A novel feature selection method for image classification

D. ASIR ANTONY GNANA SINGH[a]. S. APPAVU ALIAS BALAMURUGAN[b], E. JEBAMALAR LEAVLINE[c]
[a]*Department of Computer Science and Engineering, Anna University, BIT Campus, Tiruchirappalli – 620 024, Tamilnadu, India*
[b]*Department of Department of Information Technology, K.L.N College of Information Technology, Sivagangai-630 611, Tamilnadu, India*
[c]*Department of Electronics and Communication Engineering, Anna University, BIT Campus, Tiruchirappalli – 620 024, Tamilnadu, India*

Image classification plays a significant role in pattern recognition. In the recent past, due to the advancements in imaging technology, massive data are being generated through various image acquisition techniques. Classifying these massive images is a challenging task among the researchers. This paper presents a novel feature selection method to improve the performance of image classification. The performance of the proposed method is tested on the publically available real image dataset and compared with various state-of-the-art feature selection methods. The experimental results show that the proposed method outperforms the other state-of-the-art methods.

(Received September 27, 2015; accepted October 28, 2015)

Keywords: Image classification, Feature selection, Image recognition, Clustering, Image acquisition techniques

1. Introduction

Image recognition is an attractive area in various fields such as engineering, medical, social, economical and etc. to extract the knowledge from the images that are acquired through various image capturing devices such as digital camera, x-ray, ultrasound, spectroscope, optical coherent tomography (OCT), positron emission tomography (PET), computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography–computed tomography (PET-CT) and etc. [1]. The image recognizing mechanisms struggle in image classification due to the high dimensional (feature) space of the images. The high dimensional space degrades the performance of the image classification in terms of accuracy and time to build the classification model since it contains irrelevant and redundant features. Therefore, feature selection is employed to reduce the number of features extracted from the images by removing the irrelevant and redundant features. In image classification or image recognition, the classification algorithm is employed to build the classifier by learning the feature space of the image dataset with the class attribute. Then the classifier is used to recognize or predict the label of the unidentified or unlabeled images.

The feature selection is classified into four categories namely wrapper, embedded, filter and hybrid approaches. The wrapper approach uses the supervised learning algorithm or classification algorithm to evaluate the significance of the feature subsets. This is computationally expensive since it adopts the supervised learning algorithm and has poor generality as it gives higher classification accuracy only for the particular classification algorithm that is used in the feature selection process. Embedded method uses a part of the supervised learning algorithm for selecting the features and also be computationally expensive since it uses the supervised learning algorithm. However, it is computationally cheaper than the wrapper approach. The filter method uses any one of the statistical measures for selecting the significant features regardless of the classification algorithm. Therefore, the filter approach is computationally cheaper and offers higher generality. The hybrid approach is a combination of wrapper and filter approaches [2].

This paper proposes a novel computationally cheaper filtered-based feature selection method namely ranking with clustering based feature selection (RCFS) with high generality to improve the performance of the classifier for image classification. The performance of the proposed method is statistically analyzed and compared with various state-of-the-art feature selection methods on various real-world image datasets. The results show that the proposed method is more promising than other methods compared.

2. Related works

This section reviews the research works that are related to the proposed method. In the feature selection literature, some researchers have succeeded in effectively removing the irrelevant features but failed to handle the redundant features [3]. On the other hand, few researchers dealt with removing the irrelevant features and redundant features [4].

The process of feature selection aimed at choosing the relevant features. The best example is Relief [5] that was developed with the distance-based metric function that weights each feature based on their relevance (correlation) with the target class. However, Relief is ineffective as it can handle only two-class problems. The modified version of the Relief known as ReliefF [6] handled the multi-class problems and dealt with incomplete and noisy datasets, but it failed to reduce the redundant features. Holte [7] developed a rule-based attribute selection known as OneR which forms one rule for each feature and selects the rule with the smallest error.

R. Battiti et al. [8] developed a mutual information-based feature selection method (MIFS). In this method, mutual information measure is used to determine the relevancy between the individual feature and the target class. The attributes having similar information are considered as the redundant features that are to be removed. Fleuret et al. [9] presented a feature selection scheme namely conditional mutual info maximization (CMIM) that recursively chooses the features that have maximum mutual information with the target class for classification. Lin and X. Tang [10] introduced an information theory-based conditional infomax feature extraction (CIFE) algorithm to measure the class-relevancy and redundancy for feature selection. Gavin Brown et al. used the conditional redundancy (CondRed) [11] metric for selecting the significant features from the dataset.

From the literatures, it is observed that some of the feature selection methods do not treat the redundant features and some of them use the same selection criteria for redundant and relevant feature identification. Therefore, they are not able to produce better accuracy for the classification algorithms.

In order to overcome these limitations, the proposed RCFS method adopts the computationally cheaper high generality filter approach. RCFS has two different mechanisms for relevancy and redundancy analysis with two different feature selection metrics. The first metric is the symmetric uncertainty measure that weights each feature with respect to the correlation between the individual feature and target class, and then the weighted features are ranked to aid relevancy analysis. The second metric is the feature similarity measure that performs redundancy analysis using clustering technique. The proposed method also takes the advantages of k-mean clustering such as scalability and simplicity [12]. Thus, the proposed feature selection approach is scalable, simple, and more generic with less computational and space complexity. Further, it performs both relevancy and redundancy analysis for selecting optimal set of significant features from an image dataset so as to improve the performance of the classifiers.

3. Ranking with clustering based feature selection (RCFS)

The presence of irrelevant and redundant features in a training image dataset degrades the performance of the classification algorithm by deteriorating the classification accuracy and increasing the time to build classification model [13]. Therefore, the proposed work aims to eliminate the irrelevant and redundant features in order to improve the classifier performance. The proposed RCFS algorithm is developed with two phases: (1) Removal of irrelevant features and (2) Removal of redundant features from the training dataset.

3.1 Definitions and theoretical background

In order to give a better insight to the readers about the proposed algorithm, the following definitions according to John et al. [14] are presented. Let X denote a full set of features of the dataset D, and X_i be a feature i.e. $X_i \in X$, $Z_i = X - \{X_i\}$ and $Z'_i \subseteq Z_i$. Assume z'_i, x_i and c are the assigned value of Z'_i, X_i, and the target class C ,respectively.

Definition 1. Dataset: The dataset D contains N number of features X_i and one target class C denoted as D = (X, C) where X = $\{X_1, X_2, ..., X_N\}$, where N is total number of features and C is the target class attribute of D.

Definition 2. Feature Relevancy: Feature X_i is relevant to the target class C iff there exist the assigned values z'_i, x_i and c for which $p(Z'_i = z'_i, X_i = x_i) > 0$, $p(C = c \mid Z'_i = z'_i, X_i = x_i) \neq p(C = c \mid Z'_i = z'_i)$.

Definition 3. Feature Irrelevancy: Feature X_i is irrelevant to the target class C iff it is not relevant. The redundant features contain the information which is present in other features and hence, they do not participate in producing better accuracy to predict the target class C of the unlabeled instance and make confusion to the classifier. The redundant features are defined with Markov blanket according to Yu and liu [15] as follows.

Definition 4. Markov blanket: If a feature $X_i \in X$, and $M_i \subset X (X_i \notin M_i)$, then the M_i is said to be a Markov blanket for X_i iff

$$p(X - M_i - \{X_i\}, C \mid X_i, M_i) = p(X - M_i - \{X_i\}, C \mid M_i)$$

Definition 5.Feature Redundancy: Assume a feature set X and if $X_i \in X$ a redundant feature iff it has a Markov blanket within X.

Definition 6. Symmetric Uncertainty and Information Gain: The mutual information (MI) is a nonlinear estimation that measures the mutual dependency (i.e. correlation) of a feature and the target class or the mutual dependency between the features. The *symmetric uncertainty* (SU) measure [16] was developed using MI and standardized to calculate the entropy measure among the features or between the feature and target class. Many researchers used this SU to determine the significant features for classification applications [17] [18]. In the proposed algorithm RCFS, SU is used to measure the correlation or dependency between the feature and target class as expressed in Equation (1).

$$SU(X, Y) = \frac{2 \times Gain(X \mid Y)}{H(X) + H(Y)} \tag{1}$$

where $H(X)$ denotes the entropy of a discrete random variable X. Assuming the prior probabilities $p(x)$ for the possible values of variable X, $H(X)$ is expressed as in the Equation (2).

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \quad (2)$$

The SU of two variables X_i and X_j is represented as the $\delta(X_i, X_j)$ where $i \neq j$. Gain(X|Y) represents the decrease in the entropy of Y and reveals the information about Y given X. This is known as information gain (IG) which is expressed in Equation (3)

$$\begin{aligned} Gain(X|Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned} \quad (3)$$

where $H(X|Y)$ denotes the conditional entropy that represents the uncertainty (i.e. remaining entropy) of a given random variable X with the known value of the other random variable Y. Assume $p(x)$ is the prior probability for the possible values of X and $p(x|y)$ is the posterior probability of the random variable X given the values of Y, then $H(X|Y)$ is expressed as in Equation (4).

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log_2 p(x|y) \quad (4)$$

The specialty of the IG measure is that it is symmetric in nature, i.e., IG value of X after determining Y is the same as IG value of Y after determining X. Therefore, the IG measure is not affected by the order of a pair of variables i.e., $IG(X,Y) = IG(Y,X)$. IG tolerates the bias caused due to many variables by normalizing the values between 0 and 1. The value 1 denotes that the known value of a variable absolutely predicts the value of another variable. The value 0 denotes that the two variables are independent.

It is worth mentioning that the relevant features do have a strong correlation with the target class (i.e. dependent on the target class), but the redundant features do not. However, the redundant features have a correlation with other features (i.e. dependent on the feature). Hence, the irrelevant and redundant features are identified and eliminated. The C- Relevancy and F-Relevancy are defined as follows:

Definition 7. C-Relevancy: The relevancy (i.e. correlation) between the feature X_i and the target class C, is denoted as $\alpha_i = SU(X_i, C) = \frac{2 \times Gain(X|C)}{H(X) + H(C)}$ where α_i is the C-Relevancy weight of feature X_i such that $i \leq N$, where is the total number of features of dataset D.

Definition 8. F-Similarity: The similarity between any two features X_i and X_j based on the Euclidean distance measure that is denoted as $\beta_{ij} = (X_i - X_j)(X_i - X_j)'$ where $i \neq j$.

3.2 RCFS Algorithm

The proposed RCFS algorithm is designed with two phases. In the first phase, the irrelevant features are removed by selecting the relevancy measure symmetric uncertainty with a threshold value from a given image dataset D. In the second phase, the selected relevant features are grouped as K number of clusters, then the cluster-representative-features are selected from each cluster using the relevancy measure. Thus, the selected significant features are obtained by combining the cluster-representative-features of each cluster.

Algorithm: RCFS
Input: D, K, Γ // D is image dataset, K is number of clusters and Γ is the number of features to be selected
Output: S_{list} // S_{list} contains Γ number of selected significant features from image dataset D

====== Phase 1 Removal of irrelevant features ======
Steps:
-- *Calculate the C-Relevancy weight of each* X_i *of* D --
(1) **begin**
(2) **for** i=1 to N **do begin** // N is the total number of features in D
(3) α_i=SU(X_i,C) // Calculate C-Relevance weight α_i of i^{th} feature X_i
(4) **store** i and α_i in the array SX as SX[i, α_i] // i is the feature index
(5) **end for**
-- *Rank all the* X_i *of D based on C-Relevancy weight* --
(6) RX= Rank(SX) //Sort the feature index i based on the weight α_i in descending order
-- *Select top* β *number of ranked* X_i *of* D *as C-relevant features for feature-cluster formation* --
(7) **for** j=1 to β **do begin**
(8) R_{list}[j]=$X_{RX[j]}$
(9) **end for**
//Select top β numbers of features as relevant features from the sorted array RX and append them to R_{list}
====== Phase 2 Removal of redundant features ======
-- *Feature-cluster formation* --
(10) **randomly choose K features** from R_{list} as the initial cluster- centers
(11) **compute F-similarity** between each feature and each cluster-center.
(12) **do batch_ update** //Assign each feature to the cluster with the maximum F-similarity to the cluster-center.
(13) **do online_ update** //Individually assign features to a different cluster-center if the reassignment decreases the sum of the within-cluster F-similarities, sum-of-squares of feature to cluster-center F-similarities.
(14) **compute the average of the features** in each cluster to obtain K new cluster-center locations.
(15) **do again** step 10 to 13 until no change in cluster assignments or the stopping criteria are met.
(16) **return FC** //FC={fc_1,fc_2,...fc_K} where fc_i is i^{th} feature-cluster (i.e fc_i is the array of features of i^{th} cluster)

-- Select the C-relevant class-representative-feature from each feature-cluster fc_i *of* FC --

(17) **calculate C-Relevancy weight** of each X_i of each fc_i (using step 1 to step 9)

(18) **If** (each fc_i has same number of features)

(19) **then** select the top ranked λ number of features from any one of the fc_i and select the top ranked η number of features from other fc_i and append to $S_{list.}$ // η=floor(Γ/K) and λ= η +Γ– (η×K)

(20) **else** select the top ranked λ number of features from the fc_i which has more number of features compared to others and select the top ranked η number of features from other fc_i and append to $S_{list.}$

(21) **return S_{list}** // S_{list} is the selected Γ number of significant feature from D.

3.2.1 Algorithm description

This algorithm receives the training image dataset D which contains features from X_i to X_N where N is the total number of features with the target class C. C-Relevancy weight α_i of each feature X_i of D is calculated (Definition 7) and these weighted features are kept in a two dimensional array SX with their α_i. The weighed features in SX are ranked based on their weight α_i and stored in RX. The top β number of ranked features from RX are selected and stored in $R_{list.}$ These selected features are known as selected C-relevant features (The features most relevant to the target class than other features) of the original image dataset D. Thus, the irrelevant features are removed from the original image dataset D.

Secondly, to remove the redundant features from the C-Relevant features R_{list}, K number of feature-clusters fc_i (group of similar features) are formed from the R_{list}. In order to form the feature-clusters, K features are chosen at random from R_{list} as the initial cluster-centers, and then the F-similarity (Definition 8) is calculated between the feature and cluster-center for all features to each cluster-center. Then, Batch_update is preformed i.e., assignment of each feature to the cluster with the maximum F-similarity to the cluster-center. Online_update is then performed by independently assigning features to a different cluster-center if the reassignment decreases the sum of the within-cluster F-similarity and sum-of-squares feature to cluster-center F-similarity. The average of the features is computed in each cluster to obtain K new cluster-center locations. This process is repeated until no change in cluster assignments or the maximum number of iterations is reached. Thus, the K number of feature-clusters fc_i are combined together as FC where $FC=\{fc_1,fc_2,...,fc_k\}$ and $fc=\{X_{1k},X_{2k},...,X_{nk}\}$ and X_{nk} is the n^{th} feature of the k^{th} feature-cluster of FC.

In order to select the C-relevant cluster-representative-features (the features more relevant to the target class than other features in a feature-cluster) from each feature-cluster, the C-relevancy thresholds η and λ are used where η = floor (Γ/K) and λ = η +Γ– (η×K). Initially, all the fc_i are checked whether all the fc_i have same number of features or not. If so, select the λ number of features from any one of the fc_i and η number of features from other fc_i and append to S_{list}. If not, select λ number of features from the fc_i which has the maximum number of features than other fc_i and η number of features from other fc_i and appended to S_{list} where $S_{list}=\{X_1,X_2,...,X_\Gamma\}$ is a set of selected Γ number of significant features from the given training image dataset D.

3.3 Time complexity analysis

The major computational effort of the proposed RCFS algorithm involves the computation of C-Relevancy and F-Similarity for a given image dataset D. The first phase of the algorithm has a linear complexity O(N) in terms of the number of features N, assuming that β features (1≤ β ≤ N) are selected as C-relevant features in the first phase. In the second phase, K number of feature-clusters are formed from β number of C-relevant features and the complexity involved is O(nβKI) where n is the number of objects, β is the number of C-relevant features to be clustered, K is the number of feature-clusters, I is the number of iterations until there occurs convergence of clustering. Then, Γ number of C-relevant cluster-representative-features are chosen from all the K feature-clusters with the complexity of O(β) where the β is the number of C-relevant features that form K clusters. Thus, when 1<K≤N, the complexity of the algorithm is O(N+β (nKI+1)). In phase 1, β is calculated as $N/\log_{10}N$, and hence the complexity of the algorithm becomes $O(N+(N/\log_{10}N)(nKI+1))$. Since nKI>>1, the complexity can be reduced to $O(N+((N/\log_{10}N)(nKI))) = O(N(1+(nKI/\log_{10}N))) \approx O(NnKI/\log_{10}N)$. It is obvious that the algorithm has linear complexity in terms of the number of features N, number of objects n, number of feature-clusters K, and the number of iterations I. In general, K and I are far less when compared to N and n. Hence, the complexity can be further approximated as $O(Nn/ \log_{10}N)$. Assuming N>>n for high-dimensional data, the computational complexity involved reduces to $O(N/\log_{10}N)$.

4. Implementation and experimental setup

The proposed method is implemented and the experiments are conducted using the MATLAB12b software environment with the system specification as Processor: Intel® Core™ 2 CPU T5300 @ 1.73GHz, Memory (RAM): 4 GB and Operating system: 32-bit Windows vista Home Premium. The performance of the proposed RCFS is analyzed on the high-dimensional image datasets of face detection and object recognition as tabulated in Table 1.

Table 1. Details of image datasets

Dataset	Features	Instances	Classes
ORL10P[1]	10304	100	10
PIX10P[1]	10000	100	10
PIE10P[1]	2420	210	10
AR10P[1]	2400	130	10
ORL_32x32[2]	1024	400	40
Yale_64x64[2]	4096	165	15
COIL20[2]	1024	1440	20

1 = http://featureselection.asu.edu/datasets/
2= http://www.cad.zju.edu.cn/home/dengcai/Data/

Three different classification algorithms namely probabilistic-based Naïve Bayes (NB) classifier [19], tree-based classifier J48 [20], and instance-based classifier (kNN) [21] are used to evaluate the performance of RCFS in terms of accuracy and runtime. Further, the performance of the proposed RCFS is compared with the 6 state-of-the-art feature selection methods as discussed in Section 2 namely OneR [7], ReliefF [6], MIFS [8], CMIM [9], CIFE [10], and CondRed [11].

4.1 Experimental procedure

In RCFS, the C-relevancy threshold β is set as $N/10\log_{10}N$ features where N is the total number of features present in a training image dataset D, as presented in the literature [22]. For the proposed method RCFS, the number of clusters K is set as 4. For the experimental steps, initially the dataset is given to the respective features selection method and the Γ numbers of features are selected then the runtime is noted and the selected features are given to the classification algorithm for obtaining the accuracy. In order to get the stable results, the entire experimental steps are repeated 5 times and the obtained results of runtime to select the Γ number of significant features by each feature selection method and accuracy of each classifier for Γ number of selected significant features are averaged for each image dataset. The average classification accuracy is obtained by 10 runs of the classification algorithm with the 10-fold cross validation procedure.

4.2 Sensitivity analysis

In order to statistically analyze the performance of the state-of-the-art feature selection methods on the image datasets in terms of classification accuracy, the following statistical tests are conducted. The Friedman test [23] is performed with the null hypothesis that "*all the methods perform equivalently*". If the result of Friedman test p=0, at α=0.10 where p is the probability that the null hypothesis is accepted and α is the level of significance, the null hypothesis rejected, that means the algorithms compared are statistically different. In that case, the post-hoc Nemenyi test [24] is conducted to identify the algorithms that statistically differ in performance with α=0.10. The Nemenyi test ranks the feature selection methods with mean value and compares the methods in pairwise based on the critical distance (CD) as given in Equation (5)

$$CD_\alpha = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{5}$$

where q_α is fixed as the Studentzed range statistic divided by $\sqrt{2}$. If any two methods are significantly different to each other, then both of them can not lie within the critical distance of average rank.

5. Experimental study

In order to observe the overall performance of RCFS in terms of classification accuracy and runtime compared to other state-of-the-art feature selection methods, the experiments are conducted on the entire image datasets listed in Table 1.

5.1 Experimental results

The average accuracy of NB, J48, and kNN classifiers and runtime of various feature selection methods are shown in Fig. 1 and Fig. 2, respectively. The accuracy of NB, J48, kNN classifiers on all the datasets with respective number of selected features with the feature selection methods are shown in Fig. 3 to Fig. 5.

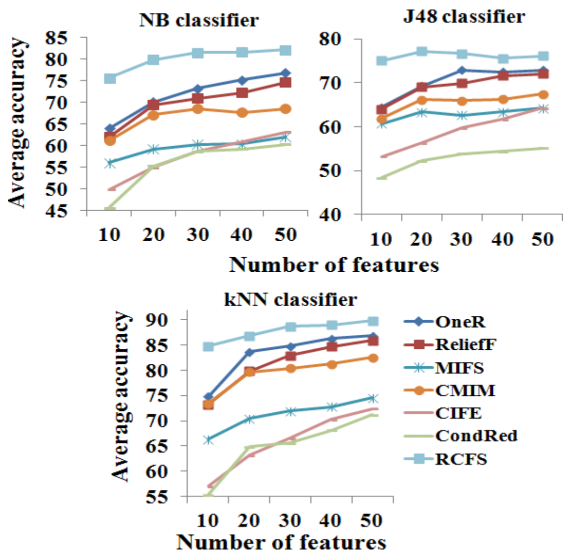


Fig. 1. Average classification accuracy of the state-of-the-art feature selection methods with various numbers of selected features from all the datasets (a) NB (b) J48 (c) kNN.

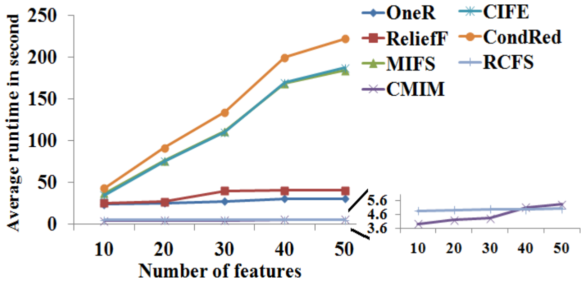


Fig. 2. Average runtime of all the state-of-the-art feature selection methods against the number of selected features on all the datasets

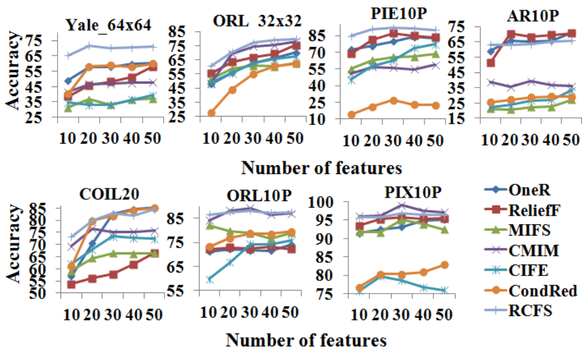


Fig. 3. Accuracy of NB classifier on all the datasets with respective number of selected features with the feature selection methods

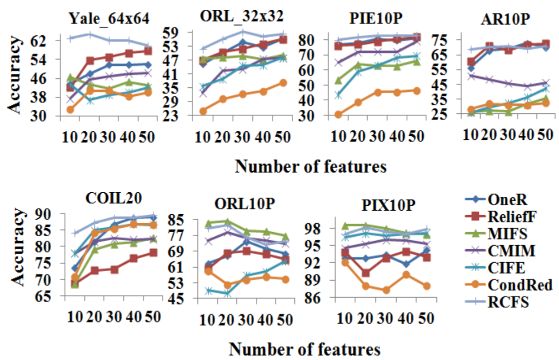


Fig. 4. Accuracy of J48 classifier on all the datasets with respective number of selected features with the feature selection methods

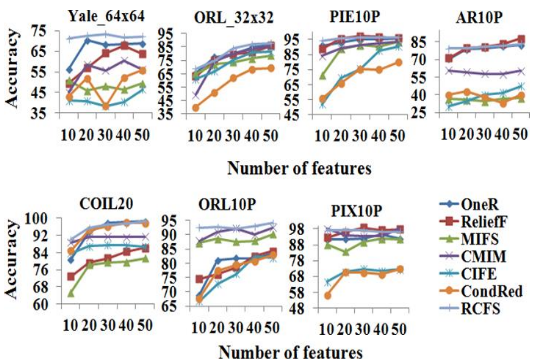
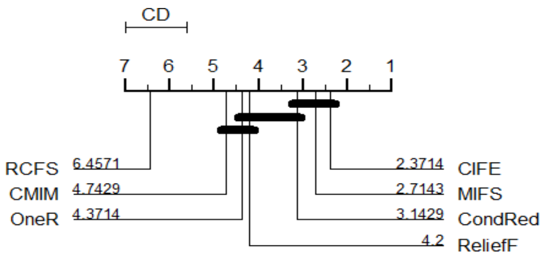


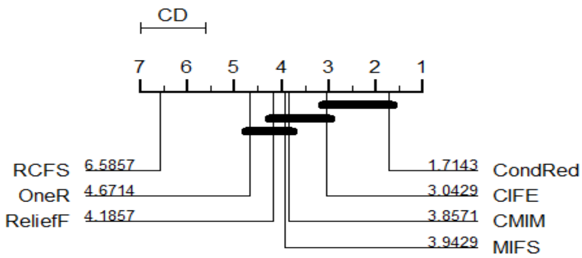
Fig. 5. Accuracy of kNN classifier on all the datasets with respective number of selected features with the feature selection methods

To analyze the statistical significance between the RCFS and other methods in terms of classification accuracy, the Friedman test is conducted on the accuracy of NB, J48, and kNN classifiers and runtime of the number of selected features varying from 10 to 50 in steps of 10 from all the image datasets with the other state-of-the-art feature selection methods.

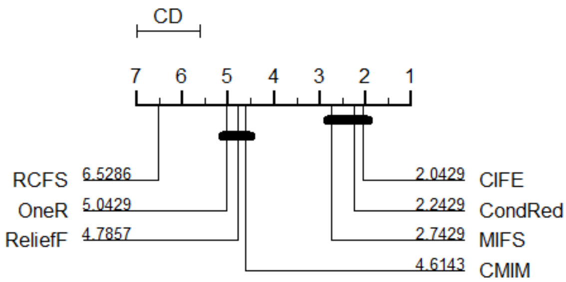
The results of Friedman test is p = 0 at α= 0.10 on the classifier accuracy of NB, J48, and kNN. Hence, the null hypothesis is rejected meaning that all the state-of-the-art feature selection methods significantly differ from each other in terms of classification accuracy. Therefore, post-hoc Nemenyi test is conducted on accuracy of NB, J48, and kNN classifiers with all the methods on all datasets, the results are recorded in Fig. 6, and the CD = 1.39 with α = 0.10.



(a)



(b)



(d)

Fig. 6. Comparison of the state-of-the-art feature selection methods with the Nemenyi test in terms of classification accuracy on all the datasets with number of selected features 10, 20, 30, 40, and 50 (a) NB (b) J48 (c) kNN classifiers

5.2 Discussion

From the Fig. 1, it is observed that the proposed method produces higher average accuracy compared to all other state-of-the-art feature selection methods for NB, J48 and kNN classifiers and CondRed is poorly performing for all the classifiers. From Fig. 2, it is evident that the proposed method takes less average runtime than all other methods except CMIM, however CMIM takes more time when the number of features are increased and unfortunately the CMIM is under performer in producing the accuracy as it is evident from Fig. 1 that CondRed takes more time compared to all other methods in terms of average runtime.

From the Fig. 3 to Fig. 5, it is observed that the proposed method consistently produces better accuracy with NB, J48 and kNN classifiers, respectively compared to other methods for all the datasets.

From Fig. 6 (a), (b), and (c) it is observed that the proposed method significantly differs from all other state-of-the-art methods compared in terms of classification accuracy for the classifiers NB, J48, and kNN respectively since it lies under separate CD for all the classifiers and obtains the first rank for the classifiers NB, J48, and kNN . From Fig. 6 (a) and (b) there is no statistical difference among the feature selection methods for classifiers NB and J48 except the proposed method since they lie on the over lapping CD. From Fig. 6 (c), it is obvious that there is no statistical difference among the feature selection methods OneR, ReliefF, CMIM in terms of classification accuracy with kNN classifier since they all lie on the same CD. There is no significant difference among the feature selection methods CIFE, CondRed and MIFS in terms of classification accuracy with kNN classifier since they all lie on the same CD.

6. Conclusions and future work

This paper proposed a novel ranking with clustering based feature selection (RCFS) for image classification. The performance of the proposed method is tested on various real-world image datasets and also its performance is compared with various state-of-the-art feature selection methods. From the experimental results, it is evident that the proposed method outperforms all other methods compared in terms of accuracy and runtime. In future, this work can be extended with different mechanisms for redundancy analysis.

References

[1] A. Luna, J. C. Vilanova, L. C. Hygino, S. E. Rossi, Functional Imaging in Oncology, Springer (2014).
[2] M. Richter, J. Beyerer IEEE Winter Conference on WACV, 2014, p. 123.
[3] G. Forman, J Mach Learn Res., 2003, p.1289.
[4] H. Peng, F. Long, C. Ding, IEEE Trans Pattern Anal Mach Intell.1226 (2005).
[5] K. Kira, L. Rendell A, Proc. of the Tenth National Conference on AI, 129 (1992).
[6] R. Šikonja, I. Kononenko, Mach Learn. 23 (2003).
[7] R. Holte, Mach Learn, 63 (1993).
[8] R. Battiti, IEEE Trans Neural Netw, 537 (1994).
[9] F. Fleuret, J Mach Learn Res., 1531 (2004).
[10] X. Tang, D. Lin, Computer Vision, ECCV, 68 (2006).
[11] G. Brown, A. Pocock, M. Zhao, M. Luján, J Mach Learn Res., 27 (2012).
[12] G. Hamerly, J. Drake, Accelerating Lloyd's Algorithm for k-Means Clustering (2014).
[13] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos. Appl Soft Comput, 136 (2015).
[14] H. John, R. Kohavi, K. Pfleger, Proc. Eleventh International Conference on Machine Learning, 121 (1994).
[15] L. Yu, H. Liu, Proc. of the Tenth ACM SIGKDD, 737 (2004).
[16] Tabakhi, Sina, P. Moradi, F. Akhlaghian. Eng Appl Artif Intel, 112 (2014).
[17] C. Freeman, D. Kulić, O. Basir, Pattern Recogn, 1812 (2015).
[18] R. Urias, S. Barigye, Mol Divers, 305 (2015).
[19] S. Atienzar, J. Martínez , J. Barba, J. Puerta, L. Ossa, Expert Syst Appl, 1531 (2015).
[20] C. Catal, S. Tufekci, E. Pirmit, G. Kocabag, Appl Soft Comput, 156 (2015).
[21] C. Mao, B. Hu, P. Moore, Y. Su, M. Wang, Advances in Knowledge Discovery and Data Mining, (2015).
[22] S. Qinbao, N. Jingjie, G. Wang. "IEEE Trans. Knowl. Data Eng., 1 (2013).
[23] G. Geoffrey Zhang, Radiat Oncol, 152 (2011).
[24] I. Brown, C. Mues, Expert Syst Appl, 3446 (2012).

*Corresponding author: asirantony@gmail.com