IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Special Issue || Dec. 2019 || PP 39-43
International Conference on Management Practices, Innovations & Research 2019

A Survey on Data Redundancy Minimization Based on Feature Extraction

[1]Ashwini Hanwate, [2]Jayant Adhikari
[1]*M.Tech Student, Department of Computer Science & Engineering, Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, Maharashtra, India.*
[2]*Assistant Professor, Department of Computer Science & Engineering, Tulsiramji Gaikwad Patil College of Engineering and Technology, Nagpur, Maharashtra, India.*

Abstract—*In high dimensional data mining feature selection is most important step. Feature selection is use to select most relevant,important and informative features from the high-dimensional dataset. It plays an important role in many scientific and practical applications, because itincreases the speed of learning process. So, it is very important to develop an efficient framework, which can improve the feature selection process. In literature, various supervised and unsupervised feature selection methods are developed. In order to utilize both local and global structures, existing system propose unsupervised local and global discriminative (LGD) feature selection criterion. Generally, supervised feature selection methods with supervision information are better than unsupervised ones without supervision information. There is another challenging task in feature selection that is, to decrease the redundancy in selected features. In this paper, we use another form of supervised feature selection method that is constrained score for feature selection. This constraintscore algorithm is used for feature selection in high dimensional dataset and also used to decrease the redundancy. To evaluate the performance of proposed system, experiments are carried out on LUNG dataset. Experimental results show that, the constrained score is better than LGD feature selection criteria, in terms of reduced redundancy and accuracy in number of feature selected.*
Keywords— *Data mining, feature selection, redundancy minimization, supervised and unsupervised feature learning.*

I. Introduction

Information technologyis the application ofcomputersandtelecommunications, used to store, retrieve, transmit and manipulate data. Application of this system leads to huge amount of data, also known as big data. This big data includes various challenges such as, analysis, capture, data restoring, searching, sharing, storage, transfer, visualization, querying and owners data privacy. Accuracy in big data require more confident decision making system.We should take better decision for analysis of such big data, which results inincreased operational efficiency, cost reduction and reduced risk. Analysis of these big data is become challenging task.

Many data mining and machine learning techniques are used to analyze the data for various applications. Among this technique, feature selection is the most important method for classification technique. Inmachine learningandstatistics,feature selection is also known as variable selection, attribute selection or variable subset selection. It is the process of selecting a relevant features used in model construction. The advantage of this technique is that, it increases the speed of learning process and decrease the running time of algorithm. The advantages of feature selection are listed below:

- Facilitating data visualization
- Facilitating data understanding
- Reducing the measurement and storage requirements
- Reducing training and utilization times
- Improve prediction performance

But there is a problem of data redundancy in selected features.There is need to minimize the redundancy between sequentially selected features.

There are different types of feature selection approaches implemented. All these mechanisms are uses different ways for feature selection, such as Variance, Laplacian Score and Fisher Score. Among them, Variance and Laplacian Score are unsupervised, while Fisher Score is supervised.

According to class labels are used or not, feature selection methods are divided into supervised feature selection and unsupervised feature selection. When labeled data is sufficient, supervised feature selection methods perform better than unsupervised feature selection methods. But in many cases class labels are expensive to obtain. Usually the amount of labeled training data is very limited. Most of traditional supervised

feature selection methods may fail on such limited labeled training data. A recent research on this problem is to use both labeled and unlabeled data for feature selection, i.e. semi-supervised feature selection.

In Existing system, an unsupervised local and global discriminative feature selection criterion is proposed. These criteriamake use of both local and global structures. But still there are some limitations. This system uses the parameter that is the size of neighborhood set. If the parameter is too small, local variance is not captured well enough. If the parameter is too large, the boundary between local and global blurs. Therefore, our system makes use of constrained score technique for feature selection.

Pairwise constraints score is a supervision information for feature selection, which specifies whether a pair of data samples belong to the same class (must-link constraints) or different classes (cannot-link constraints). Pairwise constraints arise naturally in many tasks and are more practical and inexpensive than class labels. We compare the Constraint Score with LGD feature selection method. Experiments are carried out on LUNG data sets. Experimental results show that, Constraint Score achieves higher performance than LGD, in terms of most relevant feature selection, reduced redundancy etc.

In this paper we study about the related work done on the feature selection techniques in section II, the implementation details in section III where we see the system architecture, modules description, mathematical models, algorithms and experimental setup. In section V we provide a conclusion.

II. Related Work

In this paper [1], author proposes another element feature selection to all-inclusive minimize the component repetition with expanding the given element positioning scores, which can originate from any supervised or unsupervised methods. Our new model has no parameter with the goal that it is particularly suitable for any data mining application. Test results on benchmark information sets demonstrate that the proposed strategy reliably enhances the element determination results contrasted with the first methods. In between, they present another unsupervised worldwide and local feature element choice system which can be brought together with the global feature redundancy minimization structure and shows better performance.

In this paper [2] the creator focused on that in different data examination errands; one is reliably faced with high dimensional information. Highlight determination system are proposed to locate the appropriate component subset of the first highlights which can support grouping, recovery, request. In this paper, they consider the component determination issue in unsupervised learning condition, which is especially troublesome because of the nonappearance of class, denotes that would control the imperative data. The part choice issue is basically a combinatorial advancement issue which is computationally costly. Traditional unsupervised highlight decision tree address this issue by selecting the top arranged component considering certain scores arranged self-sufficiently for every part. These methods release the conceivable relationship between specific portions and in this manner can't convey perfect subset.

In this paper [3] the author exhibited that he adds to a face affirmation calculation which is brutal to substantial variety in lighting course and outward appearance. Taking a case grouping system, authors consider each pixel in a photo as a heading in a high-dimensional space. By take purpose of enthusiasm of the discernment that the photos of a particular face, under moving light however settled posture, lie in a 3D direct subspace of the high dimensional picture space if the face is a Lambertian surface without shadowing. Then again, since appearances are not by any stretch of the imagination Lambertian surfaces and manage in actuality produce self-shadowing, pictures will wander off-track from this straight subspace. Rather than explicitly showing this deviation, straightly broaden the photo into a subspace in a way which discounts those regions of the face with large deviation. Our projection system relies upon Fisher's Linear Discriminant and delivers all around separated classes in a low-dimensional subspace, even under amazing variety in lighting and outward appearances. The Eigenface method, another system in perspective of straightly expecting the photo space to a low dimensional subspace, has practically identical computational necessities. Yet, wide trial results demonstrate that the proposed "Fisher face" methodology has screw up rates that are lower than those of the Eigen face procedure for tests on the Harvard and Yale Face Database

In this paper [4] the author considered that in various data examination endeavors; one is as often as possible confronted with high dimensional data. Highlight Constraint score for Feature Selection by means of Global Redundancy Minimization determination frameworks are planned to find the appropriate component subset of the first highlights which can encourage grouping, arrangement and recovery. In this paper, by considering the segment determination issue in unsupervised learning circumstance, which is particularly troublesome in view of the nonattendance of class, denotes that would deal with the journey for important information. The component determination issue is essentially a combinatorial streamlining issue which is computationally unreasonable. Standard unsupervised highlight determination systems address this issue by selecting the top situated parts considering certain scores figured self-ruling for each segment. These strategies neglect the possible connection between assorted components and therefore can't make a perfect component subset. Motivated from the late upgrades on complex learning and L1-regularized models for subset

determination, in this paper another philosophy, called Multi-Cluster Feature Selection (MCFS), for unsupervised component determination.

In this paper the author [5] examined that the Anti-radical (Census Transforms hISTogram), another visual descriptor for seeing topological spots or scene classes, is exhibited in this paper. Place and scene acknowledgment, especially for indoor circumstances; require its visual descriptor to have properties that are particular from other vision spaces (e.g. object acknowledgment). Against radical satisfies these properties and suits the spot and scene acknowledgment task. It is a widely inclusive representation and has strong generalization for class acknowledgment.Anti-radical generally encodes the fundamental properties within a photo and smothers point by point textural information. Our tests demonstrate that CENTRIST beats the current state of-risk in a couple spot and scene affirmation datasets, differentiated and distinctive descriptors, for instance, SIFT and Substance. In addition, it is definitely not hard to execute. It has no parameter to tune, and surveys to an extraordinary degree quick.

In this paper [6] the author showed that the most research in accelerating content mining incorporates algorithmic moves up to impelling calculations, but then for some tremendous scale applications, for instance, indexing large record storage facilities, the time spent isolating word highlights from compositions can itself surpass the starting get ready time. This paper delineates a speedy technique for substance component extraction that overlays together Unicode change, constrained lowercasing, word limit location, and string hash count. Paper show that our entire Constraint score for Feature Selection by means of Global Redundancy Minimization number hash element result in classifiers with similar true execution to those made using string word components, yet require far less estimation and less memory.

In this paper [7] author have proposed a system for overall excess minimization. The excess is reduced by applying the GRM structure, and characterization precision has upgraded in a general sense for both unsupervised also, managed highlight determination calculations. This delineate the adequacy of the GRM structure, which minimize the excess between picked features, along these lines, the picked components are depended upon to be more minimized moreover, discriminant.

In this paper [8] the writer showed that the component subset determination issue, a learning calculation is stood up to with the issue of selecting an appropriate subset of components whereupon to focus on its though, while disregarding the rest. To finish the best execution with a particular learning calculation on a particular get ready set, a segment subset determination system should consider how the computation and the readiness set interface. By explore the association between perfect part subset determination and relevance. Our wrapper system chases down a perfect segment subset redid to a calculation and an area. By contemplate the qualities and deficiencies of the wrapper approach and show a movement of improved diagrams. Also by balance the wrapper approach with affectation without highlight subset determination moreover, to Relief, a channel approach to manage highlight subset decision. Significant change in accuracy is proficient for some datasets for the two gatherings of inciting estimations used: decision trees and Innocent Bayes.

In paper [9] authors, A. Mariello and R. Battiti proposed a novel approach for feature selection based on the minimization of the neighborhood entropy, which corresponds to the maximization of the MI between the features and the output variable. The locally optimal subset of features is selected by using a greedy procedure and the LSH index for NNs. Authors also compared their proposed algorithm with some of the most effective methods for selecting features based on MI, NNs, and correlation.

In paper [10] authors, X. Cai, F. Nie, and H. Huang propose a novel robust and pragmatic feature selection approach. Unlike those sparse learning based feature selection methods which tackle the approximate problem by imposing sparsity regularization in the objective function, the proposed method only has one `2;1-norm loss term with an explicit `2;0-Norm equality constraint. An efficient algorithm based on augmented Lagrangian method will be derived to solve the above constrained optimization problem to find out the stable local solution.

In paper [11] authors, Quanquan Gu, Zhenhui Li, Jiawei Han finds a subset of features, based on which the label correlation regularized loss of label ranking is minimized. The resulting multi-label feature selection problem is a mixed integer programming, which is reformulated as quadratically constrained linear programming (QCLP). It can be solved by cutting plane algorithm, in each iteration of which a minimax optimization problem is solved by dual coordinate descent and projected sub-gradient descent alternatively.

In paper [12], authors D. P. Bertsekas considered optimization problems that were subject to constraints. These include the problem of allocating a finite amounts of bandwidth to maximize total user benefit, the social welfare maximization problem and the time of day pricing problem. They make frequent use of the Lagrangian method to solve these problems. This appendix provides a tutorial on the method.

III. Implementation Details

A. System Overview

The system consisting of following modules:

- Identify feature similarities
- Feature score calculation
- Removal of redundant features
- Top k feature selection
- KNN classification

Following fig.1 shows the proposed system architecture.

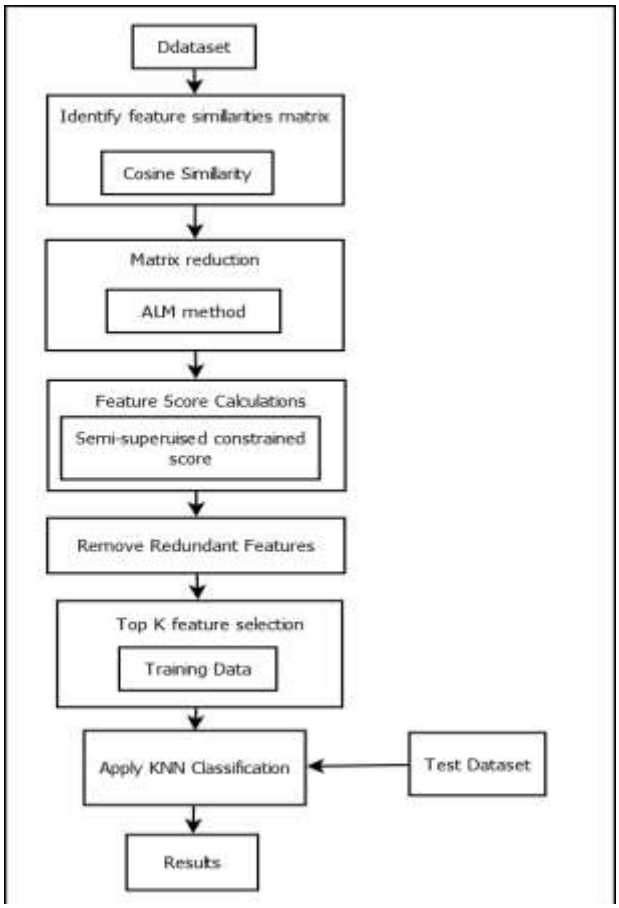

Fig. 1: System Architecture

IV. Conclusion and Future scope

The system is proposed for feature selection and redundancy minimization and also survey is done on various feature selection techniques. System uses semi-supervised constrained score algorithm for feature selection algorithm instead of LGD feature selection algorithm. The redundancy is decreasedby applying ALM method. The classification accuracyis improved significantly for semi supervised feature selection algorithm. The proposed system effectively minimizes the redundancy between selected features. The experimental results showed that the proposed system is better because it provides highly correlated results within same group.

References

[1]. De Wang, Feiping Nie, and Heng Huang, "Feature Selection viaGlobal Redundancy Minimization", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 10, OCTOBER 2015

[2]. D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data,"in Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2010, pp.333-342.

[3]. P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces:Recognition using class specific linear projection," IEEE Trans. Pattern Anal. Mach.Intell., vol. 19, no. 7, pp. 711-720, Jul. 1997.

[4]. X. Chang, F. Nie, Y. Yang, and H. Huang, "A convex formulation for semisupervisedmulti-label feature selection," in Proc. AAAI Conf. Artif. Intell., 2014,pp. 1171-1177.

[5]. J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization,"IEEE Trans. Pattern Anal. Mach. Intell., vol. 33,no. 8, pp. 1489-1501, Aug. 2011
[6]. G. Forman and E. Kirshenbaum, "Extremely fast text feature extraction for classificationand indexing," in Proc. Int. Conf. Inf. Knowl. Manag., 2008, pp. 1221-1230.
[7]. R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artif. Intell., vol.97, no. 1/2, pp. 273-324, 1997.
[8]. Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques inbioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.
[9]. A. Mariello and R. Battiti, "Feature Selection Based on the Neighborhood Entropy," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 12, pp. 6313-6322, Dec. 2018.
[10]. X. Cai, F. Nie, and H. Huang, "Exact top-k feature selection via l2,0-norm constraint," in Proc. 23rd Int. Joint Conf. Artif. Intell., 2013, pp. 1240-1246.
[11]. Quanquan Gu, Zhenhui Li, Jiawei Han, "Correlated Multi-Label Feature Selection", CIKM'11, October 24-28, 2011, Glasgow, Scotland, UK.
[12]. D. P. Bertsekas. Constrained Optimization and Lagrange Multiplier Methods. Belmont, MA, USA: Athena Scientific, 1996.