International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 01,January-2015

A Fast Clustering Based FSS Algorithm for High Dimentional Data using SVM

Snehal Shinde
Department of Computer Engineering
JSPM NTC Pune, India

Sneha Malvadkar
Department of Computer Engineering
JSPM NTC Pune, India

Sonali Waskar
Department of Computer Engineering
JSPM NTC Pune, India

Nayantara Patil
Department of Computer Engineering
JSPM NTC Pune, India

***Abstract*— In today's life HIGH-DIMENSIONAL data is the major aspect in each and every field. Although it contains useful data but still it has many irrelevant and redundant features which do not contribute to any predictive accuracy and hence leads to wastage of storage space and memory management problems. Thus our topic aims at introducing FSS which is an effective way for reducing dimensionality, removal of irrelevant data, increasing learning accuracy and improving results comprehensibility and this process is improved by cluster based FAST Algorithm. FAST Algorithm can be used for recognizing and removing the irrelevant data. FSS algorithm works under two different steps that are graph theoretic clustering methods and representative feature cluster selection. To achieve desired efficiency and accuracy we are implementing the algorithm using Support Vector Machine (SVM). SVM's are supervised learning models with associated learning algorithms that analyze data, used for classification and analysis. Likewise the model speaks to the focuses in the space and mapped in such a way, to the point that the cases of the distinctive classes are separated by an acceptable gap that is as wide as would be prudent. Therefore the redundant feature number minimizes and effectiveness is picked up.**

Keywords—Feature subset selection, filter method, feature clustering, graph-based clustering.

I. INTRODUCTION

Feature selection includes recognizing a subset of the most valuable features that delivers perfect results as the first whole set of features. A feature selection algorithm may be assessed from both the proficiency and viability perspectives. While the proficiency concerns the time needed to discover a subset of features, the viability is identified with the nature of the subset of features. Based on these criteria, a clustering-based Feature Subset Selection (FSS) algorithm is proposed and tentatively assessed here. The FSS algorithm fulfills in two steps. In the first step, features are partitioned into clusters by utilizing graph-theoretic clustering methods. In the second step, the most illustrative feature that is firmly identified with target classes is chosen from each one group to structure a subset of features. Emphasizes in distinctive clusters are generally autonomous; the clustering-based strategy of FAST has a high probability of delivering a subset of helpful and free features. To guarantee the effectiveness of FAST, we receive the Minimum Spanning Tree (MST) clustering system. The proficiency and viability of the FAST calculation are assessed through an observational study. Thus on comparing FAST and a few delegate feature selection algorithms, specifically, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with four sorts of well-known classifiers, the Naive Bayes, the C4.5, the Ib1, and the RIPPER prior and then afterward the feature determination. As a result of survey on 35 freely accessible true high-dimensional picture, microarray, and content data, exhibit that the FAST creates littler subsets of peculiarities as well as enhances the exhibitions of the four sorts of classifiers. Capable algorithms for distinctive kernel learning and best feature determination calculation are introduced. Kernel function called Gaussian Radial basis Polynomial Function (GRPF) is familiar set up which improves the characterization accuracy of Support Vector Machines (SVM's) for both straight and non-control data sets. The fact is Support Vector Machines (SVM's) with differing bits differentiated and back-inciting learning algorithm in characterization undertaking. Finally the proposed algorithm is upgraded with respect to precision and time appeared differently in relation to the existing algorithm. Section II gives short thought regarding SVM's (Support Vector Machine). Section III gives thought of the writing study and the current techniques.

A. Support Vector Machine

SVM model is the presentation of illustration as focuses in the space mapped so that the case of distinctive classifications are partitioned by a reasonable crevice that is as wide as could reasonably be expected, Support Vector Machine order is picking a suitable kernel of SVMs for a specific application, i.e. different applications require diverse kernels to get trustworthy results. It is well realized that the two typical kernel functions frequently utilized as a part of SVMs are the Radial Basis Function Kernel and

IJERT

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

polynomial Kernel. Later kernels are introduced to handle high dimensional data sets and are computationally productive when taking care of non-detachable data with multi characteristics. In any case, it is not easy to identify kernels that have the capacity to accomplish high arrangement exactness for differences of data sets. Considering the end goal to make kernel functions from existing ones or by utilizing some other more straightforward kernel functions as building constructs, the closure properties of kernel functions are crucial.

B. Problem Statement

In the existing system the irrelevant features are removed successfully, but fails in removal of redundant features .The irrelevant features does not contribute to the predictive accuracy and the redundant features gives the same meaning which are already present in another features. So there is a need to remove the redundant features. The irrelevant and redundant features cause wastage of memory and poor performance. In the medical field and industry the high dimensional data in the form of text or images are present. For a better performance, there is a need of a method or algorithm which successfully and efficiently removes irrelevant and redundant data.

C. Objectives Of Work

Our point, as the name recommends, goes for picking a subset
of great features regarding the target ideas, with the help of
FSS Algorithm. The proposed calculation subsequently makes
a Fast method for clustering the high dimensional information.
The cluster is shaped on the premise of the features that are connected with the corresponding information thus each cluster is dealt with as a unique feature and in this manner dimensionality is definitely decreased.
Thus the objectives are:

- To reduce the dimensionality
- To remove irrelevant data
- To eliminate redundant data
- To increase learning accuracy
- To reduce time complexity
- To improve result comprehensibility

II. LITERATURE SURVEY

A. Filter Method

Filters appraise a relevance index for each one feature to quantify how relevant a feature is. There the filters rank features by their relevance indices and perform inquiry as per the ranks or focused around some statistical foundation e.g. significance level. The most recognizing normal for filters is that the relevance index is ascertained built singularly in light of a solitary feature without considering the estimations of different features. Such usage suggests that filters expect orthogonally between features which generally is not valid in practice. Along these lines, filters preclude any restrictive dependence (or autonomy) that may exist, which is known to be one of the demerits of filters, since they may omit ideal subset of features. Notwithstanding, filters are productive and turned out to be more powerful to over fitting hypothetically. [2]

B. Wrapper Method

In wrapper method, Instead of ranking each and every gimmick, wrappers rank peculiarity subsets by the Prediction execution of a classifier on the given subset. Unlike filters, wrappers can be utilized to search through all conceivable subsets of features and investigate the common data between features. In the wake of picking a classifier, wrappers assess the prediction execution either by cross-approval or hypothetical execution limits. Other than the decisions of classifiers, wrappers vary in the fundamental search techniques. Thoroughly searching combinatorial subsets is NP-hard and is inclined to over fitting. Greedy search methodologies are for the most part favored, for example, successive forward determination or retrogressive disposal. Since search method is a theme critical for both wrappers and embedded methods. [3]

C. Embedded Method

Embedded methods select features focused around rules that are created amid the learning methodology of a particular classifier. As opposed to wrapper method, they don't separate the gaining from the feature determination part, i.e. the selected features are touchy to the structures of the fundamental classifiers. Hence, as a rule, the feature selected by one embedded methods may not be suitable for others. Formally, embedded methods are composed unequivocally or certainly too rough arrangements of the minimization issue concerning weights for features and the parameterization of a classifier. [4]

D. Hybrid Method

Another feature choice method called Hybrid method proposed for joining focal points of filter and wrapper methods. It is a two-phase calculation where the filter method is utilized as a part of first phase and the pursuit calculation is utilized as a part of the second. In this method, the first features are assessed by the filter method. Here, the consequence of filter gives heuristic data to the inquiry of wrapper system. Mixture method has no restriction to information sort and if the classifiers are utilized that has no uncommon confinement, and the probes different sort and estimated information sets accept that half and half method out-performs independently of vast scale information sets. The Hybrid feature determination methods can manage substantial scale information sets. [5]

III. PROPOSED WORK

Of the many algorithms, some can effectively eliminate irrelevant features but fail to handle redundant features yet some of others can eliminate the irrelevant while taking care of the redundant features. Our proposed FSS algorithm comes under second group. Traditionally, FSS research has focused on searching for relevant features.

IJERT

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

- Feature Selection Algorithm and SVM based classification algorithm. The proposed methodology discusses about the feature selection using FSS with SVM.
- The aim of Support Vector Machines (SVMs) with different kernels compared with back-propagation learning algorithm in classification task.

A. System Architecture

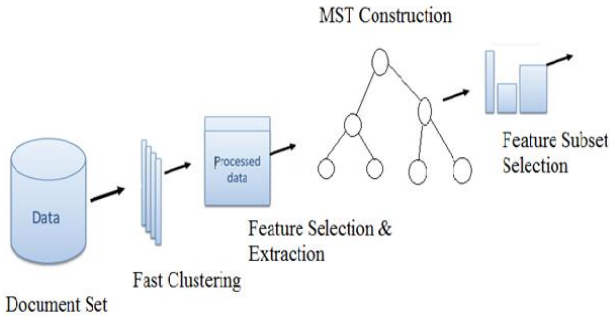

Fig. 1. Architecture

The figure demonstrates the stream of steps followed in the framework. It begins with authentication of client. Every client is furnished with username and password enlisted in framework as of now. There is a salt quality authentication alongside password. The authentication client has admittance to the database and framework has specific access rights for every client. The Anonymous database stifles and sums up the information as per information esteem. The database can be gotten to via research habitats for social affair statistical information viewing specific question, for example, prescriptions, the rate of treatable medicines. The private data of the patients are not uncovered to the research focus processing. The research persons can see the information's send by the database as per its get to right. What's more dispense research people groups to each one research information. Also forward the information to research people. Here individuals can't do any progressions or changes in database they just can utilize the database for reference reason.

The authorized database updaters can login into the system. Here likewise all the insights about the database updater are enrolled by the administrator and he gives the authentication points of interest to the specific updater in the wake of getting the authentication subtle elements, can login to the database and can begin the courses of action.

Clustering Based FSS with SVM brings about the above calculation then order the peculiarity in the information .Support vector machine based peculiarity subset choice calculation is performed to group the information in the gimmick subset. The fundamental SVM takes a set of data information as peculiarity subset result from the grouping based gimmick choice and predicts, for every given info, which of two achievable classes structures the yield, making it a non-probabilistic binary linear classifier

Support Vector Machine classification is picking a suitable kernel of SVM's for a specific application, i.e. different applications require diverse kernels to get reliable characterization results. It is well realized that the two typical kernel functions regularly utilized as a part of SVM's are the outspread premise function kernel and polynomial kernel. Later kernels are exhibited to handle high dimensional data sets and are computationally proficient when taking care of non-divisible data with multi characteristics. In any case, it is not simple to discover kernels that have the capacity accomplish high grouping precision for differing qualities of data sets. Keeping in mind the end goal to make kernel functions from existing ones or by utilizing some other more straightforward kernel functions as building obstructs, the closure properties of kernel functions are vital.[6] the irrelevant feature evacuation is direct once the right significance measure is characterized or chose, while the repetitive feature evacuation is a bit of refined [1].

Our proposed algorithm, Clustering based FSS, includes;

- The design of the Minimum Spanning Tree(MST) from a weighted complete diagram;
- The partitioning of the Minimum Spanning Tree(MST) into a forest with each one tree speaking to a cluster;
- The gathering of agent features from the clusters.

FSS calculations coherently comprises of three steps:

1) Removing irrelevant features,
2) Constructing a MST , and
3) Partitioning the MST and selecting agent characteristics.

B. Algorithm

Multi-Class SVM:

Input: Training set (v1; l1) ;:::;(vN; lN)

Output: Classifier Training:

Binary SVM's and given relevance scores
For j = 1 to (k - 1) do

- For all samples from C1 to Cj classes, set labels to (+1) and all samples from Cj+1 to Ck, set labels to (-1)
- Train jth binary SVM
- Classify the training samples
- if (j > 1), compute fuzzy scores p for all training samples v p classified as (+1) and define (j1) thresholds by split- ting the curve of sorted relevance scores into equally spaced intervals.
- if (j < k), compute fuzzy scores for all training samples vn classified as (-1) and define (kj1)thresholds by split-ting the curve of sorted relevance scores into equally intervals end for.

IV. MODULES

A. User Module

In this module, users are having authentication to access the data set which is present in the system. For accessing or searching the data set, it is necessary that the users should register themselves and create an account so that they will be authenticated to access the data set or any related details.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

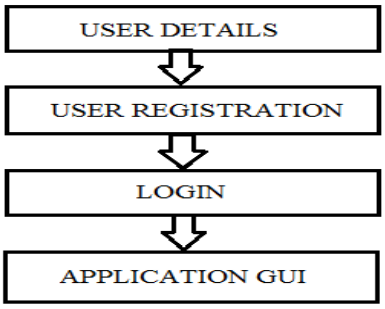

Fig. 1. User module

B. *Experimental Setup*

First, the authenticated user will login to the page or else the new user has to first register and the can login and access the data. On successful login, a message box saying "Welcome To Homepage" is popped up.

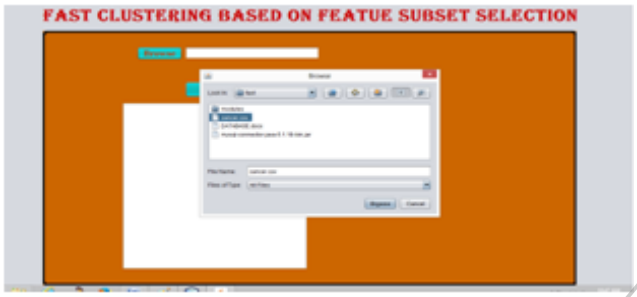

Fig. 3. Load Dataset

Fig. 4. Convert Dataset

Fig. 5. Extract Dataset

C. *Distributed Clustering*

The Distributional clustering has been used to cluster the words into groups based on their participation in particular grammatical relations with other words. It is proposed to cluster the features utilizing a special metric of separation, and afterward makes utilization of the of the resulting cluster order to pick the most relevant attributes.

D. *Subset Selection*

Accuracy of the learning machines is badly affected due to the presence of irrelevant and redundant features. Accordingly, subset choice of features is carried out in such a way, to the point that it can undoubtedly recognize and uproot however much of the irrelevant and redundant features as could reasonably be expected. In this module we design a novel algorithm which will efficiently and effectively deal with both irrelevant and redundant features, and provide us with a good subset of features

V. CONCLUSION

We have presented a novel clustering-based FSS algorithm for high dimensional data. The algorithm involves (i) removal of irrelevant features, (ii) constructing a minimum spanning tree (MST), and (iii) partitioning the MST and selecting required features. In the FSS algorithm, a cluster consists of features. The dimensionality is drastically reduced as each cluster is considered as a unique feature. Thus, the proposed algorithm obtained the best proportion of selected features. The classification algorithm based on SVM's is applied to high dimensional data for achieving accuracy. It shows that proposed clustering based feature subset selection with SVM best classification accuracy then the previous work..

IJERT

ACKNOWLEDGMENT

We might want to express our appreciation and gratefulness to every one of the individuals who guided us in our task. An extraordinary on account of Prof. Snehal Shinde our undertaking facilitator, whose help, animating recommendations and consolation, guided us for coordination in our venture. We might likewise want to recognize HOD of Computer Department, Dr. Sulochana Sonkamble,for her patient backing. To wrap things up, numerous on account of our venture guide Prof. S. S. Shinde who has provided for her full exertion in controlling us in accomplishing the objective and additionally her consolation to keep up our advancement in track.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

REFERENCES

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992.
[2] Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992. Kira K. and Rendell L.A., The feature selection problem: Traditional methods and a new algorithm, In Proceedings of Nineth National Conference on Artificial Intelligence, pp 129-134, 1992.
[3] Kohavi R. and John G.H., Wrappers for feature subset selection, Artif. Intell., 97(1-2), pp 273-324, 1997.
[4] Yu L. and Liu H., Feature selection for high-dimensional data: a fast correlation-based filter solution, in Proceedings of 20th International Conference on Machine Leaning, 20(2), pp 856-863, 2003.
[5] Yu J., Abidi S.S.R. and Artes P.H., A hybrid feature selection strategy for image defining features: towards interpretation of optic nerve images, In Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 8, pp 5127-5132, 2005.
[6] K.Saranya1 and T. Deepa Discriminative Clustering Based Feature Selection and Nonparametric Bayes Error Minimization and Support Vector Machines (SVMs) International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.
[7] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.
[8] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pear- son redundancy based filter, AdvancesinSoftComputing, 45, pp 242C249, 2008.
[9] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Se- lection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.
[10] 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.
[11] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.
[12] Demsar J., Statistical comparison of classifiers over multiple data sets, J. Mach. Learn. Res., 7, pp 1-30, 2006.
[13] Zhao Z. and Liu H., Searching for interacting features, In Proceedings of the 20th International Joint Conference on AI, 2007.
[14] Zhao Z. and Liu H., Searching for Interacting Features in Subset Selection, Journal Intelligent Data Analysis, 13(2), pp 207-228, 2009.
[15] Park H. and Kwon H., Extended Relief Algorithms in Instance-Based Feature Filtering, In Proceedings of the Sixth International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007), pp 123-128, 2007.
[16] Sha C., Qiu X. and Zhou A., Feature Selection Based on a New Depen- dency Measure, 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 1, pp 266-270, 2008.
[17] Souza J., Feature selection with a general hybrid algorithm, Ph.D, Univer- sity of Ottawa, Ottawa, Ontario, Canada, 2004.
[18] Van Dijk G. and Van Hulle M.M., Speeding Up the Wrapper Feature Subset Selection in Regression by Mutual Information Relevance and Redundancy Analysis, International Conference on Artificial Neural Networks, 2006.
[19] Demsar J., Statistical comparison of classifiers over multiple data sets. J. Mach. Learn. Res., 7, pp 1-30, 2006.
[20] Liu H., Motoda H. and Yu L., Selective sampling approach to active feature selection, Artif. Intell., 159(1-2), pp 49-74 (2004)

IJERT

(This work is licensed under a Creative Commons Attribution 4.0 International License.)