[Agrawal, 5(6): June, 2016]
IC™ Value: 3.00

ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

# AN APPROACH FOR UNSUPERVISED FEATURE SELECTION USING GENETIC ALGORITHM

**Nootan Agrawal*, Mr. Sandeep Gonnade**
* Student, Mtech(CSE), Mats University, Gullu, Aarang, Raipur, (C.G)
Assistant Professor, Department of CSE, Mats University, Gullu, Aarang, Raipur, (C.G)

**DOI**: 10.5281/zenodo.55985

## ABSTRACT

Each data mining application has widespread issue; dataset has gigantic number of features which are irrelevant or redundant to the data mining task in hand which negatively affects the performance of the elementary learning algorithms, and makes them lesser capable. There is difficulty of inadequate increase in dimension is strappingly related to fascination of cassette or measuring data at a far granular level then it was done previously. There is no doubt that this is a blistering problem. It has started gaining more magnitude recently due to surge in data. Hereafter plummeting the dimensionality of dataset is principal and imperative job for data mining applications and machine learning algorithms in order that computational burden of the learning algorithms can be minimized. In this paper we will measure up to the GFS (Greedy Feature Selection) and our proposed method and diverse unsupervised feature selection algorithms discussed in order to find out factors which influence the performance of existing algorithm. In our proposed method we have incorporated the Genetic feature selection method and GFS and TPR (True Positive Rate), FNR (False Negative Rate) estimated using KNN Classifier.

**KEYWORDS:** k-means, Muti-core, Opencl

## INTRODUCTION

Data mining is the practice of sighting and estimation of large databases of data with the aim of exploring data and rules. It can be defined as the process that starting from apparently unstructured data tries to extract knowledge and/or unknown fascinating patterns. In data mining unstructured data are evaluated using two learning methods, supervised or unsupervised learning.

- **Supervised learning:** Supervised learning is the practice of machine learning for dying a function from labeled training data. It is also called as directed data mining practice. In this technique data set values are distinguished as dependent and independent variable and the values of the dependent variable should be known for passably huge part of dataset.

**Unsupervised learning:** Unsupervised learning is the practice of discovery the hidden structure in the data which is not labeled. This practice is also called undirected data mining. In this practice the target is achieved typically by clustering technique.

By unsupervised learning we stand for unsupervised clustering. Clustering is the procedure of finding groupings by combining "similar" founded on some similarity measure objects collectively. For numerous learning domains, human being defines the features that are potentially functional. However, not all of these features may be relevant. In such a case, choosing a subset of the original features will often lead to better performance.

Feature selection is popular in supervised learning (Fukunaga, 1990; Almuallim & Dietterich, 1991; Cardie, 1993; Kohavi & John, 1997). For supervised learning, feature selection algorithms maximize some function of predictive accuracy. Because we are given class labels, it is natural that we want to maintain only the features that are interrelated to or lead to these classes. But in case of unsupervised learning, class labels not given. Which features should we keep? Why not use all the information we have? The problem is that not all features are important. Some

of the features may be redundant, some may be irrelevant, and some can even misguide clustering results. In addition, reducing the number of features increases comprehensibility and ameliorates the problem that some unsupervised learning algorithms break down with high dimensional data. Concluding two approaches have been proposed for dimension reduction feature selection, and feature extraction.

Concluding that reducing the dimension of data set has following advantages:

- It reduces the storage, time and space required.
- Elimination of multi-co linearity improves the feat of the machine learning algorithm.
- It becomes easier to think about the data when reduced to low dimensions such as 2D or 3D.

The remaining sections of the paper are organized as section II we will converse about previous work has been carried out in this field. Further in section III we will discuss about how we motivated for research work. In section IV we will identify problems in dimension reduction approach. Section V converse the comparison of some existing algorithms. Section VI concludes our survey.

## LITERATURE SURVEY

Numerous research works has been carried for dimension reduction of data instance to exploit feature.

Ahmed Elgohary, Ali Ghodsi & Ahmed K. Farahat (2013) proposes algorithm that depends on a novel recursive formula for the reconstruction error of the data matrix, which allows a greedy selection criterion to be calculated efficiently at each iteration. They have also presents an accurate and efficient MapReduce algorithm for selecting a subset of columns from a massively distributed matrix. This work enables data analysts to comprehend the insights of the data instance and explore its secreted structure. The preferred data instances can also be used for data preprocessing tasks such as learning a low-dimensional embedding of the data points.

Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel (2013) defines a generalized column subset selection problem which is concerned with the selection of a few columns from a source matrix A that best approximate the span of a target matrix B. They proposes a fast greedy algorithm for solving this problem and draws connections to different problems that can be efficiently solved using the proposed algorithm.

Carlos Vicient (2012) discussed about log jam introduced by the manual semantic mapping process. To deal with this problem, presents a domain-independent, automatic and unsupervised method to detect relevant features from heterogeneous textual resources, associating them to concepts modeled in background ontology. The method has been applied to raw text resources and also to semi structured ones (Wikipedia articles). The work has been weathered in the Tourism domain, showing promising results.

Ahmed K. Farahat (2011) presents a novel greedy algorithm for unsupervised feature selection. The algorithm optimizes a feature selection standard which measures the reconstruction error of the data matrix based on the subset of selected features. Ahmed K. Farahat proposes a novel recursive formula for calculating the feature selection criterion, which is then employed to develop an efficient greedy algorithm for feature selection. Additionally two memory and time efficient variants of the feature selection algorithm are proposed.

Yi Yang & Heng Tao Shen (2011) discussed that it is much more complicated to select the discriminative features in unsupervised learning due to be deficient in of label information. They have proposed a new unsupervised feature selection algorithm which is able to select discriminative features in batch mode. An efficient algorithm is proposed to optimize the l2,1-norm regularized minimization problem with orthogonal constraint. Different from existing

Jennifer G. Dy et. al. In this paper, author identified two issues involved in developing an automated feature subset selection algorithm for unlabeled data: the need for finding the number of clusters in conjunction with feature selection, and the need for normalizing the bias of feature selection criteria with respect to dimension. We explore the feature selection problem and these issues through FSSEM (Feature Subset Selection using Expectation-Maximization (EM) clustering) and through two different performance criteria for evaluating candidate feature subsets: scatter separability and maximum likelihood.

| S.No. | Author/Year | Name of Algorithm | Advantage | Disadvantage |
|---|---|---|---|---|
| 1. | Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE TKDE, 26(9):2138–2150, | CGSSL | Provides label information for the structured learning in optimized form | Feature correlations are not investigated explicitly |

ISSN: 2277-9655
Impact Factor: 4.116

| | | | | |
|---|---|---|---|---|
| | Sept 2014 | | | |
| 2. | Haichang Li ; Inst. of Autom., Beijing, China ; Shiming Xiang ; Zisha Zhong ; Kun Ding Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification IEEE 2015 | Unsupervised Spatial-Spectral Feature Selection Method | Best relevant features from hyper spectral image dataset are obtained with approximation | Not applicable for large datasets |
| 3. | Padungweang, P. Padungweang, P. A Discrimination Analysis for Unsupervised Feature Selection via Optic Diffraction Principle IEEE 2012 | Unsupervised Feature Selection Via Optic Diffraction Principle | The notion of physical optics is used effectively for discrimination calculation of distribution | Sometimes depends on probability density estimation which requires future search for finding optimal solution |
| 4. | Ahmed K. Farahat Ali Ghodsi Mohamed S. Kamel An Efficient Greedy Method for Unsupervised Feature Selection IEEE 2011 | Greedy Method for Unsupervised Feature Selection | Algorithm optimizes a feature selection criterion which measures the reconstruction error of the data matrix based on the subset of selected features | Less efficient for very large data instance. |

## PROBLEM IDENTIFICATION

Each data mining application has widespread issue; dataset has gigantic number of features which are irrelevant or redundant to the data mining task in hand which negatively affects the performance of the elementary learning algorithms, and makes them lesser capable. There is difficulty of inadequate increase in dimension is strappingly related to fascination of cassette or measuring data at a far granular level then it was done previously. There is no doubt that this is a blistering problem. There are some bottlenecks in dimension lessening approach.

- Physically classification of enormous amounts of training data is very prolonged; furthermore, it is hard for one data mining system to be ported across different domains. Caused by the limitation of supervised methods, some semi-supervised approaches have been recommended.
- Order selection and discriminative label detection.
- The inherent dimension.
- Data compression for data instance storage.
- Speed of learning.
- Predictive accuracy.
- Minimalism and unambiguousness of mined result.

## PROPOSED METHODLOGY

Plummeting the dimensionality of dataset is most important and significant task for data mining applications and machine learning algorithms so that computational burden of the learning algorithms can be minimized. In this paper we have compared the GFS (Greedy Feature Selection) and our proposed method and different feature selection algorithms discussed so as to find out factors which affect the performance of existing algorithm. In our proposed method we have incorporated the Genetic feature selection method and GFS and TPR (True Positive Rate), FNR (False Negative Rate) estimated using KNN Classifier.

ISSN: 2277-9655
Impact Factor: 4.116

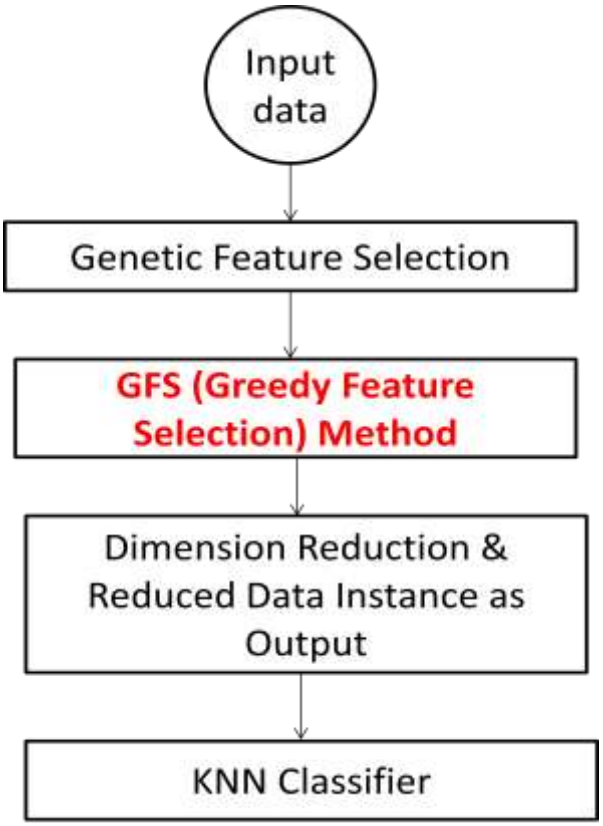


***Fig.- Proposed Layout***

Algorithm used for Greedy feature selection is same as given by Ahmed K. Farahat el al. 2011.

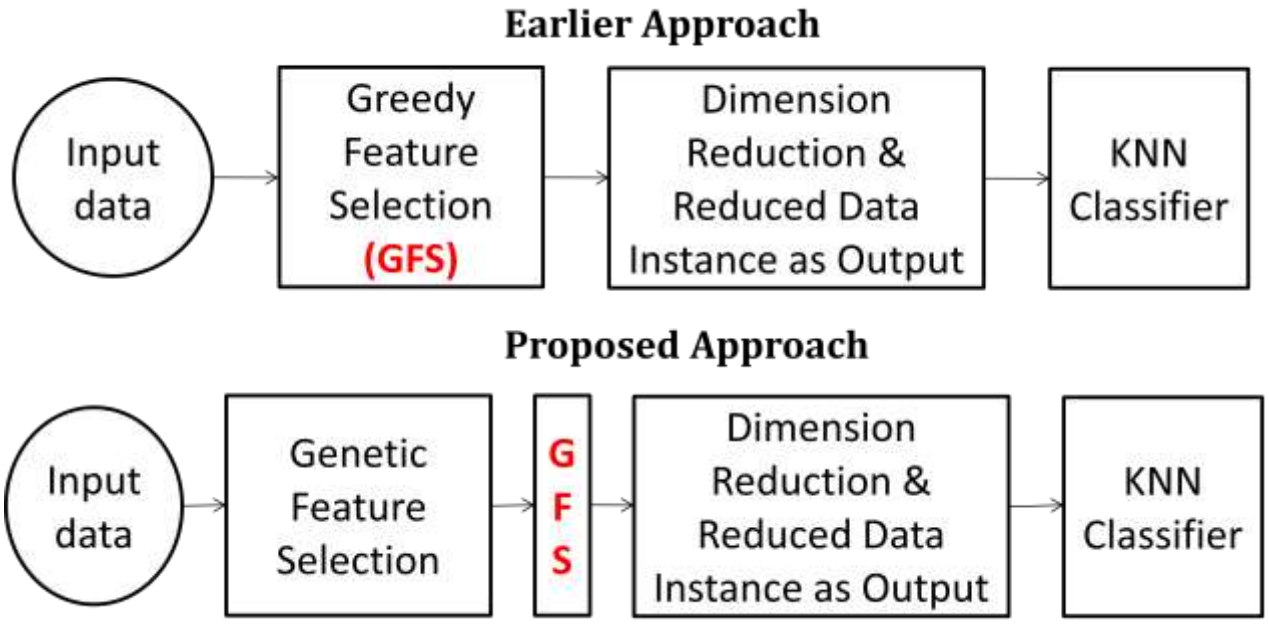


***Fig. Comparison of flow of Earlier approach vs Proposed approach***

ISSN: 2277-9655
Impact Factor: 4.116

**Greedy Feature Selection (GFS)**

In our proposed approach we have used Greedy feature selection algorithm as discussed by Ahmed K. Farahat 2011, in which PCA-like standard used which minimizes the rebuilding error of the data matrix depend on the elected subset of features.

GFS selects one feature at each iteration such that the reconstruction error for the novel set of features is in least amount.

An adolescent implementation of the greedy feature selection algorithm is to compute the reconstruction error for every candidate feature, and then select the feature with the minimum error.

Suppose we are having N data points (xi,yi) where xi is a m-dimensional feature descriptor and yi is label, and we wish for to find out which of the m dimensions are functional, or possibly rank the dimensions in order of worth. GFS (Greedy feature selection) adds one feature dimension at a time to a set of previously elected features, and checks how fine that feature is by testing and training classifiers on k cross-validation splits. The best feature (in terms of correctness) is then appended to the set of elected features, and the next iteration started.

## RESULT AND DISCUSSION

Experiments have been performed on Matlab 2015 platform. For experiment we have taken dermatologist data and dimension reduction performed then over reduced dataset classification has been done using Fine KNN, Medium KNN & Coarse KNN. Comparison table as follows.

| Sr. No. | Method | Classification | | |
|---|---|---|---|---|
| | | Fine KNN | Medium KNN | Coarse KNN |
| 1 | GFS | 94% | 94.50% | 94.10% |
| 2 | Genetic Feature Selection | 95.10% | 95.40% | 95.10% |
| 3 | Genetic Feature Selection->GFS | 95.60% | 96.40% | 96.70% |

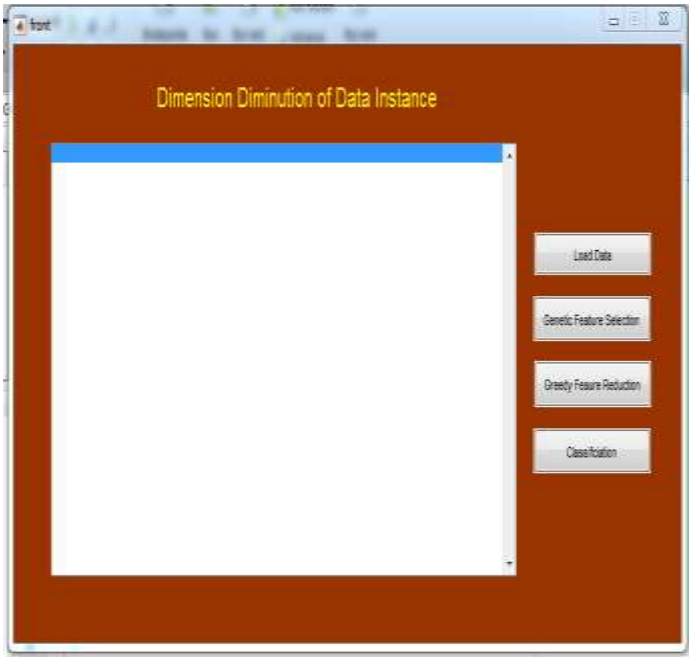


***Fig. UI of Project***

ISSN: 2277-9655
Impact Factor: 4.116

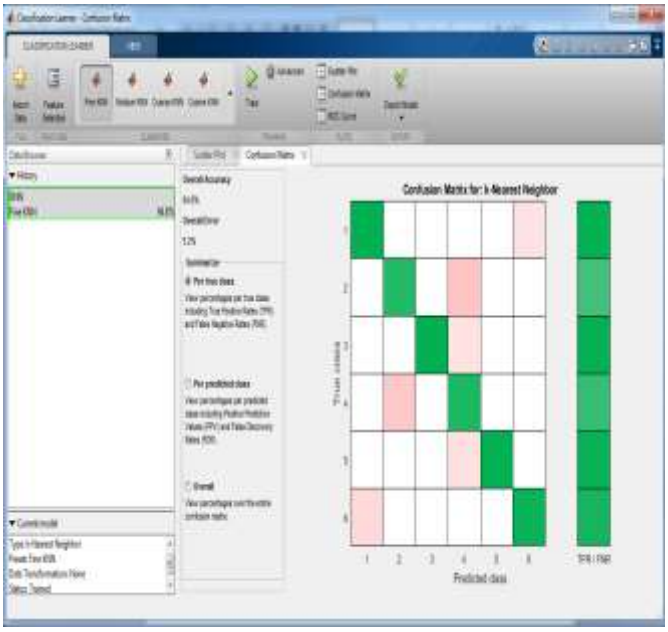


*Fig.- Classification using KNN Classifier (Confusion Matrix)*

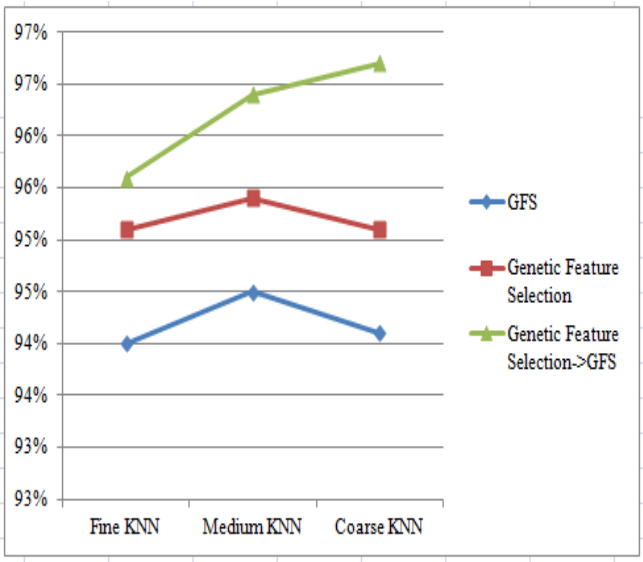


*Fig. Comparison of Algorithms over KNN Classifier*

## CONCLUSION

Plummeting the dimensionality of dataset is most important and significant task for data mining applications and machine learning algorithms so that computational burden of the learning algorithms can be minimized. In this paper we have compared the GFS (Greedy Feature Selection) and our proposed method and different feature selection algorithms discussed so as to find out factors which affect the performance of existing algorithm. In our proposed method we have incorporated the Genetic feature selection method and GFS and TPR (True Positive Rate), FNR (False Negative Rate) estimated using KNN Classifier. In future we can apply this approach to increase data prediction accuracy.

ISSN: 2277-9655
Impact Factor: 4.116

## REFERENCES

1. Shiming Xiang ; Zisha Zhong ; Kun Ding Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification IEEE 2015.
2. P.Miruthula1, S.Nithya Roopa Unsupervised Feature Selection Algorithms: A Survey IJSR 2015.
3. Wee-Hong Ong, Leon Palafox, Takafumi Koseki nvestigation of Feature Extraction for Unsupervised Learning in Human Activity Detection Volume 2, Number 1, pages 30–35, January 2013.
4. Liang Du, Yi-Dong Shen Unsupervised Feature Selection with Adaptive Structure Learning 2015.
5. Ahmed K. Farahat Ali Ghodsi Mohamed S. Kamel An Efficient Greedy Method for Unsupervised Feature Selection 2011 11th IEEE International Conference on Data Mining
6. Yi Yang1, Heng Tao Shen1, Zhigang Ma2, Zi Huang1, Xiaofang Zhou l2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence 2015.
7. L.J.P. van der Maaten * , E.O. Postma, H.J. van den Herik Dimensionality Reduction: A Comparative Review MICC, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands. 2015.
8. Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu. Clustering-guided sparse structural learning for unsupervised feature selection. IEEE TKDE, 26(9):2138–2150, Sept 2014.
9. Jennifer G. Dy Feature Selection for Unsupervised Learning School of Electrical and Computer Engineering US 2003.
10. Padungweang, P. Padungweang, P. A Discrimination Analysis for Unsupervised Feature Selection via Optic Diffraction Principle IEEE 2012.
11. Ahmed K. Farahat, Ali Ghodsi, and Mohamed S. Kamel A Fast Greedy Algorithm for Generalized Column Subset Selection 2013.
12. Jiliang Tang and Huan Liu"Unsupervised feature selection framework for social media data"IEEE trans on knowledge engg and datamining., vol 26, no.12,Dec 2014.
13. Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou, Senior Member, IEEE, and Hanqing Lu, Senior Member,IEEE" Clustering-Guided Sparse Structural Learning For Unsupervised Feature Selection"IEEE trans on knowledge engg and data mining., vol 26,sept 2014