robert, and Rolf Wigand. Electronic markets and virtual value chains on the information superhighway. Sloan Management Review 36, no. 2 (1995): 62.

[38] Moore, Andrew W., and Mary S. Lee. Efficient Algorithms for Minimizing Cross

Validation Error. In ICML, pp. 190-198. 1994.

[39] John, George H., Ron Kohavi, and Karl Pfleger. Irrelevant features and the subset selection problem. In Machine learning: proceedings of the eleventh international conference, pp. 121-129. 1994.

[40] Kanellakis, Paris C., Gabriel M. Kuper, and Peter Z. Revesz. Constraint query languages (preliminary report). In Proceedings of the ninth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, pp. 299-313. ACM, 1990".