INSTITUTO TECNOLOGICO SUPERIOR DE MISANTLA

ITSM

SYMPOSIUM FOR APPLIED
COMPUTER SCIENCE
(S A C S)

ISSN 2594-1054

Symposium for Applied Computer Science (SACS)

SACS-2017

8th and 9th of June

Misantla, Veracruz, México

SYMPOSIUM FOR APPLIED
COMPUTER SCIENCE
(S A C S)

Información Legal

SYMPOSIUM FOR APPLIED COMPUTER SCIENCE (SACS), Año 1, Vol 1, noviembre 2016 - noviembre 2017, es una publicación anual editada por el Instituto Tecnológico Superior de Misantla, Km. 1.8 carretera a Loma del Cojolite, Col. Centro, Misantla, Veracruz, México, C.P. 93821, teléfono 01 (235) 323 15 45, msc.itsm.edu.mx. Editor responsable Dr. Eddy Sánchez de la Cruz. Reserva de Derechos al Uso Exclusivo No. 04-2017-102716083000-203, ISSN 2594-1054, ambos otorgados por el Instituto Nacional del Derecho de Autor. Última actualización 25 de noviembre de 2017. Km. 1.8 carretera a Loma del Cojolite, Col. Centro, Misantla, Veracruz, México, C.P. 93821.

Directorio

Cuerpo Editorial

Comité de Revisión

Diseño Editorial

Propiedad Intelectual

Distribución

Publicación electrónica disponible en:
http://msc.itsm.edu.mx/sacs/index.php
Correo electrónico:
sacs@itsm.edu.mx

SYMPOSIUM FOR APPLIED
COMPUTER SCIENCE

Index

Methodology for the optimal assignment of class schedules in the Higher Technological Institute of Misantla 1
Ramos Salvador Antonio Aquino and Jorge Mario Figueroa Garcia

Planning schedules in the ITSM through Genetic Algorithms 5
Carmen Juliana Aguilar Fernandez and Jorge Mario Figueroa Garcia

Virtual Reality Applied in Gradual Exposure Therapy for Spider Phobia 8
Lopez Escalante, Hector and Suarez Leon Guillermo

Arquitectura de un sistema multi-agente para control de inventarios 12
Maribel Durán Salas and Galdino Martínez Flores

Analysis of hyperspectral images for the detection of pests in coffee crops 18
Arely Guadalupe Sánchez Méndez and Simón Pedro Arguijo Hernández

Architecture for the analysis of Surface damage of citrus (orange and lemon) in the region of Misantla using image processing 21
Juan Pablo Salazar and Simón Pedro Arguijo Hernández

Graphic model to evaluate human brain connectivity over time 24
María Luisa Córdoba Tlaxcalteco, Carlos Hernández Gracidas, Alejandro del Rey Torres Rodríguez and Yoselyn Nohemí Ortega Gijón

Classification of signals in multiple EEG channels. Detection of human brain maturation factors 39
Yoselyn Nohemí Ortega Gijón,Carlos Hernández Gracidas, Alejandro del Rey Torres Rodríguez and María Luisa Córdoba Tlaxcalteco

Artificial Intelligence Techniques to create Distance Learning Adaptive Styles in Moodle platform 45
Angel Gaspar May Uuh, Dr. Alejandro del Rey Torres Rodriguez and Rajesh Roshan Biswal

SYMPOSIUM FOR APPLIED
COMPUTER SCIENCE

Index

Traducción de lenguaje de señas usando algoritmos de visión por computadora y reconocimiento de patrones 50
Eduardo Mancilla Morales and Simón Pedro Arguijo Hernández

Aprendizaje automático para la clasificación de nefropatía diabética mediante KDIGO 53
Francis Susana Carreto Espinoza and José Antonio Hiram Vásquez López

Secuenciación de cadenas ADN, basado en Secuencias Frecuentes Maximales 58
Ma del Refugio Velazco Hermosillo, Luis Alberto Morales Rosales, Antonio Alejo Aquino and Mariana Lobato Báez

Siscom ETL, an alternative when versalitility in the extraction, transformation and load are factor in the quality of the data analysis..... 62
S. Juarez, E. Trujillo and H. Andrade

Classification techniques used to resolve the email overloading problema 66
José Arcángel Salazar Delgado, Alberto Méndez Torreblanca and Jorge Estudillo Ramírez

Electrocardiograms Analysis on the Cloud 70
Uriel Rubio Escamilla and Simón Pedro Arguijo Hernández

Proposal of a Multi-agent System for the generation of school hours in the ITSM 74
Francisco Adolfo Aguilar Gómez and Jorge Mario Figueroa García

Generation of a model implemented in Java that allows it Preprocessing, segmentation, and classi_cation of species of photosynthetic protists from microscopic images 76
N. Ramirez and S. Arguijo

Arquitectura de un sistema multiagente para el monitoreo a protocolos de mujeres adolescentes embarazadas de alto riesgo 81
Arcos Muñoz Sandra G

Methodology for the optimal assignment of class schedules in the Higher Technological Institute of Misantla.

Ramos Salvador Antonio Aquino [*]
Jorge Mario Figueroa Garcia [**]

[*] *Higher Technological Institute of Misantla, PC. 93820 Mexico (e-mail: 162t0083@itsm.edu.mx).*
[**] *Higher Technological Institute of Misantla, PC. 93820 Mexico (e-mail: jmfigueroag@itsm.edu.mx).*

Abstract: In all educational institutions, the process of creating an academic schedule is a challenge they face each semester. The process is not simple, since it is subject to restrictions both of teachers, number of students, number of classrooms, workshops or laboratories. The objective of this research is to propose an alternative solution for the problem of scheduling using the metaheuristic optimization technique of Ant Colony. In this paper we review some of the most used methods to solve these timetabling problems, a problem considered NP-complete, among which are, genetic algorithms, Multi-agent systems and optimization techniques such as Colony Ants and Tabu Search. Before the solution of the problem, an analysis of all the necessary information is proposed, to identify the objectives goals, restrictions on the subject, clarity of the flow of the information that is considered to be able to make the planning of schedules in an appropriate way.

Keywords: Timetabling, Scheduling, Metaheuristics, Optimization Techniques, Hard Restrictions, Soft Restrictions.

1. INTRODUCTION

In general, the problem of assigning schedules in universities consists of the assignment of specific courses to determined periods of time in an academic period, considering the teachers required in each subject, groups of students taking a set of subjects, days Or slots available, the rooms required in such a way as to optimize a set of strong and soft constraints related to the institution's education system. At the Instituto Tecnolgico Superior de Misantla (ITSM), there is also the need to have an application that allows a good planning process in the allocation of classrooms and class schedules, for this reason it is proposed to develop a methodology to solve the problem raised using a metaheuristic tool Ants' Colony (ACO for its acronym in English).

The objectives set forth in order to carry out the development of the proposed methodology are shown below.

General objective

Propose an optimization methodology for the assignment of classrooms and class schedules in the ITSM, which satisfies the restrictions caused by the resources involved.

Specific objectives

- Identify the characteristics of the scheduling process and the assignment of salons, specific to the institution that allows defining particular restrictions of the problem.
- Design a methodology that is capable of providing an efficient and effective solution to the scheduling problem at the ITSM.
- Generate an experimental design that allows to adequately evaluate the performance of the algorithms that are generated based on the methodology described above.
- Validate the results of the model, comparing them with the work done on the same problem, to define its possible application.

2. GENERALITIES

In the context of the allocation of academic hours, there are two key concepts in the literature that group the research around the problem. Concepts with timetabling and scheduling.

A problem of scheduling is one that is about the allocation of resources over time to carry out a set of tasks and aims to minimize the total cost of the resources allocated.

Timetabling is defined as the problem of allocating certain resources, subject to constraints, in a limited number of schedules and physical locations in order to satisfy a series of objectives to the greatest extent possible.

Timetabling problems contain strong constraints and weak constraints. The strong or hard restrictions are those that must be fulfilled in a mandatory way and not fulfilled disables the solution. They are generally spatial or temporal constraints that must be met.The weak or soft restrictions are those that are desirable to be met, but that their non-compliance does not disable the solution although it does of lower quality. Generally, they are preference or prioritization constraints, and it is often these constraints that are to be maximized (or minimized as the case may be) in order to get closer to the optimal solution.

The main problem is to assign classrooms and hours for the development of the classes of each of the subjects within a framework of five days a week and the time variables between 7:00 a.m. and 8:00 p.m.

The problem also presents a series of characteristics that are common to all the systems of elaboration of schedules as they are:

Hard Restrictions:

- A student should not attend two different events at the same time (student time crossing).
- A teacher can not dictate more than one class at a time (timetable for teachers).
- Respect the schedules that teachers have available.
- Each course must be assigned to a classroom with sufficient capacity for the estimated student demand for that course.
- A classroom may not be a simultaneous venue for two different events (classroom crossover).

Soft Restrictions:

- Avoid assignments where possible in the last two hours of the day.
- It should be avoided, as far as possible, to assign courses in classrooms that are considered for other important events.
- The schedules of a subject must have some coherence (consecutive hours).
- There should be no time slots in the classrooms (consecutive hours).
- A subject must be taught in one classroom only.

Below are some research papers that have been dedicated to propose a solution to the problem of scheduling using similar techniques.

Suarez and Castrillon, carried out an investigation of the problems that the educational institutions undergo, referring to the optimal allocation of schedules, mainly at university level. In addition, they point out that the investigations consulted by themselves focus on answering the questions such as: What to teach ?, how to teach ?, or where to teach?, but that the questioning of When to teach?, has not been analyzed objectively and much less tools have been developed to plan the time according to students' cognitive rhythms. The methodology they propose in this research is through metaheuristics that are intelligent strategies to design or improve very general heuristic procedures with high performance. In the results obtained using these intelligent scheduling techniques, they point out that a more adequate programming is achieved and that the development of cognitive rhythms is guaranteed, the solution is approaching 93% efficiency with respect to an optimal optimum solution.

Lozada et al., Present another solution to the problem faced by educational institutions regarding the optimal allocation of classrooms, the proposed methodology is based on heuristic techniques and on the technique of optimization by Ant Colony, indicate that institutions must have The infrastructure necessary for students and teachers to find themselves in an environment conducive to the development of their activities. In order to achieve the objective, the authors point out that some important variables must be considered, such as the number of events programmed in classrooms, the total number of students assigned to the institution, number of students per event and total number of classrooms available. It is also mentioned that it is very important to take into account the preferences of students in order to provide a clear and simple schedule that guarantees full attendance at each event. The algorithm used by Ant Colony is a metaheuristic capable of finding good quality solutions to highly complex optimization problems, is based on the ability of natural ants to find food from relatively simple individuals, but with a structure Very efficient social. Finally, in the results stage, there is no mention of the percentage achieved by implementing this algorithm, only if an optimal solution was found based on the tests performed.

Lopez mentions that the generation of class schedules is an inevitable academic-administrative activity in all educational institutions and that this problem is of very high computational complexity, therefore, it has been classified as an NP-complete problem. He points out that the solutions that have been proposed in many investigations are more oriented towards the administrative staff without considering the parties actually affected who are teachers and students. To solve this problem, a solution based on an agent or multi-agent system is proposed in order to dynamically generate student schedules. Regarding the results, the author mentions that he does not yet have quantitative results, however, he indicates that it is expected that implementing this technique will not only optimize the computational times, but will achieve greater student satisfaction.

Suarez and other authors, developed a solution to the problem of optimal scheduling in a public school in Colombia, not only consider that the allocation of classrooms and teachers is appropriate, for them the most important factor in this optimization process are the Rhythms Cognitive characteristics presented by students. In order to solve this problem they propose the Genetic Algorithm of Non-denominated Classification and Random Search. The result obtained from a reduction of 22% in the number of classes lost by the group, the algorithm used shows a greater efficiency compared to others that apply to the same problem and evaluated in the same way.

3. METHODOLOGY

The algorithm of optimization by Ant Colony is a metaheuristic able to locate a solution of approximate to highly complex problems with the case of scheduling. This method is based on the study of the behavior of the ants to locate their food by identifying the shortest path between the food and its colony; They follow that path that is most likely where the pheromone brand has more relevance.

Below is the process that natural ants follow to find food.

Figure 1 shows the example when the ants reach a point where they must decide whether to turn right or left, as initially there is no trace of pheromone on the roads the choice is made randomly.

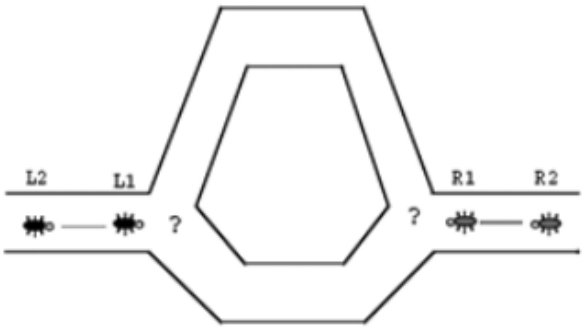


Fig. 1. Ants at decision point

Figures 2 and 3 show how the ants begin to make their way, it must be assumed that all the ants go at the same speed. And you can also see the number of dotted lines is proportional to the number of pheromones that insects have deposited in their travel.

As shown in figure 4 the lower path is shorter therefore, the number of ants that transit through it will be greater in the same time span. This indicates that on the shortest path accumulates more pheromone and after a certain time this amount will be enough for the new ants to decide which route to go.

In the implementation of an ant colony algorithm, possible solutions are called artificial ants, which can be a feasible solution or not. This alternative solution is built on the basis of rules that emulate the behavior of real ants such as exploitation, exploration and evaporation of pheromones. At the beginning there is a colony or a set of possible solutions and in each iteration the ants create alternatives

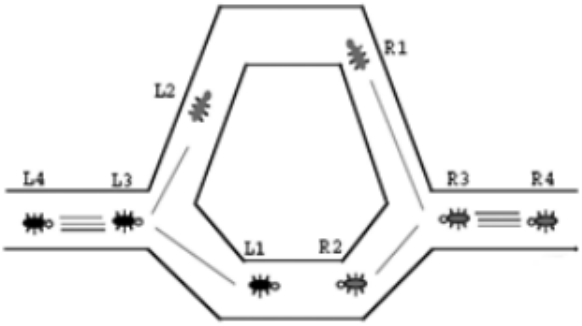


Fig. 2. . Choosing randomly between the lower and upper paths

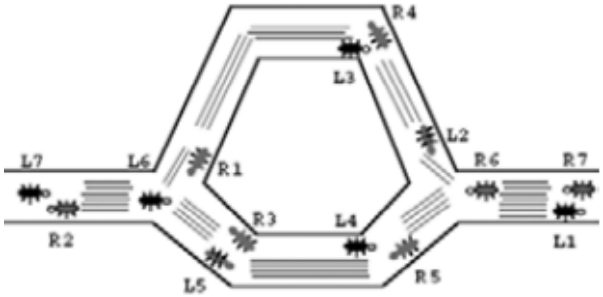


Fig. 3. Effects of the number of ants circulating in each path

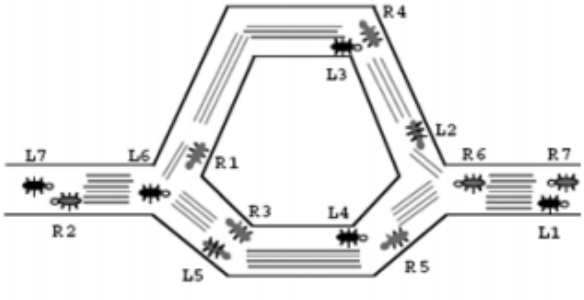


Fig. 4. Greater accumulation of pheromones in the shortest path.

based on the information collected by the predecessor solutions. The number of ants is an important parameter to be defined in the algorithm design.

Also defined is a memory structure that stores numerical data that emulate the pheromone of real ants. This structure, called pheromone trail, stores the degree of ant acceptance or previous solutions for a state variable found in an alternative solution.

This method differs from other heuristic methods in which the construction of the solutions is done based on the learning gained previously and that has been stored in the pheromone matrix, leaving in the background the impact that this solution can have on the objective function.Like other methods, the ant colony algorithm prevents it from converging to a local solution, for which there is a process called pheromone evaporation, which reduces the

pheromone trace using a probability function. Some of the advantages of this method in the scheduling problem is that they have a high probability of finding some global optimum, since conventional methods regularly converge to local optimum.Moreover, the representation is easy to analyze and understand. Among the disadvantages of the ant colony method is the high degree of uncertainty as to the convergence time of the algorithm and the high processing cost it requires. Like other heuristic methods, it does not ensure that exact solutions are found, but only optimal solutions in one neighborhood, which are often sufficient.

4. CONCLUSIONS

With the bibliographical review of several investigations that have been performed using ant colony optimization algorithms, it has been found that they are a very useful tool for solving complex problems called NP-Hard, it can also be observed that this type of algorithms Are the object of investigation of many authors which favors that constantly are being proposed improvements and new adaptations that makes them more complete and efficient.

5. BIBLIOGRAPHY

Cruz, O. L. (2015). A solution based on agents to the problem of generation of schedules. Engineering, mathematics and information sciences, 7385. Esmoris, A. V. (2013). Heuristic algorithms in optimization.PhD thesis, University of Santiago de Compostela.

Lozada, J. M., Diana Lorena Hoyos, C. A. P., and Chvez, J. J. S. (2013). Optimal heuristic tools for the assignment of class schedules. Engineering Research, 10:6874.

Nouri, H. E. and Driss, O. B. (2016). Matp: A multi-agent model for the university timetabling problem. Springer International Publishing Switzerland, pages 1122.

Suarez, V. F. and Castrilln, O. D. (2011). Design of a methodology based on intelligent techniques for the distribution of academic processes in work environments job shop. 5th International Conference on Industrial Engineering and Industrial Management, pages 485492.

Surez, V. F., lvaro Guerrero, and Castrilln, O. D. (2013). Programming of school schedules based on rhythms cognitive using a non-dominated sorting genetic algorithm, NSGA-II. Technological information, 24:103114.

Planning schedules in the ITSM through Genetic Algorithms ⋆

Carmen Juliana Aguilar Fernandez *
Jorge Mario Figueroa Garcia **

* *Instituto Tecnológico Superior de Misantla, CP 93820 MEX (e-mail: 162T0074@itsm.edu.mx).*
** *Instituto Tecnológico Superior de Misantla, CP 93820 MEX (e-mail: jmfigueroa@itsm.edu.mx)*

Abstract: The problem of scheduling is a very complex academic-administrative process of institutions due to the amount of restrictions it may have. It consists broadly in the assignment of each course to be offered in an academic period (semester, quarter, quarter, etc.) to a time band, a classroom, a teacher, subject to a set of restrictions and requirements. This paper reviews some algorithms used previously and will propose one for the assignment in an automated way.

Keywords: Scheduling Problem, Optimización, Algoritmos Genéticos, Restricciones duras y blandas, Timetabling.

1. INTRODUCCIÓN

The High Technological Institute of Misantla (ITSM) has scheduling as part of its processes. In particular, the Career Managers are responsible for assigning classrooms and teachers to the subjects of their respective careers, a process that to date is done manually, which is long, tedious and repetitive. Currently ITSM offers 9 Degree courses and two master's degrees taught in m Aulas, n Workshops by or Teachers.
GENERAL OBJECTIVE Provide an algorithmic solution for scheduling that automates and improves the time / use of classrooms and makes more effective the time the teacher invests in teaching and preparing classes.
SPECIFIC OBJECTIVES Identify the restrictions and policies applicable to the process to the process of assigning classrooms and teachers for the subjects of each of the careers offered in the institution. Design and implement two or three optimization solutions to compare the results obtained and choose the solution with the best results.

2. MODELADO DEL PROBLEMA

The problem modeling consists of two components: Data Component and Logical or Algorithmic Component.

2.1 Componente de Datos

Defines what information to include or process in the model. The matrix will be constructed at the execution time of the model, and will be constantly validated. The necessary data are as follows:

- **Academic Hours**: Three-dimensional shape matrix $H_{ndXnaXnhd}$ where: nd= number of days contemplated in the planning horizon

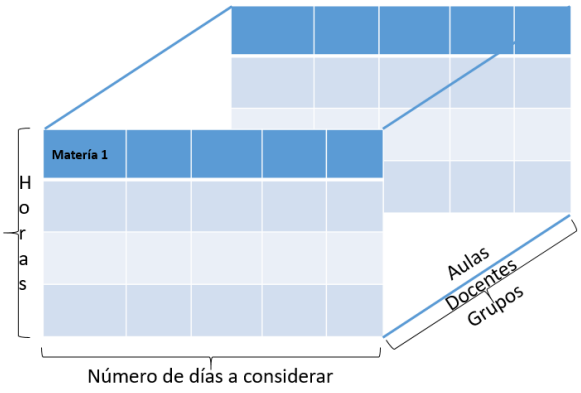


Fig. 1. Img 1. Three-Dimensional Matrix

na= number of Existing Classes
nhd= number of hours per day

- Asignaturas: Two-dimensional matrix containing number of subjects by the number of characteristics.
- Grupos: Three-dimensional matrix containing the number of existing subjects, the number of groups per subject and the number of subjects to be included in each field.
- Groups-day-time
- Subjects: suggested-lessons
- Classes
- Lessons-day-time
- Teachers:
- Teaching groups:
- Teachers-day-time:

Hard Restrictions: They are those restrictions that strictly and without exception must satisfy the model, otherwise the solution will be rejected.

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

- A group can only be assigned to a single classroom
- Two groups can be assigned to the same classroom at the same time if they take the same subject but only if the classroom capacity allows it
- A Classroom should only be scheduled in a time when it is available
- The hours programmed weekly for a group (subject-group) must be those required by the subject
- A group (subject-group) must be programmed in the classroom with sufficient capacity for its students
- The daily class hours for a group must be programmed consecutively (block) and in the same classroom.
- A subject group, either theoretical or practical, should be programmed in a classroom for this purpose
- A subject should only be programmed within the time period defined for it
- A group can only be programmed if it has assigned a teacher
- A teacher can only be assigned to a single group in a single schedule.
- A teacher can only be scheduled at a time when it is available
- A teacher can only be programmed in the subjects of his profile
- The number of hours a teacher can not exceed his number of hours for which he is hired

Soft Restrictions: They are those that desirably or as far as possible will try to satisfy the model and which will be a criterion for the selection of the most optimal solution.

- The groups for each subject should preferably be in the teacher's schedule
- The subjects in the Exact area must be taught preferably in the first hours of the day.
- No groups will be allocated in the lunch hour of a teacher

Restricted Optimization Consists of the process of Optimizing an objective function
Function Purpose: This function will try to minimize soft constraints as much as possible. To control that the objective function is met, it must be equal to 0, as it fences increasing its value
min f(H)=$\sum_i peso_k * costo_k$
where:
k= cantidad de restricciones suaves
$peso_k$= Weight defined for the restriction k
$costo_k$= Cost associated with dissatisfaction or non-compliance with the restriction k

2.2 Logical Component

This component defines the model to be used for solving the problem. In the state of the art an analysis of some articles of the existing literature is made.

3. ESTADO DEL ARTE

De la Cruz and Rodríguez propose to solve scheduling problems using genetic algorithms that consider two restrictions: two events can not be performed at the same time and there must be at least d days between two events. This is a good option for scheduling problems with additional restrictions such as "there must be at least d days between two events." The article does not include quantitative results, but has been included because of its importance and relation to the present investigation.

Castrillon proposes a method based on algorithms (evolutionary and genetic) to schedule class schedules in a university. The problem considers two types of restrictions: hard and soft. The combination of the algorithms lies creating a new parent each time the evolution stagnates in a local optimum. The methodology is an excellent technique for solving that kind of problem, which is at least 19.5 % more efficient than traditional techniques.

Nouri, & Driss propose a multi-agent model to solve the problem of university course schedule works through a set of cooperating agents that allow a highly distributed processing of tasks. The model provides a better solution that satisfies hard and soft constraints. The restrictions or preferences met indicate that it is 78 % on average per category analyzed.

Chmait & Challita use the Simulated Annealing (RS) and Ant Colony (OCH) algorithms to solve scheduling problems. A total of 15 times for each algorithm was tested in order to construct complete programs with different initial configurations. The execution times for CH are greater than those for RS, but when varying the number of ants of CH it is observed that the runtime increases or decreases and does not improve the results.

Restrepo and Moreno use a model through tabu search apply a local search procedure, preventing through the criminalization of certain movements, to fall into local optimum. The starting point for the application is an initial solution, which can be obtained through some algorithm constructor or using historical solution data. The most optimal results were using 71 Classrooms, 34 Subjects and 238 Groups with total execution times between 3m: 09seg and 3m: 51sec, this applying previously defined criteria for neighborhood selection compared to the criterion of randomness.

López in this article mentions that the generation of class schedules is an inevitable academic-administrative activity in all educational institutions and that this problem is of very high computational complexity, therefore, it has been classified as an NP-complete problem. He points out that the solutions that have been proposed in many investigations are more oriented towards the administrative staff without considering the parties actually affected who are teachers and students. To solve this problem, this article proposes a solution based on a system of agents or multi-agent to dynamically generate student schedules. Regarding the results, the author mentions that he does not yet have quantitative results, however, he indicates that it is expected that implementing this technique will not only optimize computational times, but will also result in greater student satisfaction.

4. METODOLOGÍA

In a genetic algorithm a population of candidate solutions (called individuals, creatures or phenotypes) to an optimization problem is developed towards better solutions. Each candidate solution has a set of properties (its chromosomes or phenotype) that can be mutated or altered. The solution starts from a set of individuals generated at random, in each generation the fitness of each objective is evaluated (objective function in the optimization problem). The most suitable individuals are chosen to make a new generation.

Table 1. Algoritmo Genético Simple

Requires:	Objective Function
1	Generate an Initial Population
2	.....While t < Max number of generations do
3	.....Generate a new solution by crossing and mutating
4	.....Cross
5	.....Muta
6	.....Select the best solution for the next generation
7	.....End of the loop
8	Return The best solution of the population

5. CONCLUSIONES

The work shows that Genetic Algorithms generate optimal solutions with a reduced computational time and improving the optimum solutions of an expert. The GAs operate simultaneously, which allows them to operate with different suboptimal solutions to reach the optimum. It has been demonstrated that the GAs determine better solutions than those of an expert. They allow to minimize the number of conflicts between the different resources of the institution. Depending on the design, the genetic algorithm can converge prematurely or never converge. converge.

6. HELPFUL HINTS

6.1 References

J. R. Rodriguez C. P. De La Cruz. Un algoritmo genético para un problema de horarios con restricciones especiales. Revista de Matemática: Teoría y Aplicaciones, 18(2), pags. 215229, 2011.

O. D. Castrillón. Combinación entre algoritmos genéticos y aleatorios para la programación de horarios de clases basado en ritmos cognitivos. Información tecnológica, 25(4), pags. 5162, 2014.

N. Chmait K. Challita. Using simulated annealing and ant-colony optimization algorithms to solve the scheduling problem. Computer Science and Information Technology, 1(3), págs. 208224, 2013.

Orlando López Cruz. Una solución basada en agentes al problema de generación de horarios. Ingeniería, Matemáticas y Ciencias de la Información, págs. 7385, 2015.

H. E. Nouri & O. B. Driss. Matp: A multi-agent model for the university timetabling problem. In Software Engineering Perspectives and Application in Intelligent Systems, págs. 1122, 2016.

Hernández & Martínez. Multiclasicador para predecir interacción de proteínas usando optimización de colonia de hormigas. Revista Cubana de Ciencias In- formáticas, págs. 195210, 2017.

& Perez Valladares Molina, R. A. Elaboración e implementación de un sistema informático para el Instituto Nacional San José Verapaz" del municipio de Ve- rapaz. Doctoral dissertation, Departamento de San Vicente, Universidad de El Salvador, 2013.

Aguilar, L. D. M. O., Valadez, J. M. C., Soberanes, H. J. P., González, C. L. D., Ramírez, C. L., & Soria-Alcaraz, J. A. (2015). Comparativa de algoritmos bioinspirados aplicados al problema de calendarización de horarios. Research in Computing Science, 94, 33-43.

REFERENCES

Appendix A. A SUMMARY OF LATIN GRAMMAR

Appendix B. SOME LATIN VOCABULARY

Virtual Reality Applied in Gradual Exposure Therapy for Spider Phobia

Lopez Escalante, Hector * **Suarez Leon, Guillermo** **

* *Instituto Tecnologico Superior de Misantla, Veracruz,(e-mail: 162T0007@itsm.edu.mx).*
** *Instituto Tecnologico Superior de Misantla, Veracruz,(e-mail: gsuarezl@itsm.edu.mx).*

Abstract:
The present document is related to the work of therapy of gradual exposure, in an artificial way, using Virtual Reality as support to carry it out, and will work with 20 people who have the phobia of spiders, forming two teams of 10 members To carry out the test, subjecting the team to therapy by means of a present psychologist without using the virtual support, and the team two will be carried out with the help of virtual reality.

Keywords: Virtual Reality, Therapy, Gradual Exposure, Spider Phobia.

1. DESCRIPTION OF THE PROBLEM

Every person throughout his life has experienced moments in which for some reason developed in him a fear caused by a bad experience, either with an animal, environment, activity or object that when they appear again cause him a sense of anguish .

Arachnophobia, a common fear in most people, survey data in 2005 mentions that 50 percent of women and 10 percent of men suffer from this phobia previously mentioned.

In these cases it is necessary to have the proper diagnosis by a professional, and determine if it is an "anxiety disorder" or a "specific phobia", and thus give the correct follow-up.

The treatment accustomed by psychologists, is the therapy of gradual exposure. In this method the specialists guide the patient to gradually compare their fear gradually. In documented cases of gradual exposure therapy has shown efficient results if it is completed.

2. JUSTIFICATION

Exposure to phobic stimulus is the common ingredient in techniques such as systematic desensitization, flooding, or graded exposure(Capafons, 2001).

Virtual reality (VR) is an environment of scenes or objects of real appearance. The most common meaning refers to an environment generated by computer technology, which creates in the user the feeling of being immersed in it. Said environment is contemplated by the user through normally a device known as glasses or helmet of virtual reality.

The gradual exposure with virtual reality (VR), is a methodology that is commonly observed in the treatments that are performed to patients with specific phobias, so that the raised system would be something new, in addition that allows the control of the fear with a better level , As well as providing security and confidence when implementing exposure therapy, by controlling the system with the appropriate exposure intensity.

If we focus on the technique of exposure, it should be remembered that, despite the many studies that demonstrate its effectiveness, about 25 percent of patients who are offered exposure treatment are rejected or abandoned. (Marks, 1992).

That's where VR will help patients be more receptive to participating in a more artificial way of their fear than the real part.

The characteristics of VR can contribute to this. In fact, our research group found that the majority of a sample of people, who suffered specific phobias (76.6 percent), preferred to start a virtual exposure treatment than a live exposure treatment for their problem. (Palacios, Bottle, Hoffman and Fabregat, 2007).

This technique of use with VR offers the same benefits as physical exposure therapy to patients, but the simple fact that patients know it is of a virtual nature, is less likely to abandon treatment.

The good use of VR shows us that it could be a very good tool for the treatment of this type of cases, since it allows to handle an artificial visual environment by means of modeling and without there being a need to really expose the patient to a therapy that results excessive, unnecessary and heavy.

3. OBJECTIVES

3.1 General

To design an atmosphere of arachnophobic simulation, and apply it in the psychological treatment of gradual exposure using virtual reality.

3.2 Specific

- Reduce time and cost patients have to move to take therapies with their doctor.
- create an easy-to-use interface for users (doctor and patient).

4. HYPOTHESIS

It is possible to improve the success of therapy by using the application in half-hour appointments in your own home, and to take the treatment without the presence of a psychologist.

5. SCOPE AND LIMITATIONS

The system will have a test that will help to evaluate the intensity of fear or anxiety, in addition to a system with a device composed of glasses with a small display on each lens and mobile device which will be responsible for showing the environment to show the patient; The application will add a virtual environment that will generate the phobia to the captured video sequence, in turn the result of the virtual video signal, can be sent to two places, one to the patient's glasses and another to the monitor of the therapist's computer To control the session.

In addition, it should be noted that only an environment with spiders, designed, animated and loaded to the system will be available. In order to perform the virtual reality object sample process, the proposed system creates the environment through the computer vision that the therapist uses to follow the therapy.

Also, provide the possibility of performing the test before and after each simulation, so that the therapist by graphs of the results can see in what degree of fear or anxiety the patient is provoked by the therapy. The system seeks to be a tool to help therapists with the treatment of phobias.

6. BACKGROUND

The concept of virtual reality arose in 1965, when in his article Ivan Sutherland says that The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real. He would be the creator of the first virtual reality helmet visor using cathode ray tubes (one for each eye) and a mechanical tracking system.

Later in 1968 together with David Evans they will create the first generator of scenarios with three-dimensional images, stored data and accelerators. A year later, in 1969, Myron Krueger created the so-called Artificial Reality that allowed interaction with virtually created elements.

The use of graphics through the computer had to wait still some time, and is due to the work done in the MIT by Roberts and Sutherland. Roberts wrote the first algorithm to remove dark and hidden surfaces from an image, thus opening the way to the use of 3D graphics. Sutherland's work consisted in the development of algorithms that could perform this task efficiently. One of the fruits of these efforts lies in the development by Henri Gouraud, in 1971, of a lighting algorithm that is still widely used today. This algorithm makes it possible for a surface formed by polygons to cover the appearance of a smooth, continuous surface.

At first the field in which it had a greater application was the military, in fact, in 1971, in the United Kingdom they begin to manufacture flight simulators with graphic displays, but it will be a year later, in 1972, when General Electric develops the First computerized flight simulator. These operated in real time, although the graphics were quite primitive. And a few years later in 1979, the military began experimenting with simulation helmets.

In 1977, appears one of the first documented gloves Sayre Glove, developed by Tom Defanti and Daniel Sandin. This glove, based on an idea by Richard Sayre, had on each finger a flexible fiber optic tube with a light emitter at one end and a receiver at the other. Depending on the amount of light reaching the receiver, flexion of the fingers could be calculated.

In the early 80's Virtual Reality is recognized as a viable technology. In these years Andy Lippman along with a group of researchers developed the first virtual interactive map of the city of Aspen, Colorado. The recording was made by means of four cameras, taking a photo every three meters and reproducing them at 30 frames per second, simulating a speed of 330 km/h which would later be reduced to 110 km/h.

In 1981, Thomas Furnes developed the Virtual Booth. It was the first cockpit simulator to train pilots. The initial problem consisted in the increasing complexity of the cabs of these devices, reason why Furness began to look for the form to facilitate the interaction with the pilots.

The solution was the development of a booth that provided 3D information to the pilots, who could control the device through a virtual representation of the terrain with field of view of 120 in horizontal. They first started this apparatus in September 1981 and have formed the basis for the development of military training systems created from that moment on. This same scientist only one year later presented the most advanced flight simulator that exists, contained in its entirety in a helmet and created for U.S Army AirForce.

In 1983, Dr. Gary Grimes of the Bell Labs patented the first glove that recognized the positions of the hand with the intention of creating alphanumeric characters and being able to replace the keyboards by these. This glove had bending sensors on the fingers, tactile sensors on the fingertips and sensors for positioning and positioning on the wrist.

In 1985, Mike McGreevy and Jim Humphries along with NASA developed the Vived system (Visual Environment Display system), the first low-cost stations equipped with a wide, stereo field of view with position sensors on the RV hull; Whose usefulness was focused on future astronauts at NASA. And the first practical system of stereoscopic displays will also be built.

In 1987, Tom Zimmerman developed the first RV glove to be marketed.

In 1988, Dr. Davidson works in the production of low cost displays.

In 1989, VPL, and then Autodesk demonstrate their complex VR systems, which were too expensive.

In 1990, the first commercial software company VR, Sense8, founded by Pat Gelband, emerged.

In 1991, the W. Industries company developed the Virtuality, and installed in the US recreational rooms. The team included helmets and eyeglasses.

In this same period will appear numerous models emulating cockpits of flight or conduction. In the same year, the first program aimed at users for the creation of 3D virtual environments comes out. In the same year, a TVE children's program called "The Rescue of the Talisman" was launched, in which the contestants had to guide a blindfolded partner through virtual scenarios.

In 1992, SUN made the first demonstration of its Visual Portal, the highest resolution VR environment so far. In May of this same year, leaves the first game whose perspective of the graphics was in first person Wolfenstein.

In 1995, Nintendo released the first virtual reality console called Virtual Boy whose graphics were in 3D in red and black. However it did not have a commercial success, was too big and fragile, the continuous use during several minutes could produce headache.

The Electronic Visualization Lab (EVL) of the University of Illinois, Chicago, created in 1992, the concept of a room with graphics projected from behind the walls and floor, appearing CAVE (Cave Automatic Virtual Environment).

In 1993, Silicon Graphics (SGI) announced a Virtual Reality engine. In 1994, Antena 3 is the first Spanish television channel to introduce virtual spaces in its programs.

In 1995, the first formulation of the Virtual Reality Modeling Language (VRML) appears.

In 1997, the US Army's STRICOM developed a device that allows you to walk, run and move in a small space in all directions, allowing you to experience the real movement in a cabin.

After 2000, it was Google who laid the first stone of virtual reality content, almost unintentionally, with its Street View. What claimed to be the most complete road photo on the planet, allowed users to move up and down, left and right, with their mouse for the images.

In 2003, he created the famous virtual world in 3D Second Life where through a program pc, the users or residents, can move by him, interact, modify his environment and participate in his economy.

In 2004, Google purchased Earthview, a program developed in 2001, to create Google Earth, a representation of the world that combines Google's search power with satellite imagery, maps, terrain and 3D buildings.

In 2005, Nintendo announced the launch of Nintendo's WII, (codenamed Revolution) the videoconsole that was born with the idea of getting an interaction never before experienced in a game console between the player and the video game. Just as Virtual Boy was a failure, WII to date has been a resounding success.

In 2010, a young man named Palmer Luckey, stuck in recovering the old dream, began working on a virtual reality helmet convinced that the technology was the right one. Designed the first version of Oculus Rift, which achieved a vision angle of 90 degrees, something never seen before.

In 2014, the success and media outreach of Oculus Rift has driven dozens of projects, applications and video games.

In 2015, Samsung introduced the Gear VR, a helmet-shaped accessory that mounts on their smartphones.

In 2016, our head tracking systems, gloves or specific virtual reality controls to interact with our hands.

In 2017, the year will be key to demonstrate if consumers embrace virtual reality. Canalys has made the first estimate: this year will sell approximately two million units of high-end glasses, a figure that will grow to reach 20 million devices in 2020.

7. LATEST DEVELOPMENTS

Just twenty years ago, if the subject of virtual reality had been raised as an object of study, everyone would have referred to it as if it were a matter of science fiction. However, nowadays, virtual worlds have an increasingly real presence in our lives.

The latest inventions such as HUVR or a System to combat smoking force us to think the relations between the real and the virtual from a new perspective.

7.1 HUVR

HUVR is a device created by engineers and specialists from the University of California. The machine develops an image that can be touched by users as if it were a real object, in addition to obtaining these results with a low cost of production.

The HUVR combines a high-definition 3D panel, a semi-mirrored mirror and a touch-sensitive controller, allowing users to literally touch an image generated as if it were a conventional plastic object with the three dimensions of the physical world we know.

Researchers at the University of California at San Diego point out in particular that this new RV device has a relatively low cost of production, allowing development for applications in different fields. The device could be used, for example, to visualize and manipulate a 3D image of the brain of a person taken through an MRI. In the same sense, it could be applied in the management or study of very fragile or valuable artifacts, whose physical manipulation can become dangerous.

As explained by the engineers responsible for this breakthrough, by using HUVR devices, a doctor could actually feel a flaw in the brain, rather than just watch it. Logically, the implications at this point are really unimaginable in terms of the changes that could arise in this type of activities.

7.2 VR Against Tobacco

As for the latest studies on VR, these focus on using virtual reality as a therapeutic tool, such as against smoking.

Studies published so far, regarding the use of VR in the approach to smoking, have provided reasonably positive expectations about its usefulness in the treatment of this addiction. However, most of them have not yet advanced to the stage of using these techniques as a therapeutic strategy, but are in the previous stages of validation, testing the ability of virtual environments to provoke and reduce withdrawal syndrome the patients.

After determining which specific situations or contexts produce withdrawal syndrome in smokers through a survey created for this research, it has been decided to elaborate about 6 or 7 environments that will be presented to the patients of the experimental group. Studies published so far have only validated the ability to produce abstinence syndrome from a type of virtual environment, usually a virtual bar. The environments created try to be as real as possible including interacting with virtual characters, even presenting classic risk situations like offering tobacco or seeing people smoking.

7.3 Phobia Free

Phobia Free is an application specifically designed to fight phobias to animals, specifically for arachnophobia, or phobia. It is based on systematic desensitization, an approach technique for phobias that is widely used in consultation in which the patient is progressively exposed to the phobic element.

7.4 Spider phobia cardboard

This application uses virtual reality and is designed to help people suffering from arachnophobia. Participants enter a virtual office with small spiders. The intensity of the presence of spiders increases gradually. For example, at first the spider is in a jar and then the spider is out of it. The app has been developed jointly with a group of psychologists.

7.5 Itsy

Itsy is a VR program designed for people suffering from arachnophobia to perform different tasks in a virtual world inhabited by an immense spider that stalks them every moment.

8. THEORETICAL FRAMEWORK

8.1 Virtual Reality(VR)

There are countless definitions because many researchers and companies have focused their work on it, several of which are detailed below:

"Virtual reality is a system that interacts with the user simulating a real environment in a fictitious environment" (Hctor Lpez Pombo, 2009, p.25)

"It is a three-dimensional, interactive and computer generated environment in which a person is immersed." (David C. Prez Lpez, 2009, p.20)

"Virtual reality: a computer system used to create an artificial world in which the user has the impression of being and the ability to navigate and manipulate objects in it." (Manetta C. and R. Blade, 1995)

"Virtual reality lets you explore a computer-generated world through your presence in it." (Hodder and Stoughton, s/a).

"Virtual reality is a way for humans to visualize, manipulate and interact with computers and with extremely complex information." (Aukstakalnis, 1992)

This technology generates a sensation for the person who is immersed in it, that what is happening is real, even if it is not, it is a computer generated technique that through special processes and non-visual modalities, such as auditory, Tactile, among others, also allows vertical and horizontal movements; Providing absolute freedom of movement and a great sense of synthetic realism to the user through the stimulation of the five senses.

Although, without demeaning its high level of realism and however detailed it is to perform the environment in a computerized way, the result of this does not yet match the reality.

In addition, Juan Capafons (2001) mentions that virtual reality makes the subject feel immersed in it and is not just a mere spectator, making him / her participate or interact with it, besides adding that most users indicate that When they make use of the virtual reality they consider that the generated environment is less credible than a filming, although this one has a low quality.

8.2 Immersion

It is the abstraction capability of the actual environment in which the system user is located. In the RV we try to do this by stimulating the senses so that the user feels to be within the new reality.

8.3 Presence

So that the user can interact within the RV must be able to be within it. So it becomes a fundamental feature to be present within the system and this is achieved through different input devices. (Motion sensors, dimensions, gloves, etc.).

8.4 Interactivity

The VR system is not passive, so to be able to perform actions in the system that modify it and the user obtain the answers through their senses. If this feature were removed, it would simply be to watch a movie in the front row of the cinema, perhaps with better effects than in the movies.

8.5 Specific Phobias and VR

The application of VR in the field of psychological treatments has been especially fruitful in phobias. This new

technology has proven effective in the treatment of several specific phobias such as phobia to fly, claustrophobia, insect phobia, phobia to drive and arachnophobia that will be the nucleus of this thesis.

The essential characteristics of specific phobias are fear and avoidance, related to a specific object or situation that produces significant discomfort and / or interference.

8.6 Arachnophobia

Arachnophobia is the disgust or irrational phobia of spiders. It is one of the most common phobias, and possibly the most widespread animal phobia. The reactions of arachnophobes often seem irrational to other people. They try to stay away from any place where they think spiders inhabit, or where they have observed spider webs.

8.7 Unity

Unity is a multiplatform video game engine created by Unity Technologies. Unity is available as a development platform for Microsoft Windows, OS X. The development platform has compile support with different types of platforms.

9. METHODOLOGY

For the realization of the present document we opted to carry out an investigation that has the purpose of obtaining to solve the problem raised previously.

9.1 Population

To perform the tests will take a population of 20 people who will be known to have the problem of phobia to treat, 10 of them will make up the number one team, to take the therapy in a traditional way carrying out the corresponding sessions, the other 10 People will be team number two, who will be treated with the simulation system. All this will take place within 30 days.

9.2 Instruments for therapy

For the number one team, they will have the support of a psychologist to carry out the therapy, the number two team will have to use a virtual reality glasses that in it will be placed a mobile device with the system loaded in it, besides a headset for Feel better.

9.3 Collect data

The psychologist will record the progress of his 10 patients in separate files to be able to compare. People who will receive therapy through the virtual reality glasses can answer a test that will be a self-assessment, which will help you to move forward while experiencing a new scenario. The information generated by the responses of the test will help to compare with the results of the people who took the therapy with the psychologist, and will serve to compare advances.

10. EXPERIMENTAL SECTION

After the stipulated time, the results of the number one team, which were those who underwent gradual in vivo exposure treatment, were discontinued, and 70 percent of the patients refused to continue treatment, before moving on to physical exposure.

In the case of team number two, only 10 percent of people decided not to continue treatment due to discomfort during sessions with virtual reality glasses, 90 percent of people in team number two, continued the treatment until finished , And when passing the time of entering the exhibition "in vivo", controlled their level of stress.

People were more willing to receive the virtual treatment before the exhibition "in vivo", because they knew that everything would be unreal, but in turn as close as possible to a direct situation.

11. CONCLUSION

As a conclusion we can say that, virtual reality increasingly helps us regain confidence in us and facilitates the experience of reaching many places without needing to leave home. A how expository therapy facilitates us without having to be in danger of approaching at all that which generates fear.

12. REFERENCE

Carlin, A. S., Hoffman, H. G., and Weghorst, S. (1997). Virtual reality and tactile augmentation in the treatment of spider phobia: a case report. Behaviour research and therapy, 35(2), 153-158.

Garcia-Palacios, A., Hoffman, H., Carlin, A., Furness, T. U., and Botella, C. (2002). Virtual reality in the treatment of spider phobia: a controlled study. Behaviour research and therapy, 40(9), 983-993.

Hoffman, H. G., Garcia-Palacios, A., Carlin, A., Furness Iii, T. A., and Botella-Arbona, C. (2003). Interfaces that heal: coupling real and virtual objects to treat spider phobia. International Journal of Human-Computer Interaction, 16(2), 283-300.

Paquette, V., Levesque, J., Mensour, B., Leroux, J. M., Beaudoin, G., Bourgouin, P., and Beauregard, M. (2003). Change the mind and you change the brain: effects of cognitive-behavioral therapy on the neural correlates of spider phobia. Neuroimage, 18(2), 401-409.

Parsons, T. D., and Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. Journal of behavior therapy and experimental psychiatry, 39(3), 250-261.

Leutgeb, V., Schafer, A., and Schienle, A. (2009). An event-related potential study on exposure therapy for patients suffering from spider phobia. Biological psychology, 82(3), 293-300.

Garcia-Palacios, A., Botella, C., Hoffman, H., and Fabregat, S. (2007). Comparing acceptance and refusal rates of virtual reality exposure vs. in vivo exposure by patients with specific phobias. Cyberpsychology and behavior, 10(5), 722-724.

Garcia-Palacios, A., Hoffman, H. G., Kwong See, S., Tsai, A., and Botella, C. (2001). Redefining therapeutic success with virtual reality exposure therapy. CyberPsychology and Behavior, 4(3), 341-348.

Powers, M. B., and Emmelkamp, P. M. (2008). Virtual reality exposure therapy for anxiety disorders: A meta-analysis. Journal of anxiety disorders, 22(3), 561-569.

Maldonado, J. G. (2002). Aplicaciones de la realidad virtual en psicologa clnica. Aula mdica psiquiatra, 4(2), 92-126.

Orgils, M., Rosa, A. I., Santacruz, I., Mndez, X., Olivares, J., and Snchez-Meca, J. (2002). Tratamientos psicolgicos bien establecidos y de elevada eficacia: terapia de conducta para las fobias especficas. Psicologa Conductual, 10(3), 481-502.

Arbona, C. B., Garca-Palacios, A., y Baos, R. M. (2007). Realidad virtual y tratamientos psicolgicos. Cuadernos de medicina psicosomtica y psiquiatra de enlace,(82), 17-31.

Jayaram, S., Connacher, H. I., and Lyons, K. W. (1997). Virtual assembly using virtual reality techniques. Computer-Aided Design, 29(8), 575-584.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. Journal of communication, 42(4), 73-93.

Arquitectura de un sistema multi-agente para control de inventarios ⋆

Firs A. Author L.I. Maribel Durán Salas
Second B. Author, Jr.MSC Galdino Martínez Flores

[*] *Instiruto Tecnologíco Superior de Misantla,Veracruz C.p. 93821 MEXICO (152t0733@itsm.edu.mx).*
[**] *Instiruto Tecnologíco Superior de Misantla,Veracruz C.p. 93821 MEXICO (gmartinezf@itsm.edu.mx)*

ResumenEn la actualidad hacer uso creativo de las TI para llevar a cabo todo el proceso de control de inventario, da como resultado la generación de datos veraces, confiables y oportunos, para que los que están al frente de la empresa puedan hacer mejores tomas de decisiones.
La sistematización de procesos nos facilita el trabajo manual, es decir, las tareas se realizan en menos tiempo. Los niveles de insumos estarán siempre en el stock que así se decida. Los informes estarán en cualquier momento con menos esfuerzo, los datos estarán disponibles de tal manera que a cualquier hora y el recurso no se optimizara al máximo, etc.

Keywords: TI Tecnologías de Información

1. INTRODUCCIÓN

El conocimiento de la estructura de los almacenes y los insumos que ahí se manejan; da a los encargados de las áreas e instituciones una ventaja sustancial frente a aquellas que no lo poseen. Es decir, tener un amplio conocimiento sobre los datos de todos los insumos como los son: descripción del producto, unidad de medidas, lotes, fechas de caducidades, numero de programa al cual pertenecen y existencias reales.

Todo el conocimiento anterior sobre los que se maneja en las áreas de almacenaje, trae consigo varias consecuencias: las unidades de salud pueden reducir tiempos para el desplazamiento de sus recursos entre sus almacenes, Mejora de surtimiento, realizar solicitudes de pedidos en tiempo y forma, mantener sus niveles de stock, eliminar el riego de caducidad de los productos, todo esto mejoraría la atención al usuario final.

Realizar los procesos anteriormente descritos para llevar el correcto orden y administración del inventario, no es una tarea fácil cuando se trata de realizarlo de forma manual, en cambio, si nos apoyamos en el uso de la tecnología podremos alcanzar mejores resultados, ya que por ejemplo: si hacemos uso de arquitectura de sistemas se podrán realizar el procesamiento de la información de manera rápida y segura.

⋆ Sponsor and financial support acknowledgment goes here. Paper titles should be written in uppercase and lowercase letters, not all uppercase.

2. SISTEMA MULTI-AGENTE Y MODELADO DE DATOS

2.1 Sistemas Multi-agentes

Una de las temáticas que en la actualidad ha dado solución a diferentes problemas son las arquitecturas de software que mediante tecnologías de la ingeniería de sistemas, enmarcan métodos y algoritmos apropiados que se utilizan para el desarrollo de aplicaciones, por ejemplo el desarrollo de sistemas multi agentes (SMA). La característica principal de los SMA, es que están compuestos por múltiples agentes inteligentes que interactúan entre ellos y realizan una o varias tareas específicas.

Modelado de agentes Para el diseño de un agente nos apoyamos en la Inteligencia artificial (IA), que se divide de dos maneras: la primera nos dice que los agentes realizan procesos metales y al razonamiento, es decir, sistemas que razonan y actúan como humanos al ser entrenados para realizar ciertas tareas; la segunda dice, se tiene que entrenar a los agentes a que piensen racionalmente, como por ejemplo que puedan percibir sus ambiente(recibir entradas en cualquier momento), razonar y actuar de manera inteligente. (Russell y Norvig, 2004). (**?**)

2.2 Modelado de Base de Datos

1. Modelo de los Datos Las BD pueden ser diseñadas bajo los siguientes modelos de datos, como son: el modelo Entidad-Relación, el cual está basado en una percepción del mundo real que consta de una colección de objetos básicos, llamados entidades, y de relaciones entre estos objetos; el Modelo Relacional, se utiliza un grupo de tablas para representar los datos y las relaciones entre ellos. Cada tabla está compuesta por varias columnas, y cada columna tiene un nombre único, etc.,

2. Usuarios y Administradores En una BD se tiene que definir cuáles son las personas que estarán trabajando, a las cuales se les llamara como usuarios y administradores.

3. Estructura de un sistema de BD Para tener una estructura de BD, se debe de contar con un gestor de almacenamiento, el cual es responsable de la interacción con el gestor de archivos. Los datos en bruto se almacenan en disco usando un sistema de archivos, que está disponible habitualmente en un sistema operativo convencional. El gestor de almacenamiento traduce las diferentes instrucciones LMD a órdenes de un sistema de archivos de bajo nivel. Así, este gestor es responsable del almacenamiento, recuperación y actualización de los datos en la base de datos, ésta compuesto por gestor de seguridad, archivos y memoria intermedia; y un procesador de consultas.

4. Arquitectura de aplicaciones Con relación a su arquitectura de aplicaciones, se puede utilizar la de dos capas, ésta aplicación se divide en un componente que reside en la máquina cliente, que lama a la funcionalidad del sistema de bases de datos en la máquina servidor mediante instrucciones del lenguaje de consultas, y por ultimo tenemos la arquitectura de tres capas, donde la máquina cliente actúa simplemente como frontal y no contiene ninguna llamada directa a la base de datos. En su lugar, el cliente se comunica con un servidor de aplicaciones, usualmente mediante una interfaz de formularios. El servidor de aplicaciones, a su vez, se comunica con el sistema de bases de datos para acceder a los datos.

3. ARQUITECTURA DEL SISTEMA MULTI-AGENTE PARA CONTROL DE INVENTARIO

Una de las ventajas del modelo MVC es que englobamos en la etapa controlador gran parte de las funciones que tiene por objetivo la solución a la problemática de inventarios en los almacenes de la SS, y este sistema no requiere de modificaciones constantes a las capas para su funcionamiento, por esta razón se usara dicha arquitectura, la cual tendrá agentes que estarán interactuando entre sí, para desarrollar actividades asignadas y por ultimo diseñar bases de datos seguras para el resguardo de la información **?**

Arquitectura general utilizando Modelo-Vista-Controlador, ver fig, 1:

1. Etapa Modelo.- Es aquí donde tendremos la información que se manera, así como, los accesos y consultas a dicha información. 2. Etapa Vistas.- Nos referimos al prototipo de la arquitectura del sistema, es decir, la parte con la cual estará interactuando el usuario final, teniendo en cuenta realizar un diseño amigable y fácil de manejar. 3. Etapa Controlador.- En esta última etapa, se dice que son eventos que el usuario solicita, interactuando con la etapa modelo y vistas.

4. METODOLOGÍA A UTILIZAR

La metodología que se utilizará para probar el buen funcionamiento de la arquitectura general será, la metodología de Desarrollo Rápido de Aplicaciones (DRA). Tenemos que DRA se divide en cuatro fases:

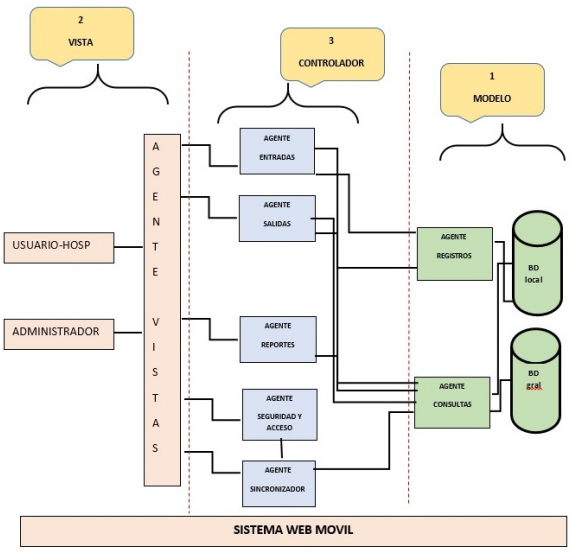


Figura 1. Arquitectura principal

Descripción del documento técnico	Descripción detallada	Herramienta Case	Observaciones	Firmas de aceptación
Sistema a Desarrollar	Control de Inventario de material de curación y medicamentos	No	-Bajo normas del manual de procedimientos de almacenes	- Responsable de la organización
Areas que abarcara el sistema	Deptos de Almacén farmacia, laboratorio y Administración de la unidad	No	-Bajo normas del manual de procedimientos de almacenes	- Responsable de la organización
Atributos del sistema	- Automatización del control de inventario mediante un sistema web-móvil. - Sistema amigable, fácil de manejar - El sistema web-móvil con alta seguridad	No	-Bajo normas del manual de procedimientos de almacenes	- Responsable de la organización
Funciones del sistema	- Registro de movimientos de Entradas y Salidas de Insumos - Reportes automatizados de informes a otras instituciones - Control de lotes de insumos por caducar (lento o nulo movimientos)	No	-Bajo normas del manual de procedimientos de almacenes	- Responsable de la organización
Especificaciones	- Catálogo de material de curación - Catálogo de Medicamentos - Catálogo de unidades del sector salud	No	-Bajo normas del manual de procedimientos de almacenes	- Responsable de la organización

Figura 2. Documentación técnica de la primera fase

4.1 Fase de Planeación

Fase de Planeación de requerimientos o fase de definición conceptual: Aquí se definen las funciones de la secretaria de salud, describe las características del software, las áreas de influencia del software y el alcance del mismo, como se muestra en la Fig. 2.

Descripción del documento técnico	Descripción detallada	Herramienta Case	Observaciones	Firmas de aceptación
Modelado de procesos	- Registro de Entradas de Insumos - Salidas de Insumos - Generación de Reportes	Si	MySQL	Responsable de la organización
Modelado de datos	- Datos de Insumos (clave, descripción, cantidad, precio, lote, fecha/cad) - Datos para generación de reportes de existencias (Nombre/unidad, fecha, no. De reporte)	Si	MySQL, PHp, HTML	Responsable de la organización
Diseño de prototipo	- Pantallas para usuarios, captura de entradas y salidas de insumos. - Pantallas de reportes	Si	MySQL, PHp, HTML	Responsable de la organización

Figura 3. Diseño Conceptual

Descripción del documento técnico	Descripción detallada	Herramienta Case	Observaciones	Firmas de aceptación
Pruebas del sistema	- Al desarrollar de sistema se tienen que someterlo a pruebas, para determinar si su función cumple con lo esperado	Si	MySQL, PHp,Java HTML	Responsable de la organización
Ayuda	- Manuales para el usuario	Si	Ayuda del sistema, manual de usuario	Responsable de la organización

Figura 4. Constucción o Desarrollo

4.2 Fase de Diseño Conceptual

Es el diseño funcional del sistema. Tenemos el modelado de procesos, como lo es la recopilación de datos, modelado de datos, que involucra datos específicos con los cuales se estará trabajando y por último el diseño del prototipo, que incluyen pantallas para que el usuario pueda validar el proceso, como se muestra en la Fig. 3.

4.3 Fase de Construcción o Desarrollo

Aquí se contempla el desarrollo del sistema, las iteraciones van arrojando componentes y se realizan pruebas de integración de los mismos de acuerdo a planes de trabajo establecido, como se muestra en la Fig. 4.

4.4 Fase de Implementación

Como se ha venido mencionando anteriormente el usuario final es el eje de este modelo, por lo tanto él mismo prueba el software y proporciona su conformidad, véase Fig. 5

5. HEURÍSTICA DE LA METODOLOGÍA

Los Servicios de Salud de Veracruz, como toda organización requieren de una constante supervisión de los Recursos Materiales que están a disposición de los almacenes de las unidades médicas. En busca de una excelencia en control de inventarios, esta institución requiere de una

Descripción del documento técnico	Descripción detallada	Herramienta Case	Observaciones	Firmas de aceptación
Entrega	- Bajo acta se realizara la Entrega-Recepción del sistema	No		Firma de aceptación del responsable de la organización
Capacitación	- Realizar capacitación a todos los usuarios y administradores que manejaran el sistema	Si	Capacitación por grupos	Asistencial
Implementación	- Instalación del software en el servidor y nodos	No	Que los altos ejecutivos estén convencidos del éxito del sistema	

Figura 5. Implementación

automatización de sus procesos de Entradas y Salidas de sus insumos y demanda con mucha premura con tiempo relativamente cortos el desarrollo de un software, porque no se cuenta con un sistema que realice tales procesos. Las características del software solicitado son: Opera en un ambiente web móvil, El desarrollo se tiene que realizar en el menor posible, el diseño será modular. En la fig. 6 Se describe de manera general el flujo de cada proceso, empezando con la descripción del sistema, atributos y funciones, estos son validados por el usuario y diseñador, si hay cambios se realizan y se regresa al inicio del proceso actual de lo contrario se continua a los siguientes procesos que incluyen la validación, planeación, especificaciones del sistema y alcances, continuando con el modelado de datos, desarrollo de prototipo, así como Desarrollo del sistema y realización de pruebas y terminando con el proceso de Implementación de cierre y Entrega-Recepción del software.

5.1 References

Use Harvard style references (see at the end of this document). With LaTeX, you can process an external bibliography database using `bibtex`, [1] or insert it directly into the reference section. Footnotes should be avoided as far as possible. Please note that the references at the end of this document are in the preferred referencing style. Papers that have not been published should be cited as "unpublished". Capitalize only the first word in a paper title, except for proper nouns and element symbols.

6. CONCLUSION

En este trabajo se ha presentado un análisis y diseño de una arquitectura multi-agente para un sistema web móvil, en donde se da solución a la problemática del control de inventarios para la secretaria de salud en segundo nivel, está diseñada para almacenar y modificar datos de insumos que se manejan en los almacenes, así como generar reportes sobre lotes de caducidades, excedentes a las diferentes unidades de salud.

Por lo tanto esta aplicación web móvil representa una reducción en problemas administrativos para consultar y editar información de todos los hospitales del estado

[1] In this case you will also need the `ifacconf.bst` file, which is part of the `ifaconf` package.

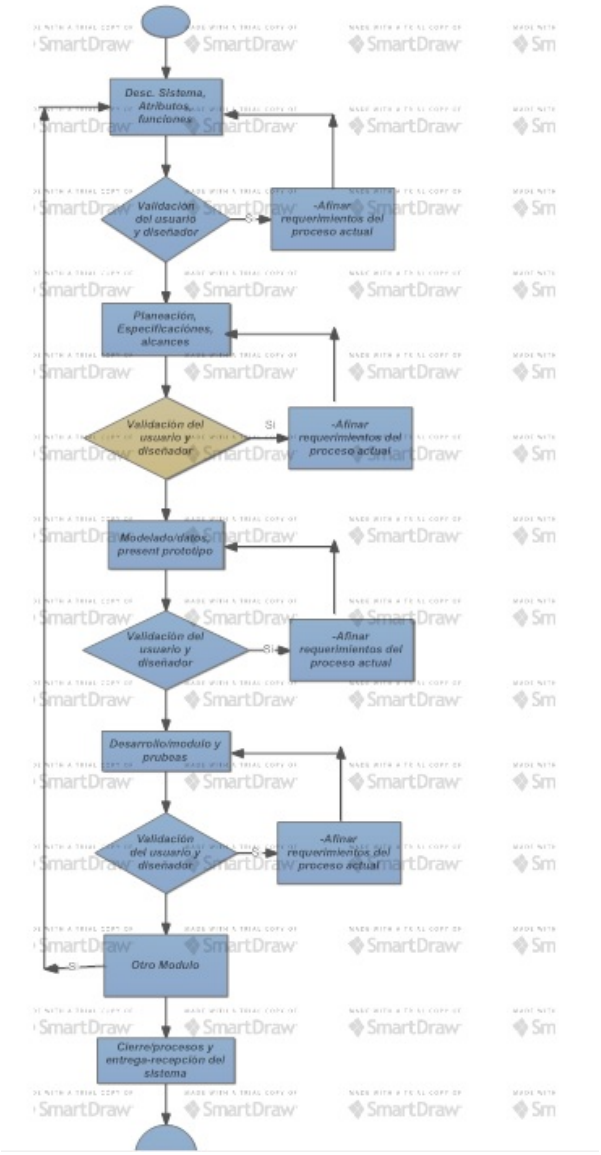


Figura 6. Heurística de la Metodología

de Veracruz, de segundo nivel desde cualquier dispositivo móvil.

Analysis of hyperspectral images for the detection of pests in coffee crops

Arely Guadalupe Sánchez Méndez [*]
Simón Pedro Arguijo Hernández [**]

[*] *Instituto Tecnológico Superior de Misantla, Veracruz, México (e-mail: 1062T0085@itsm.edu.mx).*
[**] *Instituto Tecnológico Superior de Misantla, Veracruz, México (e-mail: sparguijoh@itsm.edu.mx).*

Abstract
Coffee is one of the most commercialized agricultural products in international markets. Coffee production is an important source of employment, income and foreign exchange in many producing countries. In Mexico the organic coffee crop is affected by the rust Hemileia vastatrix. The high incidence severity occurs at uncontrollable levels of infection, severe defoliation, loss of production, weakness and death of coffee trees. Recently, several approaches based on automatic learning have built a dataset for the monitoring of the incidence of coffee rust, taking into account the climatic conditions and physical properties of the crops. In contrast to traditional methodologies, hyperspectral imaging technology emerges as a non-destructive quality assessment tool. Hyperspectral imaging (HSI) offers a high potential as a noninvasive diagnostic tool for the detection of diseases. In this context, the research is designed to develop a computer application for the use of hyperspectral imaging technology with the aim of helping to detect in time the infestation of coffee plantations by pests such as rust, providing farmers with information On the state of their lands and crops using the combination of hyperspectral images, pattern recognition and artificial intelligence.

Keywords: Hyperspectral imaging, plant disease, coffee crops, Hemileia vastatrix, remote sensing

1. INTRODUCTION

According to research carried out on the problems that coffee cultivation in Mexico has been presenting in recent years, it is considered that, besides other factors, plague such as rust are the main reasons why coffee production has Been decreasing. The rust is among the seven pests and / or diseases of plants that has left greater losses in the last 100 years, because it causes the premature fall of the leaves, causing the reduction of the photosynthetic capacity, as well as the weakening of Diseased trees, regressive death in branches and at one time the total death of the plant.

It can be said that the use of computer tools for the analysis of digital images, such as hyperspectral images, have shown great results in the implementation in the agricultural area to give solutions, since they allow the analysis from a small part of a crop , To the whole crop or more.

Using satellite imagery, such as hyperspectral images, of coffee crops in Veracruz, an analysis can be performed to assist in the early detection of pests, thus helping to avoid damaging any one or several crops that would seriously affect Many families in its economy, since the production of coffee is one of the main agricultural activities of the region.

On the other hand, the techniques used for the processing of the hyperspectral images, to realize the detection and classification of regions in the culture that are infested of some pest, by algorithms of recognition of patterns to determine the red points or points of greater incidence Of pests. These techniques are also considered to be implemented in some other work for the analysis of fruit and crop quality. That is why the computational importance of project implementation is feasible in many aspects, in addition to helping in the detection of defined problems, can be implemented in similar and undetected problems.

This article mentions the motivation to carry out the research and also shows the analysis and discussion of the main methodological aspects carried out in various research related to the analysis of images focused on the area of agriculture.

2. PROBLEM STATEMENT

The development of technically advanced tools for use in agriculture is essential to help farmers make decisions on crop health and resource management. Effective use of hyperspectral images in agriculture not only helps farmers to effectively use resources such as herbicides and pesticides, but also provides insight into the current stage of crop development and health.

A typical hyperspectral image is composed of a set of a relatively wide range of monochromatic images corresponding to continuous wavelengths which normally contain redundant information or may exhibit a high degree of correlation. In addition, the computation of the classifiers used to treat the data obtained from the images may become excessively complex and slow for these high-dimensional data sets, making it difficult to incorporate such systems into an industry that requires standard protocols Or high in the speed of the processes. Therefore, recent work has focused on the development of new systems based on this technology that are capable of analyzing crop characteristics that can not be inspected by visible images.

In Mexico the organic coffee crop is affected by the rust Hemileia vastatrix. The high severity of disease results in uncontrollable levels of infection, severe defoliation, loss of production, weakness and death of coffee trees.

Coffee is produced in Latin America, Africa and Asia, and is one of the most traded agricultural products in international markets. The coffee agro-industry has diversified throughout the world and is an important source of employment, income and foreign exchange in many producing countries. In recent years, its global supply has been affected by adverse climatological factors and pests such as rust, which has been reflected in the high volatility of international prices of this product.

The implementation of this project aims to help detect in time the infestation of coffee plantations by pests such as rust, providing farmers with information on the state of their lands and crops using the combination of hyperspectral images, pattern recognition and artificial intelligence .

2.1 Proposed Solution

Make use of high spectral and spatial resolution spectroscopic images, applying image processing techniques and pattern recognition (computational vision) in coffee crops with the intention of obtaining early specific detection of pests in which they can put the production of Coffee and the safety of its harvest.

3. PREVIOUS WORKS

In this chapter the research will focus on the analysis and processing of hyperspectral images for the detection of diseases and pests in crops and / or fruits, since it has been demonstrated that the analysis of hyperspectral images represents a great help in the field Of agriculture, because significant results are obtained compared to the analysis of digital images that show only the visible range of the electromagnetic spectrum. It is important to mention that in the investigations analyzed, they are not all related to pests and diseases in coffee crops, although, for the most part, they are focused on these; Those who do not focus on it, are considered due to the techniques used in the analysis of hyperspectral images, however, also consider the analysis of fruit quality.

Mahlein *et al.*, related leaf characteristics and spectral reflectance of leaf beet leaves diseased with Cercospora leaf spot, powdery mildew and leaf rust at different stages of development, used a spectroscope for the exploration of hyperspectral images (ImSpector V10E) With a spectral resolution of 2.8 nm from 400 to 1000 nm and a spatial resolution of 0.19 mm for the continuous detection and monitoring of disease symptoms during pathogenesis. They do not mention quantitative results, but if they concluded that the spectral reflectance in combination with the classification of the spectral angle mapping allowed the differentiation of the mature symptoms in zones that show all the ontogenetic stages from young to mature symptoms (Mahlein et al., 2012).

Luaces *et al.*, studied the fact of learning functions that could predict whether the value of a continuous target variable can be greater than a given threshold. The objective of the application that they studied was to alert about the high incidence of coffee rust, the main disease of coffee cultivation in the world. They make a comparison between the results of their confusion matrix, obtaining results where the costs of false negatives are higher than those of false positives, and both are higher than the cost of warning predictions (Oscar Luaces, 2011).

Huang *et al.*, detected in the wheat fields in winter seasons, that these are affected by the disease called yellow rust, which harms the wheat production, therefore, the objective of their study was To evaluate the accuracy of the optical spectrum, the index of photochemical reflectance (PRI) to quantify the index of this disease and its applicability in the detection of the disease by hyperspectral images. They performed PRI tests in three seasons, showing that in winter, with a determination rate of 97 %, PRI's potential is clear for quantifying yellow oxide levels in wheat and as a basis for the development of a Proximal image sensor of yellow rust in wheat fields in winter (Wenjiang Huang, 2007).

Devadas *et al.*, Evaluated ten widely used vegetation indices based on mathematical combinations of narrowband optical reflectance measurements in the range of visible and near infrared wavelengths for their ability to discriminate leaves from wheat plants Of one month infected with yellow stripes. They do not mention quantifiable results but conclude that no individual index was able to discriminate the three oxide species among themselves, however, the sequential application of the Reflectance Index Anthocyanin to separate the healthy, yellow and mixed rust and rust classes of the leaves followed By the absorption index of chlorophyll and the reflectance index to separate the classes of rust from leaves and stems, could constitute the basis of the discrimination of the oxide species in the wheat in conditions of field (R. Devadas and Backhouse, 2009).

Stefan *et al.*, Observed the plant-pathogen interaction through simultaneous measurements of reflection and transmission of hyperspectral images. These data were statistically analyzed using principal component analysis and compared with the estimation of visual and molecular disease, concluding that measurements based on reflectance facilitate early detection, and transmission measurements provide additional information to better understand and quantify The complex spatio-temporal dynamics of plant-pathogen interactions (Thomas Stefan and Anne-Katrin, 2016).

4. CONCLUSION

After carrying out the search for information related to the subject of the research and to have a better knowledge of the existing works, the next step is to obtain hyperspectral images of the area corresponding to the coffee crop to perform the image processing and analysis . According to this, it will be necessary to define the type of methodology to be used, considering mainly the process in which the following activities, mentioned in an orderly manner, are carried out: reducing the dimensionality of the images, removing information of little relevance, determining an algorithm suitable for extraction Of image characteristics, detection of problematic points in coffee cultivation, labeling of detected areas, classification of characteristics obtained, detection of crop quality, spectral decomposition, determination of pests in the analyzed crop and finally Must perform the analysis and validation of results obtained. In this last point should be enhanced the obtaining of a matrix of confusion and thus be able to compare the results obtained with those of the previously consulted works.

REFERENCES

Mahlein, A.K., Steiner, U., Hillnhütter, C., Dehne, H.W., and Oerke, E.C. (2012). Hyperspectral imaging for small-scale analysis of symptoms caused by different sugar beet diseases. *Plant Methods*, 8, 3.

Oscar Luaces, Luiz Henrique A. Rodrigues, C.A.A.M.A.B. (2011). Using nondeterministic learners to alert on coffee rust disease. *Expert Systems with Applications*, 38, 1427614283.

R. Devadas, D. W. Lamb, S.S. and Backhouse, D. (2009). Evaluating ten spectral vegetation indices for identifying rust infection in individual wheat leaves. *Precision Agriculture*, 10, 459470.

Thomas Stefan, Wahabzada Mirwaes, K.M.T.R.U. and Anne-Katrin, M. (2016). Observation of plantpathogen interaction by simultaneous hyperspectral imaging reflection and transmission measurements. *Functional Plant Biology*, 44, 23–34.

Wenjiang Huang, David W. Lamb, Z.N.Y.Z.L.L.J.W. (2007). Identification of yellow rust in wheat using in-situ spectral reflectance measurements and airborne hyperspectral imaging. *Precision Agriculture*, 8, 187197.

Architecture for the analysis of surface damage of citrus (orange and lemon) in the region of Misantla using image processing.

Juan Pablo Salazar * **Dr. Simn Pedro Arguijo Hernndez** **

* *Instituto Tecnolgico Superior de Misantla, Veracruz, CO 93821 Mxico (e-mail: 162t0084@itsm.edu.mx).*
** *Instituto Tecnolgico Superior de Misantla, Veracruz, CO 93821 Mxico (e-mail: sparguijoh@itsm.edu.mx)*

Abstract: This research work aims to develop an architecture for the identification of three common types of damage in lemon and orange, in the first section addresses the types of damages and their causes, also part of the problem and in general the way to solve it. In the state of the art are some techniques used in previous investigations to solve similar problems as curvelet transform, SVM, K-means among others and the final section provides the material and the procedures for the processing of images as segmentation and co-occurrence matrix for obtaining features and classification using neural networks for the identification of damage for diseases, chemical damage and mechanical damage.

Keywords: K-means, SVM, curvelet transform, segmentation, co-occurrence matrix, neural networks.

1. INTRODUCTION

The analysis and evaluation of the quality of great diversity of products in the food and agricultural industry derives from the need to meet the standards of consumers. Most of the classification process is carried out manually, however there is a high cost in terms of time and work with products for the detection of abnormalities. At present due to technological progress especially of systems of computer vision tools are able to carry out this work automatically. The implementation of classification algorithms and techniques of image processing and analysis seems to be the perfect solution in this area but is usually applicable in the processes of packaging and distribution of products. The producers of fruits and vegetables are facing adverse factors that are causing significant economic losses, mainly due to the damage that receives the product from before the harvest until the packaging process affecting the quality of the product. Fruits and even more citrus face pest problems, bacteria and viruses that attack the plantations directly affecting the final product, in addition to this, there are problems such as the misapplication or excess supply of chemicals such as fertilizers and/or herbicides and failures in the management in the packaging of the fruit; factors causing the damage that directly affect its quality.

The region of misantla and its surroundings have a lot of importance in the State of Veracruz and throughout the Mexican country in the production of citrus fruits like orange and lemon. One of the challenges faced by the producers of citrus fruit in this area include losses in production, caused by three types of harm which are classified in: 1. Damage for diseases. 2. Chemical damage. 3. Mechanical damage . These anomalies in citrus fruits can be detected using indicators that help to recognize different characteristics in the fruits. The features present in the shell, mainly spots defined by the color and texture found on its surface, are indicators that can be exploited via image processing and analysis for application in the detection and classification of the three types of damage.

⋆ Thanks to the Instituto Tecnolgico Superior de Misantla and to the national program CONACYT for the funding provided to be able to carry out this research.

2. STATE OF THE ART

At present the evaluation and quality control is essential to decide the price of the fruit. These must meet certain quality requirements of both external and internal to decide the time of harvest and consumption. The color and texture even without being direct indicators of maturity are parameters that are important to determine the quality of the fruit and the degree of consumer acceptance.

Khoje *et al.*, (2013) identified that one of the problems facing the automatic visual inspection is the variation of the characteristics of biological products as the shape, color, size, and the environment, as a result, the detection of damage and bruising in the shell parameters are considered to be of vital importance to its evaluation of a more practical way. The objective is the detection of defects of images of fruit by means of computer vision using textural characteristics as the entropy, low statistical characteristics such as the mean and standard deviation based on the transformed curves for their calculation and combine it with classifiers such as Support Vector Machine (SVM) and probabilistic neural network (PNN), where the recognition of healthy fruits and faulty, when applying the

classifier SVM, concludes that the method promises the 96 % accuracy [1].

Dubey *et al.*, (2014) consider that in the whole world diseases in fruits are devastating losses in the economy of the agricultural industry; due to the complexity to identify these diseases. The experimentation and validation using an adaptive approach is considered to be a good solution, taking advantage of the image processing. The method is to perform the classification by the technique of K-Means for segmentation of defects, followed by the obtaining of features and finally the images are classified in one of the classes using a multiclass Support Vector Machine applied to three varieties of apple, where the experimental result provides an accuracy of 93 % [2].

Zhang *et al.*, (2014) consider that at the time to inspect quality of fruits and vegetables the cost in time and hard work becomes one of the most common problems. The objective of the latest techniques in computer vision tends to facilitate the automated inspection. A comparison of the latest advances in computer vision traditional multispectral and hyperspectral, provides an idea of how these systems are applicable to inspection of fruits and vegetables and replace the manual inspection by a rapid assessment, accurate, objective and non-destructive manner [3].

Li *et al.*, (2013) mention that the detection of defects in the skin of fruits is difficult due to the spherical shape that have most of the fruit. In the case of the orange image acquisition of its surface for the detection of different defects, such as scars, openings, burns and cancerous becomes complex by the shape of the orange. The application of techniques as a transformation of the lighting and the method of radio images help the uniform distribution of lighting which makes it of great shape the image acquisition for processing. The result when applying these techniques contributes to an accuracy of 98.9 % in the detection range of 720 sample images [4].

Li *et al.*, (2011) propose the detection of several common defects of the orange through a system of hyperspectral images by taking advantage of the range of wavelengths from 550 to 900 nm. The method involves the use of high-dimensional spectral reflectance of images with the purpose of reducing them to optimal wavelengths forms of multispectral images by PCA (Principal Component Analysis) where the potentially optimal wavelengths were 691, 769 and 875 nm for its implementation in the detection of defects in the orange peel. The precision obtained using two lengths of radio(R875/R691) and PCA was an identification of the 93.7 % of defects on the surface of the orange without any false positive this [5].

Arivazhagan *et al.*, (2010) asserts that the use of computers for image analysis provides many applications for the automation of tasks; however the variability of forms of fruits and vegetables makes it very difficult for the adaptation of existing algorithms in the industry to the domain of agriculture. Through the combination of features such as color and texture can improve the functionality and flexibility of the systems for recognition of shape and size. The implementation of the minimum distance classifier combined with features of co-occurrence using a database of 2635 images of 15 kinds of fruit confirmed the effectiveness of this approach [6].

Devi and Vijararekha (2014) mention that the main problems facing the methods of manual classification are the work and processing time. To avoid these difficulties in the assessment and classification techniques are used in computer vision. The automatic machines employ methods of analysis and automatic inspection systems based on multispectral and hyperspectral images which enable the digital image processing [7].

3. MATERIAL AND TOOLS

To obtain features for testing purposes is performed using digital images of lemon and orange. The image capture is performed with a camera on a cell phone Sony Experia E3 of 13 Megapixels. The main tool for implementation of image processing techniques is a computer AZUZ with a 64-bit operating system and an Intel Core i3-2330M CPU and 2.20GHz. The platform for the collection of features is the programming language MatLab for segmentation and the implementation of Artificial Neural Networks.

4. METHODOLOGY

The identification of the types of damage will be done with image-processing techniques: Texture Analysis using matrix of co-occurrence and segmentation for the obtaining of the feature database and then implement the artificial neural network classifier to differentiate the type of damage. The process is shown in Figure 1 with the integration of each of the steps.

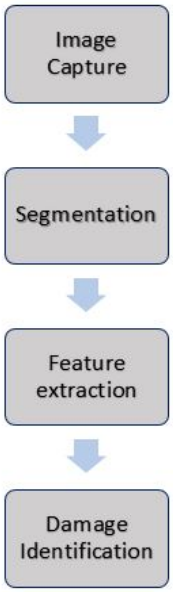


Fig. 1. Diagram of methodology for image processing.

4.1 Image Capture

The capture of the images will be developed in a controlled environment, determining a fixed distance between the camera and the fruits; in addition to a constant resolution to facilitate the processing of the fact that the camera comes from a mobile device.

4.2 Segmentation

Segmentation is the division of the image on groups of pixels, isolating regions or objects of interest in the image. The segmentation algorithms basically are carried out under two properties of the values of the gray level: discontinuity or similarity between the gray levels of the neighboring pixels; to obtain only the fruit to analyze and remove the background color of the image.

4.3 Feature extraction

Because an image contains objects, these can be characterized by lower levels of gray, color, texture, geometric properties, among other types of features, in this case the characteristics obtained will be the area of the fruit and the fund, the uniformity of the gray level between neighboring pixels, the histogram, the mean and the median calculated in the working color space.

4.4 Damage Identification

The identification of the three types of harm is done by means of the implementation of neural networks in order to decide the type of damage that presents the fruit according to the features extracted in the previous procedure.

ACKNOWLEDGEMENTS

I thank the Instituto Tecnolgico Superior de Misantla for allowing me to be a member of the postgraduate program in computer systems, also thanks to the national program CONACYT for the funding provided for the program and to be able to carry out this research.

REFERENCES

[1] Khoje, S. A., Bodhe, S. K., Adsul, A. (2013), *Automated skin defect identification system for fruit grading based on discrete curvelet transform*, International Journal of Engineering and Technology (IJET), 5(4), 3251-3256.

[2] Dubey, S. R., Jalal, A. S. (2014), *Adapted approach for fruit disease identification using images*, arXiv preprint arXiv:1405.4930.

[3] Zhang, B., Huang, W., Li, J., Zhao, C., Fan, S., Wu, J., Liu, C. (2014), *Principles, devel-opments and applications of computer vision for external quality inspection of fruits and vegetables: A review*, Food Research International, 62, 326-343.

[4] Li, J., Rao, X., Wang, F., Wu, W., Ying, Y. (2013), *Automatic detection of common sur-face defects on oranges using combined lighting transform and image ratio methods*, Post-harvest Biology and Technology, 82, 59-69.

[5] Li, J., Rao, X., Ying, Y. (2011), *Detection of common defects on oranges using hyper-spectral reflectance imaging*, Computers and Electronics in Agriculture, 78(1), 38-48.

[6] Arivazhagan, S., Shebiah, R. N., Nidhyanandhan, S. S., Ganesan, L. (2010), *Fruit recog-nition using color and texture features*, Journal of Emerging Trends in Computing and In-formation Sciences, 1(2), 90-94.

[7] PP. V., Vijayarekha, K. (2014), *Machine vision applications to locate fruits, detect defects and remove noise: a review*, RASAYAN journal, 7(1), 104-113.

Appendix A. A SUMMARY OF LATIN GRAMMAR

Appendix B. SOME LATIN VOCABULARY

Graphic model to evaluate human brain connectivity over time

María Luisa Córdoba Tlaxcalteco*,
Dr. Carlos Hernández Gracidas**,
Dr. Alejandro del Rey Torres Rodríguez***,
Yoselyn Nohemí Ortega Gijón*

* Maestría en Sistemas Computacionales. Instituto Tecnológico Superior de Misantla, 162T0077@itsm.edu.mx

** Catedrático CONCYT. Instituto Tecnológico de Ciudad Victoria. División de Estudios de Posgrado e Investigación,

***Catedrático de Tiempo Completo. Instituto Tecnológico Superior de Misantla

12 de mayo de 2017

Abstract

It has been observed during childhood, the brain structure changes towards more ordered grouping. Recently, graphics theory has been introduced to model connectivity in the brain, particularly small-world networks such as the brain, because they combine the optimal properties of both ordered and random networks. We will use theoretical concepts to examine changes in functional brain networks by comparing them between various age groups (five-year lifetimes). It is proposed to perform three electroencephalography (EEG) records with closed eyes, 80 subjects in resting state and 10-20 international assembly in three different times. Between each pair of electrodes will be calculated the probability of synchronization (SL) in three different frequency bands to obtain the graphs of weights. The mean normalized clustering index, average path length and weight dispersion will be calculated to characterize the organization of the network. The general decrease in functional connectivity (SL) could reflect the pruning of unused synapses and the preservation of strong connections resulting in more stable networks. Consequently, we intend to find it increases in mean cluster and length of trajectory and decrease in weight dispersion, which would prove that the normal maturation of the brain is characterized by a shift from random to more organized functional networks of the small world. With this development process will detect what are the characteristics that influence the maturation of the human brain.

INTRODUCTION

During childhood, the brain undergoes major structural and functional changes. Deviation from normal development can have important consequences for our abilities as an adult, and may be involved in disorders such as ADHD, autism and schizophrenia . Therefore, knowledge of normal growth and developmental trajectories of brain networks is of great importance in finding risk factors and in the treatment of neuro-psychiatric disorders. The anatomical maturation of the brain in childhood follows different growth trajectories for different regions. Cross-sectional studies examining the differences between children and adolescents showed a much weaker functional connectivity in children than in adolescents . Changes in functional connectivity appear to be changing over time and with skill development.

Recently the theory of graphs has been introduced to model complex communication systems, such as the brain, as a network consisting of nodes and links. The nodes represent some kind of processing unit and the links represent a relationship between nodes, such as an anatomical connection or a functional interaction. The way the nodes are interconnected by the links provides information about the efficiency of a network. Networks in a regular configuration are characterized by a high cluster and a long average path length. Random networks, in which there is a fixed probability that there is a link between two nodes, have a low cluster and a short average path length. The so-called small world networks show a highly efficient dispersion of information in the network due to high clustering and short pathways between clusters. The EEG measures the anatomical and functional networks of the brain and gives us information about concentration and short trajectories showing an organization in the small world network.

Materials and methods

To analyze the development of brain maturation, studies must have a longitudinal design from early age and with multiple follow-ups (minimum of three EEG shots). When performing the longitudinal study, it is intended to detect with maturity causes a change to a more structured network of the brain. For this purpose we use the probability of synchronization (SL) as a general measure for functional connectivity in resting EEG shots (The way in which the EEG registers are collected are mentioned in [YN Ortega et al.)). From functional connectivity, we constructed weighted graphs to calculate clustering index and path length and weight dispersion to examine changes in brain development and maturation.

Organization of the sample

For the study, 80 normal subjects were selected for convenience, classified into 8 age groups with Equal number of elements, that will be located in educational institutions urban and suburban of the city Of Misantla, Ver. Age groups. 8 age groups (by source) according to the characteristics of the Psychological development and maturation aspects of the central nervous system.

Group 1. Children between 5 and 6 years of age attending regular preschool urban institutions.

Group 2. Children between 6 and 9 years of age, students of urban primary

schools.

Group 3. Children between 9 and 12 years old, students of urban primary schools.

Group 4. Adolescents between 12 and 16 years of age, high school students.

Group 5. Young people between 16 and 30 years of age, students of intermediate level, superior and / or active subjects Labor market.

Group 6. Adults between 30 and 65 years of age, labor-active subjects.

Group 7. Adults between 65 and 75 years of age, work-active subjects.

Group 8. Retired adults between 65 and 75 years of age.

EEG records

Registration data. EEG recording will be performed with international parameters 10-20 in rest state, and in Functional test conditions: ocular opening and closing, hyperventilation and photostimulation. Applies Recording with an approximate 15-20 minute recording with initial calibration record [11]. The registry parameters are as follows:

1. Prevailing high cut frequency (alpha): 35-50 Hz
2. Prevailing mid-cut frequency (beta): 2-15 Hz
3. Prevailing low cut frequency (teat): 1-2 Hz
4. Sensitivity for absolute power (AP) recording: 70 muV

Assembly. Assemblies shall be of the bipolar, sagittal or transverse type, and shall be made according to the system International 10-20 [29]. The letters F, T, C, P and O represent the frontal, temporal, central, parietal lobes And occipital [29], respectively. The even numbers (2,4,6,8) refer to the positions of the electrodes in The right hemisphere, while the odd numbers (1,3, 5,7) refer to those in the left hemisphere. "Z"(Zero) refers to an electrode placed in the midline.

Power Spectrum

First, we will calculate a relative power spectrum on average for all channels, times for all subjects and for each of the age groups. When computing the relative power spectrum, the raw EEG signal from the time domain (continuous variable) will be converted to a new frequency domain (discrete variable) using Fast Fourier Transform (FFT). The power spectra will be averaged over time to obtain the average relative power spectrum for all electrode positions. Example of the localization of some electrodes and the signals recorded for a subject with intractable seizures (see Anex fig.1)[38]

SL calculation

The signal at each time will be digitally filtered in the frequency bands of interest; Theta (4-6 Hz), alpha (6-11 Hz) and beta (11-25 Hz). The probability of synchronization (SL) will be calculated as a measure of the functional connectivity between the different regions of the brain. The final result of calculating SL for all combinations of channel pairs for a specific frequency band is a square matrix N of dimension $20x20$,That is, 20 is the number of EEG channels used in this study. Each entry $N_(x, y)$ contains the value of SL for the channel combination x y y. Subsequently, the average synchronization will be calculated, resulting in a single global value SL for every time in the whole brain. Finally, this overall value of SL will be averaged over 4 seasons for each child.

Graphics model analysis approach

In this study, changes in the development of brain network characteristics measured by EEG will be analyzed. The nodes of the graph are represented by the electrodes while the links are defined by the measure of association between the nodes, in this study SL.

The clustering index of a node will represent the proportion of its neighboring nodes that are connected to each other. To calculate the clustering index of the weighted networks, the weights between the node i and other nodes j must be symmetrical ($w_{ij} = w_{ji}$) and $0 \leq w_{ij} \leq 1$. These conditions are satisfied since we use the SL value as weights:

$$C_i = \frac{\Sigma_{k\neq i}\Sigma_{l\neq i,l\neq k} w_{ij} w_{il} w_{kl}}{\Sigma_{k\neq i}\Sigma_{l\neq i,l\neq k} w_{ij} w_{il}} \tag{1}$$

In this formula $i = k$, $i = l$ y $k = l$ are not included. The average aggregation of the total network is defined as:

$$C_W = \frac{1}{N}\Sigma_{i=1}^{N} C_i \tag{2}$$

The length of an edge is defined as the inverse of the weight, that is, $L_{ij} = 1/N$ if $w_{ij} \neq 0$, and $L_{ij} = \infty$ if $w_{ij} = 0$. The shortest path between the nodes i and j is the sum of the shortest lengths between two nodes.

The average path length of the entire network is calculated as

$$L_W = \frac{1}{(\frac{1}{N(N-1)})\Sigma_{i-1}^{N}\Sigma_{j\neq 1}^{N}(\frac{1}{L_{ij}})} \tag{3}$$

The measure that describes the connectivity of the network is called weight dispersion (r_i).

$$r_i = \frac{W_{max}(i) - W_{min}(i)}{W_{max}(i) + W_{min}(i)} \tag{4}$$

W_max will represent the maximum weight and W_min the minimum weight of the edges of the node i. The average of r_i on all the nodes of the network W_r.

Section Statistical methods The inference about the characteristics detected by RP will be made using the hierarchical linear model [31]. This requires:

1. One or more continuous variables, which change over time.

2. Records of the variables in three or more instants of time, in the same subject.

3. A unit of time relevant to the phenomenon being studied, taking into account its cadence and speed exchange.

The continuous variables considered will be the characteristics detected by RP in the EEG signals. Examples of these are the Hurst index [32], [33], the likelihood of world chart synchronization Small [36], main components or independent components [34], [35]. The EEG record is Will perform three times in each subject, at intervals of six months between the first two sampling shots, And three months for the last signal taking. The analysis of the rate of change over time of Characteristics discovered by RP will be realized through a linear

model in two levels. At the first level, For each subject, a linear model of the form will be fitted:

$$Y = \beta_0 + \beta_1 * tiempo + \beta_2 * tiempo^2 + \ldots + \epsilon_i \quad (5)$$

Where Y is the pattern or characteristic of the subject's EEG signal i , *tiempo* Is the subject's age, the parameters β will be estimated by maximum likelihood, and ϵ_i represents the random error, with a distribution of Probability to be defined.

The analysis of the second level makes use of stratification by age, sex, or other characteristic of interest. Given the group j, the probability distribution of the parameter vector $(\beta_1, ..., \beta_n)$ Will be assumed the normal multivariate distribution, with vector of means μ_j , and covariance matrix D_j . In such a way that the characterization of the group j is performed through the pair (μ_j, D_j). For example, the effect of time in two age groups, say j and k, is evaluated by hypothesis testing $H_0 : \mu_j - \mu_k = 0$ [31].

Numerical experiments

From the database of patients with epilepsy described in [39], we took the EEG signals from patient 1 (in folder chb01 and chb21). The patient is of the feminine gender and had the age of 11 years. We chose this patient who is the only subject to be sampled on two different dates, with a year and a half of separation. The signals were sampled at 256 samples per second with a 16-bit resolution for one hour. Each of the files is .edf format, where the chb01 folder contains 46 EEG signal shots of one hour each. The signals were digitalized from which we only took the first take for our experiments. The location of the electrodes was under the international system 10-20 having in total 22 signals obtained between the pairs of electrodes:

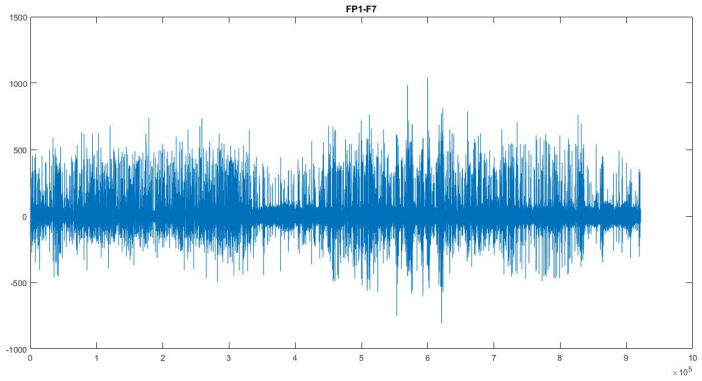


Figura 1: Signals between the electrodes: FP1-F7.

canal	
1	FP1-F7
2	F7-T7
3	T7-P7
4	P7-O1
5	FP1-F3
6	F3-C3
7	C3-P3
8	P3-O1
9	FP2-F4
10	F4-C4
11	C4-P4
12	P4-O2
13	FP2-F8
14	F8-T8
15	T8-P8
16	P8-O2
17	FZ-CZ
18	CZ-PZ
19	P7-T7
20	T7-FT9
21	FT9-FT10
22	T8-P8

The signal was filtered in the frequency bands Theta, alpha and beta; As shown in the example in figure [].

Once the EEG signals are recorded in each of the channels and in each of the bands, they are converted into a series of spatial vectors of states and have the following form:

$$X_i = (x_i, x_{i+L}, x_{i+2*L}, ..., x_{i+(m-1)*L}) \tag{6}$$

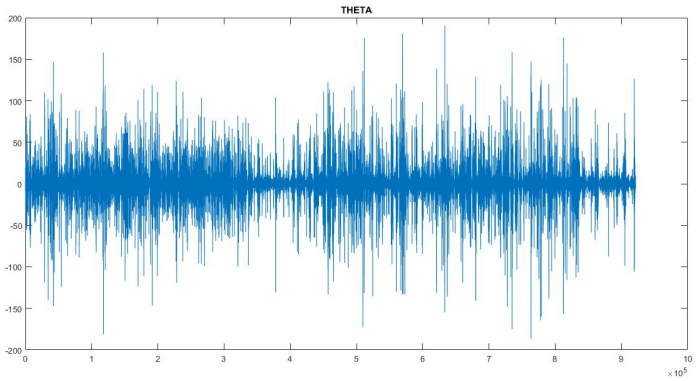


Figura 2: Signal Theta between the electrodes: FP1-F7.

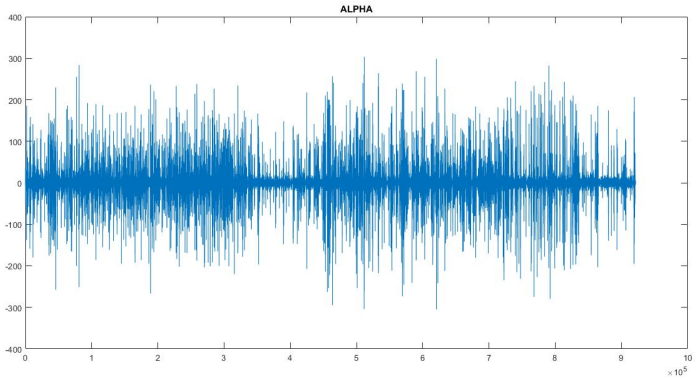


Figura 3: Signal Alpha between the electrodes: FP1-F7.

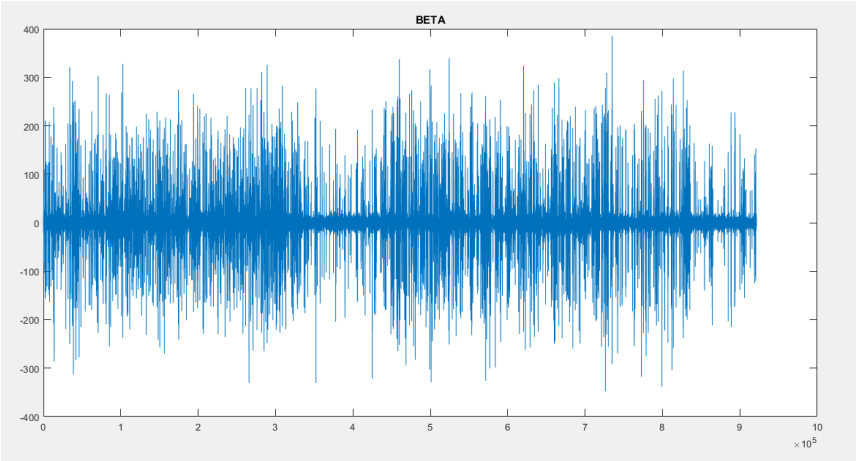


Figura 4: Signals Beta between the electrodes: FP1-F7.

Where L is the time delay and m is the inclusion dimension. The time delay (L) depends on the sampling frequency (fs) and the highest frequency of interest (HF):

$$L = fs/3xHF \tag{7}$$

For the Theta band we have that $fs = 256$, $Lf = 4$ and $HF = 8$ are respectively the lowest frequency and the highest frequency, so we calculate:

$$L = fs/3xHF = 256/3x8 = 682$$

The embedding dimension (m) depends on the lowest frequency (LF) of interest and determines the length of the embedding window :

$$L(m-1) = fs/LF \tag{8}$$

This is:

$$m = 3 * HF/LF + 1 \tag{9}$$

According to our previous data:

$$\begin{aligned} m &= 3 * HF/LF + 1 \\ &= 3 * 8/4 + 1 \\ &= 3 * 2 + 1 \\ &= 7 \end{aligned}$$

Here we can begin to calculate the spatial vectors. For example, for time i we obtain the vector X_j of j channel:

$$\begin{aligned} X_j &= (x_i, x_{i+682}, ..., x_{i+(6)*682}) \\ &= (x_{1000}, x_{1682}, ..., x_{4410}) \end{aligned}$$

In this way we calculate the 22 spatial vectors j, at the same time i but in the different channels.

$$\begin{aligned} X_1 &= (-14{,}664626, 72{,}192579, 18{,}875081, -25{,}819653, -4{,}641928, -6{,}263402) \\ X_2 &= (4{,}883106, 3{,}715114, -1{,}497538, -13{,}530475, -13{,}179306, 7{,}014671) \\ &\vdots \\ X_{22} &= (0{,}6357451, -10{,}6033441, -4{,}4791323, 9{,}8126876, -0{,}3330107, -6{,}7150066) \end{aligned}$$

The 22 vectors that have been calculated have been stored in the following dimendion matrix (22 x 6), see fig.[]

The vector X_i represents the state of the system X at time i in a time interval of length $L * (M - 1)$. In the same channel, we look for X recurrences of vector X_i at time j. Therefore, a threshold distance set in space is chosen Of states so that a fixed part (p_ref) of the compared vectors is sufficiently close to consider them to be in the same state. Five percent of the vectors X_j will be considered as recurrences of X_i given at $p_{ref} = 0{,}05$.

The vector X_i represents the state of the system X at time i in a time interval of length $L \times (M - 1)$. In the same channel, we look for X recurrences of vector X_i at time j. Therefore, a threshold distance set in space is chosen Of states so that a fixed part (p_{ref}) of the compared vectors is sufficiently close to consider them to be in the same state.

Five percent of the vectors X_j will be considered as recurrences of X_i given at $p_ref = 0{,}05$.

A window W1 is defined around time i and is called Theiler's correction for self-correlation [Theiler, 1986]. If W1 is twice the length of the inclusion vectors $[W_1 = 2 \times L \times (M - 1)]$, then two consecutive vectors only share one sample point.

To capture a sufficient number of vectors to take P_{ref} from them by defining the recurrences a second window W2 is defined:

$$N_{rec} = (W_2 - W_1 + 1) \times p_{ref}$$

Where n_{rec} is the number of recurrences.

State space vectors of the EEG signal are constructed on the Y channel and with the same value for p_ref a recurrence search is performed.

Conclusions

The graphical models are a computational and mathematical tool for the problem of detection of cerebral connectivity.

Referencias

[1] H. E. Hurst and American Society of Civil Engineers, Hydraulics Division. Long-term Storage Capacity of Reservoirs, American Society of Civil Engineers, 1950.

[2] R. Grishman, The Oxford handbook of computational linguistics, Oxford University Press, 2012.

[3] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, "Mosaic organization of DNA nucleotides", In: Phys, Rev. Ed 49, 1994, pp. 1685-1689.

[4] M. B. Priestley. .Evolutionary spectra and non-stationary processes", In: Journal of the Royal Statistical Society, Series B (Methodological),Vol. 27, No. 2, 1965, pp. 204-237.

[5] E. E. Rodríguez, E. Hernández-Lemus, B. A. Itzá-Ortiz, I. Jiménez, P. Rudomín. "Multichannel Detrended Fluctuation Analysis Reveals Synchronized Patterns of Spontaneous Spinal Activity in Anesthetized Cats", In: PLoS, 2011. ¡https://doi.org/10.1371/journal. pone.0026449¿

[6] J. R.J. Schirra, Foundation of Computational Visualistics, Dt. Unive.-Verlag, 2005.

[7] M. Staudachera, S. Telserb, A. Amannc, H. Hinterhuberb, M. Ritsch-Marte .A new method for change-point detection developed for on-line analysis of the heart beat variability during sleep"

[8] D. L. Anderson, .A Guide to Patient Recruitment: Today's Best Practices and Proven Strategies", Centerwatch Incorporated, 2001, https://books.google.com.au/books? id=aLoOAAAACAAJ.

[9] H. K. Bouma, C. Labos, G. C. Gore, C. Wolfson, and M. R. Keezer, "The Diagnostic Accuracy of Routine Electroencephalography After a First Unprovoked Seizure", In: European Journal of Neurology. Wiley Online Library, 2015.

[10] Çenso General de Población Y Vivienda, 2010.", Instituto Nacional de Estadística, Geografía E Informática, 2011.

[11] Monitoring in the NICU.Çurrent Pediatric Reviews 10(1), Bentham Science Publishers: 2-10, 2014.

[12] Force, Pharmaceutical Industry Competitiveness Task, "Pharmaceutical Industry Competitiveness Task Force: Final Report March 2001.Ünited Kingdom Department of Health, 2001.

[13] M. D. Lamblin, and A de Villepin Touzery, .EEG in the Neonatal Unit."Neurophysiologie Clinique/Clinical Neurophysiology 45 (1). Elsevier: 87-95, 2015.

[14] M. Ley, R. Vivanco, A. Massot, J. Jiménez, J. Roquer, and R. Rocamora, "Safety Study of Long Term Video Electroencephalogram Monitoring."Neurología (English Edition) 29 (1). Elsevier: 21-26, 2014.

[15] Organization, World Health, and others, "Trastornos Neurológicos: Desafíos Para La Salud Pública."Ginebra, Suiza: World Health Organization, 2006.

[16] Reglamento de La Ley General de Salud En Materia de Investigación Para La Salud", Diario Oficial de La Federación. Secretaría de Salud, 1987.

[17] R. A. Shellhaas, Çontinuous Long Term Electroencephalography: The Gold Standard for Neonatal Seizure Diagnosis.În Seminars in Fetal and Neonatal Medicine. Elsevier, 2015.

[18] R. Zafar, A. Saeed Malik, N. Kamel, S. C. Dass, J. M. Abdullah, F. Reza, and A. H. Abdul Karim, "Decoding of Visual Information from Human , 2015.

[19] "Brain Activity: A Review of FMRI and EEG Studies"Journal of Integrative Neuroscience World Scientific, 1-14.

[20] OECD Better Life Initiative, "How's life in Mexico?", OECD (Organization for Economic Cooperation and Development), Octubre 2015.

[21] G. L. Birbeck, Revising and refining the epilepsy classification system: Priorities from a developing world perspective", Epilepsia, 53 (Suppl. 2), pp: 18-21, 2012.

[22] Save the Children, "Manual para la inclusión de niños y niñas con discapacidad familiar en centros comunitarios de desarrollo infantil", Save the Children, Fundación Alfredo Harp Helú, 2014.

[23] Y. M.Mughal, A Parametric Framework for Modelling of Bioelectrical Signals Springer Singapore (2016) Paulo J.G. Lisboa, Emmanuel C. Ifeachor, Piotr S. Szczepaniak, Artificial Neural Networks in Biomedicine Springer, 2000.

[24] O. Olofsson, "The development of the electroencephalogram in normal children and adolescents from the age of 1 through 21 years", tesis de doctorado, universidad de Gotenburgo, 1970. Disponible en:

[25] A. K. Engel, P. Fries, "Beta band oscillations - signalling the status quo?", Current Opinion in Neurobiology 20(2), 156 165

[26] M. Botcharova, Çhanges in structure of EEG EMG coherence during brain development: analysis of experimental data and modelling of putative mechanisms."

[27] S. F. Farmer, J. Gibbs, D. M. Halliday, L. M. Harrison, L. M. James, M. J. Mayston, A. A. Stephens, Çhanges in EMG coherence between long and short thumb abductor muscles during human development", J Physiol. 2007 Mar 1; 579(Pt 2): 389-402.

[28] .ªgencia Española de Medicamentos y Productos Sanitarios de la guía de Buena Práctica Clínica", publicada en la página Web de la Comisión Europea.

[29] F. Sharbrough, G. E. Chatrian, R. P. Lesser, H. Lüders, M. Nuwer, T. W. Picton, .ªmerican Electroencephalographic Society. Guidelines for standard electrode position nomenclature", Journal of Clinical Neurophysiology, 8(2):200-202,Raven Press, Ltd., New York, 1991.

[30] J. K. Lindsey, Introductory Statistics: The Modelling Approach , Oxford Science Publications, 1995.

[31] J. D. Singer, J. B. Willett, Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence, Oxford University press,Inc., New York, 2003.

[32] H. Pellé, P. Ciuciu, M. Rahim, E. Dohmatob, P. Abry, V. Van Wassenhove, "Multivariate Hurst Exponent Estimation in FMRI. Application to Brain Decoding of Perceptual Learning", 13th IEEE International Symposium on Biomedical Imaging, Apr 2016, Prague, Czech Republic. 2016.

[33] N. Kannathal, U. Rajendra Acharya, C.M. Lim, P.K. Sadasivan, Çharacterization of EEG–A comparative study", Department of ECE, National University of Singapore, Singapore ECE Division, Ngee Ann Polytechnic, 2005.

[34] A. A. Putilov, "Principal component analysis of the EEG spectrum can provide yes-or-no criteria for demarcation of boundaries between NREM sleep stages", Published online 2015 Mar 10.

[35] C. Bugli, P. Lambert, Çomparison between Principal Component Analysis and Independent Component Analysis in Electroencephalograms Modelling", Institut de Statistique, Universit catholique de Louvain, Bruxelles, Belgium, 2006.

[36] M. Boersma, D. J. A. Smit, H. M. A. de Bie, G. C. M. Van Baal, D. I. Boomsma, E. J.C. de Geus, H. A. Delemarre-van de Waal, and C. J. Stam, "Network Analysis of Resting State EEG in the Developing Young Brain: Structure Comes With Maturation ", In: Human Brain Mapping 32:413-425, 2011.

[37] P. Berka, J. Rauch, D. A. Zighed, Data Mining and Medical Knowledge Management: Cases and Applications, IGI Globa, Hershey, New York, 2009.

[38] A. Shoeb, Applications of Machine Learning to Epileptic Seizure Onset Detection and Treatment, PhD Thesis,Massachusetts Institute of Technology, September, 2009.

[39] A.L. Goldberger , L.A. Amaral , L. Glass , J. M. Hausdorff , P.Ch. Ivanov , R. G. Mark , J. E. Mietus , G. B. Moody , C. K. Peng , H. E. Stanley, Components of a New Research Resource for Complex Physiologic Signals, Circulation Electronic Pages: PhysioNet 101 (23): e215-e220, 2000.

ANEXO

Recordings, grouped into 23 cases, where collected from 22 subjects (5 males, ages 3-22; and 17 females, ages 1.5-19). Each case contains between 9 and 42 continuous .edf files from a single subject.

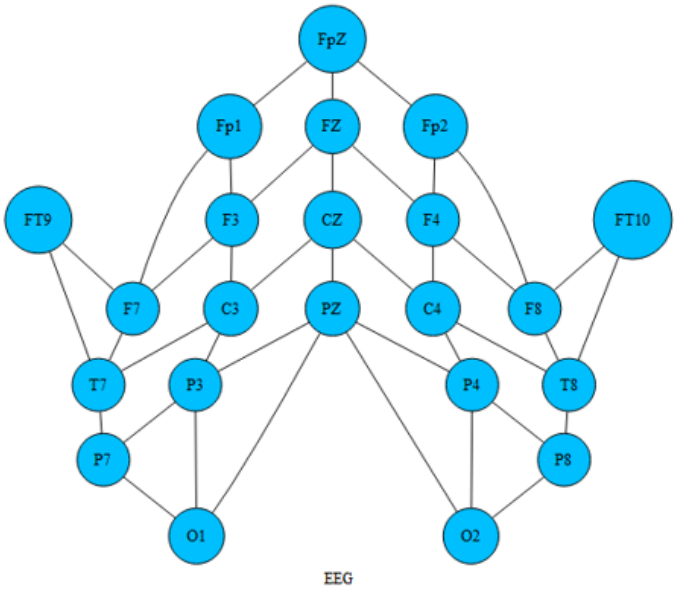


Figura 5: System 10-20.

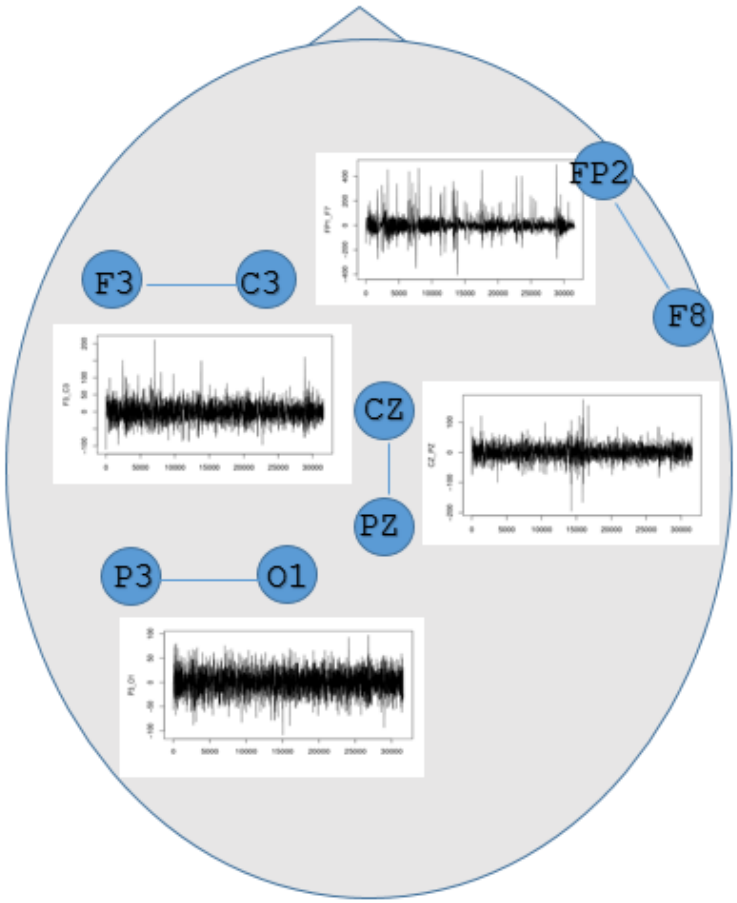


Figura 6: Signals between the electrodes: FP2-F8, F3-C3, Cz-Pz y P3-O1.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	-14.6646265	72.1925790	18.875081	-25.819653	-4.6419278	-6.2634020
[2,]	4.8831063	3.7151140	-1.497538	-13.530475	-13.1793059	7.0146714
[3,]	0.9539655	-16.1639601	1.461305	7.970282	2.6235174	-6.9410704
[4,]	-6.2392053	-4.4665476	3.282516	-6.478556	3.7145063	7.5703862
[5,]	-12.1617761	55.5537061	19.712484	-18.387022	-12.5685482	-3.7484306
[6,]	-3.4190782	5.9333572	-2.083308	-9.354666	-8.0631472	1.7916368
[7,]	-3.4190782	5.9333572	-2.083308	-9.354666	-8.0631472	1.7916368
[8,]	-4.5255167	-1.0128198	5.959708	-2.233582	9.9833625	2.1536112
[9,]	-14.5011284	37.8159122	12.665567	-19.707683	-8.9014365	-2.4759763
[10,]	-8.3694762	4.5180447	-3.072832	6.338683	0.3196079	-2.5891876
[11,]	-8.3694762	4.5180447	-3.072832	6.338683	0.3196079	-2.5891876
[12,]	-1.5678868	5.1568359	-16.754162	-43.714818	8.6373031	4.9303515
[13,]	0.6975219	26.4663980	8.525543	-31.997009	-7.4703947	0.4152987
[14,]	-1.6329084	21.3639926	2.559144	-4.424137	0.3039498	-9.1367136
[15,]	-3.3182660	-2.7076127	2.133261	3.665642	2.1529647	-0.1187928
[16,]	-12.3424633	15.3536675	-23.907088	-43.716175	5.8575570	11.5271299
[17,]	5.0952962	0.4360336	-6.447448	-16.313198	6.6755344	2.7978368
[18,]	0.2880026	2.1728270	6.160630	-1.175930	-1.6650097	1.8673999
[19,]	-0.9539655	16.1639601	-1.461305	-7.970282	-2.6235174	6.9410704
[20,]	-3.8430030	8.8445057	1.344970	-1.420164	11.1026988	1.5905404
[21,]	16.6016592	-24.9783748	-5.245338	-14.833664	-8.4415341	1.3594480
[22,]	0.6357451	-10.6033441	-4.479132	9.812688	-0.3330107	-6.7150066

Figura 7: Matrix of spatial vectors.

Classification of signals in multiple EEG channels. Detection of human brain maturation factors

Yoselyn Nohemí Ortega Gijón [*]
Dr. Carlos Hernández Gracidas [**]
Dr. Alejandro del Rey Torres Rodriguez [***]
María Luisa Córdoba Tlaxcalteco [***]

[*] *Instituto Tecnológico Superior de Misantla, Maestría en Sistemas Computacionales (e-mail: 162t0082@itsm.edu.mx).*
[**] *Tecnológico Nacional de México ,Instituto Tecnológico de Ciudad Victoria,División de Estudios de Posgrado e Investigación*
[***] *Instituto Tecnológico Superior de Misantla*

Abstract:
The application of main component analysis (PCA) to the field of data mining in the electroencephalogram (EEG) processing. The main components are calculated from the signal by eigen-decomposition of the covariance estimate of the input. Alternatively, they can be estimated using a gured neural network (NN) to extract the first major components. Instead of performing computationally complex operations for the estimation of eigenvectors, the neural network can be trained to produce the first ordered components. Possible applications include the separation of different signal components for the extraction of characteristics in the field of EEG signal processing for the detection of characteristics of brain electrical activity at each chronological stage.

1. INTRODUCTION

For centuries, humans have tried to decipher the riddles of the functioning of their brain. This curiosity has led us to advances such as being able to interact with the computer with the simple fact of thinking, with the so-called computer brain interfaces. However, only a few centuries ago, many diseases that are better documented today were erroneously diagnosed as diabolical possessions or were classified as lunatic lunatics, since their evil was believed to be related to the phases of the moon. Even today, it is very common to say that a person "is in a bad mood" when he may be showing symptoms of depression. This gives an idea of the need that still exists to develop updated techniques for the diagnosis and treatment of diseases of the nervous system.

The objective of this article is to identify by RP the characteristics of human brain electrical activity, and to evaluate the change in the time of these characteristics.

It deals with the question of the application of advanced methods for the analysis of EEG signals. First, EEG is introduced briefly. Then there is a description of the phases of EEG signal processing. The article focuses on the most important parts of this process, namely the application of the main component analysis (PCA) to the extraction of characteristics and the statistical methods to be used. The explanation is illustrated by examples of EEGs obtained.[18]

Recognition of Patterns in Artificial Intelligence

Pattern Recognition (RP) is a discipline of Computational Sciences whose purpose is to extract characteristics, from a set of data, that allow to discern, through automatic methods, those common conditions or behaviors and those considered anomalous for those data; These characteristics, once determined, can be used to represent such data for recognition, retrieval, filtering or interpretation.[1].

The RP has shown its utility to be applied in an immense variety of contexts, where it becomes necessary to identify such patterns to automate processes, optimize resources, detect faults, etc. Automatic Pattern Recognition techniques have been used in a wide variety of experimental settings with promising results. Many of these same techniques are even commonly employed in a wide range of real world applications such as Computational View and Natural Language Processing, to name a few.[2]

In the future, this type of work could help to define normal EEG for the diagnosis of possible neurodevelopmental disorders[2].

2. STATE OF ART

A. Subasi and M. Ismail Gursoy, performed the prediction of seizures is their potential for use in therapeutic epilepsy devices to trigger intervention to prevent seizures before beginning. Studies on seizure prediction vary widely in their theoretical approaches to the problem, the validation of the results, and the amount of data analyzed. They applied DWT (Wavelet Transform) for time-frequency analysis of EEG signals for classification using wavelet coefficients. EEG signals were decomposed into frequency

subbands using DWT. Then a set of statistical characteristics of these subbands were extracted to represent the distribution of the wavelet coefficients. PCA, Independent Component Analysis (ICA) and LDA (Linear Discriminant Analysis) were used to reduce the dimension of the data. The features were used as an input to a support vector machine (SVM) with two discrete outputs: epileptic seizures Or in epileptics. The best performance is achieved in the extraction of LDA features, the training process was carried out using the core RBF to PCA + SVM, ICA + SVM, and LDA + SVM. After training, they used three extraction methods of different characteristics. Using the features of PCA, ICA and LDA are extracted from the original feature sets. In addition, the number of support vectors (SV) decreased due to feature extraction. The classification rate with extraction of LDA characteristics is the highest (100 %) and the ACI came second (99.5 %). The PCA had the lowest correct classification percentage (98.75 %) compared to the LDA and ICA counterparts. Also the simulation shows that SVM by extraction of characteristics using PCA, ICA or LDA can always have better performance than without extraction of characteristics (98 %)[3]

C. Bugli and P. Lambert, compared the performance of PCA and ICA for two applications in the electroencephalogram (EEG) modeling. The first application is the reduction of dimensionality. The second application is the modeling of an interesting feature of EEG signals called event related to potential (ERP). They used PCA and ICA to reduce dimensionality. They used the JADE algorithm for these 2 applications with a Matlab1 package to perform the JADE algorithm. This package was designed to facilitate EEG analysis and to graphically visualize the estimated sources. With PCA, we found only one independent component needed to analyze the P300 peak. However, this analysis does not separate the phenomena P3a and P3b. The entire procedure is automatic and easy to set up. ICA is very promising to analyze the multidimensional biomedical signals and much more efficient than the PCA in the context of EEG analysis, to be able to separate the 2 phenomena.[4]

X. Liu et al. They investigated the complexity of EEG signals and distinguish EEG signals with different rhythms. Methods: EEG data are used in the natural mass model. The model with different parameters produces EEG signals at different rates, such as normal EEG signals and epileptiform peaks. Modified permutation-entropy is used to extract the complexity characteristics of normal EEG signals and epileptiform peaks and is compared with the permutation-entropy algorithm and Rényi's permutation-entropy algorithm. Statistical analysis is used to characterize the differences between the signals. Results: The signals corresponding to A = 3.25 mV and A = 3.6 mV are easy to distinguish according to mean values and standard deviations of the modified permutation-entropy. The statistical results reveal that the group differences of the modified permutation-entropy values for the signals corresponding to A = 3.25 mV and A = 3.4 mV are significant, manifested as F $\prec$ 1 y p $\succ$ 0,05. In comparison to the permutation entropy and Rheny's permutation entropy, the F-values of the modified permutation-entropy are the largest and the p-values of the modified permutation-entropy are the smallest. Therefore, they concluded that the permutation Modified-entropy is a measure of effective complexity for quantifying EEG signals and the modified permutation-entropy is more effective in distinguishing EEG signals with different rhythms compared to Rheny's permutation entropy and entropy-permutation.[5]

Li-Chen Shi et al. In this paper we introduce a robust algorithm of Principal Component Analysis (PCA) to reduce the dimension of EEG characteristics for surveillance estimation. The performance is compared to the PCA standard, L1-PCA standard, PCA sparse and robust PCA in feature dimension reduction in an EEG dataset of twenty-three subjects. To evaluate the performance of these algorithms, they use smoothed differential entropy features such as EEG-related surveillance features. Experimental results demonstrate that the ruggedness and robust PCA performance are better than other algorithms for estimation of off-line and online surveillance. The mean RMSE (mean square error) of the surveillance estimate was 0.158 when robust PCA was applied to reduce the dimensionality of the characteristics, while the mean RMSE was 0.172 when standard PCA was used in the same task.[6]

Kavita Mahajan. They investigated electroencephalogram (EEG) processing and analysis in a proposed framework that was performed with DWT for the decomposition of the signal into its frequency subbands and a set of statistical characteristics extracted from the subbands to represent the distribution of the Wavelet coefficients. The reduction of the data dimension is done with the help of analysis of main components and analysis of independent components. These features were then used as an input to a neural network for classification of data as normal or otherwise. They presented and compared the performance of the classification process due to different methods to show the excellent classification process. These findings are presented as an example of a method for training and for testing a normal and abnormal prediction method in small individual epileptic patient data. Method of extraction of characteristics, Accuracy (%), Sensitivity (%), Specificity (%), PCA 93.63 % 62.93 % 98.83 %. ICA 96.75 % 96.75 % 96.75 %. [7]

Seungjin et al. They developed a domain system of the brain computer interface (BCI) based on EEG. The system consists of two procedures: (1) extraction of characteristics; (2) classification. They presented two methods for the classification of EEG patterns, the first one starts with PCA and the characteristics of the main components extracted separately from the channels C3 and C4 are concatenated, then introduced into the corresponding HMM (model of Hidden Markov) (modeling the Movement to the left or right) for training. On the other hand, in the second methodology, the characteristics of the main components of each channel are entered into two separate HMMs, which results in four HMMs in total. The SVM is used in the second method to make a final decision based on the probability scores calculated by HMMs. The performance comparison for PCA characteristics, raw EEG data and Hjorth parameters reflected the following results, methodology 1: PCA 75.70, Raw 60.63, Hjorth 56.88, methodology 2: PCA 78.15, Raw 64.38, Hjorth 66.50, therefore Method-

ology 2 showed a slightly better performance compared to methodology 1. [8]

3. MATERIALS AND METHODS

3.1 Recruitment, Samples, and Records

3.2 Organization of the sample

For the study, 80 normal subjects, classified in 8 age groups with equal number of elements, were selected for convenience in urban and suburban educational institutions in the city of Misantla, Ver.

Age groups. Eight age groups (according to origin) were made according to the characteristics of the psychological development and maturation aspects of the central nervous system.

Group 1. Children between 5 and 6 years of age attending regular preschool urban institutions.

Group 2. Children between 6 and 9 years of age, students of urban primary schools.

Group 3. Children between 9 and 12 years old, students of urban primary schools.

Group 4. Adolescents between 12 and 16 years of age, high school students.

Group 5. Young people between 16 and 30 years of age, students of intermediate level, superior and / or subjects active labor.

Group 6. Adults between 30 and 65 years of age, labor-active subjects.

Group 7. Adults between 65 and 75 years of age, work-active subjects.

Group 8. Retired adults between 65 and 75 years of age.

3.3 EEG Records

Registration data. EEG recording will be performed with international parameters 10-20 in resting state, and under conditions of functional tests: ocular opening and closing, hyperventilation and photostimulation. We apply registration with an approximate 15-20 minutes record with initial calibration record [9].

The registry parameters are as follows:

1. Prevailing high cut frequency (alpha): 35-50 Hz

2. Prevailing mid-cut frequency (beta): 2-15 Hz

3. Prevailing low cut frequency (teat): 1-2 Hz

4. Sensitivity for absolute power (PA) recording: 70? V

Assembly.- Assemblies shall be of the bipolar type, sagittal or transverse, and shall be made according to the international system 10-20 [10]. The letters F, T, C, P and O represent the frontal, temporal, central, parietal and occipital lobes [10], respectively. The even numbers (2,4,6,8) refer to the positions of the electrodes in the right hemisphere, while the odd numbers (1,3, 5,7) refer to those in the left hemisphere. A "z" (zero) refers to an electrode

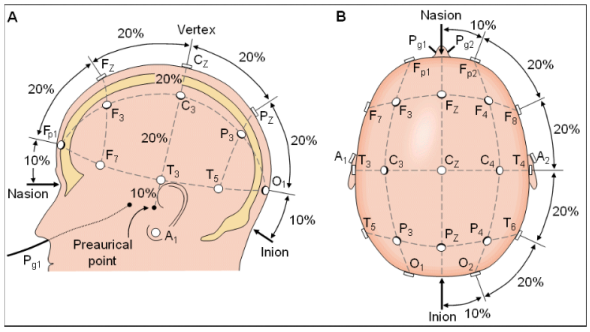


Figure 1: Sistema internacional de montaje 10-20.

montaje

Fig. 1. International assembly system 10-20

placed in the midline. Fig. Ref fig: montaje shows the assembly to occupy.

3.4 Statistical Methods

Inference about the characteristics detected by RP will be performed using the hierarchical linear model [11]. This requires:

1. One or more continuous variables, which change over time.

2. Records of the variables in three or more instants of time, in the same subject.

3. A unit of time relevant to the phenomenon being studied, taking into account its cadence and rate of change.

The continuous variables considered will be the characteristics detected by RP in the EEG signals. Example of these are the Hurst index [12], [13] and main components [14]. EEG recording will be performed three times in each subject, at six-month intervals between the first two sampling shots, and three months for the last signal taking. The analysis of the rate of change in time of the characteristics discovered by RP will be realized by means of a linear model in two levels. In the first level, for each subject, a linear model of the form will be fitted:

$$Y = \beta_0 + \beta_1 * tiempo + \beta_2 * tiempo^2 + ... + \epsilon_i$$

Where Y is the pattern or characteristic of the subject's EEG signal i, *tiempo* is the subject's age, the β parameters will be estimated by maximum likelihood, and ϵ_i represents the random error, with a probability distribution to be defined.

The analysis of the second level makes use of stratification by age, sex, or other characteristic of interest. Given the group j, the probability distribution of the parameter vectors $(\beta_1, ..., \beta_n)$ will be assumed the normal multivariate distribution, with vector of means μ_j, and covariance matrix D_j. In such a way that the characterization of the group j is performed through the pair (μ_j, D_j). For example, the effect of time on two age groups, let say j y k, is evaluated by hypothesis testing $H_0 : \mu_j - \mu_k = 0$ [11].

3.5 Procesamiento y RP en la señal EEG

The processing of the EEG signal will be performed in the following steps [15].

1. Acquisition.- The EEG signal is recorded in files with EDF or ASCII formats.

2. Pre-processing.- The purpose is to eliminate noises and artifacts that may distort PR performance. It will be done using ad hoc algorithms.

3. Digital Filters.- Elimination of unwanted frequency components is carried out by means of filter banks. Elimination of atypical or unusual signs and cases is also performed.

4. Segmentation.- The signal is divided into segments of time intervals of different sizes, using adaptive segmentation algorithms.

5. Extraction of characteristics by RP.- The signals, divided into segments and clean of noises and artifacts, are evaluated by PR techniques defined below.

6. Statistical classification and analysis.- The results of the RP will be subjected to various classification techniques (clusters, principal analysis components), and hierarchical statistical models described above.

Postulated PR techniques for the analysis of EEG signals are described below.

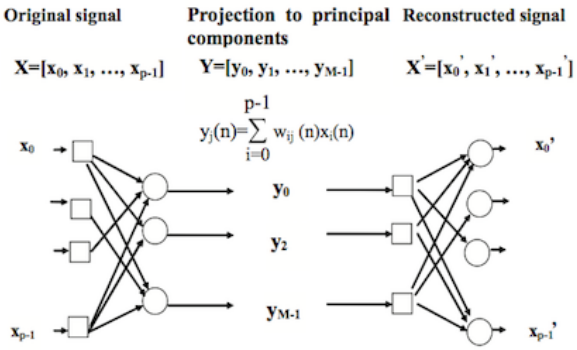
3.6 Multivariate Hurst Coefficient

In the univariate case, ie an EEG signal of a pair of electrodes, if f is the predominant frequency in the theta, alpha, or beta bands, and if $P(f)$ represents the PA of this frequency, Says the EEG signal has a Hurst exponent H if $P(f) \sim 1/f^H$, where $0 < H < 1$. The Hurst coefficient is used to evaluate the dependence-to-long-term time series. This parameter has been used to classify EEG signals under various activities. The estimation of H has been made by fluctuation analysis, linear regression in the graph $log(f)$ vs $log(P)$, or by ripples. In the case of EEG signals in several channels, there are several multivariate generalizations of the H coefficient. The present project raises the use of some of these generalizations to characterize the EEG signal in each age group [16], [17], [12].

3.7 Principal Components Analysis (PCA)

It is a technique mainly used to reduce dimensionality. In conjunction with classification techniques (such as cluster analysis), it will be used to characterize the components of each age group [14]. It deals with the truthfulness of the investigations carried out, its operation in the analysis of electroencephalogram signals when it is occupied for the segmentation of the signals, its use in the detection of epileptic patients and extraction of the EEG characteristics of patients in coma, Is assumed to be functional for the analysis and detection of human brain maturation factors. [15]

For the analysis of major components are calculated from the signal by eigen-decomposition of the covariance estimate of the input. Alternatively, they can be estimated by



estructura

Fig. 2. Structure of the multi-layer feedforward neural network

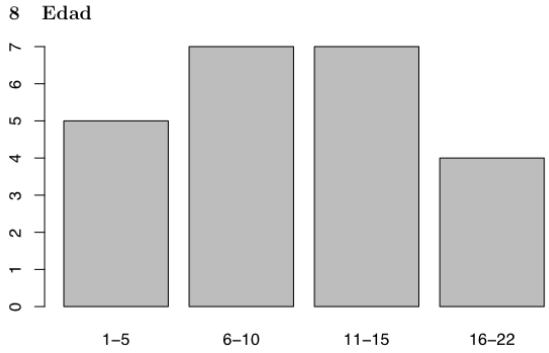


Fig. 3. Registered Data

a neural network (NN) to extract the first major components, the extracted features are used for the successive classification.

A proposed structure of this Neural Network is the one shown in Fig. 2:

4. RESULTS

As a result there is a decomposition of the signal in PCs (main components). Up to the third main component is used as a detection signal. It is square, smoothed and compared to the threshold. It is intended to obtain the PCA later by means of a neural network with the aforementioned structure.

5. EXPERIMENTS

Tests were performed to obtain the PCA and its analysis in the BD [18] of the Children's Hospital of Boston. It contains 22 subjects between the age of 1-22 years, of them 5 men and 17. It contains each taking, one hour digitalized signal in the majority of cases, occupying the assembly 10-20 to obtain it. The fig. 3 shows the grouping of the data by age and fig.4 by sex.

	Edad	Hombre	Mujer
1	1-5	2	3
2	6-10	0	7
3	11-15	1	6
4	16-22	2	2

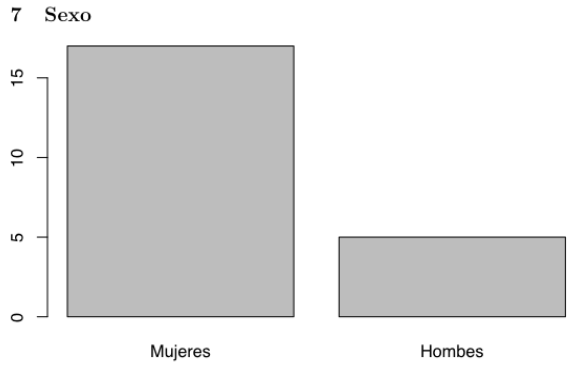


Fig. 4. Registered Data

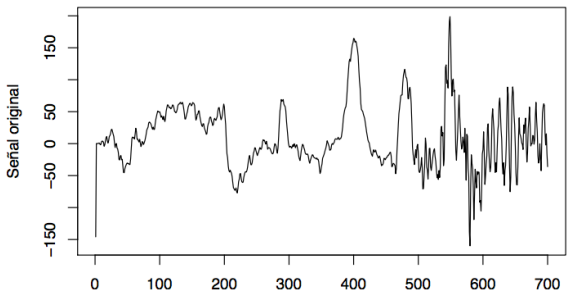


Fig. 5. Original Patient Sign 1

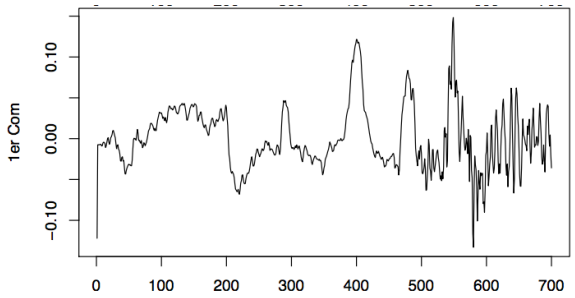


Fig. 6. 1st Principal Component of Patient 1

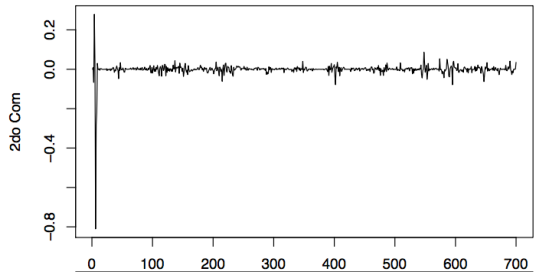


Fig. 7. 2nd Principal component of Patient 1

The images 5,6, 7 and 8 show the original signal and the main vectors of the signal of the first patient.

An analysis was made of how the main value of each patient's intake changes with respect to time. The fig. 9 shows the changes of the main value with respect to the

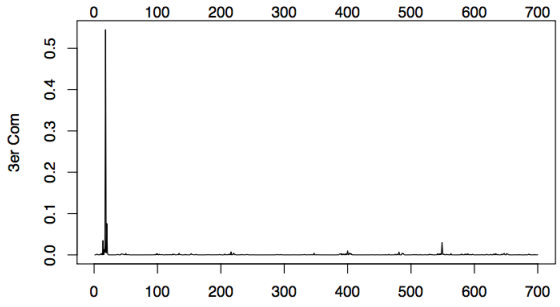


Fig. 8. 3rd Patient Principal Component of Patient 1

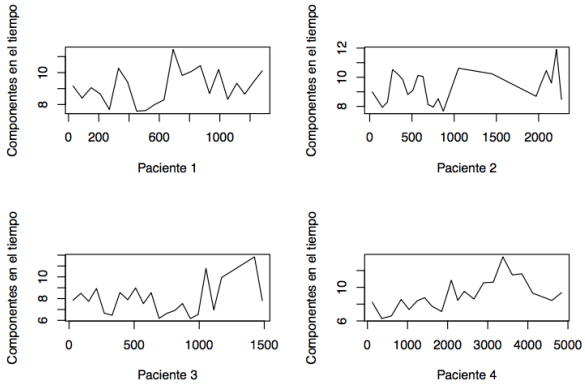


Fig. 9. Log of own values of 4 patients

time represented in minutes of 22 shots taken per patient showing the first 4.

6. CONCLUSIONES

This analysis was performed on people diagnosed with epilepsy and it is intended to perform these analyzes in healthy people to compare the results with those obtained.

By the data manipulation, alternatively, the PCA can be estimated through a neural network (NN) to extract the first main components, the extracted features are used for the successive classification.

7. BIBLIOGRAPHY

[1] R. Grishman, The Oxford handbook of computational linguistics, Oxford University Press, 2012.

[2] J. R.J. Schirra, Foundation of Computational Visualistics, Dt. Unive.-Verlag, 2005.

[3] Abdulhamit Subasi y M Ismail Gursoy. Eeg signal classification using pca, ica, lda and support vector machines. Expert Systems with Applications, 37(12):8659–8666, 2010.

[4] Celine Bugli y Philippe Lambert. Comparison between principal component analysis and independent component analysis in electroencephalograms mode- lling. Biometrical Journal, 49(2):312–327, 2007.

[5] X Liu, G Wang, J Gao, y Q Gao. A quantitative analysis for eeg signals based on modified permutation-entropy. IRBM, 2017.

[6] Li-Chen Shi, Ruo-Nan Duan, y Bao-Liang Lu. A robust principal component analysis algorithm for eeg-based vigilance estimation. En Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, p.p. 6623–6626. IEEE, 2013

[7] Kavita Mahajan, MR Vargantwar, y Sangita M Rajput. Classification of eeg using pca, ica and neural network. International Journal of Engineering and Advanced Technology, 1(1):80–83, 2011.

[8] Hyekyung Lee y Seungjin Choi. Pca+ hmm+ svm for eeg pattern classification. En Signal Processing and Its Applications, 2003. Proceedings. Seventh Interna- tional Symposium on, tomo 1, p.p. 541–544. 2003.

[9] T. Chang, T. N Tsuchida, "Conventional (Continuous) EEG Monitoring in the NICU." Current Pediatric Reviews 10(1), Bentham Science Publishers: 2-10, 2014.

[10] F. Sharbrough, G. E. Chatrian, R. P. Lesser, H. L?ders, M. Nuwer, T. W. Picton, "American Electroencephalographic Society. Guidelines for standard electrode position nomenclature", Journal of Clinical Neurophysiology, 8(2):200-202,Raven Press, Ltd., New York, 1991.

[11] J. D. Singer, J. B. Willett,Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence, Oxford University press,Inc., New York, 2003.

[12] H. Pell?, P. Ciuciu, M. Rahim, E. Dohmatob, P. Abry, V. Van Wassenhove, "Multivariate Hurst Exponent Estimation in FMRI. Application to Brain Decoding of Perceptual Learning", 13th IEEE International Symposium on Biomedical Imaging, Apr 2016, Prague, Czech Republic. 2016.

[13] N. Kannathal, U. Rajendra Acharya, C.M. Lim, P.K. Sadasivan, "Characterization of EEG–A comparative study", Department of ECE, National University of Singapore, Singapore ECE Division, Ngee Ann Polytechnic, 2005.

[14] A. A. Putilov, " Principal component analysis of the EEG spectrum can provide yes-or-no criteria for demarcation of boundaries between NREM sleep stages", Published online 2015 Mar 10. http://doi.org/10.1016/j.slsci.2015.02.004

[15] P. Berka, J. Rauch, D. A. Zighed, Data Mining and Medical Knowledge Management: Cases and Applications, IGI Globa, Hershey, New York, 2009.

[16] H. E. Hurst and American Society of Civil Engineers, Hydraulics Division. Long-term Storage Capacity of Reservoirs, American Society of Civil Engineers, 1950.

[17] E. E. Rodríguez, E. Hernández-Lemus, B. A. Itzí-Ortiz, I. Jiménez, P. Rudom?n. "Multichannel Detrended Fluctuation Analysis Reveals Synchronized Patterns of Spontaneous Spinal Activity in Anesthetized Cats", In: PLoS, 2011. https://doi.org/10.1371/journal.pone.0026449

[18] Base de datos ocupada para la realizacion de pruebas. A team of investigators from Children's Hospital Boston (CHB) and the Massachusetts Institute of Technology (MIT) created and contributed this database to PhysioNet. The clinical investigators from CHB include Jack Connolly, REEGT; Herman Edwards, REEGT; Blaise Bourgeois, MD; and S. Ted Treves, MD. The investigators from MIT include Ali Shoeb, PhD and Professor John Guttag. https://physionet.org/pn6/chbmit/

Artificial Intelligence Techniques to create Distance Learning Adaptive Styles in Moodle platform.

Angel Gaspar May Uuh[1], Dr. Alejandro del Rey Torres Rodriguez[2], y Dr. Rajesh Roshan Biswal[3]

Instituto Tecnológico Superior de Misantla, Maestría en Sistemas Computacionales
e-mail: 162t0004@itsm.edu.mx
Tecnológico Nacional de México, Instituto Tecnológico de Ciudad Victoria, División de Estudios de Posgrado e Investigación

Abstract:

The Q-Learning allows learning the optimal behavior in tasks that require the selection of sequential actions. This l earning method is based on the interactions between an agent and this environment. Through repeated interactions with the environment and receiving rewards, the agent learns what actions are associated with the highest cumulative reward, using a neural network to represent the agent's evaluation function, as well as the time difference algorithm, which is used to train the neural network.

***keywords*—Artificial Intelligence, Adaptive Learning, Q-Learning, Reinforcement Learning.**

I. Introduction

In this paper, one of the topics which in many countries and particularly in Mexico has been one of the greatest headaches of students, has been addressed: the study, understanding and application of mathematics through online courses, a theme that is most evident in the University stage where the degree of complexity and abstraction of problems is higher and requires a greater level of understanding.

The online distance education can be understood as one that is carried out outside the (physical) educational classrooms, through the use of the internet, web pages, chat, discussion forums or videoconference, although at a certain moment it could include classroom activities. [1]

It is a meeting point that allows students (apprentices) and teachers (facilitators) to interact with the purpose of achieving a common instructional goal, without the limitations of space or time, which in addition to facilitating the distribution of information, related to a particular area of knowledge, it allows for the interaction at a distance between the interconnected users through a computer network that involves the teaching-learning process. [2]

In this modality the teacher-student interaction is given by means of technological resources reducing the possibility of body language, gestural and omitting communication in real time. Also, it is important to mention that in this modality, the activities are focused on the student, while the teacher transmits his knowledge from the comments and observations made to the students. [3]

This modality has three main characteristics, i.e., its mediated by computer, it is not given in real time, and has a set of supports available. [4]

A. Online Education Resources

Online education has didactic resources that promote and facilitate the teaching-learning processes (E-A). These resources must meet certain characteristics such as clarity, simplicity, accessibility, functionality, so that the E-A process is successful.

B. Adaptive Education

The "Adaptivity" is the adjustment of one or more characteristics of the learning environment. The idea of "personalized learning" and, moreover, adaptive learning is to help meet the learning process needs of each student, allowing students to choose the steps or routes they want to take instead of being taxed. An effective tecnology can assemble and adapt the entire learning management system, not only part of the content at a time.[5]

II. Description of the virtual environment

A. Artificial Intelligence in Education

In order to carry out a virtual teaching-learning process, it requires software that integrates the main tools offered by the Internet, and allows; the development of interactive virtual courses, tutoring, follow-up for students, be flexibles, intuitive, friendly and especially adaptive where students learn, share experiences and knowledge with the rest of the virtual community.

The student's learning of mathematical knowledge is not uniform. Its acquisition is achieved by a code that the student makes from the information received.

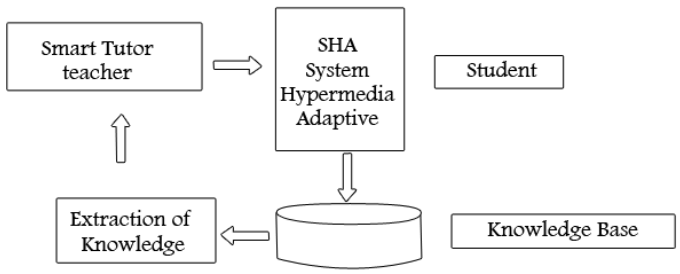


Figure 1. Architecture of Adaptive Hypermedia Systems.

In the figure 1, it is possible to observe the modules that will allow the extraction of knowledge data and its subsequent feedback to the Teacher and / or Intelligent Tutor Module, to adapt the contents through the application of Artificial Intelligence as a function of the interpretation, evaluation of the results obtained and the utilization of the discovered knowledge, achieved in the teaching-learning process of the student.

This mechanism will allow to feed back the whole learning process of the student and to update according to the obtained results, thereby facilitating continuous improvement of the teaching-learning process of the contents that will be imparted.

The AI technique that is proposed to apply is the Fuzzy Logic, which allows to store information on the student's knowledge, as well as information on the evolution within the system, which will allow the generation of material, content and activity that the student should perform according to his learning style.

III. Technological Challenge

In Mexico, it is a reality that the deficiencies in the areas of natural sciences, especially those of Mathematics, is becoming more evident, since most of the students have problem in the understanding and analysis of basic concepts in Mathematics. Upon reaching the professional level the students lack solid fundamentals and knowledge which make them less competitive. [6]

In view of this problem and with the intention of solving it, the MOOCs arise (David Wiley being the author of the first conceptual MOOC in 2007 for the University of Utah), which is the acronym in English for Massive Online Open Courses, i.e, it is a distance-course, accessible by internet to which anyone can register and practically there is no limit of participants.In addition to the traditional course materials, such as videos, lectures and questionnaires, MOOCs provide interactive user forums that help build a community for students, teachers and teaching assistants. [7]

The MOOC, although in its beginning, had a very successful acceptance, and in 2011 managed to have more than 160,000 people enrolled in an artificial inteligence course offered by Sebastian Thrun and Peter Norvig at Stanford University through a startup called Know Labs (currently Udacity). However, it has the disadvantage that the courses and the system of evaluation are the same for all students, and so, taking into consideration that each person has a particular style or method of learning, this technique falls into the same traditional method of classroom lecture.

In the search of alternative methods to solve the problem, adaptive educational platforms as well as online education and learning aids like: E-Learning and Q-Learning has become a viable solution.

The E-Learning, or Electronic Learning consists of education and training through the Internet. This type of online teaching allows the user to interact with the course material using various computer tools. The same brings together different technologies and the pedagogical aspects of teaching and learning. [8]

The E-Learning provides educational programs through educational means and learning systems (computers, cell phones, etc.) in order to provide students with educational material through tools or applications (web pages, discussion forums, etc.). [9]

While Q-Learning is a Reinforcement Learning method that allows solving sequential decision problems [10], and serves as a complement to the E-Learning method.

In early 2015 Professor Jo Boaler of Stanford University in an article criticized the traditional way in teaching mathematics in schools. Since children between 8 to 10 years are forced to memorize including multiplication tables of 11 and 12, which most of the times cause them, great deal of anxiety because of their delayed and monotonous nature, thereby causing the student to lose interest in classes. [11]

One of the main reasons for poor performance in mathematics is that students often become dis-interested by the traditional way of teaching in the classroom, or the teachers are not skilled enough.

According to Ruiz chavarria, it was observed that students in the United States perform the same basic procedure in solving most mathematical problems without entering into an understanding at a deeper level, thus revealing difficulties with problems in which further analysis and reasoning in abstract problems is needed. [12].

In Mexico, as in other countries, both public and private educational institutions, have joined efforts to address these deficiencies.

online educational platforms focused on more specific or specialized topics at the university level has surfaced. However, these platforms suggest topics, focused on any area or field focusing only on the teaching materials without focusing on the style or method of learning of the student.

For the above, the use of Artificial Intelligence is proposed, which allows solving such rigidity in terms of information processing, which is nothing more than the uncertainty that is generated in the decision-making process as a human being will do, thus giving a "human sense" to the technology whereby catching the attention to Adaptive Education.

IV. Materials and Methodology

The proposed methodology is based on the application of the algorithm Q-Learning or learning by reinforcement, which consists of learning what actions to perform, given the current state of the environment, with the aim of maximizing a numerical reward signal [13], which requires a mapping of situations of actions. The learning system must discover by itself which are the actions that give it to gain more. In the most interesting and difficult cases, actions can affect not only the immediate reward but also the following situation, and thus affect the following rewards.

These two characteristics, trial and error, and delayed rewards, are the two most outstanding characteristics of reinforcement learning. The learning is done in most of the algorithms through interaction between the agent and his environment, where the agent must exploit the knowledge he currently has to obtain rewards, but also has to explore in order to execute better actions in the future.

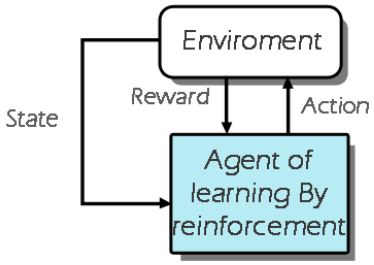


Figure 2. Basic architecture of a reinforcement learning system.

However, in Reinforcement Learning, the state transition probability and the reward function are not known, which prevents the use of any value or policy iteration algorithm, But, instead, the agent can be made to interact with the environment several times in order to know the value of the rewards obtained and the states visited, after performing different actions in different states.

The figure 2 Provides the visual idea of the flow of the algorithm by reinforcement.

This algorithm represented in the figure 3 consists of an agent interacting with an environment. The agent observes the state of the environment $s(t)$ and pursue an action $a(t)$. In the time t to go to state $s(t+1)$ in the instant $t+1$, and the environment gives a reward $r(t+1)$ at agent. Over time, the agent learns to carry out actions that lead to greater cumulative rewards.

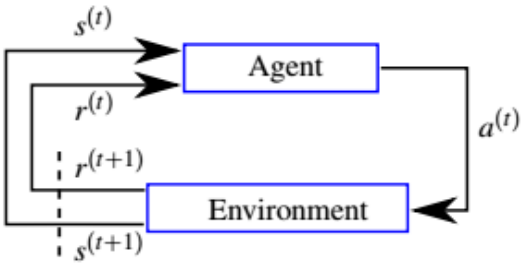


Figure 3. Methodological Environment of Q-Learning.

In this method, the states S and actions A are generally known. While to obtain the transition models given by $T(s,a,s') = P(s'|s,a)$ Which serves to obtain the initial states and from there to start to perform an action, the Reward Function is interpreted as $R(s)oR(s,a,s')$, it would be necessary to carry out tests to know the possible states and actions to be carried out, in order to find optimal policies represented by $\pi(s)$.

The sequential process of decision making consists, therefore, of a sequence of states $S = s0, s1, ..., sT$ and a sequence of actions $a = a0, a1, ..., aT$ for steps of time $T = 0, 1, ..., T$, where the state $s0$ is considered as the initial state.Feedback can be provided to the agent at each time step t in the form of a reward based on the particular action pursued.This feedback can be rewarding (positive) to pursue beneficial actions that lead to better results, or can be aversive (negative) to carry out actions that lead to worse results.

V. Conclusions

Reinforcement learning consists of learning to decide, in the face of a situation where action is most appropriate to achieve the goal. While intelligent behavior is the element present in many systems of daily life, from a clock to complex devices like a robot from a car assembly company.

For a system to be intelligent it can define two characteristics:

- The learning of a task by an agent is done through the iterative process of trial and error in the environment where the agent interacts.
- The way the environment informs the agent about whether the task he is learning is doing right or wrong.

Reinforcement learning is based on the following elements.

- A set of states S.
- A set of agent actions A.
- A set of reinforcing signals $R = 0, 1$

In the figure 4,The flow of the model diagram is presented by reinforcement, where the agent is connected in its environment via perception and action and at every instant the agent perceives from the environment I in the state are find, then the agent decides to take an action that generates

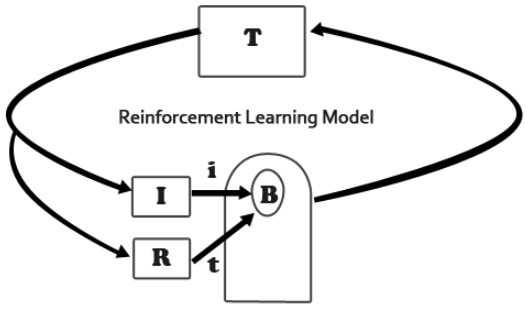


Figure 4. Algorithm Flow Model of Reinforcement Learning.

an output which changes the state of the environment and the value of that transition of the state is communicated to the agent through a signal of reinforcement. As long as the B behavior chooses actions that increase the sum of all the reinforcement signals that are received over time t. The diagram includes some functions in principle could be unknown as the input function i, which determines how the agent perceives the state of the environment,The function T that performs the state transitions and R that calculates the reinforcement that the agent receives at each moment. In General Terms, these functions determine the flow of Markov's dedication process.

This decision aims to find a policy that maximizes the long-term reinforcement measure. Although this process has the disadvantage that it assumes the environment must be redefined and implies an order between the states, Which involves the use of discretization that limits the number of states of the environment to a viable number from the point of view of memory storage and size of the set of test cases required for learning and is called state-action pairs.

For future work it is intended to use IA techniques, such as the perceptron method, which consists of inputs, connections, network function, activation and output.

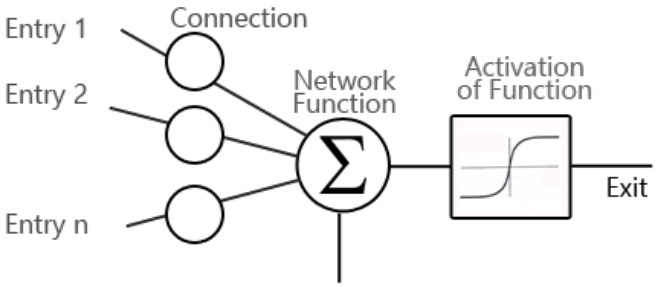


Figure 5. Example of a Simple Perceptron.

Given its characteristics, the neural networks is a method that could facilitate and efficient learning by reinforcement, due to the input patterns allows to obtain results very close to the desired, and the greater the training number, the algorithm would become more efficient and Especially adaptive.

The neural network interacts with the environment in two ways: 1) the search of actions, 2) and the learning.More specifically, the action selection procedure proceeds through the neural network observing the current state of the environment and using the state value estimates of the potential later states to select the action to be pursued. When the agent selects actions that always correspond to next state values with the largest value, it is said that the agent follows a policy of selecting greedy actions. The learning procedure uses the received rewards, which are a direct result of the actions performed, to adjust the network weights, which improves the state value estimates. In other words, the rewards received are those that guide the learning of state values.

The construction of the environment for use within a reinforcement learning framework is specific to the task at hand. The state of the environment is represented numerically by a state vector x, which serves as input to the neural network.The methodology used to train the agent to learn a domain can also have a significant effect on the efficiency with which the agent learns.

Rewards play an important role in learning temporal differences, since the letter allow the agente to learn by guiding estimates of state values toward their true values. Reward values in particular should be based on the possible outcomesof an episode of agente-enviroment interaction, and reward values should reflect the appropriateness of the results.

The agent can be represented by a neural network that acts as a function approximator. The neural network tries to learn the true values of each state interacting with the environment and refining the weights so that the outputs of the neural network become better approximations of the true state.

So, the implementation of reinforcement learning is based on a framework that comprises two main entities: an agent and an environment. The agent, represented by a neural network, to learn the approximation of state value function.Once this neural network and its learning algorithm are created, it can be easily used in many domains. The environment, on the other hand, is specific to the particular application and requires a more extensive development.

The state coding scheme is fundamental for any reinforcement learning implementation and this is what the agent (neural network) must learn. Raw coding schemes are useful to facilitate human interpretation, but may not be optimal for efficient learning through a neural network.Coding schemes that include special, expert features that are relevant to the final learning goal can be very useful for the neural network. A carefully constructed coding scheme can also reduce the state space of the environment, which facilitates the exact approximation of the function.

References

[1] Castillo, R. Q., *Evaluación del aprendizaje en la educación a distancia en linea*, Revista de Educación a Distancia, 2005, Pags. 15.

[2] Tics y Tecnología. *http://diariotecnologias.bligoo.com.mx*.

[3] George Pappas, E. L., *Monitoring Student Performance in Online Courses: New Game - New Rules* Journal of Distance, 2001, Págs. 66-71

[4] Churchill, A., *Ensuring quality in online higher education courses* Center for Education Policy, 2004, Págs.18

[5] Bill Bilic., *https://www.d2l.com/es/blog/que-es-el-aprendizaje-adaptativo*, 2015.

[6] Trujillo Lara, José Manuel., *Impacto del uso de Applets en el Aprendizaje de Física Básica en Alumnos Universitarios*, 2016, Págs 110.

[7] Universitat Autónoma de Barcelona., *Impacto del uso de Applets en el Aprendizaje de Física Básica en Alumnos Universitarios*, 2017, http://www.uab.cat/web/estudiar/mooc/-que-es-un-curso-mooc-1345668281247.html.

[8] e-ABC., *Impacto del uso de Applets en el Aprendizaje de Física Básica en Alumnos Universitarios*, 2017, http://www.e-abclearning.com/definicione-learning.

[9] Antonio Manuel Cuevas García., *El mercado de e-Learning en México*, 2012, Págs. 42.

[10] Alicia Marcela Printista, Marcelo Luis Errecalde, Cecilia Inés Montoya., *El mercado de e-Learning en México*, 2012, Págs. 42.

[11] Ana Torres Menárguez., *Los alumnos que huían de las matemáticas*, 2016, http://economia.elpais.com/economia/2016/04/24/actualidad/1461527206_970734.html

[12] Angel Ruiz., *Los Estándares en la educación Matematica de los Estados Unidos: Contexto, Reforma y Lecciones*, 2017, Págs. 16.

[13] Richard S. Sutton.,Andrew G. Barto, *Reinforcement Learning: An Introduction.*, 1998.

Traducción de lenguaje de señas usando algoritmos de visión por computadora y reconocimiento de patrones ⋆

Eduardo Mancilla Morales *
Simón Pedro Arguijo Hérnandez **

* *Instituto Tecnológico Superior de Misantla , Km. 1.8 Carretera a Loma del Cojolite, C.P. 93821, Misantla, Veracruz, México (e-mail: 152T0736@itsm.edu.mx).*
** *Instituto Tecnológico Superior de Misantla , Km. 1.8 Carretera a Loma del Cojolite, C.P. 93821, Misantla, Veracruz, México (e-mail: sparguijoh@itsm.edu.mx).*

Resumen
Las personas con problemas auditivos y/o de habla, se comunican mediante el lenguaje de señas. Sin embargo, no facilita la comunicación con las personas que desconocen el lenguaje de señas. Se va a realizar un sistema de traducción automática de señas enfocado al lenguaje de señas mexicano(LSM). El sistema no hace uso de guantes, ni de marcas en las manos, tampoco hace uso de sensores como Kinect o cámaras 3d, solo hace uso de una cámara. En primer lugar, la imagen es capturada con la cámara, posteriormente se realiza preprocesamiento de la imagen reduciendo el ruido, y mejorando la imagen con ecualización del histograma. En segundo lugar, se realiza la detección de la mano mediante algoritmos de enfoque boosting como adaboost, floatboot, gentleboost o usando algoritmos basados en color combinados con detección de movimiento. En tercer lugar, se realiza la segmentación de la imagen, para eliminar la parte del fondo, con el objetivo de mejorar la extracción de las características. Cuarto lugar se realiza la extracción de las características mediante algoritmos que detecten la orientación y la forma invariantes a escala y rotación. Por último, se realiza el reconocimiento del gesto mediante el uso del clasificador.

Keywords: detección de mano, lenguaje de señas, reconocimiento de señas, lenguaje de señas mexicano, extraccion caracteristicas.

1. INTRODUCTION

Las personas sordomudas que no pueden comunicarse, llegan a sentirse aisladas del mundo exterior. Como medio de socialización y mecanismo compensatorio, las personas con problemas auditivos y de habla han desarrollado el lenguaje de señas. Este lenguaje facilita la comunicación entre personas sordomudas, pero no facilita la comunicación con el resto de las personas, sobre todo las que desconocen el lenguaje. según Rautaray and Agrawal (2015) la semántica de los gestos depende del país en donde se encuentra, el significado depende del lugar, la misma seña puede tener diferentes significados. También el alfabeto varía de acuerdo al país, el alfabeto mexicano es muy parecido al lenguaje de señas americano, pero aun así tiene variaciones, como la incorporación de la letra ñ y ll, también cambian un poco en la representación algunas letras. Menciona Rautaray and Agrawal (2015) que hay dos tipos de gestos con las manos: los estáticos y dinámicos. Movimientos estáticos: son definidos como orientación y posición de la mano, durante algún tiempo sin mover la mano. Si la mano varía durante ese tiempo, es llamado gesto dinámico.

El primer paso del sistema es la detección de la mano. Existen los siguientes enfoques:

Color Estos usan generalmente espacios de color, según Shaik et al. (2015) se realiza la conversión para separar la cromaticidad de la luminancia, para hacerlos menos invariantes a la iluminación.

Forma Se basan en extraer los contornos del objeto, lo que permite independencia del punto de vista, del color de la piel e iluminación. Sin embargo, presenta problemas con la oclusión, o la ocultación dependiendo del punto de vista. La extracción del contorno basado en los bordes de la imagen, puede detectar la mano, pero también objetos irrelevantes del fondo.

Valores del pixel Se basan en imágenes en apariencia y textura. algunos usan escala de grises. según Rautaray and Agrawal (2015) técnicas basadas en machine learning con un enfoque llamado boosting han demostrado robustos resultados en la detección de cara y manos. Se basan en el principio de que una alta clasificación se puede generar a partir de la combinación lineal de varios inexactos o débiles clasificadores.

⋆

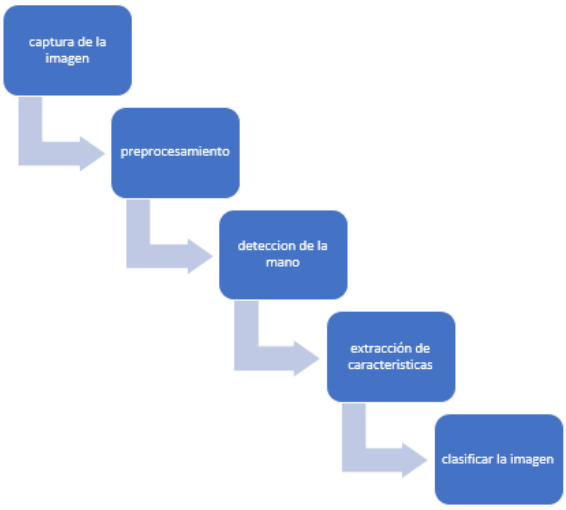


Figura 1. metodología del sistema

Modelos 3d Usan modelos 3d para la detección de la mano. Una de las ventajas de este enfoque, es la detección independiente del punto de vista. La postura de la mano es estimada entre el modelo 3d y las características observables de la imagen.

Movimiento El movimiento de una de las pistas utilizadas por varios enfoques para la detección de la mano. Debido a que requiere un entorno muy controlado, desde que asume que el único movimiento que se realiza es el de la mano. Se requiere una cámara estática, y se ha combinado con otras pistas. Es útil para distinguir objetos que tienen el color de la piel, pero pertenecen al fondo, debido a que los pixeles del fondo no tienen mucho movimiento. Las manos tienen más movimiento que otros objetos, como la cara, la ropa y el fondo.

2. PROCEDIMIENTO

2.1 Pasos a seguir

El sistema primero obtiene una imagen, ya sea almacenada en el disco duro, tomada de la cámara o un frame de un video. Fig. 5 muestra de manera general los pasos a seguir para desarrollar el sistema.

Para la detección de la mano se usará los algoritmos de tipo boosting entre los cuales están los algoritmos adaboost, floatboost. El algoritmo adaboost para el entrenamiento usa un conjunto de imágenes que consisten de ejemplos negativos y positivos, en este caso manos y fondo, que se asocian con una clase. Clasificadores débiles con agregados secuencialmente para disminuir el límite superior del error de entrenamiento. Sin embargo, este método puede tener un excesivo número de clasificadores débiles. El algoritmo floatboost extiende al original algoritmo adaboost, este remueve algún débil clasificador, si este no contribuye a disminuir el error de entrenamiento.

Para la extracción de características según Pisharady and Saerbeck (2015) se caracterizan la orientación, la forma, los ángulos de los dedos, la posición relativa al cuerpo, contexto del ambiente. Se analizarán los algoritmos: HOG, SHIFT, SURF, PCA, local topological descriptor y bordes para determinar la mejor opción para extraer características.

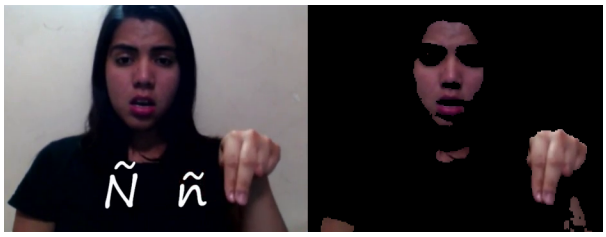


Figura 2. Izquierda: Frame tomado del video 1 con fondo de un solo color. Derecha: Segmentación por color de piel usando el espacio de color YCbCr

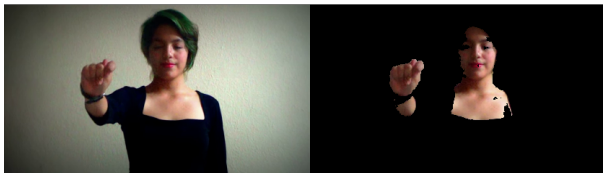


Figura 3. Izquierda: Frame tomado del video 1 con fondo de color uniforme. Derecha: Segmentación por color de piel usando el espacio de color YCbCr

Los algoritmos de reconocimiento que mas se usan son: k-means, k-nearest neightbor, Mean shift clustering, support vector machine, hidden markov model, dinamic time wraping, time delay neural network, finite state machine. Sin embargo, aún no se ha seleccionado alguno a usar, debido a que aún se están analizando las posibles alternativas.

2.2 Resultados

Detección de la piel Es dificil desarrollar un metodo uniforme para la detección de la piel de los humanos, porque el tono varia drasticamente de una region a otra. Segun Shaik et al. (2015) se han usado una gran variedad de espacios de color, sin embargo el espacio de color rgb no es muy preferible para la deteccion basada en color, porque la informacion de color e intensidad está mezclada y no tiene caracteristicas uniformes. Llegó a la conclusion que el espacio de color hsv es mejor para imagenes simples con fondo uniforme, pero yCbCr puede ser aplicado a imagenes con colores complejos y con iluminacion irregular.

La segmentacion de realiza en el espacio de color ycbcr por el umbral cr entre 150 y 200, cb entre 100 y 150

La figura (2) se usó un video que contiene un fondo de un solo color, la segmentación tiene buenos resultados, solo ignora algunos tonos de la piel que tienen mucha sombra, la figura(3) es un video con color de fondo uniforme, los resultados tambien son buenos, pero algunas partes que estan muy iluminadas son ignoradas,la figura(4) es una video con fondo estático, con unas letras que generan ruido. Muestra buenos resultados , pero aún tiene pequeñas secciones que pertenecen a las letras. La figura(5) se muestra el video con fondo parecido al color de la piel, presenta algunos problemas, ya que detecta el fondo como piel. Surgen problemas cuando el fondo tiene un color parecido a la piel.

Figura 4. Izquierda: Frame tomado del video 2 con fondo de un solo color. Derecha: Segmentación por color de piel usando el espacio de color YCbCr

Figura 5. Izquierda: Frame tomado del video 3 con un fondo un poco parecido al color. Derecha: Segmentación por color de piel usando el espacio de color YCbCr

3. CONCLUSION

Los pasos para resolver la deteccion del alfabeto de lenguaje de señas son dependientes uno del otro. Por lo tanto si un paso no se realiza correctamente puede afectar el rendimiento de los demas pasos. La segmentacion basada en el color resulta muy buena, ya que remueve gran parte del fondo de una manera muy eficiente, sin embargo cuando en el fondo se encuentran objetos de color parecidos a la piel, surgen problemas, para solucionar esto, se debe ajustar el rango de valor para la detección del color o usar algún manera de diferenciar el fondo, como la detección de movimiento.

REFERENCIAS

Pisharady, P.K. and Saerbeck, M. (2015). Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141, 152–165.

Rautaray, S.S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1), 1–54.

Shaik, K.B., Ganesan, P., Kalist, V., Sathish, B., and Jenitha, J.M.M. (2015). Comparative study of skin color detection and segmentation in hsv and ycbcr color space. *Procedia Computer Science*, 57, 41–48.

Aprendizaje automático para la clasificación de nefropatía diabética mediante KDIGO*

Francis Susana Carreto Espinoza.*
José Antonio Hiram Vásquez López.**

*Instituto Tecnológico Superior De Misantla, Km 1.8 carretera Loma del Cojolite, Misantla (e-mail: 152T0731@itsm.edu.mx).
** Instituto Tecnológico Superior De Misantla, Km 1.8 carretera Loma del Cojolite, Misantla (e-mail: jahvazquezl@itsm.edu.mx).

Resumen

La diabetes es una de las principales causas de muerte no solo en México sino a nivel mundial y una de las complicaciones que se presenta con mayor frecuencia es la nefropatía diabética. La nefropatía diabética ocasiona daño o lesión en los riñones y se ha clasificado en cinco estadios que determinan el nivel de afectación del mismo. Por lo tanto, es importante detectar la nefropatía en pacientes diabéticos cuando se encuentra en una etapa temprana (estadios I, II y III) para proporcionar un tratamiento óptimo y mejorar su calidad de vida. El objetivo de este estudio es dar a conocer un modelo predictivo de nefropatía diabética empleando técnicas de aprendizaje automático, mismo que servirá de apoyo a los médicos no especializados en esta área para proporcionar al paciente un tratamiento adecuado y oportuno. Para el desarrollo y evaluación del modelo propuesto, el conjunto de datos esta constituido por 55 pacientes que tiene 23 atributos determinados como factores de riesgo para padecer esta enfermedad, los atributos se determinaron con ayuda del experto y el conocimiento adquirido al revisar el estado del arte. Hasta el momento los resultados obtenidos mediante el software WEKA, al aplicar algunos métodos de selección de atributos han encontrado como subconjunto de atributos la urea, creatinina, proteinas_totales y filtrado_glomerular por mencionar los más representativos y un porcentaje de clasificación comprendido entre 60 % y 96 %.

Palabras clave: Selección de atributos, clasificación KDIGO, modelo predictivo, WEKA, aprendizaje automático, nefropatía diabética.

1. Introducción

La diabetes es una de las principales causas de muerte, debido principalmente a la obesidad y el sobrepeso. Hasta hace poco 7 de cada 10 individuos con edades mayores a los 20 años tenían problemas de exceso de peso (obesidad y sobrepeso) lo que significa un 72.5 % de la población [5].

Las complicaciones de la diabetes pueden ser de dos tipos: (i) macrovasculares y (ii) microvasculares; en el primer tipo (i) se encuentran las enfermedades cardiovasculares como infarto al miocardio, accidentes cerebrovasculares e insuficiencia circulatoria en los miembros inferiores, mientras que los padecimientos del tipo (ii) son retinopatía, pie diabético, neuropatía y nefropatía.

La nefropatía diabética es una enfermedad crónica que surge por daños en los riñones y es considerada una de las principales causas de muerte en pacientes diabéticos de los dos tipos (1 y 2) [2], tiene como consecuencia la Enfermedad Renal Crónica (ERC) y si no se detecta a tiempo se convierte en una Enfermedad Renal Crónica Terminal (ERCT) donde los tratamientos sustitutivos son de costos muy elevados.

Un mal control de la microalbuminuria[1], la glucosa, la obesidad y el sobrepeso, además de la hipertensión arterial son las principales causas que ocasionan que el daño renal

[1]La microalbuminuria es la primera manifestación clínica de la nefropatía, hay que destacar que es un elemento que se considera un buen predictor para la detección del daño renal

aumente progresivamente y no se detecte hasta que los daños son graves.

En los 80's se utilizaba como medida para determinar la lesión en los riñones la excreción de albuminuria[2], misma que el autor Mongensen utlizó para clasificar la nefropatía diabética en 5 estadios. Sin embargo con el paso del tiempo y de acuerdo a las guías prácticas clínicas de Enfermedad renal: Mejora de los resultados mundiales (KDIGO, por sus siglas en inglés *Kidney Disease: Improving Global Outcomes*) la medida que se utiliza para establecer el daño renal consiste en el índice de filtración glomerular (IFG) y apartir de esta se han definido los estadios como se muestra en la tabla 1 [4].

Estadio	FG	Descripción
1	>90	Daño renal con FG normal
2	60-89	Daño renal y ligero descenso de FG
3a	45-59	Descenso ligero-moderado de FG
3b	30-44	Descenso moderado de FG
4	15-29	Descenso grave de FG
5	<15	Prediálisis
5D	Diálisis	Diálisis

Tabla 1: Clasificación KDIGO

El objetivo principal de este estudio es diseñar un modelo predictivo de nefropatía diabética usando técnicas de aprendizaje automático que sirva como herramienta a los médicos para que se le proporcione al paciente un tratamiento precoz en una etapa adecuada, y se retrase el progreso de la enfermedad.

2. Materiales y Métodos

Los materiales que se utilizan como apoyo para el diseño del modelo predictivo de nefropatía diabética son un conjunto de datos extraídos de los expedientes médicos del Hospital General de Misantla mediante un estudio retrospectivo del periodo 2015-2016, las guías de prácticas clínicas KDIGO para el manejo de la enfermedad renal crónica y la ecuación(1) para calcular el filtrado glomerular. Por otra parte, se encuentra la metodología conformada por los métodos adecuados para alcanzar el objetivo de la investigación.

$$FG = 175 * Cr^{-1,154} * edad^{-0,203} * (si\ es\ mujer(0{,}741)) \quad (1)$$

[2]Esta clasificación fue realizada por el autor Morgensen en 1983

2.1. Conjunto de datos

El conjunto de datos que se tomará en cuenta para llevar a cabo esta investigación consta de las características de los pacientes con diagnóstico nefrótico que asisten a consulta al Hospital General de Misantla en el periodo 2015-2016 donde se encontraron 55 pacientes de los cuales 32 son mujeres y 23 hombres, los cuales que se han clasificado en alguno de los estadios de nefropatía diabética según la clasificación KDIGO. Cada una de las instancias cuenta con 27 atributos como son urea, creatinina, hemoglobina glicosilada, proteinuria, sexo y edad por mencionar algunos.

2.2. Metodología

Para el diagnóstico de la nefropatía diabética actualmente se realizan pruebas médicas que miden filtración glomerular, hemoglobina glicosilada o la creatinina. Sin embargo, no todos los médicos son capaces de determinar a partir de esos parámetros que el paciente tiene esta complicación.

Por lo anterior, se diseñará un modelo predictivo de nefropatía diabética que se muestra en la figura 1, el cual se trata de un gráfico enriquecido (Rich Picture) que consiste en los pasos a seguir para llevar a cabo el modelo.

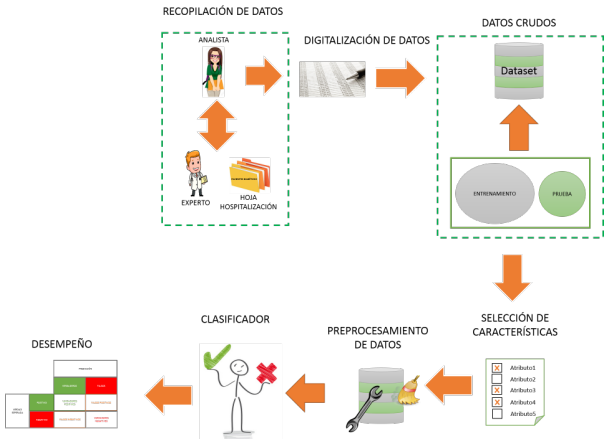


Figura 1: Etapas del modelo predictivo de nefropatía diabética.

A continuación se detallan cada una de las actividades que se van a realizar para lograr el desarrollo del modelo predictivo.

1. **Recolección de los datos**
   - Buscar una institución o médico con atención a pacientes diabéticos que padezcan la complicación nefrótica.
   - Con ayuda del experto identificar las variables que definen la nefropatía en pacientes diabéticos a partir de la hoja de hospitalización o ingreso que se realiza al asistir a la consulta.
   - A partir del conjunto de datos obtenido anteriormente una tarea importante es determinar cuales serán los atributos, instancias y clases.
2. **Digitalización de datos**
   - Realizar la digitalización de los datos que serán almacenados en la base de datos ya que la institución que los proporcionará realiza el seguimiento y control de los pacientes mediante un registro en expedientes físicos.
3. **Análisis de los datos**
   - Definir los dos conjuntos a utilizar en el modelo predictivo, es decir, seleccionar las instancias que pertenecerán al conjunto de entrenamiento y al conjunto de prueba. El conjunto de entrenamiento consiste en los datos que se utilizarán para alimentar el clasificador y logre de esta manera aprender con base en los patrones encontrados, por otra parte se tiene el conjunto de prueba que son los datos a utilizar para comparar con los patrones encontrados y determinar la clasificación a la que pertenecen.
   - Realizar una investigación sobre los trabajos relacionados en el estado del arte que sirva como ayuda para seleccionar los algoritmos o técnicas a utilizar para la resolución del problema planteado.
4. **Selección de características**
   - Investigar y analizar los métodos de selección de características y elegir uno para dar solución al problema planteado, así mismo servirá para encontrar un subconjunto óptimo que ayude a obtener mejores resultados en cuanto a la predicción de la enfermedad.
5. **Preprocesamiento de datos**
   - Realizar el preprocesamiento necesario al conjunto de características que se han encontrado y consiste en seleccionar, limpiar y transformar los datos que se van analizar, puesto que ayuda a maximizar el rendimiento del modelo. El preprocesamiento se realiza en este punto del desarrollo del modelo dado que solo se aplicará al subconjunto de características que han sido seleccionadas para dar solución al problema y no a todo el conjunto de datos, disminuyendo así el tiempo y costo computacional.
6. **Clasificación**
   - Investigar y analizar los algoritmos de clasificación y seleccionar uno para utilizar como clasificador en el modelo.
   - Utilizar el algoritmo clasificador seleccionado en el paso anterior y alimentar el algoritmo con el conjunto de datos de entrenamiento preprosesados para encontrar patrones.
   - Poner a prueba el conocimiento adquirido en el paso anterior mediante la utilización de los datos de prueba, es decir, se realiza una comparación con los patrones encontrados y se clasifica para determinar a que clase pertenecen.
7. **Desempeño del clasificador**
   - Por último, el paso que se realiza es para medir el desempeño del clasificador.

3. Resultados y Discusión

En esta sección, se presenta una discusión de los resultados obtenidos mediante las pruebas realizadas con el conjunto de datos adquirido y la plataforma WEKA acerca del comportamiento de estos datos al utilizar algunos de los algoritmos de selección de atributos y clasificadores como parte de la metodología del modelo propuesto. En la tabla 2 se muestra el conjunto de factores de riesgo que se han definido para determinar el nivel de daño renal y que se utilizan para realizar el modelo predictivo de nefropatía diabética.

Nombre	Descripción	Tipo	Valor
no_Exp	No. de expediente	numerico	
edad	Edad paciente	numérico	
sexo	Sexo paciente	numérico	f/m
talla	Estatura paciente	numérico	
peso	Peso paciente	numérico	
IMC	Índice masa corporal	numérico	

Nombre	Descripción	Tipo	Valor
CM	Otras enfermedades	texto	si/no
terapeutica	Tiene tratamiento	texto	si/no
HbA1c	Hemoglobina glicosilada	numérico	
glucosa	Glucosa paciente	numérico	
urea	Urea paciente	numérico	
Cr	Creatinina paciente	numérico	
BUN	Nitrógeno ureico	numérico	
URIC	Ácido úrico paciente	numérico	
CHOL	Colesterol paciente	numérico	
TRIG	Trigliceridos paciente	numérico	
prot_tot	Proteínas totales	numérico	
albumina	Albumina paciente	numérico	
proteinuria	Proteinuria paciente	numérico	
dep_cr	Depuración creatinina	numérico	
eritrocitos	Eritrocitos paciente	numérico	
HGB	Hemoglobina paciente	numérico	
HCT	Hematocrito paciente	numérico	
tipo	Tipo de diabetes	numérico	
evolucion	Años de diabético	numérico	
FG	Filtrado glomerular	numérico	
estadio	Nivel de daño renal	texto	G1-G5

Tabla 2: Lista de atributos clínicos utilizados en este estudio

Subconjunto	Individual
no_exp, edad, talla, albumina, eritrocitos, BUN, evolucion	6.-prot_tot 7.-sexo 8.-peso 9.-CM 10.-IMC 11.-HGB 12.-glucosa 13.-talla 14.-edad 15.-Hb1ac 16.-tipo 17.-eritrocitos 18.-proteinuria 19.-dep_cr 20.-HTC 21.-albumina 22.-TRIG 23.-CHOL 24.-URIC 25.-evolucion 26.-no_Exp

Tabla 3: Subconjunto óptimo y lista ordenada de atributos.

Para la selección de atributos se utilizaron como algoritmos evaluadores CfsSubsetEval, ConsistencySubsetEval, FilteredSubsetEval, ClassifierSubsetEval, WrapperSubsetEval en combinación con un algoritmo de búsqueda de subconjunto. Por otra parte se tienen los evaluadores que funcionan con métodos de búsqueda de tipo individual ChiSquaredAttributeEval, FilteredAttributeEval, GainRatioAttributeEval, InfoGainRatioAttributeEval, SymmetricalUncerAttributeEval, OneRAttributeEval ReliefFAttributeEval, SVMAttributeEval que devuelven una lista ordena de los atributos según su relevancia. A partir del conjunto de datos recopilado se realizaron pruebas en la plataforma WEKA implementando técnicas de selección de atributos que consisten en obtener un subconjunto o un orden de relevancia de los atributos como se muestra en la tabla 3.

Subconjunto	Individual
urea, Cr, prot_tot, sexo, terapeutica,	1.-FG 2.-Cr 3.-urea 4.-BUN 5.-terapeutica

Los algoritmos clasificadores de la plataforma Weka han demostrado un buen desempeño con el conjunto de datos adquirido en este estudio. Después de haber evaluado todos los algoritmos se ha realizado un resumen de los mismos tomando en cuenta solo los que mejor desempeño han demostrado y lo representan los algoritmos FilteredClassifier, DecisionTable y DTNB ya que superan el 95 % como se muestra en la tabla 4.

Categoria	Clasificador	Instancias correctas	% correcto	% incorrecto
bayes	BayesNet	47	85.45	14.55
function	SimpleLogistic	39	70.91	29.09
lazy	LWL	32	58.18	41.82
meta	FilteredClassifier	53	96.36	3.64
misc	VFI	49	89.09	10.91
rules	DecisionTable DTNB	53 53	96.36 96.36	3.64 3.64
trees	LADTree	52	94.55	5.45

Tabla 4: Comparación de clasificadores en WEKA

4. Conclusión

Este artículo presenta el diseño de un modelo predictivo mediante aprendizaje automático basado en la clasificación KDIGO. El modelo propuesto tiene como objetivo predecir la nefropatía diabética en etapas tempranas como son los estadios I, II y III, además de servir como herramienta de apoyo a los médicos. Actualmente los resultados que se han obtenido consisten en comparaciones de algoritmos clasificadores y técnicas de selección de atributos para observar el comportamiento del conjunto de datos adquirido al realizar estas pruebas se encontró que el desempeño del clasificador esta rebasando el 90 %.

Referencias

[1] *Técnicas de análisis de datos. Aplicaciones prácticas utiliando Microsoft y Weka*, chapter •, pages 96–158. Universidad Carlos III de Madrid, 2012.

[2] Guillermo Villatoro Godoy Carlos Hernandez Flores. Nefropatía diabética en pacientes que asisten a la consulta externa de endocrinología pediátrica del hospital escuela. *Revista Médica de los Post Grados de Medicina UNAH*, 2007.

[3] Alicia Elvert. Nefropatía Diabética. Technical report, Universidad de Buenos Aires, Buenos Aires.

[4] Kidney International. Kidney disease: Improving global outcomes (kdigo) ckd-mbd work group. kdigo clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease–mineral and bone disorder (ckd–mbd), 2009.

[5] OMENT. Cifras de sobrepeso y obesidad en méxico-ensanut mc 2016. http://oment.uanl.mx/cifrasde-sobrepesoyobesidadenmexicoensanutmc2016, 2016.

Secuenciación de cadenas ADN, basado en Secuencias Frecuentes Maximales

L.I. Ma del Refugio Velazco Hermosillo [*]
Dr. Luis Alberto Morales Rosales [**]
ISC. Antonio Alejo Aquino [***]
MSC. Mariana Lobato Báez [****]

[*] *Instituto Tecnológico Superior de Misantla, Km. 1.8 Carretera a Loma del Cojolite, C.P. 93821, México (e-mail: 152t0743@itsm.edu.mx)*
[**] *Conacyt-Universidad Michoacana de San Nicolás de Hidalgo, Av. Francisco J. Múgica S/N Ciudad Universitaria, C.P. 58030, México (e-mail: lamorales@conacyt.mx)*
[***] *Instituto Tecnológico Superior de Misantla, Km. 1.8 Carretera a Loma del Cojolite, C.P. 93821, México (e-mail: 152t0728@itsm.edu.mx)*
[****] *Instituto Tecnológico Superior de Libres, Camino Real, Barrio de Tetela, 73780 Cd de Libres Pue. (elegancia_14@hotmail.com)*

Resumen: La extracción de Secuencias Frecuentes Maximales, son una representación compacta del conjunto de las secuencias frecuentes, esto es importante porque reducen el espacio de almacenamiento y el espacio de búsqueda de información. Se sabe por la literatura que el problema de descubrir patrones secuenciales de una base de datos de documentos es un problema intratable y que la técnica de patrón de crecimiento podría acelerar la extracción de patrones secuenciales. En este artículo nos enfocamos a la extracción de Secuencias Frecuentes Maximales (SFM)de las cadenas de ADN y lo que proponemos es desarrollar un mecanismo computacional para extraer las SFM, basándonos en el algoritmo DIMASP, que tiene la característica de utilizar la técnica de crecimiento de patrones y es independiente del umbral de soporte.

Keywords: Secuencias Frecuentes Maximales, Algoritmo DIMASP, Bioinformática, Nucleótidos.

1. INTRODUCCIÓN

Dentro de las técnicas de inteligencia artificial (I A) se encuentra la minería de datos (M D), que se define como: " el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos "[1]. De acuerdo con esto, la minería de secuencias es un caso particular de la minería de datos estructurados, que enfocada a este proyecto consiste en encontrar patrones relevantes en colecciones de datos que están representados de forma secuencial[2]. Siendo las cadenas de Ácido Desoxirribonucleico (ADN) nuestro objeto de estudio, el trabajo se desarrolla en el área de Bioinformática.

La Bioinformática es una disciplina emergente que utiliza las tecnologías de la información para captar, organizar, analizar y distribuir información biológica con el propósito de responder preguntas complejas en biología. Es un área de investigación multidisciplinaria, que puede definirse como la interfaz entre dos ciencias: la biología y la computación, impulsada por la incógnita del genoma humano y la promesa de una nueva era en la que la investigación genómica puede ayudar a mejorar la condición y la calidad de la vida humana [3]. Uno de los retos de la bioinformática consistió en la secuenciación del genoma humano, lo cual originó cantidades enormes de información que debe analizarse para poder ser utilizada de manera conveniente.

La secuenciación del Ácido Desoxirribonucleico (ADN) se refiere a los métodos para determinar el orden de las bases de nucleótidos, denominados estos como Adenina (A), Guanina (G), Citosina (C) y Timina (T)[4]. El problema de trabajar con genomas es particularmente complejo por el gran tamaño de las cadenas de ADN, como es el caso de la mayoría de las eucariotas (organismos que tienen células con núcleo definido por una membrana, en el cual se almacena toda la información genética), por lo tanto se requiere investigar maneras adecuadas para la adaptación de algoritmos que mejoren la estructura, identificación, velocidad, precisión y almacenamiento de los datos. Esto representa para el área de computación un problema de tipo NP-Completo.

En particular en esta tesis, se aborda el problema de la extracción de Secuencias Frecuentes Maximales (SFM) de las cadenas de ADN. Es imprescindible encontrar la manera de comprimir los datos para analizar, almacenar y recuperar las bases cuando se requiera. El apoyo de la Inteligencia Artificial ocupa un lugar sobresaliente al

transformar los datos en información y ésta a su vez en conocimiento. Esta transformación implica el desarrollo de heurísticas para obtener, seleccionar y estructurar los datos, ya que la cantidad de información que comprende un genoma humano suscita un problema por la cuantía de combinaciones al extraer las SFM. En este trabajo se propone desarrollar un mecanismo computacional, para la extracción de las SFM, de las cadenas de ADN, basándose en un algoritmo para Descubrir Patrones de Secuencias Maximales (DIMASP del inglés Discover all the Maximal Sequential Patterns), que se basa en la técnica crecimiento de patrones, además es independiente del umbral de soporte.

2. ANTECEDENTES

El avance en la secuenciación de los ácidos nucleicos ha generado un amplio conocimiento en el campo de la genómica, estos beneficios se extienden a la fármacogenética aplicándose a la medicina personalizada. Actualmente se tiene una gran cantidad de información con aplicaciones innumerables. Entre otras cosas, la secuenciación ha permitido entender la asociación de enfermedades con la variabilidad genética, la función de genes, el patrón de expresión de genes nuevos, la similitud o variación genética entre especies diferentes, la organización de la información genética, el origen de algunos genes, etc. [5]
Las herramientas de software que facilitan la investigación en bioinformática pueden clasificarse en cuatro clases [6]. Sin embargo las que interesan para nuestro objetivo son la comparación de secuencia y las herramientas de alineación. Algunas de ellas son:

• BLAST su principal característica es la velocidad, realiza búsquedas en la totalidad de una base de datos no redundante en poco tiempo.

• GenBank y EMBL, son dos de las herramientas principales de gestión de bases de datos biológicas para alineamiento local por pares de secuencias.

• FASTA se puede utilizar para hacer una comparación rápida de proteínas o de nucleótidos. Alcanza un alto nivel de sensibilidad para la búsqueda de similitud mediante la realización de búsquedas optimizadas para alineamientos locales utilizando una matriz de sustitución.

• ClustalW para alineación de secuencias múltiples, la cual se puede utilizar para alinear las secuencias de ADN o de proteínas con el fin de dilucidar sus relaciones, así como su origen evolutivo.[6] Minería Secuencial de Patrones, tiene como objetivo encontrar todas las subsecuencias que están conteneidas al menos beta veces en una colección de secuencias, donde beta es el umbral de soporte especificado por el usuario.

Secuencias Frecuentes Maximales, es una secuencia que no es subsecuencia de ninguna otra secuencia frecuente. Estas son representaciones compactas de todo el conjunto de secuencias frecuentes.

3. TRABAJOS PREVIOS

En este apartado se mencionan trabajos de investigación referentes a las técnicas utilizadas para resolver problemáticas similares o que tienen relación con este tema, a saber: la extracción e indexación de Secuencias Frecuentes Maximales de una colección de secuencias ADN, para hacer búsquedas de cadenas de ADN solicitadas. En general, las técnicas o algoritmos que se utilizan para hacer comparaciones, alineamientos y búsquedas en bases de datos de las cadenas ADN. En lo particular, se hace referencia a trabajos relacionados con la Minería de Secuencias.

Descubrimiento de patrones secuenciales maximales de una colección de texto. Los autores proponen el algoritmo DIMASP,soporta el crecimiento de patrones con independencia del umbral que se requiera. Como primer paso, el algoritmo, asigna un número entero como identificador, para cada elemento diferente de la colección de documentos. Luego, construye una estructura de datos con cada par de elementos contiguos de los documentos de la base de datos, para acelerar la extracción de los patrones de secuencias maximales (PSM), sin perder el orden de la secuencia. Después, encuentra posibles patrones de secuencias maximales (PPSM) y verifica que no estén en el conjunto de PSM, si no están, los almacena, si un PSM es subsecuente del nuevo PSM, lo poda. Por último, se genera una estructura con todos los PSM. Los resultados muestran que DIMASP supera a los algoritmos GSP, DELISP GenPrefixSpan and SPADE, por su buena escalabilidad en cuanto al umbral, descubre secuencias más largas con un umbral mayor a dos.[7]

FINDPAT es un algoritmo que encuentra repeticiones (patrones) maximales con matching exacto dentro de secuencias de ADN (alfabeto fA, C, G, Tg), y sin límite en la longitud de las repeticiones que este arroja. Esto lo hace de una manera conocida como ab initio, ya que necesita de alguna semilla (secuencia de referencia) previa para su funcionamiento. Solamente necesita como entrada una secuencia o dos (en el caso de buscar patrones entre dos secuencias), y la longitud mínima deseada para las repeticiones resultantes. Este algoritmo puede recibir secuencias genéticas de hasta 500 Megabytes, es así que brinda la posibilidad de analizar cromosomas enteros (un cromosoma, o parejas de cromosomas). Por una cuestión de simplicidad y más que nada por cuestiones de implementación del algoritmo, se buscan repeticiones con correspondencia exacta, es decir que no se permite modificación, eliminación de caracteres para la búsqueda de los patrones. [8]

BLAST, este algoritmo permite comparar una secuencia de ADN con una, o un conjunto de bases de datos, e identificar secuencias dentro de estas con las cuales se asemeja. Introduce varios refinamientos a la búsqueda en bases de datos, que mejoran el tiempo de búsqueda. Hace énfasis en la velocidad por sobre la sensibilidad. Esto es fundamental para que el algoritmo sea práctico al buscar en las bases de datos gigantes que están disponibles hoy día. No está basado en un algoritmo que garantiza el alineamiento óptimo, sino que usa una heurística que funciona la mayoría de las veces en la práctica, así que, podría fallar con algunas secuencias poco relacionadas entre sí. Es alrededor de 50 veces más rápido que otros algoritmos que garantizan el alineamiento local de secuencias óptimo y usan programación dinámica. [9]

PrefixSpan. Es un reto ya que se tiene que examinar un número explosivo de combinatoria de patrones de las subsecuencias. Es parte del desarrollo de métodos que reducen sustancialmente el número de combinaciones que debe examinarse. Sin embargo se sigue teniendo problemas en cuanto a bases de datos muy grandes o cuando los patrones secuenciales extraídos son numerosos. Este algoritmo, propone la proyección de los prefijos en la extracción completa de las secuencias de patrones, reduce la generación de subsecuencias candidatas. La proyección de prefijos reduce el tamaño de proyección de bases de datos y conduce a un procesamiento eficiente. Supera a GSP y FreeSpan. [10]

Extracción de secuencias frecuentes con algoritmo SPADE. Utilizan propiedades combinatorias para descomponer el problema principal en pequeños subproblemas, que pueden ser resueltos independientemente en memoria principal, usando técnicas eficientes de búsqueda de enrejado, las secuencias se obtienen a través de 3 escaneos a las bases de datos. Como resultado muestran que SPADE supera a GSP con un factor de 2. [11]

4. MATERIALES Y MÉTODOS

4.1 Materiales

Para la tarea de extraer las Secuencias Frecuentes Maximales de las cadenas de ADN, se hizo uso de los siguientes elementos:
Algoritmo DIMASP, es un algoritmo de 3 pasos, basado en el crecimiento de patrones, con independencia del umbral de soporte.
Para las entradas se recopilaron los archivos en formato FASTA, que contienen las secuencias de ADN, descargados del repositorio del Centro Nacional para la Información de Biotecnología (NCBI, por sus siglas en inglés).
En la parte de la reestructuración se implementó un programa de software, para modificar la organización del contenido de los archivos .fasta, dándoles la misma estructura que los archivos de entrada del algoritmo DIMASP, para luego hacer un solo archivo con la colección de secuencias. Los archivos de entrada tienen una sola línea por cada secuencia de ADN, cada una inicia con un número consecutivo, seguido por el signo igual, antecediendo a la secuencia de nucleótidos, cada caracter después del signo igual, es separado por una coma sin dejar espacio, solo el último caracter de cada secuencia no lleva coma.

4.2 Métodos

Para resolver el problema de la extracción de Secuencias Frecuentes Maximales, se desarrolla una propuesta de un mecanismo computacional basado en el algoritmo DIMASP. Se muestra en la Fig.1.

5. EXPERIMENTOS

Uno de los pasos para resolver la problemática de extraer las Secuencias Frecuentes Maximales, es acceder a la plataforma web del NCBI para recopilar la información de las cadenas de ADN las cuales son nuestro objeto de estudio. En la Fig. 2, se muestra abajo la interfaz del sitio web NCBI.

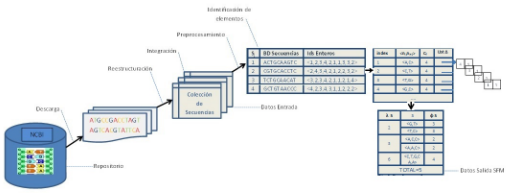


Fig. 1. Propuesta del mecanismo computacional

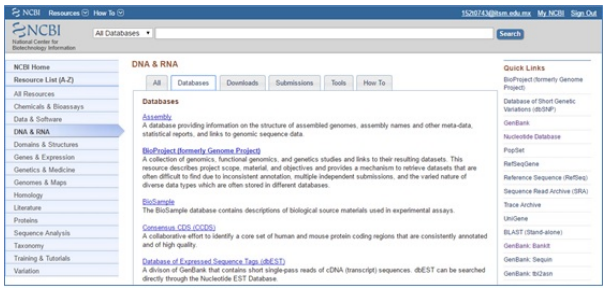


Fig. 2. Repositorio NCBI

De esta plataforma se descargaron archivos en formato .fasta, por tener una estructura en su contenido más simple y parecida a las entradas requeridas por el algoritmo DIMASP. En la Fig.3, se muestra el contenido de un archivo como ejemplo.

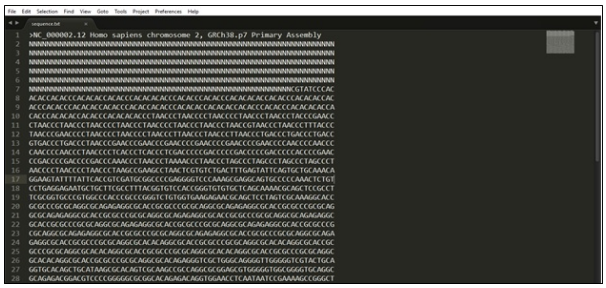


Fig. 3. Ejemplo archivo .fasta

Todas las secuencias de nucleótidos se almacenan en un solo archivo, reestructurando su contenido para que funcione como entrada del algoritmo DIMASP. Cada secuencia se identifica con un número entero seguido del signo igual y cada letra seguida por una coma a excepción de la última, todo sin espacios. Ver Fig. 4

Se extraen las Secuencias Frecuentes Maximales de la colección de secuencias de ADN y se muestra el resultado en la Fig. 5 SFM.

Fig. 4. Documento con la Colección de Secuencias

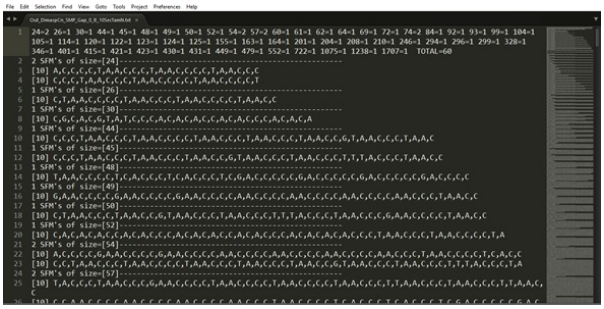


Fig. 5. Secuencias Frecuentes Maximales

6. CONCLUSION

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

[1] Witten, I. H. y Frank, E, *Data mining: Practical machine learning tools and techniques with java implementations*, segunda 2nd edition, Morgan Kaufmann Publisher. 2005.

[2] Mabroukeh, N. R. y Ezeife, C. I. *A taxonomy of sequential pattern mining algorithms*, ACM Computing Surveys 43, 1, Article 3 (November 2010), 41 pages. DOI=10.1145/1824795.1824798.

[3] Cañedo, A. R. y Arencibia, J. R., *Bioinformática: en busca de los secretos moleculares de la vida*, Acimed 2004;12(6).

[4] National Human Genome Research Institute, *¿De qué está compuesto el ADN?*,https://www.genome.gov/27562614/cido-desoxirribonucleico-adn/ NIH. Publicado el 21 octubre 2015.

[5] Hutchison, C. A., *DNA sequencing: bench to bedside and beyond*, Nucleic Acids Research, Vol. 35, No. 18 6227-6237 doi:10.1093|nar|gkm688. 2007.

[6] Meneses, E. C. A.; Rozo, M. L. V.y Franco, S. J., *Tecnologías bioinformáticas para el análisis de secuencias de ADN*, Scientia Et Technica, XVI(49) 116-121. Recuperado de http://www.redalyc.org/articulo.oa?id=84922625020. 2011

[7] Garcia, H. R. A.; Martínez, T. J. F. y Carrasco, O. J. A., *A New Algorithm for Fast Discovery of Maximal Sequential Patterns in a Document Collection*, A. Gelbukh (Ed.): CICLing, LNCS 3878, pp. 514-523, 2006.

[8] Foglino, M. A., *Identificación biológica de repeticiones en secuencias de ADN del genoma humano*, Tesis de licenciatura, Departamento de Computación Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires Argentina. 2009

[9] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W. y Lipman, D. J., *Basic Local Alignment Search Tool*, Academic Press Limited. J. Mol. Bio. 215, 403-410. 1990.

[10] Pei, J.; Han, J.; Mortazavi-Asl, B. y Pinto, H., *PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth*, Intelligent Database Systems Research Lab. Canada V5A 1S6. 1063-6382/01. IEEE. 2001

[11] Zaki, I. M., *An Efficient Algorithm for Mining Frequent Sequences*, Douglas Fisher. Kluwer Academic Publishers. Manufactured in The Netherlands. Machine Learning, 42, 31-60, 2001

Appendix A. A SUMMARY OF LATIN GRAMMAR

Appendix B. SOME LATIN VOCABULARY

Siscom ETL, an alternative when versalitility in the extraction, transformation and load are factor in the quality of the data analysis

S. Juarez, E. Trujillo, H. Andrade

***Abstract*—When the requirement and diversity of formats are a prerequisite for obtaining the best results in the Data Analysis, when the initial stage in the knowledge acquisition process in databases (KDD) and the rigidity of the existing tools do not allow the extraction, transformation and loading process (ETL). This activity requires significant amounts of time, money and effort by the professionals in their preparation before applying the techniques and algorithms of analysis in the part of the process called Data Mining, it is here when it is justified to opt for an alternative development of a tool that can be molded to the needs of each application of the KDD process. The article shows a case study in wich it is necessary the development of a tool that performs the ETL in a way that allows the diversity of output formats in the treatment and modeling of output views of DataWare Housing (DWH) itself and with the prospect of becoming a useful tool in the portfolio of the professional dedicated to Data Analysis.The work of this paper is not done in order to compare or disqualify other tools to do ETL work, it has the perspective of being an initial part of what in the future will also become a process of acquisition of knowledge KDD.**

***Index Terms*—Customers, KDD, ETL, data mining, artificial intelligence, decision making.**

I. Introduction

Data analysis without being a new topic, it is known that it is taking a boom and importance very remarkable for the companies of the world today. In each passing year, existing companies realize that they need to incorporate the competitive advantages of large companies and that the new ones, although they are already implemented, one of these advantages is the acquisition of knowledge as a basis for making their decisions, this is to analyze historical data of either accumulated years or those that are part of their daily operation and that lead to knowing the behavior patterns of their customers and other entities, ie the information generated by traditional Enterprise Resource Planning (ERP) is not longer sufficient, today it is necessary to translate information records into knowledge that allows decisions to be made so that these are the ones that produce the best dividends in the company.

It starts from a large data warehouse that may well be from different sources and formats inclusive, that is, they can be different databases, electronic sheets, etc. Thus forming a Data Warehouse (DWH), and from there start what is known as Knowledge Discovery in Databases (KDD) will be explained later.

For the purpose of this material, the which corresponding to a commercial company with subscribing clients in the alarm monitoring service in the security sector is taken as the object of application and whose interest is to know:

- On what day of each month do your subscribers customers pay as a monthly revenue budget base
- If your customers pay: early, punctual, expired or moratorium. This in relation to the comparison of the due date of each commitment with the date that the clients pay each month or in each commitment that is awarded to them for the payment in the provision of a service. This point requires for the data analysis the Discretization of values since what is counted is with date data in the DWH.
- The classification of their clients by reason of whether they are: Retirees, private sector workers, government sector workers, etc. And how good clients are in each ranking or grouping.
- Classification regarding the form of payment: cash at home, payment at convenience store, payment at a bank window or payment in bank with electronic transfer.
- Finally, the classification by location in the city: North, South, West, East, Foreign, etc.

And although the initial DWH is a database of about 70 tables, the following scheme is simplified and in line with the requirements and the DWH is as follows:

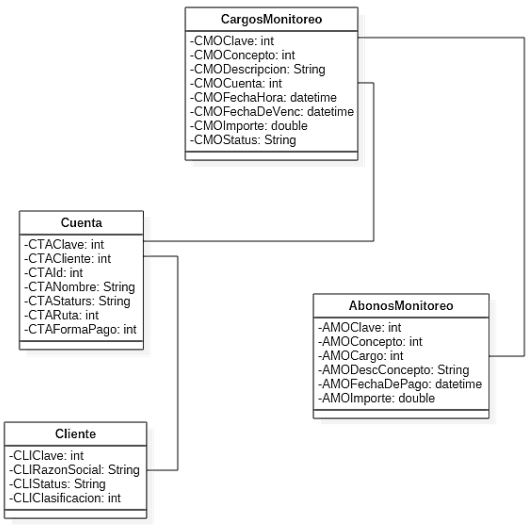


Figure 1. Initial Conceptual Company Model

KDD, The process of the 21st century, (email: senen.juarez@gmail.com)

II. The Process of Knowledge Acquisition (KDD)

Because of their relationship with other terms and probably by its orientation with other approaches, we have heard perhaps terms like BigData, Data Mining, etc. The reality is that everything converges in the idea of treatment with large volumes of data for its analysis and extraction of more than just information, knowledge. Everything in a process that involves a series of steps and where effectively one of them is the Data Mining that consists of the application of various algorithms that are selected for their application based on the objectives set for the process. This whole set of steps is known as Knowledge Discovery in Databases and abbreviated literally as KDD.

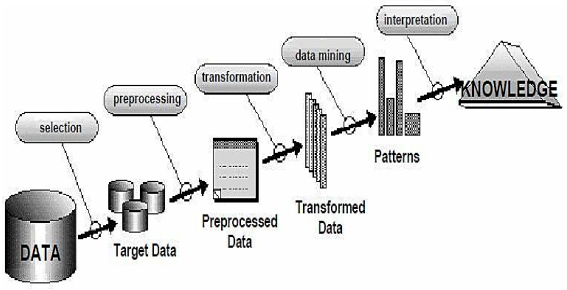


Figure 2. The KDD Process

III. ETL Process

Extraction, Transformation and Load (ETL) is the phase prior to the application of Data Mining in the KDD Process and it is that set of steps in which the Dataset is prepared in a convenient way so that the experts in the Data Mining experiment in the application of the various algorithms with their respective purposes, these steps are described as follows:

- Extraction, extract data from your source systems, different formats, fonts, etc.
- Transformation, select only certain columns or attributes, translate values for their best treatment or, if applicable, carry out what is known as data discretization, which is to transform a continuous and widely dispersed value into discrete values in a degree that is more appropriate For analysis and interpretation.
- Load, provision of the transformed values for its generation in formats susceptible to handle for the tools to carry out the Data Mining.

There are specialized ETL Process tools, however, when the objectives of the Data Analysis raised exceed the standard way of generating the formats for their entry to the tools with support of Data Mining and to carry out this preparation takes excessive treatment of the data increasing Time, money and effort in its implementation in a process also repetitive at the time, a suitable decision is to develop software that quickly and efficiently perform this work as is the case of this work. Sometimes ETL process can be carried out by lower-cost work teams and their economic treatment leads to long hours of work based on initial formats such as electronic sheet or plain text or files to produce general CSV formats Or in particular cases other formats such as the .ARFF for the well-known New Zealand Weka software, in the case of this paper the decision is made to develop our own ETL tool by investing resources in this but considering that beyond the use for these analysis objectives, can be adapted and improved for future professional work and be applied as one of the phases of the service process to offer in a market that as already mentioned there are of this type of solutions.

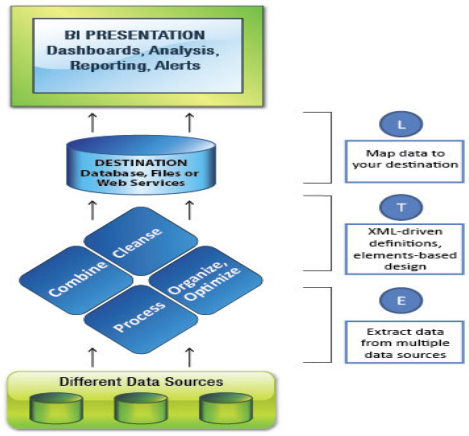


Figure 3. The ETL Process

IV. Siscom ETL Structure

Siscom ETL in its initial version is designed on the basis of a necessary structure and to the measure of the requirements established in a particular way, its structure is shown below.

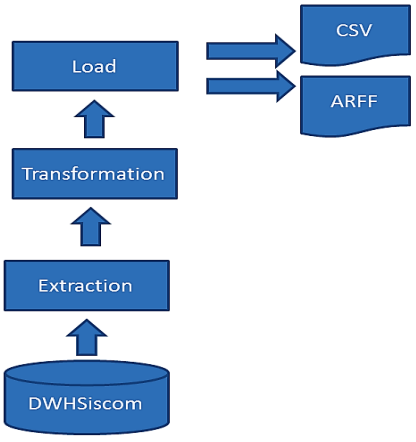


Figure 4. Siscom ETL Structure

V. Extraction

A. Selection

Starting from the initial work that was based on the objectives shown before, a selection of tables was carried out from where to extract the necessary data as previously shown as initial conceptual company model, we also want to have in real time updated the initial part of the generation chain of the Dataware House, a Materialized View was defined within the database with a list of 11 attributes as shown below:

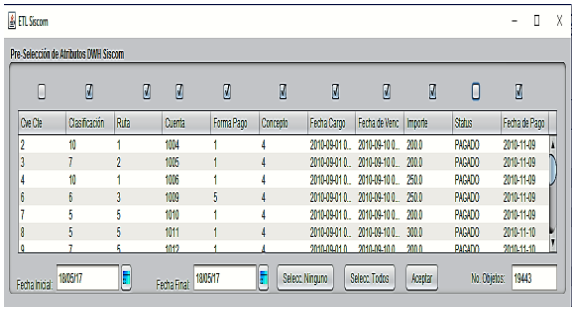


Figure 5. Selection

B. Second level of selection

Likewise, having the option to generate our dataset in the period of time that is desired to a certain moment, being able a month, a year or any other range of dates and generating a second level of selection adding versatility to our tool:

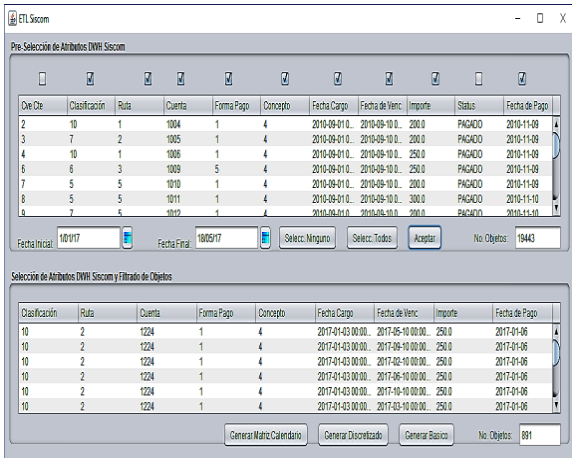


Figure 6. Second level of selection

VI. Transformation

A. Data Cleaning

The cleaning of the data is fundamental, in our case we opted for the method of elimination of instances with incomplete values, this because a previous process of revision of the origin database was made and care was taken to have updated data of The most recently active clients, because these periods of time will emphasize the analysis, although the database contains instances of a period of about 7 years, the analysis will be applied in the last 2 years, because they are the years with a more disciplined record of the ERP that has produced these records . On the other hand, contained in a field of type date for example, which includes data of hour, minute, seconds are not relevant, so they were also cleaned by the process, you can also see the view with the formats to be generated as are files with extensions .CSV which is a format considered standard in the industry for input data of various tools and .ARFF for Weka, as shown below:

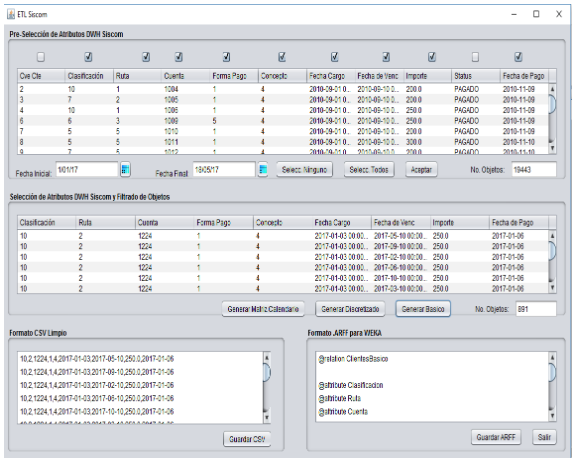


Figure 7. Data Cleaning

B. Discretized Data

There is interest on the part of the company as manifested in the objectives at the beginning of this article, to know in which day of each month do their customers pay to obtain patterns of behavior and beyond this, the principle is to know simply if The existing customers are punctual in their payments or not, in this case the data as such does not exist, then a pre-process is made in which the payment dates are compared against the due dates of the payment commitments and this way It is determined whether the customer pays: advance, punctual or expired, this transformation is shown below:

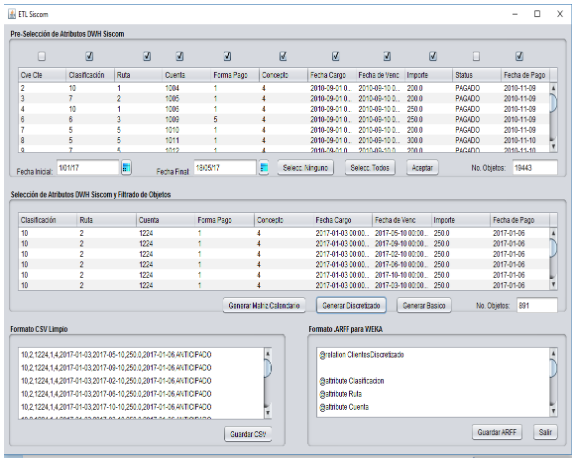


Figure 8. Discretized Data

VII. Load

From the data already treated and shown in the interface, the respective formats were generated in .CSV and .ARFF as the elements to be taken as a load in the tools through which the analysis of The datasets generated differently in the quantity and diversity required in a simple, fast and precise way, storing them in a directory for the case created or designated by the user from our tool Siscom ETL, putting the name that the user considers it says best Form the content of the dataset in a well-conducted process, the interface shows the respective buttons with the drive in the generation of each of the formats

and their storage in a precise place, as shown in the following figure:

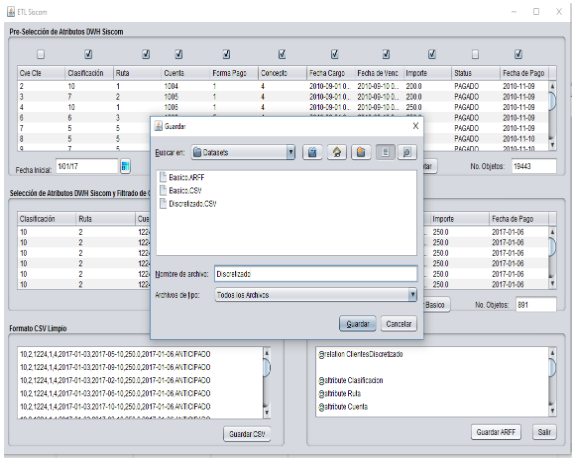


Figure 9. Saving formats

After this, the generation is completed. The process can be repeated by generating the number of formats to have a broad spectrum of analysis with the simple work of establishing criteria in the selection of attributes, date ranges, discretization and generation of formats to complete.

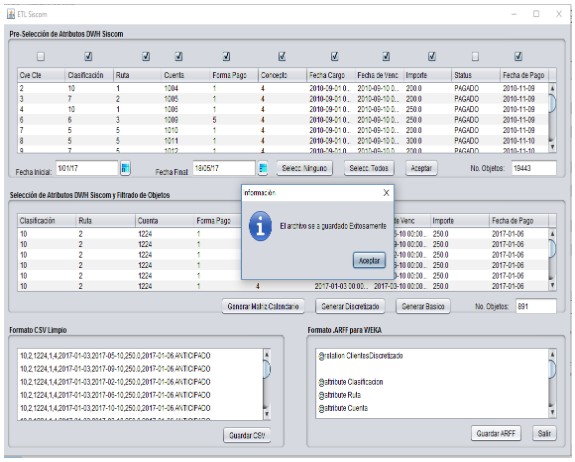


Figure 10. Completion of the process

VIII. Results

A. Format .CSV

Next, the .CSV format is shown from a tool for creating and viewing electronic sheets, it is shown how each attribute occupies a column of the sheet as illustrated in the following figure:

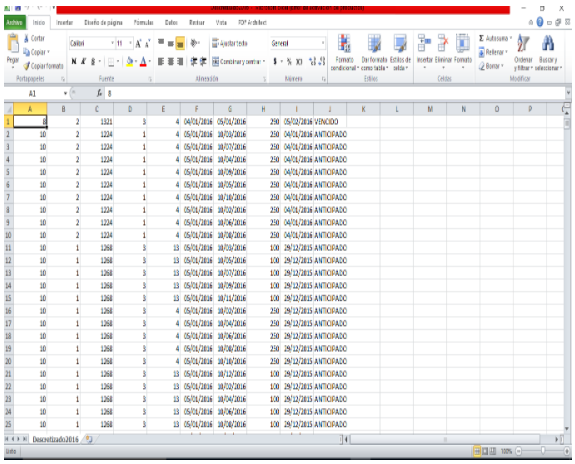


Figure 11. Format .CSV

B. Format .ARFF

On the other hand, the format for the Weka Software, shown in a tool to create and visualize text files, shows that it complies with the format requirements of the analysis tool referred to as shown in the following figure:

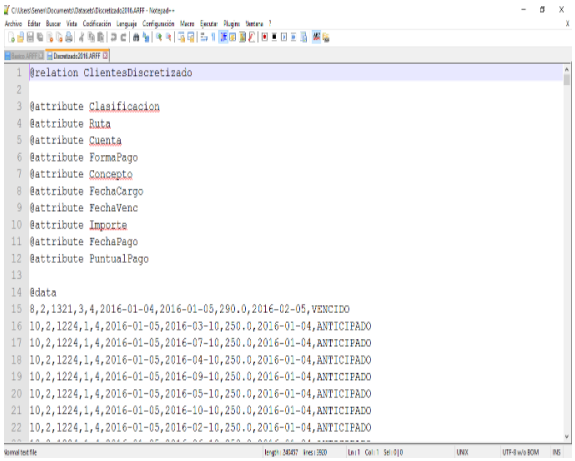


Figure 12. Format .ARFF

IX. Conclusions

There is no doubt that the time invested in developing Siscom ETL has been worth it, and the help it has provided to the process of analysis in the generation of formats in savings of time, money and effort justify it widely, in addition to that it has led to open an interesting business perspective. We think that it can be the beginning of a solid work structure in the market of data analysis, independently that will use diverse tools of Data Mining has Weka to be a free and open tool, as the option to take this product to a next step of making it a tool that implements the complete KDD process, initially manually, step by step and in the future the perspective is the complete development to operate in real time, under an architecture with the use of multiagents or simply oriented to services that allows real time can be viewed the results on a website or mobile application for the end user, becoming a professional product with a value and high chances of success in the domestic and international market inclusive.

Classification techniques used to resolve the email overloading problem

José Arcángel Salazar Delgado *
Alberto Mendez Torreblanca, Jorge Estudillo Ramírez **

* *Instituto Tecnológico Superior de Misantla, (e-mail: arcangel.salazar@gmail.com).*
** *Instituto Tecnológico de Veracruz, (e-mail: atorreblanca@gmail.com, jorgeestudillo@gmail.com).*

Abstract: In this paper we will review the techniques used to classify email in order to resolve the email overloading problem. We explain the email overloading problem a his consequences, also we give a introduction of supervised and unsupervised techniques.

Keywords: Classification, Algorithms, Natural Language Processing, Email.

1. INTRODUCTION

The email is one of the most used tools in the world and is the standard method for formal communication in most of the companies. Many companies got a lot of emails from their clients or providers and the emails needs to be processed as quickly as possible. When the quantity of emails is too big, the email management becomes painful, in other words the companies got more email that their employees can handle efficiently. This phenomenon is called "email overload". One approach widely use to resolve the email overloading problem is the classification method. This method is particularly good and consist on grouping emails in some folder to diminish the complexity of the tray. There are many techniques that implements email classification. In one technique the employees make folders to group similar emails with regard the same subject or topic and they classify the email using his personal criterion. Another technique is the use of message filters witch are part of email software clients. The email filters are configurable conditions used to identify emails, for example, emails that include some text in the body or subject of the email. The filters includes configurable actions like automatic grouping of emails in predefined folders that accomplish predetermined conditions.

A field that has interest in the solution of the email overloading problem and proposed different techniques for automatic email classification is the artificial intelligence field. This techniques includes supervised and unsupervised learning algorithms. In the classification techniques with supervised algorithms, the algorithms needs an example set to learn how to classify the new emails. The techniques with unsupervised algorithms learns at the same time they are processing the emails and identifies categories according email similarity.

This paper describe a revision of many representative techniques or contributions to automatic email classify using supervised or unsupervised learning. The contribution will be mentioned in chronological order. The purpose of this paper is give to a general view of the field and give a starting point to motive the generation of new techniques of automatic classification.

The article start explaining the problem of email overloading in detail in section 1. In the section 2, the different techniques that has been used for email classification are described. After that, a very interesting review the email classification techniques using supervised and unsupervised learning are explained in section 3 and finally, the conclusions are discussed giving point of view of the different techniques.

2. THE EMAIL OVERLOADING PROBLEM

The "information overload" is a term popularized by Toffler (1970) and is used to describe a phenomenon that happens when you have so many information about an issue that you can not make a decisions or understand it. Whittaker and Sidner (1996) observe this phenomenon in company employees email users when they got too much emails and call it "email overloading". After that, the email overloading problem was defined by Dabbish and Kraut (2006) as email users perceptions that their own use of email has gotten out of control because they receive and send more email than they can handle, find, or process effectively.

The "email overloading" has serious consequences to the affected users. From stress to anxiety, email overloading can diminish the general productivity of the user. This can represent huge losses to the companies. Every email user can be affected by this problem. In addition to that, the problem can be aggravate by the unwanted email used by persons and companies with a lot of different purpose like publicity, frauds, extortions etc. This type of email is called "spam".

All those reasons motivate the research of different methods to alleviate the "email overloading" problem. One of the most successful method is the email classification. This method consist in grouping similar emails in folders

in order to organize them for later consult. In the next section we will describe in detail many techniques that implement the email categorization methods.

3. EMAIL CLASSIFICATION TECHNIQUES

Since the identification of the email overload problem, many researchers try to implement the classification method using different techniques with distinct degree of success. Each technique has strengths and drawbacks depending on the context of application.

The techniques are divided in two categories:

Non-automatic techniques and automatic techniques.

The non-automatic techniques are all techniques that needs constant user intervention in order maintain classified the email tray. Example of these techniques are:

- Manual folder classification. A human is in charge of putting the emails in the correct folder. Most of the users just give up of this technique because is slow and error-prone.
- Filter based classification. The filters are configurable conditions used to identify emails, for example, emails that include some text in the body or subject of the email. The main drawback with filters is the requirement of time from users in order to update the conditions. This is the most popular technique because is available in many email management tools.

On contrary, automatic techniques are all techniques that doesn't need user intervention. Depending on the way the learn they can be classified in two types

- Supervised.- They need a predefined set of examples to learn how they need to do the classification. All the methods used to classify email as spam are supervised.
- Unsupervised.- They don't need any examples, the algorithm discover how they need to do the classification. The complexity of these methods make them slow and very difficult to use for a normal user. Most of these methods are implemented in email servers with specific purpose in highly specialized environments.

Most of the automatic techniques are slow and resource consuming tasks. Because of that, most of the computers programs used to manage emails doesn't include any feature to automate the email classification. The only option included in these programs to help with the classification task are the filters.

4. AUTOMATIC EMAIL CLASSIFICATION TECHNIQUES

In this section we will review the techniques in chronological order starting from the oldest. The table 1 and table 2 summarize the sequence of the works for the supervised and unsupervised techniques. We will start with the supervised techniques.

Table 1. Supervised techniques time line

2000	TF-IDF classifier .
2000	SVM.
2000	Unigram model.
2003	Ontology.
2004	SpamBayes.
2009	MBPNN.
2010	BPT.
2011	Co-training and SVM.
2012	PCADR.

4.1 Supervised techniques

The supervised techniques use a training set to deduce the rules used to classify the email. In other words, a human create a set of emails and add to some category, then we use this set to train some algorithm.

At first, **Brutlag and Meek (2000)** review three common methods to classify user email stores, support vector machines, TF-IDF classifier and the unigram model. They compare three aspects of the methods:

- Accuracy
- Performance
- Wasted resources

They obtain these conclusions:

- Accuracy varies more between email stores than between classifiers
- No classifier is consistently superior to the others
- Classification accuracy is highest for the two mail stores with the lowest proportion of sparse folders.

Using a very interesting approach, **Taghva et al. (2003)** build an ontology that applies rules for identification of features to be used for email classification with a bayesian classifier.

They conclude that this methodology can be generally applied to other classification challenges.

Trying to improve the efficiency of the spam classfication **Meyer and Whateley (2004)** build the SpamBayes classification engine using a combination of techniques (chi-squared and n-gram tiling). They classify the email with three categories using the chi-squared combined classifier: Ham, Spam and unsure.

Their approach is proven effective and achieved high praise from its users due to a combination of factors.

The use of a message scoring system that provides an unsure classification greatly reduces the users exposure to spam messages by almost completely eliminating false positives.

The tokenizer has been developed to exploit all clues that are found to improve results in a statistically significant way.

The development group continuously evaluates new techniques, such as the n-gram tiling to seek out performance improvements and to keep the product ahead of the ever-changing attacks that are being used by spammers.

Using a very complex conjunction of techniques **Yu and Zhu (2009)** proposes new email classification models using a linear neural network trained by perceptron learning algorithm and a nonlinear neural network trained by back propagation learning algorithm and they propose an advanced email classification systems based on the combination of modified back propagation neural network (MBPNN) and semantic feature space (SFS). MBPNN overcomes the slow learning speed problem in the traditional Back-propagation neural network and can escape from the local minimum.

The SFS method not only reduces the number of dimensions drastically, but also overcomes the problems existing in the vector space model commonly used for text representation.

The experimental results show that MBPNN enhances the performance of email classification and SFS method further improves its accuracy and efficiency

Smiliar to the previous model, **Ayodele et al. (2010)** proposes a new email classification model using a teaching process of multi-layer neural network to implement back propagation technique. They study how to generate accurate email categories. They analyze the characters of emails and study the email conversation structure, which they argue have not been sufficiently investigated in previous research on email classification using back propagation technique.

They classification is based on heuristic technique with the used of Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in a corpus of email messages might be more favorable to use in a query, They also implement a neural network based system for automated email classification into user defined "word classes" and their back propagation techniques implemented was able to learn technique in an associative learning approach, in which the network is trained by providing it with input and matching output patterns.

They concluded that neural networks using back propagation technique can be successfully used for semi-automated email classification into meaningful words. Thy observe that back propagation is based on learning by example and outperforms several other algorithms in terms of classification performance.

Using a simple yet interesting techniques, **Kiritchenko and Matwin (2011)** presented a learning technique introduced in Brutlag and Meek (2000) which greatly decreases the effort needed in applying machine learning on real-life data.

With a interesting combination of algorithms **Gomez and Moens (2012)** presented and evaluated a novel technique based on principal component analysis (PCA) document reconstruction, the Principal Component Analysis Document Reconstruction(PCADR), in the context of email filtering while using only text-content features.

Table 2. Unsupervised techniques time line

2005	Author-Recipient-Topic.
2008	Adaboost.
2010	Graph mining.

Their results show that PCADR performs well when separating spam from ham, and phishing from ham messages. PCADR is able to outperform a Support Vector Machine (SVM) when considering the accuracy of the classification, and in terms of F1 and Receiver Operating Characteristic curve (ROCA) for most of the experiments, with the advantage of PCADR being faster than the SVM when classifying test examples.

When computing the Principal Components (PCs) based on the Power Factorization Method (PFM), PCADR is competitive in training time in comparison to a SVM. PCADR is especially well suited for classification when training with a labeled dataset collected using a given setup and testing with a dataset collected with another setup.

Finally, using a novel approach **Park et al. (2012)** proposed automatic email categorization method, and the architecture of a system to implement it. The proposed method uses inherent feature of emails and fuzzy theory to construct email category and reorganizing email category hierarchy.

The method was tested in experiment which shows a high degree of flexibility, efficiency and effectiveness in the email categorization and category hierarchy reorganization.

4.2 Unsupervised Techniques

Evolving from the LDA(latent dirichlet allocation) algorithm, to the the Author model and the Autor-Topic model **McCallum et al. (2005)** presented the Author-Recipient-Topic model, a Bayesian network for social network analysis that discovers discussion topics conditioned on the sender-recipient relationships in a corpus of messages. This model combines for the first time the directionalized connectivity graph from social network analysis with the clustering of words to form topics from probabilistic language modeling. The model can be applied to discovering topics conditioned on message sending relationships, clustering to find social roles, and summarizing and analyzing large bodies of message data. The model would form a useful component in systems for routing requests, expert-finding, message recommendation and prioritization, and understanding the interactions in an organization in order to make recommendations about improving organizational efficiency.

Another interesting approach is from **Coussement and Van den Poel (2008)**. They build a email binary classifier using adaptative boosting techniques (Adaboost).

This study focuses on how a company can optimize its complaint-handling strategies through an automatic email-classification system and offers an automatic email-classification system that distinguishes complaints from non-complaints.

Using a completely new approach **Chakravarthy et al. (2010)** introduced a framework for effectively classifying emails into using graph mining techniques.

They have proposed a ranking formula for ordering the representative substructures generated from each document class in the training set. The characteristics as well ranking proposed in this paper is adaptive as it adjust to the size and characteristics of a folder. Various parameters that affect the ranking, and therefore classification, have been identified and analyzed in detail.

The concept of feature subset selection enables to classify data even when the amount of data available for training purpose is small. Alternative graph representations (such as tree and star) have been proposed in order to represent the documents and to incorporate useful and relevant domain information.

The experimental results validate their framework by showing significant performance improvement over Naive Bayesian approach for varied email drawn from different domains.

5. CONCLUSION

As we see, most of the methods perform very well for binary classification (spam/ham or phishing/not phishing) and Meyer and Whateley (2004) technique is very successful and widely use. Only Taghva et al. (2003), McCallum et al. (2005) and Park et al. (2012) try to solve the problem of multi classifiers for email.

Taghva et al. (2003) has the problem of the added overhead in the user work in order to give meta-data for the ontology. Also their technique needs a very complicate analysis in order to use efficiently.

Park et al. (2012) doesn't include the author or the subject in their categorization approach. This is an unnecessary information lost.

McCallum et al. (2005) is the most complete method because include the author and the recipient in their model, also their technique is unsupervised.

REFERENCES

Ayodele, T., Zhou, S., and Khusainov, R. (2010). Email classification using back propagation technique. *International Journal of Intelligent Computing Research (IJICR)*, 1(1/2), 3–9.

Brutlag, J.D. and Meek, C. (2000). Challenges of the email domain for text classification. In *ICML*, 103–110.

Chakravarthy, S., Venkatachalam, A., and Telang, A. (2010). A graph-based approach for multi-folder email classification. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 78–87. IEEE.

Coussement, K. and Van den Poel, D. (2008). Improving customer complaint management by automatic email classification using linguistic style features as predictors. *Decision Support Systems*, 44(4), 870–882.

Dabbish, L.A. and Kraut, R.E. (2006). Email overload at work: an analysis of factors associated with email strain. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 431–440. ACM.

Gomez, J.C. and Moens, M.F. (2012). Pca document reconstruction for email classification. *Computational Statistics & Data Analysis*, 56(3), 741–751.

Kiritchenko, S. and Matwin, S. (2011). Email classification with co-training. In *Proceedings of the 2011 Conference of the Center for Advanced Studies on Collaborative Research*, 301–312. IBM Corp.

McCallum, A., Corrada-Emmanuel, A., and Wang, X. (2005). The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email.

Meyer, T.A. and Whateley, B. (2004). Spambayes: Effective open-source, bayesian based, email classification system. In *CEAS*. Citeseer.

Park, S., Shin, J.W., Kwon, J., Jeong, M.A., Lee, Y., and Lee, S.R. (2012). Email categorization using inherent features and fuzzy theory.

Taghva, K., Borsack, J., Coombs, J., Condit, A., Lumos, S., and Nartker, T. (2003). Ontology-based classification of email. In *Information Technology: Coding and Computing [Computers and Communications], 2003. Proceedings. ITCC 2003. International Conference on*, 194–198. IEEE.

Toffler, A. (1970). *Future Shock*. Random House. URL `https://books.google.com.mx/books?id=-BhHAAAAMAAJ`.

Whittaker, S. and Sidner, C. (1996). Email overload: Exploring personal information management of email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '96, 276–283. ACM, New York, NY, USA. doi:10.1145/238386.238530. URL `http://doi.acm.org/10.1145/238386.238530`.

Yu, B. and Zhu, D.h. (2009). Combining neural networks and semantic feature space for email classification. *Knowledge-Based Systems*, 22(5), 376–381.

Electrocardiograms Analysis on the Cloud

Uriel Rubio Escamilla *
Dr. Simon Pedro Arguijo Hernández **

* *Student of Computer Systems Master at Superior Technology Institute of Misantla, Misantla, Veracruz, P.C. 93821 México (e-mail: rubio_esc_uriel@hotmail.com).*
** *Superior Technology Institute of Misantla, Misantla, Veracruz, P.C. 93821 México (e-mail: sparguijoh@itsm.edu.mx).*

Abstract: Today cardiovascular diseases are the first cause of death in the world. This paper specifically affront the problem of preprocessing and reduction of the dimensionality of an arrhythmia data set which contains electrocardiogram (ECG) signal data. Subsequently, the classification of different types of heart arrhythmias problem will be solved through different techniques of Artificial Intelligence (AI). It is intended to test between four and six techniques like K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN) trained by Genetic Algorithms (GA), and others, and at the end get the technique with the best accuracy to be implemented on Cloud Computing (CC).

Keywords: Heart diseases; electrocardiogram; data mining; classification; PCA;

1. INTRODUCTION

According to the World Health Organization (WHO) cardiovascular diseases (CVD) are the leading cause of death globally (Mendis, S., Puska, P. and Norrving, B. 2011), and it is estimated that by 2030 approximately 23.3 million of people will die by CVD (World Health Organization. 2017). According to the National Institute of Statistics and Geography (INEGI), in Mexico CVD are the leading cause of death too (INEGI. 2016).

It is important to search for implementation of new strategies that contribute to reduce the mortality rates attributed to CVDs, as shown in the Multinational Monitoring of Trends and Determinants of Cardiovascular Diseases (WHO MONICA Project) project that in a period of ten Years managed to reduce mortality from coronary heart disease and stroke in many of the 38 MONICA populations (Tunstall-Pedoe, H. 2003).

Several studies reveal that bioinformatics is a very important area in the fields of data mining (DM). DM is to discover interesting and novel patterns in large-scale data (Zaki, M J and Meira, W Jr. 2014). AI is a fundamental pillar of DM, and it provides techniques such as, for example, ANNs.

2. MATERIALS AND METHODOLOGY

2.1 Software and hardware used

Computing was made on a computer with *Intel(R) Core(TM) i3-3217U CPU @ 1.80Ghz* processor and *12GB of RAM* memory, *Windows 8 Enterprise* OS. Also was used *RStudio Version 1.0.136 2009-2016 RStudio, Inc.* running on a *R version 3.3.2 (2016-10-31).*

2.2 Artificial Neural Networks

ANN is an AI tool which tries to emulate the human brain working. The *perceptron* was the first ANN model, which consists of a single neuron with multiple inputs, weights, activation function and a single output (Gurney, K., 1997). See fig. 1 (Sarangapani, J., 2006). In addition to perceptron there are other models such as *back propagation*, *hopfield*, etc., an ANN can be multilayer that is having several hidden layers where each layer has one or more neurons which can be connected in different ways to each other (Gurney, K., 1997).

Fig. 1 depicts the mathematical model of a neuron which show the *inputs* x_n, the *weights* v_n, the firing threshold v_0 (also called *the bias*), the *activation function* $\sigma(.)$, and the *output* y which is expressed by the equation 1.

$$y = \sigma\left(\left(\sum_{j=1}^{n} v_j x_j\right) + v_0\right). \quad (1)$$

The ANN training is an important phase, because it is the weight adjusting process by the method of the choised model, once calibrated the ANN is ready to the classification process.

2.3 Description of data set

Arrhythmia data set used was obtained from the *UCI Machine Learning* (Guvenir, H. A., Acar, B. and Muderrisoglu, H., 2017). This data set has records from 452 patients described by 279 features, 73 features are nominal and the remaining 206 features are linear, the first four features are patient's general data like: age, sex, height and weight, the remaining 275 features are extracted from patient's ECG. Each record is classified in one of the 16 classes as show in table 1, classes 11, 12 and 13 do not have

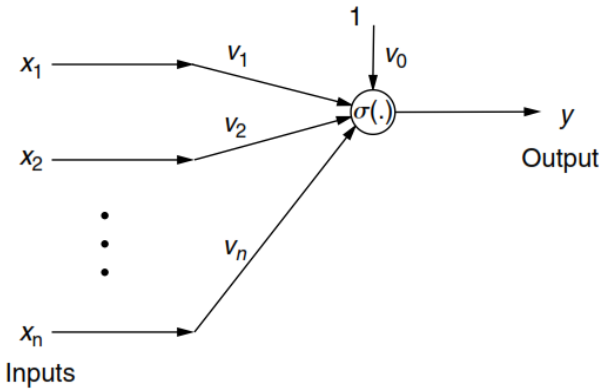


Fig. 1. Mathematical model of a perceptron (neuron)

any instance and in the *Others* class are 22 instances what are unclassified. Also about 0.33% of the feature values in this data set are missing.

Table 1. Class distribution of arrhythmia types

Class code	Class	Num. of Instances
01	Normal	245
02	Ischemic changes (Coronary Artery Disease)	44
03	Old Anterior Myocardial Infarction	15
04	Old Inferior Myocardial Infarction	15
05	Sinus tachycardy	13
06	Sinus bradycardy	25
07	Ventricular Premature Contraction (PVC)	3
08	Supraventricular Premature Contraction	2
09	Left bundle branch block	9
10	Right bundle branch block	50
11	1. degree AtrioVentricular block	0
12	2. degree AV block	0
13	3. degree AV block	0
14	Left ventricule hypertrophy	4
15	Atrial Fibrillation or Flutter	5
16	Others	22
	Total	452

Some data set features extracted from patient's ECG signal are: QRS duration, RR, P-R, and Q-T intervals, duration average in msec. of P and T intervals, heart rate per minute and others.

2.4 Data preprocessing

Missing data and others. Arrhythmia data set used contains 408 missing values distributes in 4 features as show in table 2.

Feature J has 376 missing values which represent the 83.19% of the feature data; this feature was removed from data set because there are a high percentage of missing values.

Missing values of the features T, P, $QRST$ and *Heart rate* were calculated by the mean of all values in the feature which belong into the same class of the missing value, using equation 2. However, the feature P has 5 missing data that all belong in class 15, in this case were used all non-missing feature P values to calculate the mean (using equation 2) and repair the data.

$$\frac{1}{n}\sum_{i=1}^{n} x_i. \tag{2}$$

Table 2. Missing data per feature

Num. of feature	Feature name	Total of missing data	% respect to the feature data
11	T	8	1.77%
12	P	22	4.87%
13	QRST	1	0.22%
14	J	376	83.19%
15	Heart rate	1	0.22%
	Total	408	

All values of features showed in the table 3 are 0, what means that those features do not provide enough information to separate the 16 classes and they only consume computing time. Therefore, those features were removed from data set.

Table 3. Features which all their values are 0

Num. of feature	Feature name
20	S' wave average width of channel DI
68	S' wave average width of channel AVL
70	Existence of ragged R wave average width of channel AVL
84	Existence of ragged P wave average width of channel AVF
132	Existence of ragged P wave average width of channel V4
133	Existence of diphasic derivation of P wave average width of channel V4
140	S' wave average width of channel V5
142	Existence of ragged R wave average width of channel V5
144	Existence of ragged P wave average width of channel V5
146	Existence of ragged T wave average width of channel V5
152	S' wave average width of channel V6
157	Existence of diphasic derivation of P wave average width of channel V6
158	Existence of ragged T wave average width of channel V6
165	S' wave amplitude of channel DI
205	S' wave amplitude of channel AVL
265	S' wave amplitude of channel V5
275	S' wave amplitude of channel V6

After removing 18 features the data set contains 452 records and 261 features. For more *Arrhythmia data set* information see (Guvenir, H. A., Acar, B. and Muderrisoglu, H., 2017).

Feature scaling. The range of the values in the features are different, what means that, some features with high values can dominate features with low values, then the classify model will be influenced by dominating features. To avoid it, data set is normalized using the equation 3. All data set was normalized per feature.

$$x_{norm} = \frac{x - min(X)}{max(X) - min(X)}. \tag{3}$$

Splitting the data set into the training set and test set. For splitting data set were randomly used 2/3 of records to create the training set and 1/3 of records to create the

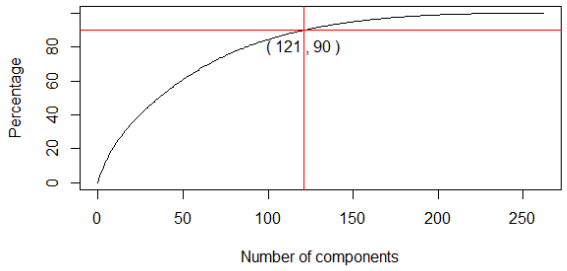


Fig. 2. Cumulative percentage of standard deviation saving 90% of information

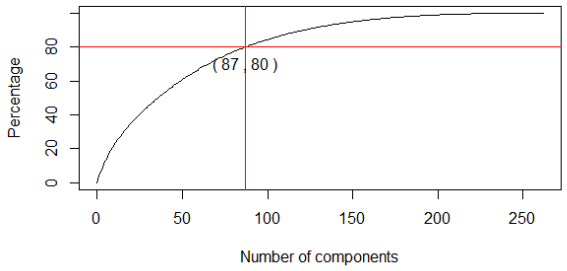


Fig. 3. Cumulative percentage of standard deviation saving 80% of information

testing set. Finally training set contains 301 records and testing set contains 101 records.

2.5 Reduction of dimensionality

The *curse of dimensionality* is a problem caused by the exponential increase of volume associated with adding extra dimensions to Euclidean space (Bellman, 1957 cited in Keogh, E. and Mueen, A., 2011). When a large multi-features dataset is analyzed *Principal Component Analysis* (PCA) is an optional technique for reduce its dimensionality:

"It replaces the p original variables by a smaller number, q, of derived variables, the principal components, which are linear combinations of the original variables. Often, it is possible to retain most of the variability in the original variables with q very much smaller than p. Despite its apparent simplicity, principal component analysis has a number of subtleties, and it has many uses and extensions. A number of choices associated with the technique are briefly discussed, namely, covariance or correlation, how many components, and different normalization constraints, as well as confusion with factor analysis." (Jolliffe, I., 2002.).

Arrhythmia data set contains 261 features (variables), those are a lot of features, and applying PCA technique in the figures 2 and 3 we can see how many components are needed to retain the variability of original variables. Fig. 2 show what 90% of the variability can be retained mapping 261 data set features to 121 variables, compressing data set to 46.36%. In the other hand, fig. 3 show what 80% of the variability can be retained mapping 261 data set features to 87 variables, compressing data set to 33.33%.

2.6 Cloud Computing

Some authors define *CC* as:

(1) "... can be dened as a new style of computing in which dynamically scalable and often virtualized resources are provided as a services over the Internet." (Furht, B., and Escalante, A., 2010).
(2) "... refers to both the applications delivered as services over the Internet and the hardware and system software in the datacenters that provide those services." (Fox et al., 2010)

Virtualization is a technology that allows to create different computing environments (simulation of software and hardware) that is expected by the guest, and it is a fundamental element of CC (Buyya, R., Vecchiola, C. and Selvi, S.T., 2013). "Virtualization confers that degree of customization and control that makes cloud computing appealing for users and, at the same time, sustainable for cloud services providers" (Buyya, R., Vecchiola, C. and Selvi, S.T., 2013).

Some *cloud services providers* are:

- *Amazon EC2* (Amazon Web Services, 2017)
- *Verizon* (Verizon Enterprise, 2017)
- *GoGrid* (Datapipe, 2017)
- *Microsoft Azure* (Microsoft, 2017)

3. CONCLUSION

In this work and arrhythmia data set with about 0.33% of missing data was preprocessed to repair and remove data, all features data set were scaled to avoid the dominate by higher features on the classify result. The scaled arrhythmia data set was splitted into training set with 301 records (66.6% of the total) and testing set with 101 records (33.4% of the total). Finally, through PCA was achieved to map and reduce from 261 features to 121 principal components, retaining 90% of the data variability.

Next activities to do are:

- try different IA techniques of classification,
- evaluate them and select the best one to be implemented under an architecture of cloud computing.

REFERENCES

©Amazon Web Services (AWS), Inc., 2017. *Amazon Web Services, Inc* [online]. Consulted on March 3th, 2017, https://aws.amazon.com/es/

Buyya, R., Vecchiola, C. and Selvi, S.T., 2013. *Mastering cloud computing: foundations and applications programming.* Newnes.

©Datapipe, Inc. 2017. *GoGrid - A Datapipe Company* [online]. Consulted on March 3th, 2017, https://www.datapipe.com/gogrid

Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A. and Stoica, I., 2009. Above the clouds: A Berkeley view of cloud

computing. *Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Rep. UCB/EECS*. 28(13), p.2009.

Furht, B., and Escalante, A., 2010. *Handbook of cloud computing*. (Vol. 3). New York: Springer.

Gurney, K., 1997. *An introduction to neural networks*. CRC press.

Guvenir, H. A., Acar, B. and Muderrisoglu, H., 2017. *UCI Machine Learning Repository* [online]. Consulted on February 23th, 2017, http://archive.ics.uci.edu/ml/datasets/Arrhythmia

INEGI (2016). *Estadísticas por tema* [online]. Consulted on January 30th, 2017, http://www3.inegi.org.mx/sistemas/sisept

Jolliffe, I., 2002. *Principal component analysis*. John Wiley & Sons, Ltd.

Keogh, E. and Mueen, A., 2011. Curse of dimensionality. In *In Encyclopedia of Machine Learning* (pp. 257-258). Springer US.

Mendis, S., Puska, P. and Norrving, B., 2011. *Global atlas on cardiovascular disease prevention and control*. World Health Organization.

©Microsoft, 2017. *Microsoft Azure* [online]. Consulted on March 3th, 2017, https://azure.microsoft.com/es-es/

Sarangapani, J., 2006. *Neural network control of nonlinear discrete-time systems* (Vol. 21) (pp. 1-74). CRC press.

Tunstall-Pedoe, H. 2003. *MONICA, monograph and multimedia sourcebook: world's largest study of heart disease, stroke, risk factors, and population trends 1979-2002*. World Health Organization.

©Verizon Enterprise, 2017. *Verizon Enterprise* [online]. Consulted on March 3th, 2017, http://www.verizonenterprise.com/products/it-solutions/cloud/

World Health Organization. (2017). *Global Health Observatory (GHO) data* [online]. Consulted on January 28th, 2017, http://www.who.int/gho/en/

Zaki, M J and Meira, W Jr. 2014 *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

Proposal of a Multi-agent System for the generation of school hours in the ITSM

Francisco Adolfo Aguilar Gómez, Jorge Mario Figueroa García *

* *Instituto Tecnologico Superior de Misantla, Veracruz, C.P. 93820 MX (web: www.itsm.edu.mx).*

Abstract: In the academic management of universities, the scheduling of schedules and the allocation of rooms is a complex problem. There are multiple variables to generate a school schedule, the most important variables to contemplate in the generation of schedules are the restrictions that are established by the institutions. This paper describes the problem of scheduling at the Instituto Tecnológico Superior de Misantla (ITSM), and presents a proposal for a Multiagent System for the generation of school schedules that could be implemented in the ITSM Integral System.

Keywords: Timetabling, Scheduling, Soft Restrictions, Hard Restrictions, Metaheuristics, Multi-Agent System, Optimization Techniques.

1. INTRODUCTION

1.1 Problem

In total there are 9 career in the ITSM, each one of which has several semesters and the race leaders establish the schedules of each group, if it is a regular group the schedule is assigned in uninterrupted sequence of hours, but there are cases in That the group is not regular and makes necessary to open a course that is marked out of school planning, these special cases are very recurrent and wreak havoc on the activities making it more difficult to create an academic load per group. But not only does the problem end in the group, there are also students who are not regular and who must individually arm their schedules without denying the courses marked by the curriculum with the few teachers shared between careers in the institution. In addition to the complications mentioned above is also the fact that the number of classrooms is barely enough to cover student demand making it necessary to optimize the allocation of each one. Classes at the ITSM Start at 7:00 am and end at 9:00 pm.

Multiagent Systems are an area of knowledge within Distributed Artificial Intelligence. But they also constitute a new methodological approach to the study and characterization of the behavior of complex systems.

As an area of knowledge it allows a new way of dealing with and approaching real systems. It consists in the development of new systems topologies that allow and take advantage of the autonomous interaction of artificial agents between them and with humans, thus linking with The most authentic essence of the origins of AI.

As methodology of analysis and study of complex systems allow us to characterize them from the individual entities that make up. That is, the analysis of the behavior that emerges from the system once it has finished the specifications of the agents involved. This approach calls the 'bottom-up', and in the Social Sciences allow the development of the third way in scientific research: the generative method, López Paredes A (2001).

Multi-Purpose Systems can be used in the two lines of work mentioned above within the IO. The first, linked to Computational Research beyond conventional operational research. The second, methodological, commonly called "Agent-Based Modeling".

2. DEVELOPMENT

The proposed model of optimization based on Multiagent Systems is presented below to improve the administrative process of programming the ITSM school schedule.

The two components in modeling the problem are as follows:

- Data Component.
- Algorithmic Component or logical.

2.1 Data Component.

The assignment matrix is a three-dimensional matrix, these dimensions are the functional hours of classrooms, days of the week and courses for all groups in the ITSM. (Dimension: Hours x Days x Courses) See Fig. 1.

The important variables to consider in the classroom assignment are the following:

- Aula
  - · Capacity.
  - · Availability.
- Curso
  - · Professor
  - · Number of students in the course.
  - · Hours required for the course per week.

Fig. 1. Representation of schedule schedules in the form of blocks

These variables are contemplated in the restrictions that the schedule must meet.

Strong Restrictions: Strong constraints, also known as hard constraints, are those conditions that, strictly and without exception, must satisfy the model, otherwise the solution will be rejected. The strong restrictions considered for this work are shown below, Moreno (2011).

(1) A group can only be assigned to a single classroom and time slot.
(2) Two groups can be assigned to the same classroom at the same time, provided that these groups are previously enabled for this.
(3) A classroom can be assigned to two or more groups at a time (day-time), only if its capacity, and if the characteristics of the groups, allow it.
(4) A classroom should only be scheduled at a time when it is available.
(5) The hours programmed weekly for a group (subject-group) must be those required by the subject.
(6) A group (subject-group) must be programmed in a classroom with sufficient capacity for its students.
(7) The daily hours of class for a group must be programmed consecutively (block) and in the same classroom.
(8) A group of a subject either theoretical or practical should be programmed in a classroom for this.
(9) A subject should only be programmed within the defined time horizon.
(10) A group can only be programmed if assigned to a teacher.
(11) A teacher can only be assigned to a single group in a single schedule.
(12) A teacher should only be scheduled at a time when it is available.
(13) A teacher can only be programmed in the subjects of his / her profile.
(14) The number of class hours scheduled for a teacher can not exceed the dedication that his category defines for this.

Weak Restrictions: Weak constraints, also known as soft constraints, are those conditions that, desirably or as far as possible, attempt to satisfy the model and which will be a criterion for the selection of the best quality solution. The following are the weak constraints considered for this work, Moreno (2011).

(1) The groups for each subject should be programmed in the teachers' preference schedule.
(2) In case a teacher has in his profile more than one subject to dictate, they should be assigned according to their preference.
(3) Groups belonging to the same level and academic program, should require the minimum displacement between classrooms.
(4) A group must be programmed in a classroom whose capacity is equal to the number of students that make up the group.

2.2 Función Objetivo.

Castrillo (2011), In order to evaluate each of the generated solutions, the following function is defined:

$$Fitness = Min\left(\sum_{j} K_j\right)$$

If it's, do not violate the restriction:

$$K = 0$$

If the restriction is less than 5:

$$K = 10 * Restriction\ rating$$

If the violated restriction is 5 or more:

$$K = \infty$$

3. CONCLUSION

As a result of the research presented, it is possible to conclude that the school scheduling model is a very complex task to perform, with many variables which must be considered, but multi-agent systems are a possible solution for optimization of The school resources in the institution, thanks to the agents can work cooperatively without supervision to avoid conflicts that may hinder the final solution, satisfying both the administrative area, teachers and students.

REFERENCES

Castrillo, V.F.S.O.D. (2011). Diseño de una metodología basada en técnicas inteligentes para la distribución de procesos académicos en ambientes de trabajo job shop. *Avances en Sistemas e Informática*, 114–116.

López Paredes A, Hernández C, P.J. (2001). Towards a new experimental socioeconomics: Complex behavior in bargaining. *Journal of Socioeconomics*, 200.

Moreno, R. (2011). Modelo para la asignación de recursos académicos en instituciones educativas utilizando técnicas metaheurísticas, búsqueda tabú. *Dpto. de Ingeniería Industria)*, 488.

Generation of a model implemented in Java that allows it Preprocessing, segmentation, and classification of species of photosynthetic protists from microscopic images ⋆

N. Ramirez * **S. Arguijo** **

* *Instituto Tecnolgico Superior de Misantla, Misantla, CO 93821 MEX (e-mail: 162t0051@itsm.edu.mx).*
** *Instituto Tecnolgico Superior de Misantla, Misantla, CO 93821 MEX (e-mail: sparguijoh@itsm.edu.mx).*

Abstract:
Currently, the treatment, segmentation, and classification of images in a large number of current scientific texts are done by segmenting each of these processes and based on different technologies such as Matlab, imageJ, multi-language frameworks, etc.
Many tools allow the processing of images in a very friendly way to the user, but when it is required to generate a complete solution from multiple technologies, this implementation is very difficult or expensive.
Therefore, a tool is presented that allows to work with the classification model of the species, but allows the results to be partialized in each of its stages (image processing, shape segmentation, and pattern-based classification), in order to be able to customize and be used as a support tool in new applications.

Keywords: images, segmentation, enhancement, analysis, matching, modelling.

1. INTRODUCTION

–

2. PROCEDURE FOR PAPER SUBMISSION

Next we see a few subsections.

2.1 Related Jobs

The processing of the images is extremely important because in a level of superior accuracy in the delimitation of the contour of the images leads to obtaining better results in the classification of them.

J. Haggerty et al. concluded that in some cases the image segmentation method for cases where color shades are very similar to the example in the segmentation of epidermal tissue with and without histopathological damage, they used a series of additional techniques to segment them, so they developed an algorithm of transformation to the color space L * a * b * with the information of the image intensity, as well as the normalization of the color to provide robustness. Finally, the optimized image is submitted to a threshold to create a binary image and subsequent to the segmentation of the same, applying all the techniques was obtained to obtain a precision of 96.5% without the interaction of the user Pereira Borges et al. (2015-08-28).

⋆ Instituto Tecnolgico Superior de Misantla's property.

Once the image is obtained in a monochrome format, Matrix with binary values to later be able to apply the techniques of learning Supervised based on a model of recognition and classification of images.

M. Kloster concluded that many of the solutions to the problem of taxonomic identification of diatoms were partially described or did not explicitly specify the tools or software framework used to implement the applied methodology, to this only a small collection of tools and source files in MATLAB, ImageJ and C are available. However, these only represent fragments of the workflow of the identification of diatom species. SHERPA (SHAPE Recognition, Processing and Analysis) was born in C #.NET which provides a complete workflow, from segmentation, extraction of traits, application of multiple segmentation methods and the coincidence of multiple contours of the objects with a set of template schemes to allow a broad taxonomic analysis, generating a with 93% ability in the segmentation of diatom species Kloster et al. (25 June 2014).

2.2 DataSet

The images were obtained from the Protistas Information Server, which expresses that the use is free of charge as long as its use is for purposes of academic or educational study.

The use of the images or other data in this database is free of charge if the objective is academic study or educational purposes. However, if you use them in public space (eg printing, meetings, internet, etc.), you must obtain permission from the copyright owner before use.

Thanks to the collaborator of the site John C. Kingston since he has the copyright of the diatom species Cymatopleura (Bacillariophyceae and Heterokonta), which will be used in the present work Kingstone (2017).

2.3 Expected results

The stages of the image correction model, presented in Figure 1, initially require user intervention at the beginning and at the end of some image processing, thus establishing the initial parameters such as image background and Palette of colors that make up the diatoms, with this information the model will perform the treatment of the image, at the end will request again the intervention of the user to evaluate the result and make the relevant calibrations, this way the model will have the possibility of being used in different scenarios besides the objective of this work.

The final goal is to obtain a two-dimensional binary matrix that can be processed by the most common classification methods, such as Vector Support Machines or Naive Bayes algorithms.

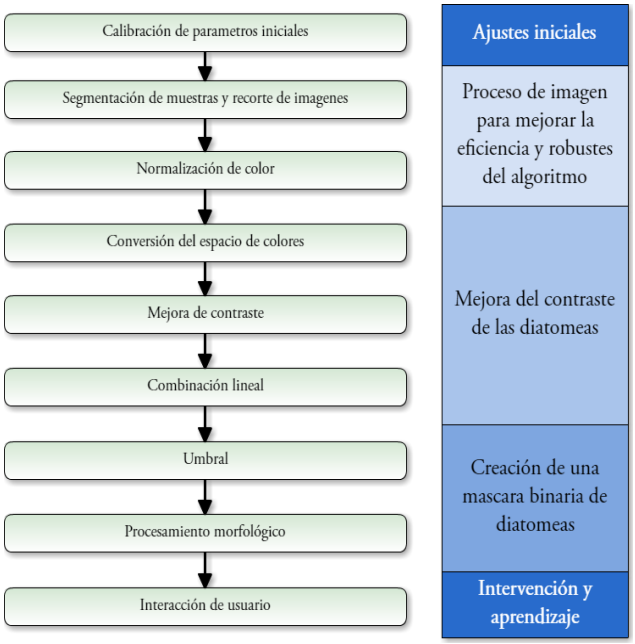


Fig. 1. Stages of image processing.

2.4 Convert to 8-bit Grayscale image

Based on the fact that every image we receive will be in color, the first thing to do is to convert the image to 8-bit grayscale, for this will be used the most traditional way to do it.

A common strategy is to use the principles of photometry or, more broadly, colorimetry to match the luminance of the grayscale image to the luminance of the original color image Poynton (July 17, 1998). This also ensures that both images will have the same absolute luminance, as can be measured in its SI units of candelas per square meter, in any given area of the image, given equal whitepoints. In addition, matching luminance provides matching perceptual lightness measures, such as $\boldsymbol{L^*}$ (as in the 1976 CIE Lab color space) which is determined by the linear luminance $\boldsymbol{Y}$ (as in the CIE 1931 *XYZ* color space) which we will refer to here as $\boldsymbol{Y_{linear}}$ to avoid any ambiguity.

To convert a color from a colorspace based on an *RGB* color model to a grayscale representation of its luminance, weighted sums must be calculated in a linear *RGB* space, that is, after the gamma compression function has been removed first via gamma expansion.

For the *sRGB* color space, gamma expansion is defined in the Eq. (1).

$$C_{linear} = \begin{cases} \frac{C_{srgb}}{12.92}, & C_{srgb} \leq 0.04045 \\ \left(\frac{C_{srgb}+0.055}{1.055}\right)^{2.4}, & C_{srgb} > 0.04045 \end{cases} \quad (1)$$

where $\boldsymbol{C_{srgb}}$ represents any of the three gamma-compressed $\boldsymbol{sRGB}$ primaries (R_{srgb}, G_{srgb}, and B_{srgb}, each in range [0,1]) and Clinear is the corresponding linear-intensity value (R_{linear}, G_{linear}, and B_{linear}, also in range [0,1]). Then, linear luminance is calculated as a weighted sum of the three linear-intensity values. The *sRGB* color space is defined in terms of the CIE 1931 linear luminance $\boldsymbol{Y_{linear}}$, which is given by Eq. (2).

$$Y_{linear} = 0.2126R_{linear} + 0.7152G_{linear} + 0.0722B_{linear} \quad (2)$$

Color images are often built of several stacked color channels, each of them representing value levels of the given channel. For example, *RGB* images are composed of three independent channels for red, green and blue primary color components; *CMYK* images have four channels for cyan, magenta, yellow and black ink plates, etc.

An example of color channel splitting of a full *RGB* color image is the Figure 2. The column at left shows the isolated color channels in natural colors, while at right there are their grayscale equivalences.

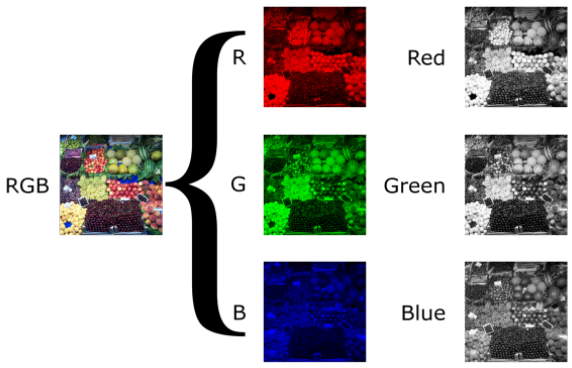


Fig. 2. Composition of RGB from 3 Grayscale images.

2.5 SobelFeldman operator

The operator uses two 3 * 3 kernels which are convolved with the original image to calculate approximations of the derivatives one for horizontal changes, and one for vertical. If we define A as the source image, and G_x and G_y are two images which at each point contain the horizontal and vertical derivative approximations respectively, the computations are as the Eq. (3) and Eq. (4)

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} * A \tag{3}$$

$$G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * A \tag{4}$$

where * here denotes the 2-dimensional signal processing convolution operation.

Since the Sobel kernels can be decomposed as the products of an averaging and a differentiation kernel, they compute the gradient with smoothing. For example, G_x can be written as Eq. (5).

$$G_x = \begin{bmatrix} +1 & 0 & -1 \\ +2 & 0 & -2 \\ +1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} [+1 \ 0 \ -1] \tag{5}$$

The x-coordinate is defined here as increasing in the "right"-direction, and the y-coordinate is defined as increasing in the "down"-direction. At each point in the image, the resulting gradient approximations can be combined to give the gradient magnitude, using the Eq. (6).

$$G = \sqrt{G_x^2 + G_y^2} \tag{6}$$

Using this information, we can also calculate the gradient's direction in Eq (7).

$$\Theta = atan\left(\frac{G_y}{G_x}\right) \tag{7}$$

where, for example, Θ is 0 for a vertical edge which is lighter on the right side.

Since the intensity function of a digital image is only known at discrete points, derivatives of this function cannot be defined unless we assume that there is an underlying continuous intensity function which has been sampled at the image points. With some additional assumptions, the derivative of the continuous intensity function can be computed as a function on the sampled intensity function, i.e. the digital image. It turns out that the derivatives at any particular point are functions of the intensity values at virtually all image points. However, approximations of these derivative functions can be defined at lesser or larger degrees of accuracy.

The Sobel-Feldman operator represents a rather inaccurate approximation of the image gradient, but is still of sufficient quality to be of practical use in many applications. More precisely, it uses intensity values only in a 33 region around each image point to approximate the corresponding image gradient, and it uses only integer values for the coefficients which weight the image intensities to produce the gradient approximation.

The result of edge search processing by the Sobel-Feldman filter generates a grayscale image with the edges in light tones [Figure 3].

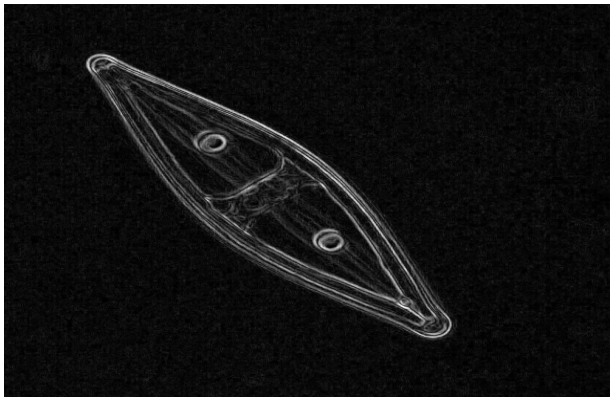


Fig. 3. Using findEdges funtion.

K₂ K²

2.6 Calibration of initial parameters in original image

Based on the similarity of the segmentation work of epidermal tissue with hematological damage in hematoxylin and eosin stained human skin images by J. M. Haggerty et al., A model is generated that is able to distinguish the taxonomy of diatoms from the family of theHeterokontas adapting the L * a * b technique of image processing. Also developed is the Java model with ImageJ recreating the processes used by the L * a * b technique developed in MATLAB Version 7.11, R2010b.

Due to the fact that the diatom's frustum often presents a pigmentation similar to the color of the background of the image, it is necessary to initially identify the color palette that intervenes both from the bottom of the image and from the most representative colors of the diatom [Figure 4]. In this way, the model uses this information as a basis to have an initial basis when dealing with the image.

2.7 Otsu Threshold

The first stage of the model is to segment the pixels in the image representing the background area. This first segmentation increases the efficiency of the algorithm by limiting the number of pixels being processed during subsequent steps. Although segmentation of theSample can be achieved using only the background pixels as they have less variation between images.

This model will be supported by two factors, first and based on the initial and historical calibration the system will obtain the value of K1, which is a data obtained by the

Fig. 4. Identification of the base colors of a diatom.

user, the value of K2 is obtained by creating a composite image from pixels , adding the intensities of red, green and blue (R, G and B) for each pixel of the RGB image, taking the most frequent K value to approximate the background color.

With these 2 approximations, we will obtain the segment of colors K which is K2 with a correction range of the color +/- of K1. Later this value is stored as the background threshold.

$$K_2 = R + G + B \tag{8}$$

$$K = K_{min}^{max} = \{\mathbb{N} \in K_2 + K_1 | \mathbb{N} \in K_2 - K_1\} \tag{9}$$

$$bg_{thresh} = \{k_{ij} | k_{ij} \text{ es un elemento de } K \text{ y } k_{ij} > 0\} \tag{10}$$

The result is the Figure 5

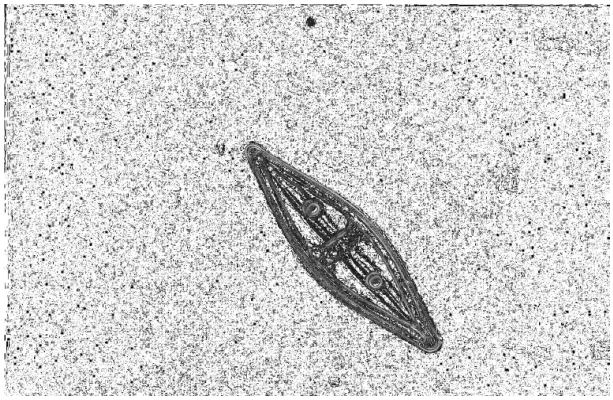


Fig. 5. Using Otsu Threshold

2.8 Erode Filter

The result is the Figure 6

2.9 Black and White Filter

The result is the Figure 7

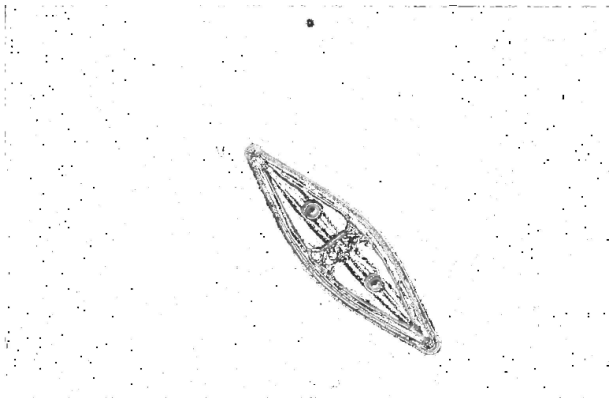


Fig. 6. Using erode filter

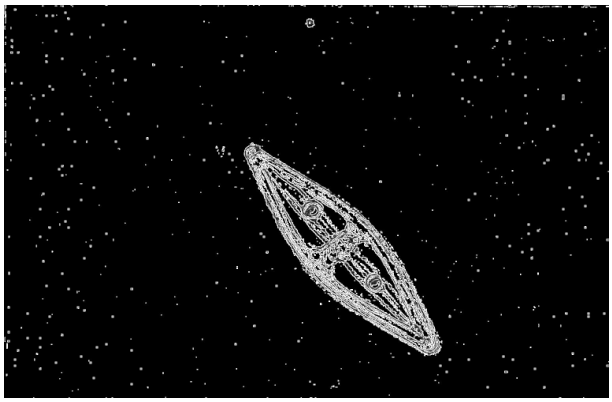


Fig. 7. Using Black and White filter

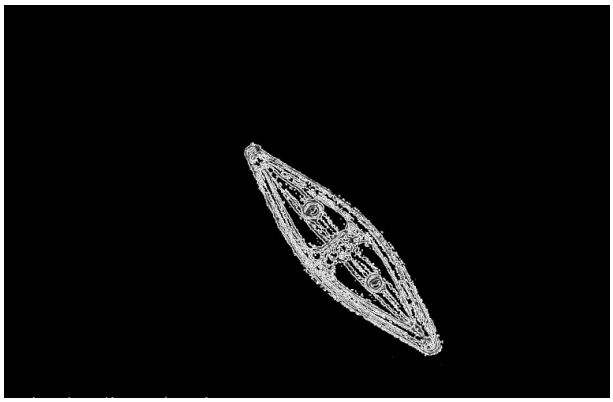


Fig. 8. Using property algorithm

2.10 Clean noise algorithm

The result is the Figure 8

3. CONCLUSION

—

ACKNOWLEDGEMENTS

—

REFERENCES

Kingstone, J.C. (2017). Protist information server. identifications of desmogonium, rhopalodia, amphipleura, diploneis, cymatopleura (bacillariophyceae, heterokonta). URL `http://protist.i.hosei.ac.jp/`.

Kloster, M., Kauer, G., and Beszteri, B. (25 June 2014). Sherpa: an image segmentation and outline feature extraction tool for diatoms and other objects. *BMC Bioinformatics*. doi:10.1186/1471-2105-15-218. URL `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-218`.

Pereira Borges, V.R., Hamann, Bernd; Silva, T.G., Vieira, A.A.H., and Oliveira, M.C.F. (2015-08-28). A highly accurate level set approach for. *Conference on Graphics, Patterns and Images*.

Poynton, C.A. (July 17, 1998). Rehabilitation of gamma. *Human Vision and Electronic Imaging III*, 232.

Arquitectura de un sistema multiagente para el monitoreo a protocolos de mujeres adolescentes embarazadas de alto riesgo.

Autores: Arcos Muñoz Sandra G,

Resumen **Este artículo presenta la descripción de una arquitectura del sistema multiagente para el monitoreo a protocolos de mujeres adolescentes embarazadas de alto riesgo de hasta 19 años, el monitorio es para ayudar a las mujeres embarazadas de alto riesgo y así puedan conservar una salud estable, utilizando programación orientada agentes.**

Palabras Claves **Sistemas Multiagentes, Inteligencia Artificial Distribuida, Programación orientada a Agentes, Salud.**

Introducción

Desde tiempos remotos ha sido común en muchos lugares del país que las mujeres se casaran y tuviera su primer bebe alrededor de los 14 o 15 años de edad, pero en la actualidad esto implica diversos riesgos sociales, psicológicos y biológicos.

Adolescentes de 15 a 19 años de edad dan a luz alrededor de 15 millones de niños o niñas cada año en el mundo [1]; y la principal causa de muerte dentro de este rango de edad es ocasionada por complicaciones durante su embarazo [2].

En la actualidad, sabemos que las mujeres embarazadas menores de 19 años tienen el doble de probabilidad de morir durante el embarazo o el parto que las mujeres de 20 a 30 años de edad [3]; sabemos también que en aquellas de 15 años de edad o menos, el riesgo aumenta 5 veces [2].

Los riesgos de las gestantes adolescentes se incrementan en los países llamados en desarrollo, ya que si bien el embarazo en la adolescencia conlleva riesgos; la muerte ocasionada por problemas relacionados con el embarazo, parto y puerperio es uno de los principales problemas de salud de las mujeres en edad reproductiva en México.

En este trabajo se desarrollará un arquitectura donde se ara el monitoreo a mujeres embarazadas de alto riesgo donde se estudiará los elementos que se involucran cuando una adolescente presenta un embarazo de alto riesgo; ya que se estima que alrededor de un 20% de los embarazos corresponde a la denominación de alto riesgo y ellos son responsables de más del 80% de los resultados perinatales adversos. [4]

El objetivo principal con este monitoreo recopilar cada una de las variables de un embarazo de alto riesgo para que el experto (médico) determine un diagnóstico con dicha información, esto se realizara tomando en cuenta las principales complicaciones, causantes de las muertes maternas y su entorno; basándose en las estadísticas se decide estudiar estos factores en las adolescentes de hasta 19 años de edad las cuales presentan un rango de mortalidad mayor.

El estudio de todas estas complicaciones determinará una alerta por cada caso que se presenta en las principales complicaciones más comunes en las embarazadas de alto riesgo.

Para generar las alertas se desarrolla una arquitectura multiagente de un sistema para el monitoreo a protocolos de las mujeres embarazadas de alto riesgo aplicando y utilizando los conocimientos de inteligencia artificial distribuida para construir un sistema multiagente que permita llevar el control y seguimiento de dichas mujeres en tiempo y forma.

La arquitectura multiagentes cuenta con siete agentes que se encargan de recopilar y enviar información que las adolescentes proporcionan con la aplicación y el doctor analiza, para dar un diagnóstico sin necesidad que las gestantes estén en el centro médico.

Materiales y Métodos

La arquitectura multiagente para el monitoreo a protocolos de adolescentes embarazadas de alto riesgo está enfocado en establecer los seguimientos y control del embarazo con ayuda de los protocolos clínicos de adolescentes embarazadas de alto riesgo, utilizando alertas y notificaciones, analizar las variables detonantes en los protocolos en mujeres embarazadas, monitorear los signos y progreso del embarazo, registrar el cuadro clínico que el especialista y el médico familiar por cada una de las gestantes.

La estructura conceptual del sistema dividida por agentes y funcionalidad de manera general. Representa los agentes divididos de acuerdo con la funcionalidad de cada uno y la interacción que existe entre ellos se muestra en la Figura 1 y se describe en la Tabla1.

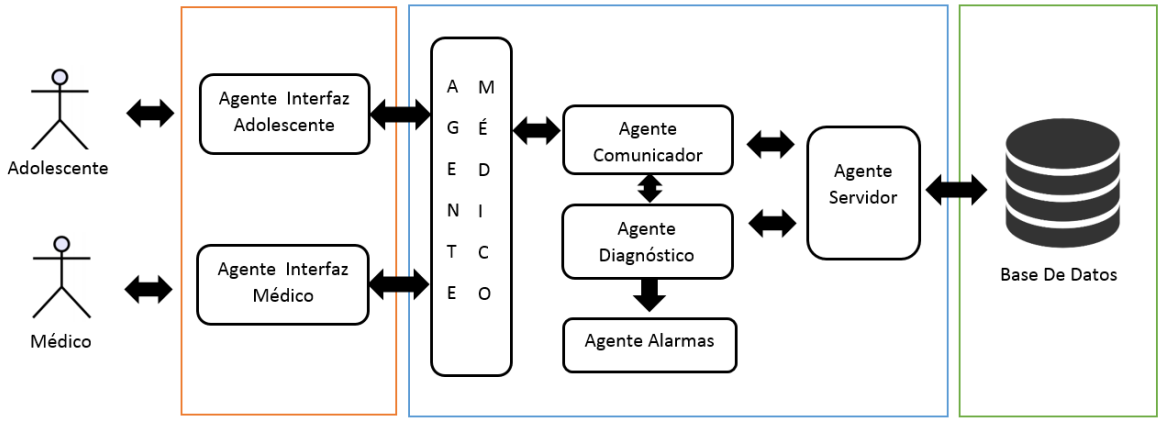


Figura 1. Arquitectura multiagente para el monitoreo a protocolos en adolescentes embarazadas.

Esta investigación propone una arquitectura multiagente para el monitoreo a protocolos en adolescentes embarazadas de alto riesgo de hasta 19 años que informa espontáneamente al médico acerca de situaciones anormales.Los agentes realizaran la comunicación con la gestante y el médico para darle seguimiento desde casa

Agente	Responsabilidad	Descripción
Interfaz Adolescente	Estable el inicio de sesión de adolescentes.	Observa e interpreta las acciones realizadas por el usuario (adolescente) en la interfaz y envía la información interpretada al agente médico.
Interfaz Médico	Estable el inicio de sesión de los médicos.	Observa e interpreta las acciones realizadas por el usuario (médicos) en la interfaz y envía la información interpretada al agente médico
Médico	Registra las actividades de los agentes interfaz médico e interfaz adolescente	Gestiona la información solicitada por el usuario desde la interfaz médico y adolescente para posteriormente enviarla al agente comunicador. Consulta y guarda registros solicitados por los agentes interfaz adolescente. Envía información al agente comunicador.

Servidor	Envía y Recibe información de todos los agentes para enviarla a la Base de Datos	Recibe y envía los datos solicitados por los agentes comunicador y diagnóstico para darle respuesta a los agentes interfaz
Comunicador	Establecer el enlace con la base de datos y el Agente Diagnóstico	Permite el envío y recepción de datos con la aplicación web, utilizando un protocolo de comunicación para enviarla al agente diagnóstico.
Diagnóstico	Monitorear toda la información que recibe el agente comunicador.	Verifica anomalías. Consulta y guarda registros solicitados por el agente médico. Mantiene informado al agente alarmas, sobre los cambios que se realicen a la información. Activa los mensajes SMS y genera historiales. Tiene comunicación con el agente comunicador
Alarmas	Enviar y generar alarmas con la información de agente Diagnóstico.	Genera alarmas, activar urgencias y notificaciones, recordar cita, consumir medicamentos; utilizando la información registrada por el usuario. Lleva el monitoreo de las alarmas. Activa notificaciones cuando el médico recibe un mensaje con alguna duda.

Tabla 2. Agentes de la arquitectura multiagente para el monitoreo a protocolos en adolescentes embarazadas

Esta investigación propone una arquitectura multiagente para el monitoreo a protocolos en adolescentes embarazadas de alto riesgo de hasta 19 años que informa espontáneamente al médico acerca de situaciones anormales.

Para ilustrar los requisitos de los usuarios, las funciones y servicios del sistema y la interacción del sistema con varios tipos de usuarios, hemos desarrollado dos diagramas de casos.

Los diagramas de casos de uso sirven como guía para el desarrollo y proporcionan una representación gráfica y simplificada de la interacción del usuario con el sistema.

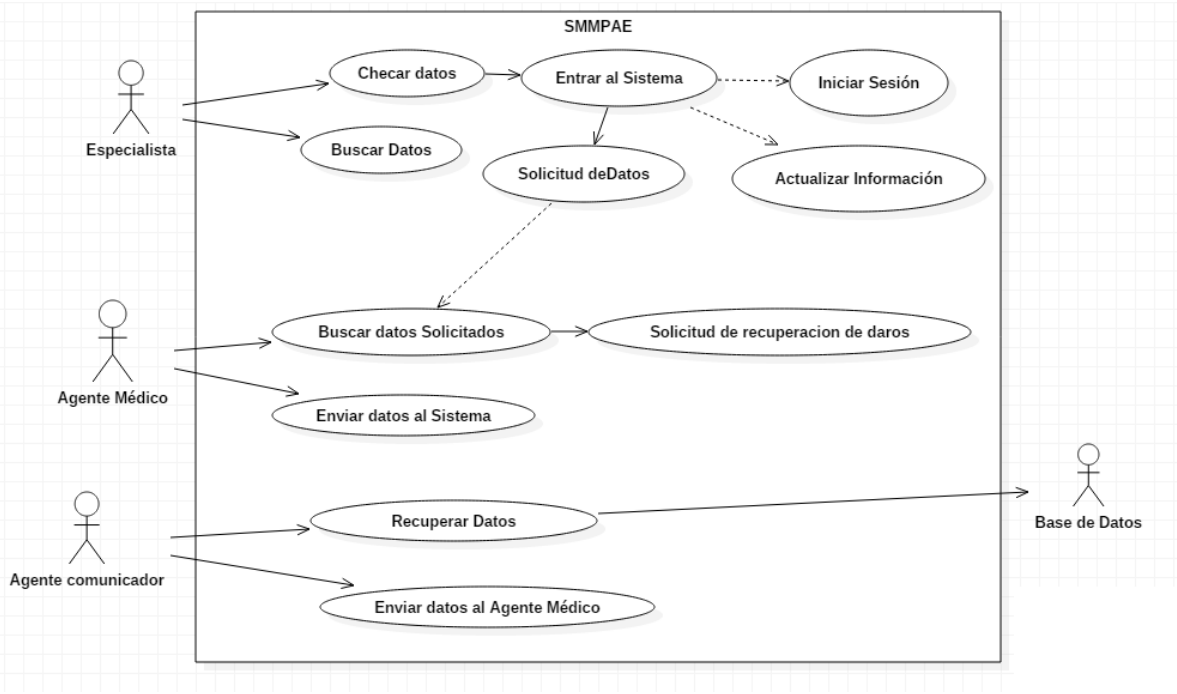


Figura 2. Diagrama de Caso de Uso desde la perspectiva de la adolescente de la Arquitectura multiagente para el monitoreo a protocolos en adolescentes embarazadas.

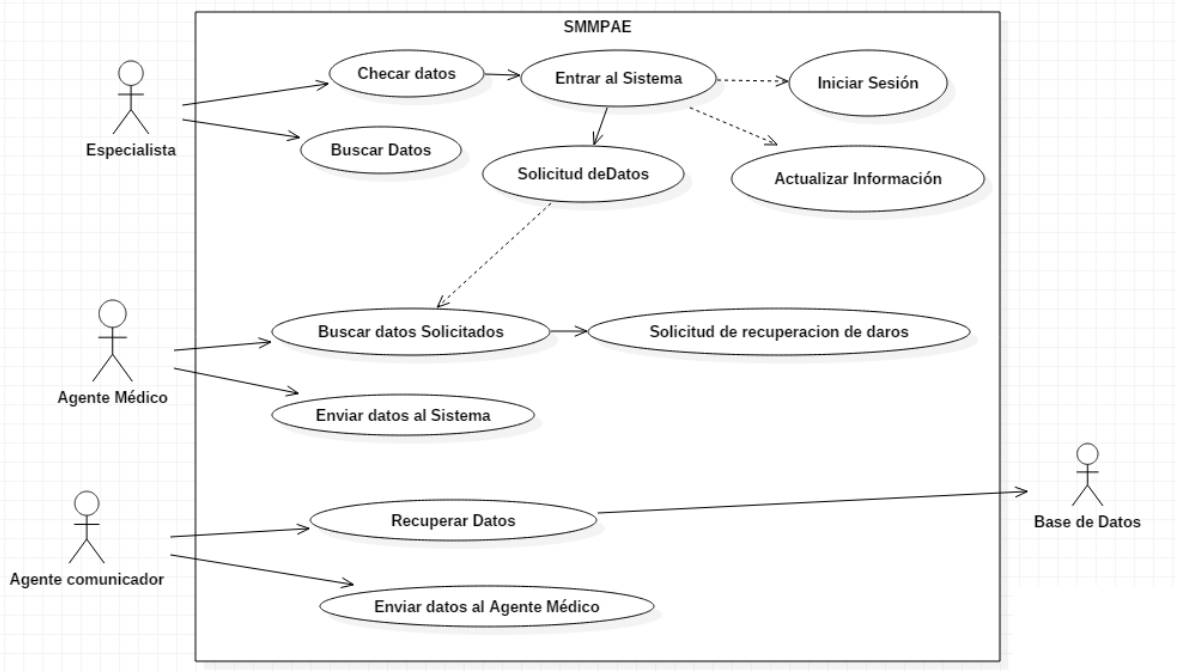


Figura 3. Diagrama de Caso de Uso desde la perspectiva de la adolescente de la Arquitectura multiagente para el monitoreo a protocolos en adolescentes embarazadas.

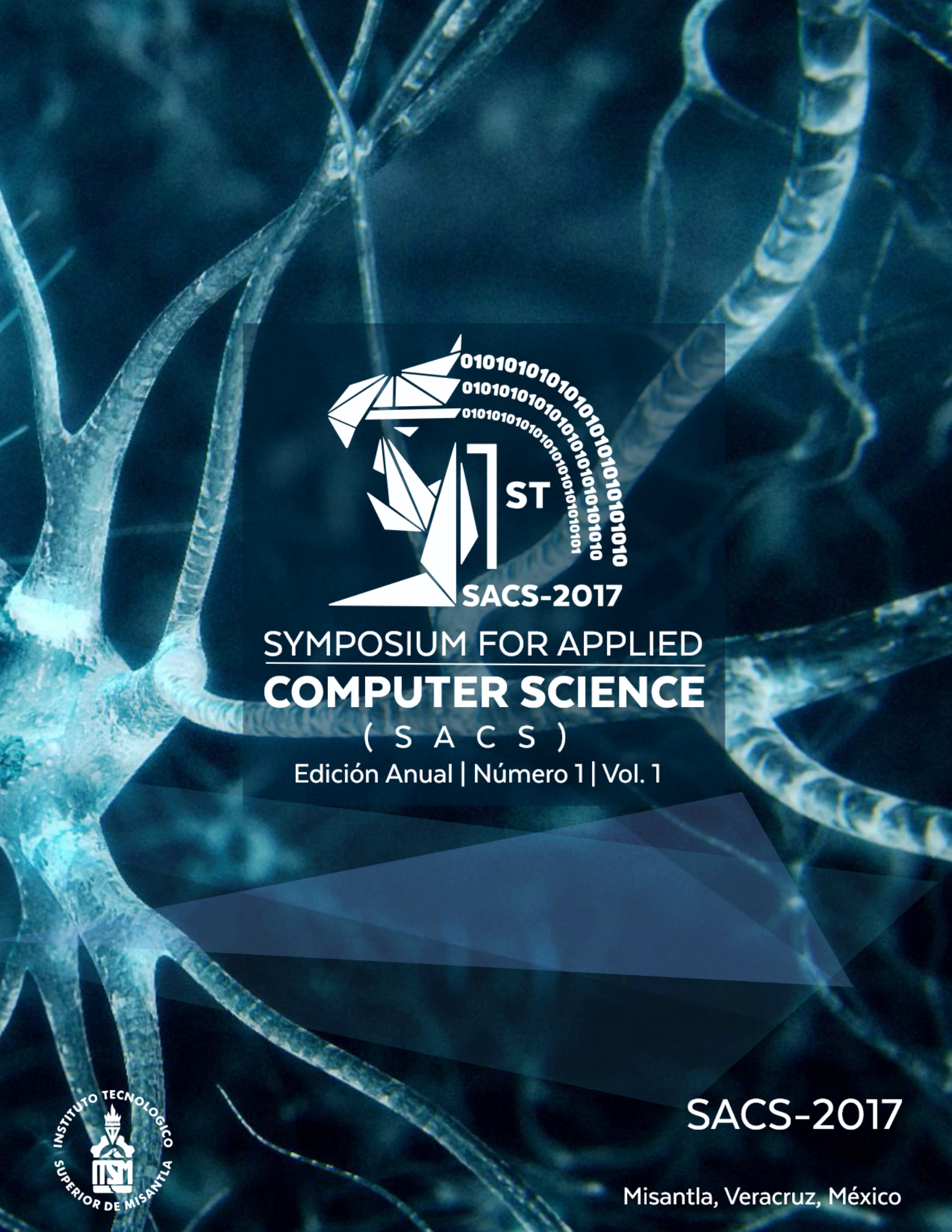
Conclusión y Trabajos Futuros

Referencias

[1] Organización Mundial de la Salud. OMS. El embarazo en la adolescencia Nota descriptiva Nº 364 Actualización de septiembre de 2014 www.who.int/mediacentre/factsheets/fs364/es.
[2] Observatorio de Mortalidad Materna en México, Indicadores 2014.
[3] INEGI. Encuesta Nacional de Dinámica Demográfica 2014.
[4]Datos de la SS en Comunicado de prensa N° 325. 23 de septiembre de 2009. 90. CONAPO, Elena Zúñiga, Secretaria Nacional. Discurso realizado el 30 de octubre del 2007 en la Ciudad de México en www.conapo.gob.mx/prensa/discursos2007.

Notas Bibliográficas

El **L.I Sandra Gabriela Arcos Muñoz** estudiante de la Maestría en sistemas Computacionales en el Instituto tecnológico Superior de Misantla, Veracruz, México, con intereses en el área Programación y Base de Datos.



SYMPOSIUM FOR APPLIED COMPUTER SCIENCE

(S A C S)

Edición Anual | Número 1 | Vol. 1

INSTITUTO TECNOLOGICO SUPERIOR DE MISANTLA

SACS-2017

Misantla, Veracruz, México