A Novel Approach for Dimension Reduction of Data Instance Using Genetic & Greedy Feature Selection Method

Ujjawal Singh
Department of CSE, FET SSTC-Shri Shankaracharya Group of Institutions, Junwani(Bhilai), India.

Prof. Megha Mishra
Department of CSE, FET SSTC-Shri Shankaracharya Group of Institutions, Junwani(Bhilai), India.

Dr. V.K.Mishra
Department of CSE, BCET ,Durg, India.

Abstract – Data mining application has massive advantages, as historical data have huge number of features. Feature selection has been broadly deliberated in supervised learning, whereas it is still comparatively infrequent researched in case of unsupervised learning. Every data mining application has common issue; dataset has huge number of features which is immaterial or redundant to the data mining job in hand which pessimistically affects the performance of the fundamental learning algorithms, and makes them less efficient. There is trouble of useless increase in dimension is strongly related to obsession of recording or measuring data at a far granular level then it was done earlier. There is no doubt that this is a hot problem. It has started gaining more importance lately due to surge in data. Henceforth reducing the dimensionality of dataset is primary and important job for data mining applications and machine learning algorithms so that computational burden of the learning algorithms can be minimized. In this paper we will compare the GFS (Greedy Feature Selection) , 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 integrated the Genetic feature selection method and GFS.

Index Terms – Supervised learning, unsupervised learning, Feature selection, GFS.

1. INTRODUCTION

Feature selection is an essential role in improving the eminence of learning algorithms in data mining and machine. This has been broadly deliberated in supervised learning, whereas it is still comparatively infrequent researched in case of unsupervised learning. As per Dunham (2002), machine learning tasks can be seen as predictive or descriptive ones. Classification is an example of predictive models. Friedman (1997) described it as a model where discrete output values (class labels) are learnt from the different variables (features) of the input data. Clustering, on the other hand, is categorized by Dunham (2002) as a descriptive task. The features of the input data are used to categorize it without supervised training. In both cases, the choice of the feature-set plays an important role in the performance of the data mining problem. Liu et al. (2010) listed three advantages for removing irrelevant and redundant features: it makes the data mining task more efficient, improves its accuracy and simplifies the inferred model, making it more comprehensible.

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.

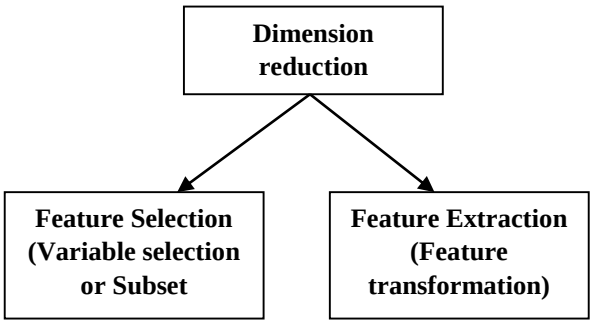

Fig. – 1 Type of feature reduction

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.

2. 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

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, Sept 2014	CGSSL	Provides label information for the structured learning in optimized form	Feature correlations are not investigated explicitly

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.

3. PROBLEM IDENTIFICATION

There are some bottlenecks in dimension reduction approach.

- Physically tagging of huge amounts of training data is very prolonged; furthermore, it is hard for one data mining system to be ported across different domains. Due to the limitation of supervised methods, some semi-supervised approaches have been recommended.
- Order selection and discriminative label identification.
- The intrinsic dimension.
- Data compression for data instance storage.
- Speed of learning
- Predictive accuracy
- Simplicity and comprehensibility of mined result

The difficulty of (nonlinear) dimensionality reduction can be explained as follows. Suppose we have dataset represented in a n×D matrix X consisting of n data vectors x_i ($i \in \{1, 2, \ldots, n\}$) with dimensionality D. Suppose auxiliary that this dataset has intrinsic dimensionality d (where d < D, and often d<<D). at this point, in mathematical stipulations, intrinsic dimensionality means that the points in dataset X are lying on or near a manifold with dimensionality d that is embedded in the D-dimensional space. Dimensionality reduction techniques transform dataset X with dimensionality D into a new dataset Y with dimensionality d, while retaining the geometry of the data as much as possible.

In general, neither the geometry of the data manifold, nor the intrinsic dimensionality d of the dataset X are known. Therefore, dimensionality reduction is an ill-posed problem that can only be solved by assuming certain properties of the data (such as its intrinsic dimensionality).

4. PROPOSED METHODOLOGY

Reducing the dimensionality of dataset is primary and important job 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 integrated the Genetic feature selection method and GFS.

Algorithm for Genetic feature selection

Begin

1. t = 0

2. initialize population P(t) / P opsize = |P| */*

3. for i = 1 to P opsize

compute fitness P(t)

4. t = t + 1

5. if termination criterion achieved go to step 10

6. select (P)

7. crossover (P)

8. mutate (P)

9. go to step 3

10. output best chromosome and stop

End

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

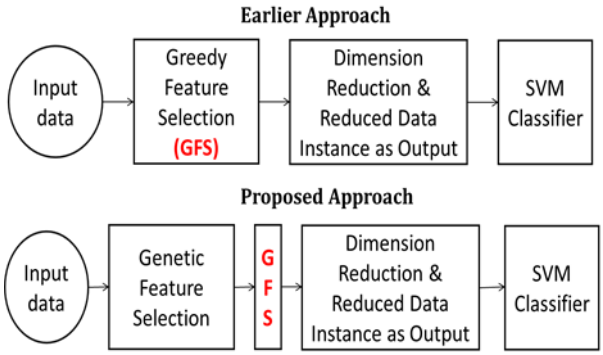

Fig. Comparison of flow of earlier approach vs proposed approach

5. 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 Linear SVM, Quadratic SVM & Cubic SVM. Comparison table as follows.

Sr. No.	Method	Classification		
		Liner SVM	Quadratic SVM	Cubic SVM
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%

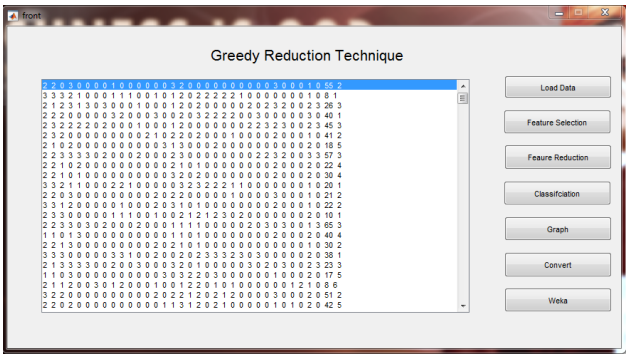

Fig. UI of Project

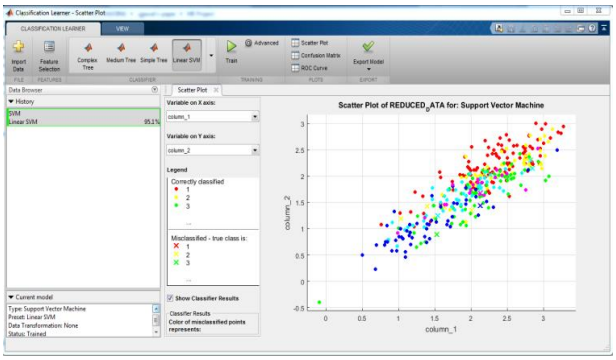

Fig Classification using Liner SVM

6. CONCLUSION

Reducing the dimensionality of dataset is primary and important job for data mining applications and machine learning algorithms so that computational burden of the learning algorithms can be minimized. Dimension reduction is based on two approach feature reduction and feature selection. In feature reduction all original features are used and the transformed features are linear combinations of the original features. In feature selection only a subset of the original features are selected. In our proposed method we have integrated the Genetic feature selection method and GFS and trained the outcome of each method with Linear SVM, Quadratic SVM and Cubic SVM, we found our proposed algorithm classification percentage is better than earlier method.

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.