Indian Journal of Science and Technology, Vol 10(25), DOI: 10.17485/ijst/2017/v10i25/101205, July 2017
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

High Dimensional Classification - An Overview

Seetha Hari, Lydia Jane Gnanasigamani and Lijo Vellaplakkal Paulose

VIT University, Vellore – 632014, Tamil Nadu, India;
hariseetha@gmail.com, lydiajane.g@vit.ac.in, lijo.vp@vit.ac.in

Abstract

Objective: A comprehensive overview of high dimensional data classification techniques is presented for the benefit of researchers, scientists and data engineers in both government and private sectors working on large dimensional data. **Methods/Statistical Analysis:** A systematic approach was followed by studying and reporting the literature review for the years 1969-2016. **Findings:** The high dimensional data classification is found to be a challenging task as the data will not fit into main memory as required by conventional classification methods. Many of the features would be irrelevant and as the dimensionality increases with limited number of samples any conventional supervised learning algorithm may over fit to noise. The present study reveals the methods to generate artificial samples to increase the size of training data for better classification performance. It is also noted that reducing dimensionality not only reduces the storage space and computational time but increases the understandability. **Applications:** Text Classification, Email Classification, Pattern Classification, Information Retrieval, Gene Expression Analysis, Health Care Analysis, Predictive Modelling.

Keywords: Dimensionality Reduction, Feature Selection, GA, LSA, PCA, Synthetic Pattern Generation

1. Introduction

The development of data gathering, data generation tools and storage media has not only led to large amounts of data but also led to explosive growth of dimensionality of each sample. The emergence of many new areas such as bioinformatics, WWW, ecommerce, health care industry, computer vision etc., gave rise to high dimensional data. In a high dimensional data the number of attributes is much larger than the number of samples. Classification of high dimensional data is useful in text categorization, pattern classification, remote sensing, credit scoring, sentiment analysis, medical diagnosis etc.[1-4] In a classification problem, one of the attributes of the sample (pattern or object) represents the target class of the sample and the other attributes define the characteristics of the sample. The goal of classification is to build a model using labelled samples. This model is then used to classify the samples (from their attribute values) for which the class label is unknown. Nearest neighbour classifiers, Naïve Bayes classifier, Decision tree and Support Vector Machine are some of the popular classifiers. These traditional classifiers were found to be extremely useful in case of low dimensional data but they break down in case of high dimensional data.

The most common challenges in classification of high dimensional data are:

- Curse of dimensionality
- decrease in the specificity of similarities between samples in high dimensional space
- existence of noise or outliers
- Need more computational time and memory

Curse of Dimensionality is a term coined by Bellman that refers to the diminishing performance of the classifier due to the increase in dimensionality. As the dimensionality increases the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with the number of

*Author for correspondence

dimensions that it comprises.[5] It was reported in literature that the number of training samples per class should be at least 5–10 times the dimensionality of the.[6,7] The performance of the classifier increases with increasing dimensionality until some optimal dimensionality is reached beyond which its performance decreases with increasing dimensionality. In order to overcome the problem of "Curse of Dimensionality" either we have to generate artificial patterns or reduce the dimensionality.

2. Synthetic Pattern Generation

Many real world data sets are small in size but large in dimensionality. So we need to generate artificial patterns to improve the classification performance. In the reported article proposed a novel pattern synthesis method called partition based pattern synthesis which can generate an artificial training set of exponential order when compared with that of the given original training set.[8] The features of the training set of each class are partitioned in such a way that the features in a partition are better correlated with each other than features in different partitions. The absolute value of the correlation coefficient was used as a similarity measurement between the features of each class for partitioning them into different blocks. This is a heuristic method. For a given number of partitions (p) it first finds the p features that are least correlated with each other and assigns each of these p features to different p blocks respectively. The remaining features are then assigned to a block such that its average correlation with features in the block is high when compared with that of other blocks. After partitioning, they used divide and conquer strategy for nearest neighbour classification. In this method, for each class, the nearest neighbours of each sub-pattern of the test pattern are determined from each block. Then synthetic patterns are generated for each class by taking Cartesian product of these nearest neighbours from each block. The k local nearest neighbours for a given test pattern are determined from each class of synthetic patterns. The union of local neighbours from each class is determined and the test pattern is classified based on majority voting.

In the proposed article a bootstrapping method that creates (not selects) new training samples.[6] It not only generates synthetic patterns but also smoothes the distribution of training samples. It was successfully applied in the design of 1NN classifier, particularly, in high dimensional spaces. Further, the bootstrap samples were generated by combining the training data locally and illustrated that the NNC (nearest neighbour classifier) based on bootstrap patterns performed better than that of k-NNC (k-nearest-neighbor classifier) based on the original data.[9]

In the article applied multiple kernel learning approach to generate synthetic approach.[10] The training set is initially partitioned class wise. The class wise data is then bootstrapped to remove variations in the data.[11] The class wise bootstrapped data is then partitioned into p blocks. Using one class SVM classifier on each block of partitions, support vectors are generated. The Cartesian product of these support vectors represents the synthetic samples for each class. The union of all these synthetic samples of all classes respectively generates synthetic training set. The classification performance of SVM classifier was shown to better using this synthetic set.

Several works were reported in the literature for online and offline handwritten character recognition using artificial patterns that showed improved recognition rate. In the article applied deformation models to character pattern image to produce pattern variations for off-line handwritten numeral recognition.[12] They applied the concept of perturbation to writing habits and instruments for off-line handwritten numeral recognition, and applied six types of linear distortion models to reverse an input image back to its standard form to solve the problem of patterns variation. In the reported article and also in another article it generated a huge number of training samples artificially in accordance with a non-linear distortion model for off-line handwritten Chinese characters recognition and showed that distorted sample generation is the most effective, followed by regularization of class covariance matrices and feature reduction using Fisher's discriminant when the dimension of the feature vector is high while the number of training samples is not sufficient.[13,14] In the article proposed a method of automatically generating deformed off-line character data by using, for each category, the pattern correspondence between the training samples and the template pattern.[15] In the article proposed a method to generate brush-written off-line patterns from on-line.[16] Notably in 2003 published work that used synthesised hand written text, generated from the distortion of real lines of text, to train an HMM-based hand written sentence recognize resulting in an overall performance improvement in the detection of handwritten sentences.[17,18] In the presented article presented an effective approach to enhance the

accuracy of on-line handwriting Japanese recognition by using a large amount of artificial patterns generated by the combination of various linear distortion models with a non-linear distortion model.[19]

Synthesis of training data for machine learning applications suffering from a shortage of initial training examples has been used in emotion recognition techniques, where speech is synthesised to train acoustic models for emotional speech recognition.[20] The use of synthesised training data has been widely used in the field of hand-written text recognition. In the article showed the use of semi-synthetic data sets for classifier generation in place of real-world data sets, specifically for detection of Mine Like objects from Sonar imagery.[21] In the article investigated the use of synthetic data in indirect determination of rock strength by employing Fuzzy C-Means (FCM) and Adaptive Neuro-Fuzzy Inference System (ANFIS) and Synthetic images were used in image classification to enable Unmanned Aerial Vehicles (UAVs) to see and avoid each other.[22,23]

In the article presented a method of water disaggregation at the meter point using a simple Hall Effect water meter and has compared the classification accuracy of ANN, SVM and KNN classifiers on predicting the fixture responsible for an event, using both labelled data collected over a two month period and synthesised data generated from only two extreme examples of labelled data per event class.[24] They show that by synthesising labelled training data, using a domain specific algorithm, an innovative water meter disaggregation system that uses Artificial Neural Networks (ANN), Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) classifiers can be trained in minutes rather than hours.

3. Dimensionality Reduction

The feature extraction and feature selection are very important in high dimensional classification. The bad performance of classifiers is caused by the error accumulation when estimating noisy features.[25] High dimensional data analysis needs large amount of memory and computational power and a classification algorithm may over fit to training sample and may not generalize well on new samples. A low dimensional representation improves the model's generalization ability and diminishes the risk of overfitting.[26,27] Feature selection is a combinatorial problem in the number of original features, and finding the optimal subset of variables is considered NP-hard.[26] Many researchers have extensively studied the dimensionality reduction methods (i.e. feature extraction and feature selection methods).[28-33] Some of the dimensionality reduction methods applied in high dimensional classification are as follows

3.1 Random Projection

Random projection is based on the article provided.[34-37]

It states that given $\in > 0$ and an integer n, let q be a positive integer such that $q \geq O(\varepsilon^{-2} \log n)$. For every set P of n points in R^d there exists $f : R^d \to R^q$ such that for all $u, v \in P$

$$(1-\varepsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\varepsilon)\|u-v\|^2 \quad (1)$$

In random projection, the original d-dimensional data is projected to a q-dimensional subspace through the origin, using a random matrix whose columns have unit lengths. Using matrix notation is the original set of n observations each of dimension d. $X^{R}_{q\times n} = R_{q\times d}\ X_{d\times n}$ is the projection of the data onto a lower q-dimensional subspace defined by[37]

$$X^{R}_{q\times n} = \frac{1}{\sqrt{q}}[R_{q\times d} X_{d\times n}] \quad (2)$$

The computational complexity of random projection is $O(dqn)$.

3.2 PCA (Principal Component Analysis)

The high dimensional data is projected by PCA onto the axes corresponding to the largest eigen values. Along these axes the variance of the data is maximum. Thus, the original data is transformed to the data that has fewer dimensions by choosing some of the eigen vectors corresponding to these eigen values.

The dimensionality reduction of the data set is achieved by projecting the data onto a subspace spanned by the most important eigenvectors

$$X^{PCA} = E_q^T X \quad (3)$$

Where the $d \times q$ matrix E_q contains q eigen vectors corresponding to the largest q eigen values.[25] The computational complexity of PCA is $O(d^2 n) + O(d^3)$.

3.3 LDA (Linear Discriminant Analysis)

The major goal of LDA is to compute an optimal transformation (projection) by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination.[38]

If the set of samples $y_1, y_2, \ldots\ldots\ldots\ldots y_n$ are divided into subsets Y_1, Y_2 and so on, and if the mean vector of samples belonging to class i is μ_i and the total mean vector is μ, where c is the number of classes. Then the within class scatter matrix S_w and between class scatter matrix S_B, are given as follows:

$$S_w = \sum_{i=1}^{c} \sum_{y \in Y_i} (y - \mu_i)(y - \mu_i)^t \text{ and } S_B = \sum_{i=1}^{c} n_i (\mu_i - \mu)(\mu_i - \mu)^t \quad (4)$$

Thus the goal of LDA is to maximize the ratio

$$\frac{|S_B|}{|S_W|} \quad (5)$$

The computational complexity of LDA is $O(dnt + t^3)$ where d is the number of features, n is the number of samples and $t = \min(n, d)$.

3.4 Non-Negative Matrix Factorization (NNMF)

NNMF seeks to find a lower rank approximation of the data matrix with non-negative elements where the factors that give the lower rank approximation are also non-negative.[39-43] Given a non-negative matrix (data set) $X_{n \times m}$ having n patterns with m attributes, NNMF finds non-negative matrices $W_{n \times q}$ and $H_{q \times m}$ that minimize the norm of the difference $X - W * H$. W and H are thus approximate non-negative factors of X i.e. $X \approx WH$. The value of rank q is selected according to the condition $q < \min(n, m)$ or according to $q < \left(m/(n+m)\right)$ in order to reduce the dimensionality.

3.5 Chi Square (CHI)

Chi-square or χ^2-distribution is a very popular feature selection technique and it is widely used in text classification.[44,45] The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation.[46]

Chi square measures dependence between a term and a category.[47] In the presented article discusses Chi square and its role in feature selection for text classification.[48]

The Chi is defined in general as follows:[49]

$$\chi^2 = \frac{\Sigma (Observed - Expected)^2}{Expected} \quad (6)$$

For a term t and category c for N number of documents

$$\chi^2(t, c_i) = \frac{N[P(t, c_i)P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i)p(\bar{t}, c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)} \quad (7)$$

the feature selection metrics CHI and some of the following metrics are the functions of four dependency tuples:

(t, c_i): presence of t and membership in c_i.
$(t, \bar{c}_i)$: presence of t and non-membership in c_i.
$(\bar{t}, c_i)$: absence of t and membership in c_i.
$(\bar{t}, \bar{c}_i)$: absence of t and non-membership in c_i.

3.6 Information Gain (IG)

Information gain is commonly used as term-suitability criterion in machine learning algorithms.[50] It estimates the number of bits of information obtained for the category prediction by checking the presence or absence of a term in a document.[51] The information gain of term t and category c is defined to be

$$G(t, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t' \in \{t, \bar{t}\}} P(t', c) \cdot \frac{\log(P(t', c))}{P(t') \cdot P(c)} \quad (8)$$

3.7 Gini Index

Gini index is a normalized version of the Gini coefficient. The Gini coefficient measures the inequality among values of a frequency distribution (for example, levels of importance). A Gini coefficient of zero (0%), one (100%) expresses perfect equality, maximal inequality respectively.[52] The Gini index is utilized well for the text classification and it is one of the most useful feature selection methods.[53-55] Gini index of items $x_1, x_2, \ldots x_n$ as follows:

$$G = \frac{\Sigma_{i=1}^{n} \Sigma_{j=1}^{n} |x_i - x_j|}{2 \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} x_j} \quad (9)$$

Gini Index for continuous probability distribution function $p(x)$, where $p(x)dx$ is the fraction of the population with value x to $x + dx$

$$G = \frac{1}{2\mu} \int_{\infty}^{\infty} \int_{\infty}^{\infty} [p(x)p(y)|x - y|dxdy] \quad (10)$$

where μ is the mean of the distribution for n observations

3.8 Correlation Coefficient (CC)

A correlation coefficient is a number that quantifies some type of correlation and dependence, meaning statistical relationships between two or more random variables or observed data values.[56] The theory and application of various information filters have been extensively discussed by the presented article have used correlation coefficient for feature extraction.[57,58] The correlation coefficient among two variables x and y with *n* observation is calculated as follows:

$$C_{xy} = \frac{\Sigma_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\Sigma_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\Sigma_{i=1}^{n} y_i^2 - n\bar{y}^2}} \quad (11)$$

In the reported article introduces a feature selection method via Correlation Coefficient, where using correlation coefficient clustering in removing similar/redundant features.[59]

3.9 Latent Semantic Analysis (LSA)

Latent semantic analysis (LSA) is a technique in natural language processing to analyse relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. LSA assumes that words that are close in meaning will occur in similar pieces of text.[60] LSA represents the theory and method for extracting and representing the contextual usage meaning of words by statistical computations applied to a large corpus of text.[61] The theory, methods and potential applications of LSA is explained in the presented article.[62] LSA is useful in automatic indexing, termed as Latent Semantic Indexing (LSI).[63,64] The LSI employs the Singular Value Decomposition (SVD) to find the lower dimensional subspace that consider the terms and document relationship in the form of term document matrix.[60] Let X be a term-document matrix and X can have a decomposition with two orthonormal matrices and one diagonal matrix.

$$X = U\Sigma V^T \quad (12)$$

Where U contains eigenvectors of XX^T and V contains eigenvectors of X^TX , and U and V are orthogonal matrices and Σ is diagonal matrix whose diagonal components correspond to singular values ordered in decreasing order. LSI assumes that there is an underlying latent semantic structure in the term usage across documents, and the basic idea is that such structure can be revealed by dropping small singular values in Σ.[65]

$$\hat{X} = U\hat{\Sigma}V^T \quad (13)$$

Where $\hat{\Sigma}$ is the matrix with n largest values set to zero. $\hat{X}$ is an optimal approximation for X in terms of mean square error.

Hofmann introduces a variation for the LSI with probabilistic approach in LSA.[66]

3.10 Mutual Information (MI)

In probability theory and information theory, the Mutual Information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" (in units such as bits) obtained about one random variable, through the other random variable.[67] The application of mutual information is illustrated in the article.[68] The MI is useful for feature selection in multilabel classification and the multivariate MI for multilabel classification was proposed by the article presented.[69,70] The variant of mutual information-based multi-label text classification using interaction information is proposed by the reported article.[71]

The mutual information of two variables x and y are defined as following:[67]

$$I(X;Y) = \sum_{y\epsilon Y}\sum_{x\epsilon X} p(x,y)log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (14)$$

3.11 Odds Ratio (OR)

In statistics, the Odds Ratio (OR) is one of main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population.[72-74] A statistical learning model for text classification uses odd ration as the feature selection for the SVM.[75] A variation Multi-class odd ratio for feature selection in text classification has been used along with Naïve Bayes by the article presented.[74]

$$OR(t,c) = \frac{\log(P(t|c_i)[1 - P(t|\bar{c}_i)])}{[1 - P(t|c_i)]P(t|\bar{c}_i)} \quad (15)$$

3.12 Genetic Algorithm

Genetic Algorithm (GA) has been widely used for feature selection.[76] Genetic algorithm is a heuristics based algorithm that mimics the process of natural evolution by inheritance, mutation, selection and crossover.

In the reported article have used a parallel implementation of Genetic Algorithm to simultaneously examine and select large sets of candidate features.[77] They have used PGAPack software, which has a single population GA algorithm implemented in parallel master/slave configuration. The master node is responsible for storing the initial population and applying the GA operations such as selection, mutation and crossover. The slave nodes evaluate the fitness function for every feature. The authors have applied the parallel GA to nine different medical datasets and showed good results. A tool for parallel GA based feature selection has also been developed and is available at http://www.cbrc.kaust.edu.sa/dwfs.

A hybrid of GA and Particle Swarm Optimization (PSO) has been used for feature selection with SVM for pixel discrimination in image processing by the presented article.[78] PSO can remember the past iterations, but in GA if a chromosome is discarded, it is last forever. However premature convergence is a problem with PSO. These both algorithms have been combined by integrating the selection, mutation and crossover of GA and standard velocity and update rules of PSO. The hybrid algorithm was applied on the Indian Pines dataset and the RGB Toronto Roads data set. The hybrid algorithm terminates automatically when the average value of the swarm is less than the predefined threshold value. The advantage of this method is there is no need to set the number of features required before starting the iterations.

In the presented article have proposed a guided hybrid genetic algorithm for cost selection which has been tailored to reduce the number of cost function evaluations for feature selection.[79] Random forests are used as the guide. The guide suggests the features that may be removed and the most suitable feature set.

A combination of GA and Support Vector Regression (SVR) has been used for effective stock selection, which can be used in stock ranking and building an investor portfolio.[80] In this GA has been used to get the best subset of input variables to be provided as input for SVR.

Hybrid Genetic Algorithm with Neural Networks (HGA-NN) has been proposed by the presented article for credit risk assessment.[81] Here they have split the feature selection into two phases. Initially the feature space is reduced using means filter techniques such as Gini index, information gain, gain ratio and correlation. The features selected by the filter method are then passed on to GA. Also before the start of each iteration in GA, the initial population, mutation and crossover type can be changed and controlled.

Genetic Algorithm has been used for both feature selection and instance selection by the presented article in two separate steps.[82] GA with SVM has been used for feature selection and classification for automatic identification of diabetic retinopathy.[83]

Feature selection using Forest Optimization Algorithm (FOA) has been proposed by the reported article for selection of more informative features.[84] Generally FOA is used for continuous space problem, it has been adopted for discrete space feature selection by resetting the age of the best tree to zero.

4. Conclusion

Generating artificial patterns increases the size of training data and improves the classification performance in case of high dimension low sample size data. Alternatively, dimensionality reduction approaches help in finding the most informative or relevant features for classification. Random Projection is computationally efficient method. PCA is a widely used technique due to its simplicity and intuitiveness. But PCA is computationally expensive when compared to Random Projection. Moreover, PCA fails in situation where the size of the data set is very much smaller than the dimensionality. In such cases, Non –negative matrix factorization is more suitable. But it works only on non-negative data sets. Moreover, there exists no guarantee that NNMF converges to global minimum and so performance may not be as good as expected. LSA depends on SVD which is computationally expensive. But it is popularly used in text classification and information retrieval as it can handle synonymy problems to some extent. The Chi square is very easy to compute and it is very suitable with data that has been measured nominal scale. All the features measured must be independent and Chi square does not give much information about the strength relationship among the features. The Odds ratio and Correlation coefficient helps to quantify the strength of the relationship among participants. The Gini index is a probability measure to get the significance of features. The computation is complex when compare to other techniques. GA's major drawback is that it cannot remember the past iterations. If a feature is discarded in one of the iterations it is lost for all the future iterations. This drawback has been overcome by combining it with PSO which could remember the past iterations. One another draw-

back of GA is it takes a lot of iterations for the algorithm to converge. Random forest algorithm has been used as guide to GA to identify the features to be removed for that iteration. This is aimed at reducing the cost factor of GA by reducing the number of iterations. Parallel GA has been shown to have good computational efficiency. This parallelization of GA can be extended to big data framework and can be used for extremely large datasets.

In this paper we presented two different approaches used in high dimensional data classification. Although synthetic pattern generation is found to be an effective approach its usage in various fields such as text classification, categorical data classification could be further explored. On the other hand, dimensionality reduction methods reduce the complexity of data structure and provide a more understandable depiction of the same information. This field is active and many more novel dimensionality methods are being studied with different types of high dimensional data such as sentiment data, twitter data, social media data, hyper spectral image data, and genomic data and so on. The future research must be focussed on dimensionality reduction methods that have low time complexity with high scalability.

5. References

1. Suganthi J, Malathi V. Fuzzy based Feature Selection Scheme through Transductive SVM Technique for Cancer Pattern Classification and Prediction. Indian Journal of Science and Technology. 2016 Apr; 9(16):1–7. Crossref
2. Venkateswaran K, Shree TS, Kousika N, Kasthuri N. Performance Analysis of GA and PSO based Feature Selection Techniques for Improving Classification Accuracy in Remote Sensing Images. Indian Journal of Science and Technology. 2016 Apr; 9(16):1–7. Crossref
3. Sang HV, Nam NH, Nhan ND. A Novel Credit Scoring Prediction Model based on Feature Selection Approach and Parallel Random Forest. Indian Journal of Science and Technology. 2016 May; 9(20):1–6.
4. Jeevanandam, Jotheeswaran, Koteeswaran S. Feature Selection using Random Forest Method for Sentiment Analysis. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–7.
5. Samet H. Foundations of multidimensional and metric data structures. Morgan Kaufmann; 2006 Aug.
6. Hamamoto Y, Uchimura S, Tomita S. A bootstrap technique for nearest neighbor classifier design. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997 Jan; 19(1):73–9. Crossref
7. Jain AK, Chandrasekaran B. Dimensionality and Sample Size Considerations in Pattern Recognition Practice. Handbook of Statistics. PR Krishnaiah and LN Kanal editions; 1991, pp.1-13.
8. Viswanath P, Murty N, Bhatnagar S. Overlap pattern synthesis with an efficient nearest neighbor classifier. Pattern Recognition. 2005 Aug 31; 38(8):1187–95. Crossref
9. Saradhi W, Murty MN. Bootstrapping for efficient handwritten digit recognition. Pattern recognition. 2001 May 31; 34(5):1047–56. Crossref
10. Seetha H, Saravanan R, Murty MN. Pattern Synthesis Using Multiple Kernel Learning for Efficient SVM Classification. Cybernetics and Information Technologies. 2012 Dec 1; 12(4):77–94. Crossref
11. Seetha H, Murty MN, Saravanan R. A Note on the Effect of Bootstrapping and Clustering on the Generalization Performance. International Journal of Information Processing. 2011; 5(4):19–34.
12. Ha TM, Bunke H. Off-line, handwritten numeral recognition by perturbation method. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1997 May; 19(5):535–9. Crossref
13. Leung CH, Cheung YS, Chan KP. Distortion Model for Chinese Character Generation. In Proc IEEE International Conference on Systems Man and Cybernetics; 1985. p. 38–41.
14. Leung KC, Leung CH. Recognition of handwritten Chinese characters by combining regularization Fisher's discriminant and distorted sample generation. 10th International Conference on Document Analysis and Recognition; 2009 Jul 26. p. 1026–30.
15. Mori M, Suzuki A, Shio A, Ohtsuka S, Schomaker LR, Vuurpijl LG. Generating new samples from handwritten numerals based on point correspondence. In Proc 7th International Workshop on Frontiers in Handwriting Recognition; 2000 Sep. p. 281–90.
16. Velek O, Nakagawa M, Liu CL. Vector-to-Image Transformation of Character patterns for On-line and Off-line Recognition. International Journal of Computer Processing of Oriental Languages. 2002 Jun; 15(02):187–209. Crossref
17. Varga T, Bunke H. Effects of training set expansion in handwriting recognition using synthetic data. In Proc 11th Conference of the International Graphonomics Society; 2003 Nov. p. 200–3.
18. Varga T, Bunke H. Generation of Synthetic Training Data for an HMM-based Handwriting Recognition System. ICDAR. 2003 Aug; 3(3):618–22. Crossref
19. Chen B, Zhu B, Nakagawa M. Training of an on-line handwritten Japanese character recognizer by artificial patterns. Pattern Recognition Letters. 2014 Jan 1; 35:178–85. Crossref

20. Schuller B, Burkhardt F. Learning with synthesized speech for automatic emotion recognition. In 2010 IEEE International Conference on Acoustics Speech and Signal Processing; 2010 Mar 14. p. 5150–3. Crossref
21. Barngrover C, Kastner R, Belongie S. Semisynthetic versus real-world sonar training data for the classification of mine-like objects. IEEE Journal of Oceanic Engineering. 2015 Jan; 40(1):48–56. Crossref
22. Sezer EA, Nefeslioglu HA, Gokceoglu C. An assessment on producing synthetic samples by fuzzy C-means for limited number of data in prediction models. Applied Soft Computing. 2014 Nov 30; 24:126–34. Crossref
23. Rozantsev A, Lepetit V, Fua P. On rendering synthetic images for training an object detector. Computer Vision and Image Understanding. 2015 Aug 31; 137:24–37. Crossref
24. Wonders M, Ghassemlooy Z, Hossain MA. Training with synthesised data for disaggregated event classification at the water meter. Expert Systems with Applications. 2016 Jan 31; 43:15–22. Crossref
25. Fan J, Fan Y, Wu Y. High-dimensional classification. High-dimensional Data Analysis; 2011. p. 3–7. PMCid:PMC3025826
26. Guyon I, Elisseeff A. An introduction to feature extraction. Feature extraction. Springer; Berlin Heidelberg. 2006. p. 1–25. Crossref Crossref
27. Maldonado S, Weber R. A wrapper method for feature selection using support vector machines. Information Sciences. 2009 Jun 13; 179(13):2208–17. Crossref
28. Amarnath B, Appavu S, Balamurugan. Metaheuristic Approach for Efficient Feature Selection: A Data Classification Perspective. Indian Journal of Science and Technology. 2016 Jan; 9(4):1–6. Crossref
29. Bikku T, Rao NS, Akepogu AR. Hadoop based Feature Selection and Decision Making Models on Big Data. Indian Journal of Science and Technology. 2016 Mar; 9(10):1–6. Crossref
30. Fodor, Imola K. A survey of dimension reduction techniques; 2002.
31. Martin, Levine D. Feature extraction: A survey. In Proceedings of the IEEE. 1969; 57(8):13–91.
32. Chandrashekar, Girish, Ferat S. A survey on feature selection methods. Computers & Electrical Engineering. 2014; 40(1):16–28. Crossref
33. Kumar, Vipin, Minz S. Feature Selection. SmartCR. 2014; 4(3):211–29.
34. Achlioptas D. Database-friendly random projections. In Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems; 2001 May 1. p. 274–81. Crossref
35. Bingham E, Mannila H. Random projection in dimensionality reduction: Applications to image and text data. In Proceedings of the 7th ACM SIGKDD International conference on Knowledge discovery and data mining; 2001 Aug 26. p. 245–50. Crossref
36. Dasgupta S, Gupta A. An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures & Algorithms. 2003 Jan 1; 22(1):60–5. Crossref
37. Paul B, Athithan G, Murty MN. Speeding up AdaBoost classifier with random projection. In Advances in Pattern Recognition. IEEE Seventh International Conference (ICAPR); 2009 Feb 4. p. 251–4. Crossref
38. 38. Duda RO, Hart PE, Stork DG. Pattern classification. John Wiley & Sons; 2012 Nov 9.
39. 39. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999 Oct 21; 401(6755):788–91. Crossref PMid:10548103
40. Guillamet D, Vitria JSB. Introducing a weighted non-negative matrix factorization for image classification. Pattern Recognition Letters. 2003; 24(14):2447–54. Crossref
41. Shahnaz F, Berry MW, Pauca VP, Plemmons RJ. Document clustering using nonnegative matrix factorization. Information Processing & Management. 2006 Mar 31; 42(2):373–86. Crossref
42. Liu W, Yuan K, Ye D. Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis. Journal of biomedical informatics. 2008 Aug 31; 41(4):602–6. Crossref PMid:18234564
43. Zhu Z, Guo YF, Zhu X, Xue X. Normalized dimensionality reduction using nonnegative matrix factorization. Neurocomputing. 2010 Jun 30; 73(10):1783–93. Crossref
44. Forman G. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research. 2003 Mar; 3:1289–305.
45. Moh'd AMA. Chi square feature extraction based SVMs Arabic language text categorization system. Journal of Computer Science. 2007; 3(6):430–5. Crossref
46. Chi-Squared Distribution [Internet]. 2016 [cited 2016 Sep 2]. Available from: Crossref.
47. Chi-Square Goodness of Fit Test [Internet]. 2016 [cited 2016 Sep 2]. Available from: Crossref.
48. Forman G. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research; 2003 Mar. p.1289–305.
49. Aggarwal CC, Zhai C. Mining text data. Springer Science & Business Media; 2012 Feb 3.
50. Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter. 2004 Jun 1; 6(1):80–9. Crossref
51. Mitchell TM. Machine Learning. WBA; 1997.
52. Gini Coefficient [Internet]. 2016 [cited 2016 Sep 2]. Available from: Crossref

53. Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. ICML. 1997 Jul 8; 97:412–20.
54. Singh SR, Murthy HA, Gonsalves TA. Feature Selection for Text Classification Based on Gini Coefficient of Inequality. FSDM. 2010; 10:76–85.
55. Park H, Kwon S, Kwon HC. Complete gini-index text (git) feature-selection algorithm for text classification. IEEE 2010 2nd International Conference on Software Engineering and Data Mining (SEDM); 2010 Jun 23. p. 366–71.
56. Glossary of Important Assessment and Measurement.
57. Billsus D, Pazzani MJ. Learning Collaborative Information Filters. ICML. 1998 Jul 24; 98:46–54.
58. Guyon I, Elisseeff A. An introduction to feature extraction. Springer Berlin Heidelberg. Feature extraction; 2006. p. 1–25.
59. Hsu HH, Hsieh CW. Feature selection via correlation coefficient clustering. Journal of Software. 2010 Jan 12; 5(12):1371–7. Crossref
60. Latent Semantic Analysis.
61. Landauer TK, Dumais ST. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review. 1997 Apr; 104(2):2–11. Crossref
62. Landauer TK, Foltz PW, Laham D. An introduction to latent semantic analysis. Discourse processes. 1998 Jan 1; 25(2-3):259–84. Crossref
63. Dumais ST. Latent Semantic Indexing (LSI): TREC-3 Report. Nist Special Publication SP; 1995.
64. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of the American society for information science. 1990 Sep 1; 41(6):1–34. Crossref
65. Shima K, Todoriki M, Suzuki A. SVM-based feature selection of latent semantic features. Pattern Recognition Letters. 2004 Jul 2; 25(9):1051–7. Crossref
66. Hofmann T. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval; 1999 Aug 1. p. 50–7. Crossref
67. Mutual Information [Internet]. 2016 [2016 Sep 2]. Available from: Crossref
68. Basu T, Murthy CA. Effective text classification by a supervised feature selection approach. 2012 IEEE 12th International Conference on Data Mining Workshops; 2012 Dec 10. p. 918–25. Crossref
69. Doquire G, Verleysen M. Mutual information-based feature selection for multilabel classification. Neurocomputing. 2013 Dec 25; 122:148–55. Crossref
70. Lee J, Kim DW. Feature selection for multi-label classification using multivariate mutual information. Pattern Recognition Letters. 2013 Feb 1; 34(3):349–57. Crossref
71. Lee J, Kim DW. Mutual information-based multi-label feature selection using interaction information. Expert Systems with Applications. 2015 Mar 31; 42(4):2013–25. Crossref
72. Mosteller F. Association and estimation in contingency tables. Journal of the American Statistical Association. 1968 Mar 1; 63(321):1–28. Crossref Crossref
73. Edwards AWF. The Measure of Association in a 2 × 2 Table. Journal of the Royal Statistical Society. Series A (General). 1963; 126(1):109–14. Crossref
74. Chen J, Huang H, Tian S, Qu Y. Feature selection for text classification with Naïve Bayes. Expert Systems with Applications. 2009 Apr 30; 36(3):5432–5. Crossref
75. Joachims T. A statistical learning learning model of text classification for support vector machines. In Proceedings of the 24th annual international ACM SIGIR Conference on Research and development in information retrieval; 2001 Sep 1. p. 128–36. Crossref
76. Abbas A, Wu Z, Siddiqui IF, Lee S. An Approach for Optimized Feature Selection in Software Product Lines using Union-Find and Genetic Algorithms. Indian Journal of Science and Technology. 2016 May; 9(17):1–8. Crossref
77. Soufan O, Kleftogiannis D, Kalnis P, Bajic VB. DWFS: A wrapper feature selection tool based on a parallel genetic algorithm. PloS one. 2015 Feb 26; 10(2). Crossref
78. Ghamisi P, Benediktsson JA. Feature selection based on hybridization of genetic algorithm and particle swarm optimization. IEEE Geoscience and Remote Sensing Letters. 2015 Feb; 12(2):309–13. Crossref
79. Jung M, Zscheischler J. A guided hybrid genetic algorithm for feature selection with expensive cost functions. Procedia Computer Science. 2013 Dec 31; 18:2337–46. Crossref
80. Huang CF. A hybrid stock selection model using genetic algorithms and support vector regression. Applied Soft Computing. 2012 Feb 29; 12(2):807–18. Crossref
81. Oreski S, Oreski G. Genetic algorithm-based heuristic for feature selection in credit risk assessment. Expert systems with applications. 2014 Mar 31; 41(4):2052–64. Crossref
82. Tsai CF, Eberle W, Chu CY. Genetic algorithms in feature and instance selection. Knowledge-based Systems. 2013 Feb 28; 39:240–7. Crossref
83. Welikala RA, Fraz MM, Dehmeshki J, Hoppe A, Tah V, Mann S, Williamson TH, Barman SA. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. Computerized Medical Imaging and Graphics. 2015 Jul 31; 43:64–77. Crossref PMid:25841182
84. Ghaemi M, Feizi-Derakhshi MR. Feature selection using Forest Optimization Algorithm. Pattern Recognition. 2016 Dec 31; 60:121–9. Crossref