Applied Soft Computing 49 (2016) 407–422

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A novel memetic algorithm for discovering knowledge in binary and multi class predictions based on support vector machine

S. Sasikala (Associate Professor)[a,*], S. Appavu alias Balamurugan (Dr., Professor and Head)[a], S. Geetha (Dr., Professor)[b]

[a] Department of I.T., K.L.N. College of Information Technology, Tamil Nadu, India
[b] School of Computing Science and Engg., VIT University – Chennai Campus, Tamil Nadu, India

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 14 October 2015
Accepted 21 August 2016
Available online 25 August 2016

Keywords:
Data mining
Classification
Svega-svm
Feature selection
Shapley values
Genetic algorithm
Memetic algorithm

ABSTRACT

In classification, every feature of the data set is an important contributor towards prediction accuracy and affects the model building cost. To extract the priority features for prediction, a suitable feature selector is schemed. This paper proposes a novel memetic based feature selection model named Shapely Value Embedded Genetic Algorithm (SVEGA). The relevance of each feature towards prediction is measured by assembling genetic algorithms with shapely value measures retrieved from SVEGA. The obtained results are then evaluated using Support Vector Machine (SVM) with different kernel configurations on 11 + 11 benchmark datasets (both binary class and multi class). Eventually, a contrasting analysis is done between SVEGA-SVM and other existing feature selection models. The experimental results with the proposed setup provides robust outcome; hence proving it to be an efficient approach for discovering knowledge via feature selection with improved classification accuracy compared to conventional methods.

1. Introduction

Mining the data to procure desired output is an important step in machine learning algorithms [1]. However, datasets consist of large number of features which are recorded during data entry for a given set of instances. It is observed that not all the features are required for classification and prediction. Hence, there arises a need to prioritise the features based on their relativeness. Dimensionality reduction technique attempts to mitigate the complexity of huge learning models.

Most real-world classification problems deal with the class probabilities of the instances, features and their associated class labels. The features may be relevant, irrelevant and redundant. The presence of irrelevant and redundant features will affect the learning accuracy of the classifier. Hence the key idea of the feature selection process is to select only the relevant features and get rid of the irrelevant features and redundant features. Reducing the number of irrelevant/redundant features can drastically reduce the running time of the learning algorithms and yield a more general classifier. This helps in gaining a better insight into the underlying concept of a real-world classification problem. In many classification problems, it is even difficult to train good classifiers without removing these unwanted features from the huge dataset.

Feature selection plays a crucial role in data analysis by selecting the subset of features which have a greater contribution ratio. It not only leads to enriched results which guarantee better accuracy but also dispatches the results at lower computational cost. As the nature of classes vary for each dataset, the feature selection method also needs to be chosen appropriately to generate results which do not underestimate the important features. Also, due to fluctuating dimensionality of datasets under study, the feature selection model must be capable of pruning the features which have little or no priority [2,3].

Feature selection methods aim to reduce the burden of the classifier. Classifier will operate only on the optimal features selected, instead of operating on the whole set of features. The inclusion of the classifier in the feature selection process is decided by the choice of feature selection method i.e. either filter or wrapper method.

On a general note, the selected features subset used for classification must satisfy the following conditions:

* Corresponding author.
E-mail addresses: nithilannsasikala@yahoo.co.in (S. S.), app_s@yahoo.com (A.a.B. S.), geethabaalan@gmail.com (G. S.).

http://dx.doi.org/10.1016/j.asoc.2016.08.038
1568-4946/

- For all the selected features the resulting class distribution must coincide with original class distribution.
- The features in the selected subset must always aim in increasing the classification accuracy.

The main issues to be addressed by the feature selection algorithm from the classification perspective are (i) dimensionality reduction, (ii) class imbalance and (iii) choice of classifier for handling the multi class problems and both linear and non-liner data.

1.1. Dimensionality reduction

The high dimensional information is commonly represented by a very large number of highly correlated features [13]. Usually the dimensionality reduction is carried out by either feature selection or feature extraction techniques [18]. Both the approaches aim in reducing the number of features without affecting classifier's accuracy. Feature selection approach uses the statistical measures to choose the best subset of features from the original data. The feature extraction methods solve the problem of dimensionality reduction by projecting the data from the original feature space onto a low-dimensional subspace, which contains most of the original information.

1.2. Class imbalance

In many real-world applications it has been observed that class imbalance often leads to poor classification performance. A data set is said to be imbalanced when the minority classes are greatly suppressed relative to the other majority classes. Also class imbalance occurs when the samples of the minority and majority classes usually represent the presence and absence of positive and negative samples. Although class imbalance has been extensively studied for binary classification problems in the last decades, only very few approaches deal with imbalanced multi-class data sets.

1.3. Choice of classifiers

Classifiers are generally categorized as linear and non-linear classifier. A linear classifier makes a classification based on the significance of a linear combination of the features. Some of the linear classifiers are naive bayes classifier, fisher's linear discriminant, logistic regression etc. A separation between the linear and non-linear classifier is accomplished by the hyper-plane that has the maximum distance to the nearest training data point of any class. In general the larger the margin, lower is the generalization error of the classifier.Some of the classifiers like decision trees have a usual extension to handle both binary class and multi class problems. Contrarily the most popular classifier like SVM has the problem of unable to be easily extensible to multi class problems.

Apart from these general issues, additional caveats that arise when dealing with the multi class problems include:

1. Whether to carry out multi classifications as a chain of binary classifications or not.
2. The initial starting point to perform a search among the different classes.
3. Dealing withnil-sampled or ill-sampled class cases in the multi classification problems

However, class binarization – decomposition of multi class problem to a several simpler binary class problems [43] can be done to combat these issues.

Though feature selection has been researched upon since the inception of data mining and numerous studies have been conducted, this field always has been refined and modified as per the evolving requirements of the user.

This research work is aimed at developing an efficient feature selection model that meets out multiple contradictory objectives – selecting the most relevant features, increasing the classifier's accuracy, and reducing the dimensionality of the dataset. This uses a novel feature selection approach – which is Shapley Value Analysis Embedded Genetic Algorithm (SVEGA), to address all the caveats of a feature selection model detailed earlier. The proposed approach is tested on SVM classifier model with different kernel setup, to show the efficacy of the model. Further, this research paper is narrowed down to bio medical health care domain with a motive of providing an efficient data mining solution to health care industry to yield efficient and effective diagnosis, which may help out in faster treatment with greater accuracy.

The succeeding sections of the paper are as follows-the Section 2 supplies an overview of the related studies that have been conducted. Section 3 deals with the proposed system and the methodologies recommended. It describes the rationale for the choice of various components in the algorithms. Section 4 presents the experimental results and analysis while with Section 5 provides the concluding remarks about the study.

2. Related work

At the outset, the feature selection algorithms are classified into – *supervised* and unsupervised methods. Supervised feature selection methods can further be broadly categorized into *filter* models, *wrapper* models and *embedded* models. In filter based feature selection methods, the bias of feature selection algorithm does not interact with the bias of a learning algorithm. Instead, it depends on the general measures (dependency, distance, consistency, correlation and information) of the training set. Some of the filter based feature selection algorithms are Information gain, GainRatio, Relief, ReliefF, Symmetrical Uncertainty, Fisher score etc. The wrapper model makes use of the predictive accuracy of a preset learning algorithm to determine the excellence of selected features. Wrapper methods are more expensive than the filter model due to processing of large and high dimensional data in classification. Due to these deficiencies of each model and in order to bridge the gap between the filter and wrapper model, an embedded model is proposed.

Embedded model performs feature selection in two steps. In the first step it acts as filter method to select the various feature subset based on the cardinality, where as in the second step it acts like wrapper method to choose the best subset which results in highest classification accuracy. In this way the wrapper based feature selection methods attain the strength of both filter and wrapper model. In other words, it accomplishes the model building and feature selection simultaneously [40].

Feature selection algorithms which perform feature selection on dataset with the absence of class labels, but depend on clustering quality measures are collectively called as *unsupervised* feature selection methods. This also results in providing many valid feature subsets. The drawback of this method is that it faces difficulty in handling high dimensional data due to its implausibility in finding the relevant features. A detailed description about the unsupervised feature selection can be found in [41]. Though unsupervised feature selection algorithms work with un-labeled data and consumes less time they fail to find the relevant features when dealing with high dimensional data. Always a good feature selector needs a labeled data even though it is time consuming to construct the model.

The combination of the characteristics of supervised and unsupervised feature selection algorithms manifests a new research

challenge of dealing both labeled and unlabeled datasets. Under the assumption, a new algorithm called semi-supervised feature selection makes use of both labeled and unlabeled data to estimate feature relevance [42].

All the three types of feature selection methods follow the four basic steps, namely subset generation, subset evaluation, stopping criterion, and result validation. A candidate feature subset will be chosen based on the search strategy in the subset generation step. In the second step, the subset generated is evaluated by specific evaluation criteria like classification accuracy, model error etc. The subset that best fits the evaluation criterion will be chosen among all the candidate's subsets that have been evaluated after the stopping criterion are satisfied. Finally, the best chosen subset will be validated by any learning methods.

In the following paragraphs, we detail the specific methods in each of these feature selection types. A feature selection method which uses SVM-RFE was proposed byDaassi-Gnaba et al. [6]. Two methods – external SVM-RFE and internal SVM-RFE methods for the feature selection,are applied for ranking the features for linear SVM classifiers for the purpose of recognizing the speaker emotions. Marina Sokolova et al. [7] made an analytical study by considering twenty four performance measures for effectively classifying the datasets with binary class, multi-class, multi-labeled, and hierarchical classification. This study resulted in variation of confusion matrix related to specific characteristics of data. Clinical dataset was studied using a combination of feature selection and classification methods [8]. The issues such as missing values, high dimensionality, and unbalanced classes for comprehending the underlying statistical characteristics of a typical clinical dataset were investigated. Supervised learning has confirmed to be more suitable for mining clinical data than unsupervised methods. Over the years, huge set of results have been presented, specifically dealing with the issue of feature selection and the development of models for diagnosing heart failure using data mining techniques by Shi et al. [9–11].

Improved F-score and Sequential Forward Search (IFSFS) [12] were proposed for feature selection to diagnose erythemato-squamous disease. This method was designed so as to improve the F-score and measured the discrimination between more than two sets of real numbers instead of measuring between only two sets of real numbers. The method's applicability to other medical data sets was not reported and hence it was a very specific system targeted at the diagnosis of erythemato-squamous disease only. Hybrid Feature Selection (HFS) [13] is presented for dimensional reduction by combining the filter and wrapper models. Hybrid schemes that combine wrapper-based and filter-based approaches are also in the literature. [18,19,25] Correlation Based Filter [14] is another strategy for feature selection.

Ganet al. [15] proposed the Filter-Dominating Hybrid Sequential Forward Feature Selection (FDHSFFS) algorithm for high dimensional feature subset selection. This method proved to be fast but demanded huge computational complexity. La Vinhet al. [16] proposed a novel feature selection method based on the normalization of the well-known mutual information measurement and utilized the information measurement to estimate the potential of the features. The method could not eclipse the impact of strongly correlated features on the classification results. Correlated features may be accounted for redundancy and hence a single representative feature from that subset may be selected for further processing.

Feature selection methods [20] tend to identify the features which are most relevant for classification and can be broadly categorized as either subset selection methods or ranking methods. The former type returns a subset of the original set of features which are considered to be the most important for classification. Ranking methods sort the features according to their usefulness in the classification task. Feng et al. [17] projected genetic algorithm (GA) for feature subset selection by including multiple feature selection criteria and find small subsets of features that perform well for the inductive learning algorithm for building the classifier. Their evaluation on the data sets resulted in higher classification accuracy. AKhan et al. [21] proposed an approach for solving multi-objective feature subset selection problem based on evolutionary algorithm. This approach applies multi-objective genetic algorithms (NSGA – II) as feature subset selection technique for solving multi-objective optimization problem. The fitness value of a particular feature subset is measured by using ID3 algorithm. This proposal revealed finally NSGA II as the best choice of performing feature selection. Yu-Jun Zheng et al. [22] made a survey about the advances in evolutionary algorithms (EAs) applied to disaster relief operations. This survey provided a detailed discussion by giving strengths, limitations and future directions in the area. SenthamaraiKannan et al. [23] presented a memetically framed novel hybrid feature selection algorithm (MA-C) for feature selection.

Ferreira [24] proposed a combined approach like unsupervised feature discretization and feature selection technique for performing the feature selection task. Uguz [26] made a study with the goal of classifying the transracial Doppler (TCD) signals with the hybrid combination of feature ranking (Information Gain – IG) and dimension reduction methods (Principal Component Analysis – PCA) to improve the classification efficiency and accuracy. The experimental results showed that using the IG and PCA methods as a hybrid improved the classification efficiency and accuracy compared with individual usage. NassimLaouti et al. [27] observed the effectiveness of radial basis function kernel based SVM for the fault detection and isolation in a variable speed horizontal axis wind turbine. This work where the sensor faults were treated by SVM was found to be a good method for pattern recognition and to be adapted for online implementation.

Nibaran Das et al. [28] made a comparison between Genetic Algorithm (GA), Simulated Annealing (SA) and Hill Climbing (HC) together with SVM based classifier to select an optimal group of local regions on recognition of handwritten Bangla digits dataset. Miguel Garcia-Torres et al. [29] presented a comparison between the familiar meta heuristics search techniques like best first (BF), genetic algorithm (GA), scatter search (SS) and variable neighbourhood search (VNS) for performing feature selection task to detect relevant peak bins in Mass spectrometry (MS) data.

Sarafrazi et al. [30] proposed a novel GA-SVM hybrid system to improve classification accuracy with an appropriate feature subset in binary problems. Evaluation on the several UCI machine learning benchmark datasets showed that the idea was capable of selecting the discriminating input features correctly and achieved high classification accuracy. Weiss et al. [31] found a novel algorithm for the feature selection named CASH (Cost-sensitive Attribute Selection algorithm using Histograms) for analysing the feature collection cost and misclassification costs. Evaluation on several datasets showed its superiority of CASH over other cost-sensitive genetic algorithms. Comprehensive descriptions of SVM and kernelmethods can be found in Burges [32], Tibshirani [33], Scholkopf and Smola [34], and Shawe-Taylor and Cristtianini [35]. Zhu and Hastie [36] designed an Import Vector Machine (IVM) based on kernellogistic regression, which is generally equivalent to SVM but has advantages in selecting fewer training points (import points) for computing and in its natural generalization to multi category classification. Zhu, Su and Chipman [37] showed that support vector machine is closely related to a radial basis network and could be used in rare-target detection problems. Tang and Zhang [38] and Liu and Shen [39] improve SVM performance on multi category classification.

Many feature selection algorithms have been proposed in the literature. All these methods search for optimal or near optimal subsets of features that optimize a given condition. Some feature

selection algorithms cannot delete the redundant features, which cause no improvement in the classification accuracy. Even though the filter model is fast, it fails to produce the optimal feature subset. The simplest form of feature selection, the wrapper models espouse the accuracy rate of the classifier as the performance measure.

This paper proposes a hybrid feature selection method reaping the advantages of all approaches with three-fold motivations – *abridging*the classifier by retaining only the relevant features; *improving* the accuracy of the classifier; and *reducing* the dimensionality of the data thus reducing the size of the features.

3. Proposed system and methodology

In this section, the various components of the algorithm, the rationale for their choice with necessary background details are presented.

3.1. Rationale for the choice of Shapley value analysis (SVA) for feature selection

Shapley Value Analysis (SVA) [46,47,48] is a game theory based technique that addresses the issue in describing and calculating the contributions made by the interactions among the group of elements in a data set with multiple features. In game theory, a cooperative game is a game where groups of players ("coalitions") may enforce cooperative behaviour and aim to obtain high total profit.

Consider a set of players denoted as N. Let N = | N | be the number of players in this set. Any non-empty set $S \subseteq N_p$ is referred as a coalition of players. Each coalition has a worth function denoted as $v(S)$, which calculates the total profit produced by the service when all the players in this coalition S are active. Let $P_i(S)$ represent the profit of player i in the coalition S, then $v(S)$ is given as follows

$$v(S) = \sum_{i \in S} P_i(S) \tag{1}$$

Shapely in presented the value as an operator that assigns an expected marginal contribution to each player in the game with respect to a uniform distribution over the set of all permutations on the set of players. Specifically, let $\prod$ be a permutation (or an order) on the set of players, i.e., a mapping exists as one-to-one function from N onto N, and let us imagine the players appearing one by one to collect their payoff according to the order $\prod$. The marginal contribution Δ_i*of* player i to a coalition S is given as follows:

$$\Delta_i(S) = v(s \cup \{i\}) - v(s) \tag{2}$$

Here function v associates with every non-empty subset S of F, a real number $v(s)$(the value of S) with$v(\{\phi\}) = 0$. The unbiased estimator for the Shapley value, for a player $i$given by the mean of marginal contributions to all possible coalitions of players in N, is given as

$$\Phi_i(v) = \frac{1}{n!} \sum_{x \in \prod} \Delta_i(S_i(\pi)) \tag{3}$$

where $\prod$ is the set of permutations over N_p and $S_i(\pi)$ is the set of players from π that appears before player i in the permutation.

The feature selection process can be analogously seen as a coalition game where many features cooperate among themselves to achieve optimal performance in a particular task like classification, in our case. Here, the set $N_p$represents all the features, n represents the individual features and $v(S)$ stands for the accuracy metric obtained by the classifier using a subset of features S. Evaluation of features using the Shapley value involves testing on all possible combinations of subsets of features.

It has become quite common in many real world applications that the number of features n could easily be in the range of thousands e.g., microarray datasets and other high dimensional datasets, and at least in hundreds eg., medical datasets like Cardiac Arrhythmia etc., calculating the interactions among all those features by computing using all $n!$ subset is computationally infeasible. Therefore, an approximation of the Shapley value may be used. Keinan et al. [47] proposed a simple yet effective approximation method that uses uniformly sampled feature subsets rather than the entire set of subsets. Even though the number of sampled subsets is said to be much less than the $n!$ value, it is quite large to produce a strong estimation.

In our proposed feature selection algorithm, we use the Shapley values to estimate the contribution value of each feature towards the task of classification. In most realistic cases we observe that the size 'd' of the significant interactions among features is relatively smaller than the total number of features, 'n', we are limiting ourselves to calculate the contribution value from the possible permutations sampling over the entire feature set i.e., with a bounded permutation size.

This technique of feasible approximation is called as d-bounded permutations, where d is the size of a permutation. Formally, approximation of the Shapley value through d-bounded permutations is defined as

$$\Phi_i(v) = \frac{1}{|\prod_d|} \sum_{x \in \prod_d} \Delta_i(S_i(\pi)) \tag{4}$$

where $\prod_d$ represents the set of d-bounded permutations. The value of d stands for the number of interactions considered during the feature evaluation process. When we use $d = 1$, each feature is evaluated individually and in turn the interactions among the features are ignored naturally. In this work, we use a'd' value of $\sqrt{n}$, which has been adapted from an equivalent work on feature selection by random forests [52].

The feature selection based on Shapley value results in pareto optimality. The resulted shapely score is not affected by any arbitrary reordering or renaming of the features, which is required. Also a feature which is dummy i.e., – irrelevant feature, and does not affect the classifier's performance in any way scores a contribution value of null. It is apparent that irrelevant features don't help the classification process positively and hence to be assigned null value. In special cases of combining two different payoff values calculated on the same set of features, the Shapley value for a particular feature, that depicts its contribution to the collective performance measure, is merely the summation of all the respective Shapley values.This property leads to the linearity of the Shapley value. i.e., if the payoff value function v is multiplied by any real number α, then all corresponding Shapley values are scaled by the same factor α like $\Phi_i(\alpha v) = \alpha\Phi_i(v)$. When put in other way, multiplying the performance metric value by any constant does not affect or change the rank of the features – a crucial property required for any scheme that ranks the features based on their 'importance'. Considering all these supporting analogies prevailing between SVA and the feature selection procedure, this paper attempts to employ SVEGAfor feature selection to attain high accuracy.

3.2. Proposed feature selector-Shapley value embedded genetic algorithm (SVEGA)

In this section, the proposed memetic algorithm, particularly, Shapley Value Embedded GA (SVEGA) is outlined. At the beginning of the SVEGA search, the population for GA [17] is initialized randomly where each chromosome in the pool encodes a candidate feature subset. In this work, each chromosome is built of a binary string whose length equals the total number of features in

the dataset of interest. In binary encoding, a bit of value '1' ('0') indicates that the respective feature is selected (omitted). The objective function for calculating the fitness of each chromosome is then obtained as follows:

$$\text{Fitness}(c) = \text{Obj\_Fun}(SF_c) \tag{5}$$

Where SF_c denotes the Selected Feature subset encoded by a given chromosome c, and the objective function for feature selection Obj_Fun (SF_c) calculates the contribution of the given feature subset SF_c.

$$Obj\_Fun(SF_c) = \alpha * (1/\tau) + \beta * Sensitivity(SF_c) + \gamma * Specificity(SF_c)$$
$$\text{where} \tau = \text{No. of once in the } F_c \tag{6}$$

and, $\alpha = 0.4, \beta = 0.3$ and $\gamma = 0.3$

We use the *Sensitivity* and *Specificity* as the metrics in our Obj_Fun (SF_c). Both of them need maximization. i.e. maximum sensitivity and maximum specificity. Further the number of features has to be as low as possible. In case of two chromosomes having same fitness value, the chromosome with smaller number of selected features is given higher priority of surviving and is moved on to the next generation. This is recommended in a feature classification problem, where a subset of features with fewer features giving higher classification accuracy is preferred over a subset of features with more features giving lower or equal classification accuracy.

We define two memetic operators in the SVEGA, namely an *Include* operator which includes/adds a feature to the elite chromosome, and a *Remove* operator which removes/omits the existing features from the elite chromosome. The key issue is deciding which features to include and which ones to omit. Preferably, the features to be removed will be the ones which provide the least contribution when considered as a whole set and the features to be included are the ones which provide highest contribution to the solution feature subset. This characteristic has to be brought in the existing GA paradigm. This requirement is fulfilled by the use of Shapley value concept. For a given chromosome encoding c of a selected subset, let Q and R be the sets of selected and omitted features encoded in c, respectively. The function of the *Include* operator is to identify and select the feature with maximum Shapley score when measured in coalition, from set R and pushes it to the set Q. On the other hand, the *Remove* operator serves to identify and select the features with minimum contribution score and deletes from set Q and moves that into the set R. Then this feature ranking information is stored for use inside *Include* and *Remove* operators, for fine-tuning the entire search of solution space by the GA process.

The computational complexity of these two memetic operators can be quantified according to the search range$L$, which specifies the upper bound for both *Include* and *Remove*. Therefore, with $'L'$ possible *Include* operations and $'L'$ possible *Remove* operations, we get a total of $'L^{2'}$ possible combinations of *Include* and *Remove* operations executed on a chromosome. The $'L^{2'}$ combinations of *Include* and *Remove* are executed on the candidate chromosome in a random sequence and once an improvement is seen either in the fitness value or reduction is seen in the number of selected features without decline in the fitness value, the procedure is stopped. The best results on all system constraints were obtained with the memetic operation range $'L'$ empirically set to 4.

The following parameter setting is custom-made in our SVEGA:

- Population size *PS*: 50
- Number of generations *gencount*: 100
- Probability of crossover Pc: 0.6
- Probability of mutation Pm: 0.005

The functional flow and the pseudo code of the proposed SVEGA feature selection method are shown in Figs. 1 and 2 respectively.

3.3. Rationale for the choice of SVM classifier

Vapnik [49] introduced Support Vector Machine (SVM) as one of the learning methods for evaluating the model and accuracy prediction. Support Vector machine (SVM) [54] is one of the best and popular machine learning techniques for accuracy prediction. Recently, SVM attracts by its modeling features like promising performance and robustness in solving many classification problems. SVM model mostly focuses on prediction with good accuracy in many applications. Moreover, by selecting a suitable kernel function SVM can find the solution for complex nonlinear relationships. SVM can handle binary class problems, multiclass problems [51,53] and nonlinear regression – estimation problem and so on by proper choice of kernels [32,34,35]. Generally the kernel parameters are called hyper parameters and they differ for each kernel. The choice of selecting the kernels highly depends on the nature of the data. The kernel function transforms the input space into a high dimensional feature space. This choice of selecting the mapping functions (kernels) must be determined experimentally by applying and validating various kernels functions and their performances. Hence by kernel selection and adjusting the kernel parameters the best prediction could be done for many real world applications.

We have experimented with the following four kernels for classification, since they are more suitable to our medical dataset nature:

The **linear kernels** of the form:

$$K(x_i, x_j) = (x_i, x_j + 1) \tag{7}$$

The **polynomial kernels** of the form:

$$K(x_i, x_j) = (x_i, x_j + 1)^p \tag{8}$$

where 'p' is the degree of the polynomial,x_i is the ith instance in the dataset.

The **RBF kernels** of the form:

$$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|^2}{\gamma}\right)\right] \tag{9}$$

where '$\gamma > 0$' is the width of the radial basis function.

The **PUK kernels** of the form:

$$K(x_i, x_j) = \frac{1}{\left[1 + \left(\frac{2\sqrt{\|x_i - x_j\|^2}\sqrt{2^{\left(\frac{1}{\omega}\right)} - 1}}{\sigma}\right)^2\right]^{\omega}} \tag{10}$$

Pearson VII Universal kernel is one type of kernel functions in SVM and also referred as PUK [50]. It is a generic kernel and can be used in the place of Gaussian mapping function. It has an excellent flexibility to change its parameters.

Table 1 presents the overall summarization of these kernels used in the proposed system along with the required parameters to justify the performance of the choice of the kernel; $C = 1.0$ is taken for all the kernels, since it yielded the best results.

The predicted accuracy of SVM depends on a good setting of hyper parameters and the other respective kernel parameters. In this paper we show examples of SVM with Sequential Minimal Optimization (SMO) [44,45] using four different kernels. Parameter C determines the substitution between the model complexity and the degree to which divergence larger than 1 are tolerated in minimization optimization. This choice of C involves a bias-variance trade off:

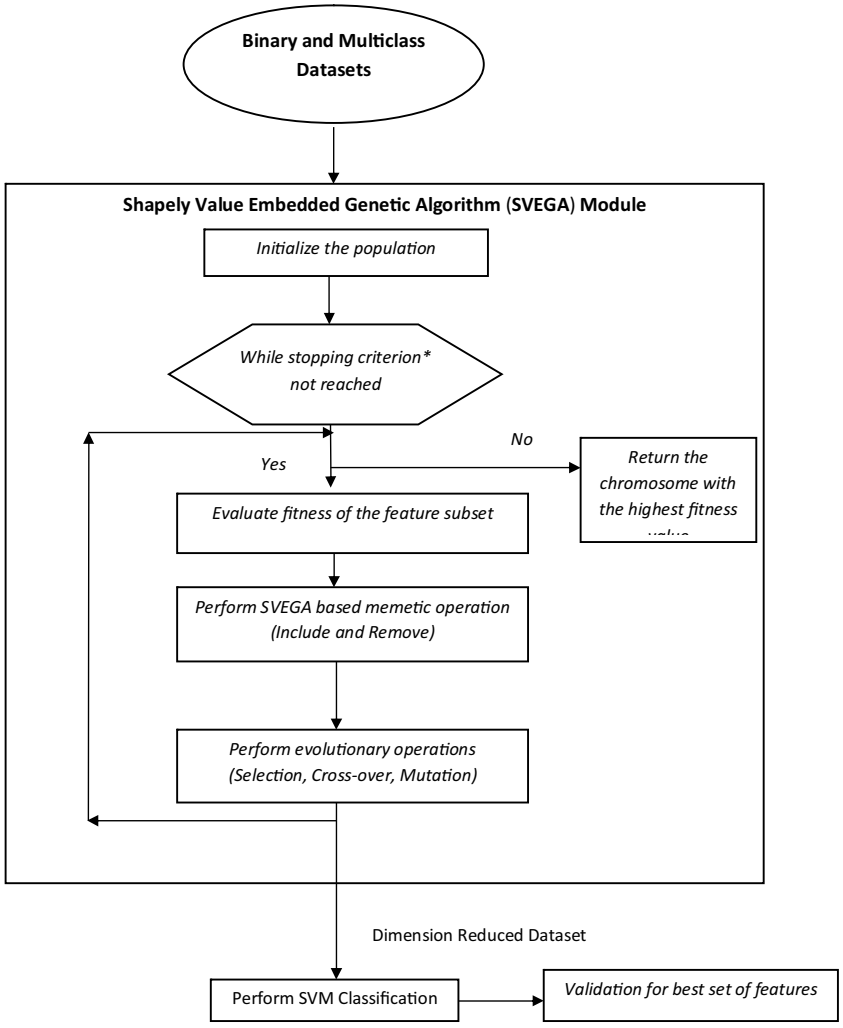

Fig. 1. General Diagram for SVEGA-SVM.

Table 1
Summary of various kernel types.

Kernel Type	Mapping function	Parameter Adjusted
Linear Kernel	$K(x_i, x_j) = (x_i, x_j + 1)^d$	'd' ['d' = 1 always for linear]
Normalized Polynomial	$K(x_i, x_j) = \frac{(x_i, x_j+1)^d}{sqrt(x_{i+1}+x_j)}$	'$d = 2.0$'
Radial Basis Function	$K(x_i, x_j) = \exp\left[-\frac{1}{2}\left(\frac{\|x_i - x_j\|^2}{\gamma}\right)\right]$	'$\gamma > 0$(here$\gamma = 0.01$)'
Pearson VII Universal kernel	$K(x_i, x_j) = \frac{1}{\left[1+\left(\frac{2\sqrt{\|x_i - x_j\|^2}\sqrt{2^{\left(\frac{1}{\omega}\right)}-1}}{\sigma}\right)^2\right]^{\omega}}$	'$\sigma = 1.0$' & '$\omega = 1.0$'

1. large C indicates low bias, high variance
2. small C indicates high bias, low variance

In a support vector machine, regularization results in a soft margin that allows a few points to traverse the optimal decision boundary. As C gets larger, the more stable the margin becomes, since it is allowing more points to determine the margin. Thus we can tune C to determine the optimal hyper-plane. The kernel determines how the model simplifies, or extrapolates to new data.

1-vs-1 approach is applied to construct all the possible two-class classifiers while handling multi-class problems. Each classifier is trained on samples only from two classes. When testing, we count the vote of each class and adopt the MaxWins algorithm to determine the final output. In our work, we combine the 1-vs-1 approach of the SVM with the SVEGA measured features to predict the accuracy for the selected class – dependent features due to its memory and time limitations.To evaluate the prediction performance of SVM kernels based on the proposed SVEGA feature selector method, 10-fold cross validation technique is used. In 10-fold cross validation strategy the complete dataset is randomly divided into 10 folds with approximately equal size. The validation is done on the both

Algorithm: Shapley Value Embedded Genetic Algorithm (SVEGA)

Input: *Encoded n-bit binary string (where n is the number of features), number of generations gencount, population size PS, crossover probability (Pc), mutation probability (Pm).*

Output: *A set of selected features that has lower cardinality and yields higher sensitivity and specificity values which in turn yields higher accuracy.*

BEGIN

(1) Population Initialization:

Initialize $\alpha = 0.4$, $\beta = 0.3$ *and* $\gamma = 0.3$, *M (total number of records in the training set), Pc and Pm.Randomly generate an initial population, which denoted* SF_c *of size PS encoded with n-bit binary string. Each gene value can be '0' or '1'. (A gene value of '1' means, the feature at that position is selected and a value of '0' means, the feature at that position is omitted).*

(2) While(not Current_fitness= Previous_fitness< 0.0001 orgencount is not reached)

a. Apply restrictive cross over and mutation operator to the chromosome at the specified probability Pc and Pm.

b. Evaluate the fitness value of all chromosomes in the population according to

$$Obj\_Fun(SF_c) = \alpha * (1/\tau) + \beta * Sensitivity(SF_c) + \gamma * Specificity(SF_c)$$

where $\tau = No.\ of\ ones\ in\ the\ SF_c$

(3) Selection:

The elite chromosome c_e *is selected and subjected to Shapley Value based memetic operations.*

BEGIN

Select the elite chromosome c_e *to undergo memetic operations.*

A. ***Include Operation*** *:*

BEGIN

a. Rank the features in R *in decreasing order of their Shapley values.*

b. Select a feature R_i *in* R *by linear ranking selection in such a way that a feature with larger Shapley value of a feature in* R *is more likely to be selected.*

c. Add R_i *to* Q.

END

B. Remove Operation :

BEGIN

a. Rank the features in Q *in decreasing order of Shapley value.*

b. Select a feature Q_i *in* Q *by linear ranking selection in such a way that a feature with larger Shapley value of a feature in* Q *is more likely to be selected.*

c. Eliminate all the features in $Q - \{Q_i\}$.

END

i. For $j = 1$ *to* L^2

ii. Generate a unique random pair of values $\{i, r\}$ *where* $0 \le i, r \le L$.

iii. Apply 'i*' times Include on the elite chromosome* c_e *and generate a new chromosome* c_e'.

ivApply 'r*' times Remove on* c_e' *and generate a new chromosome* c_e''.

v. Calculate the fitness of new modified chromosome c_e'' *based on* $Obj\_Fun(SF_c)$.

vi. If c_e'' *is better than* c_e *either on fitness value or the number of features*

vii. Replace the genotype c_e *with* c_e'' *and stop applying the memetic operation.*

viii.End If

ix. End For

END

(4) Lamarckian learning

The elite chromosome c_e *is replaced with improved new chromosome* c_e'' *by Lamarckian Learning process.*

(5) Evolutionary Operations:

The evolutionary operations like linear ranking selection, restrictive crossover and mutation operator with elitism.

(6) End While

END

Fig. 2. Pseudo code for the proposed SVEGA feature selection algorithm.

Table 2
Description of Binary and Multi Class Datasets.

Binary class datasets	Dataset description (Instances, Features, Class)	Source taken
Haberman's Survival	306,4,2,	UCI ML Repository
Liver Disorder	345,7,2	
Biomed	209,9,2	
Pima Diabetes	768,9,2	
Breast Cancer	286,10,2	
Statlog Heart	270,14,2	
Hepatitis	155,20,2	
Sick	3772,30,2	
Back ache	180,33,2	
CNS	60,7130,2	Kent ridge Repository
Leukemia	72,7130,2	
Multi class datasets	Dataset description (Instances, Features, Class)	Source taken
E-coli	336,8,8	UCI ML Repository
Post Operative patient	345,9,3	
Lymph Nodes	209,19,4	
Hypo-Thyroid	768,30,4	
Dermatology	286,35,6	
Lung Cancer	270,57,3	
Cardiac Arrhythmia	155,280,16	
Leukemia-3C	3772,7130,3	Kent ridge Repository
Leukemia-4C	180,7130,4	
SRBCT	60,2309,4	
MLL	72, 12583,3	

training and testing dataset. In each iteration, one fold is considered as test set and the remaining sets are used to train the SVM model.

4. Experimental reusults and analysis

Table 2 represents the list of binary class and multiclass datasets taken from UCI [4] and Kent ridge repository [5] for respective classification process. Java based WEKA 3.7.1 software [55] on Window 7 operating system was used for testing the proposed feature selector SVEGA on binary class datasets and multiclass datasets with different dimensions (small, medium and high) by the concerned kernels based SVM classifier. Support Vector Machine (SVM) with the use of four different Kernel function, i.e. "Normalize Polynomial Kernel (NP)", "Polynomial Kernel (PK)", "Radial Basis Function kernel (RBF)", and "Pearson VII function-based universal kernel (PUK)" on features selected by proposed SVEGA.

In our experiments, the following parameter values are used:

1. The complexity parameter 'C' = 1.0 is applied for all the four kernels.
2. For normalized polynomial kernel, 'd' is assigned the value of 2.
3. For Pearson VII Universal kernel, '$\sigma = 1.0$' & '$\omega = 1.0$' are assigned
4. For Radial Basis Function kernel, $\gamma = 0.01$

The results are compared for performance evaluation. Figs. 3 and 4 shows the performance of SVEGA on both binary and multi class problems.

In Table 3, 10–fold cross validation results for the binary class SVM obtained using four different are presented. Among four kernels such as K1-Normalized Poly kernel (exp = 1.0), K2-Linear, K3-PUK kernel, K4-RBF Kernel, for the binary class data classification problems 'binary SVM with linear kernel' classifier yields promising results. Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Relative Absolute Error (RAE) were also recorded for the individually best selected kernels. Table 4 shows the result with same track as in Table 4 for the multi class data classification problems where 'SVM with linear kernel' classifier together with 1-vs-1 pairing gives the best accuracy. Table 5 and Table 6records the optimal features selected by proposed SVEGA approach i.e. after applying the proposed feature selection technique. On evaluating the features selected after applying SVEGA approach, Normalized Poly kernel is found to be the best choice for the binary class data classification problems and linear kernel of SVM is observed

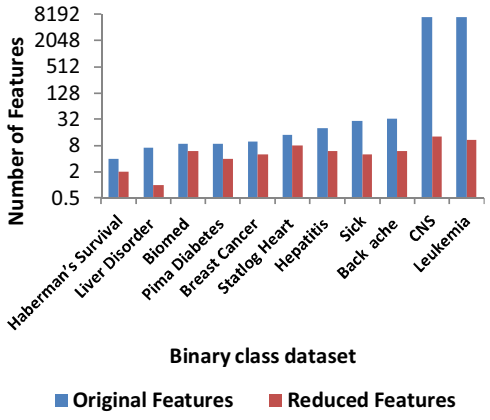

Fig. 3. SVEGA Performance on Binary Class Data.

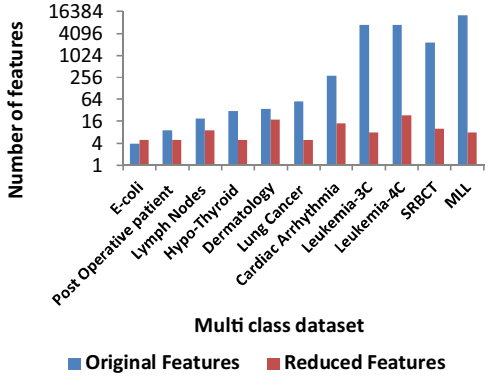

Fig. 4. SVEGA Performance on Multi Class Data.

Table 3
Performance evaluation with various SVM kernels for the binary class problems on theoriginal dataset.

BINARY CLASS PROBLEMS								Error Rates for the Best individual SVM Kernel		
S.No.	Datasets	Observations	Actual Features	SVM Performance for proposed method with choice of Kernels Parameters (C = 1.0 and γ = 0.001)				Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Relative Absolute Error (RAE) (%)
				K1-Normalized Poly kernel (exp = 1.0) $A_{ccuracy}$ (%)	K2-Linear Kernel $A_{ccuracy}$ (%)	K3-PUK Kernel $A_{ccuracy}$ (%)	K4-RBF Kernel $A_{ccuracy}$ (%)			
1.	Haberman's Survival	306	4	73.20	73.52	73.85*	73.52	0.26	0.5113	67.01
2.	Liver Disorder	345	7	61.15	58.26	70.72*	57.97	0.29	0.5411	60.06
3.	Biomed	209	9	85.64	87.55	88.99*	64.11	0.11	0.33	23.88
4.	Pima Diabetes	768	9	67.18	77.34*	76.56	65.10	0.22	0.46	49.84
5.	Breast Cancer	286	10	74.82	69.58*	69.58	70.27	0.30	0.55	72.70
6.	Statlog Heart	270	14	84.07*	84.07*	81.48	82.59	0.15	0.39	32.24
7.	Hepatitis	155	20	83.87	85.16*	78.06	79.35	0.14	0.38	44.93
8.	Sick	3772	30	93.87	93.84	94.77*	93.87	0.06	0.24	53.38
9.	Back ache	180	33	86.11	86.11*	86.11	86.11	0.13	0.37	57.26
10.	CNS	60	7130	65.00	68.33*	65.00	61.66	0.31	0.56	69.27
11.	Leukemia	72	7130	77.77	95.83*	65.27	65.27	0.04	0.20	9.14
Best kernels for Binary class problems on average basis				77.51	79.96	77.30	72.71	0.18	0.41	49.07

* indicates the maximal accuracy among the SVM Kernels.

Table 4
Performance evaluation with various SVM kernels for the multi class problems on the original dataset.

MULTI CLASSIFICATION PROBLEMS									Error Rates for the Best individual SVM Kernel		
S.No.	Datasets	Classes	Observations	Actual Features	SVM Performance for proposed method with choice of Kernels Parameters (C = 1.0 and γ = 0.001)				Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Relative Absolute Error (RAE) (%)
					K1-Normalized Poly kernel (exp = 1.0) $A_{ccuracy}$ (%)	K2-Linear Kernel $A_{ccuracy}$ (%)	K3-PUK Kernel $A_{ccuracy}$ (%)	K4-RBF Kernel $A_{ccuracy}$ (%)			
1.	E-coli	8	306	4	73.32	84.22	87.5*	42.55	0.189	0.295	103.744
2.	Post Operative patient	3	345	9	71.11*	67.77	66.66	71.11*	0.30	0.39	104.52
3.	Lymph Nodes	4	209	19	85.13	86.48*	79.05	80.40	0.26	0.33	97.62
4.	Hypo-Thyroid	4	768	30	93.37	93.61*	93.21	92.28	0.25	0.32	351.26
5.	Dermatology	6	286	35	96.72	95.35	72.67	96.99*	0.22	0.31	83.64
6.	Lung Cancer	3	270	57	71.87*	65.62	71.87*	71.87*	0.28	0.530	68.18
7.	Cardiac Arrhythmia	16	155	280	58.40	70.13*	54.20	54.20	0.10	0.23	128.17
8.	Leukemia-3C	3	3772	7130	70.83	97.22*	52.77	52.77	0.23	0.29	59.65
9.	Leukemia-4C	4	180	7130	62.5	93.05*	52.77	52.77	0.25	0.32	82.76
10.	SRBCT	4	60	2309	97.59	98.79*	34.93	93.97	0.25	0.31	69.25
11.	MLL	3	72	12583	76.38	97.22*	38.88	43.05	0.23	0.29	59.65
Best kernels for Multi class problems on average basis					77.92	86.31	64.04	68.36	0.23	0.33	109.86

* indicates the maximal accuracy among the SVM Kernels.

Table 5
Performance evaluation with various SVM $_{kernels}$ for the binary class problems on the SVEGA reduced dataset.

BINARY CLASS PROBLEMS									Error Rates for the Best individual SVM Kernel		
S.No.	Datasets	Observations	Actual Features	Selected Features by Proposed Method	SVM Performance for proposed method with choice of Kernels Parameters (C = 1.0 and γ = 0.001)				Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Relative Absolute Error (RAE) (%)
					K1-Normalized Poly kernel (exp = 1.0) $A_{ccuracy}$ (%)	K2-Linear Kernel $A_{ccuracy}$ (%)	K3-PUK Kernel $A_{ccuracy}$ (%)	K4-RBF Kernel $A_{ccuracy}$ (%)			
1.	Haberman's Survival	306	4	2	85.36*	71.56	85.36*	85.36*	0.26	0.33	99.30
2.	Liver Disorder	345	7	1	75.36*	57.97	57.97	57.97	0.21	0.50	60.16
3.	Biomed	209	9	6	79.42	90.03*	90.03*	64.11	0.23	0.30	301.69
4.	Pima Diabetes	768	9	4	65.10	76.82	86.69*	65.10	0.18	0.37	57.26
5.	Breast Cancer	286	10	5	85.36*	66.43	71.32	70.27	0.13	0.37	31.15
6.	Statlog Heart	270	14	8	84.44*	83.70	83.70	84.88*	0.15	0.39	31.24
7.	Hepatitis	155	20	6	90.51*	83.22	82.58	79.35	0.258	0.32	354.81
8.	Sick	3772	30	5	95.73*	93.84	95.46	93.87	0.22	0.31	83.77
9.	Back ache	180	33	6	88.11*	84.44	88.11*	88.11*	0.11	0.33	23.88
10.	CNS	60	7130	13	93.35*	93.35*	88.33	65	0.256	0.32	351.69
11.	Leukemia	72	7130	11	97.22*	97.22*	97.22*	65.27	0.25	0.31	69.53
Best kernels for Binary class problems on average basis					85.45	81.68	84.25	74.48	0.20	0.35	133.23

* indicates the maximal accuracy among the SVM Kernels.

Table 6
Performance evaluation with various SVM kernels for the Multi class problems on the SVEGA reduced dataset.

MULTI CLASSIFICATION PROBLEMS										Error Rates for the Best individual SVM Kernel		
S.No.	Datasets	Classes	Observations	Actual Features	Selected Features by Proposed Method	SVM Performance for proposed method with choice of Kernels Parameters (C = 1.0, γ = 0.001 and 1-VS-1 coupling)				Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	Relative Absolute Error (RAE) (%)
						K1-Normalized Poly kernel (exp = 1.0) $A_{ccuracy}$ (%)	K2-Linear Kernel $A_{ccuracy}$ (%)	K3-PUK Kernel $A_{ccuracy}$ (%)	K4-RBF Kernel $A_{ccuracy}$ (%)			
1.	E-coli	8	336	8	5	71.13	83.03	93.54	42.55	0.25	0.32	70.92
2.	Post Operative patient	3	90	9	5	71.11	67.77	76.35	71.11	0.27	0.36	62.99
3.	Lymph Nodes	4	148	19	9	82.43	87.32	82.43	79.05	0.18	0.29	103.74
4.	Hypo-Thyroid	4	3772	30	5	98.36	93.13	93.79	92.28	0.25	0.31	69.25
5.	Dermatology	6	366	35	18	98.54	97.26	88.25	98.54	0.23	0.29	59.65
6.	Lung Cancer	3	32	57	5	78.51	75	71.87	78.51	0.21	0.46	66.43
7.	Cardiac Arrhythmia	16	452	280	14	65.70	68.14	84.15	54.20	0.18	0.29	103.72
8.	Leukemia-3C	3	72	7130	8	98.61	100	97.22	52.77	0.02	0.08	7.41
9.	Leukemia-4C	4	72	7130	23	79.16	98.86	78.38	52.77	0.25	0.31	69.25
10.	SRBCT	4	83	2309	10	77.10	78.31	81.92	34.93	0.18	0.43	37.49
11.	MLL	3	72	12583	8	95.83	98.61	97.22	38.88	0.26	0.33	99.30
Best kernels on average basis						83.15	85.59	83.68	63.41	0.21	0.31	68.20

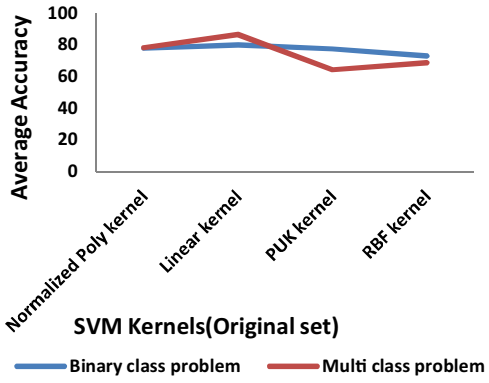

Fig. 5. SVM Kernels Performance on Binary Class Data.

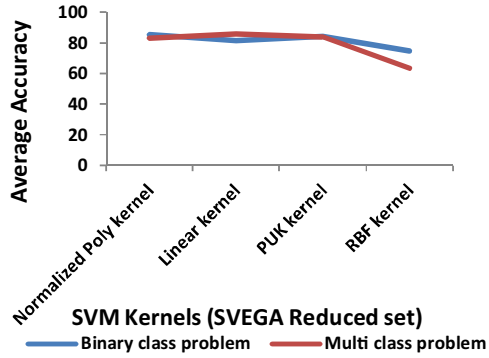

Fig. 6. SVM Kernels Performance on Multi Class Data.

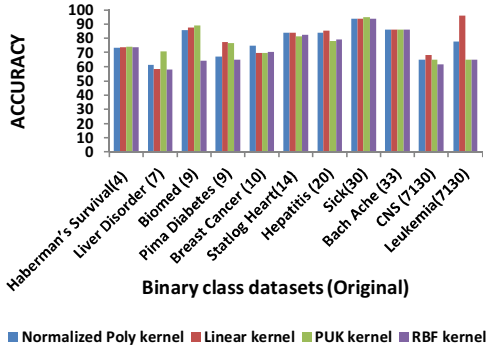

Fig. 7. SVEGA-SVM (Accuracy) Performance on Original Binary Class Data.

to be the best for the multi class data classification problems (Figs. 5 and 6, Table 6).

Linear kernel (79.96% of accuracy) provides the first best result for the binary classification on the original data followed by normalized poly kernel with 77.51% and PUK kernel with 77.30%. In case of multi classification on the original data, linear kernel (86.31% of accuracy) provides the first best result followed by the other kernels like normalized poly kernel with 77.92% and RBF gives third best result with 68.36% accuracy and PUK gives poor performance among all with only 64.04% accuracy. Contrarily, in the case of binary classification on the optimal features selected by SVEGA, normalized poly kernel (85.45% of accuracy) gives the first best performance followed by PUK and linear kernel with the accuracy of 84.25% and 81.68% of accuracy. But in multi classification on SVEGA results, the linear kernel gives the first best performance with accuracy of 85.59% followed by PUK kernel with accuracy of 83.68% and normalized kernel with accuracy of 83.15%. Figs. 7–14 show the

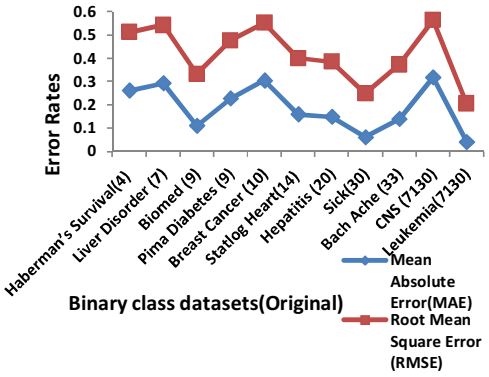

Fig. 8. SVEGA-SVM (Error Rates) Performance on Original Binary Class Data.

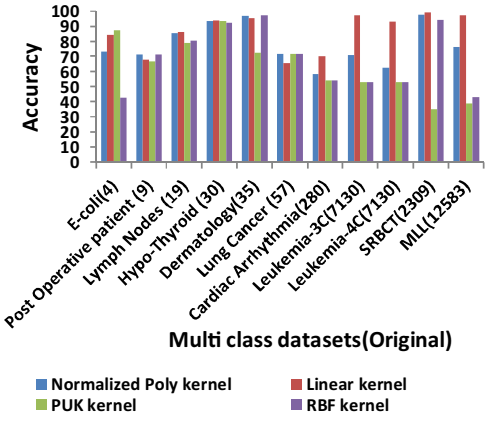

Fig. 9. SVEGA-SVM (Accuracy) Performance on Original Multi Class Data.

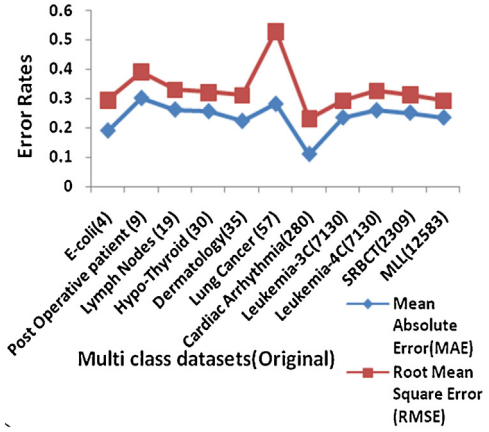

Fig. 10. SVEGA-SVM (Error Rates) Performance on Original Multi Class Data.

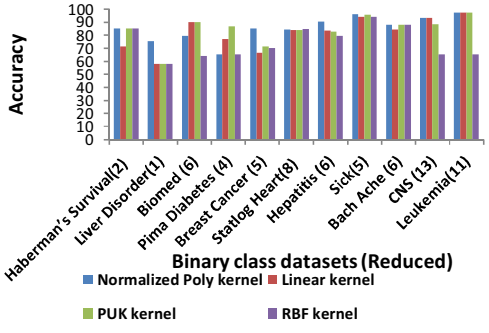

Fig. 11. SVEGA-SVM (Accuracy) Performance on Reduced Binary Class Data.

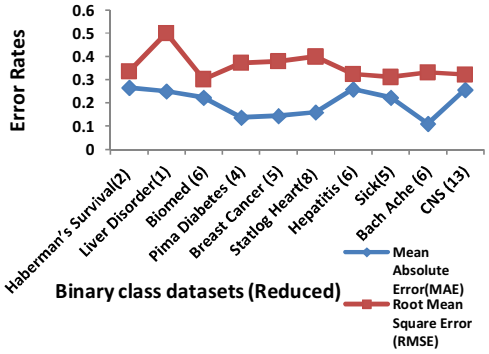

Fig. 12. SVEGA-SVM (Error Rates) Performance on Reduced Binary Class Data.

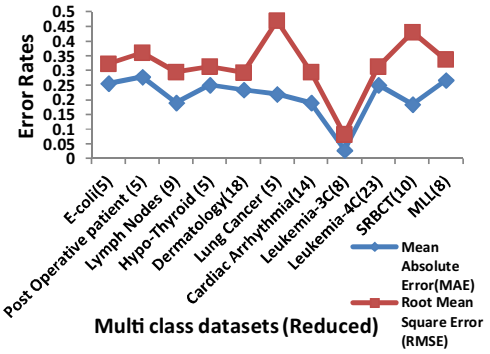

Fig. 14. SVEGA-SVM (Error Rates) Performance on Reduced Multi Class Data.

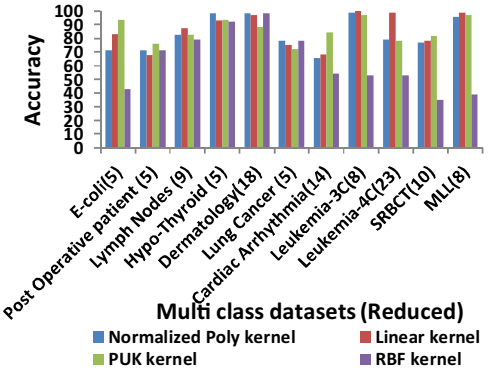

Fig. 13. SVEGA-SVM (Accuracy) Performance on Reduced Multi Class Data.

evaluation performance (accuracy and error rates) of the proposed approach on original and reduced binary and multi class datasets

Here we evaluated the effectiveness of the proposed feature selector SVEGA by SVM classifier and we made extensive comparison by extending the comparative study with other classifiers like Naïve Bayes (NB),K-Nearest Neighbor (KNN) and SVM-RFE. The performance comparison of the proposed SVEGA-SVM model against other classifiers is shown in Tables 7 and 8. The comparison is extended in three tracks: applying more different datasets (22 data sets), analysing the features selected by SVEGA method and evaluating the different classifiers performance in terms of accuracy and error rate. Our comparison study shows that combination of SVEGA approach with SVM provides the best result and could be used for performing classification on data analysis.

The testing performance of the SVEGA-SVM based diagnostic system is found to be reasonable and this system can be used in clinical studies. This application brings objectivity to the evaluation of any data classification problems and its automated nature makes it easy to be used in clinical practice. Besides the feasibility of a real-time implementation of the expert diagnosis system, diagnosis may be made more accurate with the limited features.

Table 7
Comparision of SVEGA-SVM for Binary class Datasets with existing methods.

Binary class Datasets	SVM (Actual data)		SV-RELIEFF-KNN			SVM-RFE			MA-C-NB			SVEGA-SVM Kernels		
	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}
Haberman's Survival	73.52	0.05	4	73.52	0.07	2	71.56	0.06	2	71.56	0.06	*2*	*85.36*	*0.05*
Liver Disorder	58.26	0.4	5	57.97	0.5	1	57.97	0.8	1	58.26	0.7	*1*	*75.36*	*0.6*
Biomed	87.55	0.3	6	84.68	0.2	6	88.03	0.2	6	88.56	0.1	*6*	*90.03*	*0.09*
Pima Diabetes	77.34	0.8	6	76.69	0.7	4	76.82	0.6	4	77.34	0.8	*4*	*86.69*	*0.7*
Breast Cancer	69.58	0.31	9	69.58	0.17	5	66.43	0.09	5	66.43	0.09	*5*	*85.36*	*0.08*
Statlog Heart	84.07	0.35	12	81.85	0.35	8	82.96	0.18	8	82.59	0.15	*8*	*84.88*	*0.17*
Hepatitis	85.16	0.31	14	84.51	0.24	10	83.22	0.12	10	87.10	0.10	*6*	*90.51*	*0.11*
Sick	93.84	1.8	18	93.87	0.8	5	93.87	0.8	5	93.86	0.7	*5*	*95.73*	*0.5*
Back ache	86.11	0.36	16	85	0.9	7	84.44	0.02	7	85.12	0.01	*6*	*88.11*	*0.01*
CNS	68.33	88.15	918	74.36	78.32	475	78.45	66.8	111	83.16	70.6	*13*	*93.35*	*67.1*
Leukemia	95.83	258.3	1126	95.82	195.36	678	96.32	118.0	58	96.45	98.56	*11*	*97.22*	*112.5*

Table 8
Comparision of SVEGA-SVM for Multi class Datasets with existing methods.

Multi class Datasets	SVM (Actual data)		SV-RELIEFF-KNN			SVM-RFE			MA-C-NB			SVEGA-SVM Kernels		
	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}	F_{select}	Accuracy	R_{Time}
E-coli	84.22	0.16	6	83.63	0.17	5	83.63	0.20	5	87.20	0.18	*5*	*93.54*	*0.15*
Post Operative patient	67.77	0.04	7	68.88	0.04	5	68.88	0.04	5	70	0.03	*5*	*76.35*	*0.02*
Lymph Nodes	86.48	0.09	16	85.81	0.09	9	83.78	0.09	9	86.49	0.08	*9*	*87.32*	*0.05*
Hypo-Thyroid	93.61	5.23	25	95.53	5.6	5	93.13	5.68	5	97.77	4.27	*5*	*98.36*	*5.1*
Dermatology	95.35	0.55	28	95.62	0.38	19	97.26	0.27	19	95.35	0.25	*18*	*98.54*	*0.15*
Lung Cancer	65.62	0.12	21	56.25	0.12	6	59.37	0.12	6	40.62	0.09	*5*	*78.51*	*0.05*
Cardiac Arrhythmia	70.13	4.24	103	70.13	3.78	26	68.14	3.09	20	70.57	2.34	*14*	*84.15*	*1.86*
Leukemia-3C	97.22	292.5	540	97.36	213.4	262	97	174.2	25	98.16	176.6	*8*	*100*	*170.1*
Leukemia-4C	93.05	298.6	4156	95.16	246.8	1146	95.21	218.6	58	96.76	248.4	*23*	*98.86*	*234.3*
SRBCT	97.56	530	1568	96.34	386.8	546	97.88	350.8	648	97.92	246.2	*10*	*98.79*	*232.6*
MLL	97.22	392.6	8756	97.22	350.8	6306	93.45	236.7	345	95.16	248.3	*8*	*98.61*	*240.4*

5. Discussion and conclusion

The experiments conducted in the study successfully demonstrate the beneficial factors of SVM-SVEGA over a range of other feature selection algorithms and prove it to be a promising technique. The results also show that the choices of the kernel function and feature selection technique have a profound effect on the performance of SVM for both binary and multi data classification. The effect of various SVM kernels impacted by the proposed SVEGA scheme is also investigated. Depending on the nature of the problem – binary class or multi class, appropriate kernels may be chosen. The study conducted can be comprehendingly substituted as a complete expert system model for deducing faster clinical diagnosis with better accuracy. The future work can test the proposed techniques on datasets from other domains. Further Ant Colony Optimization and Particle Swarm Analysis may be employed when low cost study is required.

Acknowledgment

This work is supported in part by the University Grant Commission (UGC), New Delhi, India – Major Research Project under grant no. F.No.: 39-899/2010 (SR).

References

[1] H. Liu, H. Motoda, Computational Methods of Feature Selection, Chapman and Hall/CRC Press, 2007.
[2] X. Wu, K. Yu, H. Wang, W. Ding, Online streaming feature selection, Proceedings of the 27th International Conference on Machine Learning (2010) 1159–1166.
[3] Z. Xu, R. Jin, J. Ye, M. Lyu, I. King, Discriminative semi-supervised feature Selection via manifold regularization in IJCAI'09, Proceedings of the 21th International Joint Conference on Artificial Intelligence (2009).
[4] S. Hettich, C. Blake, C. Merz, UCI Repository of Machine Learning Databases, 1998 http://www.ics.uci.edu/mlearn/MLRepository.html.
[5] L. Jinyan, L. Huiqing, Kent Ridge Bio-Medical Data Set Repository, 2002, pp. i2 http://datam.r.astar.edu.sg/datasets/krbd.
[6] H. Daassi-Gnaba, Y. Oussar, External vs. internal Svm-Rfe: the Svm-Rfe method revisited and applied to emotion recognition, Neural Netw. World (2015), http://dx.doi.org/10.14311/nnw.2015.25.004.
[7] Marina Sokolova, Guy Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (2009) 427–437, http://dx.doi.org/10.1016/j.ipm.2009.03.002.
[8] Nongnuch Poolsawad, Lisa Moore, Chandrasekhar Kambhampati, John G.F. Cleland, Issues in the mining of heart failure datasets, Int. J. Autom. Comput. 11 (2) (2014) 162–179, http://dx.doi.org/10.1007/s11633-014-0778-5.
[9] P. Shi, S. Ray, Q.F. Zhu, M.A. Kon, Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction, BMC Bioinf. 12 (2011) 375, http://dx.doi.org/10.1186/1471-2105-12-375.
[10] T. Mar, S. Zaunseder, J.P. Martinez, M. Llamedo, R. Poll, Optimization of ECG classification by means of feature selection, IEEE Trans. Biomed. Eng. 58 (8) (2011) 2168–2177, http://dx.doi.org/10.1109/tbme.2011.2113395.
[11] M. Sugiyama, M. Kawanabe, P.L. Chui, Dimensionality reduction for density ratio estimation in high-dimensional spaces, Neural Netw. 23 (1) (2010) 44–59.
[12] J. Xie, Using Support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases, Expert Syst. Appl. (2010), http://dx.doi.org/10.1016/j.eswa.2010.10.050.
[13] Bu Hualonga, X. Jing, Hybrid feature selection mechanism based high dimensional data sets reduction, Energy Procedia 11 (2011) 4973–4978.
[14] Y. Chen, S. Yu, Selection of effective features for ECG beat recognition based on nonlinear correlations, Artif. Intell. Med. 54 (1) (2012) 43–52, http://dx.doi.org/10.1016/j.artmed.2011.09.004.
[15] J.Q. Gan, B.A.S. Hasan, C.S.L. Tsui, A filter-dominating hybrid sequential Forwardfloating search method for feature subset selection in high-dimensional space, Int. J. Mach. Learn. Cybernet. 3 (4) (2012) 1–8, http://dx.doi.org/10.1007/s13042-012-0139-z, Springer-Verlag.
[16] L.T. Vinh, S. Lee, Y. Park, B.J. Auriol, A novel feature selection method based on normalized Mutual information, Int. J. Appl. Intell. 37 (1) (2011) 100–120, http://dx.doi.org/10.1007/s10489-011-y.
[17] Feng Tan, Xuezheng Fu, Yanqing Zhang, Anu G. Bourgeois, A genetic algorithm based method for feature subset selection, Soft Comput. 11 (1) (2008) 111–120, http://dx.doi.org/10.1007/s00500-007-0193-8.
[18] o.P Bermej, a.L.D.L Oss, J.A. Gamez, J.M. Puerta, Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking, Knowl. Based Syst. 25 (1) (2012) 35–44, http://dx.doi.org/10.1016/j.knosys.2011.01.015.
[19] V. Bolon-Canedo, N. Sanchez-Maroio, A. Alonso-Betanzos, Distributed feature selection: an application to microarray data classification, Appl. Soft Comput. 30 (1) (2015) 136–150, http://dx.doi.org/10.1016/j.asoc.2015.01.035.
[20] M.M. Jazzar, d.G Muhamma, Feature selection based verification/identification system using fingerprints and palm print, Arab. J. Sci. Eng. 38 (4) (2013) 849–857, http://dx.doi.org/10.1007/s13369-012-0524-7.
[21] A. Khan, A.R. Baig, Multi-objective feature subset selection using non-dominated sorting genetic algorithm, J. Appl. Res. Technol. 13 (1) (2015) 145–159, http://dx.doi.org/10.1016/s1665-6423(15)30013-4.
[22] Yu-Jun Zheng, Sheng-Yong Chen, Hai-Feng Ling, Evolutionary optimization for disaster relief operations:A survey, Appl. Soft Comput. 27 (1) (2015) 553–566, http://dx.doi.org/10.1016/j.asoc.2014.09.041.
[23] n.S SenthamaraiKanna, j.N Ramara, A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm, Knowl. Based Syst. 23 (1) (2010) 580–585, http://dx.doi.org/10.1016/j.knosys.2010.03.016.
[24] Artur J. Ferreira, Ma Rio, A.T. Figueiredo, An unsupervised approach to featured discretization and selection, Pattern Recogn. 45 (9) (2012) 3048–3060, http://dx.doi.org/10.1016/j.patcog.2011.12.008.
[25] P. Smialowski, D. Frishman, S. Kramer, Pitfalls of supervised feature selection, Bioinformatics 26 (3) (2010) 440–443, http://dx.doi.org/10.1093/bioinformatics/btp621.
[26] H. Uguz, A hybrid system based on information gain and principal component analysis for the classification of transcranial Doppler signals, Comput. Methods Programs Biomed. 107 (3) (2012) 598–609, http://dx.doi.org/10.1016/j.cmpb.2011.03.013.
[27] Nassim Laouti, Sami Othman, Mazen Alamir, Combination of model-based observer and support vector machines for fault detection of wind turbines, Int. J. Autom. Comput. 11 (3) (2014) 274–287, http://dx.doi.org/10.1007/s11633-014-0790-9.
[28] Nibaran Das, Ram Sarkar, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu, A genetic algorithm based region sampling for selection of local features in handwritten digit recognition application, Appl. Soft Comput. 12 (5) (2012) 1592–1606, http://dx.doi.org/10.1016/j.asoc.2011.11.030.
[29] Miguel Garcia-Torres, Ruben Armananzas, Concha Bielza, Pedro Larranaga, Comparison of meta heuristic strategies for peak bin selection in proteomic mass spectrometry data, Inf. Sci. 222 (1) (2013) 229–246, http://dx.doi.org/10.1016/j.ins.2010.12.013.
[30] Soroor Sarafrazi, Hossein Nezamabadi-pour, Facing the classification of binary problems with a GSA-SVM hybrid System, Math. Comput. Modell. 57 (1–2) (2013) 270–278, http://dx.doi.org/10.1016/j.mcm.2011.06.048.
[31] Yael Weiss, Yuval Elovici, Lior Rokach, The CASH algorithm-cost-sensitive attribute selection using histograms, Inf. Sci. 222 (1) (2013) 247–268, http://dx.doi.org/10.1016/j.ins.2011.01.035.
[32] J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. (2) (2016) 121–167, http://dx.doi.org/10.1023/A:1009715923555.
[33] T. Hastie, R. Tibshirani, Classification by pair wise coupling, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 1998, pp. 507–513.
[34] Scholkopf B. Smola, Learning with Kernels – Support Vector Machines, Regularization, Optimization, and Beyond, Cambridge, The MIT Press, Massachusetts, 2016.
[35] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, UK, 2004.
[36] Ji Zhu, Trevor Hastie, Kernel logistic regression and the import vector machine, J. Comput. Graph. Stat. 14 (1) (2005) 185–205, http://dx.doi.org/10.1198/106186005X25619.
[37] M. Zhu, W. Su, LAGO: a computationally efficient approach for statistical detection, Technometrics 48 (2006) 193–205, http://dx.doi.org/10.1198/004017005000000643.
[38] Y. Tang, H. Zhang, Multiclass proximal support vector machines, Journal of Computational l and Graphical Statistics 15 (2) (2006) 339–355, http://dx.doi.org/10.1198/106186006X113647.
[39] Y. Liu, X. Shen, Multicategory ψ – learning, J. Am. Stat. Assoc. 101 (2006) 500–509, http://dx.doi.org/10.1198/016214505000000781.
[40] A.I. Guyon, Elisseeff.A, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (1) (2003) 1157–1182.
[41] S. Alelyani, J. Tang, H. Liu, in: Charu Aggarwal, Chandan Reddy (Eds.), Feature Selection for Clustering: A Review. Data Clustering: Algorithms and Applications, CRC Press, 2013.
[42] Z. Zhao, H. Liu, Semi-supervised feature selection via spectral analysis, Proceedings of SIAM International Conference on Data Mining (2007).
[43] Anderson Rocha, Siome Klein Goldenste, Multiclass from binary: expanding one-versus-all, one-versus-one and ECOC-based approaches, IEEE Trans. Neural Netw. Learn. Syst. 25 (2) (2014) 289–302, http://dx.doi.org/10.1109/TNNLS.2013.2274735.
[44] J.C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, in: S.A. Solla, T.K. Leen, K.-R. Muller (Eds.), Advance in Neural Information Processing Systems, The MIT Press, 547-553, 2000 (12).
[45] S.K. Keerthi, C. Bhattacharyya, K.R.K. Murthy, Improvements to platt's SMO algorithm for SVM classifier design, Neural Comput. 13 (3) (2001) 637–649.
[46] S. Moretti, Danitsja van Leeuwen, Hans Gmuender, Stefano Bonassi, Joost Van Delft, Jos Kleinjans, Fioravante Patrone, Domenico Franco Merlo, Combining Shapley value and statistics to the analysis of gene expression data in children

Exposedto air pollution, BMC Bioinf. 9 (361) (2008) 1–21, http://dx.doi.org/10.1186/1471-2105-9-361.
[47] A. Keinan, B. Sandbank, C.C. Hilgetag, I. Meilijson, E. Ruppin, Fair attribution of functional contribution in artificial and biological networks, Neural Comput. 16 (9) (2004) 1887–1915, http://dx.doi.org/10.1162/0899766041336387.
[48] S.B. Cohen, Dror Gideon, Ruppin Eytan, Feature selection based on Shapley value, Proceedings of IJCAI (2005) 665–670.
[49] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.
[50] Khalid A.A. Abakar, Chongewen Yu, Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity, Indian J. Fibre Text. 39 (2014) 55–59.
[51] J. Weston, C. Watkins, Multi-class Support Vector Machines, Royal Holloway, University of London, U.K, 1998, Technical Report CSD-TR-98-04.
[52] C.W. Hsu, C.J. Lin, A comparison of methods for multi-class support vector machines, IEEE Trans. Neural Netw. 13 (2000) 415–425;
G. James, Majority vote classifiers: Theory and Applications. Ph.D. Thesis, Department of Statistics, Stanford University, Stanford, CA, 1998.
[53] J.C. Platt, N. Cristianini, J. Shawe-Taylor, Large margin DAGs for multiclass classification, in: S.A. Solla, T.K. Leen, K.-R. Muller (Eds.), Advance in Neural Information Processing Systems, 12, The MIT Press, 2000, pp. 547–553.
[54] Hui Li, Chang-Jiang Li, Xian-Jun Wu, Jie Sun, Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine, 19(1), 2014, 57–67, doi: 10.1016/j.asoc.2014.01.018.
[55] WEKA 3: Machine Learning Software in Java. The University of Waikato software documentation. http://www.cs.waikato.ac.nz/_ml/weka.