Absolute Redundancy Analysis Based on Features Selection

Ginel Dorleon, Nathalie Bricon-Souf, Imen Megdiche, Olivier Teste

Absolute Redundancy Analysis Based on Features Selection

Ginel Dorleon
Ginel.Dorleon@irit.fr
Université Toulouse 3 Paul Sabatier
IRIT (CNRS/UMR5505)
Toulouse, France

Nathalie Bricon-Souf
Nathalie.Souf@irit.fr
Université Toulouse 3 Paul Sabatier
IRIT (CNRS/UMR5505)
Toulouse, France

Imen Megdiche
Imen.Megdiche@irit.fr
Institut National Universitaire Jean François Champolion
IRIT (CNRS/UMR5505)
Toulouse, France

Olivier Teste
Olivier.Teste@irit.fr
Université Toulouse 2 Jean Jaurès
IRIT (CNRS/UMR5505)
Toulouse, France

ABSTRACT

The goal of feature selection (FS) in machine learning is to find the best subset of features to create efficient models for a learning task. Different FS methods are then used to assess features relevancy. An efficient feature selection method should be able to select relevant and non-redundant features in order to improve learning performance and training efficiency on large data. However in the case of non-independents features, we saw existing features selection methods inappropriately remove redundancy which leads to performance loss.

We propose in this article a new criteria for feature redundancy analysis. Using our proposed criteria, we then design an efficient features redundancy analysis method to eliminate redundant features and optimize the performance of a classifier. We experimentally compare the efficiency and performance of our method against other existing methods which may remove redundant features. The results obtained show that our method is effective in maximizing performance while reducing redundancy.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Feature selection**.

KEYWORDS

Machine Learning, Features Selection, Redundancy Analysis

ACM Reference Format:
Ginel Dorleon, Nathalie Bricon-Souf, Imen Megdiche, and Olivier Teste. 2021. Absolute Redundancy Analysis Based on Features Selection. In *DSIT '21: 4th International Conference on Data Mining and Big Data (DMBD 2021), Shangai, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3478905.3479002

DSIT '21, July 23–25, 2021, Shangai, China

ACM ISBN 978-1-4503-9024-8/21/07…$15.00
https://doi.org/10.1145/3478905.3479002

1 INTRODUCTION

Nowadays, many real world applications [1, 5, 9, 11, 16] deal with so called high-dimensional data. However, high-dimensional data became a major issue in the field of machine learning [13] due to its size and the amount of resources required to process it. Learning performance is impacted by high-dimensional dataset [17]. Naturally, one may believes that more features we get, the more information we get from the features, but that is far from true because it becomes more difficult to extract meaningful conclusions from a dataset as the dimensionality of the data increases. Hence there is a need to resort to dimensionality reduction techniques in order to reduce the size of these data. Feature selection is one among the methods used for dimensionality reduction [17, 18]. The goal is therefore to obtain a meaningful subset of relevant and non redundant features that is vital for improving efficiency and reducing overfitting. Generally, it is easier to remove irrelevant features than finding redundant ones. Thus, the difficulty in selecting features now is finding the ones that are redundant. Existing works for features redundancy analysis such as RABFS [19], mRMR [10], FCBF [20] introduced approaches to reduce redundancy. Yet, our study shows that these methods remove inappropriately redundancy because they required user to set a threshold. We saw several problems with these methods based on the definition of a threshold:

- A feature will be redundant depending on the fixed threshold and the user's experience.
- The performance of the model will depend on the fixed threshold.

That being said, on the same dataset, users may find different redundant features. This is practically dangerous if there is protected features in the data that we are using. By protected features we mean features that carry special importance and are of priority when making relevant decisions [4]. Therefore, there is a need for more research on feature redundancy analysis. Our study here focuses on features redundancy analysis and proposes a new method for analyzing features redundancy without the need of defining a threshold.

The rest of this article is as follows: In section 2, we define features relevancy and features redundancy. Section 3 presents the proposed redundancy criterion as well as our redundancy analysis method. Experimental results are analyzed in section 4 and section 5 is our conclusion and perspectives.

2 FEATURES RELEVANCY AND REDUNDANCY

This section introduces feature redundancy and relevance based on the definition from the literature.

2.1 Features Relevancy

In general, features relevance is done by using common features selection methods such as Filter, Wrapper, Embedded or Hybrid [2, 6]. Other works have introduced notions of ranking the level of features relevance. According to the authors in [15], the relevance of a feature can be strong, weak or totally irrelevant. They introduced three categories that are defined below. A feature that has a strong relevance should be considered for the selected subset of features. One with a weak relevance feature is not really important but it may be used under certain conditions. A totally irrelevant feature is not necessary and should be removed.

From these definitions, the authors in [15] conclude that, for an optimal result, the subset must contain features with a strong and weak relevance only. There are many works on feature relevancy [2, 6] with different strategies. However, for feature redundancy, there are not that many effective methods. Thus, research method for redundancy analysis is always of considerable importance. Our objective is to analyze features redundancy and we propose a redundancy criterion in section 3.

2.2 Features Redundancy

Here we discuss redundancy in the sense of correlation with respect to the target variable. There is redundancy when there is a high correlation between features [8]. The authors in [19] believed that redundancy could be strong, moderate or weak. Based on a chosen threshold, the correlation between features will determine if the redundancy is strong, moderate or weak. When the correlation is high, the redundancy is believed to be high and such features are removed. When it is moderate, features are considered as partially redundant and they can be kept or removed. Low correlation means low redundancy. However, the problem with this approach is that the removal of a partially redundant feature based on a fixed threshold may be excessive and may also lead to the loss of useful information for the intended task. This also means that the redundancy is subjective and it varies between users and experiments. Hence we propose a new criterion to evaluate the redundancy between relevant features.

3 THE PROPOSED REDUNDANCY ANALYSIS METHOD

In this section, we present our proposed redundancy analysis method. We use the Symmetrical Uncertainty (SU) [7] as correlation measure to assess redundancy.

3.1 Correlation Measure

The correlation measure that we used is based on information gain [12] which for two variables X,Y is given by:

$$IG(X, Y) = H(X) - H(X|Y) \tag{1}$$

Using this measurement, a feature Y is considered to be more correlated to a feature X than a feature Z if and only if: $IG(X|Y) > IG(X|Z)$. Information gain uses the notion of entropy to measure the mutual dependence between two variables. The entropy for a random variable X is:

$$H(X) = -\sum_i P(x_i) \log_2 P(x_i) \tag{2}$$

and between two variables X and Y, it is given by:

$$H(X|Y) = \sum_j P(y_j) \sum_i P(x_i|y_j) \log_2 \left(P(x_i|y_j)\right) \tag{3}$$

$P(x_i)$: probabilities for all values of X, $P(x_i|y_j)$: conditional probabilities between X given the values of Y.
However, the study in [12] showed that features with more values are favored with information gain, thus its normalized version known as Symmetrical Uncertainty (SU) is used. Using equation (1), (2), and (3), SU is defined as:

$$SU(X, Y) = 2\left[\frac{IG(X, Y)}{H(X) + H(Y)}\right] \tag{4}$$

3.2 Redundancy Criterion

Our proposed redundancy method uses symmetrical uncertainty as correlation measure. Unlike the proposed methods in [10, 19, 20], we do not have to set a threshold. Using a threshold to analyze a feature redundancy is subjective because the redundancy of that feature depends on that threshold. In our method, we focus on absolute features redundancy. We define our redundancy criterion as follows: two features F_j and F_i are redundant with respect to a class C (the output variable) if and only if they provide exactly the same amount of information for the output variable. In other words if and only if:

$$SU(F_j, C) = SU(F_i, C) \tag{5}$$

If F_j and F_i are redundant, the least relevant one needs to be deleted. $SU(F_j, C)$ refers to the symmetrical uncertainty between a feature F_j and the class C and $SU(F_i, C)$ refers to the symmetrical uncertainty between a feature F_i and the class C.

3.3 Redundancy Analysis Algorithm

The algorithm defined below (Algorithm 1) can be literally translated as follows: from a list **F** of relevant features resulting from a feature selection method, we choose the most important or relevant feature F_j (line 1). Then, the symmetrical uncertainty between F_j and the next remaining feature F_i in the list **F** (line 2 to 6) is calculated. If the redundancy criterion is true (line 7), the feature F_i is deleted from the list **F** (line 7) then the next feature F_i in **F** is used (line 8) until all the remaining features in **F** have been used. Then we start over by varying F_j in the list (line 11) and so on until we have considered every remaining feature as F_j (line 12). After removing all the redundant ones, we add all the remaining features to the list **F'**, (**F'** $\leq$ **F**). This list, **F'**, containing only relevant and non-redundant attributes, will be used for the desired learning task.

4 EXPERIMENTAL APPROACH

The diagram below(Figure 1) reflects our experimental approach to perform the redundancy analysis.
Features Importance: we determine the importance of the features using a wrapper feature selection method and then rank the

Algorithm 1: Absolute Redundancy Algorithm

Input: $\mathbf{F}(F_1, F_2, ..., F_n)$, C // Features list **F** and target class C
Output: **F'** //Final List of non-redundant attributes

$F_j \leftarrow$ getFirstElement (**F**)
do begin:
    $F_i \leftarrow$ getNextElement(**F**)
    while F_i <> NULL:
        compute $SU(F_j, C)$, $SU(F_i, C)$
        if $SU(F_j, C) = SU(F_i, C)$:
            $\mathbf{F} \leftarrow \mathbf{F} \setminus \{F_i\}$
        end if
        $F_i \leftarrow$ getNextElement (**F**)
    end while
    $F_j \leftarrow$ getNextElement(**F**)
end until (F_j == NULL)
F' $\leftarrow$ **F**

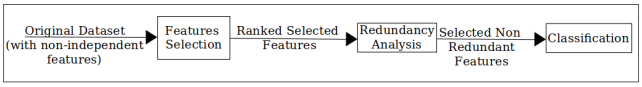

Figure 1: Design of our experimental approach

features in order of importance from the most important one to the least. Importantly, this ordered list constitutes the "Ranked Relevant Features".

Redundancy Analysis: to obtain the best subset of non-redundant features, we use the "Ranked Relevant Features" list to proceed to the redundancy analysis using the redundancy criterion that we defined in section 3.2 and the algorithm 1.

Classification: the redundancy analysis produces a final reduced list of features, "Selected Non-Redundant Features". This list is used to perform a supervised learning task using SVM [14] and C4.5 [15] classifiers. These classifiers were used for comparison purpose with other existing methods.

4.1 Experiments

We carry out our experiments in a way so that we can compare our results with other existing methods. The existing methods that can reduce redundancy and to which we compare our results are RABFS, mRMR and FCBF respectively. RABFS [19] uses maximum information coefficient to establish a threshold and analyze features redundancy and build a subset of features for training. In mRMR [10] the aim is to select features with a high relevance with the target and a low redundancy between themselves. FCBF uses symmetrical uncertainty as correlation measure and approximate Markov blanket to remove redundancy [20].

4.2 Datasets

To evaluate the performance of our method in finding redundant features and improving the performance of the learning task, 6 datasets including biological and text data from the UCI [3] were used. In Table 1 below, we give details of those datasets. Those datasets were chosen based on their differences, the number of features varying from 325 to 22283. Plus, this choice will help us to compare the result of our method against other proposed methods that have used the same datasets.

Table 1: Experimental Datasets used

Dataset	Observations	Nb of Features
Colon	62	2000
ALLAML	71	7129
PCMAC	1943	3289
Prostage-GE	102	5966
GLI-85	85	22283
lung_small	73	325

4.3 Results

On a classification task, results of our method were compared to others based on the number of selected features and the classification accuracy. We have used SVM with Gaussian kernel [14] and C4.5 [15] as classifiers. SVM is a supervised machine learning model that uses classification algorithms for two-group classification problems. It has many obvious advantages in solving large-dimensional. C4.5 is an algorithm used to generate a decision tree that can be used for classification.

Table 2: Number of Selected Features by method

Dataset	Our Method	RABFS	FCBF	mRMR
Colon	7	3	9	3
ALLAML	5	3	6	3
PCMAC	54	28	112	28
Prostage-GE	21	16	4	16
GLI-85	13	4	5	4
lung_small	52	34	112	34

Table 3: SVM Classification accuracy by method

Dataset	Our Method	RABFS	FCBF	mRMR
Colon	**92.03**	91.66	90.0	78.57
ALLAML	**97.8**	96.07	92.85	97.14
PCMAC	**83.01**	80.91	77.51	56.25
Prostage-GE	93.83	**94.0**	91.99	**94.0**
GLI-85	**94.01**	92.77	90.69	89.30
lung_small	**88.0**	84.82	59.88	84.64
average	**91.44**	90.03	87.71	79.42

In tables 2, 3 and 4, we report the results obtained during our experiment including results reported by the other methods. To select our features, we used a wrapper features selection method. Then, with the list of obtained features, we apply our redundancy criterion in order to obtain the list of the best features without redundancy. And finally, we perform a classification task using SVM

Table 4: C4.5 Classification accuracy by method

Dataset	Our Method	RABFS	FCBF	mRMR
Colon	**92.04**	91.90	75.47	91.78
ALLAML	**96.56**	96.07	95.71	94.28
PCMAC	**84.01**	82.50	77.81	59.42
Prostage-GE	**91.73**	90.09	86.18	84.18
GLI-85	**96.60**	95.13	84.58	85.69
lung_small	**89.01**	87.76	81.14	78.61
average	**91.65**	90.58	83.49	82.33

and C4.5 so we can compare our results with other existing methods. To assess the effectiveness of our result in term of accuracy, we applied cross validation techniques on each dataset. The results above show that our method performs well. Table 3 and Table 4 show results obtained by the four methods on SVM and C4.5. Compared to the others methods, we clearly see that our method performed better. In table 3 with SVM, our algorithm has a higher accuracy than other methods on 5 datasets. Only the Prostage-GE dataset, RABFS has a higher accuracy than our method. In table 4 with C4.5, the accuracy of our method is better on all the six datasets than all the other methods.

In Table 2, the other algorithms found fewer features than our proposed method. This can be understood by the fact that the other methods used a Filter strategy by setting a threshold to select the features while we used a wrapper approach.

5 CONCLUSION

This article presents a redundancy analysis criterion based on symmetrical uncertainty which is a measure of correlation between features. We then design a redundancy analysis algorithm designed according to this criterion. Unless other proposed redundancy methods, our algorithm does not require users to set a threshold. The performance of our method was experimentally compared to other methods such as RABFS, FCBF and mRMR on six different data sets. The comparative results show that our method finds satisfactory results.

Therefore, we intend to conduct further research in the future to incorporate our redundancy criterion into a hybrid feature selection method in order to select both relevant and non-redundant features.

REFERENCES

[1] Thushara Amarasinghe, Achala Aponso, and Naomi Krishnarajah. 2018. Critical Analysis of Machine Learning Based Approaches for Fraud Detection in Financial Transactions. In *Proceedings of the 2018 International Conference on Machine Learning Technologies* (Jinan, China) *(ICMLT '18)*. Association for Computing Machinery, New York, NY, USA, 12–17. https://doi.org/10.1145/3231884.3231894

[2] Girish Chandrashekar and Ferat Sahin. 2014. A Survey on Feature Selection Methods. *Comput. Electr. Eng.* 40, 1 (Jan. 2014), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[3] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[4] Boli Fang, Miao Jiang, P. Cheng, J. Shen, and Yi Fang. 2020. Achieving Outcome Fairness in Machine Learning Models for Social Decision Problems. In *IJCAI*.

[5] Bahar Farahani, Mojtaba Barzegari, and Fereidoon Shams Aliee. 2019. Towards Collaborative Machine Learning Driven Healthcare Internet of Things. In *Proceedings of the International Conference on Omni-Layer Intelligent Systems* (Crete, Greece) *(COINS '19)*. Association for Computing Machinery, New York, NY, USA, 134–140. https://doi.org/10.1145/3312614.3312644

[6] I. Gheyas and L. Smith. 2010. Feature subset selection in large dimensionality domains. *Pattern Recognit.* 43 (2010), 5–13.

[7] Mark Andrew Hall. 1999. Correlation-based feature selection for machine learning. (1999).

[8] George H. John, Ron Kohavi, and Karl Pfleger. 1994. Irrelevant Features and the Subset Selection Problem. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (New Brunswick, NJ, USA) *(ICML'94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 121–129.

[9] Ioannis K. Paparrizos, B. B. Cambazoglu, and A. Gionis. 2011. Machine learned job recommendation. In *RecSys '11*.

[10] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 8 (Aug. 2005), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

[11] Zhiwei (Tony) Qin, Jian Tang, and Jieping Ye. 2019. Deep Reinforcement Learning with Applications in Transportation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discover & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 3201–3202. https://doi.org/10.1145/3292500.3332299

[12] Laura Elena Raileanu and Kilian Stoffel. 2004. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence* 41, 1 (2004), 77–93.

[13] G. Thippa Reddy, M. Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. 2020. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* 8 (2020), 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942

[14] Matthias Ring and Bjoern M. Eskofier. 2016. An Approximation of the Gaussian RBF Kernel for Efficient Classification with SVMs. *Pattern Recogn. Lett.* 84, C (Dec. 2016), 107–113. https://doi.org/10.1016/j.patrec.2016.08.013

[15] S. Salzberg. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* 16 (1994), 235–240.

[16] Nirmalya Thakur and Chia Y. Han. 2020. An Approach for Detection of Walking Related Falls During Activities of Daily Living. In *2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*. 280–283. https://doi.org/10.1109/ICBAIE49996.2020.00066

[17] Adnan Ullah, Usman Qamar, Farhan Hassan Khan, and Saba Bashir. 2017. Dimensionality Reduction Approaches and Evolving Challenges in High Dimensional Data. In *Proceedings of the 1st International Conference on Internet of Things and Machine Learning* (Liverpool, United Kingdom) *(IML '17)*. Association for Computing Machinery, New York, NY, USA, Article 67, 8 pages. https://doi.org/10.1145/3109761.3158407

[18] B. Venkatesh and J. Anuradha. 2019. A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies* 19 (2019), 26 – 3.

[19] Mei Wang, Xinrong Tao, and Fei Han. 2020. A New Method for Redundancy Analysis in Feature Selection. In *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence* (Sanya, China) *(ACAI 2020)*. Association for Computing Machinery, New York, NY, USA, Article 21, 5 pages. https://doi.org/10.1145/3446132.3446153

[20] L. Yu and H. Liu. 2004. Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res.* 5 (2004), 1205–1224.