# FEATURE SELECTION: A REVIEW

*P. Mohana Chelvan*[1] *Dr.K. Perumal*[2]

**ABSTRACT**

In everyday life, we use a large number of web applications which result in the storage of a huge volume of data both column-wise and row-wise. Along with the explosion of web applications, the emergence of IoT resulted in the creation of a mammoth volume of dynamic data. This may result in the inevitable use of dimensionality-reduction techniques which is a vital part of the preprocessing of data for data mining applications. The mostly used dimensionality-lesseningmethods are feature selection and feature extraction, of which feature selection is superior in functionality. The heftiness of feature selection algorithms for a slight amendment of data is called feature selection stability. Selection stability is deliberated as one of the significant criteria of feature selection algorithms along with data utility. The significance of feature selection stability is very much considered for dynamic data like incremental micro data as well as privacy-preserved micro data, as feature selection stability is mostly data reliant.

***Keywords:*** *data mining, stability measures, privacy preservation, feature selection, feature selection stability.*

## 1. INTRODUCTION

Every organization needs data mining to get tactical statistics from the colossal volume of archived operational data for their day to day activities. Today's micro data is mostly high dimensional and hence dimensionality reduction technique is an essential preprocessing step for data mining applications. There will be two types of dimensionality-lesseningpractices called feature selection and feature extraction. Feature selection is superior as it selects a small subset of micro data for improved accuracy and has reduced processing time and better model inter-pretability along with less storage space. In the case of feature extraction technique, the extracted features are projected into a new space and hence there is a relation of the extracted features in the new space and the space of the original micro-data, and hence feature selection is preferable to feature extraction.

[1]Associate Professor, Department Computer Science
Karpagam Academy of Higher Education, Coimbatore. India.

[2]Professor, Department of Computer Applications
Madurai Kamaraj University, Madurai. India.

## 2. FEATURE SELECTION

Higher dimensionality of micro-data leads to soaring of noisy and irrelevant data with the relevant data which are required for predicting accuracy. Feature selection eliminates the noisy, redundant, and irrelevant data which are not required for predicting accuracy. The statistically relevant traits which are joined with each other by the characteristics are chosen by feature selection [1].

Feature selection will be of three types called Filter, Wrapper, and Embedded method. The Filter model does not use any algorithm and it uses the statistical properties of the trait values and hence it is faster than other methods. However, this model is not as accurate as the Wrapper model. The Wrapper model uses a classifier for iteratively selecting the traits and hence it is more accurate than the Filter model. However, this model is computationally more expensive depending on the classifier algorithm. The Embedded model makes use of the advantages of both the models.

## 3. FEATURE SELECTION PROCEDURE

The feature selection procedure begins with search direction and strategy. The search direction may be forward search, backward elimination or random search [2, 3]. The forward search starts with an empty set of traits and searches forward iteratively. The backward elimination starts with a full set of traits and then eliminates irrelevant traits iteratively. The random search starts at a random point and

selects or eliminate straits iteratively depending on the importance of traits.

Following the stage of search direction and strategy, using evaluation criteria, the best traits are selected [4]. Grounded on the evaluation criteria, feature selection algorithms are classified as Filter, Wrapper, Embedded and Hybrid methods.

The next stage is the stopping criteria. The feature selection process is stopped after an optimum number of traits are selected with a low computational cost while avoiding over fitting problems. The stopping criteria may be the optimum number of chosen features, iterations and percentage of advancement in successive iterations, or based on evolution function.

For validating the results, various feature set validation measures are used such as Confusion matrix, Cross-validation, etc. The validation method most commonly used is Cross-validation (CV). The Cross-validation method gives an unbiased error estimate and it is the main advantage for the method. For the appraisal of the classifier Confusion Matrix is created. The most commonly used clustering and classification measures are as follows:

| Clustering Measures | Classification Measures |
|---|---|
| Jaccard index | ROC (Receiver Operating Characteristic) Curve |
| Dunn Index | TP Rate/ Recall / Sensitivity |
| Dice index | Precision |
| Davies-Bouldin Index | Error Rate |
| Fowlkes-Mallows index | F-Score / F-Measure |
| F-Measure | Specificity |

## 4. FEATURE SELECTION STABILITY

The heftiness of feature selection algorithms for a slight variation of sample data is called selection stability. If the researcher gets different results in different cycles of experiments, he does not rely on his experimental conclusions [5]. The problem will be acute for high dimensional data with small samples like researches in DNA samples. If each iteration of the experiment of the feature selection algorithm selects a different subset of traits, the researcher may be very much confused about his findings. Hence, selection stability is now well-thought-out as one of the significant criteria of feature selection algorithms along with data utility. Selection stability has currently emerged as a sizzling topic of research.

Researchers have considered earlier that selection stability depends on feature selection algorithms. Now they realize that the causal physiognomies of the micro-data impact the selection stability. The sample size, the number of selected traits, the dimensionality of the micro-data, and data variance among different folds of the micro-data are the factors that disturb selection stability. Selection stability increases with the upsurge of sample size. Selection stability improves up to the most favorable number of selected traits and then declines. Selection stability is adversely interrelated with the dimensionality of the micro-data. Data variance affects selection stability. Hence, selection stability is generally data-centric but not entirely algorithm-independent [6 - 11].

Selection stability is measured by comparing the selected subsets in the consequent iteration of feature selection algorithms. Selection stability measures will be of three types, i.e., stability by index, stability by rank, and stability by weight. Stability by index measure produces a subset of traits by comparing the trait index values. Examples of stability by index methods are Average Normal Hamming Distance ANHD, Percentage of Overlapping Gene POG, Kuncheva Index KI, Tanimoto Distance, Symmetrical Uncertainty SU, Jaccard's Index, Consistency Measure, and Dice's Coefficient. The stability by rank measure producesa ranked list of all the traits. The examples of stability by the rank measure are Canberra Distance CD and Spearman's Rank Correlation Coefficient SRCC. The stability by weight measure produces a list of traits based on their weights. The example of stability by weight measure is Pearson's Correlation Coefficient PCC.

## 5. FEATURE SELECTION STABILITY VERSUS PRIVACY PRESERVING DATA MINING

Selection stability is more prominent in dynamic micro-data like incremental micro-data and privacy-preserved micro-data, as the data variance mostly affects

selection stability. For safeguarding the privacy of the personages, the micro-data is modified. Sensitive data along with quasi-identifiers are altered to defend the privacy of the personages. Even with strong background knowledge, the intruder should not be able to identify the record of a person in micro-data. This privacy-conserving perturbation affects the selection stability, as it is generally data-centric. The modification in the statistical properties' numeric attribute values of micro-data should be at the slightest for enhanced selection stability. Hence, privacy-preserving perturbation should be lowest for improved selection stability along with data utility [12].

## 6. FEATURE SELECTION STABILITY MEASURES VERSUS FEATURE SELECTION ALGORITHMS

`Some feature selection algorithms produce ranked lists of traits. The list contains all the traits and their ranks. Some feature selection algorithms produce a subset of traits. Selection stability improves up to the ideal number of traits and then declines. Rank-based and weight-based selection stability measures result in a full set of features, and index-based stability measures result in an optimum number of subsets of features. Also, in the experimental sample, the different folds of the samples produce different results for the same feature selection algorithm, as the characteristics of the sample differ among the different folds of the same sample. Hence, the characteristics of the sample in different folds should be considered for selecting an appropriate feature selection algorithm. However, selecting the appropriate selection stability measure for the feature selection algorithm is considered challenging [13].

## 7. CONCLUSION

This paper explains the feature selection techniques along with selection stability. The data perspective nature of selection stability and the influence of privacy-preservation on selection stability is explained. The selection of appropriate selection stability measure for the feature selection algorithms along with the sample characteristics is also explained.

**Reference**


1. Gheyas, I. A., L. S. Smith. Feature Subset Selection in Large Dimensionality Domains. – Pattern Recognit., Vol. 43, No 1, pp. 5-13, January 2010.

2. Gutkin, M., R. Shamir, G. Dror. SlimPLS: A Method for Feature Selection in Gene Expression-Based Disease Classification. – PLoS One, Vol. 4, No 7, p. e6416, July 2009.

3. Ang, J. C., A. Mirzal, H. Haron, H. N. A. Hamed. Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection. – IEEE/ACM Trans. Comput. Biol. Bioinforma., Vol. 13, No 5, pp. 971-989,September 2016.

4. Dash, M., H. Liu. Feature Selection for Classification. – Intell. Data Anal., Vol. 1, No 1-4, pp. 131-156, January 1997.

5. Alexandros Kalousis, Julien Prados, and Melanie Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowledge and Information Systems, 12(1):95–116, http://link.springer.com/article/10.1007/s10115-006-0040-8, May 2007.

6. Salem Alelyani, Huan Liu., The Effect of the Characteristics of the Dataset on the Selection Stability, IEEE DOI 10.1109/International Conference on Tools with Artificial Xiniun Intelligence, 2011.167, 1082-3409/11, http://ieeexplore.ieee.org/document/ 6103458, 2011.

7. Salem Alelyani, Zheng Zhao, Huan Liu., A Dilemma in Assessing Stability of Feature Selection Algorithms, IEEE DOI 10.1109/ International Conference on High Performance Computing and Communications, 2011.99, 978-0-7695-4538-7/11, http:// ieeexplore.ieee.org/document/6063062, 2011.

8. Salem Alelyani, On feature selection stability: a data perspective, Doctoral Dissertation, Arizona State University, AZ, USA, ISBN: 978-1-303-02654-6, ACM Digital Library, 2013.

9. Barbara Pes, Feature Selection for High-Dimensional Data: The Issue of Stability, Proceedings of the 26th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE 2017), June 21–23, 2017.

10. Jundong Li, Huan Liu, Challenges of Feature Selection for Big Data Analytics, Special Issue on Big Data, IEEE Intelligent Systems, eprint arXiv:1611.01875, 2017.

11. Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. Feature Selection: A Data Perspective. ACM Comput. Surv. 50, 6, Article 94, 45 pages. DOI: https://doi.org/10.1145/3136625, 2018.

12. Mohana Chelvan P, Perumal K, "Correlation between Privacy Preserving Data Publishing and Feature Selection Stability", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), ISSN: 2278-6856, Volume 4, Issue 5(2), pp. 001-003, 2015.

13. Mohana Chelvan P, Perumal K, "A Study on Selection Stability Measures for Various Feature Selection Algorithms", IEEE International Conference on Computational Intelligence and Computing Research, IEEE Xplore Digital Library, E-ISSN: 2473 – 943X, SCOPUS Indexed, DOI: 10.1109/ICCIC.2016.7919544, 2016