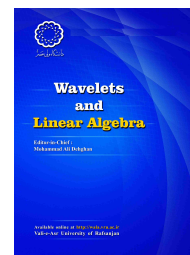


Vali-e-Asr University
of Rafsanjan

Wavelets and Linear Algebra

<http://wala.vru.ac.ir>



Application Of Wavelets To Improve Cancer Diagnosis Model In High Dimensional Linguistic DNA Microarray Datasets

Nasibeh Emami^{a,*}

^aDepartment of computer science, faculty of basic sciences, kosar university of bojnord, bojnord, Iran.

ARTICLE INFO

Article history:

Received 9 December 2020

Accepted 27 April 2021

Available online 20 July 2021

Communicated by Ali Tavakoli

Keywords:

fuzzy numbers,
linguistic data,
diagnosis cancer,
Wavelet decomposition,
high dimensional data.

2000 MSC:

65F05, 46L05, 11Y50.

ABSTRACT

DNA microarray datasets suffer scaling and uncertainty problems. This paper develops a model that manages DNA microarray datasets challenges more precisely by using the advantages of Wavelet decomposition and fuzzy numbers. For this aim, the proposed method is utilized to classify linguistic DNA microarray datasets set, where datasets can be given as linguistic genes. Linguistic genes are represented by using triangular fuzzy numbers provided as LR (left-right) fuzzy numbers. Then the WABL method is applied as the defuzzification method. Also, a set of orthogonal wavelet detail coefficients based on wavelet decomposition at different levels is extracted to specify the localized genes of DNA microarray datasets. Three DNA microarray datasets are used to evaluate this method. The experiments are shown that the proposed model has better diagnostic accuracy than other methods.

© (2021) Wavelets and Linear Algebra

*Corresponding author

Email address: nasibeh.emami@kub.ac.ir (Nasibeh Emami)

1. Introduction

Classification of the DNA microarray datasets plays an important role in early detection of cancer, treatment, prevention, and drug development[16]. A major challenge with DNA microarray datasets analysis is the huge amount of data that are often related to high-dimensional data. This is due to $d \gg n$ (where d is the number of genes and n is the number of samples)[1]. The high dimensional nature of DNA microarray datasets with a small number of samples presents a well-known phenomenon named the curse of dimensionality[12]. As the dimensionality increase, the computational cost increases too[11]. Usually, this growth is exponential. On the other side, observations and estimations are usually imprecise or uncertain when the number of genes increasing [3].

For overcoming these problems, it is necessary to find a way to reduce the number of genes and uncertainty. This work aims to handle the curse of dimensionality and uncertainty /imprecise by gene reduction technique and fuzzy logic concept respectively. Some methods manage these two issues coming in A and B with details. For overcoming these problems, it is necessary to find a way to reduce the number of genes and uncertainty. This work aims to handle the curse of dimensionality and uncertainty /imprecise by gene reduction technique and fuzzy logic concept respectively. Some methods manage these two issues coming in A and B with details.

A. handling the curse of dimensionality Concern with DNA microarray datasets is that how to earn valuable information. Figure 1 shows a scheme of handling high-dimensional DNA microarray datasets.

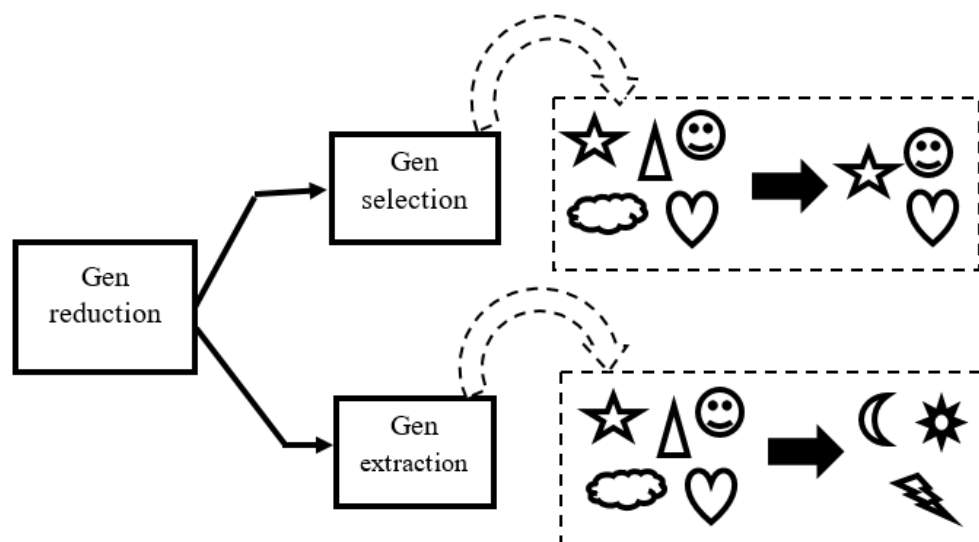


Figure 1: a scheme of handling high-dimensional DNA microarray datasets

There are two general methods for gene reduction: gene selection and gene extraction. There are two general methods for gene reduction: gene selection and gene extraction.

Gene selection is a process that selects only the most important and relevant genes of an object. Redundant or irrelevant genes are discarded before they have a chance to reduce the performance of the classification system. The gene selection methods based on heuristic search algorithms, ensemble, filter, and wrapper are in this category [1, 14].

Gene extraction methods transform the original genes set from a higher dimensional space to a lower one. The genes are not selected at all but the given gene space is projected to a new gene space. The effort to reduce the redundant information from the DNA microarray datasets and improve the performance of the learning [14].

Here are some popular methods used to reduce genes. Principal component analysis (PCA) is the most well-known of them. It finds principal components (PCs) in the DNA microarray datasets from the covariance matrix. PCs are uncorrelated eigenvectors which representing some proportion of variance in the DNA microarray datasets. The research shows when PCA and its variations compute the principal components (PCs) of a dataset there is no guarantee that the PCs will be related to the class variable [9].

Independent Components Analysis (ICA) can find the correlation between the DNA microarray datasets and de correlates the DNA microarray datasets by maximizing or minimizing the contrast information. This is called whitening. The whitened matrix is then rotated to minimize the Gaussianity of the projection and in effect retrieve statistically independent data [9].

Wavelet decomposition is a transformation method that shows a trend (approximation coefficients) and discontinuities (detail coefficients). Classification can be done using a Wavelet basis based on compactness and finite energy characteristic of the Wavelet function. PCAs and ICAs based methods do not detect localized genes. Also, their gene transform schemes depend on a training dataset of DNA microarray datasets. When the training dataset change, the basis vector change too. In other words, a new basis vector must construct a new training dataset. PCA and ICA have a large computation basis for DNA microarray datasets. Wavelet transform can be effective to extract a set of Wavelet bases that aims to detect the localized genes contained in DNA microarray datasets. Also, operations such as adding, deleting, and changing on the training dataset do not affect the computation of other sample vectors. [17].

B. Handling uncertainty/imprecise DNA microarray datasets suffer scaling problems, which are owing to a large number of gene measurements. When the number of genes increases, observations and estimations are ordinarily incomplete or uncertain that is coming from missing values or wrongly measured genes [3, 9]. In these conditions modeling and handle the imprecise information, Fuzzy logic can reach more informative results for the classification and decision-making.

A scope of fuzzy logic is Interval analysis. It is a field proposed by R. E. Moore [10] which suppose that, in the real world, measurements and estimations are ordinarily incomplete or ambiguous and, consequently, they do not represent the datasets accurately. According to this field, if accuracy is required, datasets be measured as intervals enclosing the actual quantities. Interval analysis presents techniques to control errors in numeric calculations dealing with intervals [3].

Given the challenges of A and B in DNA microarray datasets, a method introduced in this paper use the advantage of the uncertainty and Wavelet decomposition to extract a subset of genes by a more accurate and low cardinality. The rest of this paper is organized as follows: preliminaries

of the proposed method are given in Section 2, the proposed method is explained in Section 3, implementation and the results are provided in Section 4, and finally, conclusions are given in Section 5.

2. Preliminaries

In this section, the preliminaries of the proposed method are stated. These preliminary involve Linguistic variables, Weighted Averaging Based on Levels (WABL), Wavelet decomposition, and support vector machines as a classifier.

2.1. Linguistic variables

A linguistic variable is defined by a quintuple $(x, T(x), U, G, M)$: [8, 13] X is the name of the variable. U call variables universe of discourse that shows the range of possible values of a linguistic variable. $T(x)$ is the term set (linguistic values) of x with each value being a fuzzy variable (approximate values) defined on U . G is a set of syntactic rules that generate T from a set of primitive terms. M is a semantic rule, which assigns to each linguistic term its meaningan appropriate fuzzy number defined on the range of the base variable (which is a fuzzy subset on the universe of discourse U).

Linguistic variables use fuzzy variables as their values. A fuzzy variable is characterized by a triple $(X, U, R(X))$ in which: X is the name of the variable, U is a universe of discourse, and $R(X)$ is a fuzzy subset of U .

definition1: A fuzzy number with membership function is in the form

$$\mu_A(x) = \begin{cases} \left(\frac{x-a}{b-a}\right)^{\frac{1}{s}} & x \in [a, b) \\ \left(\frac{c-x}{c-b}\right)^{\frac{1}{s}} & x \in [b, c) \\ 0 & \text{o.w} \end{cases} \quad (2.1)$$

Where s is a parameter and greater than zero. It will be named as a parametric triangular fuzzy number $A=(s; a,b, c)$. Figure 2 shows different forms of parametric triangular fuzzy numbers based on the different range of s .

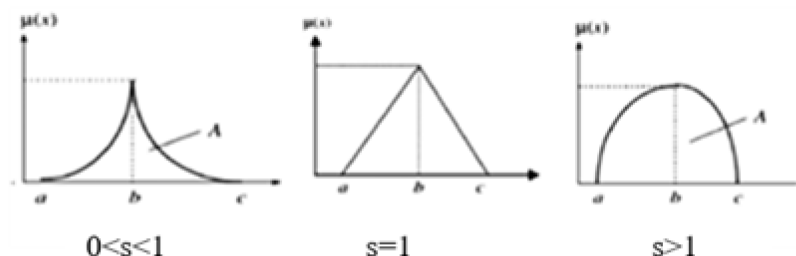


Figure 2: different forms of parametric triangular fuzzy numbers based on the different range of s [13]

In this paper, following representation of a fuzzy number is used

$$A = \bigcup_{\alpha \in [0,1]} (\alpha, A^\alpha) \quad (2.2)$$

Where $A = [L_A(\alpha), R_A(\alpha)]$ and functions $L : [0, 1] \rightarrow E \equiv (-\infty, \infty)$ and $R : [0, 1] \rightarrow E$ determine the left and right sides of the fuzzy number (the LR representation of the fuzzy number) respectively [13].

2.2. Weighted Averaging Based on Levels (WABL)

WABL is a defuzzification method that is based on level sets (level-based methods) of the fuzzy numbers [6, 5]. It utilizes aggregation of the values of averaged representatives of all level sets. Let A be a fuzzy number shown by the LR representation. The average representative of this fuzzy number is determined by the following equation:

$$WABL(A) = \int_0^1 (c_L L_A(\alpha) + c_R R_A(\alpha)) p(\alpha) d\alpha \quad (2.3)$$

Where $c_L \geq 0, c_R \geq 0, c_L + c_R = 1$. Here, the coefficients c_L and c_R reflect the degree of importance of the left and right sides, respectively, and function p is the distribution function of the importance of the level sets that can be:

$$p(\alpha) = (k + 1)\alpha^k \quad (2.4)$$

According to what is mentioned, the WABL equation is :

$$WABL(A) = c_R \left(c - \frac{k+1}{k+s+1} (c-b) \right) + c_L \left(c - \frac{k+1}{k+s+1} (b-a) \right) \quad (2.5)$$

2.3. Discrete Wavelet Transform (DWT)

Grossmann and Morlet [2] proposed the Wavelet transform method. It analyzes a signal by transforming its input time domain into a time-frequency domain [15].

$$DWT_x^\psi(u, 2^j) = \frac{1}{\sqrt{2^j}} \int x(t) \psi^* \left(\frac{t-u}{2^j} \right) dt \quad (2.6)$$

Where the discrete Wavelet transformed signal ($DWT_x^\psi(u, 2^j)$) is a function of the signal itself ($x(t)$), mother Wavelet (ψ), the scale parameter (j), and translation parameter (u); t stands for time. Discrete Wavelet transform (DWT) can represent a gene expression vector as a sum of Wavelets at different time shifts and scales. The DWT of a pattern is calculated by passing it through a series of filters. At each step, the pattern is decomposed to approximations (A) and details (D). Approximations show the general trend of the signal, i.e., low frequencies, and details focus on its detailed events, i.e., high frequencies. At each level $j + 1$ the approximation part from the previous level (A_j) is used for decomposition. This signal is convoluted to a low pass (g_p) and a high pass filter (h_p) to yield the new approximation (A_{j+1}) and detail (D_{j+1}). Figure 3 shows the Wavelet decomposition tree at three levels. The process of decomposition continue until desirable level

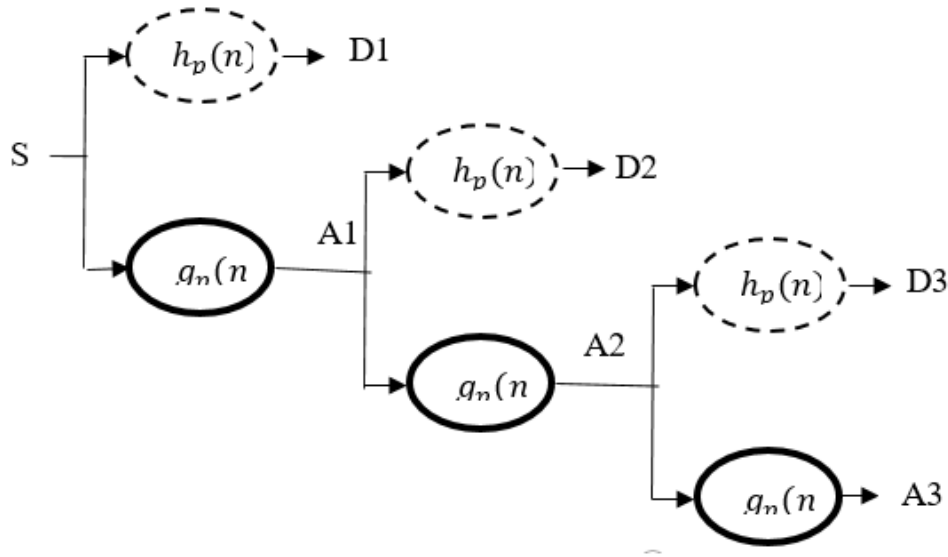


Figure 3: Wavelet decomposition tree, Variables s , $A1A3$ and $D1D3$ show original microarray data, approximation coefficients at three levels, respectively

reach. Convolution is denoted by:

$$A_{j+1}(n) = A_j \times g_p(n) = \sum_{k=-\infty}^{\infty} A_j(n)g(n - 2k) \quad (2.7)$$

$$A_{j+1}(n) = A_j \times h_p(n) = \sum_{k=-\infty}^{\infty} A_j(n)h(n - 2k) \quad (2.8)$$

At the end, according to Eq 2.7 and 2.8 one can reconstruct the original dataset by aggregation of detail decomposed parts (D_i) and the last level approximation decomposed part (A_m)[15].

$$signal(x_t) = A_m + \sum_{i=1}^m D_i \quad (2.9)$$

In the proposed method, one-dimensional Wavelet analysis apply on each pattern vector to obtain the detailed coefficients. These coefficients help to separate the classes.

2.4. Support Vector Machines(SVM)

SVM is a classifier that developed in the computer science community in the 1990s. Given a set of points $[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$ as input vector of i^{th} pattern and $y_i = \{-1, 1\}$ is corresponding output. The SVM try to find a maximal margin hyper plane $x : f(x) = x^T \beta + \beta_0 = 0$ which is the optimal separating hyper plane that is farthest from the training observations of both classes, for the distance of an observation from the hyper plane can be seen as a measure of our confidence that the observation was correctly classified. A classification rule induced from $f(x)$ is: $G(x) = \text{sign}[x^T \beta + \beta_0]$. Points on the border of the margin are called support vectors [7].

2.5. *K- Nearest Neighbors(kNN)*

KNN is a classifier that use distance metrics to predict classes of samples that still are not seen. It computes the distance between new sample and every object in the training set, and then K closest training object to new sample is selected. The predicated class, for new sample, is specified by majority voting on K-nearest neighbors [18].

2.6. *Decision Tree(DT)*

Tree Classifier works in three steps: 1. Choose the finest feature to partition the records. 2. Select the attribute as a decision node and divide the dataset into subsets of smaller size. 3. Begin tree building process recursively for every child till at least one of the following conditions match: All the tuples belong to the same attribute value, There are no more remaining attributes and There are no more instances[19].

3. Proposed method

This paper proposed the LR-Wavelet-SVM, LR-Wavelet-KNN and LR-Wavelet-DT methods for the classification of linguistic DNA microarray datasets, where they can be given as linguistic genes. Linguistic genes are defined by using triangular fuzzy numbers given as LR fuzzy numbers. Each gene applied in DNA microarray datasets formed the following phases:

Phase 1: Each gene's value is the center (c).

Phase 2: Each center value is multiplied with a random number generated between 0 and 0.20(R).

Phase 3: left value is computed with the subtraction of R with c.

Phase 4: the right value is computed with the summation of R with c.

So we have a fuzzy number for each measured gene. Then weighted averaging based on levels (WABL) method is applied as the defuzzification method for each gene in DNA microarray datasets. These operations help to control the imperious and uncertainty coming from the measurement. For controlling the domain values of each gene, they are normal in the range of 0 and 1 by Eq 3.1

$$x = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.1)$$

In the next step, important genes extract from DNA microarray datasets. Due to the DNA microarray datasets are high dimensional, many of the measured genes do not have fundamental changes during the experiment. To make it simple to discover the considerable genes, the genes filter with less than 30% profile variance was put aside. Also to improve the classification performance of DNA microarray datasets, a One-dimensional Wavelet decomposition transform was performed on each pattern from filtered DNA microarray datasets sets. Detail coefficients use to disclose the difference between classes in datasets. For this purpose, Daubechies basis 7 (db7) which has 7 non-zero coefficients of orthogonal wavelet basis is applied. Thus, the decomposition was done up to 4 levels. After wavelet genes extract, SVM use to classify the DNA microarray datasets. Figure 4 shows the process of classification by the proposed method that is based on linguistic DNA microarray datasets.

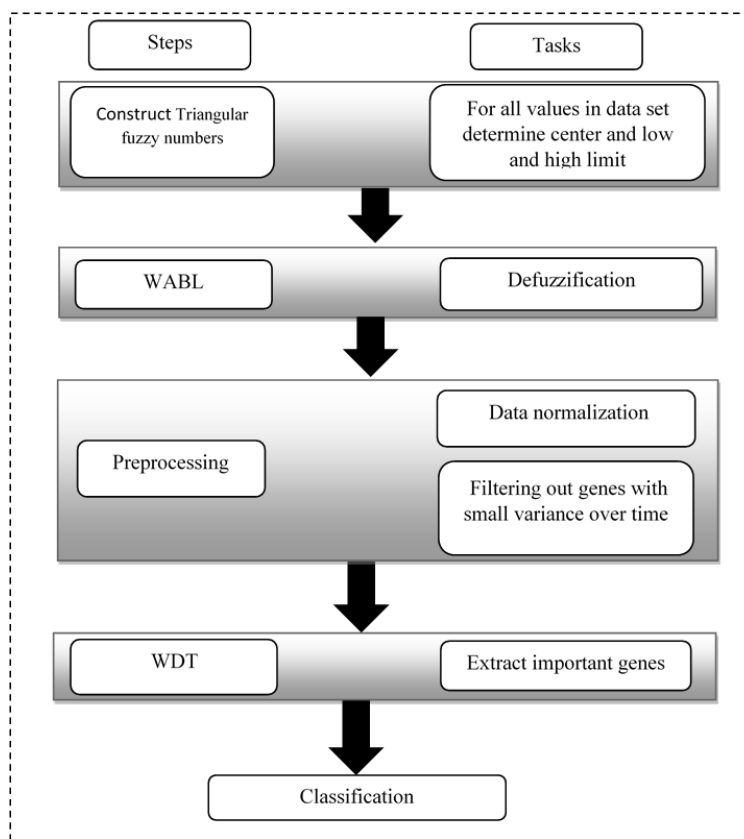


Figure 4: process of classification by LR-wavelet-SVM/KNN/DT

4. Implementation and result evaluation

The proposed algorithms were performed in MATLAB 2015b (intel i7, 1.80 GHz, 16 Gb RAM). These methods applied on three biological datastes namely, Colon, Lung and Leukemia have been selected from ASU dataset repository. The specifications of all datasets are given in Table 1. In this experimental study, three methods are used. First, the SVM, KNN and DT with-

<i>Dataset</i>	sample	genes	number of classes
<i>Colon</i>	62	2000	2
<i>Lung</i>	203	3312	5
<i>Leukemia</i>	72	7070	2

Table 1: data sets

out gene reduction on numerical datasets are used. Second, the Wavelet-SVM, Wavelet-KNN and Wavelet-DT algorithms based on numerical data sets are applied. Third is the LR-Wavelet-SVM, LR-Wavelet-KNN, LR-Wavelet-DT algorithms based on linguistic datasets. In all cases, the value of parameter $K=3$ (K is the number of nearest neighbors). This paper aims to show that fuzzy numbers and wavelet gene extraction make a significant results. The results shows composition of fuzzy numbers and wavelet decomposition can be effective. Proposed methods use the WABL

method that has some parameters. Table 2 shows the setup of these parameters. Parameters are set by trial and error.

<i>parameteres</i>	s	k	c_L	c_R
<i>value</i>	0	1	0.5	0.5

Table 2: setting up

Fivefold cross-validation and classification accuracy are used to evaluate the performance of the proposed methods. Let TP; TN; FP, and FN be the number of true positive (cancer), true negative (control), false positive, and false-negative samples. This criterion defines as follow:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.1)$$

For better comparison, experiments run 10 times. Table 3 shows comparison average classification accuracy between LR-Wavelet-SVM and SVM without gene reduction and Wavelet-SVM. There is average KNN classification accuracy in Table 4 based on LR-Wavelet-KNN and KNN without gene reduction and Wavelet-KNN. Also the results based on DT classifier are in Table 5.

<i>dataset</i>	# genes	SVM	Level	#extractedgenes	Wavelet-SVM	LR-Wavelet-SVM
<i>Colon</i>	2000	75.51	1	706	77.44	79.92
			2	359	77.56	81.64
			3	186	77.44	77.97
			4	99	76.03	78.30
<i>Lung</i>	3312	74.37	1	1165	74.37	74.48
			2	589	74.37	74.30
			3	301	73.38	74.38
			4	175	73.40	73.99
<i>Leukemia</i>	7070	84.29	1	2477	97.33	97.79
			2	1245	97.33	96.74
			3	629	94.57	94.34
			4	321	91.62	92.80

Table 3: average SVM classification accuracy result

dataset	# genes	KNN	Level	#extractedgenes	Wavelet-KNN	LR-Wavelet-KNN
Colon	2000	69.62	1	706	72.44	75.13
			2	359	71.03	77.83
			3	186	62.49	73.23
			4	99	75.64	70.97
Lung	3312	74.87	1	1165	75.37	75.56
			2	589	74.90	75.51
			3	301	74.40	75.02
			4	175	73.91	74.88
Leukemia	7070	76.38	1	2477	76.10	78.55
			2	1245	72.00	80.19
			3	629	78.95	78.46
			4	321	77.62	79.91

Table 4: average KNN classification accuracy result

dataset	# genes	DT	Level	#extractedgenes	Wavelet-DT	LR-Wavelet-DT
Colon	2000	73.85	1	706	59.74	66.99
			2	359	74.36	76.41
			3	186	67.95	61.15
			4	99	59.87	63.56
Lung	3312	68.51	1	1165	67.00	71.25
			2	589	68.01	69.51
			3	301	64.11	69.98
			4	175	65.54	68.42
Leukemia	7070	93.14	1	2477	75.05	73.90
			2	1245	68.10	77.10
			3	629	69.33	67.16
			4	321	72.38	65.95

Table 5: average DT classification accuracy result

Table 3, 4 and 5 shows that the proposed algorithm has better results in terms of average classification accuracy for DNA microarray datasets. It can be seen that fuzzy numbers help to control imprecisely and wavelet gene extraction by extracting orthogonal basis help to manage scaling problem in DNA microarray datasets. So results show that composition of Linguistic variables and wavelet decomposition improve the classification accuracy effectively. Wavelet-SVM and LR-Wavelet-SVM have better accuracy than SVM at each level For colon and leukemia. In lung dataset, LR-Wavelet-SVM is the best. Also, Wavelet-SVM gets a result equal to SVM but in lower genes. In relation to the KNN classifier, the proposed algorithm was effective. Although the accuracy rate obtained for the DT with all genes in leukemia dataset is the best, nevertheless fuzzy numbers have helped to increase the accuracy rate obtained in the dimension reduction with wavelet. So the accuracy rate obtained with the DT classifier in the proposed algorithm is better than the wavelet-DT. In general, LR-wavelet-SVM classification has achieved better results than LR-wavelet-KNN and LR-wavelet-DT. In LR-Wavelet-SVM, on average, Lung and Leukemia have 74.48% and 97.79% accuracy in level one with 1165 and 2477 genes respectively. For the colon dataset this result is coming in level 2 with 81.67% accuracy and 359 genes. Table 6 has been created to compare average SVM classification accuracy with previous works. In SVM classifier, this comparison shows that in addition to managing dimension reduction, controlling uncertainty can improve outcomes too.

5. Conclusion

In this paper, the LR-Wavelet-SVM, LR-Wavelet-KNN and LR-Wavelet-DT methods are proposed. These methods try to control inaccuracy in the DNA microarray datasets to increase classification accuracy. So LR fuzzy number and WABL are applied. To overcome the curse of dimensionality Wavelet decomposition is used. Each pattern represented by wavelet basis. It can extract

Method	Leukemia	Colon
<i>LS – SVMlinearkernel(noregularization)</i> [4]	87.39	51.37
<i>kPCARBF + FDA(unsupervisedPCselection)</i> [4]	89.50	75.11
<i>kPCARBF + FDA(supervisedPCselection)</i> [4]	92.02	64.07
<i>kPCAlin + FDA(supervisedPCselection)</i> [4]	92.44	76.84
<i>LS – SVMRBFkernel</i> [4]	93.56	81.39
<i>PCA + FDA(supervisedPCselection)</i> [4]	93.56	76.84
<i>PCA + FDA(unsupervisedPCselection)</i> [4]	94.40	76.84
<i>kPCAlin + FDA(unsupervisedPCselection)</i> [4]	97.40	76.84
<i>Wavelete + SVM</i> [17]	97.33	77.56
<i>LR + Wavelete + SVM</i>	97.79	81.67

Table 6: comparison average accuracy with previous work

localized genes. In addition, detail coefficient can detect the situation of change for classification. LR-Wavelet-SVM has better result than LR-Wavelet-KNN and LR-Wavelet-DT. so it compares with other methods. The results show that the proposed method by SVM has a significant performance by considering inaccuracy in datasets and dimension reduction. Also, detail coefficients at the first and second levels are robust to identify the genes DNA of microarray datasets.

References

- [1] A. Alkuhlani, M. Nassef and I. Farag, Multistage feature selection approach for high-dimensional cancer data, *Soft Comput.*, **21** (2017), 6895–6906.
- [2] A. Grossmann and J. Morlet, Decomposition of hardy functions into square integrable wavelets of constant shape, *SIAM J. Math. Anal.*, **15**(4) (1984), 723–736.
- [3] A.M.S Roque, C. Mate, j. Arroyo and A. Sarabia, Impl: applying multi-layer perceptrons to interval-valued data. Neural processing letters, *SIAM J. Math. Anal.*, **25** (2007), 157–169.
- [4] D.S Huang and CH. Zheng, Independent component analysis-based penalized discriminant Method for tumor classification using gene expression data, *Biostat. Bioinform. Biomath.*, **22**(15) (2006) 1855–1862.
- [5] E. Nasibov and A. Mert, On methods of defuzzification of parametrically represented fuzzy numbers, *Autom. Control Comput. Sci.*, **41** (2007), 265–273.
- [6] E. Nasibov, Aggregation of fuzzy information on the basis of decompositional representation, *Cybern. Syst. Anal.*, **41**(2), (2005), 309–318.
- [7] L. Yang and Z. Xu, Feature extraction by pca and diagnosis of breast tumors using SVM with de-based parameter tuning, *International Journal of Machine Learning and Cybernetics*, **10** (2017), 591–601.
- [8] L.A. Zadeh, The concept of linguistic variable and its application to approximate reasoning, *Inf. Sci.*, **8**(3) (1975), 199–249.
- [9] Z.M. Hira and D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, *Advances in Bioinformatics*, **2015**(5) (2015), 1–13.
- [10] R.E. Moore, R.B. Kearfott and M.J. Cloud, *Introduction to Interval Analysis*, Society for Industrial and Applied Mathematics Philadelphia, 2009.
- [11] S. Sarbazi-azad, M. Saniee Abadeh and M.E. Mowlaei, Using data complexity measures and an evolutionary cultural algorithm for gene selection in microarray data, *Soft computing letters*, DOI:10.1016/j.socl.2020.100007, 2020.
- [12] S. Tabakhi, A. Njafi. R. Ranjbar and P. Moradi, Gene selection for microarray data classification using a novel ant colony optimization, *Neurocomputing*, **168** (2015), 1024–1036.
- [13] S.K Sava and E. Nasibov, A fuzzy id3 induction for linguistic data sets, *International journal of intelligent systems*, **33**(4) (2018), 1–21.
- [14] S.M Saqlainshah, F. Alishah, S.A Hussain and S. Batool, Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods, *Comput. Electr. Eng.*, **84**(10) (2020), DOI:10.1016/j.compeleceng.2020.106628.

- [15] S.O.M kasha and M. A Akbarzadeh, A framework for short-term traffic flow forecasting using the combination of wavelet transformation and artificial neural networks, *J. Intell. Transp. Syst.*, **23**(1) (2019), 1–12.
- [16] X. Zheng, W. Zhu, CH. Tang and M. Wang, Gene selection for microarray data classification via adaptive hypergraph embedded dictionary learning, *Gene*, **706** (2019), 188–200.
- [17] Y. Liu, Wavelet feature extraction for high-dimensional microarray data, *Neurocomputing*, **72** (2009), 985–990.
- [18] X. Wu and et al., Top 10 algorithms in data mining, *Knowledge and Information Systems*, **14** (2008), 1–37.
- [19] J. Basavaiah and A.A. Anthony, Tomato Leaf Disease Classification using Multiple Feature Extraction Techniques, *Wireless Personal Communications*, **115** (2020), 633–651.