Annals of Emerging Technologies in Computing (AETiC)
Vol. 7, No. 1, 2023

Research Article

The Theory of Probabilistic Hierarchical Learning for Classification

Ziauddin Ursani[1,*] and Ahsan Ahmad Ursani[2]

[1]University of Bristol, UK
ziaursani@yahoo.com
[2]Mehran University of Engineering and Technology, Pakistan
ahsan.ursani@faculty.muet.edu.pk
*Correspondence: ziaursani@yahoo.com

Received: 31st May 2022; Accepted: 21st December 2022; Published: 1st January 2023

Abstract: Providing the ability of classification to computers has remained at the core of the faculty of artificial intelligence. Its application has now made inroads towards nearly every walk of life, spreading over healthcare, education, defence, economics, linguistics, sociology, literature, transportation, agriculture, and industry etc. To our understanding most of the problems faced by us can be formulated as classification problems. Therefore, any novel contribution in this area has a great potential of applications in the real world. This paper proposes a novel way of learning from classification datasets i.e., hierarchical learning through set partitioning. The theory of probabilistic hierarchical learning for classification has been evolved through several works while widening its scope with each instance. The theory demonstrates that the classification of any dataset can be learnt by generating a hierarchy of learnt models each capable of classifying a disjoint subset of the training set. The basic assertion behind the theory is that an accurate classification of complex datasets can be achieved through hierarchical application of low complexity models. In this paper, the theory is redefined and revised based on four mathematical principles namely, principle of successive bifurcation, principle of two-tier discrimination, principle of class membership and the principle of selective data normalization. The algorithmic implementation of each principle is also discussed. The scope of the approach is now further widened to include ten popular real-world datasets in its test base. This approach does not only produce their accurate models but also produced above 95% accuracy on average with regard to the generalising ability, which is competitive with the contemporary literature.

Keywords: ***Classification; Hierarchical learning; Probabilistic learning; Set partitioning; Supervised learning***

1. Introduction

Main faculty of human consciousness is its ability to classify. It is our ability of classification through which we become aware of things in existence by distinguishing them from each other. Providing this ability to machines now constitutes major part of artificial intelligence. Its application has now touched every aspect of human life, including but not limited to renewable energy e.g. [1], chemometrics e.g. [2], cyber security e.g. [3], natural language processing e.g. [4], finance e.g. [5], microbiology e.g. [6], ecology e.g. [7], and healthcare e.g. [8]. Therefore, proposal of any novel way through which machines could learn to classify has an enormous potential of application in wide variety of areas. The theory of probabilistic hierarchical learning for classification was introduced in a conference paper [9]. It was moulded into a theory after a series of earlier papers [10-13]. This theory introduces the hierarchical model of learning. In the hierarchical learning, multiple models are learnt hierarchically over their subdomains, each containing elements from several flat classes. The model in each hierarchy is learnt on a subset of the training set. The configuration of that subset is also decided during training in that corresponding hierarchy. Therefore, a model and its application subdomain are learnt altogether. Since this subdomain may contain elements from several classes therefore, this subset is not a part of class hierarchy, this is a part of a hierarchy of

Ziauddin Ursani and Ahsan Ahmad Ursani, "The Theory of Probabilistic Hierarchical Learning for Classification", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 61-74, Vol. 7, No. 1, 1st January 2023, Published by International Association for Educators and Researchers (IAER), DOI: 10.33166/AETiC.2023.01.005, Available: http://aetic.theiaer.org/archive/v7/v7n1/p5.html.

subdomains corresponding to a hierarchy of learnt models. The interesting thing about these models is that they are errorfree over their respective subdomains.

The hierarchical learning should not be confused with a hierarchical classification or any of its contexts such as a natural hierarchical classification (NHC) or a methodological hierarchical classification (MHC). The hierarchical learning does not have hierarchical classes like NHC, such as classification of all biological organisms on earth e.g. [14] and hierarchical classification of diseases by world health organization[1]. The classes in the hierarchical learning are flat and cannot be represented through a directed acyclic graph e.g. [15] as can be done in the NHC [16]. The hierarchical learning is not even an artificial hierarchy of classes such as one done in MHC where classes are flat, rather, classification itself is performed hierarchically. One of its kind are Hierarchical Support Vector Machines [17], where hierarchies are decided manually. The MHC is also done by automated generation of meta-classes such as in a handwriting character recognition system [18].

The hierarchical learning should not be confused with ensemble learning [19] which does also deal with multiple models but differs from hierarchical learning in several diverse ways based on the method of learning, the domains of models, the error handling, and the application methodology. The ensemble learning models are not hierarchically learnt, their domain covers the whole training set, their errors are averaged, and they can be applied simultaneously in parallel. Whereas in the hierarchical learning, each model has a domain which is a unique subset of the training set, it is errorfree, and models can only be applied sequentially.

The hierarchical learning is a supervised classification learning method, but it should not be confused with other methods of supervised classification learning such as Neural Networks e.g. [20-22], Decision trees e.g., [23] and Naïve Bayes e.g. [24], as none of them follows the scheme of hierarchical learning. The multiple hierarchies in the hierarchical learning should not be confused with multiple layers of learning such as in Deep Learning e.g. [20-21]. This is because the number of layers in these networks are set prior to learning, however, number of hierarchies are not decided prior to learning, they are part of the learning process instead. Therefore, hierarchical structure in hierarchical learning is a learnt model rather than a predefined structure such as in Recurrent Neural Network e.g. [22]. Since the hierarchical learning uses the probabilistic model for the class discrimination as most of the linear e.g. [25] and nonlinear e.g. [26] discriminants do, it is termed as the theory of probabilistic hierarchical learning. The interesting thing about its probabilistic model is its flexibility that allows for the negative probabilistic values, a concept that was first introduced in an entirely different field i.e., Quantum mechanics [27].

Earlier, the theory was emphasized through mathematical means. In this paper, the theory is redefined and revised by introducing four mathematical principles including the principle of successive bifurcation, the principle of two-tier discrimination, the principle of class membership and the principle of selective data normalization. The first principle provides structure for the hierarchical application of learning, the second principle supports this hierarchical structure with mathematical logistics, the third principle tailors the rule of probabilistic class membership to suit the hierarchical structure, and the fourth principle applies data normalization in a selective way which again helps in hierarchical learning. These four strong mathematical principles provide more fidelity to theory that its application is now extended from 5 to 10 datasets.

The rest of the paper divides into eight sections. Section 2 introduces the principle of successive bifurcation, section 3 covers principle of two-tier discrimination, section 4 discusses principle of class membership, and section 5 details the principle of data normalization. The theory's ability to develop errorfree models is tested in section 6. Section 7 presents experimental results on generalization ability of theory. The generalizing ability of theory is compared with literature in section 8. Lastly, conclusions and future work are orchestrated in section 9.

2. Principle of Successive Bifurcation

The concept behind the hierarchical learning is creating a set of disjoint subsets whose union is the full training set and a developing a simple model corresponding to each of these subsets. The model is learnt using the unclassified samples of the training set and then domain is assigned to the model consisting of

[1] WHO (2016) International statistical classification of diseases and related health problems. - 10th revision, Fifth edition, 2016. 3 v.

the samples which are classified by it correctly. The rest of the samples are retained in the unclassified set of the training set to train another model in the next hierarchy. This means that at each hierarchy, the learnt model bifurcates the training set into classified and unclassified subsets. Therefore, the hierarchical learning follows the principle of successive bifurcation of the training set. This principle is visualised in Figure 1.

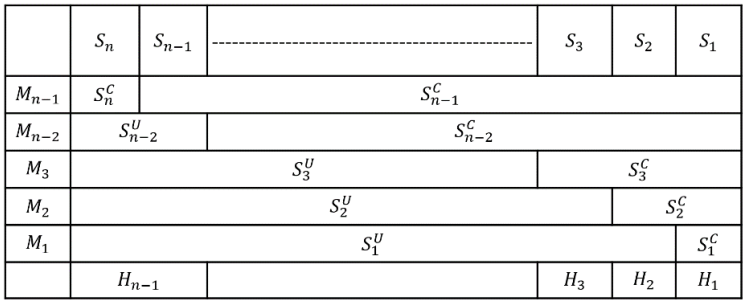

Figure 1. Principle of successive bifurcation

Figure 1 depicts the training set, which is partitioned into subsets $S_1, \ldots\ldots, S_n$ by the models $M_1 \ldots\ldots M_{n-1}$. Each model M_i in hierarchy H_i bifurcates the training set into classified S_i and unclassified S_i^u subsets. The last subset S_n is either NULL subset or the subset containing samples from one class only, illustrating the end of hierarchical classification procedure.

The principle of successive bifurcation described above generalises the hierarchical learning procedure. At any given point during the heirarchical learning the four statements $(a - d)$ of equation 1, must be satisfied.

$$\forall_{i=m} H_i : \begin{cases} a: M_{i-1} \prec M_i : S_i \to C \\ b: \quad U = S_i^c \cup S_i^u \\ c: \quad S_i^c = \bigcup_{1 \le j \le i} S_j \\ d: \quad m = n - 1 \end{cases} \tag{1}$$

Where,

m = Total number of hierarchies

n = Total number of subsets

H_i = Level i in the hierarchy

M_i = Model achieved at H_i

S_i = Subset of training set classified by model M_i at H_i

C = Class set

U = the training set

S_i^c = Set of classified samples at the end of H_i

S_i^u = Set of unclassified samples at the end of H_i

Therefore, the four statements $(a - d)$ of the principle of successive bifurcation in equation 1, can be described as follows.

a. At H_i, M_{i-1} precedes M_i, whose domain is the subset S_i and class set C is its codomain.
b. At H_i, the training set U is the set of all samples including S_i^c and S_i^u
c. At H_i, S_i^c is the union of all sets classified on levels 1 through i
d. At H_i, total number of hierarchies is always one less than total number of subsets.

2.1. Postulate 1

The principle of successive bifurcation can be used as a tool to replace a high complexity model with several simpler models each with a constrained domain representing a unique subset of the training set such that union of all constrained domains equals the training set.

2.2. Implementation

The Figure 2 presents flowchart showing implementation of principle of successive bifurcation. The algorithm learns from the training set which results in its bifurcation into the classified and the unclassified subsets. If the unclassified subset is not empty and has samples belonging to more than one class, then the unclassified set is used as the training set in the next hierarchy. In the figure, X refers to subset containing members from only one class.

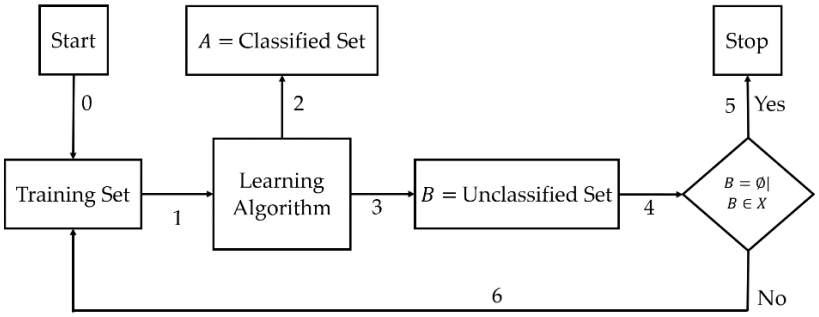

Figure 2. Principle of successive bifurcation implementation

3. Principle of two-tier Discrimination

The phrase two-tier discrimination refers to the two-step discrimination function of each model. Each model must perform two-step discrimination. The two-step discrimination means the discrimination in terms of classification of samples and then the discrimination in terms of partitioning the training set into two subsets that is, the subset within its domain and subset outside its domain. The expression 2 encapsulates this whole idea.

$$\begin{cases} q_h \in \{C_k \cap S_i\} & P(i,h,k) > \forall_{j \neq k} P(i,h,j) + \psi \\ q_h \in S_i^u & otherwise \end{cases} \quad (2)$$

where,

$P(i,h,k)$ = Probability that the sample q_h is the member of class C_k w. r. t. model M_i.

ψ = Size of the set partition

The expression 2 says if the Probability $P(i,h,k)$ is greatest among all the classes by the minimum value equal to size of set partition, then sample q_h is the member of class C_k and lies within classified subset S_i, otherwise it belongs to unclassified subset S_i^u. Whereas, the size of set partition is the greatest margin by which the M_i could misclassify the sample during a training session. The size of set partition can be calculated through equation 3.

$$\psi = max\left(\forall h\, q_h \in \{C_j \cap S_i^u\} \left[P(i,h,k) - max\left(\forall_{k \neq j} P(i,h,j)\right)\right]\right) \quad (3)$$

From equation 3, it can be seen that ψ represents the size of set partition which is maximum difference among all the training samples between the largest and second largest probabilities of membership, whereas largest probability belongs to a wrongly assigned class.

The model in expression 2, describes a state in hierarchical learning at any given hierarchy H_i, where there are only two possibilities available that either the sample under the test is classified correctly, or its classification is postponed to the next hierarchy level. This process continues until the last hierarchy where it attains a state that either $S_i^u = \emptyset$ or $S_i^u \in X$ where X contains samples belonging to one class only. The class X can be called a remainder class. The remainder class does not have potential to introduce errors in the classification as it is distinguishable by the second tier of discrimination based on set partitioning. Therefore, this process can only end up in accurate classification of all samples. However, in the case of $S_i^u \in X$, the set partitioning $\psi > 0$ otherwise if $S_i^u = \emptyset$ then $\psi = 0$.

3.1. Postulate 2

Incorporating a set-partition in the framework of probabilistic class membership is here referred to as the principle of two-tier discrimination. The principle can eliminate misclassification of samples completely during hierarchical training, making the hierarchical learning model errorfree.

3.2. Implementation

Please refer to Figure 2. The flowchart between arrow 1 and arrow 4 doesn't show any discriminatory rules for the classified and unclassified set. Replace it with the flowchart in Figure 3, which shows two tier discrimination.

It can be seen in Figure 3, that in the first tier it is checked that whether the sample under investigation obeys expression 2. If it doesn't obey, then 'otherwise' clause of expression 2 is materialized, where it is sent to the unclassified set to postpone its classification to the next hierarchy. However, if it obeys expression 2

then in the second tier it is checked whether it is classified correctly. If it is classified correctly then it is sent to the classified set otherwise the set partitioning margin is reset for the learning algorithm to restart the classification for the current hierarchy again. The symbols used in the figure refer to the symbols described in expression 2 and equation 3.

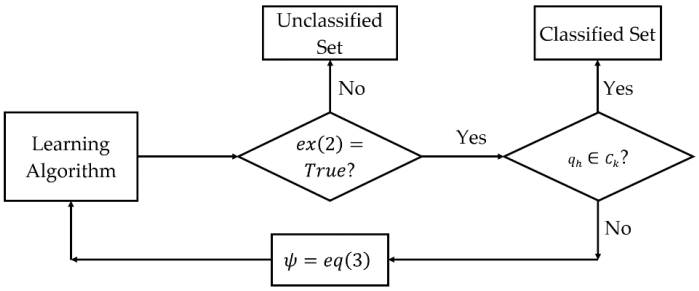

Figure 3: Principle of two-tier discrimination implementation

4. Principle of class membership

From the descriptions presented in section 3 about principle of two-tier discrimination (expression 2), it can be seen that the decision about the class membership in hierarchical learning is largely based on the probability. Now the question arises how to model this probability so that it could be useful in assignment of membership of a class. Equation 4 provides such a framework.

$$P_{(i,h,k)} = \begin{cases} \frac{\mu_{(i,h)} - \gamma_{(i,k,min)}}{\gamma_{(i,k,mean)} - \gamma_{(i,k,min)}}, & \mu_{(i,h)} \leq \gamma_{(i,k,mean)} \\ \frac{\gamma_{(i,k,max)} - \mu_{(i,h)}}{\gamma_{(i,k,max)} - \gamma_{(i,k,mean)}}, & otherwise \end{cases} \quad (4)$$

where

$\mu_{(i,h)}$ = Value of sample s_h w. r. t. model M_i

$\mu_{(i,k,min)}$ = Estimated minimum value of samples belonging to class C_k w. r. t. model M_i

$\mu_{(i,k,max)}$ = Estimated maximum value of samples belonging to class C_k w. r. t. model M_i

$\mu_{(j,k,mean)}$ = Estimated mean value of samples belonging to class C_k w. r. t. model M_i

The mean value of model M_i of class C_k can be estimated as follows.

$$\mu_{(i,k,mean)} = \frac{\sum_{h \in C_k} \mu_{(i,h)}}{n_k} \quad (5)$$

Where,

n_k = Number of samples in the training set belonging to class C_k

The maximum and minimum value among samples of class C_k w. r. t. model M_i can be estimated as,

$$\mu_{\left(i,k,\substack{max \\ min}\right)} = \mu_{(i,k,mean)} \pm \rho * \mu_{(i,k,sd)} \quad (6)$$

where

ρ = Range parameter, computed in theorem later in appendix.

$\mu_{(i,k,sd)}$ = Estimated standard deviation of samples belonging to class C_k, w. r. t. model M_i

The standard deviation $\mu_{(i,k,sd)}$ can be estimated as,

$$\mu_{(i,k,sd)} = \sqrt{\frac{\sum_{h \in C_k} \left(\mu_{(i,h)} - \mu_{(i,k,mean)}\right)^2}{n_k}} \quad (7)$$

It should be noted that in equation 4, if any of the conditions either $\mu_{(i,h)} < \mu_{(i,k,min)}$ or $\mu_{(i,h)} > \mu_{(i,k,max)}$ is true then such a condition will push probability towards negative zone, which shows flexibility of this model to accept negative probabilities - the concept quite well established in quantum mechanics [27]. However, this shows that probabilistic model of equation 4 depends on relative closeness of sample with respect to class means. It should be noted that the measure of relative closeness heavily depends on the computation of class boundaries. On the other hand, the computation of class boundary entirely depends on the estimates of the minimum and maximum of the class. Alternately, the minimum and maximum estimates of a class largely depend on the range parameter mentioned in equation 6 of the model. Obviously, this range parameter is different at different hierarchical levels as size of the training subset continues to become smaller with each subsequent hierarchical level. Therefore, this parameter should be

set according to the size of a training subset. Furthermore, as we move to the higher levels of hierarchies the subset not only become smaller but also their sample spread becomes larger, as remaining samples are only those who failed to fit in the earlier models. This necessitates to estimate maximum possible value of this parameter to capture the structure of the subset. We have proved in our theorem presented in appendix that range parameter for maximal spread is equal to $\sqrt{n-1}$. Please see appendix, where we have calculated its value.

The experiments have shown that the value of range parameter as computed is only advantageous at tail end hierarchies where subset sizes are substantially curtailed. For the rest of the hierarchies its value trends around the value of π. Therefore, final value of range parameter is shown in equation 8.

$$\rho = min\left(\pi, \sqrt{n-1}\right) \tag{8}$$

4.1. Postulate 3

The principle of class membership based on the relative closeness of a sample to the class mean provides a convenient estimate of probability of class membership provided class boundaries configured carefully according to sample size.

4.2. Implementation

Follow the steps below:

- Compute mean based on equation 5.
- Compute standard deviation based on equation 7.
- Compute range parameter based on equation 8
- Compute minimum and maximum of the class based on equation 6
- Compute probability from equation 4.
- Use this probability to assign class to samples according to equation 2.

5. Principle of selective data normalisation

The principle of selective data normalisation involves two steps. First is the development of a mechanism that decides whether data normalisation is needed. The second step decides how it should be done. Let us introduce the notion of range ratio, which can help with the decision whether to normalise data. The range ratio is the ratio between maximum and minimum of data belonging to a feature. The range ratio γ can be computed as shown in equation 9.

$$\gamma = \frac{max\left(abs(min), abs(max)\right)}{min\left(abs(min), abs(max)\right)} \tag{9}$$

Now let us consider a feature as potent if its range ratio is greater than or equal to two, else consider it as impotent. Now if the dataset has minimum of 50% features as potent then there is no need to normalise the data. However, data normalization is needed if this is not the case. The equation 10, provides the method of data normalisation.

$$v_m = \frac{v_0 - v_{min}}{v_{max} - v_{min}} \times (u_{max} - u_{min}) + u_{min} \tag{10}$$

Where,

v_0 = original value of a feature of a sample

v_m = modified value of a feature of a sample

v_{min} = minimum value of a feature among all samples

v_{max} = maximum value of a feature among all samples

u_{min} = minimum value of all features among all samples

u_{max} = maximum value of all features among all samples

The equation 10, proportionally distributes the values from u_{min} to u_{max} and it is applied only to features which satisfy the condition $v_{max} = u_{max}$.

5.1. Postulate 4

The principle of selective data normalisation is based on the range of values of different features in the dataset, which can then be utilized to decide whether the dataset needs data normalisation procedure.

5.2. Implementation

Follow the steps below.

- Compute the range ratio γ according to equation 9 for each of the features.
- Compute the number of potent features f_p with $\gamma \geq 2$.
- If $f_p \geq \frac{f_t}{2}$ then stop (f_t = total number of features).
- Choose the features to be normalised f_n satisfying the condition $v_{max} = u_{max}$.
- Apply data normalisation to features f_n according to equation 10.

6. Errorfree model test

An errorfree model test is the test prepared to verify whether the models developed through hierarchical learning could be errorfree as claimed in postulate 2 represented by expression 2. An errorfree model should accurately classify the whole dataset, or it should not misclassify any of the samples of the dataset. Therefore, accurate classification of some of the well-known datasets through a chain of simpler models trained in hierarchical order may validate postulate 2 of the theory. Some of the challenging real-world datasets from the UCI repository[2] were chosen to test this hypothesis. The details of those datasets are given in Table 1 for feature and class description. In Table 1, column 1 gives name of the dataset, its domain and reference. Column 2 provides feature description in sequence as they appear in the dataset. Column 3 contains class names and finally column 4 states number of samples in each class. It also gives the total number of samples in the dataset. The datasets are alphabetically sorted.

Table 1. Class and Feature Description of datasets

Dataset	Feature List	Class Name	Nr. of samples
(1)	(2)	(3)	(4)
Acute Inflammations Nephritis (Medical [28])	1. Temperature, 2. Nausea 3. Lumbar pain 4. Urine pushing 5. Micturition pains 6. Burning of urethra itch: swelling of urethra outlet	Positive	50
		Negative	70
		Total	120
Acute Inflammations Urinary (Medical [28])	Same as above	Positive	59
		Negative	61
		Total	120
Balance Scale (Psychological [29])	1. Left weight 2. Left distance 3. Right weight 4. Right distance	Balanced	49
		Left tipped	288
		Right tipped	288
		Total	625
Banknote Authentication (Image[3])	1. Variance of wavelet transformed image (WTI) 2. Skewness of WTI 3. Curtosis of WTI 4. Entropy of image	True	610
		False	762
		Total	1372
Breast Cancer Wisconsin (Diagnostic) (Image [30])	1. Radius (mean distance of points on the perimeter from the center) 2. Texture (standard deviation of gray-scale values) 3. Perimeter 4. Area 5. Smoothness (local variation in radius lengths) 6. Compactness (perimeter^2 / area - 1.0) 7. Concavity (severity of concave portions of the contour) 8. Concave points (number of concave portions of the contour) 9. Symmetry 10. Fractal dimension ("coastline approximation" - 1)	Malignant	212
		Benign	357
		Total	569
Car Evaluation	1. Buying cost 4. Number of seats	Unacceptable	1210

[2] https://archive.ics.uci.edu/ml/datasets.php

[3] D. Dua and C. Graff, (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

(Decision making [31])	2. Maintenance cost 3. Number of doors 5. Size of lug-boot 6. Level of safety	Acceptable	384
		Good	69
		Very good	65
		Total	1728
Iris (Botanical [25, 32])	1. Sepal length 2. Sepal width 3. Petal length 4. Petal width	Setosa	50
		Verginica,	50
		Versi-colour	50
		Total	150
Seeds (Image [33])	1. Area 2. Perimeter 3. Compactness 4. Length of kernel k 5. Width of k 6. Asymmetry coefficient 7. Length of k groove	Kama	70
		Rosa	70
		Canadian	70
		Total	210
User Knowledge Modelling (Educational [34])	1. Degree of study time for goal object materials (GOM) 2. Degree of repetition number for GOM 3. Degree of study time for related objects with GOM 4. Exam performance for related objects with GOM 5. The exam performance for GOM	Very Low	50
		Low	129
		Middle	122
		High	130
		Total	403
Wine (Chemical [35])	1. Alcohol 2. Malic acid 3. Ash 4. Alcalinity of ash 5. Magnesium 6. Total phenols 7. Flavanoids 8. Nonflavanoid phenols 9. Proanthocyanins 10. Color intensity 11. Hue 12. OD280 / OD315 of diluted wines 13. Proline	Class 1	59
		Class 2	71
		Class 3	48
		Total	178

It should be noted that in the Table 1 only 10 features are shown for the dataset of Breast Cancer Wisconsin Diagnostic [30], which are computed from a fine needle aspirate digitized image of breast mass. The dataset comprises 10 image features to characterize the cell nuclei. However, the dataset has been augmented to comprise 30 attributes, including mean attribute values, standard deviations, and largest deviation from the mean values.

An evolutionary algorithm [10-13] was used to train a model proposed hierarchical learning method realized in Microsoft Visual Studio C/C++. It should be noted that the data normalisation procedure was not applied for errorfree model test. The program was tried on the 10 datasets described in Table 1. The method was tried 30 times on each dataset using random seeds. It should be noted that whole dataset was taken as the training set. The trained models were later used to classify the same dataset. Each random trial ended up in accurate classification of the dataset. The description of models is given in Table 2.

Table 2 presents one model with least number of hierarchies from 30 random trials for all the ten datasets (Table 1). The column 1 gives the name of the dataset. The column 2 provides the number of hierarchies. The Column 3 shows the corresponding hierarchy level. The mathematical model in that hierarchy is presented in column 4. The column 5 reveals the number of samples classified by the model, whereas the column 6 mentions either the number of unclassified samples or the number of samples belonging to one last remaining class. Finally, the column 7 states the size of set partition. The datasets are presented in the same order as that of Table-1. The feature numbers in the models (column 4), correspond to feature numbers shown in Table-1. The following procedure can be followed to verify the models presented in the Table 2:

- Model value of each sample in the relevant subset of training set should be calculated by putting feature values of each sample in the models shown in Table 2.
- Classification of each sample should be done using principle of class membership. Its implementation procedure is given in the section 4.2.

If the readers find that any of the models given in the Table 4, misclassify any of the samples of the datasets, they should report to author along with details.

From the results shown in Table 2, one finds that the proposed method is able to classify four datasets namely Acute Inflammation Nephritis, Acute Inflammation Urinary, Balance Scale and Wine dataset in just one hierarchy, whereas Banknote Authentication, Iris and Breast Cancer Wisconsin Diagnostic are classified in only two hierarchies and finally Seeds, User knowledge modelling and car evaluation are classified in

three, five, and 13 hierarchies respectively. The accurate classification of the datasets in multiple hierarchies shows the ability of the proposed theory to model complex non-linear datasets with a high precision.

Table 2. Accurate models of classification datasets

DS	NH	HL	Model Description	No. of cl. Sp.	ucl/ rcl	S/Part.
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Acute Inflam. Nephritis	1	1	$0.8668f_2 + 0.4732f_3 + 0.2610f_1 - 0.4767f_6 - 0.9315f_5 + 0.9398f_4$	120	0	0.0000
Acute Inflam. Urinary	1	1	$0.7452f_3 - 0.7190f_4 + \frac{0.0020f_1}{0.7499f_6} - 0.7284f_5 \times 0.7855f_2$	120	0	0.0000
Balance Scale	1	1	$\frac{0.4469f_1}{0.2260f_3} - \frac{0.0823f_4}{0.0416f_2}$	619	6	0.0184
Banknote Authenti-cation	2	1	$0.4567f_3 - 0.0528f_4 + 0.3640f_2 + 0.7121f_1$	1340	32	0.8526
		2	$(0.1890f_4 - 0.2364f_3) \times (0.0904f_2 - 0.5019f_1)$	32	0	0.0000
Breast Cancer Wisconsin Diagnostic	2	1	$\frac{11.9558f_{12}}{2.1969f_{19}+2.3684f_{15}} \times \frac{0.3705f_{11}\times 0.8895f_5}{0.1627f_{17}} + (0.0871f_{23} + 497.222f_7) \times (0.496f_3 - 15.5294f_6) - \frac{0.4083f_4}{0.0643f_2\times 0.4172f_{28}\times 0.0431f_{14}} - \left\{\left(\frac{23.4939f_{20}}{1.1706f_8} - 9.1887f_{25} \times 0.1537f_1 - \frac{0.982f_{27}}{2.2949f_{16}} - 0.0065f_{13} - 0.5138f_{21}\right) \times \left(\frac{19.2656f_{29}+0.4549f_{22}}{0.4412f_{10}+0.1184f_{26}} + 1.6674f_{30} \times 3.5946f_{24}\right)\right\}$	549	20	0.1474
		2	$\frac{(1.9206f_{15}+0.1401f_6)\times 0.7436f_{27}\times(0.7285f_{25}-0.1565f_{28})\times 0.8941f_{22}\times 0.8295f_4}{0.0928f_{10}\times\left(\frac{0.6147f_{26}}{0.1009f_{19}}+\frac{0.8566f_2}{0.1397f_9}-0.5595f_{21}-0.4672f_5-\frac{0.651f_{20}}{0.1248f_{18}}\right)} - (0.9158f_7 + 0.2982f_{29} - 0.6528f_{13} \times 0.3787f_8) \times \frac{0.3687f_3\times 0.0367f_{12}}{0.1825f_1+0.6836f_{23}} \times (0.74f_{24} - 0.4713f_{14} + 0.1267f_{11} \times 0.9811f_{30} + 0.9386f_{17} \times 0.2252f_{16})$	20	0	0.0000
Car Evaluation	13	1	$(0.0254f_3 + 1.0307f_6) \times (1.2495f_4 - 0.3656f_2) + 0.1328f_5 - 0.7774f_1$	1003	725	0.9772
		2	$0.9312f_1 + 0.9311f_2 - 0.0450f_6 + 0.0038f_4 - 0.0432f_5$	124	601	1.2129
		3	$(0.4148f_3 + 0.9397f_2) \times (0.6600f_1 + 0.2111f_6 + 0.2247f_5 + 0.0188f_4)$	150	451	1.4296
		4	$\frac{0.2071f_2 + 0.3115f_1}{0.0026f_3 + 1.8128f_5} - 1.1642f_6$	70	381	0.9502
		5	$(0.3339f_2 + 0.9655f_1) \times 0.1604f_6 \times 0.4692f_5 + 0.4905f_3 \times 0.0016f_4$	75	306	1.3791
		6	$\frac{(1.6270f_2 - 0.5243f_6) \times 0.2967f_3}{1.1159f_5} - 0.4037f_4 - 0.0001f_1$	39	267	1.6433
		7	$\left((0.8728f_2 \times 0.5461f_5) - 0.5702f_6 - 1.8730f_1\right) \times (0.8159f_3 + 0.2953f_4)$	27	240	0.7261
		8	$\frac{0.5375f_5 + 1.4298f_3 + 0.1258f_6 - 0.1129f_2}{0.9633f_1 + 4.7259f_4}$	31	209	0.8590
		9	$\frac{0.6127f_1 - 0.1598f_3 - 0.1375f_5 + 0.6102f_2}{0.0141f_6 \times 1.1527f_4}$	50	159	1.6065
		10	$0.0301f_4 - 0.0049f_3 - 1.1130f_6 - 1.0766f_5 + 0.0972f_2 \times 0.7656f_1$	57	102	0.5555
		11	$\frac{0.6220f_6 - 0.8985f_5}{0.3331f_2 + 0.1390f_1} - 0.1048f_4 \times 0.1458f_3$	53	49	0.9945
		12	$0.1655f_3 + 2.8962f_6 + 0.9130f_5 + 0.1253f_4 - 0.0773f_1 + 0.0046f_2$	40	9	0.0578
		13	$(0.4115f_6 + 0.0820f_4) \times \frac{0.9167f_2 - 0.7246f_1}{0.4177f_3 - 0.2257f_5}$	9	0	0.0000
Iris	2	1	$0.2559f_3 + 0.0018f_4 - 0.6265f_1 - 0.5043f_2$	138	12	0.6973
		2	$1.2035f_2 + 0.1410f_3 - 0.4938f_4 + 0.4061f_1$	6	6	0.6741
Seeds	3	1	$\frac{0.1673f_6}{0.3916f_5} + (0.0619f_2 - 0.0489f_7) - (0.1594f_3 \times 1.4004f_4 \times 0.2024f_1)$	130	80	1.5480
		2	$\frac{0.0724f_2}{0.7171f_5} \times \frac{0.0080f_3}{0.0139f_6} - \frac{1.6947f_7}{0.3063f_4} - 0.0478f_1$	42	38	1.0749
		3	$\frac{\frac{0.8065f_5}{0.4834f_6} + 0.7505f_1 \times 0.3210f_4}{\frac{1.6162f_3}{0.0911f_2} + 1.3856f_7}$	37	1	0.1038
User Knowledge Modeling	5	1	$0.6115f_4 + 0.1941f_1 - 0.1072f_2 + 0.1395f_3 + 1.3571f_5$	287	116	1.4598
		2	$0.0871f_2 + 0.0378f_3 + 0.1894f_4 + 0.0096f_1 + 0.6614f_5$	84	32	1.0208
		3	$\frac{0.0009f_2}{0.1408f_4 \times 0.3473f_5 \times (0.7899f_1 - 0.1434f_3)}$	13	19	0.7728
		4	$\frac{0.1034f_4 - 0.5578f_5}{10.7284f_2} + 0.0103f_1 - 0.2243f_3$	7	12	0.3946
		5	$(1.2328f_2 + 0.1893f_3 + 0.4515f_4) \times (0.0286f_1 + 1.1009f_5)$	12	0	0.0000
Wine	1	1	$\left(0.7565f_{13} + 25.5745f_2 + \frac{9.595f_8}{0.0365f_9}\right) - (0.1767f_{10} + 0.1541f_6) \times \frac{1.8678f_5}{0.392f_7} - 0.7482f_1 \times (2.4062f_4 - 13.9003f_{12} - 19.4186f_3)$	177	1	0.0387

7. Generalising ability test

Models are normally tested on two criteria. First is about their accuracy on the data on which they are trained (Section 6). Second is about their generalising ability i.e., their accuracy on the data on which they are not trained. This section deals with the second scenario. Experiments to evaluate the generalising ability of the hierarchical learning involve using two disjoint subsets of the dataset for the purpose of training and testing the model, respectively. The training sets are used to learn models, which are later tried on the test set. This division of training and test set is done entirely at random. Furthermore, to cross validate the results, the roles of training and test sets are also interchanged to remove any bias. In addition to this, it is done several times to produce average performance results. In this work this standard procedure is strictly followed to test the generalizing ability of hierarchical models. The datasets are randomly divided into training and test sets of equal size, their roles are also reversed, and this procedure is repeated for 30 independent random trials. The 10 datasets used in the above-described experiments were the same as those used in experiments described in section 6. Results of these experiments are reported in Table 3, in which, column 1 and 2 identify the dataset. Column 3 and column 4 list the best and the worst results, respectively, achieved among 30 random trials. Column 5 lists average results of 30 random trials. Column 6 states what percentage of results were 100% accurate classification among the 30 random trials. The average results reveal that the proposed method achieves more than 90% correct results on each dataset.

Table 3. The Classification Results

No.	Dataset	Best Results %	Worst Results %	Average Results %	%age of Accurate Results	Average Execution Time (Seconds)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1st	Acute Inflammations Nephritis	100.00	98.33	99.83	90.00	0.00
2nd	Acute Inflammations Urinary	100.00	94.17	99.36	76.67	0.00
3rd	Balance Scale	100.00	97.92	99.47	16.67	0.72
4th	Banknote Authentication	100.00	98.25	99.60	3.33	2.75
5th	Breast Cancer Wisconsin (Diag.)	96.66	92.62	95.00	0.00	141.76
6th	Car Evaluation	94.85	91.09	92.88	0.00	83.24
7th	Iris	96.67	91.33	93.87	0.00	0.11
8th	Seeds	93.33	85.71	90.17	0.00	1.31
9th	User Knowledge Modelling	95.53	87.59	92.69	0.00	1.99
10th	Wine	95.51	83.71	90.82	0.00	2.67
	Average	97.25	92.07	95.37	18.67	23.45

8. Comparison with contemporary methods

We compare the proposed method with state-of-the-art methods for which the results on all the above datasets have been reported. Unfortunately, there is none that has been tried on all the ten datasets. However, one recently published paper about Random Forest [36] has tested five popular ensemble methods on nine of the ten datasets (except Breast cancer Diagnostic). We also found another paper about classification trees [37], which has been tested on nine out of ten datasets (except user knowledge modelling). Table 4 presents the comparison of the results of proposed method with those of methods. Column 1 and 2 of the table identify the dataset, columns 3-9 state average results of the seven methods (as referred in the table) on ten datasets in terms of number of correct predictions in percentage. At the bottom of the table there are three additional rows i.e., rows (11-13). Row 11 provides overall average of all the ten datasets. Number of random trials is mentioned in row 12 and finally row 13 reveals size of the training set as the percentage of the size of the complete dataset. The results show that the hierarchical learning achieves overall average of 95.37% against the best average of other schemes 94.68%. Furthermore, the hierarchical learning was trained on a much smaller proportion of the datasets, i.e., 50%, as compared to the competing methods trained on 75%-90% of the datasets. In addition to this, computational time of other schemes range within 5-15 minutes against average of less than one minute in case of hierarchical learning. Further to add, other schemes used much faster application-optimized cloud-computing environments as compared to personal laptop used to produce results presented here with the proposed method. Accounting all these facts, one may regard the results produced by the proposed method as competitive.

Table 4: Comparison with Literature

No.	Dataset	BA-C4.5	BA-CDT	RF	CRF	RCRF	OCT	Hierarchical Learning
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1st	Acute Inflam. Nephritis	100.00	100.00	100.00	100.00	100.00	100.0	99.83
2nd	Acute Inflam. Urinary	100.00	99.42	100.00	100.00	100.00	100.0	99.36
3rd	Balance Scale	81.56	82.41	80.30	81.94	82.76	87.6	99.47
4th	Banknote Authentication	98.95	98.77	99.34	99.31	99.26	98.7	99.60
5th	Breast Cancer (Diagnostic)	-	-	-	-	-	94.0	95.00
6th	Car Evaluation	94.33	93.55	94.70	94.44	93.3	87.5	92.88
7th	Iris	94.47	95.07	94.53	94.6	94.87	95.1	93.87
8th	Seeds	92.71	91.19	93.57	93.57	93.71	91.3	90.17
9th	User Knowledge Modelling	90.33	89.98	91.31	90.79	90.60	-	92.69
10th	Wine	95.34	95.84	97.74	97.51	97.51	91.6	90.82
11th	Overall Average	94.18	94.03	94.61	94.68	94.67	93.98	95.37
12th	No. of Random trials	10x10-fold x-validation					5x4-fold x-validation	30x2-fold x-validation
13th	Size of Training Set %					90.00	75.00	50.00

9. Conclusion and Future Work

This is sixth paper in the series of papers on the hierarchical learning. It theorises the method of hierarchical learning through 4 principles, i.e., principle of successive bifurcation, principle of two-tier discrimination, principle of class membership and the principle of selective data normalization. The first principle proposes that the multiple simpler models can emulate the effect of a more complex model, when put together hierarchically. The second principle separates the datapoints in terms of classes as well as domain. The third principle establishes class membership rule at different hierarchy levels. The last principle articulates the rules for the data normalisation. The presented method is not only supported on the mathematical grounds but also on the empirical results on ten popular real-world classification datasets taken from UCI repository. On these datasets accurate classification nonlinear discriminant models were produced, details of some of those models are given in section 6. The procedure to evaluate the accuracy of those models is also detailed. The generalising ability of the hierarchical method was also tested on the same datasets. The technique produced more than 95% correct results on average while trained on only 50% of the samples. Interestingly, the average of worst results in 30 random trials on all the datasets also turns out to be greater than 92%, which is a commendable result. The method performs competitively when compared with the results from other state of the art methods. Despite all the above success the technique still needs further theoretical enhancements for its wider applicability on the large spectrum of datasets, which are currently under investigation.

Appendix

Maximal Spread Theorem

The maximal spread theorem calculates the value of the range parameter for the maximum possible spread of data points in terms of the size of the dataset, i.e., the number of data points present in the dataset under investigation.

Theorem Statement: The range parameter for estimating minima/maxima of a set of points with maximal spread is equal to $\sqrt{n-1}$.

Please refer to equation 6 where range parameter is used for estimation of minima and maxima of the set of points. In statistics this parameter is roughly taken as 3. However, this value is suitable for larger datasets. In the theory of probabilistic hierarchical learning, training set is continually bifurcated in each hierarchy. Therefore, farther the hierarchy level smaller the dataset. Therefore, it was thought necessary that this parameter should be the function of the size of the subset. Furthermore, as we move to the higher levels of hierarchies the subset not only become smaller but also their sample spread becomes larger, as remaining samples are only those which failed to fit in the earlier models. This necessitates to estimate maximum possible value of this parameter to capture the structure of the set of points. For solution purpose, let us take only minima part of equation 6, as shown in equation 11.

$$\mu_{min} = \mu_{mean} - \rho * \mu_{sd} \quad (11)$$

Rearranging the variables:

$$\rho = \frac{\mu_{mean} - \mu_{min}}{\mu_{sd}} \quad (12)$$

Applying limits: as minima approaches 0.

$$\lim_{\mu_{min} \to 0} \rho = \frac{\mu_{mean}}{\mu_{sd}} \quad (13)$$

The samples of the set will be maximally spread when one of the samples is closest to the mean while rest of the samples are at farthest point from the mean. Let us assume that maximum distance between the points is a unity. In the case of equation 13 since minima lies at 0, therefore, maxima should be at 1. To further follow our assumption of maximal spread let us consider one sample lies at minima 0 and n-1 samples lie at maxima 1. Therefore, by substituting these values in equation 5, the mean of the point set can simply be calculated, as shown in equation 14.

$$\mu_{mean} = \frac{n-1}{n} \quad (14)$$

We can compute standard deviation by substituting the value of mean from equation 13 in equation 7, as shown in equation 15.

$$\mu_{sd} = \sqrt{\frac{\left(\frac{n-1}{n}\right)^2 + (n-1) \times \left(\frac{1}{n}\right)^2}{n}} = \frac{\sqrt{n-1}}{n} \quad (15)$$

Substituting the values of the mean (equation 14) and the standard deviation (equation 15) in equation 12, we get the value of the range parameter as shown in equation (16).

$$\rho = \sqrt{n-1} \quad (16)$$

Equation 16 proves the theorem statement.

References

[1] Maria Pérez-Ortiz, Silvia Jiménez-Fernández, Pedro A. Gutiérrez, Enrique Alexandre, César Hervás-Martínez *et al.*, "A Review of Classification Problems and Algorithms in Renewable Energy Applications", *Energies*, ISSN: 1996-1073, pp. 1-27, Vol. 9, No. 8, 2 August 2016, Published by Multidisciplinary Digital Publishing Institute (MDPI), DOI: 10.3390/en9080607, Available: https://www.mdpi.com/1996-1073/9/8/607.

[2] Jan Luts, Fabian Ojeda, Raf Van de Plas, Bart De Moor, Sabine Van Huffel *et al.*, "A Tutorial on Support Vector Machine-based Methods for Classification Problems in Chemometrics", *Analytica Chimica Acta*, Print ISSN: 0003-2670, Online ISSN: 1873-4324, pp. 129-145, Vol. 665, No. 2, 30 April 2010, Published by Elsevier, DOI: 10.1016/j.aca.2010.03.030, Available: https://www.sciencedirect.com/science/article/pii/S0003267010003132.

[3] Shan Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning", *ACM SIGMETRICS Performance Evaluation Review*, ISSN:0163-5999, pp. 70-73, Vol. 41, No. 4, March 2014, Published by ACM, DOI: 10.1145/2627534.2627557, Available: https://dl.acm.org/doi/10.1145/2627534.2627557.

[4] Mahdieh Labani, Parham Moradi, Fardin Ahmadizar and Mahdi Jalili, "A Novel Multivariate Filter Method for Feature Selection in Text Classification Problems", *Engineering Applications of Artificial Intelligence*, ISSN: 0952-1976, pp. 25-37, Vol. 70, 3 February 2018, Published by Elsevier, DOI: 10.1016/j.engappai.2017.12.014, Available: https://www.sciencedirect.com/science/article/pii/S0952197617303172.

[5] Yi Peng, Guoxun Wang, Gang Kou and Yong Shi, "An Empirical Study of Classification Algorithm Evaluation for Financial Risk Prediction", *Applied Soft Computing*, Print ISSN: 1055-6788, Online ISSN: 1029-4937, pp. 2906-2915, Vol. 11, No. 2, 1 January 2011, Published by Taylor & Francis, DOI: 10.1080/10556789808805680, Available: https://doi.org/10.1080/10556789808805680.

[6] Begüm D. Topçuoğlu, Nicholas A. Lesniak, Mack T. Ruffin IV, Jenna Wiens and Patrick D. Schloss, "A Framework for Effective Application of Machine Learning to Microbiome-based Classification Problems", *MBio*, Print ISSN: 2161-2129, Online ISSN: 2150-7511, pp. 1-13, Vol. 11, No. 3, 09 June 2020, Published by American Society for Microbiology, DOI: 10.1128/mBio.00434-20, Available: https://doi.org/10.1128/mBio.00434-20.

[7] Gerald E. Rehfeldt, Nicholas L. Crookston, Cuauhtémoc Sáenz-Romero and Elizabeth M. Campbell, "North American Vegetation Model for Land-Use Planning in a Changing Climate: A Solution to Large Classification Problems", *Ecological Applications*, Print ISSN: 2161-2129, Online ISSN: 2150-7511, pp. 119-141, Vol. 22, No. 1, January 2012, Published by Wiley, DOI: 10.1890/11-0495.1, Available: https://pubmed.ncbi.nlm.nih.gov/22471079.

[8] Fathima Fajila and Yuhanis Yusof, "Incremental Search for Informative Gene Selection in Cancer Classification", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-

029X, pp. 15-21, Vol. 5, No. 2, 1st April 2021, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2021.02.002, Available: http://aetic.theiaer.org/archive/v5/v5n2/p2.html.

[9] Ursani Ziauddin and Dicks Jo, "Introducing the Theory of Probabilistic Hierarchical Learning for Classification", *Lecture Notes in Computer Science*, Print ISBN: 978-3-030-22998-6, Online ISBN: 978-3-030-22999-3, pp. 628-641, vol: 11606, 15 June 2019, Published by Springer, DOI: 10.1007/978-3-030-22999-3_54, Available: https://link.springer.com/chapter/10.1007/978-3-030-22999-3_54.

[10] Ursani Ziauddin and David W. Corne, "Use of Reliability Engineering Concepts in Machine Learning for Classification", in *Proceedings of the 4th International Conference on Soft Computing & Machine Intelligence (ISCMI 2017)*, 23-24 Nov. 2017, Mauritius, Published by IEEE, eISBN: 978-1-5386-1314-6, DVD ISBN: 978-1-5386-1313-9, Print on Demand ISBN: 978-1-5386-1315-3, DOI: 10.1109/ISCMI.2017.8279593, Available: https://ieeexplore.ieee.org/document/8279593.

[11] Ziauddin Ursani and David W. Corne, "A Novel Nonlinear Discriminant Classifier Trained by an Evolutionary Algorithm", in *Proceedings of the 10th International Conference on Machine Learning and Computing (ICMLC 2018)*, pp. 336–340, February 26-28, 2018, China, Published by ACM, ISBN: 978-1-4503-6353-2, DOI: 10.1145/3195106.3195132, Available: https://doi.org/10.1145/3195106.3195132.

[12] Ziauddin Ursani and David W. Corne, "A Hierarchical Nonlinear Discriminant Classifier Trained through an Evolutionary Algorithm", in *Proceedings of the 3rd International Conference on Big Data, Cloud and Applications – BDCA18*, pp 273–288, Vol. 872, April 4-5, 2018, Kenitra, Morocco, Published by Springer, eISBN: 978-3-319-96292-4, DOI: 10.1109/ICAIBD.2018.8396159, Available: https://link.springer.com/chapter/10.1007/978-3-319-96292-4_22.

[13] Ziauddin Ursani and David W. Corne, "A Hierarchical Set-Partitioning Nonlinear Discriminant Classifier Trained by an Evolutionary Algorithm", in *Proceedings of the International Conference on Artificial Intelligence and Big Data (ICAIBD 2018)*, pp. 15-20, May 26-28, 2018, China, Published by IEEE, eISBN:978-1-5386-6987-7, USB ISBN:978-1-5386-6986-0, Print on Demand ISBN:978-1-5386-6988-4, DOI: 10.1109/ICAIBD.2018.8396159, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8396159.

[14] Roy Albert Crowson, *Classification and Biology*, 1st ed. New York, USA: Taylor and Francis, 1970, Available: https://doi.org/10.4324/9781315081090.

[15] Richard P. Stanley, "Acyclic Orientations of Graphs", *Discrete Mathematics*, pp. 171–178, Vol. 5, No. 2, 1973, Published by Elsevier, DOI:10.1016/0012-365X(73)90108-8, Available: https://doi.org/10.1016/0012-365X(73)90108-8.

[16] Carlos N. Silla Jr. and Alex A. Freitas, "A Survey of Hierarchical Classification across Different Application domains", *Data Mining and Knowledge Discovery*, pp. 31–72, Vol. 22, No. 1–2, 2011, DOI: 10.1007/s10618-010-0175-9, Available: https://doi.org/10.1007/s10618-010-0175-9.

[17] Yangchi Chen, Melba M Crawford and Joydeep Gosh, "Integrating Support Vector Machines in a Hierarchical Output Space Decomposition Framework", in *Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing*, vol: 2, pp. 949–952, 2004, Published by IEEE, Print ISBN:0-7803-8742-2, DOI: 10.1109/IGARSS.2004.1368565, Available: https://ieeexplore.ieee.org/document/1368565.

[18] Cinthia O. A. Freitas, Luiz S. Oliveira, Simone B. K. Aires and Flávio Bortolozzi, "Metaclasses and Zoning Mechanism Applied to Handwriting Recognition", *Journal of Universal Computer Science*, Print ISSN: 0948695X, Online ISSN: 09486968, pp. 211–223, Vol. 14, No. 2, Jan 1, 2008, Published by Technische Universitat Graz, Austria, DOI: 10.3217/jucs-014-02-0211, Available: https://lib.jucs.org/articles.php?id=28939.

[19] David Opitz and Richard Maclin, "Popular Ensemble Methods: An Empirical Study", *Journal of Artificial Intelligence Research*, ISSN: 1076-9757, pp. 169-198, Vol. 11, 1 August 1999, Published by Association for the Advancement of Artificial Intelligence, DOI: 10.1613/jair.614, Available: https://www.jair.org/index.php/jair/article/view/10239/24370.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, Deep learning, 1st ed. Cambridge, UK: MIT press, 2016, Available: https://www.deeplearningbook.org.

[21] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno and Tetsuya Ogata, "Audio-Visual Speech Recognition using Deep Learning", *Applied Intelligence*, Electronic ISSN: 1573-7497, Print ISSN: 0924-669X, pp. 722–737, Vol. 42, June 2015, Published by Springer, DOI: 10.1007/s10489-014-0629-7, Available: https://doi.org/10.1007/s10489-014-0629-7.

[22] Yi-Chung Chen and Jeen-Shing Wang, "A Hammerstein-Wiener Recurrent Neural Network with Frequency-Domain Eigensystem Realization Algorithm for Unknown System Identification", *Journal of Universal Computer Science*, ISSN: 0948695X, Vol. 15, No. 13, 2009, Published by Technische Universitat Graz, DOI: 10.3217/jucs-015-13-2547, Available: https://www.jucs.org/jucs_15_13/a_hammerstein_wiener_recurrent.html.

[23] Sreerama K. Murthy, "Automatic Construction of Decision Trees from Data: a Multi-Disciplinary Survey", *Data Mining and Knowledge Discovery*, Electronic ISSN: 1573-756X, Print ISSN: 1384-5810, Vol. 2, No. 4, 1998, pp. 345–

389, Published by Kluwer Academic Publishers, DOI: 10.1023/A:1009744630224, Available: https://link.springer.com/content/pdf/10.1023/A:1009744630224.pdf.

[24] Dewan Md. Farid, Li Zhang, Moiz Rahman Chowdhury, M. Alamgir Hossain and Rebecca Strachan, "Hybrid Decision Tree and Naïve Bayes Classifiers for Multi-Class Classification Tasks", *Expert Systems with Applications*, ISSN: 0957-4174, Vol. 41, 2014, pp. 1937–1946, DOI: 10.1016/j.eswa.2013.08.089, Available: https://doi.org/10.1016/j.eswa.2013.08.089.

[25] R. A. Fisher , "The Utilization of Multiple Measurements in Taxonomic Problems", *Annals of Eugenics*, pp. 179–188, Vol. 7, No. 2, September 1938, Print ISSN: 2050-1420, Online ISSN: 2050-1439, DOI: 10.1111/j.1469-1809.1936.tb02137.x, Available: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x.

[26] Michael L. Raymer, Travis. E. Doom, Leslie A. Kuhn and William F. Punch, "Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm", *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Print ISSN: 1083-4419, eISSN: 1941-0492, Vol. 33, No. 5, Oct 2003, DOI: 10.1109/TSMCB.2003.816922, Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1232717.

[27] Paul Adrien Maurice Dirac, "Bakerian Lecture: The Physical Interpretation of Quantum Mechanics", in *Proceedings of the Royal Society of London A: Mathematical and Physical Sciences*, 180 (980): pp. 1–39, 1942, Published by Royal Society, Print ISSN:0080-4630, Online ISSN:2053-9169, DOI: 10.1098/rspa.1942.0023, Available: https://doi.org/10.1098/rspa.1942.0023.

[28] Jacek Czerniak and Hubert Zarzycki, "Application of Rough Sets in the Presumptive Diagnosis of Urinary System Diseases", in *Proceedings of the 9th International Conference on Artificial Intelligence and Security in Computing Systems*, *ACS'2002*, October 23–25, 2002, Międzyzdroje, Poland, Print ISBN:978-1-4613-4847-4, Online ISBN:978-1-4419-9226-0, Published by Springer Boston MA, pp. 41-51, 2003, DOI: 10.1007/978-1-4419-9226-0_5, Available: https://link.springer.com/chapter/10.1007/978-1-4419-9226-0_5.

[29] Robert S. Siegler, "Three Aspects of Cognitive Development", *Cognitive Psychology*, Published by Elsevier, Vol. 8, No. 4, pp. 481-520, 1976, ISSN: 0010-0285, DOI: 10.1016/0010-0285(76)90016-5, Available: https://www.sciencedirect.com/science/article/abs/pii/0010028576900165.

[30] W. Nick Street, William H. Wolberg and Olvi L. Mangasarian, "Nuclear Feature Extraction for Breast Tumor diagnosis", *International Symposium on Electronic Imaging: Science and Technology*, IS&T/SPIE, 29 July 1993, vol: 1905, San Jose, CA, pp. 861-870, DOI: 10.1117/12.148698, Available: https://doi.org/10.1117/12.148698.

[31] Marko Bohanec and Vladislav Rajkovic, "Knowledge Acquisition and Explanation for Multi-Attribute Decision Making", in *Proceedings of the 8th International Workshop on Expert Systems and their Applications*, Avignon, France. Vol. 1, pp. 59-78, 1988, Available: https://kt.ijs.si/MarkoBohanec/pub/Avignon88.pdf.

[32] Edgar Anderson, "The Irises of the Gaspe Peninsula", *Bulletin of the American Iris Society*, Vol. 59, pp. 2–5, 1935, Available: https://cir.nii.ac.jp/crid/1571980073972926080.

[33] Małgorzata Charytanowicz, Jerzy Niewczas, Piotr Kulczycki, Piotr A. Kowalski, Szymon Łukasik *et al.*, "A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images", *Information Technologies in Biomedicine: Advances in Intelligent and Soft Computing*, Ewa Pietka, Jacek Kawa, Eds., Print ISSN: 978-3-642-13104-2, Online ISSN: 978-3-642-13105-9, pp. 15-24, vol. 69, 2010, Published by Springer-Verlag, Berlin-Heidelberg, DOI: 10.1007/978-3-642-13105-9_2, Available: https://link.springer.com/chapter/10.1007/978-3-642-13105-9_2.

[34] H. Tolga Kahraman, Seref Sagiroglu and Ilhami Colak, "Developing Intuitive knowledge Classifier and Modeling of Users' Domain Dependent Data in Web", *Knowledge Based Systems*, ISSN: 9507051, pp. 283-295, Vol. 37, 2013, Published by Elsevier, DOI: 10.1016/j.knosys.2012.08.009, Available: https://www.sciencedirect.com/science/article/abs/pii/S0950705112002225.

[35] Stefan Aeberhard, Danny Coomans and Olivierde Vel, "Comparative Analysis of Statistical Pattern Recognition Methods in High Dimensional Settings", *Pattern Recognition*, ISSN 0031-3203, pp. 1065-1077, Vol. 27, No. 8, August 1994, Published by Elsevier, DOI: 10.1016/0031-3203(94)90145-7, Available: https://www.sciencedirect.com/science/article/pii/0031320394901457.

[36] Joaquín Abellán, Carlos J. Mantas, Javier G. Castellano and Serafín Moral-García, "Increasing Diversity in Random Forest Learning Algorithm via Imprecise Probabilities", *Expert Systems With Applications*, ISSN: 9574174, pp. 228-243, Vol. 97, May 2018, Published by Elsevier, DOI: 10.1016/j.eswa.2017.12.029, Available: https://www.sciencedirect.com/science/article/pii/S0957417417308515.

[37] Dimitris Bertsimas and Jack Dunn, "Optimal Classification Trees", *Machine Learning*, ISSN: 0885-6125, pp. 1039-1082, Vol. 106, No. 7, 2017, Published by Springer, DOI: 10.1007/s10994-017-5633-9, Available: https://rdcu.be/cWVOF.