Prediction System of Larynx Cancer

Benjamín Moreno-Montiel and Carlos Hiram Moreno-Montiel
Posgrado Ciencias y Tecnologías de la Información
Universidad Autónoma Metropolitana - Iztapalapa,
México D.F., México
opelo1209@yahoo.com, hiramoreno@gmail.com

Abstract—In the task of data classification, there exist many uses and applications such as Credit assignment, Business, games Development, gene Research in public health problems, among others. In this research there is a large collection of data for treatment and prevention of some diseases, the most complex is the study of Cancer. The databases there have provided valuable knowledge useful for study of this disease that in many cases is unknown. An example of these databases is at the Centro Medico Nacional Siglo XXI, with information of human laryngeal carcinoma (LaCa). In this paper, we propose a Prediction System of Larynx Cancer (PSLC) to apply the task of classification of this type of databases to obtain novel knowledge for LaCa. The prediction system has two components, one component is the transformation and selection of data, the second component is a set of classifiers to obtain the prediction of life of sample patients with this type of cancer. With this prediction system, we found that, when there is an increase in CRBP-1 gene, it was correlated with patient survival; this allowed us to implement a Hybrid Classifier of Decision Rules (HCDR). The HCDR obtained the highest predictive value using genes CRBP-1 and provided a better degree of accuracy, with more than 90%, in comparison with different classifiers, indicating that the PSLC has a high degree of reliability.

Keywords-Data Minning; Classification; Classifier; Biochip genetic; Larynx Cancer.

I. INTRODUCTION

One of the most recent advances in genomic science is the search of genes responsible for alterations in different organisms, generating new research, in areas such as Agriculture, Medicine, Biomedical Sciences, among others. Some techniques, such as comparative genomic hybridization [6] and immunohistochemical assays on tissues [14] can determine the possible origin of different diseases such as cancer. By exhaustive search techniques and use of decision trees [2], there have been developed systems that make use of these techniques, finding relationship between genetic patterns of patients with some types of cancer, to provide prevention and early treatment.

An example of these systems is OncoTree [11] (for renal carcinoma progression), which finds relationship between genetic patterns of Renal Cancer using decision trees. Although having good results on this type of cancer, this system has two limitations: on the one hand, it can only do an analysis of renal cancer, and, on the other hand, limiting the use of this system is reserved only for some businesses due to their high prices.

There exist a small number of these and other predictive systems in only some countries; with high prices for use and analysis of results, limiting research in this area worldwide, this is a main problem with cancer prediction systems.

Because of this problem in some institutions [4], studies have been generated on different types of cancer, seeking to develop prediction systems like Onco Tree. An example of these studies was conducted in the Centro Medico Nacional Siglo XXI in Mexico City, which obtained a database (DB) with information about Larynx Cancer. This DB has the study of 21 patients with this type of cancer for over five years, which represents one of the largest repositories in Mexico.

The problem with this DB is that only specialists in the area of genetics have studies of how this may originate this type of cancer at the chromosomal level. In a study conducted with the Universidad Autónoma Metropolitana (UAM) and Instituto de Ciencias y Tecnologías de la Información (ICyTDF), we had access to this DB, for which we proposed as a way to find these patterns using Data Mining, such as those used in Onco Tree.

Data Mining [10] is the exploration and analysis to identify patterns non trivial (knowledge) within large amounts of data, which may be valid, novel, potentially, useful and understandable. With Data Mining we solve many tasks [13] as Prediction, Classification, Identification, Grouping and Association.

In this paper, we propose a Prediction System of Larynx Cancer (PSLC), with which we use the first and second Data Mining tasks. The prediction we use to find the correlations of the attributes within the DB's of Genetic Markers (dbGM). The task of classification is used to make the analysis and diagnosis of potential patients using the correlations found in the prediction task.

The PSCL is composed of two components; the first component is the transformation and selection of data, because dbGM has a special format called ISCN (An International System for Human Cytogenetic Nomenclature), which corresponds to all nomenclature that is handled in the research area of Human Genetics.

For the second component, we mix the prediction and classification tasks to get a prediction of laryngeal cancer in patients who are predisposed to have this type of cancer. At this stage we call the Engine of Operation (EM) of PSCL, which we use to obtain the experimental results of the classification of dbGM.

 ISBN: 978-1-61208-277-6

This paper is organized as follows. In Section 2, we will discuss previous work on prediction systems and the main algorithms used for constructing such systems. In Section 3, we describe how we build the PSLC. In Section 4, the tests that were performed to a DB of genetic biochips, will be discussed along with the results obtained using several Data Mining classifiers. Finally, we will present a conclusion and future work steps.

II. PREVIOUS WORK

A. Background of LaCa

Laryngeal cancer (LaCa) represents an important public health problem mainly affecting people over fifty years worldwide. Several methods are used for treatment and prevention. Comparative genomic hybridization (CGH) has been widely used in cancer research [8], [10], [14] the detection of specific patterns of chromosomal imbalances, for example, loss of chromosome 13 in all carcinomas cell carcinomas of the larynx [9], [10].

However, in the spatial resolution of CGH not much is known about the identity of specific genes that could be targeted chromosomal regional imbalances. The CGH array solves this problem by increasing the sensitivity for detection of changes in DNA copy number in specific loci (which are the specific location of a gene or DNA sequence on a chromosome) through the use of genomic DNA fragments is defined by a mapping location.

It is known that these arrays extend over a solid surface, resulting in a resolution of the imbalances in a number of copies of a single gene level [12]. To refine patterns of chromosomal imbalances present in squamous cell carcinoma of the larynx, and especially to identify the specific genes that could target the of copy number changes in this tumor type, array CGH is applied with oncogenes, tumor suppressor genes or some other genes associated with cancer.

This is achieved by determining the relation of level chromosomal patterns of samples obtained using different classification models in data mining, to determine which genes are involved, below some methods are described for classifying data. In the next section we review some data mining classifiers.

B. Several Methods for classification

In the literature, there have been a large number of classifiers that allow us to differentiate a set of samples according to the category or as usually called the class to which they belong.

There is a large number of classifiers [5] for instance the classifiers based on decision trees as adultery, C4.5 and ID3, classifiers based on decision rules as Decision Tables and Decision List, ensemble-based systems such as Bagging and Boosting, classifiers based on separating classes by hyperplanes such as Support Vector Machine (SVM) classifier, among others.

In this paper, we developed a Hybrid Classifier of Decision Rules (HCDR) [2], which was incorporated in the information on the main genetic markers within the dbGM, which allow us to obtain better performance measures. Since this classifier has the better results, in this section we focus on the previous work of this classifier.

There are three ways [13] of how to build a classifier based on decision rules, which are described below:

- **Decision Trees:** With this method we create a set of rules, each of them for each leaf of the tree, which are easy to interpret.
- **Specific algorithms for rule induction:** The language of representation of decision rules is essentially propositional. Is this method, each rule is learned one by one, so each time it selects a rule, the examples that are covered by the selected rule are removed from the training set.
  The process is repeated iteratively until a stop condition is fulfilled. To learn a rule begins with rules as general as possible, then these records are being added to maximize classification accuracy of this rule.
- **Models based on association rules:** The aim of the association rules is to find associations or correlations between items or objects in the database. Wanted the best association rules, in order to overcome some limitations of the models based on decision trees, which consider only the attributes one by one partially.

Once we review the main strategies on how to build classifiers based on decision rules, in the next section we describe how we build the PSLC we propose.

III. PREDICTION SYSTEM OF LARYNX CANCER (PSLC)

A. PSLC components

The PSLC has two components [7], the component of Transformation and Selection of Data (TSD) and the Engine of Operation (EO).

In the TSD component receiving data input, either read from a file or entered manually by people who will use the PSLC, which have a special format ISCN. The function of this component will convert this format to a matrix format for easier the processing in EO component.

In the EO component, once the data are in the matrix format we perform the classification of each patient tested. This component is implemented in the HCDR, which will allow us to make a prediction of what kind of life has one patient. The following sections explain in more detail each component of PCS_EME.

 ISBN: 978-1-61208-277-6

B. Transformation and Selection of Data

For this component, we implement techniques to apply the pre-processing and transformation phases on dbGM. For these two phases we implement two algorithms of Digitization or Labeling and Discretization, which perform the pre-processing and transformation phases. First, we show the ISCN format, taking an example of genetic biochip BD; then, later, we describe each algorithm of the TSD component.

1) ISCN Format

The DB has some special format called ISCN, which has an entire nomenclature that is handled in the area of Human Genetics, for the specific case of the dbGM.

These nomenclatures are reduced to a subset, which contains the information about possible genetic alterations associated with certain cancers. In one study of Centro Médico Nacional Siglo XXI of Instituto Mexicano Del Seguro Social, they obtained a dbGM with 19 records of patients with laryngeal cancer and two patients with cervical cancer [4], over five years.

They obtained this dbGM from progenetix (this database containing an overview of the anomalies in the number of copies of Human Cancer of Comparative Genomic Hybridization (CGH)), to generate a karyotype and a grouping of data; one record of this base is as follows:

L1 rev ish enh(1p36.22, 1p13.1, 1qtel, 2q14, 3p14.2, 3q26.3, 5q21q22, 5q33q35, 7q32q34, 8p22, 8p22q21.3, 8q24qter, 9q22.3, 14qtel, 15qtel, 16qtel, 17ptel, 17q23, 20q13.2, 20qtel, Xq12) dim enh(19ptel) amp enh(1q21, Xp22.3)

As we can see, this information is totally unknown because it is in a specialized format, such as the ISCN, but each acronym has a special meaning, which we are going to describe for each acronym that appears on this record:

- *L1:*Label assigned to the patient.
- *rev, ish, enh, dim and amp:* Techniques used to obtain genetic information, *rev* is the technique called Reverse, *ish* is the technique called In Situ Hybridization, *enh* is the technique called Enhanced, *dim* is the technique called Diminished and finally *amp* is the technique called Amplified Signal.
- *1p36.22:*It is the first record in the parentheses, the number at the start leading *1* is the number of chromosomes, *p* is the kind of arm that is in this case *p* is the short arm, if occupied *q* it would be the long arm, and finally *36.22* refers to the area of the chromosome where there is a change.

We can see that each record has a set of implicit information, which is difficult to handle for most classifiers, which is why we implement a method of transformation for these data. To achieve this objective, we implement digitization and discretization algorithms to apply on this dbGM and we obtain a format more appropriate for the incorporation of classifiers, then we show how we apply these algorithms on dbGM

2) Implementing Labeling and Discretization techniques.

In the example of the previous subsection, there are terms that are used exclusively in the ISCN format, which represent the acronym of each element that compose a record of this type of the dbGM.

With these acronyms, the labeling and discretization techniques were applied for change to integer values, which are handled by classifiers more directly, as a first element each acronym that are present in the dbGM is listed.

Techniquesfor each patient

- rev ish enh
- dim enh
- amp enh
- amp
- dim

Chromosomes

- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, Y

Type of arm on chromosome

- p: Short arm.
- q: Long arm.

Region where the genetic alteration is present:

- Real values associated with each of the chromosomes that can go terminal region (ter) at the telomeric (tel).

So, when we apply the labeling and discretization techniques, we have to assign a nomenclature to each of these acronyms, the nomenclature used is as follows:

Techniques:

- rev ish enh → 101
- amp enh → 102
- dim enh → 103
- amp → 104
- dim → 105

Chromosomes:

- From 1, 2, …, 22
- X → 23
- Y → 24

Arm:

- P → 300
- Q → 301

Region:

- Real values multiplied by 100
- ter → 4000
- tel → 4001

The allocation of these values was arbitrary, having this way the transformation of all acronyms of ISCN format. In Table I, we can see an example of the application of these techniques on the record shown in Section 1).

 ISBN: 978-1-61208-277-6

The allocation of these values was arbitrary, having this way the transformation of all acronyms of ISCN format. In Table I, we can see an example of the application of these techniques on the record shown in Section 1.

Record in ISCN format:

L1, rev ish enh (1p36.22, 1p13.1, 1qtel, 2q14, 3p14.2, 3q26.3,5q21q22, 5q33q35, 7q32q34, 8p22, 8p22q21.3, 8q24qter, 9q22.314qtel, 15qtel, 16qtel, 17ptel, 17q23, 20q13.2,20qtel, Xq12) dim enh(19ptel) amp enh(1q21, Xp22.3)

TABLE I. RECORD AFTER WE APPLY THE LABELING AND DISCRTETIZATION TECHNIQUES.

1011	300	3622	1	300	1310	1
	301	4000	2	301	1400	3
	300	1420	3	301	2630	5
	301	2100	5	301	2200	5
	301	3300	5	301	3500	7
	301	3200	7	301	3400	8
	300	2200	8	300	2200	8
	301	2130	8	301	2400	8
	301	4000	9	301	2230	14
	301	4000	15	301	4000	26
	301	4000	17	300	4000	17
	301	2300	20	301	1320	20
	301	4000	23	301	1200	104
	19	300	4000	101	1	301
	2100	23	300	2230		

In the record of Table I, there are only integer data so makes it easier handling it for the classifiers, however exist a problem because the size of registration is large and each record has a different length. For this reason we propose a way of how to make the registers have the same size, we do this by separating techniques which have in a register.

To show how we apply this change in records to have the same size, we use another record of dbGM, which contains the following information:

L8T2N0 amp (19ptel, 22q11.21, 22q11.2, 22q13.3, 22qtel) dim (19ptel)

As we can see in the previous record, there are two techniques, which are the amp and dim, this example will help to illustrate each phase that must be performed to transform the format ISCN to the format that handles PSLC.

The operation of transformation has four phases, which are as follows.

Phase 1:

In the record above, there are two techniques, amp and dim, in this phase we identify the techniques, are divided, and new records are formed according to the number of techniques, the new records are creating as follows:

- amp(19ptel, 22q11.21, 22q11.2, 22q13.3, 22qtel)
- dim(19ptel)

Phase 2:

For each new record, we must separate it into each genetic alteration, retaining the order for each technique, after which we apply this phase to the new records and they change as follows:

- amp
    1. 19ptel
    2. 22q11.21
    3. 22q11.2
    4. 22q13.3
    5. 22qtel
- dim
    1. 19ptel

Phase 3:

Since they have separate genetic alterations, we will proceed to separate in chromosome number, type of arm and area of the chromosome where the alteration is present, respecting each new nomenclature established, we carry out as follows:

- amp -> 104

Order	*Chromosome*	*Arm*	*Region*
1	19	300	4001
2	22	301	1121
3	22	301	1120
4	22	301	1330
5	22	301	4001

- dim-> 105

Order	*Chromosome*	*Arm*	*Region*
1	19	300	4001

Phase 4:

Once we have coded both records, all information representing a new record in the dbGM, so now instead of

 ISBN: 978-1-61208-277-6

having two records, it will have six records, but all the same size; the example obtained from this phase can be seen in Table II:

TABLE II. FIRST CODING OF RECORD

Technique	*Chromosome*	*Arm*	*Area*	*Order*
104	19	300	4001	1
104	22	301	1121	2
104	22	301	1120	3
104	22	301	1330	4
104	22	301	4001	5
105	19	300	4001	6

We can see that in this matrix format, there are the different techniques (amp and dim), with all information in them, but we add new information of each record, for this case we add two new attributes, number of changes and sequence.

The number of changes attribute, has the techniques used in some patient. In the sequence attribute, since each technique was encoded in a range of 101 to 105, for this new attribute the last digit is used and depending on whether one or more techniques used they are strung together, returning to the previous example used the technique amp and dim which have associated values 104 and 105 respectively, so that the new attribute would be 45. Finally, adding these two attributes, the final record we can see in Table III.

With this example, we explain how we performed the transformation the format of dbGM, for having a matrix format to carry out the EO component of PSLC.

3) *Operation Engine*

When performing the task of data classification, there are two sets, training and testing. With the training set the models of classification algorithms are constructed. Using the test sets the classification is performed for each of the records in the DB, which are carried out individually, i.e. record by record.

This is the traditional way of how to perform data classification, however for the dbGM is not possible to carry out the classification in the traditional way, since as we can see in Table III; a set of six records in the matrix format is equivalent to one record in the ICSN format.

In this paper, we develop a module called Engine of Operation (EO) of PSLC. The EO implements a set of classifiers as Naive Bayes, C4.5, k-NN and the HCDR, among others to perform the task of classification. The classification performed with these classifiers is the tradition; therefore we find a way to change it, for this we implement a voting criterion as used in classifiers based on ensembles.

A classifier based on ensembles [1], joins one or more types of classifiers to obtain improvements in performance measures, which is why we need a criterion to designate the classification of each example in the test set, there have been two voting criteria the Majority and Weighted voting criterion.

TABLE III. MATRIX FORMAT OF A RECORD

Tech nique	*Number ofchanges*	*Sequence*	*Chro moso me*	*Arm*	*Area*	*Order*	*Class*
104	2	45	19	300	4001	1	Survival
104	2	45	22	301	1121	2	Survival
104	2	45	22	301	1120	3	Survival
104	2	45	22	301	1330	4	Survival
104	2	45	22	301	4001	5	Survival
105	2	45	19	300	4001	6	Survival

In Majority Voting Criterion, each of the classifiers votes to decide to which class each example belongs to for the test set, eventually counting the votes and assigning the class majority. For Weighted Voting Criterion, each classifier has a weight, so each vote has a different weight, so that in the end has a weighted voting, but the class is assigned given the most voting weight.

With these two criteria, we decided to use the majority-voting criterion [3] to adapt the classifiers of EO, to exemplify this incorporation we can see the diagram in Figure 1, which shows the traditional way to carry out the classification (Tclass) and the applying of majority voting criterion for classification (MVclass).

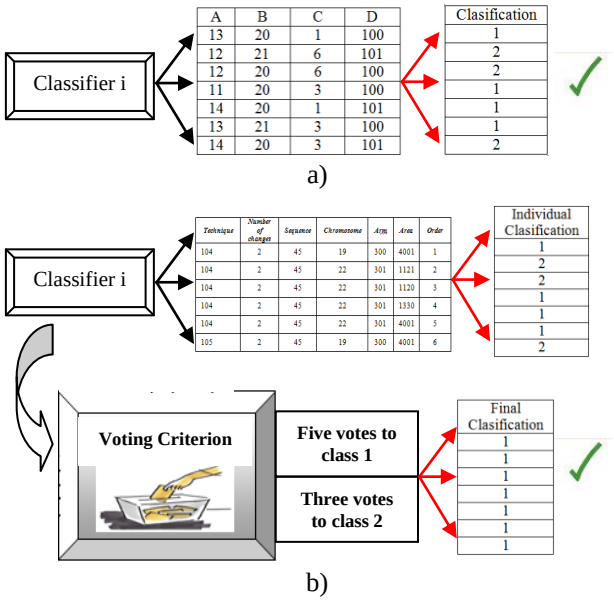

Figure 1. a) Traditional way to perform the classification. b) Classification with a majority voting criterion.

As we can see in Figure 1, the difference between Tclass and MVclass is the incorporation of the majority-voting criterion, in which the votes are counted there for the class 1

and the class 2. The class 1 has the most votes, this set of examples that represent a single example in the ISCN format, they assign the majority class, respecting the original meaning that we have for each record, we conclude this component; in the next section, we review the experiments and results obtained by the PSLC.

IV. EXPERIMENTS AND RESULTS

In this section, we show the results obtained by performing a series of tests with the BD of larynx cancer to measure the performance of the PSLC. The tests consisted in generating a number of test and training sets with different sizes; we can see these sets in Table IV.

TABLE IV. SELECTED TRAINING SETS.

Cases	Number of patients	Number of records
1	3	49
2	5	92
3	7	135
4	9	151
5	11	160
6	13	224
7	15	299
8	17	332
9	19	382
10	Full DB (21)	431

With each training set of Table IV, the MVclass were performed according to the following steps:

1. We take two sets of dbGM, a set classless and another set with the class of each record; these will be the test and training set respectively.
2. The test sets we use by some classifiers, to perform the MVclass of each example.
3. Once it gets the MVclass of each record of the test set, we compared them to their real classes.
4. At the end of this comparison, we obtained the performance measures that will allow us to evaluate each classifier.

At the end of the testing process, we choose the accuracy as the performance measure to evaluate each classifier. Accuracy provides information on the percentage of correctly classified examples, out of the entire test set; this performance measure is formally defined as follows:

$$Accuracy = \frac{(a + d)}{(a + b + c + d)} \quad (1)$$

In equation 1 has the following:

- (a, d) is the correctly classified examples and
- (c, b) is the classified incorrectly examples

It is noteworthy that for each of the tests we used cross-validation to determine the validity of each classifier we implement. Figure 2 and Table V show the final results we obtain in each of the tests that were performed.

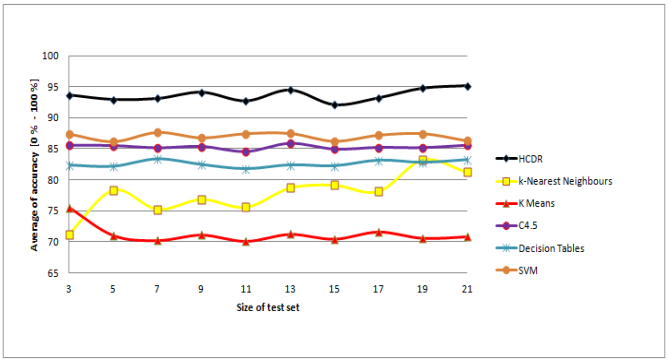

Figure 2. Graph of the final results.

TABLE V. FINAL RESULTS FOREACH CLASSIFIER

Size of test set	HCDR	k-NN	K Means	C4.5	Decision Tables	SVM
3	93.67	71.21	75.5	85.54	82.32	87.37
5	91.95	78.34	71.06	85.48	82.12	86.19
7	93.1	75.24	70.25	85.12	83.33	87.61
9	94.1	76.85	71.13	85.34	82.41	86.78
11	92.72	75.56	70.13	84.54	81.73	87.41
13	94.45	78.69	71.24	85.83	82.37	87.45
15	92.12	79.14	70.48	84.93	82.21	86.25
17	91.21	78.13	71.61	85.19	83.13	87.17
19	94.74	83.26	70.62	85.15	82.77	87.42
21	95.13	81.26	70.85	85.54	83.21	86.31
Average of accuracy	93.619	77.768	71.287	85.266	82.56	86.996

We can see in Figure 2 and Table IV that the HCDR showed better results than traditional classifiers, obtaining an average accuracy of 93,619. The HCDR was strengthened with genetic markers found in the database of the study after five years in patients with larynx cancer. By using new technologies such as scanning AXON, the targets are searched must be related to cancer of the larynx and cervical cancer.

Table VI shows the class for each patient in the study over five years. In these patients, we determined the main patterns present at the genetic level, which are present in ISCN format.

TABLE VI. PATIENTS WITH CANCER OF LARYNX, WITH THEIR RESPECTIVE CLASS AND PATTERNS.

Patient	Class	Pattern
L3	Survival	3q21q22 and 3q27q29
L4	Survival	1q25q31, 3q21q22 and 7qtel
L5	Poor survival	17q11.2

 ISBN: 978-1-61208-277-6

L6	Poor survival	5q21q22, 5q33q35, 8p22q21.3, 8q29qter and 7q32q34
L7	Poor survival	8p22q21.3 and8q29qter
L8	Survival	22q11.2 and 22q13.3
L9	Survival	3q21q22
L10	Survival	2q31q32
L11	Survival	5q33q35
L12	Survival	22q11.2 and 22q13.3
L13	Poor survival	1q25q31, 3q21q22, 3q27q29 and 7q21q22
L14	Poor survival	3q21q22 and 7p12.3p12.18p22q21.3
L15	Poor survival	1q25q31 and 3q21q22
L16	Poor survival	3q21q22 and 7p12.3p12.1
L17	Survival	3p12p13 and 17q21q22
L18	Survival	12q13q14
L19	Survival	1q25q31, 3q21q22, 3q27q29 and 7p12.3q12.1
L20	Poor survival	3q21q22, 3q27q29,7p12.3p12.1 and7q21q22
L21	Poor survival	5q33q35

With Table VI, we locate the presence of CRBP-1 and EGFR genes. CRBP-1 is a protein involved in the transport of retinol from its storage sites of the liver to peripheral tissues. Vitamin A plays an important role in a variety of cellular processes associated with epithelial tissue proliferation and differentiation. This protein could improve the condition of some form of cancer, when analyzing our database we observed that it actually improved the conditions of patients with larynx cancer.

In contrast, EGFR amplification could be associated with poor survival of patients with larynx cancer. A gain of CRBP-gene may have a protective effect and increased survival. These data suggest that alteration CRBP-1 gene and its expression in carcinomas of the larynx squamous provide prognostic information with greater potential patient survival.

Previously, Peralta et al. [4] reported the behavior of these two genes but with techniques in microarray data analysis. For this work became the chromosomal level analysis of each of the patients involved, and seek the relationship patterns of each sample to determine the genes present. In the same way, it was determined that the gene CRBP-1 was present in patients who survived the EGFR gene was present in poor survival. In this way it was shown at the chromosome level and table's decision techniques and data mining obtained the correct result.

By identifying these genetic patterns, we could implement the HCDR, which is able to incorporate decision rules that endorse the existence of these patterns in any of the branches of the decision tree constructed, which resulted an average accuracy of over 15% compared to traditional classifiers.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a Prediction System of Larynx Cancer (PSLC), for exploration and analysis of the databases of genetic markers (dbGM). The PSLC has two components, the Transformation and Selection of Data (TSD) component and the Operation Engine (EO).

With the TSD component, we perform data transformation that was in ISCN format to a matrix format for better data handling. This component was necessary since most of the classifiers are better suited to this type of data format. Otherwise if we had taken the raw ISCN format, many changes should have been made in the operation of each classifier. That is why we decided to not make these changes and implement the TSD module.

With the EO component implemented as an alternative way to perform the classification by incorporating a majority voting criterion (MVclass). With this alternative form of classification, we performed a series of tests with a set of classifiers. In Table V and Figure 2, the accuracy grew 15% with respect to the other classifiers considered for the tests; this gives us the result that the HCDR performs better than traditional classifiers.

This classifier considers the incorporation of decision rules with tree structures, together with the genetic markers, the CRBP-1 to survival of patients and EGFR to poor survival of patients, located in a series of preliminary tests.

With PSLC we proved that by incorporating two areas of knowledge (Artificial Intelligence and Genetics), different from one another, there may be generated more complex algorithms; such is the case of HCDR. With this model of classification we obtained a better accuracy for each one of the tests we made to the dbGM, comparing it with different traditional classifiers. The PCSL is an example of the new interdisciplinary projects of Science and Technology that combine more than one area of current research.

For the PSLC we consider its seminal work on the issue, since if we want to develop new versions incorporating other types of cancer, it will be directly because in tests we found that the genetic information of two different types of cancer, in this case larynx and cervical, show many similarities, so we left the base ready to scale this project.

ACKNOWLEDGMENT

Special thanks to Instituto de Ciencia y Tecnologia Del Distrito Federal - ICyTDF and Universidad Autónoma Metropolitana - UAM for their support for the realization of this system we propose in this paper. We also thank Dr. Mauricio Salcedo-Vargas, which belongs to the Centro Médico Nacional Siglo XXI of Departamento de Oncología for their valuable collaboration is aspects of genetics.

REFERENCES

[1] B. Moreno Montiel and R. Mac Kinney Romero,"A Hybrid Classifier with Genetic Weighting," in Proceedings of the Sixth International Conference on Software and Data Technologies, July 2011, vol. 2, pp. 359–364.

[2] T. Gunnar Houeland and A. Aamodt, "An efficient hybrid classification algorithm: an example from palliative care,"

Proceedings of the 6th international conference on Hybrid artificial intelligent systems, September 2011, vol. II, pp. 197–204.

[3] E. Hüllermeier and S. Vanderlooy, "Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting," Pattern Recognition, January 2010, vol. 43, pp. 128–142, doi:10.1016/j.patcog.2009.06.013.

[4] R. Peralta, M. Baudis, G. Vazquez, S. Juárez, H. Decanini, D. Hernandez, F. Gallegos, A. Valdivia, P. Piña, and M. Salcedo, "Increased expression of cellular retinol-binding protein 1 in laryngeal squamous cell carcinoma," Journal of Cancer Research and Clinical Oncology. January 2010, vol. 136, pp. 931–938, doi: 10.1007/s00432-009-0735-9. Epub 2010 Jan 7.

[5] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. Yu, ZH. Zhou, M. Steinbach, D. Hand, and D. Steinberg, "Top 10 algorithms in data mining," Knowledge and Information Systems, April 2009, vol. 14, pp. 1–37, doi:10.1007/s10115-007-0114-2.

[6] M. Shinawi and SW. Cheung. "The array CGH and its clinical applications,". Drug Discov Today, September 2008, vol. 13, pp. 760–770, doi:10.1016/j.drudis.2008.06.007. PMID 18617013.

[7] I. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques," Morgan Kaufmann Publishers, January 2005, 2nd edition, pp. 560.

[8] D . Albertson, and D. Pinkel,"Genomic microarrays in human genetic disease and cancer," Hum Mol Genet, October 2003, no. 2:R, pp. 145-52, doi: 10.1093/hmg/ddg261.

.

[9] M. Kujawski, M. Rydzanics, M. Sarlom-Rikala, and K. Szyfter, "Rearrangements involving the 13q chromosome arm committed to the progression of laryngeal squamous cell carcinoma," Cancer Genet Cytogenet, August 2002, vol. 137, No 1 pp.54-58.

[10] S. Struski, M. Doco-Fenzy, and P. Cornillet-Lefebvre "Compilation of published comparative genomic hybridization studies", Cancer Genet Cytogenet, May 2002, vol. 135, No 1 pp. 63-90.

[11] F. Jiang, R. Desper, C. H. Papadimitriou, R. A. Schäffer , O. Kallioniem, J. Richter, P. Schraml, G. Sauter, M. J. Mihatsch, and H. Moch, "Construction of evolutionary tree models for renal cell carcinoma from comparative genomic hybridization data,"Cance research, November 2000, vol. 60, No22 pp. 6503-6509.

[12] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W. L. Kuo, C. Chen, Y. Zhai, S. H. Dairkee, B. M. Ljung, J. W. Gray, and D. G. Albertson, "High resolution analysis of DNA copy number variation using compartive genomic hybridization to microarrays," Nature Genetics, October 1998, vol. 20, pp. 207-211,, doi: 10.1101/gr.2012304.

[13] T. M. Mitchell, "Machine Learning", McGraw-Hill Science/Engineering/Math, March 1997

[14] S. Solinas-Toldo, S. Lampel, S. Stilgenbauer, J. Nickolenko, A. Benner, H.Döhner, T. Creme, and P.Lichter, "Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances," Genes Chromosomes Cancer, December 1997, vol. 20, No 4 pp. 399-407, doi: 10.1002/(SICI)1098-2264(199712)20:4<399::AID-GCC12>3.0.CO;2-I.

 ISBN: 978-1-61208-277-6