The Use of the Pattern Recognition and Classification Techniques within an Assisted Research System for the Vegetal Genetics

Nicolae Morariu

Sorin Vlad

“Ştefan cel Mare” University of Suceava,

Universităţii Street, no.9, 720225, tel:+4023021614,

Romania,

nmorariu@eed.usv.ro

sorinv@seap.usv.ro

Abstract: The paper presents the general architecture of a documentation and assisted research informatics system for vegetal genetics, dedicated to the educational and research institutions in this field. The main components are: the design and management of a data and knowledge base, database querying, statistical data processing, artificial intelligence applications. The plant description (the phenotype), the genetic hereditary heritage (the genotype) as well as the results of the experiments and researches performed in the vegetal field are fed into the system’s database within a predefined evolutionary self developing structure. Among the specific research problems in this field one can mention: population recognition and classification, plant disease diagnostication (barley species). The basis concept was inspired by the activity of Vegetal Gene Bank of Suceava.

Keywords: assisted research,, vegetal genetics, phenotype, genotype, data and knowledge base, evolutionary structure, artificial intelligence, pattern recognition.

Nicolae Morariu graduated from “Alexandru Ioan Cuza” University of Iaşi, Mathematics – Mechanics Faculty, section Computing Machines, 1972. He has obtained PhD degree from “Ştefan cel Mare” University of Suceava, 2004, PhD thesis: Contribution to the development of data and knowledge bases. His postgraduate activity includes: The design and implementation of the applications and informatics systems within the Regional Electronic Computing Center of Suceava and Informatics Services Society of Suceava (1972-1993), research projects within the national research programs (1993-2002), SSI Suceava design-research manager department (1998-2002), associate professor “Stefan cel Mare” University of Suceava, Electrical Engineering Faculty (1991-1998), He is lecturer at the Economic Sciences and Public Administration Faculty, “Stefan cel Mare” University of Suceava (2002-present). His research interests include Databases: FoxPro, Access, Oracle administration and SQL programming, Deductive databases. Artificial intelligence: expert systems, pattern recognition, neural networks, vegetal infogenetics.

Sorin Vlad graduated from “Ştefan cel Mare” University of Suceava, Electrical Engineering Faculty, section Computer and System’s Science, 1988. He is Phd Candidate in Computer Science, field of research – Chaotic systems behavior modeling, University “Ştefan cel Mare” of Suceava, since 2004. He serves as Assistant lecturer, University “Ştefan cel Mare” of Suceava, Economic Sciences and Public Administration, Informatics department. His research interests include: Artificial intelligence: neural networks, expert systems, Logic programming, Chaotic time series analysis and prediction.

1. General Description of the System

The data and knowledge that are to be stored and processed in a documentation and assisted research system for vegetal genetics, refers resources of the vegetal genetics representing all the vegetal life forms: wild plants, varieties and local populations, lines, hybrids, weeds, improved forms, etc., all of these being subject to the genetic erosion phenomena, pathogenic elements aggression (phytopatology) and environment factors. The preservation of the vegetal genetic resources, especially those of the endangered species and the identification of new species among the wild ones, susceptible of becoming new tilling plants, implies the collection, assessment and vegetal genetic preservation, a role played by international institutions such as: Gene Banks, The International Institute of Genetic Vegetal Resources (IPGRI) of Rome, etc.

The plant performed description and experiments generate two data categories: descriptive data and experimental data. The data describing genetic material (descriptive data) consist in knowledge which is represented in the database within a universal evolutionary data structure. Data can be mixed with images (plant, leaf, root, seed etc.), genetic maps. The main components of a knowledge-based system in vegetal genetics are: the design and management of a database and knowledge base, database querying, statistical data processing, artificial intelligence applications. The general architecture of the system is shown in the Figure 1. [1]

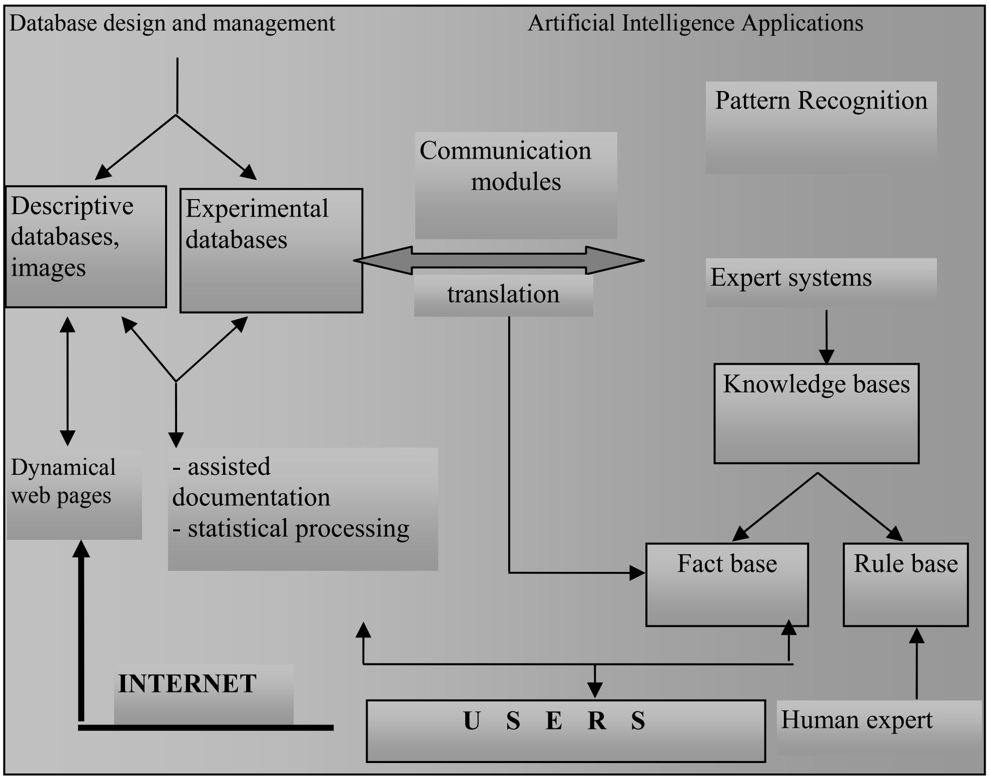


Figure 1: General system architecture

The presented model represents a tool allowing for the design of a complex informatics system using an evolutionary bottom – up strategy, the system's data being organized into a database which permanently extends as the system develops, say a dynamical database, the extension being created when new data is loaded into the database. When adopting this approach, the design of an informatics system starts from a predefined database with a universal structure which accumulates significance during the design of the system by means of the loading – updating operation of the databases with general use programs. Within the model, the attributes are not included in the entity definition described as it's usually the case with any of the data models [2] lying at the basis of the DBMS system, being treated as a distinct block, hence the maximum flexibility in the definition of the data structure, and eases the approach of the problems where the data being mostly knowledge because it allows for the description of data semantics.

2. System's Database

As a result of the collection, assessment and preservation activities in the gene bank passport data, assessment data and preservation data are obtained. The passport and preservation data (the warehouse record) are common to all the species but the assessment and characterization differs from one species to another and results from experimental measurements and laboratory analysis. The collecting and storing activities are described by the same attributes for all the species, and the characterization and assessment are described by different attributes from one species to another.

Passport data: Input number, Input name, taxonomycal classification (Genre, Species, Subspecies, Variety), Origin (Town, Department (area), Country), Collection date, Collection Source, Geographical Data (Altitude, Latitute, Longitude).

Preservation Data: Storing Code, Germination, Seed reserve, Humidity, 1000 berries weight.

Assessment data: (example: zea mays): assessment place (Country, Research Center, The person who made the assessment), Plant data (Total plant height, Stipes medium diameter, Total leaf number), Cob measurements (Length, Base diameter, Number of berries rows, Cob's weight).

Each variety is defined by values corresponding to the attributes of the species the variety belongs to. For each of the three data types some examples are presented, mentioning that the evolutionary data structure automatically extends due to the database loading operation.

The importance of the model can be justified if we consider a data structure that implies the definition of a large number of entities, each entity being described by a large number of different attributes, as an example we can mention here the genetic material description (the description of a relatively large number of breeds, each of them being characterized by a different evaluation descriptor set, to be extended subsequently).

In the relational data model, an extensive simplified database structure is defined as follows:
SPECIES (Code_s, Name_s) – the plant species catalogue
DESCRIPTORS (Code_d, Name_d, Type_d) – the attributes list
D_SPECIES (Code_s, Code_d) – species definitions
INPUTS (Code_s, Code_i, Name_i) – the input species list (varieties)
D_INPUTS (Code_i, Code_d, Val_d) – inputs definition (varieties)
A_GEN (Son_Code, Parent_Code) – inputs genealogy (genealogical three)
Where R1 – R7 relations are implicitly defined through key propagation.

The features of an organism defines the phenotype, (each of these features being established by certain genes) but while the phenotype modifies during the life of the organism the genotype is relatively constant (an organism has the same genes all of its life). In the database structure of a knowledge based vegetal genetics system are represented the two main concepts: genotypology and phenotypology.

Data structure for the phenotype description was previously presented. Regarding the genotype, several ways of representation have been used, such as: genes dictionary (tabular description), the genetic map (images), the symbolical representation of the genetic code. The symbolical representation of the genetic code is carried out [3] inside of a system resembling the Morse code, starting from an alphabet defined by the following symbols: A(adenine), G (guanine), C (cytosine) or U (uracyl) for riboviruses representing the four nitrate bases composing the DNA macromolecule (A, G, C, T) or the RNA macromolecule (A, G, C, U). The combination of the symbols forms the codons (corresponding to the words in the Morse code) and the associations of codons establish the genes (the phrases in the Morse code). The database diagram is shown in Figure 2.

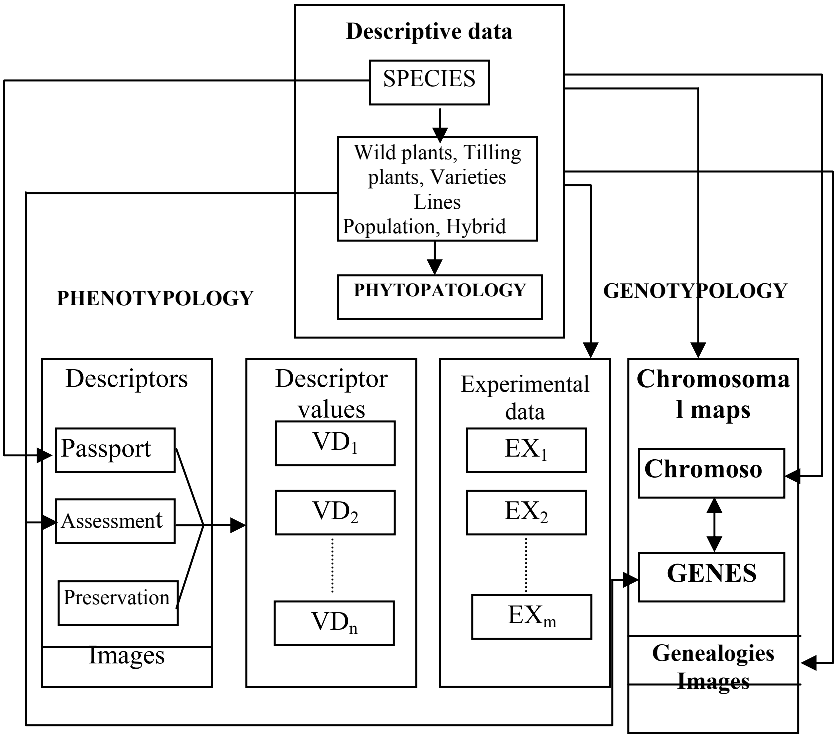


Figure 2: The database diagram

For the documentation regarding the genetic material the following requests can be formulated to query the database, for example: plant species catalogue, specified species description, input list corresponding to certain species, documentary specified input report, the input list satisfying querying specified conditions, the input genealogy (genealogical tree), species genotypology.

3. The Specific Research Problem

The problems that occur in this research field, debated inside the documentation and vegetal genetic assisted research system are: population recognition and classification, the establishment of the optimum size of the reproduction population for the preservation of a witness variety (for zea mays species), plant disease diagnostication (barley species). The classification and recognition are solved using artificial intelligence specific methods (pattern recognition [4], neural networks [5], knowledge base and expert systems [6]) and optimization problems are solved by defining and using the regressional models.

Taking into account that for corn unlike other tilling plants, there is a crossed pollination and therefore the genetic erosion process is much more emphasized, some experiments and assessments were performed at the Agricultural Research Center of Suceava for some corn varieties – Hângănesc, Cincantin, Suceava 1 – resulting an important data set corresponding to the measurements taken for 30 descriptors (assessment data) for 20 multiplication alternatives. Each of multiplication alternatives used inside the performed experiments is a form. Using pattern recognition techniques outputs in different representations are obtained: list, graph describing the affiliation to a class, grouping into classes. The determination of the optimum alternative multiplication for the preservation of the witness variety features was performed by building regressional models.

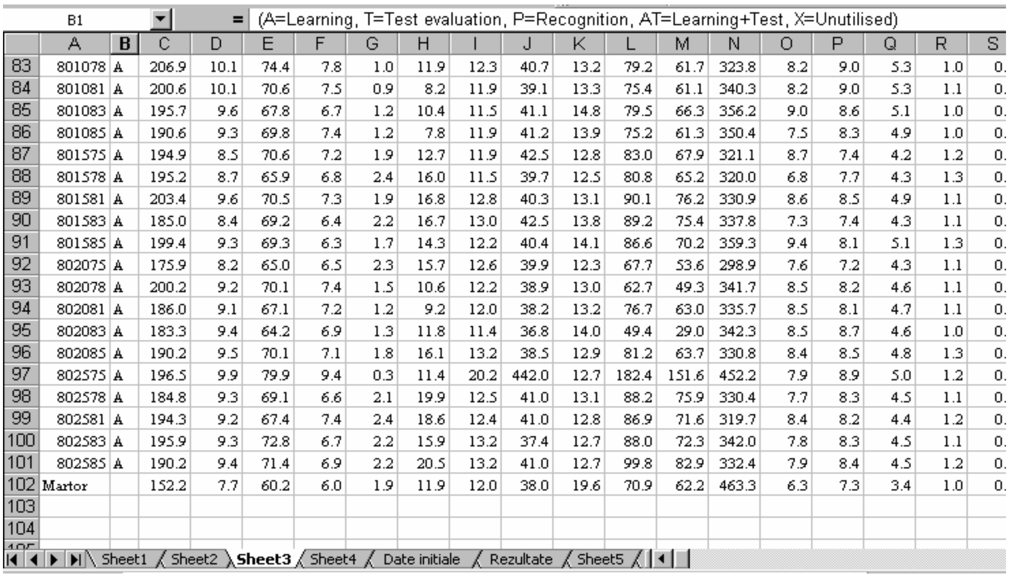
For the diagnostication of barley diseases, neural networks were defined and trained that use the data from experiments performed within the European project EU – GENRES CT98 – 104 “Evaluation and Conservation of Barley Genetic Resource to Improve Their Accessibility to Breeders in Europe” [4] supervised by IPK Genbank Gatersleben from Germany. The Vegetal Gene Bank of Suceava took part in this project as partner during 2000-2001 to determine the barley horizontal tolerance to diseases. The data and the obtained results were used in the present paper to design and train neural networks for barley diseases diagnostication (infection score), using our own software named REFORME [8] and the NeuroShell program for multilayer perceptron networks as well as MATLAB product for RBF networks.

4. The Classification and the Recognition of the Zea Mays Population. The Determination of the Multiplication Alternative Closest to the Witness Breed.

For the pattern classification and recognition the REFORME program is used [8]. The results obtained using the pattern recognition module for the Hângănesc breed are presented.

Each form is a row in a Excel spreadsheet as presented in Figure 3.

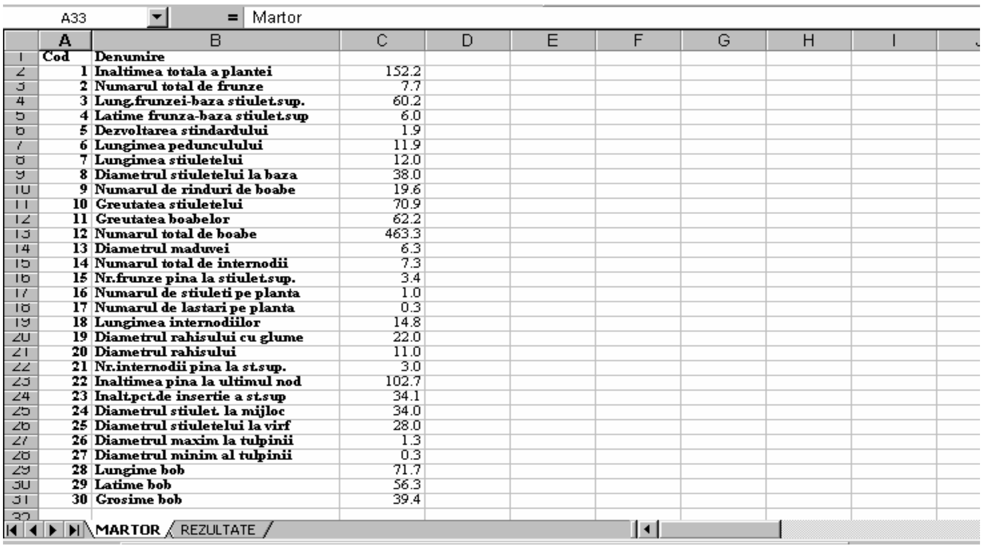
The description of the witness breed is presented in Figure 4.



B1 = (A=Learning, T=Test evaluation, P=Recognition, AT=Learning+Test, X=Unutilised)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
83	801078	A	206.9	10.1	74.4	7.8	1.0	11.9	12.3	40.7	13.2	79.2	61.7	323.8	8.2	9.0	5.3	1.0	0.
84	801081	A	200.6	10.1	70.6	7.5	0.9	8.2	11.9	39.1	13.3	75.4	61.1	340.3	8.2	9.0	5.3	1.1	0.
85	801083	A	195.7	9.6	67.8	6.7	1.2	10.4	11.5	41.1	14.8	79.5	66.3	356.2	9.0	8.6	5.1	1.0	0.
86	801085	A	190.6	9.3	69.8	7.4	1.2	7.8	11.9	41.2	13.9	75.2	61.3	350.4	7.5	8.3	4.9	1.0	0.
87	801575	A	194.9	8.5	70.6	7.2	1.9	12.7	11.9	42.5	12.8	83.0	67.9	321.1	8.7	7.4	4.2	1.2	0.
88	801578	A	195.2	8.7	65.9	6.8	2.4	16.0	11.5	39.7	12.5	80.8	65.2	320.0	6.8	7.7	4.3	1.3	0.
89	801581	A	203.4	9.6	70.5	7.3	1.9	16.8	12.8	40.3	13.1	90.1	76.2	330.9	8.6	8.5	4.9	1.1	0.
90	801583	A	185.0	8.4	69.2	6.4	2.2	16.7	13.0	42.5	13.8	89.2	75.4	337.8	7.3	7.4	4.3	1.1	0.
91	801585	A	199.4	9.3	69.3	6.3	1.7	14.3	12.2	40.4	14.1	86.6	70.2	359.3	9.4	8.1	5.1	1.3	0.
92	802075	A	175.9	8.2	65.0	6.5	2.3	15.7	12.6	39.9	12.3	67.7	53.6	298.9	7.6	7.2	4.3	1.1	0.
93	802078	A	200.2	9.2	70.1	7.4	1.5	10.6	12.2	38.9	13.0	62.7	49.3	341.7	8.5	8.2	4.6	1.1	0.
94	802081	A	186.0	9.1	67.1	7.2	1.2	9.2	12.0	38.2	13.2	76.7	63.0	335.7	8.5	8.1	4.7	1.1	0.
95	802083	A	183.3	9.4	64.2	6.9	1.3	11.8	11.4	36.8	14.0	49.4	29.0	342.3	8.5	8.7	4.6	1.0	0.
96	802085	A	190.2	9.5	70.1	7.1	1.8	16.1	13.2	38.5	12.9	81.2	63.7	330.8	8.4	8.5	4.8	1.3	0.
97	802575	A	196.5	9.9	79.9	9.4	0.3	11.4	20.2	442.0	12.7	182.4	151.6	452.2	7.9	8.9	5.0	1.2	0.
98	802578	A	184.8	9.3	69.1	6.6	2.1	19.9	12.5	41.0	13.1	88.2	75.9	330.4	7.7	8.3	4.5	1.1	0.
99	802581	A	194.3	9.2	67.4	7.4	2.4	18.6	12.4	41.0	12.8	86.9	71.6	319.7	8.4	8.2	4.4	1.2	0.
100	802583	A	195.9	9.3	72.8	6.7	2.2	15.9	13.2	37.4	12.7	88.0	72.3	342.0	7.8	8.3	4.5	1.1	0.
101	802585	A	190.2	9.4	71.4	6.9	2.2	20.5	13.2	41.0	12.7	99.8	82.9	332.4	7.9	8.4	4.5	1.2	0.
102	Martor		152.2	7.7	60.2	6.0	1.9	11.9	12.0	38.0	19.6	70.9	62.2	463.3	6.3	7.3	3.4	1.0	0.
103																			
104																			

Figure :3: Initial data (1 witness, 100 multiplication alternatives

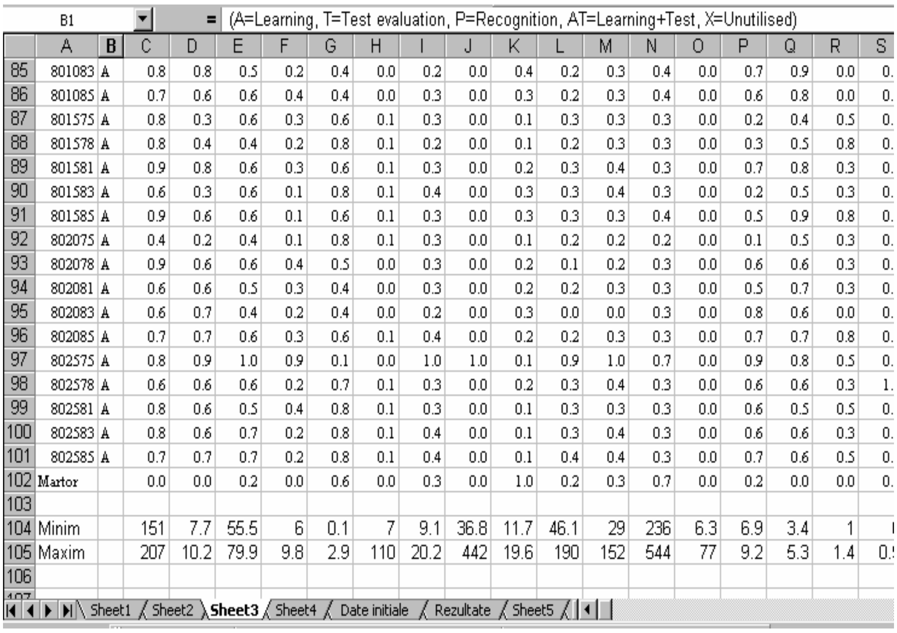


A33 = Martor

	A	B	C	D	E	F	G	H	I
1	Cod	Denumire							
2	1	Inaltimea totala a plantei	152.2						
3	2	Numarul total de frunze	7.7						
4	3	Lung.frunzei-baza stiulet.sup.	60.2						
5	4	Latime frunza-baza stiulet.sup	6.0						
6	5	Dezvoltarea stindardului	1.9						
7	6	Lungimea pedunculului	11.9						
8	7	Lungimea stiuletelui	12.0						
9	8	Diametrul stiuletelui la baza	38.0						
10	9	Numarul de rinduri de boabe	19.6						
11	10	Greutatea stiuletelui	70.9						
12	11	Greutatea boabelor	62.2						
13	12	Numarul total de boabe	463.3						
14	13	Diametrul maduvei	6.3						
15	14	Numarul total de internodii	7.3						
16	15	Nr.frunze pina la stiulet.sup.	3.4						
17	16	Numarul de stiuleti pe planta	1.0						
18	17	Numarul de lastari pe planta	0.3						
19	18	Lungimea internodiilor	14.8						
20	19	Diametrul rahisului cu glume	22.0						
21	20	Diametrul rahisului	11.0						
22	21	Nr.internodii pina la st.sup.	3.0						
23	22	Inaltimea pina la ultimul nod	102.7						
24	23	Inalt.pct.de insertie a st.sup	34.1						
25	24	Diametrul stiulet. la mijloc	34.0						
26	25	Diametrul stiuletelui la virf	28.0						
27	26	Diametrul maxim la tulpinii	1.3						
28	27	Diametrul minim al tulpinii	0.3						
29	28	Lungime bob	71.7						
30	29	Latime bob	56.3						
31	30	Grosime bob	39.4						

Figure 4: The description of the witness breed

The input data are normalized using the domain adjustment method, the result is illustrated in Figure 5.



B1 = (A=Learning, T=Test evaluation, P=Recognition, AT=Learning+Test, X=Unutilised)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
85	801083	A	0.8	0.8	0.5	0.2	0.4	0.0	0.2	0.0	0.4	0.2	0.3	0.4	0.0	0.7	0.9	0.0	0.
86	801085	A	0.7	0.6	0.6	0.4	0.4	0.0	0.3	0.0	0.3	0.2	0.3	0.4	0.0	0.6	0.8	0.0	0.
87	801575	A	0.8	0.3	0.6	0.3	0.6	0.1	0.3	0.0	0.1	0.3	0.3	0.3	0.0	0.2	0.4	0.5	0.
88	801578	A	0.8	0.4	0.4	0.2	0.8	0.1	0.2	0.0	0.1	0.2	0.3	0.3	0.0	0.3	0.5	0.8	0.
89	801581	A	0.9	0.8	0.6	0.3	0.6	0.1	0.3	0.0	0.2	0.3	0.4	0.3	0.0	0.7	0.8	0.3	0.
90	801583	A	0.6	0.3	0.6	0.1	0.8	0.1	0.4	0.0	0.3	0.3	0.4	0.3	0.0	0.2	0.5	0.3	0.
91	801585	A	0.9	0.6	0.6	0.1	0.6	0.1	0.3	0.0	0.3	0.3	0.3	0.4	0.0	0.5	0.9	0.8	0.
92	802075	A	0.4	0.2	0.4	0.1	0.8	0.1	0.3	0.0	0.1	0.2	0.2	0.2	0.0	0.1	0.5	0.3	0.
93	802078	A	0.9	0.6	0.6	0.4	0.5	0.0	0.3	0.0	0.2	0.1	0.2	0.3	0.0	0.6	0.6	0.3	0.
94	802081	A	0.6	0.6	0.5	0.3	0.4	0.0	0.3	0.0	0.2	0.2	0.3	0.3	0.0	0.5	0.7	0.3	0.
95	802083	A	0.6	0.7	0.4	0.2	0.4	0.0	0.2	0.0	0.3	0.0	0.0	0.3	0.0	0.8	0.6	0.0	0.
96	802085	A	0.7	0.7	0.6	0.3	0.6	0.1	0.4	0.0	0.2	0.2	0.3	0.3	0.0	0.7	0.7	0.8	0.
97	802575	A	0.8	0.9	1.0	0.9	0.1	0.0	1.0	1.0	0.1	0.9	1.0	0.7	0.0	0.9	0.8	0.5	0.
98	802578	A	0.6	0.6	0.6	0.2	0.7	0.1	0.3	0.0	0.2	0.3	0.4	0.3	0.0	0.6	0.6	0.3	1.
99	802581	A	0.8	0.6	0.5	0.4	0.8	0.1	0.3	0.0	0.1	0.3	0.3	0.3	0.0	0.6	0.5	0.5	0.
100	802583	A	0.8	0.6	0.7	0.2	0.8	0.1	0.4	0.0	0.1	0.3	0.4	0.3	0.0	0.6	0.6	0.3	0.
101	802585	A	0.7	0.7	0.7	0.2	0.8	0.1	0.4	0.0	0.1	0.4	0.4	0.3	0.0	0.7	0.6	0.5	0.
102	Martor		0.0	0.0	0.2	0.0	0.6	0.0	0.3	0.0	1.0	0.2	0.3	0.7	0.0	0.2	0.0	0.0	0.
103																			
104	Minim		151	7.7	55.5	6	0.1	7	9.1	36.8	11.7	46.1	29	236	6.3	6.9	3.4	1	[illegible]
105	Maxim		207	10.2	79.9	9.8	2.9	110	20.2	442	19.6	190	152	544	77	9.2	5.3	1.4	0.
106																			

Figure 5: Input normalization result

Figure 6 shows the result of the unsupervised classification using threshold algorithm for the threshold = 1.5.

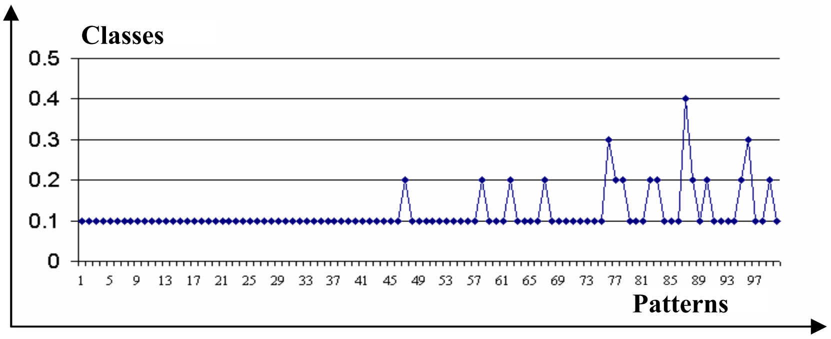
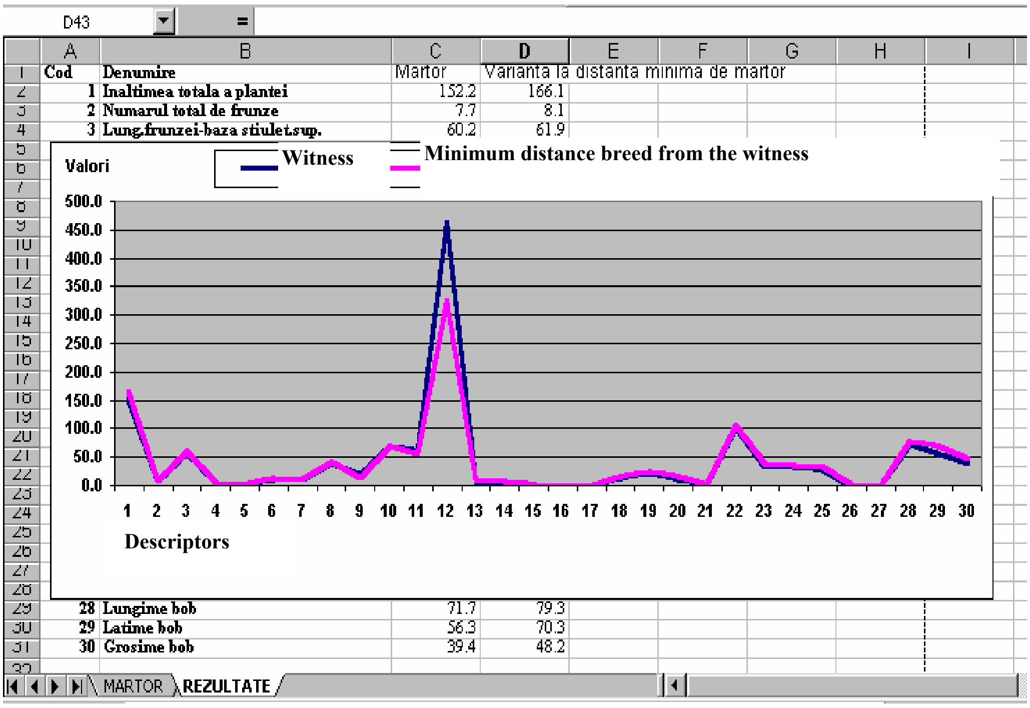


Figure 6: Unsupervised classification using threshold algorithm with Euclidian distance

The forms (100) were grouped into 4 classes, 85 percent of them being included in 0.1 class.

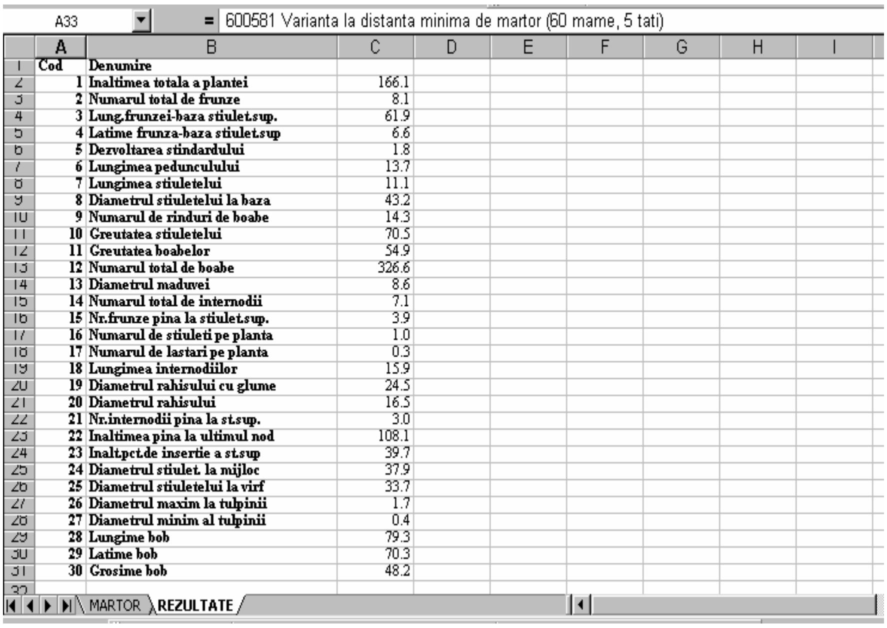
To determine the closest class to the witness breed using the nearest neighbor rule is obtained the class 0.1 and the variant closest to the witness breed is showed in Figure 7.



Cod	Denumire	Martor	Varianta la distanta minima de martor
1	Inaltimea totala a plantei	152.2	166.1
2	Numarul total de frunze	7.7	8.1
3	Lung.frunzei-baza stiulet.sup.	60.2	61.9
28	Lungime bob	71.7	79.3
29	Latime bob	56.3	70.3
30	Grosime bob	39.4	48.2

Figure 7: The breed at the minimum distance from the witness

Figure 8 shows the witness breed and the minimum distance breed from the witness vs the 30 descriptors values.



600581 Varianta la distanta minima de martor (60 mame, 5 tati)

Cod	Denumire	
1	Inaltimea totala a plantei	166.1
2	Numarul total de frunze	8.1
3	Lung.frunzei-baza stiulet.sup.	61.9
4	Latime frunza-baza stiulet.sup	6.6
5	Dezvoltarea stindardului	1.8
6	Lungimea pedunculului	13.7
7	Lungimea stiuletelui	11.1
8	Diametrul stiuletelui la baza	43.2
9	Numarul de rinduri de boabe	14.3
10	Greutatea stiuletelui	70.5
11	Greutatea boabelor	54.9
12	Numarul total de boabe	326.6
13	Diametrul maduvei	8.6
14	Numarul total de internodii	7.1
15	Nr.frunze pina la stiulet.sup.	3.9
16	Numarul de stiuleti pe planta	1.0
17	Numarul de lastari pe planta	0.3
18	Lungimea internodiilor	15.9
19	Diametrul rahisului cu glume	24.5
20	Diametrul rahisului	16.5
21	Nr.internodii pina la st.sup.	3.0
22	Inaltimea pina la ultimul nod	108.1
23	Inalt.pct.de insertie a st.sup	39.7
24	Diametrul stiulet. la mijloc	37.9
25	Diametrul stiuletelui la virf	33.7
26	Diametrul maxim la tulpinii	1.7
27	Diametrul minim al tulpinii	0.4
28	Lungime bob	79.3
29	Latime bob	70.3
30	Grosime bob	48.2

Figure 8: The witness breed and the breed closest to the witness representation

REFERENCES

1. NICOLAE MORARIU, VLAD SORIN, VANCEA ROMULUS, **Sistem bazat pe cunoştinţe destinat cercetării asistate în genetica vegetală**, Economia românească – prezent şi perspective, Univ. "Ştefan cel Mare" Suceava, 2003.

2. NICOLAE MORARIU, SORIN VLAD, **Online documentation and assisted research knowledge based system for vegetal genetics resources**, Proceedings of the Forth International Conference "Internet Education Science IES-2004", Baku State University Azerbaijan,Vinnytsia National Technical University Ukraine, St. Cyril and St. Methodius University of Veliko Turnovo Bulgaria, oct. 2004 Vinnytsia Ukraine, ISBN 966-641-104-0, Tom 2, pp. 513-516.

3. NICOLAE MORARIU, **Artificial Inteligence Techniques in an Evaluation and Decision System for Economic Activity**, SACI 2004, Budapest Polytechnic Nepszinhaz, Budapest, Hungary, Timişoara, 2004.

4. ROMUL VANCEA, ŞTEFAN HOLBAN, DAN CIUBOTARIU, **Recunoaşterea Formelor. Aplicaţii**, Ed. Academiei R.S.R., 1989.

5. D. DUMITRESCU, HARITON COSTIN, **Inteligenţa Artificială. Reţele neuronale teorie şi aplicaţii**, Ed. Teora, Buc. 1996.

6. T. CRĂCIUN, I. TOMOZEI, N. COLEŞ, GALIA BUTNARU, **GENETICĂ VEGETALĂ**, Ed. Didactică şi Pedagogică, 1991.

7. **Evaluation and Conservation of Barley Genetic Resource to Improve Their Accessibility to Breeders in Europe**. Evaluation methods, EU Project GENRES CT98-104, Genbank, IPK, Gatersleben, Germany, 2001 http//barley.ipk-gatersleben.de

8. NICOLAE MORARIU, **REFORME – A software product for pattern clasification and recognition by joint use of pattern recognition techniques and multi-layer perceptron**, The Proceedings of the Central and East European Conference in Business Information Systems, Cluj-Napoca, May 2004, Ed. Risoprint, ISBN 973-656-648-X, pp. 100-105.