# DIFFICULTY IN HANDLING HIGH DIMENSIONAL DATA- PLEASE STAND UP

**Shikha Agarwal[1], R. Rajesh[2]**

[1]*Department of Computer Science, Central University of Bihar, BIT Campus, Patna*
[2]*Department of Computer Science, Central University of Bihar, BIT Campus, Patna*

## Abstract

*With the advancement of technology huge amount of high dimensional data are getting generated. This paper present's the high dimensional data as a serious problem for computational analysis. Broadly high dimensional data could be handled in two ways, firstly with dimensionality reduction, secondly without dimensionality reduction using classical machine learning. Two main categories of dimensionality reduction are feature selection and extraction. Feature selection and feature extraction. Feature selection selects a subset of features based on some criteria while feature extraction method transforms the data into the lower dimensional space. Experiment result of the high dimensional data shows the performance improvement of classifier before and after dimensionality reduction.*

***Keywords—*** *Feature Selection, Feature Extraction, Wrapper Method, Embedded Methods, IBPSO*

----------***----------

## 1. INTRODUCTION

Huge amount of information in form of data is need of modern century and with the advancement of technology data generation is very simple and various methods have been proposed in literature to handle the high dimensional data [1]. The difficulties regarding handling high dimensional data can also be seen in the paper written by Richard Bellman [2], the inventor of curse of dimensionality where he explained the cases where the dimensionality grows exponentially. In high dimensional data, total number of features is very high as compare to the number of samples. Hence converting the high dimension data into the low dimension is very essential for classifier performance. According to kumar et. al. [3] there is lower limit in the numbers of features to be selected and below which classifier will degrade rather than improve. In most of the cases the information loses due to this dimensionality reduction is compensated by choosing the most accurate features or more accurate mapping in lower dimension.

This paper discusses about the different sources of high dimension data and few methods to reduce the dimensionality of data. Experimental studies are done on the micro array gene expression data, to show; the classification performance.

## 2. SOURCE OF HIGH DIMENSIONAL DATA

High dimensional data matrix of N rows and P columns, Rows represents the samples and columns corresponds different attributes or variables.

Web Term-Document Data is good example of high dimensional data, which is the collection of frequency of appearance of given key word in the given document, in a suitable normalization [3]. In Sensor Array Data sensor records observations over a period of seconds, at a rate of X thousand samples, second. Some other data includes, but not limited to Gene Expression Data, obtained from the huge number of genes relative to very few number of different cell lines [4]-[5].Exchange rate of foreign currencies for very long period of time [6].

## 3. DIMENSIONALITY REDUCTION

Feature selection and feature extraction are confusing term in dimension reduction. Feature selection method selects the subset of features. Feature selection is for a given feature set of size N, find a subset of size M, where M<N which is capable to optimizes an objective function f(X).

Feature extraction reduces the dimensionality by projection of higher dimension vector into lower dimension vector. Feature extraction for a given feature set X of size N find a transform feature vector Y such that

$$Y=f(X),$$

$$Y= \{Y_i | i = 1....M\},$$

Where M<N

Two methods of dimensionality reductions have been discussed here, namely feature selection and feature extraction.

### 3.1 Feature Selection

The Feature selection can be classified into three categories (1) Filter Feature selection, (2) Wrapper Feature Selection and (3) Embedded Feature Selection.

### 3.1.1 Filter Feature Selection

These methods do not count the effect of the selected features on the accuracy of classifier. Most commonly used measures include the mutual information [8], point wise mutual information [9] and Pearson product-moment correlation coefficient [10]. Few popular filter feature selection methods are Information Gain (IG) [8], t-test [11] and Relieff method [12].

The main advantage of filter method is that it can be easily implemented on the high dimensional datasets, computationally simple and fast and do not depends on the classifier. This independence from the classifier is also a big draw backs of this method. Ignorance of feature dependencies is also another problem.

### 3.1.2 Wrapper Feature Selection

In these methods features are selected according to their fitness given by classifier. In this data is partitioned into three sets namely; Training, validation, and test sets. First predictor need to be trained for each feature subset on training data. Select the best feature subset based on validation data. Repeat the procedure to reduce variance (cross validation). Finally perform testing. Exhaustive search [13], Genetic Algorithm [14] and Branch-and-Bound algorithm [15] are few examples.

The advantage of wrapper method is dependence between selected features and predictor method. It also considers the dependencies of the features. The disadvantage is the risk of over fitting and it is computationally costly while building the classifier.

### 3.1.3 Embedded Feature Selection Methods

Built-in feature selection methods are named as embedded feature selection. Consider decision tree [16] which selects the optimal feature subset while creating the tree. Another example of this type of feature selection is pruning step of apriori algorithm. They perform variable selection as part of the learning procedure. These methods are less computationally costly than wrapper methods, but these methods are specific to a learning machine.

### 3.2 Feature Extraction

Feature extraction is used to project the data in low dimension space. Few Feature extraction techniques includes: Principle Component Analysis [17], Isomer [18], Kernel PCA [19], Partial Least Squares [20] analysis etc.

### 3.2.1 Principal Component Analysis (PCA)

PCA is method which projects the high dimensional data onto a search space of small dimensionality. Variance of the original data is handled by the first principal component, second principal component and so on in decreasing order.

## 4. MACHINE LEARNING METHODS

In this section some of the classical classification methods are discussed.

**Probabilistic Neural Network (PNN):** PNNs [21] belong to the Radial Basis Function (RBF) neural network which is single hidden layer feed-forward neural networks. No weight is involved while passing the input to the hidden layer. The weights between hidden and the output layer are adjusted using a least squares optimization algorithm. A key advantage of RBF networks is that training of RBF is much more efficient than NNs.

**K-nearest Neighbour (KNN):** The key principle of KNN [22] is that it represents all the samples as points in the search space and an unseen sample is classified based on distance from K-nearest neighbours instances as determined by some distance measure such as Euclidean Distance.

**Extreme Learning Machine (ELM):** ELM is second name of single-hidden layer feed forward neural networks. The learning of ELM [23] ends in few seconds for many applications. The ELM over perform the gradient-based learning such as back propagation in most cases [15].

**Binary Particle Swarm Optimization:** Particle swarm optimization (PSO) [24] is an optimization technique, developed by Kennedy and Eberhart in 1995 which mimic the swarming behaviour of organisms, such as birds in a flock and fish in a school. A improved binary PSO (IBPSO) was given by L.-Y. Chuang in 2008 which resets the global best to the zero if for three iteration value of global best does not change.

## 5. EXPERIMENTS AND RESULTS

In this section, a systematic study of performance of dimensionality reduction followed by classification of high dimensional data is performed.

The datasets [15] consist of seven gene expression profiles datasets (http://www.gemssystem.org). The normalization of data was done by dividing the data with the maximum value in the matrix.

Gene expression data used in this experiment had 2–26 different categories of class, 50–308 samples (patients) and 2308–15009 features. In all cases leave one out cross validation (LOOCV) is performed since the number of records one are less when compared to the number of features. In PNN the optimized spread value between 0.01 to 1 is used. In case of ELM RBF kernel regularization coefficient is set to one. In IBPSO k-NN is used as the fitness evaluation function, the inertia weight is 0.5, and learning rate c1 and c2 having value 2 and the number of iterations is set to 100.

The result of experiment is shown in table 1.The average highest classification accuracy of k-NN, PNN, ELM and IBPSO-kNN are 79.40, 85.59, 88.51 and 90.79 respectively.

Classification with dimensionality reduction showed better performance than without dimensionality reduction.

Figure 1 shows the comparison of the classification accuracy of different methods. From the results it is clear that performance of classifier increases with dimensionality reduction.

**Table 1:** Classification accuracy of different classical classification methods and new nature inspired dimensionality reduction method

| Datasets | Without D-Reduction | | | With D-Reduction |
|---|---|---|---|---|
| | *K-NN* | *PNN* | *ELM* | *IBPSO k-NN* |
| 9_Tumors | 53.33 | 55.0 | 56.34 | 66.67 |
| Leukemia1 | 87.50 | 90.28 | 89.65 | 95.83 |
| Leukimia2 | 70.00 | 97.22 | 98.91 | 98.61 |
| Lung Cancer | 90.15 | 85.66 | 93.01 | 94.78 |
| SRBCT | 91.57 | 93.98 | 93.98 | 98.62 |
| Prostate_Tumor | 76.47 | 82.18 | 90.56 | 86.27 |
| DLBCL | 87.01 | 94.81 | 97.18 | 94.81 |
| Average | 79.4 | 85.59 | 88.51 | 90.79 |

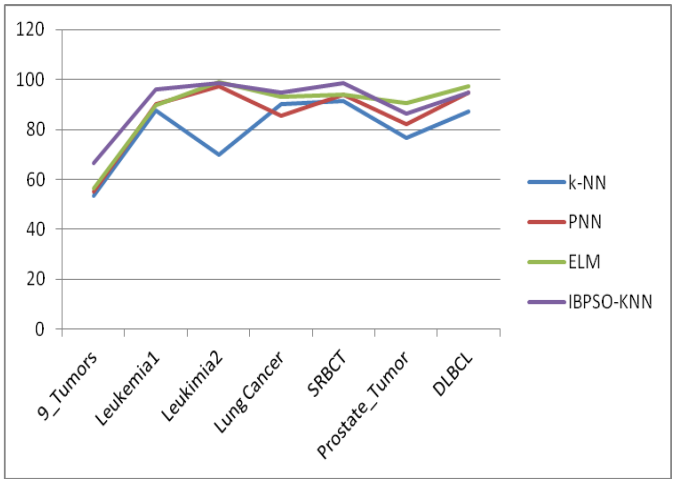


**Fig 1** Graphical comparison of classification accuracies obtained via different methods.

## 6. CONCLUSION

Handling high dimensional data with the classical methods does not fulfill the requirements of modern trend of data analysis. Classification through classical machine learning methods and new nature inspired method IBPSO implemented on gene expression data sets. IBPSO has shown best performance in terms of dimensionality reduction, classification accuracy and time complexity among all implemented methods.

## REFERENCES

[1] M. Verleysen, *Learning high-dimensional data*, Limitations and Future Trends in Neural Computation, S. Ablameyko et al. (Eds.), IOS Press, pp. 141-162, 2003.

[2] R. Bellmann, Adaptive *Control Processes: A Guided Tour*. Princeton University Press, 1961.

[3] A. P. Kumar, P. Valsa*la*, *"Feature Selection for high Dimensional DNA Microarray data using hybrid approaches*" Bioinformation, vol. (19) 6. 2013.

[4] F. Murtaugh, J. L. Starck, M. W. Berry, "*Overcoming the curse of dimensionality in clustering by means of the wavelet transform*," Computer Journal, vol. 43, pp. 107-120, 2000.

[5] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. Mckeown, V. Iragui, T. Sejnowksi, "*Removing Electroencephalographic Artifacts by blind source separation*," Psychophysiology, vol. 37, pp. 163-178, 2000

[6] D. E. Bassett, M. B. Eisen, M. S. Boguski, "*Gene expression informatics – it's all in your mine*," Nature Genetics Supplement, vol. 21, pp. 51-55, 1999.

[7] C. Dunis, B. Zhou, *Nonlinear Modelling of High-Frequency Financial Time Series*, J. Wiley, New York, 1998.

[8] M. R. Fazlollah, "*An Introduction to Information Theory,"* Dover Publications, Inc., New York, 1961.

[9] K. W. Church, P. Hanks, *"Word association norms, mutual information, and lexicography,"* Comput. Linguist, Vol. 16 (1), pp. 22–29, March 1990.

[10] P Karl, *"Notes on regression and inheritance in the case of two parents,"* Proceedings of the Royal Society of London, vol. 58 , pp. 240–242, June 20, 1895.

[11] R Mankiewicz, *The Story of Mathematics*, Princeton University Press, pp. 158.

[12] K. Kira, and L Rendell, *"The Feature Se l ection Problem: Traditional Methods and a New Algorithm,"* AAAI-92 Proceedings, 1992.

[13] C. Paar, J. Pelzl, B. Preneel, "*Understanding Cryptography: A Textbook for Students and Practitioners*," Springer, pp. 7.

[14] D. Goldberg, *"Genetic Algorithms in Search, Optimization and Machine Learning*," Reading, MA: Addison-Wesley Professional, 1989.

[15] J. Clausen, *Branch and Bound Algorithms—Principles and Examples* Technical report, University of Copenhagen, 1999.

[16] J. R. Quinlan, "*Simplifying decision trees*". *International Journal of Man-Machine Studies,"* vol. 27 (3) , pp. 221, 1987.

[17] K. Pearson, *"On Lines and Planes of Closest Fit to Systems of Points in Space,"* Philosophical Magazine, vol. 2 (11), pp. 559–572, 1901.

[18] J. B. Tenenbaum, V. de Silva, J. C. Langford, "*A Global Geometric Framework for Nonlinear Dimensionality Reduction,"* Science, vol. 290, pp. 2319–2323, 2000.

[19] F. Husson, L. Sébastien & P. Jérôme, *Exploratory Multivariate Analysis by Example Using R*, Chapman & Hall/CRC The R Series, London, pp. 224, 2009.

[20] M. Tenenhaus, E. V. Vinzi, Y-M Chatelinc, C. Lauro, *"PLS path modeling,"* Computational Statistics & Data Analysis, vol. 48 (1), pp.159–205, Jan, 2005.

[21] D. F. Specht, *"Probabilistic neural networks,"* Neural Networks, vol. 3, pp. 109–118, 1990.

[22] D. Coomans, D.L. Massart, "*Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules,*" Analytica Chimica Acta, vol. 136, pp. 15–27.

[23] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, "*Extreme learning machine: Theory and applications,"* Neurocomputing, vol .70, pp. 489–501, 2006.

[24] A. Statnikov, C. F. Aligeris, L. Tsamardinos, D. Hardin, S. Levy, "*A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis,"* Bioinformatics, vol. 21 (5), pp. 631–643, Sep 2004.

[25] L.-Y. Chuang, H.-W. Chang, C.-J. Tu, C.-H. Yang, "*Improved binary PSO for feature selection using gene expression data,"* Computational Biology and Chemistry, vol. 32, pp. 29–38, 2008.