Network Modeling Analysis in Health Informatics and Bioinformatics (2021) 10:39
https://doi.org/10.1007/s13721-021-00313-7

ORIGINAL ARTICLE

A hybrid feature selection model based on improved squirrel search algorithm and rank aggregation using fuzzy techniques for biomedical data classification

Gayathri Nagarajan[1] · L. D. Dhinesh Babu[1]

Received: 14 October 2020 / Revised: 30 April 2021 / Accepted: 6 May 2021 / Published online: 2 June 2021

Abstract
Feature selection has gained its importance due to the voluminous nature of the data. Owing to the computational complexity of wrapper approaches, the poor performance of filtering techniques, and the classifier dependency of embedded approaches, hybrid approaches are more commonly used in feature selection. Hybrid approaches use filtering metrics to reduce the computational complexity of wrapper algorithms and are proved to yield better feature subset. Though filtering metrics select the features based on their significance, most of them are unstable and biased towards the metric used. Moreover, the choice of filtering metrics depends largely on the distribution of data and data types. Biomedical datasets contain features with different distribution and types adding to the complexity in the choice of filtering metric. We address this problem by proposing a stable filtering method based on rank aggregation in hybrid feature selection model with Improved Squirrel search algorithm for biomedical datasets. Our proposed model is compared with other well-known and state-of-the-art methods and the results prove that our model exhibited superior performance in terms of classification accuracy and computational time. The robustness of our proposed model is proved by conducting experiments on nine biomedical datasets and with three different classifiers.

Keywords Hybrid feature selection · Biomedical data classification · Linguistic fuzzy modeling · Rank aggregation

1 Introduction

A huge amount of data is generated in biomedical datasets and one of its biggest challenges is its volume. The term 'Volume' and 'Value' of data are always not directly proportional. Hence, there is a need to understand the value of the data in the preprocessing step and discard the features that do not add any value to it. Dimensionality reduction is the process of reducing the dimensional space of datasets by feature selection or feature extraction techniques. While feature selection identifies the features that do not contribute to the classification and discard them, feature extraction transforms the data in high dimensional space into low dimensional space. Feature selection is generally preferred over feature extraction as it does not alter the data and hence offers better interpretability. Feature extraction loses the relation with the original features and also suffers from computational complexity despite resulting in irrelevant and redundant features (Senawi et al. 2017). These techniques are preferred when only discrimination is required (Jain et al. 2000). Hence, feature extraction may not be a good choice for biomedical datasets when interpretation is required.

Feature selection techniques are broadly classified into three categories viz., Filter approaches, Wrapper approaches, and embedded approaches. Filtering uses statistical metrics like Information Gain (IG), entropy, consistency-based measures, and correlation metrics to evaluate the features. Wrapper approaches use induction algorithms to select good feature subsets based on the accuracy of the classifier used (Kohavi and John 1997) while embedded approaches learn features during the construction of the model. Hybrid approaches are also developed where filtering metrics are used initially followed by wrapper or embedded approaches. Different feature selection techniques and their drawbacks

✉ Gayathri Nagarajan
gayunagarajan1083@gmail.com

L. D. Dhinesh Babu
lddhineshbabu@gmail.com

[1] School of Information Technology and Engineering, VIT university, Vellore, India

are discussed in Shardlow (2016) and Guyon and Elisseeff (2003). High dimensional datasets like genomic datasets are quite common in biomedical datasets. Learning model for such high dimensional datasets has to be constructed carefully as they are more prone to overfitting. Feature selection approaches such as wrapper approaches suffer from computational complexity. It involves more subset evaluation by a repeated combination of features, and checking for the accuracy of the classifier. Embedded approaches are biased towards the learning algorithm and lack generalization capability. Lot of research works help to reduce the computational burden of the wrapper and embedded approaches through heuristic and optimal search algorithms. Yet, these approaches are still considered to be computationally expensive for high dimensional biomedical datasets. Filtering approaches score well in terms of computational and statistical scalability (Inza et al. 2007). They are independent of the learning algorithm used. The result of filtering approaches can be used for different mining algorithms. Yet, filtering techniques are based on specific statistical measures they use and hence are biased towards the metric they use. They are also not proved to yield better feature subset and hence affect the classification accuracy when used in the learning model. Moreover, the choice of right filtering metric acts as a stand alone problem. Both classification accuracy and computational complexity are important for learning models of biomedical datasets. Hence, hybrid approaches that use filtering metrics before the wrapper approaches are preferred more commonly in biomedical datasets as both the computational complexity and classification accuracy are managed effectively.

As discussed, hybrid approaches use filtering techniques to derive the initial subset that is fed as input to the wrapper approaches. Filtering techniques are classified into univariate and multivariate methods. Filtering approaches can also be classified into feature weighing and feature subset selection approaches. Univariate methods consider the relevancy information but fail to consider the dependency among the features. Multivariate methods consider dependencies but are computationally slow and hence less scalable than univariate methods (Canedo et al. 2013). Few commonly used univariate methods include IG, Gain Ratio (GR), Mutual Information (MI), Symmetric Uncertainty (SU), relief, correlation, chi-square, and Fisher Score (FS). Few commonly used multivariate methods include Minimum Redundancy Maximum Relevance (mRmR), correlation-based Feature Selection (CFS), Fast Correlation-Based Filter (FCBF), and feature selection based on clustering. These methods can further be categorized into information-based, distance-based, consistency-based and correlation-based filtering approaches depending on the statistical measures used. Feature weighing approaches rank the features based on the statistical measure (e.g., IG) whereas feature subset selection filtering approaches evaluate the subset of features and obtain the best subset based on the statistical measure (e.g., CFS).

Though filtering approaches reduce the complexity in high dimensional biomedical datasets, they suffer from stability as they are based on the specific metric used. Also, most of the univariate filtering approaches fail to consider redundancy information among the features as discussed earlier. Moreover, they are sensitive to different variations in data and the distribution of training data especially in high dimensional space (Yang et al. 2012) such as genomic datasets. The use of multivariate methods is also computationally expensive in hybrid techniques. It is difficult to choose the best filtering method for a particular problem and dataset. This problem is called selection trouble. There is no solution than trying with all the methods and apply the best (Waad et al. 2014). Hence, to ensure the stability of the filtering technique and to relieve the user from the burden on the choice of the right filtering metric, rank aggregation approaches with univariate filters are used before wrapper algorithms in hybrid feature selection techniques. Rank aggregation methods take advantage of individual selectors, overcome their weakness, and also introduce diversity. Although several rank aggregation approaches are used, our method performs rank aggregation using fuzzy systems that are flexible and handles uncertainty. Our model also deals with both relevancy and redundancy information. Hence, our work proposes a Fuzzy-Based Rank Aggregation (FBRA) filter before using wrapper algorithm, Improved Squirrel Search Algorithm (ISSA) for feature selection in biomedical datasets. ISSA is proved to be one of the best optimization algorithms when compared with other well-known algorithms (Jain et al. 2019) and hence our model uses ISSA. The major contributions of our model are as follows:

- A hybrid feature selection model is proposed using rank aggregation based on fuzzy techniques and Improved Squirrel search optimization algorithm. The filtering metrics we use in our rank aggregation model are selected based on different measures, such as correlation, distance and information theory. Redundancy and relevancy are also considered. There is no specific weighing among the individual methods used.
- The proposed model uses an efficient algorithm ISSA that solves the problem of early convergence to a local optimum.
- The proposed model is tested on biomedical datasets of different volumes with three different classifiers and is proved to be robust.

- Extensive experimentation is performed by comparing our proposed model with other different rank aggregation approaches, univariate filtering metrics and different optimization algorithms. Our proposed model exhibits superior performance in terms of classification accuracy and computational time.

The rest of the paper is organized as follows: Related work is discussed in Sect. 2. Section 3 gives a summary about the methods used in our model. Section 4 explains our proposed model H-FBRA+ISSA. Section 5 demonstrates the experimental framework and discusses the results. Section 6 discusses the limitations and scope for future work and Sect. 7 winds up with the conclusion.

2 Related work

Hybrid feature selection approaches are proved to improve the classification accuracy as discussed earlier and are the focus of our research. There are few recent works that propose hybrid feature selection approaches. Few hybrid feature selection approaches are also proposed exclusively for biomedical datasets. MI filtering with Genetic Algorithm (GA) is proposed in Hoque et al. (2014) whereas MI with Particle swarm optimization (PSO) is proposed in Han and Ren 2015. mRmR with Artificial Neural Networks (ANN) is proposed for brain tumor classification in Huda et al. (2016). mRmR with Artificial Bee Colony (ABC) algorithm is proposed for cancer classification in Alshamlan et al. (2015). SU with harmony search algorithm is proposed for microarray data classification in Shreem et al. (2016). IG with fuzzy rough sets and GA is proposed for cancer microarray data classification in Chinnaswamy and Srinivasan (2017) whereas IG with Binary Differential Evolution (BDE) is proposed for microarray data in Apolloni et al. (2016). A hybrid approach that uses CFS filtering with iterative binary PSO is proposed in Jain et al. (2018). Different filtering metrics are used before different optimization algorithms. CFS with sequential search algorithm is proposed for breast cancer, diabetes and hepatitis data classification in Tomar and Agarwal (2015). The correlation coefficient with PSO is proposed for microarray data classification in Chinnaswamy and Srinivasan (2016). CFS with improved BPSO is proposed for cancer classification in Jain et al. (2018). FS with Ant Colony Optimization (ACO) is proposed for microarray data classification in Sharbaf et al. (2016).

Though these methods identify the feature subset, they are biased towards a single metric. Few metrics are based on correlation, few based on relevancy information, and few on the discriminative capability of the feature. Different biomedical datasets possess different characteristics and not all the metrics are expected to perform well for all kinds of biomedical datasets as discussed earlier. Indeed, the major problem with these single metrics is the overestimation of a feature's significance in spite of the fact that each statistical measure suffers from its own drawback. For example, MI is found to overestimate a feature's significance, especially when a feature is correlated with one or a subset of features, but completely independent from the rest (Bennasar et al. 2015). Filtering metrics such as IG is simple, but it assumes independency between the features, which is not always the case (Bennasar et al. 2015). Technique such as CFS is multivariate approach and is computationally complex.

To overcome this drawback of the individual filtering approaches, rank aggregation approaches are proposed. Rank aggregation approaches are used to induce diversity and make use of the advantages of the individual approaches. They also avoid the overestimation of features. But, rank aggregation is itself a stand-alone problem owing to the fact that different filtering metrics may give disjoint ranks to the same features (Waad et al. 2014). The common aggregation methods include mean, median, highest rank, lowest rank, weighed aggregation, and voting. Simple methods, such as mean and median (Wang et al. 2019a) aggregation, find the mean value and the median value of all the individual filtering metrics for a particular feature and assigns it as the final feature weight or rank for that respective feature. Though simple, they suffer from drawbacks, such as tied feature ranking and disjoint ranking (Waad et al. 2014). Approaches, such as weighed aggregation, highest rank, and lowest rank, are again biased towards a specific filtering metric for a particular feature. Simple voting (Bolón-Canedo et al. 2012; Abut et al. 2019) is an efficient method but its performance is based on the filtering metrics used and hence may yield biased results. Other aggregation methods, such as Borda and robust rank aggregation, (Najdi et al. 2016) are used, but they are again based on the mean and performance of other measures and are not proved to yield good results. An aggregation method, MeLif, based on linear combination of individual filtering metrics is proposed in (Smetannikov et al. 2016) and parameters of the classifier are tuned accordingly. But this method is classifier-dependent.

There are few other research works performed for rank aggregation. Five filtering metrics are merged in (Bolon-Canedo et al. 2014) using different classifiers. The results of the individual filters are combined by integrating them one by one and checking the accuracy of the classifiers. Though the results are stable, this method is computationally expensive, especially with high dimensional datasets such as microarray datasets as different subsets of features have to be tested on different classifiers. A distributed feature selection that computes the relevancy using ReliefF, FS and IG and redundancy using distance, correlation, and CS is proposed in (Ebrahimpour and Eftekhari 2018). Both the relevancy and redundancy are represented as clusters and the final subset is chosen based on the intersection of the clusters. Both homogenous and heterogeneous ensemble approaches are proposed in (Pardo et al. 2017). The homogenous ensemble approach uses the same filtering metric in different training data and heterogeneous ensemble approach uses different filtering metrics on the same training data. While the former method is biased towards the metric used, the latter method takes care of diversity. Five ranking methods ReliefF, mRmR, IG, SVM rank, and ANN rank are used and the results are combined using six combination methods minimum, median, mean, geometric mean, Stuart, SVM rank, and Robust Rank Aggregation (RRA). The downsides of the minimum, median, and mean are discussed earlier whereas the other methods used in this work, mRmR, SVM rank, and NN rank suffer from computational complexity with a large number of features. An ensemble feature selection method using GR, IG, ReliefF and CS is proposed in (Hoque et al. 2018). The results are aggregated using greedy search technique. This method of aggregation is again computationally expensive with a large number of features despite the fact that the results of the experiments conducted on this method are not strong and the benefits are not obvious. Multiple filters combined with fusion approach are used in (Bonilla-Huerta et al. 2015) to select the initial subset of genes for GA feature selection approach. A new ensemble feature selection approach based on Sort Aggregation (SA) is proposed in (Wang et al. 2019a) and a new ensemble technique Majority Voting Feature Selection (MVFS) is proposed in (Abut et al. 2019). SA method uses three-feature ranking methods—Maximum information coefficient, XGBoost and chi-square. Arithmetic and geometric means are used for sorting the results. MVFS uses three-feature ranking methods—Relief-F, mRmR and Maximum Likelihood Feature Selection (MLFS). It uses majority voting or correlation score or assigns priority to a particular ranking method to obtain the final feature rank.

The major drawback with these aggregation filtering approaches is: few of them use univariate methods and aggregate them without considering the dependency information while few other methods use techniques like greedy search, GA or clustering that are computationally expensive with high dimensional datasets. Few methods also suffer from disjoint ranking and tied ranking problems. Biomedical datasets are voluminous with many lakhs of patient's observations measuring many thousands of their information especially when genetic information is recorded. Moreover, different information would be recorded for different purposes. For example, the prediction of cancer requires different information when compared with the information required for the prediction of respiratory disease. The changes are over legislation too. Few features are categorical whereas few others are continuous. Hence, the aggregation method used for such datasets should be unbiased, computationally efficient and should not suffer from problems, such as disjoint ranking and tied ranking. To cope with this need, we propose the idea of aggregating filtering metrics using fuzzy systems that are flexible and extensible (Tal and Muntean 2012). Moreover, the use of fuzzy systems for rank aggregation is also not computationally complex. Hence, a structure-free dynamic aggregation approach can be established with fuzzy systems. Indeed, fuzzy systems are well suited for imprecise data and represent knowledge with uncertainty (Nguyen et al. 2018). In spite of the underlying fact that several aggregation approaches are proposed, very few works are carried out using fuzzy systems for aggregation. Our work incorporates the idea of using fuzzy system for rank aggregation. This, in turn, is fed as input to the wrapper algorithm ISSA. Squirrel Search Algorithm (SSA) is a recently proposed algorithm (Jain et al. 2019) and ISSA is proposed in (Zheng and Luo 2019) that improves the global convergence capability. SSA is proved to be the best when compared with several different optimization algorithms (Jain et al. 2019) and hence is used in our proposed hybrid model with rank aggregation (FBRA). It has been observed that this hybrid approach exhibits superior performance in terms of classification accuracy, dimensionality reduction and computational time.

3 Methods

The proposed model uses ranking metrics, a linguistic fuzzy system for rank aggregation and ISSA for feature selection. This section reviews the important concepts of them.

In biomedical data classification, let 'X' represent the dataset with 'n' observations and 'm' features with a class label 'C' $\in C = \{C_1, C_2, \dots C_l\}$. 'l' takes the value of 2 in case of binary classification and > 2 in case of multi-classification. Let $X_i \in X$ represent the ith observation of 'X'. Our aim is to find a subspace of 'm' features,'S', that contributes more to the target class 'C'. Let 'F' represent the feature set $F = \{F_1, F_2, \dots F_m\}$ and 'S' represent the subset of features $S = \{S_1, S_2, \dots S_k\}$ with 'k' < 'm'.

3.1 Ranking metrics

As discussed earlier, there are different feature ranking metrics each with its own strengths and weakness. For example, MI is the most commonly used ranking metric. This is due to the fact that it is a good information-based measure as it does not assume linearity between the variables and can work with both categorical and numerical variables (Wang et al. 2015). Yet, MI calculates redundancy by estimating the MI of the feature with the selected subset but forgets to measure the MI between the feature and the class label which may affect the model accuracy (Bennasar et al. 2015). Maximization of minimum criteria-based feature selection using MI is also found to be unstable (Bennasar et al. 2015). ReliefF is attractive as it has low bias and can capture local dependencies among the features that other methods miss. It is also stable but high stability does not always imply high accuracy in classification. On the other hand, methods like consistency-based and correlations based yield good performance but are not stable (Bolón-Canedo et al. 2012). Other metrics, such as IG (Bolón-Canedo et al. 2015), Entropy and Fisher score are also used. IG, as discussed earlier is simple but it assumes independency between the features which is not always true (Bennasar et al. 2015). Correlated features, on the other hand, are redundant and fail to provide mutual information that helps in data mining tasks (Wang et al. 2015). Few novel metrics are also proposed. For example, a novel feature ranking metric, maximum relevancy maximum distance is proposed for bioinformatics data classification (Zou et al. 2016). But this method does not consider the discriminative power of the features. Chen et al. (2018) uses Subspace clustering for feature weighting. Yet, it suffers from the downside that the clustering results depend largely on the initial cluster centers and other parameters. Univariate filtering feature selection approaches MI, ReliefF, and autocorrelation are compared with multivariate approach CFS in (Koprinska et al. 2015) for electricity load forecasting. The results state that there is no significant difference in accuracy. Yet, this result cannot be generalized as their work focusses only on electricity load dataset.

Our model uses three different categories of metrics, Correlation-Based (CB), Distance-Based (DB) and Information theory-Based (IB) measures. The metrics we use for our proposed model are correlation, neighborhood-based quality of information, rough MI and component co-occurrence information. Though it has its own limitations, correlation is found to be one of the best metrics in identifying the relationship between the features and the class label. It measures the dependency between the features that help to identify both the relevant and redundant features (Hsu and Hsieh 2010). Feature-class correlation and Feature-feature correlation help to identify the important features. Few recent works using correlation filtering in feature selection include (Dahiya et al. 2016; Hsu and Hsieh 2010; Low et al. 2016; Kim and Chung 2017; Zou et al. 2016; Xu et al. 2016; Senawi et al. 2017). Low et al. (2016), Kim and Chung (2017), Zou et al. (2016), Xu et al. (2016) and Senawi et al. (2017) are tested on biomedical datasets. Most of the correlation methods are designed to work on a particular data type. As biomedical dataset consists of different data types, our model computes the correlation between the nominal and continuous data types as (Senawi et al. 2017)

$$CB(F_a, C) = \left(1 - \frac{E(var(F_a|C))^{\frac{1}{2}}}{var(F_a)}\right) \tag{1}$$

where $F_a \in F$, $E[var(F_aC)]$ is the expected value of conditional variance that represents the average variability within outcomes. The correlation between nominal variables is computed using the chi-squared test.

Our second metric, neighborhood-based quality of information is a new metric proposed by Liu et al. (2017). Unlike other distance-based metrics that calculate the distance between the features and then rank the features, this metric calculates the discriminative ability of a feature. The concept of maximum nearest neighbor is introduced by calculating the maximum nearest neighbor entropy, conditional entropy and joint entropy. Any two samples having distance less than the maximum nearest neighbor and belonging to the same class are consistent, otherwise they are inconsistent. Hence, two samples belong to different classes if there exists a feature that can differentiate the maximum nearest neighbors of these two samples. Thus, a formula is derived for the quality of a feature. This metric is tested with many filtering metrics in different classifiers and is proved to be one of the best metrics. Few recent works using this metric in their experiments include (Liu et al. 2018; Zheng et al. 2019; Suo et al. 2019).

The neighborhood-based quality of a feature $F_a \in F$ where $c \in C$ and $X[c]' = X - X[c]$, is given by (Liu et al. 2017)

$$Q(F_a, C) = \frac{MH_{\eta(F_a|c)}}{count_{\eta(F_a, X[c]')}} * MH_{\eta(F_a)} \tag{2}$$

where

$$MH_{\eta}(F_a|c) = -\frac{1}{n}\sum_{i=1}^{n} log\frac{\| \eta_{F_aUC(X_i)} \|}{\| \eta_{C(X_i)} \|} \tag{3}$$

$$MH_{\eta(F_a)} = -\frac{1}{n}\sum_{i=1}^{n} log\frac{\| \eta_{F_a(X_i)}}{n} \tag{4}$$

and $count_{\eta(F_a,X[c]')}$ is the number of observations that can be distinguished from sample X_i by F_a, $\eta(X_i)$ is the maximum nearest neighbor of X_i over feature space F given by

$$\eta(X_i) = \{X_i' | \delta F(X_i, X_i') <= d(X_i), X_i \in X\} \tag{5}$$

where $d(x) = \max(\delta(x, NM(x)), \delta(x, NH(x)))$. NM($x$) is called the nearest miss and depicts the nearest observation that belongs to different class of x, NH(x) is called the nearest hit that depicts the nearest observation that belongs to the same class of x. $\delta(x, NM(x))$ and $\delta(x, NH(x))$ are the distance between x and its nearest observation that belongs to the different class and the same class, respectively.

Our third metric is MI. Though MI has its own limitations, it is robust to noise and is not limited to linear dependencies. On the other hand, rough sets theory, introduced by Pawlak (1982), handles imprecision and uncertainty. The rough set concept is introduced into MI to handle the uncertainty of knowledge that cannot be handled by Shannon's entropy(Shannon 1948), one of the commonly used metric. Moreover, rough sets are known for reducing the redundancy by preserving the discrimination power of the dataset. Few works using rough set concepts with information measures include (Zeng et al. 2014; Maji and Pal 2009; Foitong et al. 2009; Yang et al. 2014; Qian and Liang 2008). Rough sets concept used with MI for feature selection is found to improve the cancer classification accuracy (Xu et al. 2009).

The rough MI between two features $F_a, F_b \in F$ is given by Zeng et al. (2014)

$$\mathrm{RI}(F_a, F_b) = \frac{1}{n}\sum_{i=1}^{n} log\frac{n \mid [X_i]_{F_aUF_b} \mid}{\mid [X_i]_{F_a} \mid\mid [X_i]_{F_b} \mid} \tag{6}$$

where $[X_i]_{F_a}$ represents the equivalence class of X_i with respect to the feature F_a, $[X_i]_{F_b}$ represents the equivalence class of X_i with respect to feature F_b and 'U' represents union operation.

Our fourth metric, the Component Cooccurrence Feature Information (CCFI) is proposed by Wang and Feng (2018). It overcomes the disadvantages of MI and other similarity measuring metrics, such as Cosine Similarity (CS) and Jaccard Similarity (JS). MI does not normalize the output values and it fails to handle the case when the probability of the component values of both the features is zero. CS and JS ignore the fact that two features with different component values on each dimension may be highly relevant. The CCFI overcomes these drawbacks by taking into account the conditional occurring probabilities of two feature components than considering the vector similarity information. Recent work using CCFI includes (Wang and Feng 2019). CCFI between two features $F_a, F_b \in F$ is given by Wang and Feng (2018)

$$\mathrm{CCFI}(F_a, F_b) = \sum_{F_{bc}\in\omega, F_{ac}\in\omega} (p(F_{ac}|F_{bc}) * p(F_{bc}|F_{ac}) * p(F_{ac}, F_{bc})) \tag{7}$$

where F_{ac} and F_{bc} represent the possible values of F_a and F_b, respectively; $p(F_{ac} \mid F_{bc})$ represents the conditional probability that F_{ac} occurs when F_{bc} exists; $p(F_{bc}|F_{ac})$ represents the conditional probability that F_{bc} occurs when F_{ac} exists; $p(F_{ac}, F_{bc})$ is the feature component-based normalization coefficient which represents the probability that F_{ac} and F_{bc} exist together in the dataset.

$$p(F_{ac}|F_{bc}) = \frac{n(F_{ac}, F_{bc})}{n(F_{bc})} \tag{8}$$

$$p(F_{bc}|F_{ac}) = \frac{n(F_{bc}, F_{ac})}{n(F_{ac})} \tag{9}$$

$$p(F_{ac}, F_{bc}) = \frac{n(F_{bc}, F_{ac})}{n} \tag{10}$$

where $n(F_{ac})$ represents the number of observations in which F_{ac} exists; $n(F_{bc})$ represents the number of observations in which F_{bc} exists; $n(F_{bc}, F_{ac})$ denotes the number of observations in which F_{ac} and F_{bc} exists together.

3.2 Linguistic fuzzy modeling

Fuzzy rule-based systems are extensions of classical rule-based systems with their antecedents and consequents composed of fuzzy logic statements (Fernandez et al. 2017). It finds its application in uncertain and imprecision problems. Fuzzy Rule-Based Classification Systems (FRBCS) is composed of the inference system and knowledge base. The knowledge base is composed of membership functions of fuzzy partitions associated with the input features. The rule base is composed of fuzzy rules (del Río et al. 2015). Fuzzy rules are of the following form.

$$\begin{aligned} RuleR_j :\ IF \quad x_1 \quad is \quad A_j^1 \quad AND \\ x_2 \quad is \quad A_j^2 .... \\ AND \quad x_n \quad is \quad A_j^n \\ THEN \quad Class = C_j \\ with \quad RW_j \end{aligned} \tag{11}$$

where $x_1, x_2, ..x_n$ are the inputs, R_j is the jth rule, A_j^1, A_j2 ..... A_j^n are the linguistic labels of the fuzzy sets. The IF part constitutes the antecedent, THEN part constitutes the consequent and RW_j is the rule weight of rule R_j.

The rule weights are computed for each rule and the winner rule is determined to derive the final class. There are different ways for computing rule weight and the winner rule is the one with the maximum weight. One of the commonly used methods for computing rule weight is penalized certainty factor.

$$\mathrm{RW}_j = \frac{\sum_{x_p \epsilon C_j} \mu A_j(x_p) - \sum_{x_p \notin \epsilon C_j} \mu A_j(x_p)}{\sum_{p=1}^{m} \mu A_j(x_p)} \tag{12}$$

There are several methods for fuzzification and defuzzification. Few common fuzzification methods include inference, intuition, rank-ordering, using neural networks, using GAs, and deduction. Few common defuzzification methods include weighted average, centroid, the center of sums, and min–max. Linguistic labels are used to represent and categorize the membership function. A linguistic fuzzy concept is introduced in Zadeh (1973). Linguistic fuzzy systems are simple and efficient rank aggregation approaches.

3.3 Squirrel search algorithm

The squirrel search algorithm is a novel nature-inspired optimization algorithm developed in Jain et al. (2019) and is proved to yield high convergence rate when compared with other swarm intelligence optimization algorithms, such as PSO, ABC, Bat Algorithm (BA), and FireFly (FF) algorithm. This algorithm imitates the dynamic behavior in locomotion of squirrels called gliding. The flying squirrels in the forest are assigned random initial location (random solutions) using uniform distribution. The user-defined fitness function corresponds to the fitness of location for each flying squirrel. The fitness value represents the optimal food source, normal food source or no food source (quality of solutions). Depending on the fitness value, few flying squirrels move towards the normal and optimal food source (Exploration). Predator presence probability is also considered during this behavior (Exploitation). Flying squirrels from normal food source move towards optimal food source using the following equation (Jain et al. 2019)

$$\mathrm{FS}_{at}^{t+1} = \begin{cases} \mathrm{FS}_{at}^{t} + d_{\mathrm{g}} G_c (FS_{ht}^t - FS_{at}^t), & R_1 \geq p_{\mathrm{dp}} \\ \text{Random location}, & \text{otherwise.} \end{cases} \tag{13}$$

where FS_{at} are the flying squirrels on acorn nut tree (normal food source), FS_{at}^{t+1} is the new location of the squirrels, d_{g} is the random gliding distance, R_1 is the random number in the range of [0,1], FS_{ht} is the location of the flying squirrel that reached the hickory nut tree (optimal food source) and t denotes the current iteration. The balance between exploration and exploitation is achieved with the help of gliding constant Gc in the mathematical model. Gc is considered as 1.9.

Flying squirrels from no food source move towards the normal food source using the following equation (Jain et al. 2019)

$$\mathrm{FS}_{nt}^{t+1} = \begin{cases} \mathrm{FS}_{nt}^{t} + d_{g} G_c (FS_{at}^t - FS_{nt}^t), & R_2 \geq p_{\mathrm{dp}} \\ \text{Random location}, & \text{otherwise.} \end{cases} \tag{14}$$

where FS_{nt} are the flying squirrels on normal tree (no food source) and R_2 is the random number in the range of [0,1]

Flying squirrels from no food source move towards the optimal food source using the following equation (Jain et al. 2019)

$$\mathrm{FS}_{nt}^{t+1} = \begin{cases} \mathrm{FS}_{nt}^{t} + d_{g} G_c (FS_{ht}^t - FS_{nt}^t), & R_3 \geq p_{\mathrm{dp}} \\ \text{Random location}, & \text{otherwise.} \end{cases} \tag{15}$$

where R_3 is the random number in the range of [0,1]. Predator presence probability P_{dp} is 0.1 in all cases.

The gliding distance is calculated using the formula

$$d_{\mathrm{g}} = \left(\frac{h_{\mathrm{g}}}{\tan\Phi} \right) \tag{16}$$

where h_{g} is the loss in height after gliding (approximately 8m—corresponds to its original paper (Jain et al. 2019)). and Φ is the glide angle.

Another concept, seasonal monitoring change is also introduced to maintain the balance between the exploitation and exploration capability of the algorithm apart from the gliding constant. This concept is based on the fact that seasonal changes and the presence of a predator can affect the locomotion path of the flying squirrels. This behavior is modeled using

$$S_c^t = \sqrt{\sum_{k=1}^{d} (FS_{at,k}^t - FS_{ht,k})^2} \tag{17}$$

where S_c is the seasonal constant and $t = 1, 2, 3$. The seasonal monitoring condition is checked if $S_c^t < S_{\min}$ where $S_{\min}$ is the minimum value of seasonal constant and is calculated as

$$S_{\min} = \frac{10E^{-6}}{(365)^{t/(t_m/2.5)}} \tag{18}$$

where t and t_m are the current and maximum iteration values, respectively. The value $S_{\min}$ affects the exploration and exploitation capabilities.

Exploration capability is further improved by randomly relocating few flying squirrels that have not explored the optimal food source but still survive. It is modeled using the equation

$$\mathrm{FS}_{nt}^{\mathrm{new}} = \mathrm{FS}_L + \mathrm{Levy}(n) \times (FS_U - FS_L) \tag{19}$$

where Levy(n) explores the search space efficiently. FS_L and FS_U are lower and upper bounds, respectively, of ith flying squirrel in jth dimension.

Function tolerance, maximum execution time or the maximum number of iterations is considered as the stopping criterion for the algorithm. Algorithm 1 depicts the algorithm for SSA.

Algorithm 1 Squirrel Search Algorithm

```
Begin
Define input parameters
Generate random locations for n number of flying squirrels
Evaluate fitness of each flying squirrel's location
Sort the locations of flying squirrels in ascending order
depending upon their fitness value
Declare the flying squirrels on hickory nut tree, acorn nuts trees and normal trees
Randomly select some flying squirrels which are on normal trees to move towards
hickory nut tree and the remaining will move towards acorn nuts trees
while the stopping criteria is not satisfied do
    for $t = 1$ to $n1$(number of squirrels on acorn trees moving towards hickory trees) do
        if $R1 >= P_{dp}$ then
            $FS_{at}^{t+1} = FS_{at}^t + d_g \times G_c \times (FS_{ht}^t - FS_{at}^t)$
        else
            $FS_{at}^{t+1}$ = a random position of search space
        end if
    end for
    for $t = 1$ to $n2$(number of squirrels on normal trees moving towards acorn trees) do
        if $R2 >= P_{dp}$ then
            $FS_{nt}^{t+1} = FS_{nt}^t + d_g \times G_c \times (FS_{at}^t - FS_{nt}^t)$
        else
            $FS_{nt}^{t+1}$ = a random position of search space
        end if
    end for
    for $t = 1$ to $n3$(number of squirrels on normal trees moving towards hickory trees) do
        if $R3 >= P_{dp}$ then
            $FS_{nt}^{t+1} = FS_{nt}^t + d_g \times G_c \times (FS_{ht}^t - FS_{nt}^t)$
        else
            $FS_{nt}^{t+1}$ = a random position of search space
        end if
    end for
    Calculate seasonal constant $S_c$
    if Seasonal monitoring condition is statisfied then
        Randomly relocate flying squirrels
    end if
    Update the minimum value of seasonal constant $S_{min}$
end while
The location of squirrel on hickory tree is the final optimal solution
End
```

An improved SSA is introduced in Zheng and Luo (2019). The pseudocode of ISSA is shown in Algorithm 2.

SSA's global convergence capability is improved in the following ways.

Algorithm 2 Improved Squirrel Search Algorithm

```
$Begin$
$Set\ Iter_{max}, NP, n, P_{dpmax}, P_{dpmin}, sf, G_c, FS_U, FS_L$
$Generate\ random\ locations\ for\ n\ number\ of\ flying\ squirrels$
$Evaluate\ fitness\ of\ each\ flying\ squirrel's\ location$
while $Iter < Iter_{max}$ do
    $[sorted\_f, sorte\_index] = sort(f)$
    $FS_{ht} = FS(sorte\_index(1))$
    $FS_{at}(1:3) = FS(sorte\_index(2:4))$
    $FS_{nt}(1:NP-4) = FS(sorte\_index(5:NP))$
    $Generate\ new\ locations$
    $P_{dp} = (P_{dpmax} - P_{dpmin}) \times (1 - \frac{Iter}{Iter_{max}})^{10} + P_{dpmin}$
    for $t = 1$ to $n1(number\ of\ squirrels\ on\ acorn\ trees)$ do
        if $R1 >= P_{dp}$ then
            $FS_{at}^{new} = FS_{at}^{old} + d_g G_c (FS_{ht}^{old} - FS_{at}^{old})$
        else
            $FS_{at}^{new} = C_x(FS_{at}^{old}, En, He)$
        end if
    end for
    for $t = 1$ to $n2(number\ of\ squirrels\ on\ normal\ trees\ moving\ towards\ acorn\ trees)$ do
        if $R2 >= P_{dp}$ then
            $FS_{nt}^{new} = FS_{nt}^{old} + d_g G_c (FS_{at}^{old} - FS_{nt}^{old})$
        else
            $FS_{nt}^{new} = C_x(FS_{nt}^{old}, En, He)$
        end if
    end for
    for $t = 1$ to $n3(number\ of\ squirrels\ on\ normal\ trees\ moving\ towards\ hickory\ trees)$ do
        if $R3 >= P_{dp}$ then
            $FS_{nt}^{new} = FS_{nt}^{old} + d_g G_c (FS_{ht}^{old} - FS_{nt}^{old})$
        else
            $FS_{nt}^{new} = C_x(FS_{nt}^{old}, En, He)$
        end if
    end for
    $S_c^t = \sqrt{\sum_{k=1}^{d} (FS_{at,k}^t - FS_{ht,k})^2}$
    $S_{min} = \frac{10E-6}{(365)^{Iter/(Iter_{max})/2.5}}$
    if $S_c^t < S_{min}$ then
        $FS_{nt}^{new} = FS_L + Levy(n) \times (FS_U - FS_L)$
    end if
    $Calculate\ fitness\ of\ new\ locations$
    $f_i^{new} = f_i(FS_{i,1}^{new}, FS_{i,2}^{new}, .., FS_{i,n}^{new}), i = 1, 2, ...NP$
    if $f_i^{new} < f_i$ then
        $FS_i = FS_i^{new}$
        $f_i = f_i^{new}$
    end if
    $Enhance\ internsive\ dimensional\ search$
    $FindFS_{best}, f_{best}$
    for $j = 1 : n$ do
        $FS_{best,j}^{new} = C_x(FS_{best,j}, En, He), j = 1, 2..., n$
        $calculate\ fitness\ of\ new\ solution$
        $f_{best}^{new} = f(FS_{best,1}, FS_{best,2}, ...FS_{best,j}^{new}, ...FS_{best,n})$
        if $f_{best}^{new} < f_{best}$ then
            $FS_{best,j} = FS_{best,j}^{new}$
            $f_{best} = f_{best}^{new}$
        end if
    end for
    $Iter = Iter + 1$
end while
$End$
```

First, to enhance the exploitation capability of the algorithm, an adaptive predator presence probability is adopted as follows (Zheng and Luo 2019):

$$P_{\mathrm{dp}} = (P_{\mathrm{dpmax}} - P_{\mathrm{dpmin}}) \times (1 - Iter/Iter_{\max})^{10} + P_{\mathrm{dpmin}} \tag{20}$$

where P_{dpmax} and P_{dpmin} are the maximum and minimum predator presence probability, respectively.

Second, a normal cloud model generator is used instead of uniformly distributed random functions to reproduce the new location of the flying squirrels adopting the fuzziness in the foraging behavior of the squirrels. Hence, the Eqs. (13), (14) and (15) are replaced as follows:

$$\mathrm{FS}^{\mathrm{new}}_{at} = \begin{cases} \mathrm{FS}^{\mathrm{old}}_{at} + d_g G_c (FS^{old}_{ht} - FS^{old}_{at}), & R_1 \geq p_{\mathrm{dp}} \\ C_x(FS^{old}_{at}, En, He), & \text{otherwise.} \end{cases} \tag{21}$$

$$\mathrm{FS}^{\mathrm{new}}_{nt} = \begin{cases} \mathrm{FS}^{\mathrm{old}}_{nt} + d_g G_c (FS^{\mathrm{old}}_{at} - FS^{old}_{nt}), & R_2 \geq p_{\mathrm{dp}} \\ C_x(FS^{\mathrm{old}}_{nt}, En, He), & \text{otherwise.} \end{cases} \tag{22}$$

$$FS^{\mathrm{new}}_{nt} = \begin{cases} FS^{\mathrm{old}}_{nt} + d_g G_c (FS^{old}_{ht} - FS^{old}_{nt}), & R_3 \geq p_{\mathrm{dp}} \\ C_x(FS^{\mathrm{old}}_{nt}, En, He), & \text{otherwise.} \end{cases} \tag{23}$$

where En is the Entropy that represents the uncertainty measurement of a qualitative concept and *He* is the Hyper Entropy and is the uncertain degree of entropy *En*.

Third, in order to not deviate from the optimal path, a comparison between the successive positions is introduced. The flying squirrels update them with the new position only when its fitness value is better than the previous position. This is adopted as follows:

$$\mathrm{FS}_i = \begin{cases} \mathrm{FS}^{\mathrm{new}}_i & iff^{\mathrm{new}}_i < f^{\mathrm{old}}_i \\ \mathrm{FS}^{\mathrm{old}}_i, & \text{otherwise.} \end{cases} \tag{24}$$

Finally, to enhance the dimensional search and to prevent the negative effect caused in one dimension because of the changes incorporated in other dimensions, solutions are generated based on individual dimensions. This behavior is adopted as follows:

$$\mathrm{FS}^{\mathrm{new}}_{\mathrm{best},j} = C_x(\mathrm{FS}^{\mathrm{old}}_{\mathrm{best},j}, En, He),\ j = 1, 2 \ldots, n \tag{25}$$

Few recent algorithms that use SSA include Basu (2019), Wang et al. (2019b, c) and Hu et al. (2019).

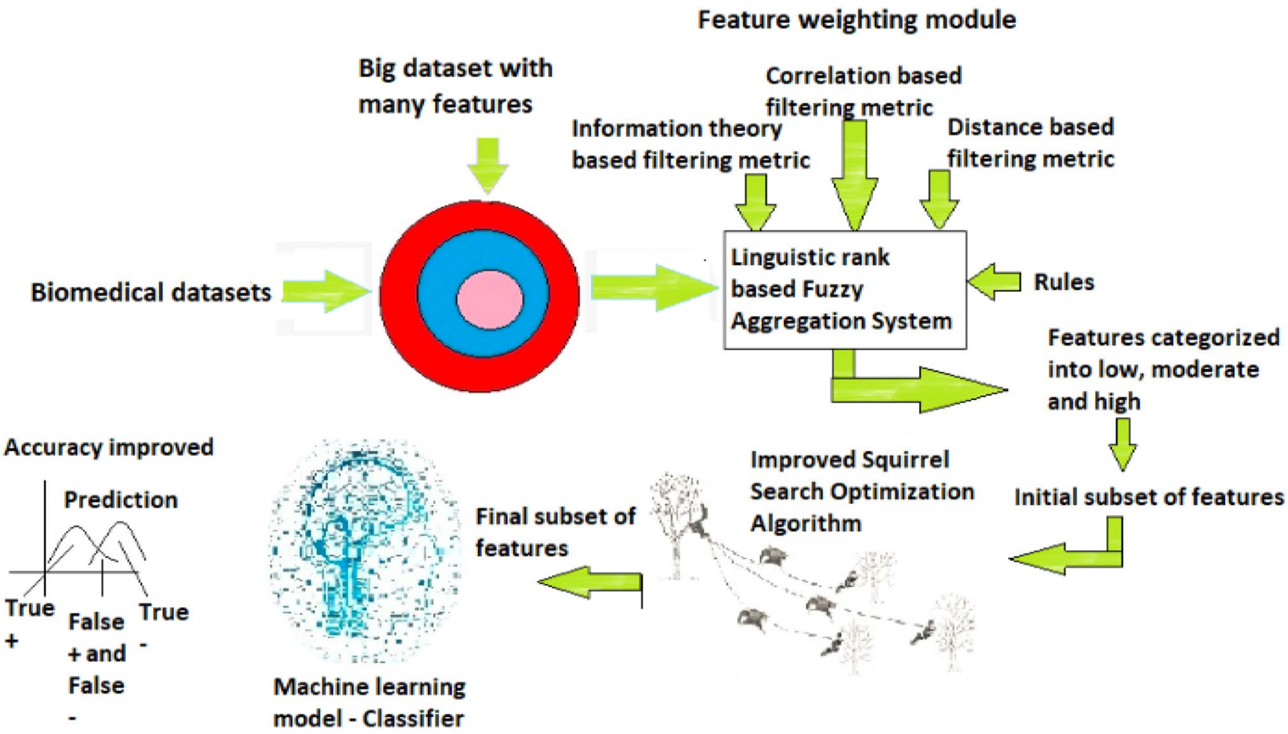

Feature weighting module
Big dataset with many features
Correlation based filtering metric
Information theory based filtering metric
Distance based filtering metric
Biomedical datasets
Linguistic rank based Fuzzy Aggregation System
Rules
Features categorized into low, moderate and high
Initial subset of features
Improved Squirrel Search Optimization Algorithm
Final subset of features
Machine learning model - Classifier
Accuracy improved
Prediction
True +
False + and False -
True -

Fig. 1 Graphical abstract of H-FBRA+ISSA

4 Proposed methodology

The proposed model H-FBRA + ISSA is explained. The graphical abstract of our proposed model is shown in Fig. 1.

4.1 Definitions

Definition 1 (*Relevant feature*): A feature $F_a \in F$ is said to be 'relevant' to the class label C if it provides some new information on class C.

$$I(F, C) > I(F', C) \text{ where } lF' = \{F - F_a\} \tag{26}$$

Definition 2 (*Redundant feature*): A feature $F_a \in F$ is said to be 'redundant' if it does not provide any new information on class C.

$$I(F, C) = I(F', C) \text{ where } F' = \{F - F_a\} \tag{27}$$

Definition 3 (*Highly ranked features*): A feature F_a is classified as a highly ranked feature if the feature as a single feature and in combination with other features gives more information about the class label C.

$$F_a \in HR \text{ if } I(F_a, C) > \lambda \tag{28}$$

Definition 4 (*Moderately ranked features*): A feature F_a is classified as a moderately ranked feature if the feature may give more information about the class label C when combined with other features.

$$F_a \in MR \text{ if } I(F_a, C) < \lambda \text{ and } > \theta \tag{29}$$

Definition 5 (*Low ranked features*): A feature F_a is classified as a low ranked feature if the feature does not give information about the class label C either as a single feature or when combined with other features. λ, θ are constants.

$$F_a \in LR \text{ if } I(F_a, C) < \theta \tag{30}$$

4.2 Computing relevancy information

The normalization of the dataset is done initially. The dataset is normalized with mean 0 and standard deviation 1 to avoid the influence of high-value features. FBRA computes the relevancy information of a feature with the class label using the three measures CB, DB and IB as discussed in Sect. 3. We have chosen the filtering metrics based on different criteria to ensure diversity.

FBRA computes the $CB(F_a, C)$ measure using Eq. (1). FBRA computes the $DB(F_a, C)$ using Eq. (2). Hence, $\mathrm{DB}(F_a, C)$ is given by

$$\mathrm{DB}(F_a, C) = Q(F_a, C) \tag{31}$$

FBRA computes $IB(F_a, C)$ by calculating the rough MI and component co-occurrence information for each feature using Eqs. (6) and (7), respectively. An average of both these measures is calculated for each feature as

$$\mathrm{IB}(F_a, C) = \frac{\mathrm{RI}(F_a, C) + \mathrm{CCRI}(F_a, C)}{2} \tag{32}$$

Now, each feature has three relevancy information $\mathrm{CB}(F_a), \mathrm{DB}(F_a), \mathrm{IB}F_a)$ based on correlation, distance and information-theoretic measures, respectively.

4.3 Computing redundancy information

Generally, redundancy is calculated by measuring the MI between the candidate features and features within the selected subset without considering the class label. Sometimes certain features are identified as more significant but it may be highly correlated with few features and may be completely discarded. Yet, it may be completely independent of the remaining features and thus contributes to the classification accuracy. Our approach addresses this problem by calculating the redundancy information of a feature with all other features and computing the average. In this way, we are considering the global redundancy as we consider the redundancy of each feature with all features and not with another single feature. Our approach computes the redundancy information of a particular feature using the three measures—correlation-based, distance-based and information theoretic-based. The correlation between continuous features is computed using Pearson's correlation and between nominal variables is computed using chi-square test. The correlation between nominal and categorical variables is computed using Eq. (1). The correlation is calculated for each feature with the rest of the features. The average correlation redundancy information of a feature F_a is computed as

$$\mathrm{CBR}(F_a) = \sum_{b=1}^{n} \frac{CB(F_a, F_b)}{n} \; and \; F_b \neq F_a \tag{33}$$

The redundancy information of a feature F_a based on distance-based measure is computed using Eq. (2) as described in the previous section. The neighborhood quality of feature is calculated for each feature with all the other features. The final redundancy information based on distance measure is computed as

$$\mathrm{DBR}(F_a) = \sum_{b=1}^{n} \frac{Q(F_a, F_b)}{n} \; and \; F_b \neq F_a \tag{34}$$

The information-based redundancy information of a feature F_a is calculated using Eqs. (6) and (7), respectively. The final redundancy measure based on information theory-based approaches is computed as

$$\mathrm{IBR}(F_a) = \sum_{b=1}^{n} \frac{\frac{RI(F_a,F_b)+\mathrm{CCRI}(F_a,F_b)}{2}}{n} \text{ and } F_b \neq F_a \tag{35}$$

Now, each feature has three redundancy information $\mathrm{CBR}(F_a)$, $\mathrm{DBR}(F_a)$, $\mathrm{IBR}(F_a)$ based on correlation, distance and information-theoretic measures, respectively.

4.4 Rank aggregation using fuzzy linguistic modeling

The algorithm of FBRA is given in algorithm 3. The main idea behind this work is to rank the features by combining different metrics to avoid bias and induce diversity. A particular filtering technique constraints the optimal search space due to the representational power of a particular feature. But aggregation technique avoids this problem by combining the results of different filtering techniques. The individual filtering method leads to local optimal subsets but rank aggregation feature selection approximates the optimal ranking of features (Waad et al. 2014).

As discussed in the earlier sections, most of the existing aggregation methods suffer from problems, such as tied ranking and disjoint ranking, whereas few others suffer from computational complexity. Hence, we propose fuzzy rule-based system for aggregation. It overcomes these problems as it uses defuzzification to calculate the weight. Besides, fuzzy rule-based system is very flexible and the rules can be extended or updated easily. For example, if the user wishes to give more importance to the information theory-based metrics than the correlation-based and distance-based metrics, the rule base can be updated easily. This flexibility opens the scope for different variations of our proposed model. Moreover, the fuzzy rule-based systems involve human reasoning and decision-making. This helps to provide specific solutions to the different types of problems. In our feature selection problem, this opens the scope to integrate or modify the rule base of our system according to the recommendations from the domain knowledge experts. Fuzzy aggregation model does not require any parameter tuning

Algorithm 3 FBRA - Fuzzy based rank aggregation

```
Step 1: Gets input data
$x \leftarrow Inputdata$

Step 2: Weight calculation and formulating rank matrix
$Normalizedvalue(x) = \frac{x-min(x)}{max(x)-min(x)}$
Compute relevancy information using equations 1, 31, 32
Compute redundancy information using equations 33, 34, 35
$Weights1 \leftarrow$ Compute FWCB using equation 36
$Weights2 \leftarrow$ Compute FWDB using equation 37
$Weights3 \leftarrow$ Compute FWIB using equation 38
$Rankmatrix \leftarrow bind(weights1, weights2, weights3)$
Normalize(rankmatrix)

Step 3:Develop linguistic fuzzy based inference system
$Var = \{CB, DB, IB, FR\}$
$CB = fuzzypartition\{L = VC_l, M = VC_m, H = VC_h\}$
$DB = fuzzypartition\{L = VD_l, M = VD_m, H = VD_h\}$
$IB = fuzzypartition\{L = VI_l, M = VI_m, H = VI_h\}$
$FR = fuzzypartition\{L = VR_l, M = VR_m, H = VR_h\}$
{%comment - Define fuzzy rules - 27 rules defined - sample given%}
$IF\ CB\ is\ VC_l\ \&DB\ is\ VD_l\ \&IB\ is\ VI_l\ THEN\ FR\ is\ VR_l$
$IF\ CB\ is\ VC_m\ \&DB\ is\ VD_l\ \&IB\ is\ VI_h\ THEN\ FR\ is\ VR_h$
$IF\ CB\ is\ VC_h\ \&DB\ is\ VD_h\ \&IB\ is\ VI_h\ THEN\ FR\ is\ VR_h$

Step 4: Use fuzzy inference system and get the final rank of the features
$Sys \leftarrow fuzzysystem(var, rules)$
for $i = 1$ to $n$ do
    $FR(F_i) \leftarrow Fuzzyinference(Sys, rankmatrix(F_i))$
end for

Step 5: Defuzzification and finding the significant features
for $i = 1$ to $n$ do
    $FR(F_i) \leftarrow defuzzify(F(F_i, centroid))$
end for
$Classify\ into\ low,\ moderate\ and\ high\ ranked\ features.$
```

too. The aggregation problem in feature selection involves uncertainty in ranking the input features based on different metrics. Fuzzy rule-based system can handle this uncertainty well. Indeed, in our aggregation problem, the data to be combined are also of lower dimension and hence the number of rules to be generated is also less. These reasons motivate us to use fuzzy rule-based system for aggregation of the different ranking metrics.

The rank for each feature is generated according to the fired fuzzy rule. Initially, the relevancy and redundancy information is combined to obtain the feature weight based on each filtering measure. A feature with more relevant information and less redundant information is considered to be important. Hence, the feature weight value based on correlation measure, distance measure and information-theoretic measure for each feature is calculated as

$$FWCB(F_a) = \alpha(CB(F_a)) + \beta(CBR(F_a)) \tag{36}$$

$$FWDB(F_a) = \alpha(DB(F_a)) + \beta(DBR(F_a)) \tag{37}$$

$$FWIB(F_a) = \alpha(IB(F_a)) + \beta(IBR(F_a)) \tag{38}$$

where α and β represent the weight assigned to relevancy information and redundancy information, respectively, in calculating the feature weight.

A rank matrix is formulated with the three different weights for each feature. The rank matrix is normalized. The three weights form the input fuzzy sets. The data distribution of the inputs is considered to define linguistic labels Low (L), Medium (M) and High (H) for each input. A total of $3^3 = 27$ fuzzy rules corresponding to each input variable and linguistic label values are generated. The weights of each feature are fed as input to the fuzzy inference system. The fuzzy rules are weighed and a rule with maximum weight is fired. Since it is a normal aggregation with few values, we use simple techniques like triangular membership function for fuzzification and centroid method for defuzzification. Fuzzy rule generation is also simple by computing membership values and obtains the grade of the uncertainty of the rule. Other methods include using decision tree, neuro-fuzzy or rough set technique to generate rules. Other membership functions include GA, Self-Organizing Map (SOM) and few other techniques but are not required in our case owing to its simplicity. The triangular membership function is given by

$$\mu_a(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{m-a}, & a < x \le m \\ \frac{b-x}{b-m}, & m < x < b \\ 0, & x \ge b \end{cases} \tag{39}$$

One of the fuzzy systems used in our proposed model is shown in Fig. 2. As stated above, we use triangular membership function for fuzzification.The other two popular methods used for fuzzication in linear problems are trapezoidal

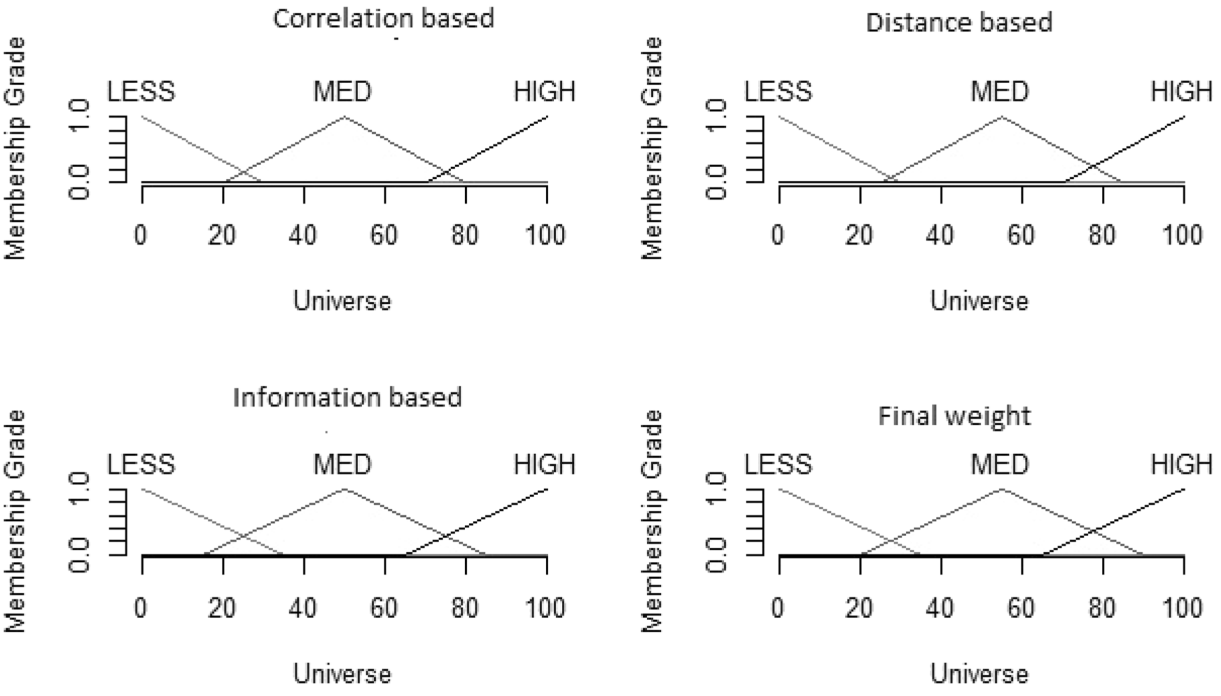


Fig. 2 Sample fuzzy system–triangular membership function

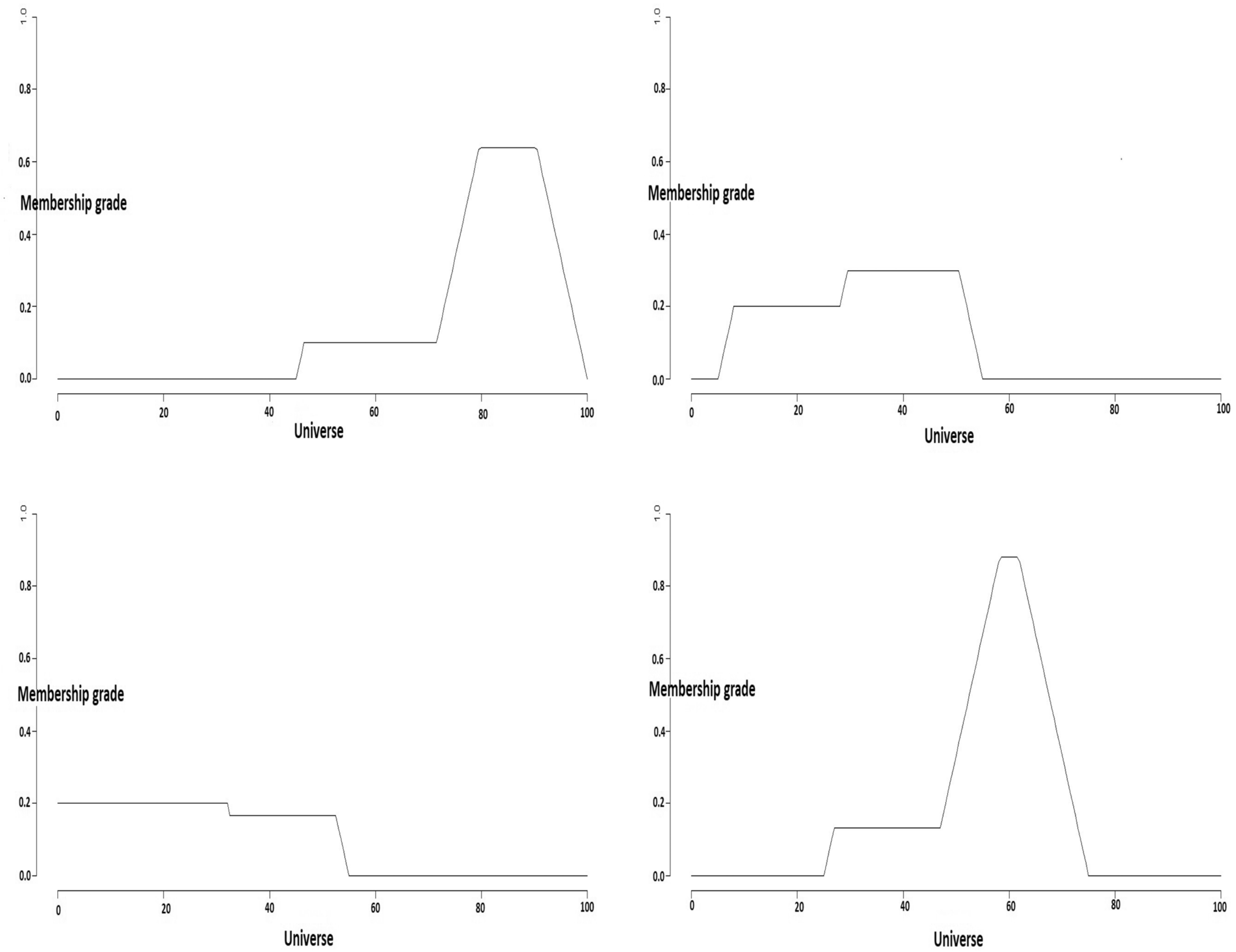


Fig. 3 Sample defuzzification of the features in our proposed model–centroid method

and guassian while non-linear problems employ few other membership functions, such as Generalized bell, TT-shaped and S-shaped membership functions. Generally, triangular membership functions are tried as the start function for each problem as they are straight lines. This property makes it simple to implement. The necessity to use advanced membership functions is avoided when triangular membership function yields satisfactory solution to the problem. While trapezoidal membership functions represent fuzzy intervals, triangular membership functions represent fuzzy numbers. Hence, triangular membership functions are highly preferred for simple problems. Moreover, the triangular membership function is easier for taking parameter values than trapezoidal or gaussian (Rahim et al. 2017). Also, triangular membership function is found to take less memory size for variables and program size than the other membership functions (Princy and Dhenakaran 2016). These reasons motivate us to use triangular membership function for fuzzification.

The output is then defuzzified using the centroid method. It is given by

$$Z^* = \frac{\int \mu_c(z)zdz}{\int \mu_c(z)dz} \tag{40}$$

where $c(z)$ is the degree of membership of the aggregated fuzzy set for the output z.

Figure 3 shows few sample defuzzification of features in our proposed model. As stated above, we use centroid method for defuzzification. The most commonly used defuzzification methods include middle of maximum, largest of maximum, and smallest of maximum, centroid and bisector (center of area). The bisector method divides the fuzzy region into regions of equal area and this line is used for defuzzification. The centroid method is the center of gravity point of the fuzzy region. The middle of maximum, largest of maximum and smallest of maximum returns the middle point, largest point and smallest point, respectively, of the maximum region. Centroids are more generous and cheap as they limit the defuzzified value. For example, when the output fuzzy set covers a range between 10 and 40, the defuzzified value is between 15 and 35 when centroid is used. The middle of maximum, smallest of maximum or the largest of maximum does not impose such limits and the defuzzified value can fall more than 35 or less than 15 which may yield poorer results in certain problems. The bisector line in the center of area method is found to yield approximately the same output as the centroid method (Uraon and Kumar 2016). Hence, there is no major performance difference between these two methods and either of them can be the choice. Besides, centroid method is found to be sensitive to all rules and hence incurs a smooth change when compared with the other defuzzification methods. It is also computationally fast as it involves simple operations (Saletic and Popovic 2006). Owing to these reasons, we chose centroid method for defuzzification.

The output variable final rank is assigned to one of the three linguistic labels Low (L), Medium (M) and High (H). Features that belong to L categories form the low-rank features. Features belonging to M category form the moderately ranked features and the features belonging to H categories form the highly ranked features.

4.5 Hybrid FBRA + ISSA algorithm

The algorithm of hybrid FBRA + ISSA is depicted in algorithm 4. Hybrid feature selection approaches as discussed in earlier sections combine filter and wrapper feature selection approaches. The initial subset of features is selected based on the filter approaches. The final subset of features is selected by identifying the subset of features that maximizes the optimizing function. The optimizing function is usually a function that maximizes the accuracy of a classifier. No single optimization algorithm is found to be effective for all the problems and hence different algorithms or improvements in existing algorithms are proposed (Jain et al. 2019). ISSA algorithm is used as the wrapper algorithm in our proposed model owing to its advantages discussed in the previous section. The initial subset of features is selected using the Fuzzy-Based Rank Aggregation (FBRA) technique. The result from FBRA is fed as input to the ISSA algorithm. The initial set of solutions is formed by including the high-rank features and then selecting different combinations from the moderate-rank features. The low-rank features are discarded completely. This decreases the computational time by reducing the feature space. Moreover, the bias towards the specific metric is avoided as our proposed model FBRA ranks features by combining different metrics that aid in the selection of the best feature subset.

Algorithm 4 Hybrid FBRA+ISSA

```
Preprocess dataset
Apply FBRA to obtain the feature ranks
Set parameters for ISSA and initialize random locations for flying squirrels
Initial locations of squirrels ← H Union different combinations of M from FBRA
repeat
for all flying squirrels do
    Construct Dataset $D'$ for each squirrel with features from initial solutions
    Learn the classifier
    Calculate the fitness value
end for
Update the positions of the flying squirrels
until stopping criteria is met
Return the best feature subset S
Return the best classification accuracy
```

5 Experimental framework

5.1 Dataset description

Nine biomedical datasets are considered for experimentation out of which five datasets are high dimensional. To study the behavior of our model in low-dimensional datasets and for a better understanding of our model, the experiments are conducted on four low-dimensional datasets too. Table 1 shows the descriptive summary of each dataset. It includes number of observations, number of features and number of classes and the source of each dataset. These datasets are binary or multiclass biomedical classification tasks and are appropriate to show the effectiveness of our model.

The first dataset is part of RNA seq PANCAN dataset (Fiorini 2016). This dataset contains the gene expression of patients with different types of tumors extracted randomly. The different type of tumors include Breast invasive Carcinoma (BRCA), Kidney Renal clear cell Carcinoma (KIRC), Colon Adenocarcinoma (COAD), Lung Adenocarcinoma (LUAD) and Prostate Adenocarcinoma (PRAD). The focus is on classifying the samples into five different classes (type of tumors) based on the gene expressions.

The second dataset, Cancer gene represents the gene expression levels corresponding to Acute Myeloid Leukemia (AML) and Acute Lymphoblastic Leukemia (ALL) (Golub et al. 1999). The tissue samples are collected from Bone Marrow and Peripheral Blood. The focus is on classifying the samples by monitoring gene expression into one of the two classes AML or ALL.

The third dataset, Lymphoma represents the gene expression of Diffuse Large B-Cell Lymphoma (DLBCL) patients (Alizadeh et al. 2000). The variability in the tumor is to be identified by classifying them into different classes. Hence, the focus is on classification of the different B-cell groups. The different stages of B-cell determines the survival chances of the patient. Thus this classification helps to determine the survival chances of the patient with Lymphoma.

The fourth dataset, Colon represents the gene expression consisting of samples both from healthy persons and colon cancer affected patients (Alon et al. 1999). It consists of two thousand genes. The focus is on classifying the normal persons and the colon cancer affected patients based on their gene expression.

The fifth dataset, microRNAs represent the microRNA expression profiling to detect breast cancer (Matamala et al. 2015). Proper classification helps to discriminate breast cancer and the intrinsic molecular subtypes. The focus is on classifying the normal and breast cancer tissues. This helps to treat the cancer at an earlier stage.

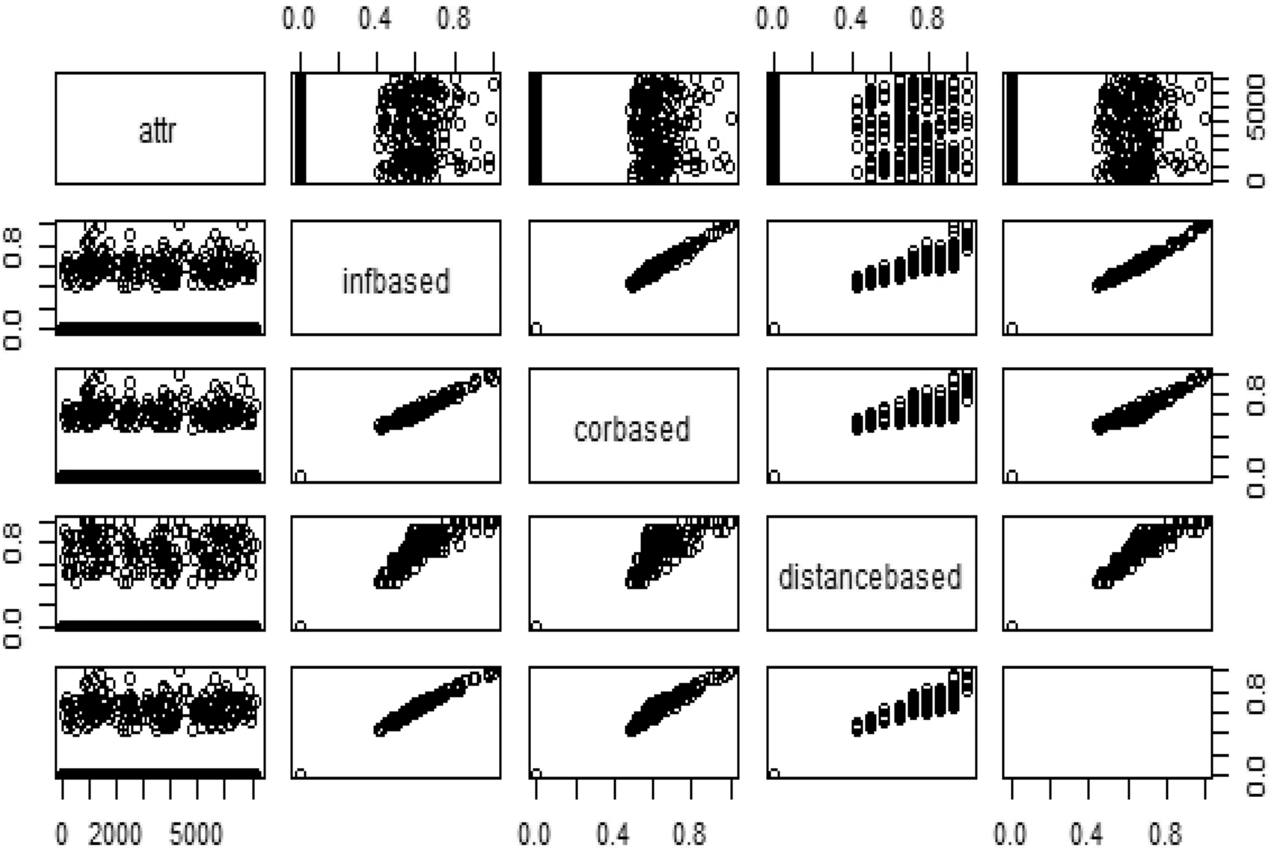
The sixth dataset, Chronic kidney is collected from patients in India. Around twenty four features of them such as their age, blood pressure, albumin, sugar, wc cells count, sodium, potassium, hypertension are recorded. The focus is on classifying the normal and the chronic kidney disease affected patients based on these features. The seventh dataset, Spine is collected in an Orthopaedic center in France. Around twelve features of patients such as pelvic tilt, pelvic incidence, sacral scope are recorded. The focus is on classifying them into normal or abnormal based on these features. The eighth dataset, heart contains thirteen features of patients such as age, fasting blood sugar, resting blood pressure and chest pain type. These features help to identify the presence or absence of heart disease in the patients. Hence, the focus is on classifying them into two classes corresponding to the presence or absence of heart disease. The ninth dataset, cancer describes the characteristics of cell nuclei in the breast. Around thirty one features such as cell radius, texture, concavity, fractal dimension, smoothness are recorded. The focus is on classifying if the cancer type is benign or malignant based on these features. Further details about these datasets and sources can be obtained from the public data repository mentioned in Table 1 corresponding to the each dataset.

5.2 Parameter setting

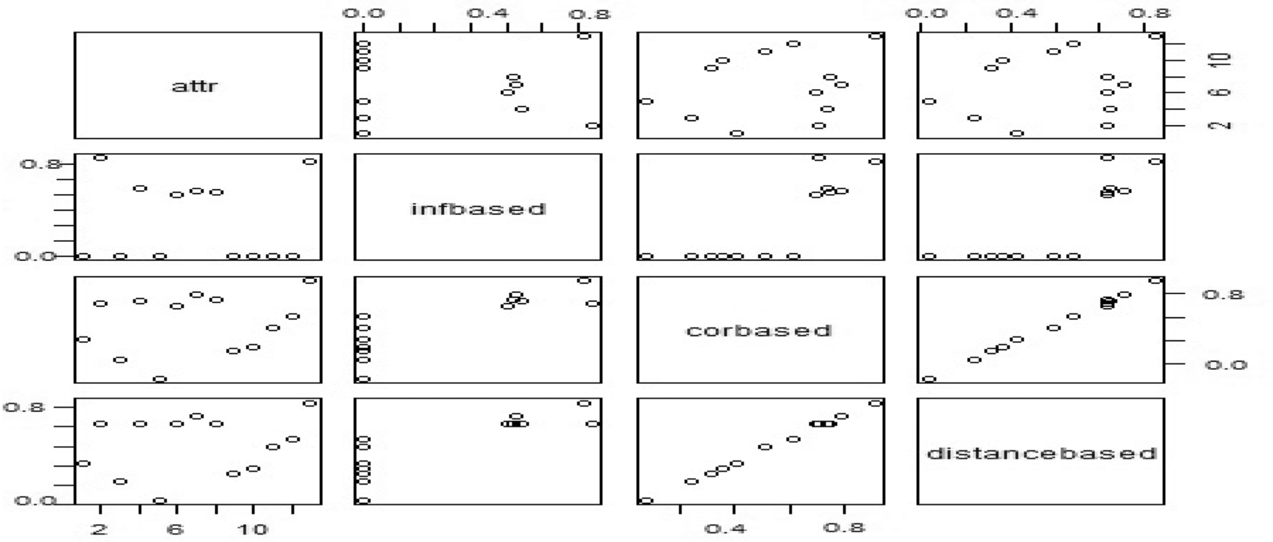
The main parameters of ISSA algorithm include maximum number of iteration $I_{\max}$, the population size P_s, the number of decision variables n, the maximum predator presence

Table 1 Dataset description

Dataset no	Dataset	Source	Rows	Features	Classes	Feature types	Sample proportion
1	Pancan	UCI repository	801	20531	5	Continuous	17–18–18–10–37%
2	Cancer Gene	Kaggle	72	7129	2	Continuous	65–35%
3	Lymphoma	llmpp.nih.gov	96	4026	9	Continuous	10–11–48–9–2–2–6–4–6%
4	Colon	Kaggle	62	2000	2	Continuous	65–35%
5	MicroRNAs	Kaggle	133	1928	2	Continuous	92–8%
6	Chronic kidney	UCI repository	400	24	2	Mixed	62–38%
7	Spine disease	Kaggle	310	12	2	Continuous	68–32%
8	Heart disease	UCI repository	270	13	2	Mixed	55–45%
9	Cancer disease	Kaggle	569	31	2	Continuous	37–63%



(a) Dataset:Dataset2



(b) Dataset:Dataset7

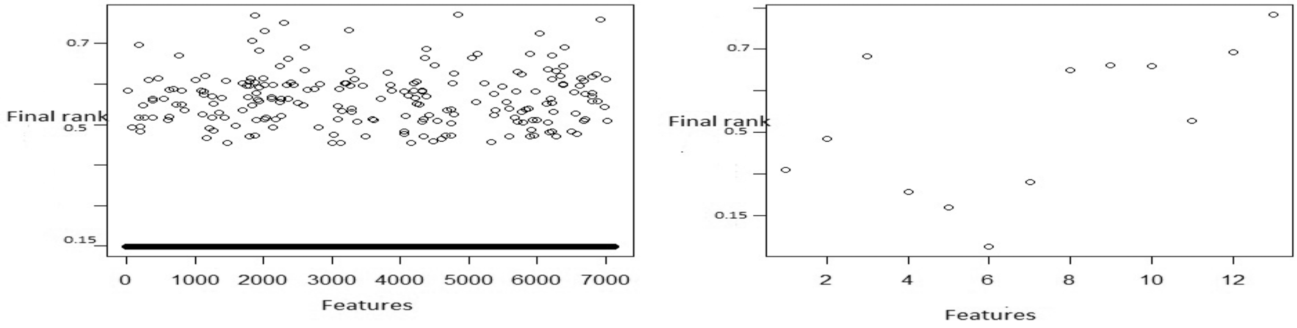
Fig. 4 Weight matrix of the datasets based on the three different measures—information, correlation, and distance

probability P_{dp}(max), the minimum predator presence probability P_{dp}(min), the scaling factor S_f, the gliding constant G_c, and the upper and lower bounds for decision variable FS_u and FS_l. The initial values are set according to the original literature (Zheng and Luo 2019). I_{max} to 100,000, P_s to 50, P_{dp}(max) to 0.1,P_{dp}(min) to 0.001,S_f to 18 and G_c to 1.9. The parameters α and β in the Eqs. 36, 37 and 38 are determined empirically and the best value obtained is 0.6, 0.4 after trying with different values in the range [0.3,0.8]. The classifiers used in our experimentation include Support Vector Machine (SVM), Random Forest (RF) and Deep Neural Networks (DNN). A radial basis kernel with degree 3 or a polynomial kernel is used in SVM. As our datasets are not linearly separable, we tried with radial basis kernel function as it is the first and preferred choice for non-linear data separation. While it yields satisfactory performance for most of the datasets, the performance of it is not satisfactory for few datasets used in our experimentation. Hence, we tried with polynomial kernel for these datasets and obtain satisfactory performance. The optimal values of the model parameters $C, \in$ are determined using the grid search technique. In case of random forest, the less number of trees increases variance and the more number of trees increases computational burden. In our datasets, though the features are more, the number of rows is relatively lesser and there is not much significant difference with respect to the number of rows among the different datasets considered in our experimentation. Hence, 50 trees are grown in RF and it yields satisfactory performance. Deep learning classification is executed with Keras on top of TensorFlow. Adaboost is used to optimize the network weights and rectifier is used as the activation function. The DNN parameters are determined by randomizedsearchcv as gridsearch is expensive with DNN. A 10-fold Cross-validation is used in all the classifiers. GA, PSO and Whale Optimization Algorithm (WOA) are used in our experiments for comparison. The parameters of these algorithms are set according to Nagarajan and Babu (2019). The population size is set between 60 and 100 in GA and PSO. The acceleration factors of PSO is set to 2.025 and the inertia weight is set to 0.625. The crossover and mutation ratios are set to 0.9 and 0.1 in GA. The parameter values of GA and PSO are determined by sensitivity analysis. The value of a is decreased from 2 to 0 over iterations in WOA as per the original literature (Mirjalili and Lewis 2016).

5.3 Results and discussion

The datasets are standardized to have a mean equal to zero and standard deviation equal to one. This is done to avoid the influence of high-value features. The relevancy and redundancy information for the features is computed using metrics based on three different measures—correlation, distance and information as discussed in Sect. 4. The weights of the features for these three measures are calculated using Eqs. 36, 37 and 38 respectively. Though the experimental results are shown for all the nine datasets, for illustrative purposes, we are showing the results of our proposed rank aggregation approach FBRA only on two datasets—2 and 7 (one high dimensional and one low dimensional).

A sample plot of the weight matrix for datasets 2 and 7 based on the three measures used in FBRA is shown in Fig. 4. For example, consider Fig. 4a that corresponds to dataset 2. Most of the features have the information based measure value between 0.4 and 0.8. The density value for a correlation-based measure is high between 0.5 to 0.7. Most of the features have a distance based measure value between 0.4 and 1. Few features have the value 0 for the information-based measure, correlation and distance-based measure indicating that they don‘t contribute to the classification. In Fig. 4b, that corresponds to dataset 7, five features—degree spon, pelvic incidence, lumbar angle, sacral scope and pelvic radius are found to have high value with respect to information-based measure. Apart from these five features, pelvic tilt is also found to have high value when the correlation-based measure is considered. But with respect



(a) Dataset:Dataset2

(b) Dataset:Dataset7

Fig. 5 Final Feature ranking by FBRA

to dependence based measure, features such as cervical tilt is also found to have high value.

The linguistic labels are defined for the input and output variables and the fuzzy system is constructed for rank aggregation. The final feature weights are computed with this fuzzy system. A sample plot of the final weight of the features is shown in Fig. 5 for datasets 2 and 7. For example, consider Fig. 5a which corresponds to dataset 2. As seen in Fig. 5a , each feature is found to have different values based on the three different measures. But with rank aggregation using linguistic fuzzy system, the final rank of the features is derived as shown in Fig. 5a. Figure 5b shows the final rank derived for dataset 7.

The selected features can be biologically interpreted. For dataset2 as displayed in Fig. 5a, the contributing features are identified and finally twenty five genes such as gene 4847 (Zyxin), 804 (Macmarcks), 1882 (CST3 Cystatin C), 6855 (TCF3 Transcription factor 3) are selected by FBRA as the high rank and moderate rank features. Hence, these genes play a major role in determining the type as AML or ALL. For dataset 7 as displayed in Fig. 5b, five features—feature 7, 1, 3, 5, 2 that corresponds to degree spon, pelvic incidence, lumbar angle, pelvic rad and pelvic tilt are selected by FBRA as high rank features. Three features thoracic slope, sacral scope, cervical tilt are selected as moderate rank features by FBRA. The remaining four features are classified as low rank features by FBRA and can be ignored. The high rank features with different combinations of moderate rank features can be tried with the optimization algorithms for feature selection.

Similarly for dataset 1, forty two genes such as TTN, RPS20, RPLP2, TRBC2, RPSA, RPS11, RPL11, RPS16 are selected as high and moderate rank features. For dataset 4, around eighteen genes such as 704 (Human tyrosine kinase (HTK) mRNA, complete cds), 581 (CALGIZZARIN), 267 (Human Cysteine-Rich Protein (CRP) gene), 1873 (Human MXI1 mRNA, complete cds.), 377 (H.sapiens mRNA for

Table 2 Comparison of H-FBRA + ISSA with Individual filtering metrics—SVM

Approaches	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9
FS	81 [38]	83.95 [20]	75.58 [19]	87.2 [17]	84.48 [16]	87.83 [14]	75.59 [8]	78.48 [7]	82.8 [16]
CHS	82 [40]	82.78 [22]	74.86 [21]	86.84 [19]	83.76 [15]	87.38 [10]	74.06 [6]	78.21 [8]	82.35 [17]
ReliefF	83.2 [39]	83.77 [27]	75.4 [18]	85.76 [22]	85.2 [17]	86.3 [15]	74.69 [4]	79.74 [10]	83.43 [20]
IG	84.2 [37]	85.84 [27]	74.32 [18]	88.1 [20]	84.75 [18]	87.02 [9]	75.5 [5]	76.68 [8]	86.04 [17]
MRMR	85 [38]	85.66 [26]	75.58 [16]	87.65 [18]	85.29 [19]	89 [8]	75.77 [7]	79.74 [9]	85.86 [16]
CFS	81.9 [40]	85.39 [21]	75.49 [17]	88.1 [19]	84.3 [17]	89.7 [9]	74.78 [8]	78.66 [10]	84.15 [15]
MRMD	84.2 [35]	85.94 [18]	76.3 [20]	85.85 [20]	85.56 [16]	88.28 [12]	73.89 [4]	80.01 [12]	84.78 [17]
SFR	83.7 [42]	85.48 [17]	75.85 [18]	87.47 [22]	84.48 [17]	87.47 [17]	75.5 [5]	79.56 [9]	83.79 [15]
NQF	85.1 [35]	84.49 [19]	74.86 [17]	86.66 [20]	85.2 [15]	88.19 [12]	73.79 [6]	74.88 [7]	82.62 [14]
CCI	84 [40]	86.02 [15]	75.31 [16]	87.65 [19]	83.4 [14]	86.84 [11]	75.77 [7]	78.75 [9]	83.34 [17]
RMI	85 [39]	85.48 [17]	75.85 [18]	86.84 [17]	84.75 [13]	87.29 [8]	74.78 [7]	79.47 [8]	84.33 [16]
FBRA	**86.85** [42]	**88.2** [25]	**78.3** [17]	**89.1** [18]	**88.56** [13]	**90** [15]	**77.4** [8]	**80.1** [8]	**86.31** [17]

The bold values represent the best solution corresponding to the evaluated metric

Table 3 Comparison of H-FBRA + ISSA with Individual filtering metrics—RF

Approaches	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9
FS	82.5 [38]	84.58 [20]	73.87 [19]	86.39 [17]	83.58 [16]	86.48 [14]	72.8 [8]	78.66 [7]	88.83 [16]
CHS	83 [40]	83.59 [22]	74.68 [21]	87.38 [19]	84.3 [15]	87.47 [10]	74.87 [6]	79.11 [8]	84.78 [17]
ReliefF	83.5 [39]	84.94 [27]	73.6 [18]	85.85 [22]	84.39 [17]	85.4 [15]	73.52 [4]	80.19 [10]	88.02 [20]
IG	84.2 [37]	85.66 [27]	74.14 [18]	86.39 [20]	83.67 [18]	86.66 [9]	74.69 [5]	78.57 [8]	86.85 [17]
MRMR	85.1 [38]	85.93 [26]	74.86 [16]	87.2 [18]	83.49 [19]	87.83 [8]	75.41 [7]	80.1 [9]	87.39 [16]
CFS	83.1 [40]	85.48 [21]	74.59 [17]	86.84 [19]	84.3 [17]	87.65 [9]	76.22 [8]	79.65 [10]	85.86 [15]
MRMD	84.8 [35]	85.21 [18]	74.05 [20]	85.85 [20]	84.57 [16]	85.58 [12]	74.6 [4]	80.64 [12]	86.58 [17]
SFR	84.1 [42]	85.48 [17]	74.59 [18]	86.39 [22]	84.3 [17]	87.29 [17]	72.89 [5]	80.1 [9]	86.49 [15]
NQF	85.9 [35]	84.85 [19]	74.87 [17]	85.76 [20]	83.76 [15]	87.47 [12]	74.06 [6]	78.66 [7]	87.57 [14]
CCI	84.1 [40]	86.02 [15]	74.86 [16]	86.48 [19]	84.48 [14]	88.1 [11]	74.6 [7]	78.84 [9]	88.47 [17]
RMI	85 [39]	85.48 [17]	75.31 [18]	87.2 [17]	84.12 [13]	87.29 [8]	75.5 [7]	80.28 [8]	89.28 [16]
FBRA	**86.85** [42]	**88.2** [25]	**78.3** [17]	**88.38** [18]	**88.2** [13]	**89.1** [15]	**77.4** [8]	**81.81** [8]	**90** [17]

The bold values represent the best solution corresponding to the evaluated metric

Table 4 Comparison of H-FBRA + ISSA with Individual filtering metrics–DNN

Approaches	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9
FS	83 [38]	83.66 [20]	77.38 [19]	87.83 [17]	83.2 [16]	81.8 [14]	71.27 [8]	73.89 [7]	80.19 [16]
CHS	84.1 [40]	82.58 [22]	76.57 [21]	87.74 [19]	82.38 [15]	82.07 [10]	67.58 [6]	75.78 [8]	80.91 [17]
ReliefF	84.1 [39]	84.02 [27]	77.38 [18]	86.48 [22]	84.2 [17]	83.6 [15]	70.46 [4]	78.48 [10]	81.09 [20]
IG	84.8 [37]	85.15 [27]	76.66 [18]	87.83 [20]	84.14 [18]	82.88 [9]	72.17 [5]	77.67 [8]	82.17 [17]
MRMR	85.3 [38]	85.38 [26]	78.64 [16]	87.11 [18]	83.56 [19]	84.86 [8]	71.63 [7]	78.3 [9]	81.81 [16]
CFS	84.6 [40]	84.2 [21]	77.65 [17]	87.2 [19]	84.2 [17]	85.4 [9]	70.91 [8]	76.77 [10]	81 [15]
MRMD	85 [35]	85.03 [18]	78.46 [20]	86.57 [20]	82.74 [16]	84.59 [12]	71.18 [4]	77.58 [12]	80.91 [17]
SFR	84.7 [42]	84.66 [17]	77.74 [18]	87.56 [22]	84.2 [17]	83.6 [17]	70.19 [5]	76.86 [9]	78.48 [15]
NQF	86 [35]	84.12 [19]	78.55 [17]	86.84 [20]	85.1 [15]	84.23 [12]	69.74 [6]	78.3 [7]	80.28 [14]
CCI	85 [40]	85.92 [15]	78.46 [16]	87.38 [19]	83.1 [14]	82.79 [11]	70.46 [7]	77.58 [9]	81.18 [17]
RMI	86 [39]	84.38 [17]	77.56 [18]	87.29 [17]	83.75 [13]	84.5 [8]	67.85 [7]	76.86 [8]	81.45 [16]
FBRA	**87.03** [42]	**87.92** [25]	**81.09** [17]	**89.1** [18]	**87.82** [13]	**86.04** [15]	**73.8** [8]	**79.2** [8]	**82.35** [17]

The bold values represent the best solution corresponding to the evaluated metric.

GCAP-II/uroguanylin precursor) are classified as high and moderate rank features by FBRA. For dataset 5, thirteen miRNAs such as miR-505-5p, miR-125b-5p, miR-21-5p, and miR-96-5p are classified as high and moderate rank features for the classification task. For dataset 3, seventeen genes such as 203, 1963, 1760, 3353, 2395 are selected as high and moderate rank features by FBRA. Inspite of several endeavors, we were unable to get the biological interpretation of these genes (dataset 3) from the literatures available. Therefore explanations about these genes are not provided in this paper. For dataset 6, fifteen features such as haemoglobin, red blood cell count, hypertension, albumin, blood glucose random are classified as high and moderate rank features. For dataset 8, seven features such as age, fasting blood sugar, Trestbps, Cholestrol are classified as high and moderate rank features. For dataset 9, seventeen features such as perimeter-largest-worst, radius-largest-worst, concave-points-mean, texture-mean are classified as high and moderate rank features. Thus FBRA helps to reduce the dimension of each dataset to a greater extent by choosing the contributing features. More information about the final subset of features selected by our proposed model after using FBRA as the initial filtering is discussed in the subsequent sections.

5.3.1 Comparison with individual filtering metrics

The first set of experimentation compares the performance of our proposed rank aggregation approach FBRA with eleven other approaches on nine datasets. The proposed FBRA is compared with six well-known filtering metrics—FS, Chi-Square (CHS), ReliefF, IG, mRmR and CFS. It is also compared with five novel filtering metrics—Maximal Relevance Maximal Distance (MRMD) (Zou et al. 2016), Subspace clustering Feature weighing (SFR) (Chen et al. 2018), Neighborhood-based Quality of Feature (NQF) (Liu et al. 2017), Component Cooccurrence Information (CCI) (Wang and Feng 2018) and Rough Mutual Information (RMI) (Zeng et al. 2014). The subset of features selected by each metric is used as input to the classifier. The performance is evaluated on the basis of the number of features selected and the classification accuracy. We used Kneedle algorithm proposed in Satopaa et al. (2011) that uses the concept of 'elbow" point in the cost-benefit curves, to determine the optimal number of features and provides a satisfactory trade-off between selected number of features and classification accuracy with individual metric approaches. The classification accuracy achieved by FBRA and other individual filtering metrics is shown in Tables 2, 3 and 4.

Table 5 Comparison of H-FBRA + ISSA with other aggregation approaches—SVM

Approaches	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9
Borda	84 [38]	84.9 [29]	76.2 [21]	87 [23]	85.1 [16]	87 [15]	73 [10]	79 [10]	85.6 [20]
RRA	84.9 [38]	85.6 [28]	76 [26]	87.8 [21]	84.3 [19]	87.6 [17]	72 [9]	80[12]	86.1 [18]
SA	85 [38]	86.1 [26]	77 [19]	87.2 20]	86 [15]	88.1 [19]	76 [8]	79.3 [9]	84 [19]
MVFS	85.5 [38]	87.3 [27]	77.8 [20]	88 [21]	8 7[14]	89 [17]	74 [6]	80[8]	86 [18]
FBRA	**86.85 [38]**	**88.2 [25]**	**78.3 [17]**	**89.1 [18]**	**88.56 [13]**	**90 [15]**	**77.4 [8]**	**80.1 [7]**	**86.3 [17]**

The bold values represent the best solution corresponding to the evaluated metric

Table 6 Comparison of H-FBRA + ISSA with other aggregation approaches—F

Approaches	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9
Borda	84 [38]	86 [29]	75.9 [21]	85.8 [23]	87 [16]	87 [15]	73 [10]	80 [10]	88.2 [20]
RRA	85.5 [38]	87.1 [28]	76.5 [26]	86.3 [21]	85.4 [19]	86 [17]	72.4 [9]	80.5 [12]	89 [18]
SA	85 [38]	86.9 [26]	77 [19]	87 [20]	86.3 [15]	88.1[19]	76.2 [8]	81 [9]	86.3 [19]
MVFS	86 [38]	87.6 [27]	7 8[20]	87.4 [21]	87.1 [14]	88.4 [17]	74 [6]	81.5 [8]	89.4 [18]
FBRA	**86.85 [38]**	**88.2 [25]**	**78.3 [17]**	**88.38 [18]**	**88.2 [13]**	**89.1 [15]**	**77.4 [8]**	**81.81 [8]**	**90[17]**

The bold values represent the best solution corresponding to the evaluated metric

Table 7 Comparison of H-FBRA + ISSA with other aggregation approaches—DNN

Approaches	DS1	DS2	DS3	DS4	DS5	DS6	DS7	DS8	DS9
Borda	85 [38]	87.19 [29]	78.5 [21]	87.4 [23]	86.6 [16]	85 [15]	71 [10]	79 [10]	80 [20]
RRA	85.6 [38]	87.18 [28]	79 [26]	88 [21]	87.1 [19]	84.4 [17]	68 [9]	78.4 [12]	82 [18]
SA	85.9 [38]	87.6 [26]	78.9 [19]	87.9 [20]	87.19 [15]	85.3 [19]	69 [8]	79 [9]	81.3 [19]
MVFS	86 [38]	86.9 [27]	80 [20]	88.1 [21]	87.5 [14]	85.1 [17]	70 [6]	79.2 [8]	82 [18]
FBRA	**87.03 [38]**	**87.9 2[25]**	**81.09 [17]**	**89.1 [18]**	**87.82 [13]**	**86.04 [15]**	**73.8 [8]**	**79.2 [8]**	**82.35 [17]**

The bold values represent the best solution corresponding to the evaluated metric

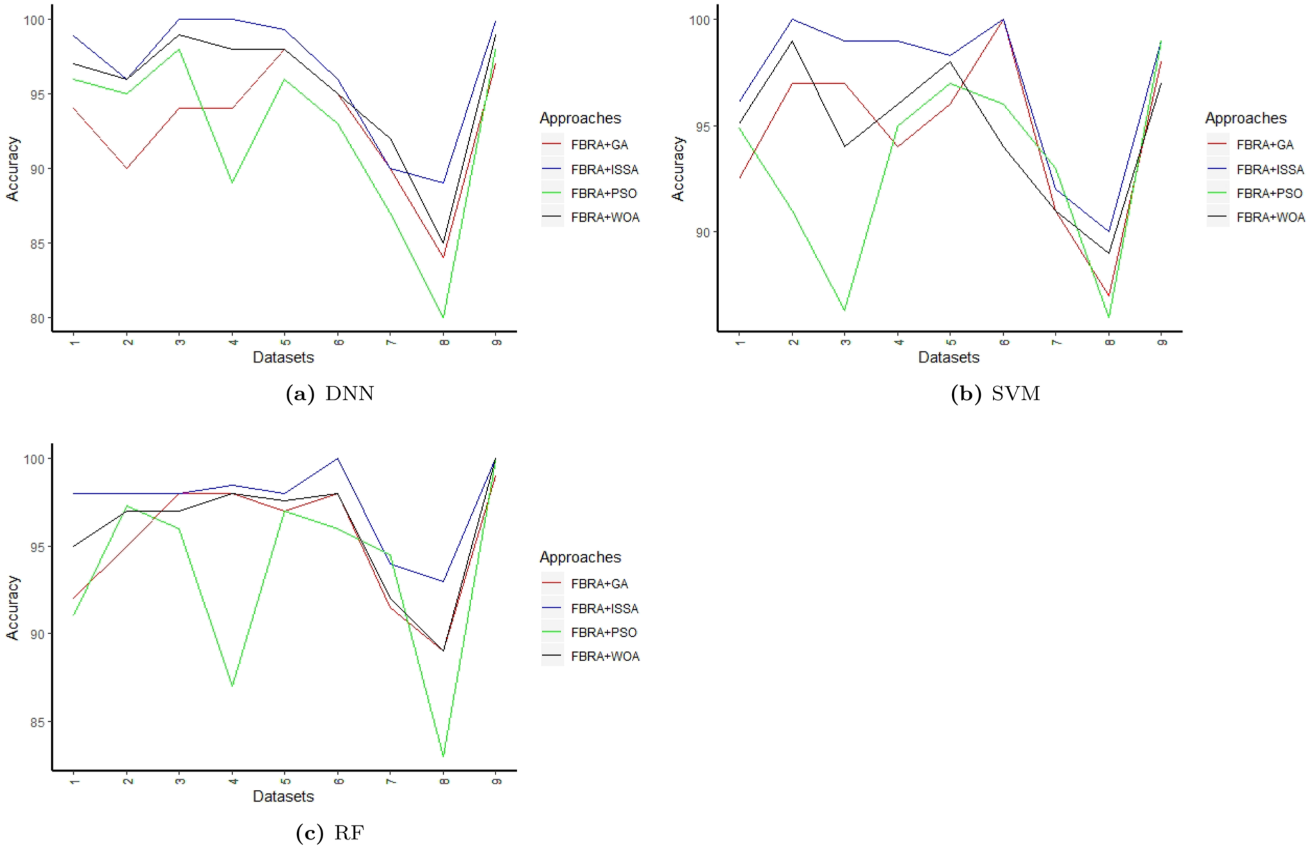
The number of features selected by each metric is specified in the brackets against the classification accuracy.

Tables 2, 3 and 4 show that FBRA performs better than the other approaches used for experimentation in terms of classification accuracy. The results are very supportive, especially in high dimensional datasets. Since there is huge amount of diversity in high dimensional datasets, rank aggregation approach FBRA performs much better than individual approaches. The difference in classification accuracy is not very high with FBRA in low dimensional datasets. Yet, FBRA exhibits superior performance than individual approaches in these datasets too. There is no specific individual approach that performed well or poorly in all the datasets. For example, mRmR exhibits superior performance in dataset 1 whereas CCI exhibits superior performance in dataset 2. There is an overlap in the performance of different individual approaches. The reason for this was discussed in earlier sections. Depending on the nature of the features in each dataset, different metrics perform well on different datasets. But FBRA exhibits exceptional performance in all the datasets as it considers several metrics for ranking the features.

Tables 2, 3 and 4 show the number of features selected by each approach. The feature dimensionality is reduced to a greater extent by filtering techniques. Though individual approaches select lesser number of features than our proposed approach in most of the datasets, the best classification accuracy is achieved by FBRA. The reason for the individual approaches to select lesser number of features than FBRA is obvious. FBRA is an aggregation technique and the aggregation techniques usually end up with more number of features than individual measures as they combine many metrics to induce diversity. Yet, there is a significant reduction in the number of features and they perform well in terms of classification accuracy. It is observed that there are variations in the classification accuracy among the three classifiers. This happens as different classifiers perform well on different datasets due to several reasons such as nature of the data types, number of features, and parameters for the classifiers Discussion on the performance of the classifiers is outside the purview of this paper. Though there are differences in the performance of the classifiers, it is clear that our method improves the classification accuracy in all the classifiers. Irrespective of the best classifier for each dataset, our experiments prove that FBRA yields the best classification accuracy with the best classifier in each dataset by reducing the dimensionality to a greater extent. This is possibly due to the diversity in the selection of filtering metrics by rank aggregation. While other approaches might have missed out some contributing features based on their metric used, FBRA selects the contributing features effectively due to the rank aggregation.

5.3.2 Comparison with other rank aggregation approaches

The next set of experiments compares our proposed rank aggregation method FBRA with other rank aggregation approaches. FBRA is compared with four rank aggregation approaches—Borda, RRA (Kolde et al. 2012) and the two recently proposed aggregation approaches viz. ensemble feature selection using SA (Wang et al. 2019a) and MVFS



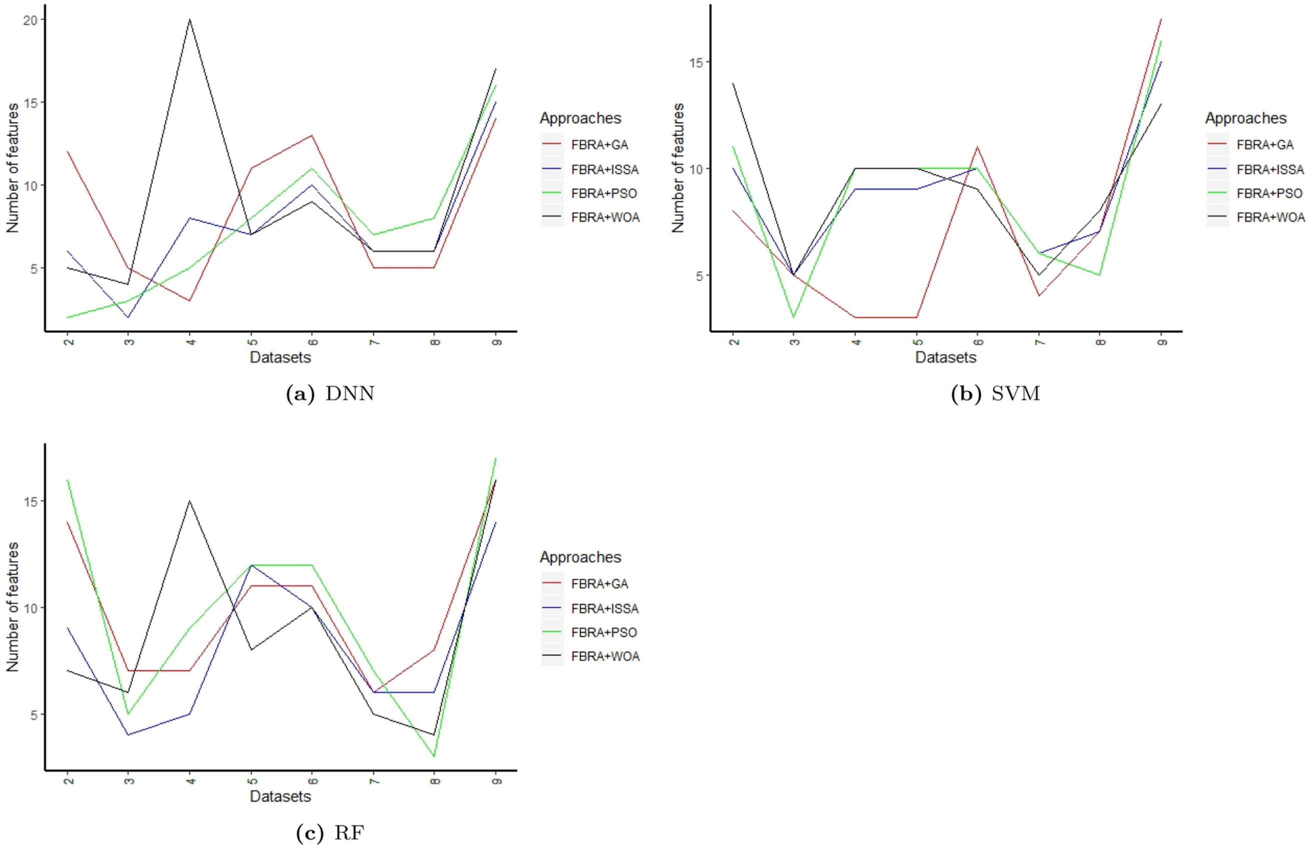
(a) DNN

(b) SVM

(c) RF

Fig. 6 Classification accuracy of optimization algorithms on different classifiers

(Abut et al. 2019). Borda aggregation method outputs the rank based on the mean position of each feature in different ranking methods (Najdi et al. 2016). RRA compares the results of different ranking methods with randomly generated ranking list. It compares the rank of each feature in different ranking methods and creates a rank order list where features are ranked based on the dominance of their ranks in rank order. A brief summary of Wang et al. (2019a) and Abut et al. (2019) are discussed in the earlier sections. Satopaa et al. (2011) is used to determine the cut off in selecting the optimal number of features for the rank aggregation methods used. The classification accuracy achieved by FBRA and other approaches used for comparison is shown in Tables 5, 6, and 7. The number of features selected by the rank aggregation approaches is mentioned in the brackets. It is evident from Tables 5, 6 and 7 that most of the rank aggregation approaches perform better than the individual approaches in terms of classification accuracy though not in terms of lesser number of features selected. Yet, there are some exceptions to this conclusion. For example, IG and mRmR perform better than the rank aggregation approach SA for dataset 9. MRMD exhibits superior performance than Borda and RRA for dataset 7. Low dimensional datasets such as dataset 7 and 9 don‘t induce much diversity in ranking the features by different metrics and hence rank aggregation approaches are not performing exceptionally well for these datasets. Indeed, this happens in high dimensional datasets too. For example, in dataset 1, NQF exhibits superior performance in terms of classification accuracy when compared with rank aggregation approaches Borda, RRA and SA. Yet, FBRA exhibits superior performance than all the individual and rank aggregation approaches. This happens because when the number of features is more, redundancy and relevancy play a major role and since individual filtering approaches are based on specific metrics, the best feature set is not identified properly. Many of the rank aggregation approaches also fail to consider redundancy and relevancy information during aggregation. But, the consideration of redundancy and relevancy information of the features in our model led to improved classification accuracy especially with high dimensional datasets. Moreover, the use of fuzzy systems for aggregation helps to rank the features efficiently in our proposed model by handling uncertainty in the ranks. From Tables 5, 6 and 7 it is clear that FBRA exhibits superior performance in terms of both classification accuracy and the number of features selected when compared with the other



(a) DNN

(b) SVM

(c) RF

Fig. 7 Number of features selected by optimization algorithms on different classifiers

rank aggregation approaches. But, the performance difference among the rank aggregation approaches deteriorates as the number of features decreases. For example, in dataset 8, MVFS performs equally well as FBRA for DNN classifier though FBRA exhibits superior performance than MVFS for RF which is the best classifier for this dataset. It is also evident from the Tables 5, 6 and 7 that MVFS is the second-best performer next to FBRA in terms of classification accuracy. This is because, though MVFS ranks the features based on majority voting, it calculates correlation score when there is a tie. The performance of SA also improves with low dimensional datasets. Borda and RRA perform equally well but their classification accuracy is comparatively lesser than other approaches. The poor performance of Borda owes to its dependency on all the feature ranking methods used from best to worst. RRA, SA and MVFS work on the basis of

Table 8 Comparison of H-FBRA + ISSA with state-of-the-art methods—classification accuracy

Dataset	Jain et al. (2018)	Apolloni et al. (2016)	Bonilla-Huerta et al. (2015)	H-FBRA + ISSA	Original number of features
1	94 [47]	95 [67]	97[94]	**98.9 [35]**	20531
2	99.3 [**5**]	96 [9]	95.8 [11]	**100** [10]	7129
3	99.4 [3]	**100 [2]**	100 [24]	**100 [2]**	4026
4	98.7 [7]	93.8 [**3**]	99 [9]	**100** [8]	2000
5	94 [8]	97[9]	98 [**7**]	**99.3** [10]	1928
6	99 [11]	99.2 [**9**]	100 [12]	**100** [10]	24
7	91 [6]	90 [**5**]	92 [8]	**94** [6]	12
8	87 [**5**]	90 [9]	89.1 [6]	**93** [7]	13
9	92.4 [**12**]	95.6 [15]	**100** [19]	**100** [14]	31

The bold values represent the best solution corresponding to the evaluated metric

the position of ranks in each feature subset. FBRA exhibits superior performance as global redundancy is considered and the use of linguistic fuzzy model handles uncertainty efficiently in ranking the features. Moreover, FBRA discards the low-rank features completely leading to lesser number of features than other rank aggregation approaches that work based on cutting rule technique.

5.3.3 Comparison with other optimization algorithms

To evaluate the performance of our hybrid model H-FBRA+ISSA, we conducted an experiment to compare H-FBRA + ISSA with few other well-known optimization algorithms. GA, PSO, and WOA are used along with FBRA to build the hybrid approaches. The performance of these hybrid approaches is evaluated against the performance of H-FBRA+ISSA. The parameter settings for these algorithms are discussed in the previous section. The classification accuracy of different hybrid models on each classifier is shown in Fig. 6. From the Fig. 6, it is evident that the classification accuracy has improved in all the datasets when hybrid models are used. The dimension of the datasets is also reduced to a greater extent especially with high dimensional datasets. This is expected in hybrid models as high-quality features selected by the filtering or rank aggregation of filters are fed to the wrapper algorithm to yield the best classification accuracy with minimum number of features. It is evident from Fig. 6 that H-FBRA + ISSA exhibits superior performance against other hybrid models on DNN classifier except for dataset 7. FBRA+WOA exhibits superior performance than H-FBRA + ISSA for dataset 7. But RF is the best classifier for dataset 7 and H-FBRA + ISSA is the second-best performer in RF for this dataset. Dataset 7 is very low dimensional dataset with only 12 features and the difference in the use optimization algorithms is not expected to have a huge impact on it. FBRA+PSO also yields good results in this dataset. The classification accuracy achieved by FBRA+GA and FBRA+PSO is lesser when compared with FBRA + ISSA for all the other datasets. This is because of few specific characteristics of ISSA in its working principle. ISSA uses three different strategies to update its solutions. Moreover, ISSA uses behaviorally inspired random variations using gliding distance. The concept of predator presence probability improves the exploration capability of ISSA. The seasonal conditioning concept in ISSA also prevents it from converging into local optimal solution which is not present in other optimization algorithms (Jain et al. 2019). WOA is the second-best performer in most of the cases and is also found to perform equally well with ISSA but when the dimensions increase, ISSA performs better

Table 9 Comparison of average execution time in seconds

Dataset	Jain et al. (2018)	Apolloni et al. (2016)	Bonilla-Huerta et al. (2015)	H-FBRA + ISSA
1	5900	4200	3900	1058
2	208	195	177	121
3	347	280	217	201
4	102	95	67	32
5	123	108	72	35
6	32	27	19	11
7	16	13	9	8
8	18	11	8	6
9	31	21	16	13

Table 10 Subset of features selected by H-FBRA + ISSA

Dataset no	Dataset	Subset of features
1	Pancan	ZNF193, SF3A3, UBE2Z, G0, AGTPBP1, B4GALT3, RIPPLY1, DLGAP3, LOC399815, MAK16, NDUFS6, LPIN2, MBOAT2, ADRA1B, RNF185, ARL2BP, RIPPLY 1, .
2	Cancer Gene	4847(Zyxin), 804(Macmarcks), 1882(CST3 Cystatin C), 6855(TCF3 Transcription factor 3) 6919(RNS2 Ribonuclease 2),2348(ACADM Acyl-Coenzyme A dehydrogenase),461, 1962, 5552, 2131
3	Lymphoma	390,3066
4	Colon	377(H.sapiens mRNA for GCAP-II/uroguanylin precursor), 765, 590, 384, 266, 1058(H.sapiens a-L-fucosidase gene),1541, 1873(Human MXI1 mRNA, complete cds.)
5	MicroRNAs	miR-505-5p, miR-125b-5p, miR-21-5p, miR-96-5p,miR-3613-3p,miR-4668-5p,miR-4516,miR-3656,miR-4488,miR-5704
6	Chronic kidney	Specific gravity, albumin, blood glucose random, potassium, haemoglobin, packed cell volume, red blood cell count, hypertension, serum creatinine, anaemia
7	Spine disease	Degree spon, pelvic incidence, lumbar angle, pelvic rad and pelvic tilt, cervical tilt
8	Heart disease	Age, fasting blood sugar, Trestbps, Cholestrol, Thal, Slope,Cp
9	Cancer disease	Perimeter-largest-worst,area-largest-worst,smoothness-largest-worst,compactness-mean, concave-points-mean, texture-largest-worst, texture-mean, symmetry-mean, concavity-largest-worst, concavity-mean, fractal-dimension-largest-worst, perimeter-mean, radius-se, area-mean

than WOA. This is due to the fact that the balance between the exploration and exploitation capability of ISSA is better than WOA thereby leading to faster convergence.

Figure 7 shows the number of features selected by the different hybrid models. Though H-FBRA + ISSA doesn‘t yield the best result with respect to a lesser number of features for all the datasets, the dimensionality of the datasets is reduced to a greater extent by H-FBRA+ISSA. Moreover, the difference in the number of features selected by the different hybrid models is also very meager. Hence, H-FBRA + ISSA reduces the dimensionality to a greater extent with better classification accuracy.

5.3.4 Comparison with state-of-the-art methods

Our proposed model H-FBRA + ISSA is compared with three state of the art methods to evaluate its performance—a hybrid framework with multiple filters and embedded approach for efficient feature selection in microarray data (Bonilla-Huerta et al. 2015), two hybrid wrapper filter feature selection algorithms (Apolloni et al. 2016) and correlation feature Selection based improved-Binary PSO for Gene Selection and Cancer Classification (Jain et al. 2018). Bonilla-Huerta et al. (2015) uses five statistical measures like sum square, within square, MI, signal to noise ratio, Wilcoxon test, T-statistic and then uses fusion method to combine and calculate the fusion score. The initial gene set is selected according to the fusion score and GA with Tabu search is used to improve the performance for feature selection with the initial gene subset. Apolloni et al. (2016) uses IG the ranking method and Binary differential evolution is used as a wrapper approach for feature selection. Jain et al. (2018) uses Multivariate filter technique Correlation-based feature selection to select the initial gene subset and an improved BPSO for gene optimization. Table 8 shows the results of our experimentation. The classification accuracy obtained by the best classifier for each dataset is considered. The table values represent the classification accuracy and the figures in the brackets indicate the number of features selected.

The best classification accuracy and the least number of features selected for each dataset are represented in bold. It is obvious that our method yields the best classification accuracy in comparison with the other state-of-the-art methods. In few datasets, such as datasets 3 and 7, other methods also perform equally well but the classification accuracy achieved by our proposed model is not lesser than these methods for these datasets too. The salient features of our proposed model help to yield better classification accuracy than other state-of-the-art methods. Bonilla-Huerta et al. (2015) achieves poorer results than our proposed model probably because redundancy and relevancy metrics are not considered by this model. Though Apolloni et al. (2016) uses hybrid approach, it yields poorer results than our model because no local search techniques are used to improve the exploitation capability in Apolloni et al. (2016). Moreover, it is biased towards the metric IG whereas our model uses a rank aggregation method to obtain the best quality features. Jain et al. (2018) is again biased towards the metric correlation and hence failed to consider contributing features based on other metrics.

But our method is not the best in selecting the least number of features. Yet, it has reduced the dimensionality of the datasets to a greater extent which is evident from Table 8. The minimum dimensionality reduction achieved by our proposed model is for dataset 8 and the maximum is for dataset 3.

In Table 9, we have reported the average execution time (in seconds) for each of the models on nine datasets. It can be summarized that our proposed H-FBRA + ISSA model takes less execution time than the other models. This happens because our rank aggregation approach effectively reduces the dimension of the datasets as the low-rank features are discarded completely. Hence, the search space in which our wrapper algorithm ISSA operates is lesser than the search space used by wrapper algorithms in other state-of-the-art methods (Bonilla-Huerta et al. 2015; Apolloni et al. 2016) and Jain et al. (2018). This reduces the computational time of our proposed model. Moreover, a proper balance between the exploitation and exploration capability of ISSA helps to converge faster. Thus our proposed model ensures faster and more reliable feature selection for classification process without increasing complexity.

Table 10 shows the subset of features selected for the different datasets by our proposed model. The complete sets of selected features are shown in the Table 10 for all the datasets except for dataset 1. As thirty-five features are selected for this dataset, we have shown around fifteen features in the table and the list continues. The biological interpretation is given for most of the genes and feature number is shown for the others.

Hence, our proposed model is a generalized feature selection model for different kinds of classification problems with biomedical datasets (e.g., prediction of the type of a disease like cancer, diagnose a disease using genomic dataset, etc.). The salient features of our proposed method include its capability to work with mixed feature types (e.g., categorical, discrete, continuous) and also with different classifiers. Our model has also proved its performance with different state-of-the-art methods. Use of ISSA algorithm with an enhanced exploitation and exploration capability yields faster convergence rate when compared with other optimization algorithms.

Yet, as suggested by the reviewers, we like to extend this subsection by discussing our proposed model in comparison with few other state-of-the-art works. Many other research works are also proposed with embedded approaches and neural networks for feature selection in high dimensional biomedical classification tasks. For instance, an embedded approach is proposed that uses binary coral reefs optimization algorithm with simulated annealing to select the features in high-dimensional biomedical datasets (Yan et al. 2019). But this approach does not use any filtering metric to select the initial feature subset to be provided as input to the optimization algorithm. Hence, the optimization algorithm is expected to search on the complete feature space that may incur high computational time than our proposed model. Another embedded approach that uses grasshopper optimization algorithm to select the best features and optimize the parameters for SVM is proposed for biomedical datasets (Ibrahim et al. 2019). But this approach is tested only on two datasets—Iraq cancer patients dataset and University of California Irvine datasets. Moreover, it suffers from the same drawback of computational complexity as the search space will be very high especially in case of genomic datasets. Neural networks, specifically deep networks are proved to be one of the best classifier for high-dimensional biomedical classification tasks. Though deep neural networks perform in-built feature extraction and yield better classification accuracy, it suffers from the major drawback of transparency. As discussed in the earlier section of this paper, feature extraction is a black box technique where the selected features cannot be interpreted. Hence, deep neural networks suit well for image classification tasks where major interpretation is not required but for certain biomedical data classification tasks (such as classification of disease) which requires to identify the major contributing features, application of feature selection techniques like our proposed model helps in better interpretation. It has also been proved that use of feature selection techniques before deep neural networks help to improve the classification performance in bioinformatics (Chen et al. 2020). Hence, though deep neural networks perform well on feature abstraction, there is a scope for performance improvement using feature selection models. Few state of the works also use feature selection models with deep neural networks. For instance, a work is proposed that applies mRmR feature selection over the extracted features from DNN to improve the classifier's accuracy in detection of lung cancer(Toğaçar et al. 2020). Another such work uses binary gray wolf optimization and binary particle swam optimization over the extracted features from DNN to select the best feature subset for diagnosis of COVID-19 (Canayaz 2021). Our experimentation also uses our proposed feature selection model H-FBRA + ISSA on deep neural networks and the results show that DNN classifier yields good classification accuracy.

5.3.5 Evaluating and validating results

All the experiments on classifiers are performed using tenfold cross-validation to avoid overfitting (Kohavi 1995). The experiments are also repeated five times to avoid bias in the results and the average of the results is considered. Oversampling is performed in the imbalanced dataset, dataset 5 to avoid bias. To compare the results statistically, a paired t test is performed at 95% confidence interval on classification accuracy. The results confirmed that our proposed model H-FBRA+ISSA outperforms other methods used for comparison. The experiments are implemented in R.

6 Limitations and scope for future work

The results depict that our proposed model is an efficient feature selection method for biological classification tasks. Nevertheless, there is more scope for future work in our proposed model. This work can be extended to high dimensional data of different domains to study the generalization of the proposed model. The use of linguistic hedges to the fuzzy inference system can also be explored in the future. Our proposed model works exclusively on the specific datasets. In future, our proposed model can be integrated with Ontology for better biological interpretations. Different variations of our proposed model can be tried by updating or extending the rule base in fuzzy aggregation. This can be tried in accordance with the input from the domain experts.

The datasets we considered in our experimentation are exclusively used for classification tasks such as predicting the presence or absence of a disease, classifying the type of disease and classifying normal and abnormal tissues. Yet, our proposed model can be applied for drug-related classification tasks too. For example, our proposed model can be used to select the best features from a set of patient-related features (such as age, sex, blood pressure, cholestrol, sodium to potassium ratio) to predict the outcome of a drug in a particular patient. Our model can also be used to select the best features in predicting alcohol-abused patients (Kumari et al. 2018). The performance of our proposed model can be explored for such datasets and classification tasks in the future. Besides, this study can be extended further for drug discovery problems. For instance, tasks, such as detection of active components, prediction of drug-target protein interactions, can be converted to classification task and our proposed model can be applied on it to study their performance.

The parameters of the ISSA and SSA algorithms are set from the original literatures in our proposed model. Though our proposed model yields satisfactory performance, future studies can analyze the performance of the model with different parameter values.

7 Conclusion

In this paper, we proposed a computationally efficient Hybrid model (H-FBRA + ISSA) for feature selection in biomedical data classification tasks. This model combines linguistic fuzzy rule-based rank aggregation and ISSA algorithm for efficient feature selection. FBRA aggregates ranks from different filtering metrics using linguistic fuzzy model and discards the least significant features. The subset of features selected by FBRA is fed as input to ISSA algorithm. ISSA optimization algorithm selects the final subset of features that yields the best classification accuracy. We conducted extensive experiments on both high-dimensional and low-dimensional datasets with three different classifiers. Our proposed model is compared with individual filtering metrics, rank aggregation methods, other optimization algorithms and state-of-the-art methods. The experiments show that H-FBRA + ISSA outperforms other models in terms of classification accuracy and computational time. It is also proved to reduce the dimensionality to a greater extent. Thus, the proposed model could be an efficient feature selection technique for biomedical data classification.

References

Abut F, Akay MF, George J (2019) A robust ensemble feature selector based on rank aggregation for developing new vo (2) max prediction models using support vector machines. Turkish J Electr Eng Comput Sci 27:3648–3664

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X et al (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature 403(6769):503–511

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 96(12):6745–6750

Alshamlan H, Badr G, Alohali Y (2015) mrmr-abc: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. Biomed Res Int 9:1–15

Apolloni J, Leguizamón G, Alba E (2016) Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments. Appl Soft Comput 38:922–932

Basu M (2019) Squirrel search algorithm for multi-region combined heat and power economic dispatch incorporating renewable energy sources. Energy 182:296–305

Bennasar M, Hicks Y, Setchi R (2015) Feature selection using joint mutual information maximisation. Exp Syst Appl 42(22):8520–8532

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2012) An ensemble of filters and classifiers for microarray data classification. Pattern Recogn 45(1):531–539

Bolon-Canedo V, Marono NS, Betanzos AA (2014) Data classification using an ensemble of filters. Neurocomputing 135:13–20

Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2015) Distributed feature selection: an application to microarray data classification. Appl Soft Comput 30:136–150

Bonilla-Huerta E, Hernandez-Montiel A, Morales-Caporal R, Arjona-López M (2015) Hybrid framework using multiple-filters and an embedded approach for an efficient selection and classification of microarray data. IEEE/ACM Trans Comput Biol Bioinform 13(1):12–26

Canayaz M (2021) Mh-covidnet: diagnosis of covid-19 using deep neural networks and meta-heuristic-based feature selection on x-ray images. Biomed Signal Process Control 64:102257

Canedo VB, Marono NS, Betanzos AA (2013) A review of feature selection methods on synthetic data. Knowl Inform Syst 34:483–519

Chen R, Sun N, Chen X, Yang M, Wu Q (2018) Supervised feature selection with a stratified feature weighting method. IEEE Access 6:15087–15098

Chen Z, Pang M, Zhao Z, Li S, Miao R, Zhang Y, Feng X, Feng X, Zhang Y, Duan M et al (2020) Feature selection may improve deep neural networks for the bioinformatics problems. Bioinformatics 36(5):1542–1552

Chinnaswamy A, Srinivasan R (2016) Hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data. In: Innovations in bio-inspired computing and applications, Springer, New York, pp 229–239

Chinnaswamy A, Srinivasan R (2017) Hybrid information gain based fuzzy roughset feature selection in cancer microarray data. In: 2017 Innovations in power and advanced computing technologies (i-PACT), IEEE, pp 1–6

Dahiya S, Handa S, Singh N (2016) A rank aggregation algorithm for ensemble of multiple feature selection techniques in credit risk evaluation. Int J Adv Res Artif Intell 5(9):1–8

del Río S, López V, Benítez JM, Herrera F (2015) A mapreduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules. Int J Comput Intell Syst 8(3):422–437

Fernandez A, del Rio S, Bawakid A, Herrera F (2017) Fuzzy rule based classification systems for big data with mapreduce: granularity analysis. Adv Data Anal Classif 11:711–730

Ebrahimpour MK, Eftekhari M (2018) Distributed feature selection: a hesitant fuzzy correlation concept for microarray high-dimensional datasets. Chemom Intell Lab Syst 173:51–64

Fiorini S (2016) Pancan dataset source. https://www.synapse.org/#!Synapse:syn4301332

Foitong S, Rojanavasu P, Attachoo B, Pinngern O (2009) Estimating optimal feature subsets using mutual information feature selector and rough sets. In: Pacific-Asia conference on knowledge discovery and data mining, Springer, New York, pp 973–980

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar):1157–1182

Han M, Ren W (2015) Global mutual information-based feature selection approach using single-objective and multi-objective optimization. Neurocomputing 168:47–54

Hoque N, Bhattacharyya D, Kalita J (2014) Mifs-nd: a mutual information-based feature selection method. Exp Syst Appl 41(14):6371–6385

Hoque N, Singh M, Bhattacharyya DK (2018) Efs-mi: an ensemble feature selection method for classification. Complex Intell Syst 4:105–118

Hsu HH, Hsieh CW et al (2010) Feature selection via correlation coefficient clustering. JSW 5(12):1371–1377

Hu H, Zhang L, Bai Y, Wang P, Tan X (2019) A hybrid algorithm based on squirrel search algorithm and invasive weed optimization for optimization. IEEE Access 7:105652–105668

Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G, Buckland M (2016) A hybrid feature selection with ensemble classification for imbalanced healthcare data: a case study for brain tumor diagnosis. IEEE Access 4:9145–9154

Ibrahim HT, Mazher WJ, Ucan ON, Bayat O (2019) A grasshopper optimizer approach for feature selection and optimizing svm parameters utilizing real biomedical data sets. Neural Comput Appl 31(10):5965–5974

Inza I, Larranaga P, Saeys Y (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517

Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell 22(1):4–37. https://doi.org/10.1109/34.824819

Jain I, Jain VK, Jain R (2018) Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. Appl Soft Comput 62:203–215

Jain M, Singh V, Rani A (2019) A novel nature-inspired algorithm for optimization: squirrel search algorithm. Swarm Evol Comput 44:148–175

Kim JC, Chung K (2017) Depression index service using knowledge based crowdsourcing in smart health. Wirel Pers Commun 93(1):255–268

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324

Kohavi R et al (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai, Montreal, Canada 14:1137–1145

Kolde R, Laur S, Adler P, Vilo J (2012) Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics 28(4):573–580

Koprinska I, Rana M, Agelidis VG (2015) Correlation and instance based feature selection for electricity load forecasting. Knowl Based Syst 82:29–40

Kumari D, Kilam S, Nath P, Swetapadma A (2018) Prediction of alcohol abused individuals using artificial neural network. Int J Inform Technol 10(2):233–237

Liu J, Lin Y, Lin M, Wu S, Zhang J (2017) Feature selection based on quality of information. Neurocomputing 225:11–22

Liu J, Lin Y, Li Y, Weng W, Wu S (2018) Online multi-label streaming feature selection based on neighborhood rough set. Pattern Recogn 84:273–287. https://doi.org/10.1016/j.patcog.2018.07.021

Low YS, Gallego B, Shah NH (2016) Comparing high-dimensional confounder control methods for rapid cohort studies from electronic health records. J Compar Effective Res 5(2):179–192

Maji P, Pal SK (2009) Feature selection using f-information measures in fuzzy approximation spaces. IEEE Trans Knowl Data Eng 22(6):854–867

Matamala N, Vargas MT, Gonzalez-Campora R, Minambres R, Arias JI, Menendez P, Andres-Leon E, Gomez-Lopez G, Yanowsky K, Calvete-Candenas J et al (2015) Tumor microrna expression profiling identifies circulating micrornas for early breast cancer detection. Clin Chem 61(8):1098–1106

Mirjalili S, Lewis A (2016) The whale optimization algorithm. Adv Eng Softw 95:51–67

Nagarajan G, Babu LD (2019) A hybrid of whale optimization and late acceptance hill climbing based imputation to enhance classification performance in electronic health records. J Biomed Inform 94:103190

Najdi S, Gharbali AA, Fonseca JM (2016) A comparison of feature ranking and rank aggregation techniques in automatic sleep stage classification based on polysomnographic signals. In: International conference on bioinformatics and biomedical engineering, Springer, New York, pp 230–241

Nguyen TT, Nguyen MP, Pham XC, Liew AWC (2018) Heterogeneous classifier ensemble with fuzzy rule-based meta learner. Inform Sci 422:144–160

Pardo BS, Diaz IP, Canedo VB, Betanzos AA (2017) Ensemble feature selection: Homogeneous and heterogeneous approaches. Knowl Based Syst 118:124–139

Pawlak Z (1982) Rough sets. Int J Comput Inform Sci 11(5):341–356

Princy S, Dhenakaran S (2016) Comparison of triangular and trapezoidal fuzzy membership function. J Comput Sci Eng 2(6):46–56

Qian Y, Liang J (2008) Combination entropy and combination granulation in rough set theory. Int J Uncertain Fuzziness Knowl Based Syst 16(02):179–193

Rahim R et al (2017) Comparative analysis of membership function on mamdani fuzzy inference system for decision making. J Phys Conf Ser 930:012029

Saletic DZ, Popovic U (2006) On possible constraints in applications of basic defuzzification techniques. In: 2006 8th seminar on neural network applications in electrical engineering, pp 225–230. https://doi.org/10.1109/NEUREL.2006.341218

Satopaa V, Albrecht J, Irwin D, Raghavan B (2011) Finding a “kneedle” in a haystack: Detecting knee points in system behavior. In: 2011 31st international conference on distributed computing systems workshops, IEEE, pp 166–171

Senawi A, Wei HL, Billings SA (2017) A new maximum relevance-minimum multicollinearity mrmmc method for feature selection and ranking. Pattern Recogn 67:47–61

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Sharbaf FV, Mosafer S, Moattar MH (2016) A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Genomics 107(6):231–238

Shardlow M (2016) An analysis of feature selection techniques. https://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf

Shreem SS, Abdullah S, Nazri MZA (2016) Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. Int J Syst Sci 47(6):1312–1329

Smetannikov I, Deyneka A, Filchenkov A (2016) Meta learning application in rank aggregation feature selection. In: 2016 3rd international conference on soft computing and machine intelligence (ISCMI), IEEE, pp 120–123

Suo M, Zhang Z, Chen Y, An R, Li S (2019) Knowledge acquisition and decision making based on bayes risk minimization method. Appl Intell 49(2):804–818

Tal I, Muntean GM (2012) Using fuzzy logic for data aggregation in vehicular networks. In: 2012 IEEE/ACM 16th international symposium on distributed simulation and real time applications, IEEE, pp 151–154

Toğaçar M, Ergen B, Cömert Z (2020) Detection of lung cancer on chest ct images using minimum redundancy maximum relevance feature selection method with convolutional neural networks. Biocybern Biomed Eng 40(1):23–39

Tomar D (2015) Agarwal S (2015) Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes. Adv Arti Neural Syst. https://doi.org/10.1155/2015/265637

Uraon KK, Kumar S (2016) Analysis of defuzzification method for rainfall event. Int J Comput Sci Mobile Comput 5(1):341–354

Waad B, Ghazi BM, Mohamed L, LARODEC I, LARIME E (2014) A new feature selection technique applied to credit scoring data using a rank aggregation approach based on: optimization, genetic algorithm and similarity. In: Knowledge discovery process and methods to enhance organisational performance , pp 347–376

Wang D, Nie F, Huang H (2015) Feature selection via global redundancy minimization. IEEE Trans Knowl Data Eng 27(10):2743–2755

Wang Y, Feng L (2018) Hybrid feature selection using component co-occurrence based feature relevance measurement. Exp Syst Appl 102:83–99

Wang Y, Feng L (2019) A new hybrid feature selection based on multi-filter weights and multi-feature weights. Appl Intell 49(12):4033–4057

Wang J, Xu J, Zhao C, Peng Y, Wang H (2019a) An ensemble feature selection method for high-dimensional data based on sort aggregation. Syst Sci Control Eng 7(2):32–39

Wang P, Kong Y, He X, Zhang M, Tan X (2019b) An improved squirrel search algorithm for maximum likelihood doa estimation and application for mems vector hydrophone array. IEEE Access 7:118343–118358

Wang Y, Shang D, Yuan X (2019c) A correction method for the proportion of key components in basic hysys library based on an improved squirrel search algorithm. In: 2019 12th Asian Control Conference (ASCC), IEEE, pp 236–241

Xu F, Miao D, Wei L (2009) Fuzzy-rough attribute reduction via mutual information with an application to cancer classification. Comput Math Appl 57(6):1010–1017

Xu J, Tang B, He H, Man H (2016) Semisupervised feature selection based on relevance and redundancy criteria. IEEE Trans Neural Netw Learn Syst 28(9):1974–1984

Yan C, Ma J, Luo H, Patel A (2019) Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. Chemom Intell Lab Syst 184:102–111

Yang F, hang Lu W, kai Luo L, Li T (2012) Margin optimization based pruning for random forest. Neurocomputing 94:54–63

Yang SM, Yan YM, Wang K, Xie ZY (2014) A new improved attribute weight algorithm based on rough sets theory for one command information system. Adv Mater Res 989:2029–2032

Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst Man Cybern 1:28–44

Zeng Z, Zhang H, Zhang R, Zhang Y (2014) A hybrid feature selection method based on rough conditional mutual information and naive bayesian classifier. ISRN Appl Math. https://doi.org/10.1155/2014/382738

Zheng T, Luo W (2019) An improved squirrel search algorithm for optimization. Complexity. https://doi.org/10.1155/2019/6291968

Zheng Y, Li G, Zhang W, Li Y, Wei B (2019) Feature selection with ensemble learning based on improved dempster-shafer evidence fusion. IEEE Access 7:9032–9045

Zou Q, Zeng J, Cao L, Ji R (2016) A novel features ranking metric with application to scalable visual and bioinformatics data classification. Neurocomputing 173:346–354

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.