Feature Selection: Filter Methods Performance Challenges

Marianne Cherrington
Department of Computer Science
University of Huddersfield
Huddersfield, UK
marianne.cherrington@hud.ac.uk

Fadi Thabtah
Digital Technologies
Manukau Institute of Technology
Auckland, New Zealand
fadi.fayez@manukau.ac.nz

Joan Lu
Department of Computer Science
University of Huddersfield
Huddersfield, UK
j.lu@hud.ac.uk

Qiang Xu
Department of Engineering
University of Huddersfield
Huddersfield, UK
Q.Xu2@hud.ac.uk

***Abstract*— Learning is the heart of intelligence. The focus in machine learning is to automate methods that achieve objectives, improve predictions or encourage informed behavior. Feature selection is a vital step in data analysis that often reduces dataset dimensionality by eliminating irrelevant and/or redundant attributes to simplify the learning process or improve outcomes' quality. This research critically analyses different filter methods based on ranking procedures (Information Gain (IG), Chi-square (CHI), V-score, Fisher Score, mRMR, Va and ReliefF) and identifies possible challenges that arise. We particularly concentrate on how threshold determination can affect results of different filter methods based on ranked scores. We show that this issue is vital, especially in the era of big data in which users deal with attributes in the magnitudes of tens of thousands with only a limited number of instances.**

***Keywords*— data mining, feature selection, filter methods, machine learning, feature ranking**

I. INTRODUCTION

Data mining will intensify with the growth of cloud computing and the internet of things. In diverse fields such as microarray analysis, text categorisation, high-frequency financial data and phishing detection, data have extremely high dimensionality with sparse training samples, undermining statistical significance. Feature selection is a pre-processing phase that typically reduces data dimensionality and training time by eliminating possible noise and overfitting of models therefore improving user interpretation [1, 2].

Removing irrelevant and redundant features is important; many existing datasets contain thousands and sometimes tens of thousands of attributes. Consequently, the role of feature selection techniques is to exponentially decrease hypothesis space dimension [3]. For instance, in a training dataset that contains N features (independent variables) with a binary target class (dependent variable), the hypothesis space is massive [4]. A subset of useful predictive features can increase robustness of classification models, but redundant yet relevant variables might be excluded. This task can be efficiently accomplished using filter methods based on ranked scores. Feature selection methods based on ranking features can find potentially relevant suboptimal features, especially if the data have redundant variables [5].

Filter methods such as CHI and IG can be utilised on supervised and unsupervised learning tasks [6,7]. In supervised tasks such as classification, filter methods typically classify based on the target class and work well when there is large data representation. On the other hand, when we are dealing with descriptive learning task such as clustering, the process of feature selection is more challenging as no target class is available to guide the filter methods [8]. Typically, search methods used are categorised into wrapper, filter or embedded methods (that perform feature selection and classification simultaneously) [4]. Wrapper methods evaluate feature subsets based on performance by evaluating different sets of features' combinations using a classification algorithm. The resulting classifier is evaluated based on evaluation metrics such as predictive accuracy or error rate. Wrappers are exhaustive and tend to be slow, 'forceful' and resource intensive [9].

Filter methods use data characteristics to evaluate features without the use of classification algorithms. Filter measures can be based on information theory, correlation, distance, consistency, fuzzy-set and rough-set [10]. Firstly, features are selected and ranked, independently of feature space in the univariate case and in a batch in the multivariate case (naturally treating redundancies). The second step uses a performance criterion to choose the features with the highest ranking [11]. Some common univariate ranking methods are IG [6], CHI [7], multiple scores aggregation [3,8] and Fisher score [12]; multivariate methods include minimum redundancy maximum relevance [13] and ReliefF [14,15]; see Table 1 on the following page. Most methods are generalised but are often chosen to suit the problem type and/or to improve predictive reliability [1,16,17].

Most users, including domain experts, tend to favour using the default value given by the filter method to differentiate amongst influential and useless features whilst performing feature selection. This is because many of the users have no detailed knowledge of the training dataset or even the dataset's characteristics. Nevertheless, using the default threshold value may not be the favourable option as outcomes vary substantially from one dataset to another. A more realistic approach is to develop a new search procedure that can be integrated with any filter method. This search procedure will come up with a ranked feature along with corresponding scores based on the filter method used; notably, the number of variables can be chosen by the user.

This paper critically analyses common ranked filter methods to reveal their challenges, particularly how threshold determination can affect results. This obvious challenge impacts the filter methods, commonly resulting in

978-1-5386-8125-1/19/$31.00 ©2019 IEEE

high discrepancies between outcomes, regardless of the filter method used [1,18,19]. Future work will consider novel threshold free filter methods. This paper is structured as follows: Section 2 reviews common filter methods that produce ranked scores. Section 3 discusses the challenges of filter methods. Section 4 highlights conclusions and future work.

II. Literature Review

IG is commonly used in a range of applications and is based on an entropy metric. Pre-defined thresholds can rank and select relevant features [6]. IG reveals how informative an attribute in the input dataset is with respect to the dependent variable (class label). The information gained between the i^{th} feature f_i with entropy $H(f_i)$ and the class labels C is given in Table 1.

CHI filter method explores significant relationships using discretisation for mixed attribute and multiclass data. Phase 1 generalised ChiMerge uses an incremental χ^2 threshold to retain original data fidelity; phase 2 refines phase 1 [7]. If X is the count of feature a and class c occurring together, W counts class c without feature a, Y counts feature a without class c, Z counts neither a nor c occurring together and N is the total training set size, then the resulting CHI score is given in Table 1.

IG, CHI and similar filter methods often generate different scores for the set of variables in the input and can be criticised for deriving discrepant scores and the domain expert might get confused which filter methods to utilise. To minimise this problem, V-score filter method was proposed wherein IG and CHI scores are combined. V-score improves classification algorithms by improving consistency and reducing variability of ranking features [1]. The IG and CHI scores are first normalised, then combined in a single feature score vector, whose magnitude becomes a true scalar metric. The result is a mathematical structure for analysing the space of combined scores that compares features to one another, aiding dimensionality reduction. See Table 1.

A similar method to V-score was applied on behavioural and cybersecurity applications and successfully revealed that unifying scores of different filters can boost the performance of the phishing detection models [1,8,18].

Fisher score is a widely used supervised method that selects features separately by computing individual Fisher scores that span the data space. It cannot detect combined effects or manage redundant features but under certain orthogonality assumptions [12], Fisher's criterion produces optimal predictors [5]. Table 1 gives the Fisher score S_i for the i^{th} feature in the j^{th} class using mean μ_{ij}, variance ρ_{ij} and n_j instances in the j^{th} class and mean μ_i of the i^{th} feature.

Maximum Relevance Minimum Redundancy (mRMR) is a multivariate method that finds relationships amongst features (univariate methods consider each feature separately). The initial stage measures maximum statistical dependency based on mutual information, followed by more sophisticated feature selection [13]. mRMR measures relevancy-to-class using heuristic selection algorithms, while minimising feature similarities. Between continuous features *fi* and *fj*, *with* mutual information *I(i,j)*, correlation *C(i,j)*, target class *h* and F-statistic *F(i,h)*, the minimised redundancy W_c and maximised relevancy V_c are given in Table 1.

TABLE I. COMMON FILTER METHODS

	Mathematical Model
IG [6]	$IG(f_i, C) = -\sum_j p(x_j) log_2(p(x_j)) - H(f_i\|C)$ $where\ H(f_i\|C) = -\sum_k p(c_k) \sum_j p(x_j\|c_k) log_2[p(x_j\|c_k)]$ (1)
CHI [7]	CHI (a,c) = N(XZ –YW) /{(X+W)(Y+Z)(X+Y)(W+Z)} (2)
V-score [1,18]	$\| \upsilon_a \| = [(IG_a)^2 + (CHI_a)^2]^{½}$ (3)
Fisher score [12]	$F(x_i) = \sum_{k=1}^{K} n_j (u_{ij} - \mu_i)^2 \left\{ \sum_{k=1}^{K} n_j \rho_{ij}^2 \right\}^{-1}$ (4)
mRMR [13]	$W_c = \|s\|^{-2} \sum_{i,j \in s} \|C(i,j)\| \quad and \quad V_c = \|s\|^{-2} \sum_i F(i,h)$ (5)
ReliefF [14,15]	$S_i = \frac{1}{K} \sum_{k=1}^{l} \{ \frac{-1}{m_k} \sum_{x_j \in M_k} d(X_{ik} - X_{ij}) +$ (6) $\sum_{y \neq y_k} \frac{1}{h_{ky}} \frac{p(y)}{1 - p(y)} \sum_{x_j \in H_k} d(X_{ik} - X_{ij}) \}$
FCBF [20]	$SU(x, y) = 2\left\{ IG(x, y)\left(H(x) + H(y)\right)^{-1} \right\}$ $and\ IG(x/y) = H(y) + H(x) - H(y)$ (7)

ReliefF is a multiclass extension of Relief; it selects features to separate instances from classes, enhancing performance in many applications [14,15]. For *l* randomly sampled instances where M_k and H_{ky} are the sets of nearest point to $\mathbf{x}_k$ with the same class (size m_k) and class y (size h_{ky}); *p(y)* is the probability of, and instance from, class y [11], the i^{th} feature score is given in Table 1.

Fast correlation-based filter (FCBF) eliminates pairwise correlation analysis to improve efficiency for high-dimensional data; it employs a threshold symmetrical uncertainty value to determine relevancy and redundancy of features (as does Relief) [21]. Symmetrical uncertainty is the ratio between IG (numerator) and entropy (H) of features x and y, without interaction assessment [22].

III. Challenges of Ranked Filters

High-throughput techniques continue to generate ultrahigh dimensionality data requiring additional computational power, statistical accuracy and stability of algorithms. Ranked filter methods and domain experts can be beneficial, as prior knowledge and partitioning can lead to further efficiencies [23]. Social media sites generate linked data, which violates the basic and common assumption of independent and identically distributed data. It creates data that is massive, noisy, and incomplete and intensifies the magnitude of the feature selection issues already discussed. Research in this applied area is ongoing [11].

An extension of multi-source feature selection, sparse learning is the study of how to appreciate data from multiple knowledge sources that inherently possess a high degree of interpretability and low computational expense [11]. Feature selection is combined with joint model fitting so as to require fewer samples to attain reliability, but a trade-off between the goodness-of-fit measure and sparsity of the result occurs. These methods are useful with large collections of text documents for example, where improved understanding is desired without the need for user expertise.

Explanation-based feature selection (EBFS) is used when a class is formed from diverse attributes. EBFS attempts to overcome the resulting correlation issues and data irregularities in such data. EBFS can also create high-level concepts that elucidate and advance understanding of feature relationships [22]. New methods tend to have specific areas of application; embedded methods allow simultaneous feature selection and classification while other techniques improve with the use of hybrid and ensemble methods [24].

Many feature selection methods are dependent on domain experts, especially filter based ones [3,8]. Methods can be sensitive to noise, redundant variables, changes to initial conditions or 'instability of inputs' [5]. Noisy data can be pre-processed by applying domain-specific cleansing and normalisation techniques, or features that are less noisy can be selected as a condition. Suboptimal prediction issues are not necessarily a shortcoming; in microarray analysis, feature selection is used mainly for discovery purposes (as filter methods with $\geq 10^5$ features do not scale well) [10]. Feature ranking is ubiquitous; it is simple, practical and improves results. While multivariate methods have more universal predictors than univariate, poorer performance may result due to overfitting and regularisation techniques support better performance [25,26].

A vital and challenging issue arises when utilising filter methods. Variables below the predefined user threshold are not considered relevant and are therefore discarded. Discarding these variables is generally based on the user's predefined threshold or the default threshold which in many cases can be questionable. For example, the user can define the default threshold for IG, V-Score or Va methods before the feature selection phase initiates. Other methods use added constraints that can preclude near-trivial splits to further reduce the hypothesis space and reduce dimension [27]. Other thresholds can be employed, such as choosing the threshold to maximise gain, which may over-fit the learning models [28]. Furthermore, CHI normally uses a manually set α to terminate using discretisation but again, it is not easy to find an ideal value for α [10]. CHI and IG methods use a decrement rate between successive features to determine cut-off points. In certain fields, convention and expertise can guide the determination of the number of features to be selected. For example, in phishing data [8], it has been shown that fewer than nine features can be chosen using rule-based classification [1] but indeed this required researchers to create a new phase for evaluation of features.

The V-score method is indicative of stabilising methods and a prevalence of combined methods exist. The V-score and Va methods not only produce scalar metrics from the normalised scores, they are a relatively simple method. As well, analysis of variables can be undertaken using the resulting mathematical structure. It is important to note that they can also provide a rationale and basis from which to validate the number of features to choose. Still, it is suggested that in the trade-off between number of features and classifier accuracy, individual preferences can be used to improve overfitting, providing generalisation balance [1,8].

Fisher score commonly computes a score for each feature independently and uses a heuristic to select the top ranked features. Interactions and redundancies are therefore ignored and the number of features can be set to be 50% of the dimensionality of the data [29]. This 50% threshold may not work well for many datasets. Moreover, ReliefF is claimed to be a more robust and noise tolerant method that can detect conditional dependencies, however less prominent features can be underestimated at times. Additionally, the number of key features to be chosen is highly dependent on having a sufficient number of examples to sufficiently cover the hypothesis space [14]. A threshold θ is suggested for relief, where $0 < \theta \leq (\alpha m)^{-½}$ where α is the probability of accepting an irrelevant feature as relevant and m is the number of iterations. The upper bound for θ is very loose, so smaller values tend to be used [14].

Users tend to follow the default threshold to differentiate amongst features when applying filter methods in pre-processing datasets. However, besides the default threshold value, there are additional factors that affect the scores of the features. These include the dataset variables (distribution, types, structure, etc.) and more importantly the mathematical formula used by the filter method to calculate the scores. The outcomes of the filter methods may vary depending on these factors. An optimisation method to guide the search of the features would be advantageous, as it may offer the domain expert supporting evidence for the optimal number of variables that should be chosen. This will indeed improve the efficiency of the feature selection and provide a legitimacy of the set of features selected by the end user.

Realistically, although default thresholds exist in filter methods, determining the number of features to retain often requires a domain expert's opinion to balance functionality (fast, efficient training without undue loss of performance) while avoiding the curse of dimensionality. Pre-setting default thresholds or default mechanisms tend to be sub-optimal and conservative, thereby enlarging the hypothesis space. By providing improved methods with a clear recommendation on the number of variables needed, users would gain valuable knowledge and analysis would be simplified.

IV. Conclusions and Further Research

Feature selection is a dynamic field. As data quantity, complexity and diversity intensify, improved methods will continue to be generated. The use of different feature types and combinations will continue to expand in real applications, especially with the explosion of technological advances in areas such as cloud computing, the Internet of Things, intelligent automation and data streaming. Feature selection will continue to play an important role in dimensionality reduction as applications become more extensive and as data collection techniques become more automatic, instantaneous and as data storage capacity continues to expand. Exciting advances are being made and the scope for machine learning almost defies imagination; the need to work towards a unifying theoretical framework becomes ever more apparent because of the need for clarity amidst the formidable complexities of decision-making.

This paper identified some of the challenges of common feature selection techniques that relate to filter methods. Specifically, we reveal that most filter methods suffer from the problem of setting up the default threshold used to distinguish influential feature from redundant features; this issue continues to make machine learning heavily dependent on domain experts, as results can be inconsistent. This problem also leads to more serious performance issues related to the quality of the final features set offered to the end-user besides the additional inefficiency when the user is required to manually check the features set. Therefore, novel optimisation techniques continue to be generate; they are needed in both single- and multi-objective feature selection for large scale data. These search methods will be able to stop adding features into the final features set when adding such features have no added value to both the filter method performance and the end-user. These novel methods also suggest new and expanding fields of application where sensitivity analysis can support a course of action and highlight areas of deficiency. In the near future, we intend to propose a new search method to tune default thresholds in different filter methods and perform extensive experimentations using sensitivity analysis.

REFERENCES

[1] F. Kamalov and F. Thabtah, “A feature selection method based on ranked vector scores of features for classification,” Annals of Data Science, 2017.

[2] R. AlShboul, F. Thabtah, N. Abdelhamid, M. Al-diabat, “A visualization cybersecurity method based on features' dissimilarity”, Computers & Security 77, 289-303. 2018

[3] F. Thabtah, F. Kamalov, K. Rajab, “A new computational intelligence approach to detect autistic features for autism screening”. International Journal of Medical Informatics, Volume 117, pp. 112-124.

[4] H. Liu and H. Motoda, *Computational methods of feature selection.* New York: CRC Press, 2007, p. 4.

[5] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research 3, March 2003.

[6] J. Quinlan, “Induction of decision trees,” Machine Learning 1:1, pp. 81-106, 1986.

[7] H. Liu and R. Setiono, ”Chi2: Feature selection and discretization of numeric attribute,” in Proc. The Seventh IEEE International Conference on Tools with Artificial Intelligence, Herndon, 1995, p. 388.

[8] F. Thabtah and N. Abdelhamid, “Deriving correlated sets of website features for phishing detection: a computational intelligence approach,” Journal of Information and Knowledge Management, 15, No. 4, p.1650042-8, November 2016.

[9] L. Wolf and A. Shashua, “Feature selection for unsupervised and supervised inference: the emergence of sparsity in a weight-based approach,” Journal of Machine Learning Research 6, p. 1855, November 2005.

[10] M. Hall, “Correlation-based feature selection for machine learning,” Thesis, Department of Computer Science, Waikato University, New Zealand. p. 28, April 1999.

[11] J. Tang, S. Alelyani, H. Liu, “Feature selection for classification: a review,” Data Classification: Algorithms and Applications, CRC Press, 2013.

[12] R. Duda, P. Hart and D. Stork, *Pattern Classification*, New York: John Wiley & Sons, 1999.

[13] H. Peng, F. Long and C. Ding, “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp 122-1238, August 2005.

[14] M. Robnik-Šikonja and I. Kononenko, “Theoretical and empirical analysis of Relief and ReliefF,” Machine Learning, 53:23-69, 2003.

[15] J. Novaković, P. Štrbac, and D. Bulatović, “Toward optimal feature selection using ranking methods and classification algorithms,” Yugoslav Journal of Operations Research, 21, 2011.

[16] F. Thabtah, “Machine learning in autistic spectrum disorder behavioral research: A review and ways forward”, Informatics for Health and Social Care 43 (2), 1-20.

[17] F. Thabtah, D. Peebles, “A new machine learning model based on induction of rules for autism detection”. Health Informatics Journal, 1460458218824711.

[18] K Rajab, “New Hybrid Features Selection Method: A Case Study on Websites Phishing”, Security and Communication Networks 2017 (Article ID 9838169).

[19] F. Thabtah, “An Accessible and Efficient Autism Screening Method for Behavioural Data and Predictive Analyses”. Health Informatics Journal. 19:1460458218796636. doi: 10.1177/1460458218796636. 2018.

[20] L. Yu and H. Liu. Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the twentieth International Conference on Machine Learning, pages 856–863, 2003.

[21] N. Sánchez-Maroño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter methods for feature selection a comparative study, in: H. Yin, P. Tino, E. Corchado, W. Byrne, X. Yao (Eds.), Intelligent Data Engineering and Automated Learning —IDEAL 2007. Lecture Notes in Computer Science, vol. 4881, Springer, Berlin, Heidelberg, 2007, pp. 178–187.

[22] Z. Zhao and H. Liu, “Searching for interacting features”, In Proceedings of the 20th International Joint Conference on AI, 2007.

[23] Z. Zao, F. Morstatter, S. Sharma, S. Alelyani, A. Anand and H. Liu. “Advancing feature selection research - ASU feature selection repository”, TR-10-007, School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, 2010.

[24] V. Bolón-Canedo, N. Sánchez-Marono, A. Alonso-Betanzos, “Recent advances and emerging challenges of feature selection in the context of big data,” Knowledge Based Systems, 86, C, September 2015.

[25] Y. Yu, “SVM-RFE Algorithm for Gene Feature Selection,” Technical report. University of Delaware.

[26] I. Guyon, H. Bitter, Z. Ahmed, “Multivariate non-linear feature selection with kernel multiplicative updates and Gram-Schmidt Relief,” In: Proceedings of the BISC FLINT CIBI 2003 Workshop, Berkeley, 2003.

[27] J. Quinlan, “C4.5: Programs for Machine Learning,” Morgan Kaufmann, San Mateo, 1993.

[28] J. Quinlan, “Improved use of continuous attributes in C4.5,” Journal of Artificial Intelligence Research, 4, 77–90, 1996.

[29] Q. Gu, Z. Li, and J. Han. “Generalized fisher score for feature selection,” In: Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, pages 266–273, 2011.