A SURVEY ON DATA DIMENSIONALITY REDUCTION TECHNIQUES

Fouziya Banu.H[*1], Thenmozhi.N[2]

[*1]Department of Information Technology, Government Arts College, Coimbatore,Tamilnadu,India.
[2]Department of Information Technology,Government Arts College,Coimbatore,Tamilnadu, India.

ABSTRACT

Data mining and machine learning methods face a formidable problem when dealing with high-dimensional data. Generally, the number of input variable is reduced to speed up and enhance decision making in data mining and machine learning methods. This can be achieved by dimensionality reduction technique. Dimensionality reduction is the analysis of methods to reduce the dimension which characterize the data. The main intention of dimensionality reduction technique us to remove the redundant and irrelevant data in order to minimize computing costs and avoid over-fitting data, and to enhance the quality of data for effective data-intensive processing tasks. This paper presents a detailed survey of different dimensionality reduction techniques. At first, different techniques developed by previous researchers for dimensionality reduction are studied in detail. Then, a comparative analysis is carried out to know the limitations of each technique and provide a suggestion for further improvement in dimensionality reduction.

Keywords: Machine learning, data mining, dimensionality reduction, data-intensive processing tasks.

1. INTRODUCTION

In the big data environment, huge volume of data generated from every minute. It is more complex to analysis such data. High dimensional data increases cost storage, requires lot of computing resources and it also affects the performance of data mining and machine learning algorithms. There exists a low-dimensional structure in high high-dimensional data, which capture the latent features of the high dimensional data. Dimensionality reduction [1] is applied in different applications such as regression analysis, influential observation, microarray gene expression data analysis, document indexing, image retrieval, etc.

Different dimensionality reduction techniques have been proposed to extract important features and data to help analyze high dimensional data. One of the easiest ways to reduce the dimensionality of data is by feature selection. It selects the most significant features for solving the particular problem. Feature extraction is another way to reduce the dimensionality of data which develops a transformation of the input space onto the low-dimensional subspace that preserves most of the relevant information. Feature selection and feature extraction [2] methods are used isolated or in combination with the intention to enhance performance such as comprehensibility of learned knowledge, estimated accuracy and visualization.

In this article, an analysis of different techniques related to dimensionality reduction is carried out to find a more efficient technique for dimensionality reduction. The main intention of this article is studying in detailed information on different techniques for dimensionality reduction. In addition, their limitations are addressed to further improve the dimensionality reduction process.

2. SURVEY ON DIMNENSIONALITY REDUCTION TECHNIQUES

Shanthi & Bhaskaran [3] proposed a Modified Artificial Bee Colony based Feature Selection (MABCFS) to select the predominant feature set from mammogram images. Each employed bee in MABCFS initialized with number of features and then it was investigated the new food source. This knowledge was communicated with the onlooker bees during it exploited the food sources that the employed bees discover. The best global solution of MABCFS was considered to enhance the use of Artificial Bee Colony (ABC) algorithm for feature selection. The selected features were used in classifier for classification of breast lesion.

Sasikala et al. [4] proposed a Shapely Quality Embedded Genetic Algorithm (SVEGA) based feature selection for improved survivability diagnosis of breast cancer. In SVEGA, two memetic operators were included in the embedded Shapely value and eliminate features that made the genetic algorithm solution possible. The system arranged the genes based on their class differentiation capability. It selected the genes that fine tuned the potential of different classes to discriminate. This reduced the dimensionality of features and significantly improved the classification accuracy rate.

Peralta et al. [5] proposed MapReduce for Evolutionary Feature Selection (MR-EFS) for feature selection to classify big data. Initially, a MapReduce algorithm was designed where the original data was split into number of blocks which is equal to the number of mappers. Then in the mapper phase, EFS process was carried out and the solution of each mapper was combined in the reducer phase. It allowed the feature selection process to be implemented flexibly using a threshold which evaluated the selected features. Support Vector Machine (SVM), Logistic Regression (LR) and Naïve Bayes (NB) were processed the selected features for big data classification.

Suji & Rajagopalan [6] proposed Multi Ranked Feature Selection Algorithm (MRFSA) based feature selection for efficient breast cancer detection. MRFSA was developed based on FOREST algorithm and Enhanced Multiclass SVM (EMSVM). In MRFSA, information gain ratio of all features was calculated. Then, the features were ranked based on the calculating feature weights by FOREST algorithm. It returned a best subset of features which was given as input to EMSVM for breast cancer detection.

Galván-Tejada et al. [7] proposed a multivariate feature selection for breast cancer diagnosis. The breast cancer diagnosis model was built using K Nearest Neighbor (KNN), Nearest Centroid (NC) and Random Forest (RF) strategies. The result of these models was processed as cost function in a genetic algorithm. In the multivariate model, two texture descriptor features were extracted which had a similar or better ability to predict breast cancer. It identified the data result compared to the multivariate model composed of all the features based on the fitness value. This model thus reduced the radiologist's workload.

Wang et al. [8] proposed weighted feature selection strategy for feature selection of microarray gene expression cancer data. The weighted feature selection strategy distinguished the features by their classification performances, occurrence frequency in population based on two matrices. In the weighted feature selection strategy, different objectives such as minimizing the computational cost, minimizing number of features and maximizing the performance was considered to fine tune the features through bacterial colony optimization algorithm.

Shi et al. [9] proposed an Unsupervised Multi-view Feature Extraction with Dynamic Graph Learning (UMFE-DGL) for feature extraction. A unified learning framework was devised to concurrently performed dynamic graph learning and feature extraction. The dynamic graph learning adaptively captured the intrinsic multiple view-specific relations of samples. Feature extraction learned the projection matrix which consequently preserved the dynamically adjusted sample relations modeled by graph into the low-dimensional features.

Zhang et al. [10] proposed low-rank affinity matrix based feature extraction for biological recognition. The affinity matrix was designed to better preserve the underlying low-rank structure of data representation revealed by Low-Rank Representation (LRR). The main intention of LRA-DP is to enhance the method by optimizing the affinity matrix of LRR. It considered that the more block-diagonal the affinity matrix is, the better discriminative projection obtained. For each iteration, K max singular values were selected and Inexact ALM algorithm was processed to calculate the affinity matrix of LRR.

Viegas et al. [11] presented a Genetic Programming approach for high efficient feature selection technique that an efficient selection of the significant features was offered. Here, two main challenges such as curse of dimensionality and skewed data classification were considered for Automatic Document Classification (ADC). The proposed solution used the space of possible combinations of features selected via basic metrics to establish an unbiased estimator of the features ' discriminative power. Numerous feature space projections were combined with the proposed approach, optimizing classification accuracy and capturing the strongest feature-class relationships. In this method, due to data skewness, the problem of weighting and combining numerical values ranging from different scales to poor feature choice was avoided.

Zheng et al. [12] proposed two formulations of Harmonic mean based Linear Discriminant Analysis (HLDA) and HLDA pairwise (HLDAp) for dimensionality reduction. The HLDA used the harmonic mean based pairwise between-class distance for dimensionality reduction. The HLDAp was an extended version of HLDA that used for multi-label classification problems. HLDA and HLDAp ensured that there are no small between-class distances in subspace, thus enhanced the classification performance.

3. RESULT AND DISCUSSION

A comparative analysis of the merits and demerits of different dimensionality reduction techniques whose functional information is discussed in the above section is presented. The following Table 1 gives the merits and demerits of the above mentioned dimensionality reduction techniques.

Table.1 Comparison of Dimensionality Reduction Techniques

Ref. No.	Methods Used	Merits	Demerits	Performance Metrics
[3]	MABCFS	Enhance classification quality	It was applicable in a clinical environment to small databases.	For Mammographic Image Analysis Society (MIAS) database: Accuracy = 96.89% For Digital database for screening mammography (DDSM) database: Accuracy = 97.17%
[4]	SVEGA	Reduced dimensionality of data significantly improved the classification accuracy	Classification accuracy needs to be improved further	Classification accuracy: J48 = 93.81% SVM = 91.75% NB= 88.5% KNN = 82.48%
[5]	MR-EFS	Flexible for high dimensional data	Threshold value highly influences the classification accuracy	Area Under Curve (AUC): LR = 0.7 NB = 0.7127 SVM = 0.6865 Training runtime: LR = 367.29 sec NB = 605.14 sec SVM = 334.18 sec
[6]	MRFSA, FOREST, EMSVM	Better accuracy	Proper selection of kernel function for EMSVM is more difficult	Classification Accuracy = 95.98
[7]	multivariate feature selection, KNN, NC, RF	Reduce workload	High false positive rate which affect the prediction accuracy	False Positive: RF = 10 KNN =8 NC = 13 False Negative: RF = 5

				KNN 19 NC = 23
[8]	weighted feature selection strategy, bee colony optimization	Reduce computational complexity	It has to confront with the challenge to determine an appropriate search space for high classification accuracy without prior knowledge of datasets	For 9_Tumor s (5920) dataset: Classification accuracy = 0.9222
[9]	UMFE-DGL	Converge efficiently	Has parameter sensitivity problem	For MSRC-v1 dataset: Purity = 0.7095 For YouTube dataset: Purity = 0.3668 For outdoor scene dataset: Purity = 0.4337
[10]	Low-rank affinity matrix	Underlying low-rank structure of data representation preserved by LRA-DP is helpful for classification problem	High computational complexity	Recognition rate = 99%
[11]	Genetic programming approach	Poor feature choice is avoided	Has convergence problem	For Top-42096 Features of Collection ACL-BIN: Standard deviation = 0.21 For Top-16280 Features of Collection 20NG: Standard deviation = 0.41

[12]	HLDA, HLDAp	Better performance by using arithmetic mean based between-class distance	Most time expensive computation comes from the initialization part of HLDA and HLDAp	For PIE dataset: Average Precision: HLDA = 0.9007 HLDAp = 0.8805 For MediaMill dataset: Average Precision: HLDA = 006975 HLDAp = 0.6943 For Barcelona dataset: Average Precision: HLDA = 0.8946 HLDAp = 0.8870

4. CONCLUSION

In this paper, a detailed analysis on different dimensionality reduction techniques was presented. Evidently, it shows all researchers tried to enhance their techniques for dimensionality reduction than the conventional dimensionality reduction techniques. Based on the analysis, it is known that the HLDA and HLDAp based dimensionality reduction method has better performance than other dimensionality reduction methods. However, most time expensive computation comes from the initialization part of HLDA and HLDAp methods. In future, this problem is considered to further enhance the performance of dimensionality reduction process.

REFERENCES

[1] Tian, H., Lan, L., Zhang, X., & Luo, Z. (2019). Neighbors-Based Graph Construction for Dimensionality Reduction. *IEEE Access*, *7*, 138963-138971.

[2] Chao, G., Luo, Y., & Ding, W. (2019). Recent Advances in Supervised Dimension Reduction: A Survey. *Machine Learning and Knowledge Extraction*, *1*(1), 341-358.

[3] Shanthi, S., & Bhaskaran, V. M. (2014). Modified artificial bee colony based feature selection: a new method in the application of mammogram image classification. *International Journal of Science, Engineering and Technology Research*, *3*(6), 1664-1667.

[4] Sasikala, S., alias Balamurugan, S. A., & Geetha, S. (2015). A novel feature selection technique for improved survivability diagnosis of breast cancer. *Procedia Computer Science*, *50*, 16-23.

[5] Peralta, D., del Río, S., Ramírez-Gallego, S., Triguero, I., Benitez, J. M., & Herrera, F. (2015). Evolutionary feature selection for big data classification: A mapreduce approach. *Mathematical Problems in Engineering*, *2015*, 1-11.

[6] Suji, R. J., & Rajagopalan, S. P. (2016). Multi-ranked feature selection algorithm for effective breast cancer detection. *Biomedical Research*.

[7] Galván-Tejada, C. E., Zanella-Calzada, L. A., Galván-Tejada, J. I., Celaya-Padilla, J. M., Gamboa-Rosales, H., Garza-Veloz, I., & Martinez-Fierro, M. L. (2017). Multivariate Feature Selection of Image Descriptors Data for Breast Cancer with Computer-Assisted Diagnosis. *Diagnostics*, *7*(1), 9.

[8] Wang, H., Jing, X., & Niu, B. (2017). A discrete bacterial algorithm for feature selection in classification of microarray gene expression cancer data. *Knowledge-Based Systems*, *126*, 8-19.

[9] Shi, D., Zhu, L., Cheng, Z., Li, Z., & Zhang, H. (2018). Unsupervised multi-view feature extraction with dynamic graph learning. *Journal of Visual Communication and Image Representation*, *56*, 256-264.

[10] Zhang, N., Chen, Y., Xi, M., Wang, F., & Qu, Y. (2018). Feature extraction based on Low-rank affinity matrix for biological recognition. *Journal of computational science*, *27*, 199-205.

[11] Viegas, F., Rocha, L., Gonçalves, M., Mourão, F., Sá, G., Salles, T., ... & Sandin, I. (2018). A Genetic Programming approach for feature selection in highly dimensional skewed data. *Neurocomputing*, *273*, 554-569.

[12] Zheng, S., Ding, C. H., Nie, F., & Huang, H. (2018). Harmonic Mean Linear Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering*.