International Journal of Advanced Research in

Computer Science Engineering and Information Technology

Volume: 2 Issue: 3 08-Apr-2014,ISSN_NO: 2321-3337

Fast and Improved Feature subset selection Algorithm Based Clustering for High Dimensional Data

K. Vijayalakshmi[1] , Dr .S. Anithaa [2] , Dr.B.Raghu [3]

PG Scholar, Department of CSE, Sri Ramanujar Engineering College, Chennai, India
Professor, Department of CSE, Sri Ramanujar Engineering College, Chennai, India
Professor, Department of CSE, Sri Ramanujar Engineering College, Chennai, India
kviji28@yahoo.co.in, anithabennett@yahoo.co.in , raghubalraj@gmail.com

ABSTRACT—*The Clustering is a method of grouping the information into modules or clusters. Their dimensionality increases usually with a tiny number of dimensions that are significant to definite clusters, but data in the unrelated dimensions may produce much noise and wrap the actual clusters to be exposed. Attribute subset selection method is frequently used for data reduction through removing unrelated and redundant dimensions (or attribute). The Ant colony optimization technique is used for solving computational problems which can be reduced to find good path during graphs in the minimum spanning tree problem and traveling salesman problem. In my paper the feature subset selection algorithm and Ant colony optimization algorithm are employed to improve the feature subset selection. The proposed method helps us in improved feature subset selection algorithm based on hierarchical cluster and it's minimizes redundant data set and improves the attribute subset accuracy.*

Keywords— Hierarchical clustering based algorithm-filter technique,Graph-based cluster.

1. INTRODUCTION

1.1 DATA MINING

Data mining refers to the procedure or technique that extracts the data from large amounts of information, Information mining is frequently treated like a synonym added traditionally used term for Knowledge Discovery in Databases (KDD or IDD)[5].

The main steps of KDD are information maintenance, information combination, records selection, Data conversion, data removal, prototype evaluation, data presentation and knowledge discovery. The nontrivial procedure of identifying suitable, new, potentially helpful and finally reasonable patterns of importance in information mining is the main process of IDD. Data classification consists of supervised learning and non-supervised learning.

1. In supervised learning, the class label of every preparation tuple is provided,

2. In unsupervised learning the class label of each training tuple is not known

A major aim is to forecast the group $f_i = g(x_1,\ldots,x_n)$, where $x_1,\ldots x_n$ are input attributes. There is one well-known feature called as responsibility attribute. The enter to the feature subset selection algorithm is a data set of instruction records through a number of attributes .The nontrivial procedure of identifying suitable, new, potentially helpful and finally reasonable patterns of importance in data.

1.2 Data Reduction

The data set determination is expected to be vast, imagine that you have chosen data from the all electronic data warehouse for study. The Complex data examination and removal on vast amount of data is time consuming process, such an analysis is feasible to overcome the complexity a Novel information reduction techniques which can be practical to increase a reduced symbol of the data set that is much smaller in dimensions, until now closely maintains the reliability of the novel data. That is removal on the reduced dataset should be more well-organized and produce the same (or almost the same) logical outcome an information reduction strategies comprises dimensionality reduction, numerosity reduction and information compression. The Dimensionality reduction (or) attribute subset selection is the method to reduce the amount of variables or attributes under kindness, in which irrelevant, weakly relevant or unneeded or dimensions are detected and removed.

1.3 Attribute Subset selection

The Attribute division collection is used to decrease the information set volume by removing unrelated or unneeded attributes (or dimensions).The goal in element division collection is to discover a least set of attributes such that the possibility of allocation of data classes is as close as possible to the unique distribution obtained using all elements. Removal on a summary position of attribute has a supplementary advantage, in summary the amount of attributes appearing in the exposed pattern, that are selected to create the patterns and easier to recognize.

1.4 Literature Review

a) Finding the nearest neighbour graph using Fast Agglomerative clustering.

Pasic Frantic et al have introduced a new algorithm for Fast Agglomerative clustering using an approximate k-nearest neighbour graph for reducing the number of distance calculations [1].The time complexity of the algorithm is improved at the cost of a slight increase in deformation. There are no theoretical grounds for how to fix the exact neighbourhood size optimally, but there is one guideline that should be followed: The connectivity of the vectors within the clusters should be preserved. The bottleneck of the algorithm is the graph creation and it remains a challenge to invent a practical algorithm for creating a reasonably accurate k-NN graph in sub quadratic time. The main drawback of this procedure is due to less number of several data set tests are quite different and small neighbourhood size is sufficient to continue the value close to that of the full search.

b) Statistical comparison of many feature subset selection methods over Multiple Data Sets

Sinbad song et al they have compared Many feature subset selection methods incorporates feature selection and is embedded methods as a part of the preparation process that are usually definite to prearranged information algorithms, as an effect it may be further well-organized than the other three category. The straight part of set of tools learns algorithms alike to selection grass or imitation neural network are representation of predetermined travel towards. The covering methods use the analytical correctness of a prearranged knowledge algorithm to conclude the integrity of the particular subsets, the correctness of the knowledge algorithms is frequently high, but the computational difficulty is large and it is simplification of the particular features is restricted.

A good majority through a filter process are self-determining of knowledge algorithms. The correctness of the education algorithms is not guaranteed,_other than the low computational difficulty. Their hybrid methods are a grouping of filter and wrapper methods through with a filter process to decrease explore gap that will be measured by the successive wrapper. The mostly focus on combining filter and wrapper process to achieve the best feasible presentation with an exacting learning algorithm with parallel time difficulty of the filter process. The main disadvantage of the system is it determine to low correctness and common of the assured type is limited along with the computational difficulty is huge.

c) The Global Kernel k-means Algorithm for Clustering in Feature Space

Grigorios et al they proposed the global kernel K-means clustering algorithm that method to maps data points from input space to a higher dimensional feature space through the use of a kernel function and optimizes the clustering error in the feature space by locating near-optimal solutions[2]. The deterministic nature is the advantages of this method which makes it independent of cluster initialization and the ability to identify nonlinearly separable clusters in input space the high computational complexity is main drawback of the method.

d) Integrating clustering with supervised learning for Categorical Data Analysis

Ujjwal maulik et al they proposed differential evolution based fuzzy c-medoids clustering. It faces many problems in algorithm effectively optimizes the FCMdd error function globally. For demonstrating the superiority of the algorithm, its performance has been compared with those of DEFCMdd clustering, GAFCMdd clustering, and SAFCMdd clustering for four synthetic and four real life data sets [4].

Statistical significance tests based on Wilcoxon’s rank sum test have been conducted to judge the statistical significance of the clustering solutions produced by different algorithms. To improve the performance of clustering further, an SVM classifier is trained with a fraction

of data points selected from each clusters based on the proximity to solve clustering problems where the number of clusters is not known, the clustering problem of categorical data can be modelled as a multi objective optimization problem using the concept of amount of dominance. A sensitivity analysis of the developed technique with respect to different setting of the parameters, including the fraction of the points to be used for training the SVM, needs to be carried out. It is recalculates all distances at each iteration of the algorithm and its slowness does not generalized to higher dimensional data which reads to disadvantage of the system.

e) Subspace and Projected Clustering Algorithms of Clustering High Dimensional Data.

Rahmat et al have presented the dissimilar difficulty declaration: Clustering has a quantity of techniques that have been developed in information, prototype discovery, information removal, and other ground [3]. Subspace cluster, catalogues cluster of substance in all subspaces of a dataset. It tends on the way to produce several more than lapping clusters, move towards: Subspace clustering and future clustering areas are possible to explore and used for clustering in big dimensional places, Projected and subspace clustering can compute both disjoint and overlapping clusters in high dimensional data included both feature transformation and feature selection techniques. The main drawback is these methods is that it is difficult to differentiate similar data points from dissimilar once and limitation of distance between any two data points becomes almost the same.

d) Ant Colony Based On Color, Text Reduction

[9]Zhong et al., applied irregular set through Heuristics (RSH) and unbalanced set from beginning to end Boolean study (RSBR) for feature selection and discretization of valid-respected attribute. The main drawback is these methods is that it is difficult to differentiate similar attribute data from dissimilar once.

1.4.1 Solution for the above problem

Using Ant Colony optimization techniques To improve the performance of algorithm for cluster based feature subset selection algorithms by considering High dimensional data and To reduce the time complexity and Improve Accuracy of clustering for High dimensional data.

The rest of the paper is planned since follows: division 2 briefs as regards the information set used and records grounding for this study. Division 3 describes the feature reduction algorithm using Improved Feature subset selection Algorithm and its implementation. Division 4 describe the Ant Colony Algorithm in the framework of organization as Outcome are discussed in division 5 and the paper is concluded in division 6.

2. INFORMATION PREPARATION

The various Organization data sets were. Obtained from machine learning repository of UCI [2]. In addition we have composed and used Organization Database Set which are used for this study. The Organization database consists of information collected from the different Organizations .The improvement of this information set is that it includes enough number of records of dissimilar category of individuals affected by missing values. The set of descriptors present every one, the necessary information about Employees.

It contains the files of 5000 employees. The file of every Employees contains 50 attributes and this has summary of 23 attributes subsequent to verify with the unacceptable values. The permanent attributes of all the data sets are discredit previous to apply it to the Ant colony optimization ("Miner").

3. IMPROVED FEATURE SUBSET SELECTION ALGORITHMS

A Feature subset selection algorithm is a method to identify and remove as much unrelated and disused information as potential. This reduces the dimensionality of the information and may possibly agree to train algorithms, to work closer and more efficiently.

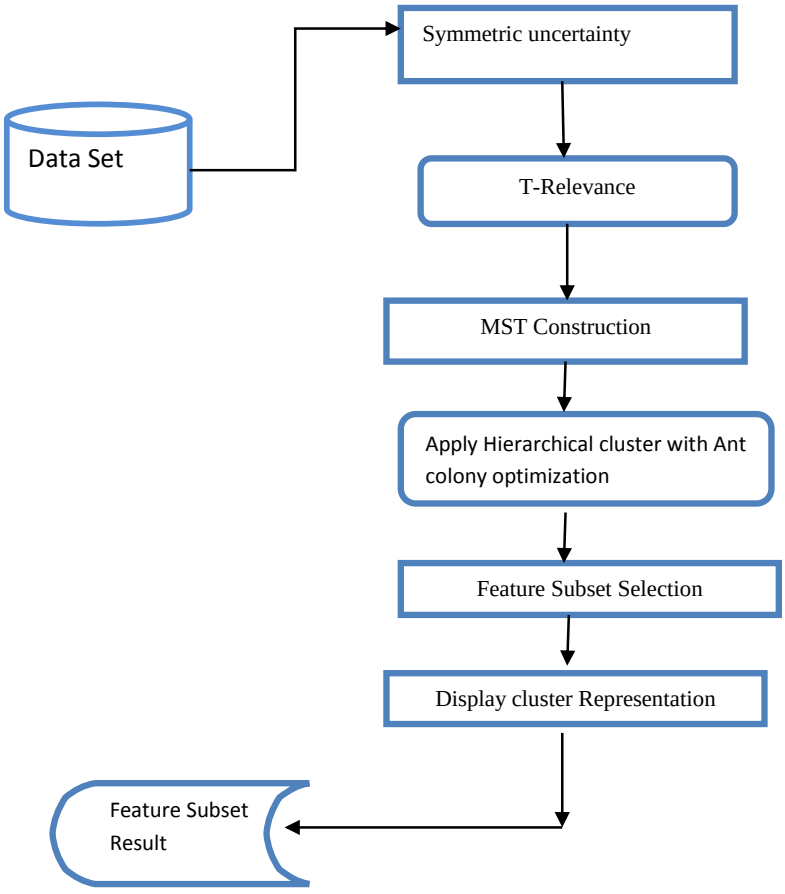

Figure 1: Framework of the Hierarchical based

In some cases correctness on potential arrangement can be better, in others the result is more compressed, easily interpreted the demonstration of the objective idea. Unrelated features, along with redundant description, severely affect the correctness of the knowledge equipment. Thus, attribute subset selection should be able to recognize and eliminate as much of the unrelated and redundant information as feasible solution. The group indexing and text assignments are repeated regularly to provide reverse combine and to continue an up-to-date clustering resolution. The distance-based clustering technique to develop a real time and online system for a particular company to predict sales in various annual seasonal cycles. The classification was based on approximate. At some stage in order to initiate algorithm to more accurately, the designed feature subset selection framework involves unrelated feature elimination and redundant feature elimination.

Attribute subset selection algorithm, unrelated features along with disused features strictly affect the accuracy of the learning machines. The attribute dividing up collection should be able to be capable to recognize and eliminate as much of the unrelated and disused information as potential. Moreover, "good attribute subsets contain features greatly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other", To develop a new algorithm ,which can resourcefully and successfully deal with both unrelated and out of work character, and gain a good quality attribute division to accomplish this during a novel attribute collection structure which collected of the two related workings of unrelated attribute elimination and redundant attribute removal ,It obtains character related towards the objective concept by eliminating unrelated ones, as well as the last one removes redundant character a and choosing representatives beginning from dissimilar attribute clusters as a result producing the ultimate division. An unrelated attribute deletion is simple, once the accurate bearing determine is definite or certain, at the same time the disused attribute removal is small complicated

In our proposed IFAST algorithm, which involves (a) determining the structure of the Minimum spanning tree (MST) starting from weighted whole graph, (b) the partition of the MST into a reforest with each hierarchy demonstrating a cluster; and (c) the collection of delegate quality as the cluster, In addition to additional specifically establish the algorithm, and since our planned attribute sub set selection structure involves unrelated attribute elimination and redundant feature elimination, we first present the traditional definition of significant and disused quality, then supplying our definition based on unpredictable correlation as follows .John et al., obtains an explanation of related character and understand to be the full set of features Si=F-{Fi }.Let Si ' be a value-assignment of all features in Fi 'T , a value-assignment of feature , and a value-assignment of the target concept .

The definition can be formalized as follows. For a set D with m features F= {F1, F2,…F_m} and class C and compute the T-Relevance SU (F_i, C) value for each feature F_i (1 < I < m) in the first step. The features Whose SU(F_i, C) values are greater than a predefined threshold Θ comprise the target-relevant feature subset ,F'={ F '$_1$,F'$_2$,…,F'$_K$} (K<M).In the second step is to calculate the F-Correlation and the complete graph G reflects the correlations among all the target-relevant features.

Definition: (Relevant feature) "is relevant to the target concept if and only if there exists some Otherwise, feature is an irrelevant feature", Definition -1 indicate so as to readily obtain are two kinds of related types outstanding to different from the definition we can know that is directly relevant to the target concept.

4. THE ANT COLONY OPTIMIZATION - (ACO)

The Ant Colony Optimization (ACO) [10] is a subdivision of recently developed swarm intelligence which has been used for classification". Swarm intellect is a ground which is studied for developing combined intellect of groups of straightforward agents, In collection of insect, which exist inside colony, such as ants and bees, a person can only do effortless odd jobs on its hold, although the colony's mutual employment is the major motive formative the intellectual performance .It shows that the majority real ants are monitored. Nevertheless, every ant although is on its foot, deposits a chemical material on the earth called pheromone. The goal of Ant-colony optimization is to extract rules from improved attribute subset collection algorithm.

5. RESULTS AND DISCUSSION

The UCI data repository has five data sets such as Anonymous Microsoft Web, Insurance Company Benchmark, Buzz in social media, Human Activity Recognition Using Smart phones, OPPORTUNITY Activity Recognition, which were used for this study (table.1).The attributes are reduced using FAST and Improved FAST is in table.2.

Table 1: Data set Description

Data Set	Total No. of. Attributes	Categorical Attributes	Continuous Attributes	Classes
Anonymous Microsoft Web	291	30	60	16
Insurance Company Benchmark	86	10	8	15
Buzz in social media	77	16	5	17
Human Activity Recognition Using	561	-	20	10

Smartphones				
Opportunity Activity Recognition	242	19	18	11

The attributes summarize using feature subset selection algorithm and improved feature subset selection algorithm is in table. 2.

Table 2: Reduced Data Sets

Data Sets	Instances	No. of. attributes	Fast	Improved Fast
Anonymous Microsoft Web	3771	291	14	7
Insurance Company Benchmark	9000	86	45	23
Buzz in social media	140000	77	4	2
Human Activity Recognition Using Smartphones	10299	561	7	4
Opportunity Activity Recognition	2551	242	30	15

We evaluated proportional presentation of the proposed technique along with Ant colony optimization -Miner using ten-fold cross-validation. We have evaluated comparative performance of the proposed method and Improved presentation of the proposed technique along with Ant colony optimization using ten-fold cross-validation. Each Database is divided into ten partitions, and each method is run ten times, using a different partition as test set each time, with the other nine as training set. The attribute subset selection algorithm and improved attribute subset selection algorithm have been implemented using ASP for database accessible in the information repository of UCI [2], and the Organization data set.

Table 3: Improved Test Set Accuracy Rate (%)

	Employee		Department	
Run Number	Existing - FAST	Proposed- Ant Miner with improved FAST	Existing -FAST	Ant Miner with improved FAST
1	91.05	94.32	68.42	72.63
2	92.15	93.15	75.79	80.00
3	91.67	91.67	74.74	81.05

4	94.59	97.06	65.26	74.74
5	89.41	92.75	73.68	75.79
6	91.20	95.65	68.42	67.37
7	89.77	93.84	71.28	82.97
8	95.55	96.55	73.40	72.34
9	90.04	92.54	67.37	78.94
10	91.86	95.71	71.58	80.00

This is because improper proportion of the selected features results in a large number of features is retained, and further affects the classification efficiency. Just like the default values used for IFAST in the experiments are often not the optimal in terms of classification accuracy, the default threshold values used for FAST and ACO might be so. In order to discover whether or not IFAST still outperforms while best entrance standards are used for the comparing algorithms, Ant colony optimization methods were firstly used to determine the optimal threshold values and then were employed to conduct classification for each of the two classification methods with the different feature subset selection algorithms upon the 10 data sets.

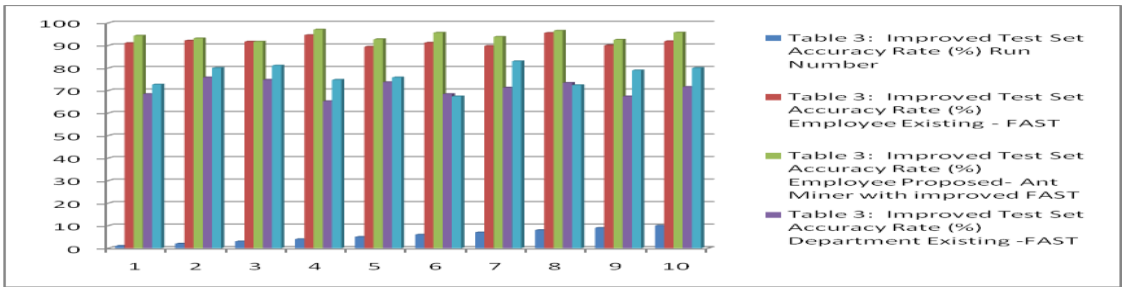

Fig. 2: Accuracy differences between FAST and the IFAST Comparing algorithms.

In order to discover whether or not IFAST still outperforms while best entrance standards are used for the comparing algorithms, Ant colony optimization methods were firstly used to determine the optimal threshold values and then were employed to conduct classification for each of the two classification methods with the different feature subset selection algorithms upon the 10 data sets.

The results reveal that IFAST still outperforms both FAST and ACO for all the two classification methods, Fig.2 shows the full details. The IFAST is significantly better than both FAST and (please refer to Table 3 for details).

6. CONCLUSION AND FUTURE WORK

It is demonstrated that Ant colony optimization algorithm with improved feature subset selection algorithm of clustering produces higher accuracy rate and fewer rules than the original Ant mining algorithm. In this paper, a new method called an improved feature subset selection algorithm, based on a variant of Random sort is proposed. We have

compared the results of Ant mining and the Ant Miner with Improved feature subset selection algorithm. The presentation of the ant mining algorithm is greater than before when it is used with better attribute subset selection algorithm.

The experimental results on text data sets demonstrate that the algorithm ACOIFAST can get hold of improved categorization accurateness other than it had a less important feature set than other similar methods. The plan was to discover changed types of association method, and learning some recognized Properties of attribute space and image for the future work.

References

[1] Pasi Franti.,Fast Agglomerative clustering using a K-Nearst Neighbor Graph,IEEE vol 28,November 2006.

[2] Qinbao song .,FAST-cluster based Feature subset selection Algorithm for High Dimensional Data.IEEE 2008.

[3] Grigorious F.Tzortzis, The Global Kernel K-Means Algorithm for clustering in feature space,IEEE Transactions on neural networks,vol 20,No.7,July 2009.

[4] Ujjwal maulik .,Integrating clustering and Supervised Learning for Categorical Data Analysis, IEEE Transactions onSystem man ,cybernetics ,vol,40,July 2010.

[5] Rahmat widia Sembiring., Clustering High Dimensional Data Using Subspace and Projected clustering Algorithms ,IJCSIT Vol.2,No.4,August 2010.

[6] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.

[7] Chanda P., Cho Y., Zhang A. and Ramanathan M.,Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.

[8] Cardie, C., Using decision trees to improve casebased learning, In Pro-ceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.

[9] Z. Pawlak, “Rough Sets”, International Journal of Computer and Information Sciences, Vol.11, No.5, pp. 341-356, 1982.

[10] Z.Pawlak, “Rough Sets: Theoritical Aspects and Reasoning about Data”, Kluwer Academic Publishers, Dordrecht, 1991.

[11] J. F. Peters and A. Skowron (eds.), “Transactions on Rough Sets 1”, Springer-Verlag,Berlin,2004.