Article:

Rajab, M. and Wang, D. orcid.org/0000-0003-0068-1005 (2020) Practical challenges and recommendations of filter methods for feature selection. Journal of Information & Knowledge Management, 19 (01). 2040019. ISSN 0219-6492

https://doi.org/10.1142/s0219649220400195

Journal of Information & Knowledge Management
Vol. 19, No. 1 (2020) 2040019 (15 pages)

DOI: 10.1142/S0219649220400195

Practical Challenges and Recommendations of Filter Methods for Feature Selection

Mohammed Rajab* and Dennis Wang†
The University of Sheffield, UK
*mdrajab@gmail.com
†dennis.wang@sheffield.ac.uk

AQ: Please check the affiliation for both the authors.

Published

Abstract. Feature selection, the process of identifying relevant features to be incorporated into a proposed model, is one of the significant steps of the learning process. It removes noise from the data to increase the learning performance while reducing the computational complexity. The literature review indicated that most previous studies had focused on improving the overall classifier performance or reducing costs associated with training time during building of the classifiers. However, in this era of big data, there is an urgent need to deal with more complex issues that makes feature selection, especially using filter-based methods, more challenging; this in terms of dimensionality, data structures, data format, domain experts' availability, data sparsity, and result discrepancies, among others. Filter methods identify the informative features of a given dataset to establish various predictive models using mathematical models. This paper takes a new route in an attempt to pinpoint recent practical challenges associated with filter methods and discusses potential areas of development to yield better performance. Several practical recommendations, based on recent studies, are made to overcome the identified challenges and make the feature selection process simpler and more efficient.

Keywords: Feature selection; filter methods; machine learning; data imbalance; ranking methods.

1. Introduction

The curse of dimensionality is one of the challenges that domain experts often face when dealing with massive amounts of data (Town and Thabtah, 2019). Feature selection is a critical processing step that directly affects the success of machine learning algorithms by reducing space dimensionality through identifying the relevant set of features to be used (Hall, 2000). It also involves simplifying the classification process by strengthening the decision rules of the feature selection algorithm (Kamalov and Thabtah, 2017). Feature selection plays a vital role in classification because a robust feature selection mechanism can reduce the computational complexity associated with the learning process and improve its generalisation capabilities (Maldonado *et al.*, 2014). Domains characterised with a large number of features and a small number of samples benefit immensely through feature selection mechanisms. For instance, domains such as biochemistry, bioinformatics, text mining, medical diagnosis, and biomedicine require robust feature selection algorithms to improve the performance and comprehensibility of the models; these are often established based on a few samples and a large number of features (Yu and Liu, 2004a; Saeys *et al.*, 2008; Thabtah and Peebles, 2019).

Filter, wrapper and embedded are the three primary types of feature selection methods used for learning purposes. The filter method is the most common and involves selecting features without utilising a classification algorithm. Basically, this method involves filtering out irrelevant features using various selection principles such as information gain (IG) (Rajab, 2017). Filter methods use selection criteria to assign scores for the available features in the training dataset and then invoke a ranker search method to rank each individual feature based on the computed scores (Tang *et al.*, 2014). Informative features usually gain higher scores and uninformative features gain lower scores. Finally, the complete features, ranked on computed scores, are offered to the end user for subset selection. Based on the selection principles used, there are various filter-based feature selection methods such as IG (Quinlan, 1986), Pearson's correlation (Hall, 1999) and Fisher's score (Gu *et al.*, 2012), among others. Wrapper methods consider using a machine learning algorithm to identify classifiers for each possible subset in the input dataset. Hence, this kind of feature selection offers the best outcome yet suffers from a lengthy, exhaustive search, particularly when the input data are highly dimensional (Thabtah *et al.*, 2018). Lastly, embedded methods use a combination of filter and wrapper methods to select an ideal set of features. This research is concerned only with filter-based methods.

Several research studies have evaluated filter-based methods, i.e. Thabtah *et al.* (2011, 2018), Rajab (2017), Zhang *et al.* (2014), Estevez *et al.* (2009), Hall (2000), Zhao *et al.* (2018), Kamalov and Thabtah (2017), and Hancer *et al.* (2017). However, most of these investigated functional issues with filter methods such as the impact on predictive performance, or enhancing training efficiency; few covered practical challenges related to the basis on which features are selected and how results can be interpreted (Cherrington *et al.*, 2019). For example, a drawback of the filter methods, such as result dependencies, which make it hard for the end user to decide which features to choose prior to the learning process, has been investigated by few scholars. These combine results of multiple filter-based methods to reduce results variability, i.e. Labani *et al.* (2018); Gao *et al.* (2018); Rahmaninia and Moradi (2017). Despite this effort, recent research (Cherrington *et al.*, 2019) pinpointed that there is a need for a domain expert to manually check the outcomes of filter-based methods to recommend the final set of features needed; this can be resource-demanding. More importantly, the authors indicated that there is no fine line to discriminate among features in the results sets which can also be a serious issue. Hence, this research covers practical challenges in filter-based methods and presents viable recommendations to overcome these issues. Particularly, this research builds upon previous efforts and ~~critically analyses crucial~~ possible research directions rarely covered including feature ranking, results discrepancies, thresholding, feature-to-feature correlation, domain expert involvement, and data imbalance.

AQ: Please check whether the suggested running head is correct.

The paper consists of ~~five~~ main sections. The Introduction section provides an overall understanding of the feature selection process, filter-based methods, aims, objectives, and the outline of the paper. The second section further explains the research problem and previous related work by various scholars. Discussion, the third section, critically analyses the potential challenges of filter-based feature selection methods with practical recommendations to overcome identified challenges. The conclusion wraps up the information provided with suggestions on future work.

2. Problem and Literature Review

Filter-based feature selection is a research topic that has attracted the attention of many scholars and experts in multiple domains. Figure 1 shows filter methods in the learning process. The filter method involves carrying out feature selection as a preprocessing step without an induction algorithm. Training data are processed through a mathematical criterion to compute and assign scores to features in the training dataset; then a feature score is used to rank the features. These feature scores vary based on the type of the filter method used, and all the feature scores/rankings are offered to the end-user to make relevant decisions. Domain experts, or the end-user, decide the features to be used in the learning process based on their computed scores. The optimum threshold between selected and eliminated features is determined by the end-user based on knowledge and experience. Finally, a machine learning approach is employed to process the results set of the features and produce the classifier. The accuracy and the performance of the established classifier are evaluated by applying the model on sample data.

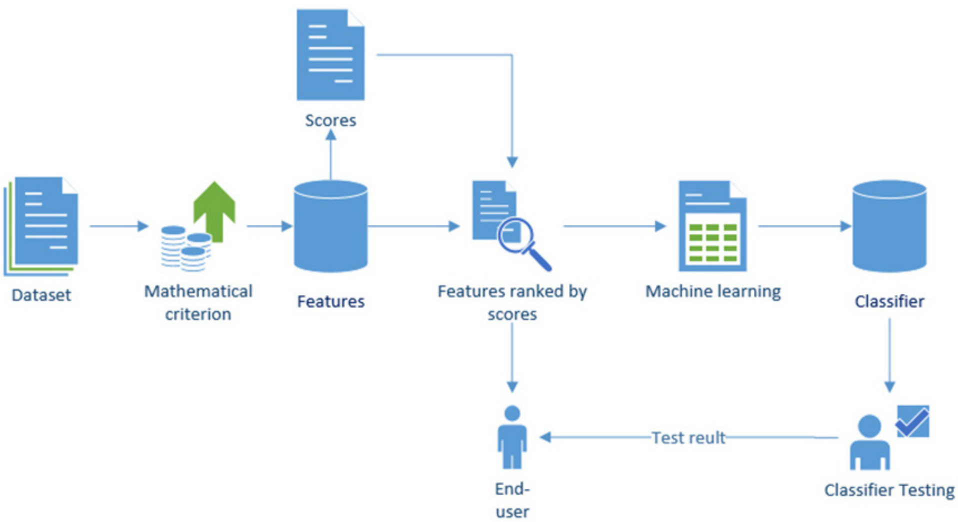

Fig. 1. Filter method as part of the learning process.

Thabtah *et al.* (2019a) introduced an observed frequency-based feature selection method called Least Lost (L2) to reduce the dimensionality of data by eliminating noisy data from the datasets while maintaining a healthy classifier performance. It is a more simplified and in-built approach that involves ranking of each variable in ascending order based on the L^2 distance between observed and expected variables and class labels. The scores are computed based on observed and expected probabilities of the available features. Tests conducted using datasets from the University of Irvine Repository (UCI) reported that L^2, when applied in the pre-processing phase, results in fewer features being obtained. When these are further processed by a machine learning algorithm, they derive competitive classifiers in terms of accuracy. L^2 implementation in Java can be accessed at https://github.com/suhel-hammoud/L2.

Zhao *et al.* (2018) proposed the redundant penalty between the feature mutual information algorithm (RPFMI), a filter-based feature selection mechanism, to identify optimal features in terms of redundancy, relationship between classifier and the selected features, and the correlation between selected features and the class labels and small data samples. The experimental results of the study suggested that the proposed RPFMI is highly effective in selecting an optimal set of features for intrusion detection as it demonstrated a high accuracy.

Gao *et al.* (2018) introduced the dynamic change of selected feature (DCSF), with the class a linear filter feature selection method, which takes dynamic information changes of the selected features with the class labels into account in the feature selection process; this to yield more accurate and efficient results. This novel model uses conditional mutual information between candidate features and class labels to identify the most informative features; the other conventional filter methods use mutual information to compute the relevancy of the candidate features to the select optimal feature subset. The experimental results implied that DCSF has the highest average classification accuracy of all the other compared methods.

Another filter mechanism presented by Hancer *et al.* (2017) is quite unique. These authors focus on selecting features based on their true rankings obtained by applying ReliefF (Robnik-Šikonja and Kononenko, 2003) and Fisher Score (Bishop, 1995) rather than focusing on their mutual redundancies. MIRFFS (Mutual Information, ReliefF, and Fisher Score), the proposed mechanism, used differential evolution (DE) (Marinaki and Marinakis, 2013) as the search strategy and it has two parts: one mechanism to be applied on single-objective problems and the other on multi-objective problems.

Labani *et al.* (2018) introduced multivariate relative discrimination criterion (MRDC), a novel filter-based feature selection mechanism to enhance the performance of the text classification process. This is accomplished by diminishing the dimensionality in feature space using minimal-redundancy and maximal-relevancy (mRmR) (Peng *et al.*, 2005). MRDC involves identifying the most relevant features using relative discrimination criterion (RDC) (Rehman *et al.*, 2015). Since RDC is not capable of classifying the irrelevant features, it utilises the Pearson correlation matrix to perform that task.

Kamalov and Thabtah (2017) used three robust filter methods in combination to produce a new feature selection mechanism (vectors of scores/V-score) to select the most relevant features of a given dataset while eliminating the shortcomings and maximising the advantages. They used information gain (Quinlan, 1986), chi-squared statistic (Liu and Setiono, 1995), and inter-correlation methods (CFS) (Hall, 1999) together to stabilise each feature's ranking score; they were able to reap more accurate prediction results rather than when applying them individually.

OSFSMI (Online Stream Feature Selection Method based on Mutual Information) and OSFSMI-k is another mutual information-based online streaming feature selection method, presented by Rahmaninia and Moradi (2017), to distinguish between the most informative and uninformative features. This is done by computing the correlation between features and their relevancy to the class labels where the number of instances increases exponentially (for example, social networks, finance analysis applications, and traffic network monitoring systems). The general framework followed by the proposed OSFSMI model comprises two unique phases: online relevancy analysis to compute the relevancy of each newly arriving feature, and online redundancy analysis to estimate the effectiveness of each selected feature and eliminate any with effectiveness below the average. OSFSMI-k is a modified version of OSFSMI, developed to address the issues arising due to the continuously increasing nature of features. To end this, OSFSMI-k keeps selecting the correlated features until the size of the selected feature subset reaches a constant value (k).

A research by Estevez *et al.* (2009) proposed a normalised mutual information feature selection (NMIFS), to evaluate the relevancy and redundancy in the features of a given dataset. Researchers have used three mutual information-based feature selection methods: Battiti's mutual information feature selector (MIFS), MIFS-U (Battiti, 1994), and min-redundancy max-relevance (mRMR) (Peng *et al.*, 2005) criteria to develop NMIFS by enhancing their individual strengths and minimising their weaknesses. They also present the Genetic algorithm, guided by mutual information for feature selection (GAMIFS), a hybrid version of both the filter and wrapper methods that combines NMIFS and genetic algorithms to fine-tune their performance.

3. Filter Methods Challenges

High dimensional data have made feature selection difficult as it necessitates dealing with a large number of features during data processing creating multiple challenges related to efficiency and quality. These challenges can be opportunities to learn and investigate new intelligent techniques to generate a meaningful concise set of features. In this section, we discuss various challenges that researchers and domain experts may face when designing, employing, or developing filter methods for data processing.

3.1. *Results discrepancies*

Results discrepancy is one of the obvious challenges in filter methods as different results may be obtained from the same dataset when applying different methods. To demonstrate this issue, we applied three different filter methods: IG, Correlation, and ReliefF (keeping Ranker as the search method) on a nursery database (Bohanec *et al.*, 1997) using WEKA 3.8 (Hall *et al.*, 2009). Table 1 shows the features extracted by the three considered filter methods and their ranks based on the assigned weights.

Table 1 clearly shows differences in the results generated by the filter methods, especially the ranking. For instance, if we consider the results derived by the IG and correlation methods, after the third ranked feature, there is a discrepancy in the results for the remaining features ranked 4–8. This discrepancy arises mainly because of the different mathematical models used by the considered filter methods to compute the weights per feature in the dataset. All these mathematical models primarily employ a contingency table that holds the frequency of the feature and that of the feature-class together, besides observed and expected probabilities, among others. For example, IG uses entropy as a base metric to compute the weights; this relies on the information of the feature and the class in the dataset, whereas the chi-square method uses the observed and expected probabilities. These differences in computing the weight assigned to each feature in the mathematical model can clearly impact the order in which the final features sets are offered to the end-user. Consequently, when these features sets are processed by the learning algorithm, performance may also be impacted such as the predictive accuracy of the models derived.

Few studies have addressed this issue and presented viable solutions to stabilise the knowledge discovery process through robust feature selection methods. For example, Kamalov and Thabtah (2017) pinpointed the results discrepancy in filter methods and showed that this problem can lead to selecting the wrong feature subsets, thus impacting the performance of the classification models derived by the learning algorithm. The authors suggested a filter mechanism that involves combining and normalising IG, inter-correlation, and CHI feature scores to produce one

Table 1. Ranking results generated by each feature selection method.

Ranking	IG features	Correlation features	ReliefF features
1	Health	Health	Health
2	Has_nurs	Has_nurs	Has_nurs
3	Parents	Parents	Parents
4	Social	Housing	Housing
5	Housing	Social	Social
6	Children	Finance	Finance
7	Form	Children	Form
8	Finance	Form	Children

unified score that can be assigned to each available feature. The term "normalising" refers to the introduction of one unified feature score range instead of several that vary according to the feature selection method used. For instance, feature selection methods like IG produce data scores ranging from 0 to 1, whereas methods like CHI produce feature scores between (-1) and ($+1$). The experimental results demonstrated that the normalisation of feature scores, and then integrating these into one unified score, is highly effective in reducing the volatility in the feature selection outcomes.

A similar approach that deals with the results discrepancy of filter methods was proposed by Rajab (2017). The author presented a method that combines the score of IG and CHI after normalising the initial scores computed by both methods. The new feature selection method was applied on a cybersecurity application for detecting phishing websites and contrasted with other common filter methods. Results reported that Rajab's (2017) method indeed reduced the dimensionality of the dataset and selected features sets, and when processed, using decision tress and rule induction classification techniques, improved the detection rate of phishing websites.

3.2. *Feature ranking*

Feature ranking refers to the process of selecting "n" number of features based on their computed weights/scores. The weights are normally computed based on a feature's relevancy to the class variable. According to Duch *et al.* (2003), feature ranking is an independent evaluation process of the available features as per their importance to eliminate potentially irrelevant features. All filter-based feature selection methods use a "Ranker" to evaluate the features based on scores computed using statistics, information theory, or some functions of the classifier's output. IG, gain ratio (GR), symmetrical uncertainty (SU), CHI, IG and ReliefF methods are examples of filter methods that use Rankers in feature selection. IG ranks the features based on amount of information relevant to the class variable, reflected by each candidate feature, whereas GR uses the prediction capabilities of each candidate feature to determine their individual rankings (Novakovic *et al.*, 2011).

Feature ranking is used by domain experts as a basic way of determining the best feature subsets; however, Ranker search methods do not provide the number of features to be selected, instead leaving the domain expert to decide. Most existing ranking search methods employ an elementary approach to display features along with their rank. More importantly, they leave the decision of which features to select up to the users' experience and knowledge, which subsequently requires time, care, and accuracy. Therefore, there is a need to develop a new intelligent Ranker search method that specifically recommends the features that should be chosen and the ones to ignore. The new Ranker should act as a recommendation to the feature selection process, be totally independent, and not filter-based method-specific. This will enable the Ranker to be embedded with any filter methods without dependency or data sensitivity and thus act as a generic search method.

A number of research studies have evaluated the performance of available feature ranking methods. Most concluded that there is no one Ranker method that is intelligent enough to distinguish influential features from redundant ones without domain expert involvement (Hu *et al.*, 2003; Duch *et al.*, 2004; Novakovic *et al.*, 2011; Cherrington *et al.*, 2019). Further, none of the studies found an intelligent solution for ranking within filter methods; hence, more research and investigation is needed to develop more advanced Rankers that can be used effectively with any feature selection method.

3.3. *Optimum threshold and domain expert involvement*

Determining the optimal threshold between good and useless features is another vital issue related to feature selection. Most of the available filter methods do not distinguish the cut-off value which could help these methods provide a small subset of features rather than relying on the domain expert. Distinguishing between features is a difficult task because of the diverse nature of datasets, their characteristics, and filter methods' mathematical metrics used to calculate weights for each feature, among others (Thabtah *et al.*, 2018). This difficult task relies on the knowledge of the domain expert, requiring additional time, care, and resources.

Let us assume that there is a dataset with over 1,000 features, and IG or CHI is used to determine the influential features. Both these filter methods will return a feature set of 1,000 ranked on the assigned weights of the filter methods. Then, the user will have to choose possibly the top 5, top 10, top 30, top 100, etc. based on his/her requirements and experience, the process of selecting which features is lengthy and difficult with a high chance that the user may miss prominent features. Having an automated threshold embedded within the filter method to offer the domain expert a small subset of features would be advantageous. This threshold is important as it represents a boundary between features to be selected and features to be eliminated. Using irrelevant features and eliminating relevant features would negatively impact the performance of learning algorithms and possibly lead to confusing and false predictions.

More research and development is recommended to establish an automated feature selection technique that has an inbuilt metric to identify the optimal threshold between informative and uninformative features without having to rely on a domain expert, dataset characteristics, and mathematical equations as used in the filter method.

3.4. *Feature-to-feature correlation*

Most of the available feature selection-based filter methods do not consider feature-to-feature correlation when determining the optimal subsets during feature analysis. Valuating this is important as it helps to reduce the number of features and then offers a set that does not overlap in data instances and is different from each other yet correlated with the class. One of the successful methods that dealt with this issue was mRMR (Peng *et al.*, 2005) and its extensions. mRMR ranks each candidate feature based on its relevancy to the class identifying the redundant features (those correlated with each other). According to Cai *et al.* (2012), mRMR defines relevant features as those with minimum redundancy with each other while maintaining the maximum relevance with the class label. Mutual information (MI) is the parameter used by mRMR to measure the mutual dependencies between features and class labels to identify the redundant and the relevant features. Fast-mRMR and mRMRe (Jay *et al.*, 2013; Ramírez-Gallego *et al.*, 2016) are extensions of mRMR that were developed to overcome computational complexities of traditional mRMR and make it more efficient.

AQ: Please add the below references to the list with full publication details: Jay et al., 2013;

Limited research investigations have been conducted to highlight the importance of identifying feature-to-feature correlation to enhance the performance of the overall feature selection process. The study by Yu and Liu (2004a) is one such attempt that addressed the need to incorporate a redundant feature analysis process as relevancy is insufficient to determine the best subsets. The authors introduced a novel mechanism called fast correlation-based filter (FCBF). This involves first selecting relevant features and then identifying predominant features from the selected set to enhance the selection process through a relevance and redundancy analysis. Yu and Liu (2004b) also discussed the importance of identifying and eliminating redundant features in gene expression microarray data analysis to classify diseases or phenotypes accurately.

Various studies have used different mathematical metrics to identify the inter-correlation among the features to produce optimal feature subsets. Radovic *et al.* (2017) proposed the temporal mRMR (TmRMR), a filter approach which uses the value of F-statistics across different time steps as the parameter to compute the temporal information and relevancy among feature; this is by applying a dynamical time-warping approach to handle temporal gene expression data in an effective manner. F-statistics values determine redundant features by identifying features with small and large inter-class variances.

Another research by Gu *et al.* (2012) presented a novel approach called more relevance less redundancy (MRLR) that uses mathematical metrics such as information amount, conditional mutual information, and relevance degree to eliminate redundant features. Mutual information is one of the most common parameters used in identifying feature-to-feature correlation in most of the literature. Cai *et al.* (2012) also used the mutual information value to rank features and identify redundant features. In a former study by Yu and Liu (2004a,b), the linear correlation coefficient is suggested as a viable mathematical metric to determine the goodness of the features. The authors describe this as a successful method as it helps to identify the features with near zero correlation with the class and it helps to eliminate the redundant features through identifying those with high correlation to each other. Table 2 shows mathematical metrics used to identify feature-to-feature relevancy.

Table 2. Mathematical metrics used in feature selection approaches to derive feature-to-feature correlation.

Literature	Filter method	Mathematical metrics	Equation
Radovic *et al.* (2017)	TmRMR	F-statistics	$F(g_j, c) = \frac{1}{T} \sum_{t=1}^{T} F(g_j^{(t)}, c)$
Gu *et al.* (2015)	MRLR	Information amount, conditional mutual information, and relevance degree	$NMI(f_i; f_s) = \frac{MI(f_i;f_s)}{\min\{H(f_i) \cdot H(f_s)\}}$
Cai *et al.* (2012)	mRMR	Mutual information	$I(X,Y) = \int\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
Yu and Liu (2004a,b)	FCBF	Linear correlation coefficient	$r = \frac{\sum_i (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_i (x_i - \overline{x_i})^2} \sqrt{\sum_i (y_i - \overline{y_i})^2}}$

3.5. *Data imbalance*

The class imbalance is a critical challenge observed in datasets with extremely different class distributions, often encountered in the classification tasks, which may result in generating results that favour the dominant class in the dataset (the class label with higher frequency) (Japkowicz and Stephen, 2002). Data is said to be imbalanced when the majority of the classification instances belong to one class and only a few instances belong to a minority class, especially in medical applications (Thabtah *et al.*, 2019b). For instance, if we have data of 1,000 instances, where only 10 of them have been diagnosed with autism, if we consider "Autism" and "No Autism" as two class values, this dataset is highly imbalanced. It will be imperative to distinguish the features that are related to autism in this dataset, which is difficult as most instances belong to the "No Autism" class. Hence, scholars proposed a solution that is mainly data-driven to balance the data before feature selection and learning phases such as under-sampling and oversampling (Wasikowski and Chen, 2010; Yin *et al.*, 2013).

Machine learning algorithms are sensitive to data with imbalanced class labels since they produce classifiers that are biased to the majority class and overlook the minority class label. This is because data instances fed into the learning algorithm tend to assume the unavailable points to make predictions by generalising the available points to the entire population. Because of that, the classifier would demonstrate a poor prediction accuracy on the minority class (Wasikowski and Chen, 2010).

A study by Wasikowski and Chen (2010) compared different schemes that include sampling and feature selection techniques to evaluate which technique performed better in dealing with imbalanced class data. The study revealed that feature selection with signal-to-noise correlation coefficient (S2N) (Gailey *et al.*, 1997) and feature assessment by sliding thresholds (FAST) (Chen and Wasikowski, 2008) techniques are highly effective on class imbalanced data. But feature selection methods used for balanced data may not perform as well on the imbalanced data, so the feature selection method should focus more on identifying features that help to predict the minority classes rather than the majority classes. A major issue that is encountered is locating a threshold to distinguish between relevant and irrelevant features. In feature selection, various ratios are used to rank the features based on their relevancy to the target class labels, but when most of the data belongs to one class, the results tend to be biased towards the features relevant to the majority class, ignoring those with more potential to predict the minority classes (Pant and Srivastava, 2015).

Many studies have been conducted on determining the most appropriate feature selection method to be used on class imbalanced data to yield a better classifier performance (Japkowicz and Stephen, 2002; Wasikowski and Chen, 2010; Yin *et al.*, 2013; Maldonado *et al.*, 2014; Thabtah *et al.*, 2019b). Most of them investigated the impact of class imbalance data on classifier performance, but little research addresses the impact on the feature selection process of imbalanced classes. Yin *et al.* (2013) addressed this problem and presented two feature selection approaches to overcome the issue. One approach is based on class decomposition (Maimon and Rokach, 2002), which involves the partition of majority classes into small class subsets before feature selection, and the other is based on Hellinger distance (Beran, 1997); this measures the distribution divergence of each class to evaluate its goodness for feature selection. The results showed that the proposed two approaches outperformed most of the available conventional feature selection methods. In an experiment carried out on protein function data, Al-Shahib *et al.* (2005) showed that under-sampling the majority class prior to feature selection significantly increases the classifier performance on imbalanced data.

4. Recommendations and Conclusions

A high level of noise is a major problem that makes managing data difficult, and most often this noise is generated from the technology used in collecting data or the source of data itself. Dimensionality reduction through filter-based feature selection is a commonly used solution to eliminate this problem. However, in the era of big data in which we have different feature types, sparse data, and unstructured data, among others, filter methods face practical challenges that have been rarely addressed in recent research. This paper critically analysed challenges of filter-based methods associated with results quality and performance including results discrepancies, ranking of features in the results set, absence of clear threshold between good and bad features, handling imbalanced data, and feature-to-feature correlation.

Different feature selection methods deliver different selection outcomes as a result of the mathematical models used to compute the feature scores based on feature-to-feature frequencies, feature-to-class frequencies, and expected and observed frequencies of the features. Therefore, if two different feature selection methods are employed on the same dataset, the end user can get two different outcomes for the most relevant feature subsets. The paper highlights the importance of addressing this challenge as the credibility and reliability of the final learning algorithm depend enormously on the feature subsets selected through the employed filter method. Use of normalised feature scores is recommended to yield more static, reliable, feature selection outcomes. Further research to develop more normalised advanced feature scoring mechanisms is vital.

All the filter methods use simple rankers to weigh the features based on their importance or the relevancy to the class labels. These rankers are very primitive and do not provide information on how many features are to be selected or eliminated. Therefore, the number highly depends on the end-user's knowledge and level of expertise, requiring an excessive amount of time, effort, and care. Hence, there is a need for an advanced Ranker that intelligently offers the subset of features by creating a fine line to differentiate good features from useless ones. Hence, the end user will not have to scan the entire features within the results set, rather just take that offered by the Ranker.

Absence of a clear threshold between good and bad features is also another challenge pinpointed in the paper that makes conventional filter-based feature selection over-dependent on the end-user/domain experts' involvement. Determining the cut-off between relevant and irrelevant features is essential as using irrelevant features in induction models can hinder the learning process significantly. Hence, the importance of developing an automated threshold embedded into traditional filter methods is emphasised.

Disregarding the feature redundancies is one of the main drawbacks of filter-based feature selection. Identifying the feature-to-feature correlation is of utmost importance as it helps to eliminate features that overlap. Therefore, to overcome this challenge, a viable approach that determines the feature-to-feature correlation and automatically eliminates the redundant features should be embedded into existing filter methods.

Some data characteristics such as uneven distribution can also make the feature selection process biased and inaccurate. Feature selection requires data that is perfectly balanced to generate unbiased accurate results. But it is not always practical to have perfectly balanced data, therefore, the paper highlighted the need for a valid mechanism to balance imbalanced data prior to the feature selection process to yield better results. Smart automated sampling techniques are recommended to be integrated into filter methods to identify class imbalanced data and to balance this without changing the original data.

Further research and investigation are advised to produce more intelligent automated feature selection techniques that mitigate the identified challenges and make the feature selection process more effective and efficient. In the near future, we are going to examine a number of filter methods on pathological datasets related to dementia in order to determine high effective attributes that may have correlations with dementia at different levels. Feature selection can provide a bottom-up approach of exploring datasets to reveal hidden useful patterns; in the case of diagnosing dementia, features that are hidden from the eyes of a pathologist but have clear impact on detecting dementia can be identified. This bottom-up approach of recommending features to domain-experts, such as pathologists, must also demonstrate that the features are interpretable to clinicians and can reduce observer bias. Features that achieve this are much more likely to be adopted by the clinical community and used as valuable biomarkers for diagnosing and stratifying patients into subgroups. Further work is needed to investigate the determinants of influential features, especially within application domains to pinpoint factors that influence feature interpretability and bias. While we highlight general best practices for feature filtering, understanding their impact in different research domains will be critical for these to have true value.

AQ: Please provide volume number for all journal-type references.

References

Al-Shahib, A, R Breitling and D Gilbert (2005). Feature selection and the class imbalance problem in predicting protein function from sequence. *Applied Bioinformatics*, 195–203.

Battiti, R (1994). Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 537–550.

Beran, R (1997). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 445–463.

Bishop, C (1995). *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.

Bohanec, M, V Rajkovic and B Zupan (1997). Applications of qualitative multi-attribute decision models in health care. *International Journal of Medical Informatics*, 58-59, 191–205.

Cai, Y, T Huang, L Hu, X Shi, L Xie and Y Li (2012). Prediction of lysine ubiquitination with mRMR feature selection and analysis. *Amino Acids*, 1387–1395.

Chen, X-W and M Wasikowski (2008). FAST: A ROC-based feature selection metric for small samples and imbalanced data classification problems. In *Proceeding KDD '08 Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 124–132.

Cherrington, M, F Thabtah, J Lu and Q Xu (2019). Feature selection: Filter methods performance challenges. In *International Conference on Computer and Information Sciences (ICCIS)*, doi: 10.1109/ICCISci.2019.8716478.

Duch, W, T Wieczorek, J Biesiad and M Blachni (2004). Comparison of feature ranking methods based on information entropy. In *IEEE International Joint Conference on Neural Networks*, 10.1109/IJCNN.2004.1380157.

Duch, W, T Winiarski, J Biesiada and A Kachel (2003). Feature ranking, selection and discretization. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 10.1109/IJCNN.2004.1380157.

Estevez, P, M Tesmer, C Perez and J Zurada (2009). Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 189–201.

Gailey, P, A Neiman, J Collins and F Moss (1997). Stochastic resonance in ensembles of nondynamical elements: The role of internal noise. *Physical Review Letters*, 79(23), 4701–4704.

Gao, W, L Hu, P Zhang and F Wang (2018). Feature selection by integrating two groups of feature evaluation criteria. *Expert Systems with Applications*, 11–19.

Gu, Q, Z Li and J Han (2012). Generalized Fisher score for feature selection. *Machine Learning*, 256–269.

Hall, M (1999). *Correlation-Based Feature Selection for Machine Learning*, Hamilton: Waikato University, New Zealand.

Hall, M (2000). *Correlation-Based Feature Selection for Discrete and Numeric Class*, Hamilton: Waikato University, New Zealand.

Hall, M, G Holmes and E Frank (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.

Hancer, E, B Xue and M Zhang (2017). Differential evolution for filter feature selection based on information theory and feature ranking. *Knowledge-Based Systems*, 1–17.

Hu, K, Y Lu and C Shi (2003). Feature ranking in rough sets. *AI Communications — Volume Pre-press*, Issue Pre-press, pp. 41–50.

Japkowicz, N and S Stephen (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429–449.

Kamalov, F and F Thabtah (2017). A feature selection method based on ranked vector scores of features for classification. *Annals of Data Science*, 483–502.

Labani, M, P Moradi, F Ahmadizar and M Jalili (2018). A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, 25–37.

Liu, H and R Setiono (1995). Chi2: Feature selection and discretization of numeric attribute. In *Proceedings of the Seventh IEEE International Conference on Tools with Artificial Intelligence*, pp. 388–391.

Maimon, O and L Rokach (2002). Improving supervised learning by feature decomposition. *Foundations of Information and Knowledge Systems*, 178–196.

Maldonado, S, R Weber and F Famili (2014). Feature selection for high-dimensional class-imbalanced data. *Information Sciences*, 286, 228–246.

Marinaki, M and Y Marinakis (2013). An island memetic differential evolution algorithm for the feature selection problem. *Nature Inspired Cooperative Strategies for Optimization (NICSO)*, pp. 29–42.

Novakovic, J, P Strbac and D Bulatovic (2011). Towards optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research*, 119–135.

Pant, H and R Srivastava (2015). A survey on feature selection methods for imbalanced datasets. *International Journal of Computer Engineering and Applications*, IX(II), 197–204.

Peng, H, F Long and C Ding (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1226–1238.

Quinlan, J (1986). Induction of decision trees. *Machine Learning*, 1, 81–106.

Radovic, M, M Ghalwash, N Filipovic and Z Obradovic (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, doi: 10.1186/s12859-016-1423-9.

Rahmaninia, M and P Moradi (2017). OSFSMI: Online stream feature selection method based on mutual information. *Applied Soft Computing*, 1568–4946.

Rajab, K (2017). New hybrid features selection method: A case study on websites phishing. *Security and Communication Networks*. doi: 10.1155/2017/9838169.

Ramírez-Gallego, S, L Lastra, D Martíne, V Bolón-Canedo, J Benítez, F Herrera and A Alonso-Betanzos (2016). Fast-mRMR: Fast minimum redundancy maximum relevance algorithm for high-dimensional big data. *International Journal of Intelligent Systems*, 134–152.

Rehman, A, K Javed, H Babri and M Saeed (2015). Relative discrimination criterion — A novel feature ranking method for text data. *Expert Systems with Applications*, 3670–3681.

Robnik-Šikonja, M and I Kononenko (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53(1–2), 23–69.

Saeys, Y, I Inza and P Larranaga (2008). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2507–2517.

Tang, J, S Alelyani and H Liu (2014). Feature selection for classification: A review. *Data Classification: Algorithms and Applications.*

Thabtah, F, F Kamalov, S Hammoud and S Shahamiri (2019a). *A New Feature Selection Method Based on Simplified Observed and Expected Likelihoods Distance*. Available at https://github.com/suhelhammoud/L2.

Thabtah, F, N Abdelhamid and D Peebles (2019b). A machine learning autism classification based on logistic regression analysis. *Health Information Science and Systems*, 7(1), 12.

Thabtah, F and D Peebles (2019). A new machine learning model based on induction of rules for autism detection. *Health Informatics Journal*, doi.org/10.1177/1460458218824711.

Thabtah, F, F Kamalov and K Rajab (2018). A new computational intelligence approach to detect autistic features. *International Journal of Medical Informatics*, 1386–5056.

Thabtah, F, W Hadi, N Abdelhamid and A Issa (2011). Prediction phase in associative classification. *Journal of Knowledge Engineering and Software Engineering*, 21(6), 855–876.

Town, P and F Thabtah (2019). Data analytics tools: A user perspective. *Journal of Information and Knowledge Management*, 1950002.

Wasikowski, M and X-W Chen (2010). Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1388–1400.

Yin, L, Y Ge, K Xiao and X Wang (2013). Feature selection for high-dimensional imbalanced data. *Neurocomputing*, 3–11.

Yu, L and H Liu (2004a). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205–1224.

Yu, L and H Liu (2004b). Redundancy based feature selection for microarray data. In *KDD '04 Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 737–742.

Zhang, X, Y Hu, K Xie, S Wang, E Ngai and M Liu (2014). A causal feature selection algorithm for stock prediction modelling. *Neurocomputing*, 48–59.

Zhao, F, J Zhao, X Niu, S Luo and Y Xin (2018). A filter feature selection algorithm based on mutual information for intrusion detection. *Journal of Applied Science*, 1–20.