A review of Feature Selection Stability

Mohana Chelvan P[1], Dr. Perumal K[2]
[1]Department Computer Science, Karpagam Academy of Higher Education (KAHE), Coimbatore, India, Email:pmohanselvan@rediff.com
[2]Department of Computer Applications, Madurai Kamaraj University, Madurai, India, Email:perumalmala@gmail.com

Abstract

Business establishments for getting tactical information on their day to day business activities use data mining as an indispensable technique. For data mining applications, feature selection is crucial as these days' microdata is mostly high-dimensional. The sturdiness of feature selection algorithms in a succeeding rehearsal of feature selection experiments even for small perturbation of microdata is called feature selection stability. Selection stability considered one of the indispensable criteria for feature selection algorithms by the researchers. This paper describes selection stability, the data perspective nature of selection stability along with selection stability measures. The selection stability is related to privacy conserving data mining as the selection stability is generally data reliant which is also explained in this paper.

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

1. Introduction

In almost all business establishments use data mining is used for attaining tactical statistics for competitive advantage. The selection of a small subset of pertinent traits end eliminating irrelevant traits for improved accuracy, faster processing speed and better model interpretability is called feature selection. Because of the high dimensional microdata of today's business world, feature selection technique is an indispensable part of data mining methods as the small subset of relevant traits of the experimental microdata works much enhanced than a full set of traits of the microdata.

The heftiness of feature selection algorithms for trivial discomposure of the microdata is called selection stability. The researcher gets confused about their research conclusion if a subsequent iteration of feature selection experiments gives a different subset of traits. If it is so, the researcher doesn't rely on his research findings. Recently, selection stability is has emerged as a blistering subject of research. Now, selection stability is well thought-out as one of the essential criteria of feature selection algorithms along with accuracy [1–4]. Selection stability is more significant for research experiments with extremely high dimensional microdata with the least number of samples like the samples in the field of Bioinformatics.

2. Data perspective nature of Feature Selection Stability

ISSN: 2005-4297 IJCA

Researchers on selection stability rely on that the selection stability is algorithm dependent. The scholars did not consider the underlying characteristics of the microdata and their consequences on selection stability. However, recently, the data perception nature of selection stability is considered by the researchers. Selection stability generally relies on the physiognomies of the microdata but it is not entirely algorithmic sovereign.

The aspects that disturb the selection stability include the dimensionality of the microdata, the number of chosen features, and distribution of data of microdata among various folds along with sample size. Selection stability is increased up to the optimal number of chosen traits and then declines. Selection stability is positively correlated with sample size and negatively correlated with dimensionality. Data variance i.e. change in the characteristics of the microdata affects selection stability [5-10].

3. Techniques for improving Feature Selection Stability

10-fold cross-validation is utilized to improve stability against input data perturbation. The Cumulative Ranking Score (CRS) of the considerable number of traits is determined in every cycle that is utilized to ascertain the significance of each trait in making a class difference. The ranking of traits from various subsets is joined in this parameter. The precise set of genes that are answerable for the malady as the traits with a high aggregate ranking are a vigorous [11].

The connection between the traits is ascertained by the feature relatedness. Adaptable parameterization is exploited by the supervised factor investigation model and reliance on existing latent factors shown in the model. Catch of low dimensional space is based on dependence. The feeble correlation was applied to the decorrelated data after the alteration or latent effect. Better performance is prompted by the decorrelation of factor change. Fewer traits within the sight of the balanced data factor were chosen by LASSO. Factor alteration improves the stability and blunder of the forecast of chosen variables which helps to hinder the impact of heterogeneity. [12, 13].

Various weights are assigned to every single sample grounded on the impact of the sample on the importance of the trait is the guideline underlying the system of sample weighting. To gauge the impact of each sample, the local profile of trait importance is utilized [14]. Mutual Information Quotient (MIQ) and Mutual Information Difference (MID) are the two distinct criteria of most extreme Relevance Minimum Redundancy (MR). The essential for a stable selection of features is by adjusting the trade-off between least redundancy and greatest relevance. Control the stability of the MRMR is helped by this weighting parameter. A trait may have various weights for redundancy and relevance which may result in the determination of the traits [15].

The technique is to choose a vigorous set of parameters that is used to upgrade the present optimization procedure. To tackle the instability issue, the technique for choosing a trait is utilized during the procedure of optimization of the parameter. A cross-entropy blunder function with a logistic regression model and a squared error along with a nonlinear regression model is examined by the author [16].

The method that grosses the mediocre of numerous learning techniques works after arbitrary subsamples of the original microdata which is called the bagging method is used by the ensemble learning systems. Ensemble feature selection is another procedure to improve selection stability [17].

The assembling of the profoundly correlated traits present in high-dimensional microdata is the method behind the group feature selection strategy which is impervious to discrepancies in training samples. If this gathering is viewed as a solitary element, the stability of the selection procedure can be enhanced [14].

The gathering of traits utilizing either a density estimate or a cluster analysis is perceived by data-driven group generation [18]. Bunches are framed constructed on information confined in input data, rather than depending on biology domain knowledge. When various gatherings are framed by correlated traits, group-lasso is pertinent [19].

4. Feature Selection Stability versus Privacy Preserving Data Mining

Privacy conserving data mining imposes alterations of the attributes of the microdata for safeguarding the privacy of the personages and simultaneously the perturbation of the microdata should not affect the data utility. High data utility and extreme privacy preservation is the aim of privacy conserving data mining. The alteration of the microdata by privacy conserving data mining affects selection stability as selection stability is generally data reliant. Amendment in the statistical properties of numerical traits of microdata ought to be least for improved selection stability. There is a negative correlation between selection stability and data utility. So, there is a tradeoff among privacy conservation, modification in the statistical belongings of the numerical traits of the microdata, selection stability and data utility [20].

5. Feature Selection Stability Measures

Estimation of selection stability may be merely by comparing subsequent result aftermaths. Hence, if the resemblance is more prominent consequently the stability has higher values. The three necessities deemed necessary for the estimation of stability [21] are in keeping with the following:

- Monotonicity: If the overlap amongst the picked subsets is massive, the aftermath should be of horrendous stability.
- Limits: The limiting values of the magnitude of every stability valuation method ought to be among the range of [-1, 1] or [0, 1]. Similar to the number of picked traits k or the dimensionality of the microdata m, these bounds are independent of any microdata factor. Once the sets are delicate and moved toward becoming the fuzziest or unwavering, these cut-off points have to be at any degree.
- Correction for chance: Because of the high-dimensional picked subset, the measure should have a diligent refinement that encourages an intersection trendy unintentionally. The more significant the probability for a larger

intersection amongst the subsets is, the larger the cardinality of the picked subsets.

There are three types of stability measures called index-based stability measure, rank-based stability measure and weight-based stability measure. The index-based stability measure group of stability measures is based on the values of the index. In this classification, the selected traits have no particular order or corresponding relevance weight. The list of stability measures in this category are as follows:

- Average Normal Hamming Distance ANHD [22]
- Dice's Coefficient [23]
- Tanimoto Distance
- Jaccard's Index [2]
- Kuncheva Index KI [21]
- Percentage of Overlapping Gene POG [24]
- Consistency Measure [25]
- Symmetrical Uncertainty SU [26]

The rank-based stability measure category of stability measures is based on ranking vectors of trait importance. The examples of stability measures by the ranking are as follows:

- Spearman's Rank Correlation Coefficient SRCC [2]
- Canberra Distance CD [27]

The weight-based stability measure category of stability measures is based on the weight assigned to the traits. There is only one type of measure in this category as follows:

- Pearson's Correlation Coefficient PCC [2]

Among the stability measures, the index-based stability measure category results in a subset of traits that are fewer traits than the total number of traits. The other category of measures, i.e. the rank-based stability measure and weight-based stability measure will have resulted in a full set of traits.

The two subsets of traits retain a vast majority in the stability measures. With the hazard of overlap, the significant cardinality is that the list of elected traits is ardently unswerving. The intersection, which is inadvertent amid the picked subsets of traits leads to the disadvantage and it was defeated by Kuncheva Index KI, which is suggested in [21] to which includes an amendment term to stopover away from the impairment. The provisions acknowledged in [21], i.e., monotonicity, limitation, and chance correction are contracted by the only measure, that is, KI.

$$\mathrm{KI}\,(\mathcal{F}'_1, \mathcal{F}'_2) = \frac{\left|\mathcal{F}'_1 \cap \mathcal{F}'_2\right| .\ \mathrm{m} - \mathrm{k}^2}{\mathrm{k}\,(\,\mathrm{m} - \mathrm{k})} \qquad (1)$$

ISSN: 2005-4297 IJCA

The KI consequences encompass [-1, 1]. The cardinality of the crossing set is k once KI attains 1 which results in Ƒ’1 and Ƒ'2 are fuzzy. There is no crossing amongst the lists and k= m/2 results in somewhere the value of KI is -1. KI value is zero when appropriate for austere obscuring lists. The number of traits that have elected k will have affect stability in the category of the Jaccard Index along with several other measures. Once k grows higher and nearer to m, it may result in greater stability reverences. The overhauling of the term, that is, a correction term is designed in KI so that it ultimately glosses with an entreaty. Notwithstanding that, the hostile properties as a consequence of the correction term are not practiced by KI, which stretches negative weight to k.

6. Conclusion

Feature selection is indispensable for today’s high dimensional microdata. For evaluating feature selection algorithms, selection stability is considered as a significant criterion along with accuracy. The techniques for improving selection stability are evaluated. The data perception nature of selection stability addressed in this paper. The effects of privacy conserving data mining on selection stability are evaluated as the selection stability is data-centric. The paper comprehensively analyses the selection stability along with selection stability measures.

Reference

[1] Mark, A., Hall, Correlation-based Feature Selection for Machine Learning, Dept of Computer science, University of Waikato. http://www.cs.waikato.ac.nz/ mhall / thesis.pdf, 1998.

[2] 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.

[3] Zengyou He and Weichuan Yu, Stable feature selection for biomarker discovery, 2010.

[4] Kalousis, A., Prados, J., and Hilario, M., Stability of feature selection algorithms, page 8, Nov. 2005

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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

[10] 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, Jan. 2018

[11] Lahmiri, S., Shmuel, A., Detection of Parkinson's disease based on voice patterns ranking and optimized support vector machine. Biomed. Signal Process Control. 49, 427–433. https://doi.org/10.1016/j.bspc.2018.08.029, 2019.

[12] Grollemund, P.M., Abraham, C., Baragatti, M., Pudlo, P., Bayesian functional linear regression with sparse step functions. Bayesian Anal. 14, 111–135. https://doi.org/10.1214/18-BA1095, 2019.

[13] Ramondta, S., Ramírezb, A.S., Assessing the impact of the public nutrition information environment: adapting the cancer information overload scale to measure diet information overload. Patient Educ. Couns. 102, 37–42. https://doi.org/10.1016/j.pec.2018.07.020, 2019.

[14] Li, Y., Li, T., Liu, H., Recent advances in feature selection and its applications. Knowl. Inf. Syst. 53 (3), 551–577. https://doi.org/10.1007/s10115-017-1059-8, 2017.

[15] Gulgezen G, Cataltepe Z, Yu L. Stable and Accurate Feature Selection. In: Proc 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I. 2009:455-468. https://doi.org/10.1007/978-3-642-04180-8_47, 2009.

[16] Isachenko, R.V., Strijov, V.V., Quadratic programming optimization with feature selection for nonlinear models. Lobachevskii J. Math. 39 (9), 1179–1187. https:// doi.org /10.1134 /S199508021809010X, 2018.

[17] Diren, D.D., BoranIhsan, S., Selvi, H., Hatipoglu, T., Root cause detection with an ensemble machine learning approach in the multivariate manufacturing process. Industr. Eng. Big Data Era, 163–174. https://doi.org/10.1007/978-3-030-03317-0_14, 2019.

[18] Jeitziner, R., Carrière, M., Rougemont, J., Oudot, S., Hess, K., Brisken, C., Two-Tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz052, 2019.

[19] Jacob, L., Obozinski, G., Vert, J.P., Group lasso with overlap and graph lasso. In: Proc 26th international conference on machine learning. ACM, pp. 433–440, 2009.

[20] P. Mohana Chelvan, K. Perumal, "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.

[21] L. I. Kuncheva, A stability index for feature selection, In Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi Conference: artificial intelligence and applications, pages 390 - 395, Anaheim, CA, USA, ACTA Press, 2007.

[22] Kevin Dunne, Padraig Cunningham, and Francisco Azuaje, Solutions to instability problems with sequential wrapper-based approaches to feature selection, Technical Report TCD-CD-2002-28, Department of Computer Science, Trinity College, Dublin, Ireland, 2002.

[23] Lei Yu, Chris Ding, and Ste1ven Loscalzo, Stable feature selection via dense feature groups, In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 803–811, New York, NY, USA, ACM, 2008.

[24] M. Zhang, L. Zhang, J. Zou, C. Yao, H. Xiao, Q. Liu, J. Wang, D. Wang,C. Wang, and Z. Guo, Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes, Bioinformatics, 25(13):1662 - 1668, Jul 2009.

[25] P. Somol and J. Novovivcova, Evaluating the stability of feature selectors that optimize feature subset cardinality, 2010.

[26] L. Yu, C. Ding, and S. Loscalzo, Stable feature selection via dense feature groups, In KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 803 - 811, New York, NY, USA, 2008. ACM.

[27] G. Jurman, S. Merler, A. Barla, S. Paoli, A. Galea, and C. Furlanello, Algebraic stability indicators for ranked lists in molecular profiling Bioinformatics, 24(2): 258 - 264, Jan 2008.

AUTHORS BIOGRAPHY

Mr. P. Mohana Chelvan is currently working as an Assistant Professor in the Department of Computer Science at Karpagam Academy of Higher Education (KAHE), Coimbatore, India. His educational qualifications are MCA, NIELIT C Level (IT), MPhil. (CS) and UGC NET. He is currently a Ph.D. research scholar in computer science from Madurai Kamaraj University, Madurai, India in the area of Privacy Preserving Data Mining.

Dr. K. Perumal working as Professor in Department of Computer Applications at Madurai Kamaraj University, Madurai, India since 1990. He awarded his Ph.D. degree in computer science from Madurai Kamaraj University in the area of digital image processing. He has contributed more than 80 research papers in International Journals and Conferences. His research interest includes Data Mining, Big Data and Image Processing especially in Medical Image Processing.