UNIVERSIDAD DEL PAPALOAPAN

Campus Loma Bonita

INGENIERÍA EN COMPUTACIÓN

SELECCIÓN DE CARACTERÍSTICAS USANDO ALGORITMOS GENÉTICOS PARA CLASIFICACIÓN DE CÁNCER DE MAMA

TESIS

QUE PARA OBTENER EL TÍTULO DE:
INGENIERA EN COMPUTACIÓN

PRESENTA:

MARLENY JUÁREZ CAYETANO

ASESOR DE TESIS:

DR. EDUARDO SÁNCHEZ SOTO

CO-ASESOR DE TESIS:

DR. SERGIO IVVAN VALDEZ PEÑA

LOMA BONITA, OAXACA. 2016

UNIVERSIDAD DEL PAPALOAPAN

Campus Loma Bonita

INGENIERÍA EN COMPUTACIÓN

LA PRESENTE TESIS TITULADA **"SELECCIÓN DE CARACTERÍSTICAS USANDO ALGORITMOS GENÉTICOS PARA CLASIFICACIÓN DE CÁNCER DE MAMA"** PRESENTADA POR LA SUSTENTANTE DE LICENCIATURA **C. MARLENY JUÁREZ CAYETANO** BAJO LA DIRECCIÓN DEL DR. EDUARDO SÁNCHEZ SOTO Y CO-DIRECCIÓN DEL DR. SERGIO IVVAN VALDEZ PEÑA, HA SIDO REVISADA Y ACEPTADA POR EL COMITÉ EXAMINADOR PARA SER DEFENDIDA EN EL EXAMEN PROFESIONAL Y OBTENER EL TÍTULO DE INGENIERA EN COMPUTACIÓN.

DR. EDUARDO SÁNCHEZ SOTO
ASESOR

DR. SERGIO IVVAN VALDEZ PEÑA
CO-ASESOR

M.C. ARIEL LÓPEZ RODRÍGUEZ
PRESIDENTE

M.C. EDUARDO ORTIZ HERNÁNDEZ
SECRETARIO

DRA. BEATRIZ C. LUNA OLIVERA
VOCAL

LOMA BONITA, OAXACA. 2016.

Dedicado a mis padres.

Agradecimientos

En el presente escrito expreso mi agradecimiento a Dios por permitirme la vida que tengo, a mi mamá Blanca por siempre estar para mi cuando la he necesitado, por su amor y sus regaños que me ayudaron a ver lo que hacía mal, a mi papá Eliud que siempre ha puesto su confianza en mis hermanos y en mi y nos ha dado amor y apoyo incondicionalmente, a mis hermanos Eliud, Elisama y Zury por hacer tanto ruido cuando necesitaba estudiar para un examen, a mi mamá Estela, a mi tía Juana, a mi tía Norma, a mi papá Felipe, a mi primo Juan y demás tíos que me regalaron cachibaches para hacer mis experimentos. Un agradecimiento especial a Alonso que desde que lo conocí ha estado cerca de mi, brindańdome su amistad, amor, comprensión y apoyo.

Agradezco a los profesores que fueron parte de mi desarrollo académico, al Dr. Ivvan Valdez por brindarme la asesoría necesaria para desarrollar mi trabajo de tesis, por su paciencia al explicar y por apresurarme para que acabara, al Dr. Eduardo Sánchez por la comprensión, paciencia, consejos y correcciones, al M. C. Ariel López por las enseñanzas, por su apoyo y sus bromas que por instantes nos hacían olvidar el estrés, a los revisores de mi tesis: el M. C. Eduardo Ortiz por la dedicación al hacer la revisión y a la Dra. Breatriz Carely Luna. A mis compañeros de generación por ayudarme a crecer personal y académicamente.

A la UNPA por permitirme ser parte de la generación 2010-2015 y al CIMAT por proporcionarme una beca para el desarrollo de mi tesis.

Resumen

El cáncer de mama, entre muchas otras enfermedades, es un problema a nivel mundial que causa miles de muertes anualmente, por lo tanto es necesario combatir esta enfermedad desde diferentes ángulos.

La asignación de un diagnóstico puede ser realizada por medio de la clasificación binaria, considerando ciertas características tomadas del paciente. Sin embargo, puede ocurrir que el número de características sea demasiado grande o que todas las características no sean necesarias para asignar un diagnóstico correcto, lo que podría insertar errores en la clasificación y/o solicitar pruebas innecesarias al paciente.

Por esta razón, esta tesis aborda el problema de la selección de características de cáncer de mama, con las que será posible proporcionar un diagnóstico acertado sin la necesidad de utilizar todas las características. Para este trabajo se toma en cuenta la Base de Datos de Cáncer de Mama en Wisconsin, que contiene características tomadas de pacientes con tumores y un diagnstico para cada uno de ellos. Para verificar la eficiencia de la metodología desarrollada se realizaron pruebas con la base de datos de libros del Mago de Oz, la cual ha sido utilizada en diversos trabajos de clasificación. Bases de datos como estas almacenan muchas características.

Para lograr este objetivo se implementó un algoritmo genético utilizando la calidad de clasificadores binarios como función objetivo, esta calidad fue medida por medio de la precisión de clasificación. Los clasificadores implementados son el *análisis discriminante lineal* y el *método del k-vecino más cercano*. El criterio de optimización del algoritmo genético fue de maximización, es decir, se esperaba que en cada iteración la función objetivo fuera mayor a la anterior. Un algoritmo genético toma distintos subconjuntos de características, los evalúa, selecciona a los subconjuntos con un valor de función objetivo mayor, entonces los recombina y los muta; el proceso es repetido hasta cumplir un criterio de paro. Finalmente, devuelve el subconjunto de características con el mejor valor de función objetivo. El algoritmo se implementó en lenguaje C.

Si bien, la meta principal de la tesis es la selección de características para el diagnóstico de cáncer de mama por medio de clasificadores binarios, la metodología desarrollada se puede utilizar para tratar otro tipo de problemas, tal y como se muestra en esta tesis.

Abstract

Breast cancer, among other diseases, is a worldwide problem that causes thousands of deaths annually; therefore it is necessary tackle this disease from different angles.

The allocation of a diagnosis can be made by means of binary classification, considering certain features taken from the patient. However, it could happen that the number of features is too large or that all features are not necessary to assign a diagnosis, which could insert errors in the classification and/or request unnecessary tests to the patient.

For this reason, this thesis tackles the problem of features selection of breast cancer, in order to provide an accurate diagnosis without the need of using all the features. This study takes into account the database of Wisconsin Diagnostic Breast Cancer, that contains features taken from patients with tumors and a diagnosis for each of them. To verify the efficiency of the methodology developed some tests are performed with the database of books of the wizard of OZ, which has been used in different case studies reported in literature. These databases as these contain a lot of features.

To achieve this objective, a genetic algorithm is implemented using the quality of the binary classifiers as target function, this quality was measured by means of the accuracy of the classification. The classifiers implemented are the Linear Discriminant Analysis (LDA) and the k-Nearest Neighbor (k-NN). The optimization criterion for the genetic algorithm is of maximization,

that is to say, it is expected that each iteration the objective function was larger than the previous. A genetic algorithm takes different sets of features, it evaluates them, selects the sets with a larger value of the objective function, then they are recombined and mutated; the process is repeated until a stopping criterion is reached. Finally, it returns the set of features with the best value of the objective function. The algorithm is implemented in C-language.

Even though, the main objective of this thesis is the feature selection for the diagnosis of breast cancer through binary classifiers, the developed methodology can be used to address other types of problems, such as it is shown in this thesis.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

De acuerdo con datos reportados por el Instituto Nacional de Estadística, Geografía e Informática (INEGI) y la Organización Mundial de la Salud (OMS), el cáncer más frecuente entre las mujeres es el de mama, que a nivel mundial representa 16 % de todos los cánceres femeninos y se estima que cada año se detectan 1.38 millones de casos nuevos. En México durante el año 2011, 30 de cada 100 mujeres que salen de un hospital por tumores malignos padecen cáncer de mama, siendo la tercer causa de muerte por cáncer en las mujeres mexicanas con 14.7 % de mortalidad. Las tasas de morbilidad hospitalaria más altas haciendo referencia a esta enfermedad se presentan en las mujeres de 60 a 69 años de edad, le sigue el grupo de mujeres de 50 a 59 años y por último el de 70 a 79 años. Resulta preocupante que el pronóstico general de la enfermedad es poco alentador [1, 4].

Entre los principales factores relacionados con la aparición del cáncer de mama se encuentran: la edad (a mayor edad se tiene más alto riesgo de padecer cáncer de mama), el inicio temprano de la menarca o menopausía tardía, el inicio de la vida reproductiva después de los 30 años de edad, la lactancia materna nula o de corta duración, el uso de anticonceptivos orales por más de cinco años, la obesidad y la exposición a la radiación [2].

Según datos reportados por el Instituto Nacional de Cancerología (INCAN) y del Registro Histopatológico de Neoplasias Malignas (dependiente de la Dirección de General de Estadística de la Secretaría de Salud) la incidencia del cáncer mamario ha ido en aumento, reportándose 11,000 nuevos casos cada año [3].

Es claro que los factores mencionados pueden darnos un indicio de si una persona puede o no tener cáncer de mama, pero también existen medidas tanto de laboratorio como las realizadas físicamente al paciente; aunque en muchos casos depende del estado de avance del padecimiento. Todas estas medidas aportan la información necesaria para definir un diagnóstico. Los diagnósticos que reciben los pacientes solo varían entre un diagnóstico benigno o un diagnóstico maligno.

En este trabajo se aborda el problema de decisión entre un diagnóstico benigno o un diagnóstico maligno, para lo que se requiere traducir las medidas obtenidas de las pruebas de laboratorio a características; el enfoque principal es seleccionar las características que aportan la información necesaria para determinar un diagnóstico acertado.

Los motivos para realizar la selección de características son la alta dimensionaldad de los datos, eliminar características redundantes y/o irrelevantes, tener capacidad para diagnósticar una enfermedad por medio de pocas características, disminuir el costo computacional; reducir el número de características puede ayudar a bajar costos al determinarse que no todas las pruebas tomadas al paciente son necesarias.

Existen muchos métodos para realizar la reducción de dimensionalidad, como lo es el PCA (Principal Component Analysis), pero a diferencia de la selección de características estos no conservan las características originales, ya que obtienen nuevas dimensiones a partir de las originales.

Para hacer la selección de características se implementó un algoritmo de optimización, que evaluó una medida de calidad obtenida de la clasificación hecha con cierto subconjunto de características hasta hallar la medida de calidad óptima; es decir, el subconjunto de caracteríscticas que posee la medida de calidad óptima es el seleccionado como el mejor para hacer una toma de decisión acertada.

Esta medida de calidad se obtuvo mediante la utilización de clasificadores sencillos que fueron entrenados con el 70 % de los datos totales, los cuales son llamados conjunto de entrenamiento (training set), y probados con el 30 % de ellos, llamado conjunto de prueba (test set). Para calcular esta medida se toma como entrada el conjunto de prueba, que al ser parte del conjunto original de datos se conoce la clase de procedencia de cada uno de los datos. Teniendo este conocimiento es posible calificar la calidad de la clasificación.

Los clasificadores utilizados fueron los siguientes: el Análisis Discriminante Lineal, mejor conocido por sus siglas en inglés LDA y el método del k-vecino más cercano, los cuales son detallados en el capítulo 4.

El entrenamiento de los clasificadores es necesario para que estos definan las clases o grupos que hay en el conjunto de datos de estudio y posteriormente, cuando un nuevo dato de procedencia desconocida ingrese al clasificador se le pueda asignar correctamente una etiqueta con la clase a la que pertenece.

Para la evaluación de estos clasificadores se utilizaron dos bases de datos: la primera es la Base de Datos de Cáncer de mama en Wisconsin, considerada en este trabajo porque ha sido utilizada en diferentes tipos de pruebas con diversos algorítmos, además contiene información real de

casos de pacientes que han padecido cáncer de mama; la segunda es la base de datos de Libros del Mago de Oz, considerada para hacer los experimentos de este trabajo porque contiene información de un problema que ha sido objeto de prueba en diversas ocasiones, se conoce cual es el resultado final, se hayan dos clases al igual que en la primera; conocer cual es el resultado de otras pruebas realizadas con ellas sirve para determinar si la metodología desarrollada en este trabajo es eficiente o no.

1.1. Antecedentes

Para abordar el problema que implica tener datos con alta dimensionalidad se han realizado diversos trabajos aplicando el uso de distintos algoritmos de clasificación, como se hace en el trabajo *Minería de datos aplicada a la detección de Cáncer de Mama* [7], aquí se aborda el problema tomando en cuenta la base de datos de cáncer de mama en Wisconsin (WBCD) y varios clasificadores que fueron comparados entre ellos por medio de una tasa de error, uno de los clasificadores probados fue el k-vecino más cercano con k=3, con una tasa de error de 3.14 % que se consideró entre las mejores, otro de los clasificadores que se probó fue una máquina de soporte vectorial (SVM) con una tasa de error de 3.0 %, después se probó un árbol de decisiones con el que se obtuvo una tasa de error de 5.4363 %; entre otros. Teniendo como conclusión que la SVM resultó con una mejor tasa de error.

En el trabajo *Estudio comparativo de técnicas de selección de características para la clasificación de lesiones de mama en ultrasonografía* [8] se realiza la selección de características de ultrasonografías de distintos diagnósticos de cáncer de mama, mediante el clasificador de análisis de componentes principales con una tasa de error máxima de 2 %.

En el trabajo *Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks* [9] se toman en cuenta datos clínicos y microarrays que contienen cerca

de 25000 genes en distintos pacientes, primero se fue reduciendo la cantidad de variables de los microarrays quitando los genes que tenian valores de 0.01 en más de 3 pacientes, lo cual dejó un total de 5000 genes, después por medio de una selección de correlaciones con un valor no menor a 0.3 se redujo la cantidad de genes a 232 y por ultimo por medio del k-vecino más cercano se hizo la recuperación de valores perdidos dando lugar a un conjunto de datos con menos cantidad de variables y esto fue recibido como entrada en una red bayesiana que fue entrenada con el 70 % de los datos totales y medida con una curva ROC con el 30 %.

En el trabajo *Estudio comparativo de descriptores de textura para el desarrollo de un método computacional de segmentación automática de lesiones de mama en ultrasonografías* [10] se procesan las imágenes de las ultrasonografías de las mamas llevando a la obtención de las características que pueden servir de entrada en los clasificadores como el análisis discriminante lineal y en máquinas de soporte vectorial (SVM) que fueron evaluados con las métricas de calidad de precisión, sensibilidad y especificidad.

Para realizar el trabajo *Selección de Características usando Algoritmos Genéticos para Clasificación de Vinos Chilenos* [20] implementaron el algoritmo genético con el análisis discriminante lineal como función objetivo, proporcionándole los químicos que se hallan en los vinos que fueron elegidos para realizar las pruebas y al final se obtuvieron los químicos que aportaban la información más relevante para realizar la clasificación.

Con las técnicas implementadas en los diversos trabajos presentados en esta sección, se tomó en cuenta el LDA para ser implementado y evaluado, así como se hizo en el trabajo *Estudio comparativo de descriptores de textura para el desarrollo de un método computacional de segmentación automática de lesiones de mama en ultrasonografías* [10] y el otro clasificador que fue considerado para implementar y evaluar es el k-vecino más cercano, el cual también fue evaluado en el trabajo *Minería de datos aplicada a la detección de Cáncer de Mama* [7]; para la

selección de características se consideró implementar un algoritmo genético.

El motivo para elegir estas técnicas para seleccionar y evaluar las características es que son sencillos de implementar y el costo que implica implementarlos es bajo, además el algoritmo genético es capaz de funcionar sin la asistencia de un *experto humano*, es decir que de manera "automática" hace la selección de las características y dependiendo de la evaluación determina al conjunto de características que son mejores para clasificar.

Para realizar las pruebas se emplea la base de datos del cáncer de mama en Wisconsin [5], ya que contiene casos reales de pacientes con cáncer de mama. También fue empleada la base de datos de libros del Mago de Oz que fue utilizada como referencia para comparar los resultados de las técnicas implementadas, la comparación se hizo con las palabras que se seleccionaron en el trabajo de José Nilo G. Binongo [6].

Capítulo 2

Marco teórico

En este capítulo se definen los conceptos que son utilizados durante el desarrollo del documento. Para poder entender de lo que trata el problema primero es necesario conocer lo que es la clasificación, mencionada en la sección 2.1, unos de los criterios de clasificación que utilizan los clasificadores que se emplean en este trabajo es el de clasificación supervisada mencionado en la sección 2.2, incluyendo los clasificadores supervisados que han sido utilizados para resolver problemas con datos multidimensionales pero en este trabajo solo se implementaron dos de ellos. El problema presentado es binario y por lo tanto es necesario conocer acerca de los clasificadores binarios y estos se describen en la sección 2.3; estos clasificadores serán evaluados por medio de una medida de calidad explicada en la sección 2.4 y esta medida de calidad será recibida por un algoritmo de optimización descrito en la sección 2.5, los algoritmos de optimización son muy efectivos en este tipo de problemas, por esa razón se implementó un algoritmo genético para hacer la selección de las mejores características.

2.1. Clasificación

La clasificación es una forma de sintetizar la información contenida en una tabla multidimensional y es mediante la conformación y caracterización de grupos o clases.

Los grupos o clases se conforman de manera que los elementos dentro de cada grupo sean lo más homogéneos posibles y que, en cambio, los elementos de diferentes grupos sean lo más diferentes posibles; estos grupos serán medidos por las mismas características [11].

Hay distintos criterios de clasificación, entre ellos [12]:

Clasificación supervisada: La clasificación supervisada consiste en tener conocimiento a priori de un conjunto de clases.

Clasificación no supervisada: No existe conocimiento a priori de las clases y se determinan mediante un procedimiento estadístico proporcionándole un número de clases deseadas.

2.2. Clasificación supervisada

El tipo de clasificación utilizada en este trabajo es la clasificación supervisada y en las siguientes subsecciones se describen algunos de los métodos de este tipo. Las descripciones de estos son simples y cortas ya que no todos son implementados en este trabajo.

2.2.1. Máquina de soporte vectorial

Una máquina de soporte vectorial es una técnica que mapea el conjunto de datos recibidos y selecciona el vector que separe estos datos en clases por el máximo margen m [13].

2.2.2. Árbol de decisión

Un árbol de decisión es un modelo de predicción que recibe un conjunto de *r* datos de entrada $\{E_1, E_2...E_r\}$ llamados raíz de los cuales se derivan *d* cantidad de decisiones llamadas ramas; estas decisiones son tomadas en base a valores de probabilidad considerados en cada decisión a tomar; y las decisiones finales son llamadas hojas [14].

2.2.3. Redes bayesianas

Una red bayesiana es un modelo de grafos acíclico que representa a un conjunto de *k* variables $\{ V_1, V_2...V_k \}$ conocidos como nodos y sus respectivas dependencias entre sí, conocidas como aristas [15].

2.2.4. Análisis de componentes principales (PCA)

El análisis de componentes principales es una técnica cuyo objetivo es la reducción considerable de características o variables del caso de estudio. Esto es posible por medio de la combinación de las características o variables originales y a esto se le llama componente; los componentes con valores más altos (por lo regular son los primeros componentes) guardan información más relevante de las variables o características originales y bajo este criterio se hace la reducción de la dimensión de los datos [16].

2.2.5. Análisis discriminante lineal (LDA)

Esta es una técnica en la que gráficamente puede observarse una línea que separa a dos clases o conjuntos de datos A y B tomando en cuenta que dicha línea de separación es un eigenvector que fue obtenido de una matriz de covarianza de los datos en conjunto; pueden tomarse en consideración 1 o más eigenvectores sin dejar de lado que los primeros obtenidos son los que proporcionan un mejor criterio de separación de las clases [17]. Este método es descrito de manera detallada en el capítulo 4.

2.2.6. Método del k-vecino más cercano

Este método consiste en tener un vecindario de puntos sobre un plano n-dimensional, en el cual se define un punto de referencia y de acuerdo a este punto se localizan a los k-puntos con los que hay una menor distancia de separación [18]. Este método es descrito de manera detallada en el capítulo 4.

2.2.7. Metodología boosting

Esta técnica consiste en obtener un clasificador a partir de las mejores cualidades de m clasificadores. Básicamente se combinan los m clasificadores con sus respectivos pesos para así producir un nuevo clasificador esperando que sea igual o mejor que los m clasificadores. Dicho proceso se lleva a cabo por medio de la ecuación 2.1.

$$NV = \Sigma_{j=1}^{n} \Sigma_{i=1}^{m} w_i C_j e_i \tag{2.1}$$

Donde:

$NV =$ Nuevo vector clasificador.

$m =$ Número de clasificadores seleccionados

$w_i =$ Peso correspondiente al clasificador i.

$n =$ Número de casos de la base de datos.

$C_j =$ Caso j de la base de datos.

$e_i =$ Clasificador i.

2.2.7.1. Ecuación de peso

Existen diferentes formas para la obtención del peso, pero para este caso se eligió hacerlo como se muestra en la ecuación 2.2.

$$w_i = \frac{p_i}{\Sigma_{j=1}^{m} p_j} \tag{2.2}$$

Donde

$w_i =$ Peso.

$p_i =$ Precisión del clasificador i

$m =$ Número de clasificadores seleccionados.

$p_j =$ Precisión que corresponde al clasificador j.

Esta metodología se implementó con los resultados del clasificador de la sección 2.2.5; ya que para este clasificador cada eigenvector obtenido hace una clasificación diferente y en este trabajo se tomaron en cuenta los 5 primeros eigenvectores.

Fueron implementados 3 de los clasificadores mencionados: el análisis discriminante lineal y el k-vecino más cercano, utilizados como clasificadores binarios. Se trabajó con estos clasificadores en específico porque son sencillos algorítmicamente, rápidos de programar o implementar, poco costosos computacionalmente y son de los más conocidos y utilizados.

2.3. Clasificador binario

Como ya se ha mencionado, el problema presentado es binario y por esto es importante conocer la definición de un clasificador binario.

Un clasificador binario es un clasificador que basa su respuesta en solo dos clases; es decir, se tiene una serie de datos representados por individuos n-dimensionales, cada individuo pertenece a una clase de entre dos opciones y pueden ser representadas por las etiquetas 1 y -1.

La clasificación consiste en que por medio de una función se determine la clase a la cual pertenece un individuo de entrada. Esta función es la siguiente: f : **X** $\subseteq$ **R**^{n} $\rightarrow$ **R**. Para esta función se tiene la entrada de un individuo: x^i=$\{x_1^i, x_2^i, ..., x_n^i\}$ al cual se le asigna la etiqueta 1 si $f(x^i) \geq 0$ ó -1 en caso contrario [19].

2.4. Medidas de calidad para clasificadores

Las medidas tomadas en este trabajo son: *Sensibilidad*, *especificidad* y *precisión*; éstos son conceptos utilizados para verificar la consistencia de pruebas médicas realizadas a los pacientes y dependen de los siguientes términos: Verdadero Positivos (**TP** por sus siglas en inglés), Falso Positivo (**FP**), Verdadero Negativo (**TN** por sus siglas en inglés) y Falso Negativo (F**N**).

Hay un TP cuando un paciente tiene una enfermedad y la prueba realizada indica la presencia de esta. Igualmente si el paciente no tiene la enfermedad y la prueba realizada lo confirma, entonces hay un TN. Si la prueba que se le realiza a un paciente indica la presencia de la enfermedad cuando este no la padece se tiene que el resultado es FP. Igualmente si la prueba realizada indica que no existe la enfermedad cuando el paciente si la padece se tiene que el resultado es FN. Estos términos están sujetos a la consistencia entre la prueba realizada y el diagnóstico que tiene el paciente y son definidos en la tabla 2.1.

	Estado respecto a la enfermedad según el estándar de referencia		
Resultado de la prueba en estudio	Positivo	Negativo	Suma total de cada diagnóstico
Positivo	**TP**	**FP**	**TP**+**FP** (Total de casos con diagnóstico positivo)
Negativo	**FN**	**TN**	**FN**+**TN** (Total de casos con diagnóstico negativo)
Total de resultado de la prueba	**TP**+**FN** (Total de casos con esta condición)	**FP**+**TN** (Total de casos con esta condición)	**N**=**TP**+**TN**+**FP**+**FN** (Número total de casos estudiados)

Tabla 2.1: Términos que definen la sensibilidad, especificidad y precisión.

El valor numérico de sensibilidad representa la probabilidad de una prueba de diagnóstico para identificar pacientes que en efecto tienen una enfermedad. Cuanto mayor sea el valor numérico de la sensibilidad, es menos probable que la prueba devuelva resultados falsos positivos. Se calcula de la siguiente manera:

$$Sensibilidad = \frac{TP}{(TP+FN)} = \frac{(Verdaderos\_Positivos)}{(Total\_de\_Positivos)}$$

El valor numérico de especificidad representa la probabilidad de una prueba diagnóstico de que no dará resultados falsos positivos de una enfermedad y se obtiene de la siguiente manera:

$$Especificidad = \frac{TN}{(TN+FP)} = \frac{(Verdaderos\_Negativos)}{(Total\_de\_Negativos)}$$

Una prueba puede ser muy específica sin ser sensible, o puede ser muy sensible sin ser específica. Ambos factores son igualmente importantes. Una buena prueba es una que tiene tanto una alta sensibilidad y como especificidad.

El valor numérico de la precisión representa la proporción de los resultados acertados en la población [21]. La precisión se obtiene de la siguiente manera:

$$\textit{Precisión} = \frac{(TN+TP)}{(TN+TP+FN+FP)} = \frac{(Resultados\_Correctos)}{(Total\_de\_Resultados)}$$

Estos conceptos son utilizados con la finalidad de conocer la eficiencia de una prueba y el valor máximo que pueden tomar 1. Ahora, para conocer si un clasificador es eficiente, este es entrenado, probado y posteriormente evaluado por medio de estas medidas, especialmente por medio de la precisión; estas medidas son utilizadas para identificar al conjunto de características que hacen que la clasificación realizada sea eficiente.

De estas 3 medidas de calidad se le da más importancia a la precisión, porque sin importar el valor numérico de la sensibilidad y especificidad, ésta determina la calidad de una prueba dado que su cálculo involucra los resultados acertados sobre el total de las pruebas.

2.5. Optimización

De acuerdo a lo dicho en *Algoritmos evolutivos y algoritmos genéticos* [22]: Un problema como el que se aborda en este trabajo es un problema de optimización, ya que lo que se busca es hallar un conjunto de caracteríticas con el que se optimice cierto criterio de calidad, es decir, maximizando o minimizando una función objetivo $f(x)$ dada.

Los algoritmos evolutivos son especialmente útiles para atacar problemas definidos en variables binarias, no derivables, y que no existe un método eficiente para encontrar una buena aproximación al óptimo. Además no requieren de conocer la expresión de la función a optimizar ni asumen un único óptimo.

2.5.1. Algoritmo genético

Los algoritmos genéticos son un tipo de algoritmo evolutivo. Los algoritmos evolutivos trabajan con una población de individuos, que representan soluciones candidatas a un problema. Esta población se somete a la selección de mejores soluciones y sobre las soluciones candidatas

que resultan de esta selección se aplican ciertas operaciones para generar nuevas soluciones candidatas. Cada ciclo de selección y variación constituye una generación, de forma que, después de cierto número de generaciones se espera que el mejor individuo de la población esté cerca de la solución óptima. Los algoritmos evolutivos combinan la búsqueda aleatoria, dada por las transformaciones de la población, con una búsqueda dirigida dada por la selección.

Un algoritmo genético trabaja con una población de cadenas binarias para la representación del problema, dicha población es sometida a una serie de variaciones para encontrar las soluciones óptimas; estas variaciones son tal y como se observan en los organismos vivientes: cruce y mutación. Existen diferentes métodos para realizar la selección: el método de la ruleta, es decir, que es como si se tuviera una ruleta y ésta fuera dividida en porciones de acuerdo a una medida de aptitud de cada individuo de la población, y el método de selecćon elitista, es decir, que se seleccionan a los mejores individuos para pasar sus genes a la siguiente generación [22], siendo este útimo el utilizado en este trabajo. El algoritmo es descrito en el capítulo 5.

Capítulo 3

Definición del Problema

En este capítulo se describe el problema que se aborda en este trabajo y la estrategia que se desarrolla para solucionarlo.

El problema de la selección de características consiste en seleccionar un subconjunto de m características de entre un conjunto original de n características, donde $m < n$ [24].

Entre los propósitos de la selección de características se cuentan:

- Reducir la complejidad del clasificador y su implementación en hardware/software.
- Compresión de información (eliminar características redundantes e irrelevantes).
- Reducir el costo de medición al disminuir el número de características.
- Proveer una mejor clasificación debido a efectos por tamaño finito de la muestra.

Se tiene un conjunto de datos con una alta dimensionalidad del que se requiere obtener dos clases para posteriormente hacer futuras clasificaciones. La alta dimensionalidad de los datos puede ocasionar que existan características que aporten información poco relevante o redundante, lo que lleva a que las clasificaciones realizadas con estos datos sean poco confiables.

Entonces, es necesario llevar a cabo la selección de las características que aportan información relevante y no redundante para hacer clasificaciones que sean confiables. Por lo tanto, la estrategia será implementar un algoritmo genético con un conjunto de clasificadores sencillos para la obtención de la función objetivo, de estos clasificadores se obtendrá una medida de calidad que reflejará la eficiencia de estos; dicha medida será de utilidad para definir el subconjunto de m características que proporcione una mejor clasificación, esta medida será empleada como la función objetivo que recibirá el algoritmo genético.

Específicamente, el algoritmo genético tomará un conjunto de características de alguna de las bases de datos, le proporcionará dichas características a cualquiera de los clasificadores y este hará su labor de clasificación, devolviendo un valor de precisión. De entre todas las precisiones proporcionadas en una generación se elegirá a la mayor y así sucesivamente hasta que las poblaciones generadas por el algoritmo tengan poca variación. Elegir la precisión de mayor valor significa que posiblemente el subconjunto de características con el que se obtuvo dicha precisión son las que pueden aportar información más importante para que la clasificación sea eficiente.

Para dicha implementación de la estrategia se tomarán como objetos de prueba dos bases de datos: la base de datos de cáncer de mama en Wisconsin y la base de datos de Libros del Mago de Oz, las dos son utilizadas para realizar las pruebas y determinar un subconjunto de características con el que se logre una precisión alta, pero la segunda además es utilizada con el fin de comprobar que lo realizado en este trabajo es eficiente, ya que es utilizada en diferentes trabajos de selección de características. Estas bases de datos son descritas completamente en la sección 4.1. Se implementarán dos clasificadores sencillos: Análisis discriminante lineal y el método del k-vecino más cercano, descritos en el capítulo 4. Cada clasificador será probado con cada una de las bases de datos.

Capítulo 4

Aplicación de clasificadores sencillos a dos bases de datos

En la sección 4.1 este capítulo se describen las bases de datos utilizadas como objeto de prueba, en la sección 4.2 se describe a detalle la técnica del análisis discriminante lineal, su implementación y los experimentos realizados, en la sección 4.3 se describe la metodología boosting aplicada a clasificadores LDA y en la sección 4.4 se describe detalladamente el método del k-vecino más cercano, su implementación y los experimentos realizados con este.

4.1. Descripción de las bases de datos empleadas

En esta sección están descritas las bases de datos que son utilizadas como objeto de prueba para los clasificadores. Se describe el origen de estas, sus principales características y los datos que fueron utilizados para realizar los experimentos.

4.1.1. Base de datos de Cáncer de Mama en Wisconsin

Esta base de datos fue creada en 1989 por el Dr. WIlliam H. Wolberg de University of Wisconsin Hospitals Madison, Wisconsin, USA y contiene 699 casos; con 9 características, correspondientes a observaciones hechas a imágenes de tumores, más dos atributos correspondientes a la identificación del caso y a la clasificación del tumor (benigno o maligno). De esta sección en adelante se hace referencia a esta base de datos como *BD Diagnósticos*. Sus principales caracterśticas son:

- Se miden 9 características.
- Tiene 2 clases:

  Clase 1 o clase que contiene los diagnósticos benignos, contiene 458 casos.

  Clase 2 o clase que contiene los diagnósticos malignos, contiene 241 casos.

Los datos utilizados son:

- Se utilizan las 9 características.
- De las clases se eliminaron algunos casos, dado que en la base de datos original existen casos con algunas características que no contienen valores numéricos, en su lugar hay NA y si se le asignara algún valor numérico sería causa de ruido en los datos ya que por alguna razón no fueron colocados en el caso, por este motivo son eliminados de la base de datos original, quedando 683 de los 699 casos iniciales.

  Clase 1 Inicialmente cuenta con 458 casos pero por la eliminación quedan 444.

  Clase 2 Inicialmente cuenta con 241 casos pero por la eliminación quedan 239.

- Para el conjunto de entrenamiento se utilizó el 70 % del total los datos.
- Para el conjunto de prueba se utilizó el 30 % del total de los datos.

4.1.2. Base de datos de Libros

Esta base de datos contiene 198 bloques, es decir registros; con 20,217 palabras o características; correspondientes a todas las palabras diferentes encontradas en todos los libros; cada bloque fue obtenido del análisis hecho a cada 5 000 palabras que estaban en los libros de R. Plumly Thompson y de L. Frank Baum creador de los libros del Mago de Oz, más las palabras de un libro que corresponde al mundo del Mago de Oz del cual se duda su autoría, ya que este fue editado y publicado por Thompson después de la muerte de Baum. De esta sección en adelante se hace referencia a esta base de datos como BD Libros. Sus principales características son:

- Se miden 20,217 palabras.
- Tiene 2 clases:

  Clase 1 o clase que contiene los libros de Baum, contiene 93 bloques de palabras medidas.

  Clase 2 o clase que contiene los libros de Thompson, contiene 105 bloques de palabras medidas.
- El libro del que se duda su procedencia contiene 7 bloques o registros de palabras medidas. Cada bloque será tomado como un nuevo elemento a clasificar.

Los datos utilizados son:

- Por limitaciones del hardware se llevó a cabo una reducción inicial de características a un total de 100, quedando las palabras que tuvieron mayor frecuencia relativa.
- Para el conjunto de entrenamiento se utilizó el 70 % del total los datos.
- Como conjunto de prueba se utilizó el 30 % del total de los datos.
- Finalmente para corroborar que las evaluaciones sean correctas se utilizan los bloques del libro del que se duda su autoría.

4.2. Técnica de análisis discriminante lineal

Esta sección esta dividida en dos subsecciones. En la subsección 4.2.1 se describe la técnica del análisis discriminante lineal y en la subsección 4.2.2 se explica con detalle el proceso seguido para realizar los experimentos con esta técnica.

4.2.1. Descripción

De acuerdo a lo descrito en *Linear Discriminant Analysis* [17], el análisis discriminante lineal es una técnica inferencial, típicamente multivariante porque suele usarse en contextos donde se tienen varias variables, pero evidentemente puede aplicarse con pocas variables, incluso con una sola variable, aunque no es lo habitual.

Una característica esencial de esta técnica es que se definen previamente dos o más poblaciones o clases; que están separadas por vectores que permiten una clara identificación entre individuos que pertenecen a una u otra clase.

El objetivo básico de esta técnica es preparar la información, seleccionarla y trabajarla con una finalidad clasificadora, haciendo que entre las distintas clases exista la mayor diferencia posible y entre los datos de una misma clase existan las mayores similitudes posibles.

De *Linear Discriminant Analysis* [17] se toma la técnica del análisis discriminante lineal aplicada a dos conjuntos de datos, como se muestra a continuación:

1. Se tienen dos conjuntos:

   C_1, es un conjunto de datos almacenados en una matriz de $m_1 x n$, conocido como clase; donde m_1 es la cantidad de datos y n es el número de características o variables medidas.

C_2, es un conjunto de datos almacenados en una matriz de m_2xn, conocido como clase; donde m_2 es la cantidad de datos y n es el número de características o variables medidas.

2. Se obtiene la media de cada una de las clases. Llamándose $\mu 1$ y $\mu 2$, siendo de dimension $1xn$; correspondientes a C_1 y C_2, respectivamente.

3. Se obtiene una tercera media que se define de la suma de la multiplicación de cada media por la probabilidad a priori (P); quedando como resultado un vector $\mu 3$ con una dimensión de $1xn$; expresado en la ecucación 4.1.

$$\mu 3 = (P_1 \cdot \mu 1) + (P_2 \cdot \mu 2) \tag{4.1}$$

   Para este caso $P_1 = P_2 = 0{,}5$.

4. Se obtiene una matriz de covarianzas que será llamada Sw; expresado en la ecuación 4.2.

$$Sw = (P_1 \cdot cov(C_1)) + (P_2 \cdot cov(C_2)) \tag{4.2}$$

5. Se calcula la dispersión entre los conjuntos, con la ecuación 4.3.

$$Sb = \Sigma_{j=1}^{2}(\mu j - \mu 3) \cdot T(\mu j - \mu 3) \tag{4.3}$$

6. Se obtiene una matriz de criterios de clasificación entre los conjuntos; con la ecuación 4.4.

$$criterio = inv(Sw) \cdot Sb \tag{4.4}$$

7. Se obtienen los eigenvectores de la matriz $criterio$, para este trabajo se utilizó el método de la potencia. Estos eigenvectores servirán para tener criterios de separación entre los conjuntos de datos.

Para hacer la clasificación de un nuevo individuo, se requiere que contenga las mismas variables medidas en los conjuntos de entrada, posteriormente, se normaliza para tener datos de acuerdo a los existentes. Al realizar la normalización del nuevo individuo sus valores serán en su mayoría $>$ ó $<$ que 0. De esta manera será posible asignar una etiqueta correspondiente a cualquiera de las clases a las que correspondan sus características.

Para saber a cual clase pertenece este individuo a clasificar, es necesario verificar con cual de las dos clases hay más similitud; es decir, si la función objetivo del individuo > 0 entonces se le asigna la etiqueta 1 y pero si el resultado de la función es < 0, entonces se le asigna la etiqueta -1. Las etiquetas de las clases fueron asignadas por medio del entrenamiento.

4.2.2. Uso de LDA para clasificación sin selección de características

Para este método, el experimento consiste en entrenar el clasificador con el conjunto de entrenamiento y posteriormente, evaluar sus resultados utilizando el conjunto de prueba, calculando las medidas de calidad.

El clasificador LDA utiliza el mejor eigenvector para realizar la separación entre las dos clases que conforman a la base de datos, pero para este trabajo se utilizaron los 5 vectores asociados a los eigen vectores de mayor valor, el motivo por el cual se obtienen los 5 es porque serán de utilidad para implementar la metodología boosting en la sección 4.3. Que halla 5 vectores asociados a los mejores eigenvectores quiere decir que se tienen 5 diferentes clasificadores, por lo tanto se tienen 5 diferentes valores para las medidas de calidad.

En las siguientes subsecciones se muestran medidas de calidad obtenidas mediante cálculos hechos en lenguaje C y gráficas elaboradas en R.

En general, el proceso a seguir en este experimento es el siguiente:

1. Obtener las medidas de calidad de los 5 vectores asociados a los eigenvectores de mayor valor.

2. Demostrar gráficamente la eficiencia de los vectores.

4.2.2.1. LDA aplicado a la Base de datos de cáncer de mama en Wisconsin

La tabla 4.1 muestra las medidas de calidad correspondientes a las clasificaciones hechas con los 5 vectores asociados a los eigenvectores de mayor valor.

Medidas de calidad					
	e1	**e2**	**e3**	**e4**	**e5**
Sensibilidad	0.8450704	0.8965517	0.7286822	0.7685185	0.9185185
Especificidad	0.7936508	0.9500000	0.4868421	0.4845361	0.8714286
Precisión	**0.8292683**	**0.9121951**	0.6390244	0.6341463	**0.9024390**

Tabla 4.1: Medidas de calidad. LDA con BD Diagnósticos.

Se observa que la mejor medida de precisión que corresponde al vector 2, seguido por los vectores 1 y 5; en la figura 4.1 se muestran los datos proyectados sobre los mejores vectores de clasificación, demostrándose que los vectores 1, 2 y 5 si hacen clasificaciones eficientes.

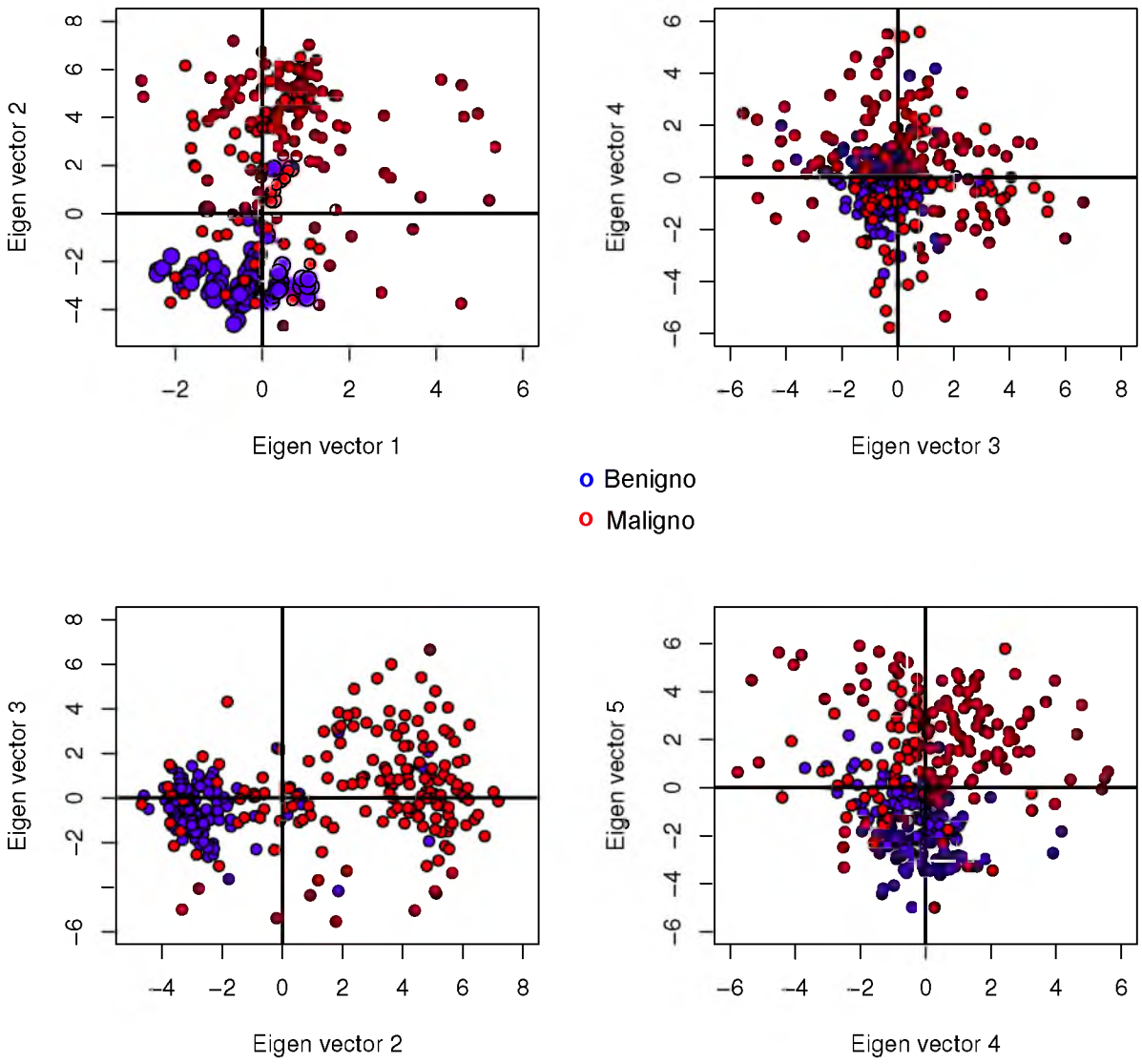


Figura 4.1: Gráfica de datos proyectados. BD Diagnósticos.

4.2.2.2. LDA aplicado a la Base de datos de Libros

En la tabla 4.2 se muestran los valores de las medidas de calidad correspondientes a las clasificaciones hechas con los 5 vectores asociados a los eigenvectores de mayor valor.

Medidas de calidad					
	e1	**e2**	**e3**	**e4**	**e5**
Sensibilidad	1.0	0.9583333	0.7419355	0.9259259	0.4814815
Especificidad	0.8648649	0.8611111	0.8275862	0.9090909	0.5454545
Precisión	**0.9166667**	**0.9000000**	0.7833333	**0.9166667**	0.5166667

Tabla 4.2: Medidas de calidad. LDA con BD Libros.

Se observa que la mejor medida de precisión que corresponde al vector 1, seguido por los vectores 2 y 4; en la figura 4.2 se muestran los datos proyectados sobre los mejores vectores de clasificación, demostrándose que los vectores 1, 2 y 4 si hacen clasificaciones eficientes.

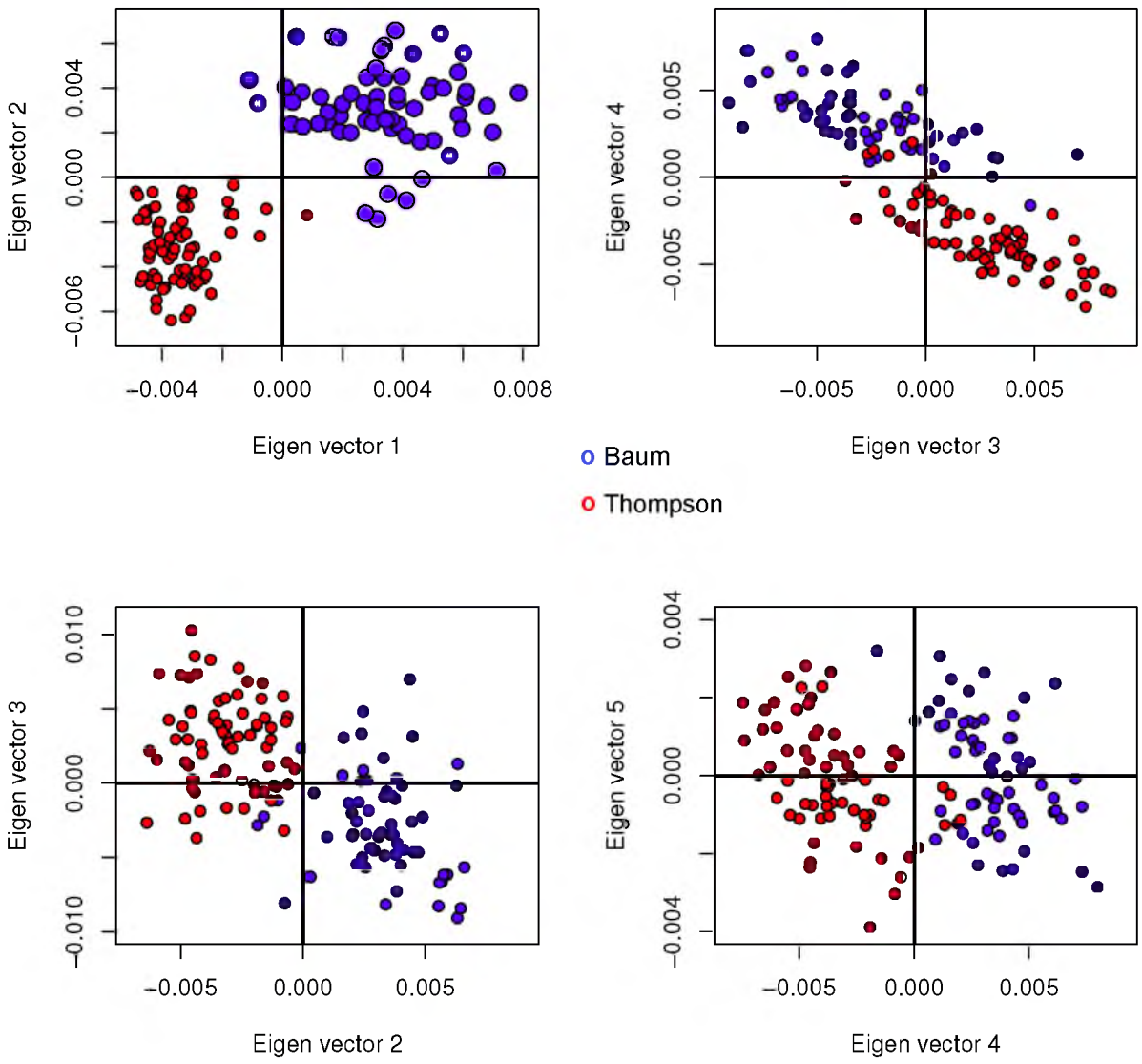


Figura 4.2: Gráfica de datos proyectados. BD Libros.

4.3. Metodología boosting aplicada a clasificadores LDA sin selección de características

Como caso preliminar se implementó una metodología boosting sencilla, con pesos sin entrenar. Para aplicar alguna de las metodologías boosting más usadas como por ejemplo la metodología AdaBoost falta entrenar los pesos, entre otras cosas.

Para la aplicación de esta metodología es necesario tener previamente clasificadores obtenidos con alguna técnica, la utilizada en este trabajo es la técnica del análisis discriminante lineal. Se tomaron los 3 mejores clasificadores, seleccionados de las tablas 4.1 y 4.2, para cada base de datos; posteriormente se pretende obtener un clasificador que reúna las mejores cualidades de estos.

En general, el proceso a seguir implementando esta metodología, sin importar la base de datos que sea objeto de prueba, es el siguiente:

1. Calcular el peso de cada clasificador seleccionado.
2. Obtener el nuevo clasificador mejorado.
3. Evaluar la eficiencia del nuevo clasificador y comparar si es mejor que los obtenidos en la sección 4.2.

4.3.1. Boosting aplicado a la Base de datos de Cáncer de Mama en Wisconsin

En la tabla 4.3 se muestran los valores de las 3 mejores medidas de calidad. Por medio de la precisión se determinó que los vectores 1, 2 y 5 son los mejores clasificadores.

Mejores medidas de calidad			
	e1	**e2**	**e5**
Sensibilidad	0.8450704	0.8965517	0.9185185
Especificidad	0.7936508	0.9500000	0.8714286
Precisión	**0.8292683**	**0.9121951**	**0.9024390**

Tabla 4.3: Mejores medidas de calidad. Boosting con BD Diagnósticos.

Para implementar la metodología boosting primero se calcula el peso de los 3 mejores clasificadores con la ecuación 2.2. Teniendo que p_1 es la precisión del vector 1, p_2 es la precisión del vector 2 porque es el segundo vector asociado al segundo mejor valor y p_3 es la precisión del vector 5 porque es el tercer vector asociado al tercer mejor valor. Los pesos se muestran en la tabla 4.4.

i	**Precisión (p_i)**	w_i
1	0.8292683	0.3136531
2	0.9121951	0.3450184
3	0.9024390	0.3413284

Tabla 4.4: Pesos. Boosting con BD Diagnósticos.

Una vez obtenido el peso de cada clasificador se implementó la ecuación 2.1. Con esta escuación se determina un nuevo clasificador con el que se calcularon nuevas medidas de calidad, mostradas en la tabla 4.5.

Nuevas medidas de calidad	
Sensibilidad	0.8965517
Especificidad	0.9500000
Precisión	0.9121951

Tabla 4.5: Nuevas medidas de calidad. Boosting con BD Diagnosticos.

4.3.2. Boosting aplicado a la Base de datos de Libros

En la tabla 4.6 se muestran los valores de las 3 mejores medidas de calidad. Se determinó por medio de la precisión que los vectores 1, 2 y 4 son los mejores clasificadores.

Mejores medidas de calidad			
	e1	**e2**	**e4**
Sensibilidad	1.0	0.9583333	0.9259259
Especificidad	0.8648649	0.8611111	0.9090909
Precisión	0.9166667	0.9000000	0.9166667

Tabla 4.6: Mejores medidas de calidad. Boosting con BD Libros.

Para implementar la metodología boosting primero se calcula el peso de los 3 mejores clasificadores con la ecuación 2.2. Teniendo que p_1 es la precisión del vector 1, p_2 es la precisión del vector 2 porque es el segundo vector asociado al segundo mejor valor y p_3 es la precisión del vector 4 porque es el tercer vector asociado al tercer mejor valor. Los pesos se muestran en la tabla 4.7.

i	**Precisión (p_i)**	w_i
1	0.9166667	0.3353659
2	0.9000000	0.3292683
3	0.9166667	0.3353659

Tabla 4.7: Pesos. Boosting con BD Libros.

Una vez obtenido el peso de cada clasificador se implementó la ecuación 2.1. Con esta ecuación se obtiene un nuevo clasificador con el que se calcularon nuevas medidas de calidad, mostradas en la tabla 4.8.

Nuevas medidas de calidad	
Sensibilidad	1.0
Especificidad	1.0
Precisión	1.0

Tabla 4.8: Nuevas medidas de calidad. Boosting con BD Libros.

4.4. Método del k-vecino más cercano

Es un método de clasificación en el que se mide la distancia entre los datos, que en este método son llamados vecinos. Al ingresar un nuevo dato al clasificador este se mide contra todos los datos pertenecientes a la base de datos, se toman a los k-vecinos con los que tiene las distancias más cortas, y por útimo se determina la clase de acuerdo al criterio de clasificación elegido. Para este trabajo se tomaron dos criterios distintos de clasificacion: 1) Asignar la clase del k-vecino más cercano y 2) Tomar la clase promedio de los k-vecinos más cercanos. Esto debido a que la distribución de los datos es desconocida y de una u otra manera es posible tener errores de clasificación y para esquivar y/o reducir estos errores se consideraron los criterios de clasificación mencionados.

La distancia utilizada en este trabajo es la distancia euclidiana. En matemáticas, la distancia euclidiana o euclídea es la distancia *ordinaria* (que se mediría con una regla) entre dos puntos de un espacio euclídeo, la cual se deduce a partir del teorema de Pitágoras.

En general, la distancia euclidiana entre los puntos $P = (p_1, p_2, ..., p_n)$ y $Q = (q_1, q_2, ..., q_n)$, del espacio euclídeo $n-$dimensional, se define como [18]:

$$dE(P,Q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\Sigma_{i=1}^{n}(pi - qi)^2}$$

4.4.1. Uso del k-vecino más cercano para clasificación sin selección de características

El experimento con este método y este criterio de clasificación consiste en introducir $k = 3$; lo que significa que se obtendrán los 3 vecinos con los que el nuevo punto tenga las menores distancias

y posteriormente, determinar la clase tomando en cuenta el k-vecino más cercano, es decir, el tercer vecino más cercano.

4.4.1.1. K-vecino más cercano aplicado a la base de datos de Cáncer de Mama en Wisconsin

Las medidas obtenidas son las que se muestran en la tabla 4.9.

Medidas de calidad	
Sensibilidad	0.984615
Especificidad	0.933333
Precisión	0.965854

Tabla 4.9: Medidas de calidad. K-vecino más cercano con DB Diagnósticos.

4.4.1.2. K-vecino más cercano aplicado a la base de datos de Libros

Las medidas obtenidas son las que se muestran en la tabla 4.10.

Medidas de calidad	
Sensibilidad	0.928571
Especificidad	0.937500
Precisión	0.933333

Tabla 4.10: Medidas de calidad. K-vecino más cercano con BD Libros.

4.4.2. Uso del promedio de los k-vecinos más cercanos para clasificación sin selección de características

El experimento con este método y este criterio de clasificacón consiste en introducir $k = 3$; lo que significa que se obtendrán los 3 vecinos con los que el punto a clasificar tenga las menores distancias y posteriormente determinar la etiqueta que será asignada al nuevo punto. Esta asignación se hace considerando la clase promedio de los 3 vecinos.

4.4.2.1. Promedio de los k-vecinos más cercanos aplicado a la base de datos de Cáncer de Mama en Wisconsin

Las medidas obtenidas son las que se muestran en la tabla 4.11.

Medidas de calidad	
Sensibilidad	0.942446
Especificidad	0.969697
Precisión	0.951220

Tabla 4.11: Medidas de calidad. Promedio de los k-vecinos más cercanos con BD Diagnósticos.

4.4.2.2. Promedio de los k-vecinos más cercanos aplicado a la base de datos de Libros

Las medidas obtenidas son las que se muestran en la tabla 4.12.

Medidas de calidad	
Sensibilidad	1.0
Especificidad	1.0
Precisión	1.0

Tabla 4.12: Medidas de calidad. Promedio de los k-vecinos más cercanos con BD Libros.

Capítulo 5

Algoritmo genético

El algoritmo genético tiene como función básica es emular la evolución de los organismos naturales; o sea, hacer que los mejores individuos sobrevivan en futuras generaciones.

Dentro del algoritmo genético se tiene una población, la cual está formada por individuos, de tales individuos solo pasarán a la siguiente generación los que proporcionen las mejores soluciones; para avanzar a la siguiente generación de individuos es necesario atravesar una selección de los mejores individuos, una cruza entre estos mejores y pasar por una mutación hecha al azar para tener más semejanza con la naturaleza y por último tomar al mejor individuo de la población y pasarlo a la siguiente generación para que así sus genes se propaguen.

Algorithm 1 Algoritmo genético.

1: $X^t \leftarrow$ generaPoblacion(nPob,nVar)
2: minOP $\leftarrow$ 1
3: **while** count $<$ 5 **do**
4: $F^t \leftarrow$ Eval(X^t)
5: $[X^t_{best}, f^t_{best}] \leftarrow$ Elite(X^t, F^t)
6: $S^t \leftarrow$ Select(F^t)
7: $X^{nt} \leftarrow$ Crossover(X^t, S^t)
8: $X^{-t} \leftarrow$ Mutation(X^{nt})
9: $X^{t+1} \leftarrow$ Replacement(X^{-t}, X_{best})
10: pi $\leftarrow \Sigma_{j=1}^{npob} X_{ij}$
11: op $\leftarrow \frac{4}{nvar}\Sigma_{i=1}^{nvar}(0{,}5 - pi)^2$
12: op $\leftarrow$ 1$-$op
13: **if** minOP$<$op **then**
14: minOP $\leftarrow$ op;
15: count $\leftarrow$ 0;
16: **else**
17: count++;
18: **end if**
19: **end while**

Descripción del algoritmo

1 Se genera una población donde $nPob$ es el número de individuos dentro de la población y $nVar$ es el número de características o genes que tiene esa población.

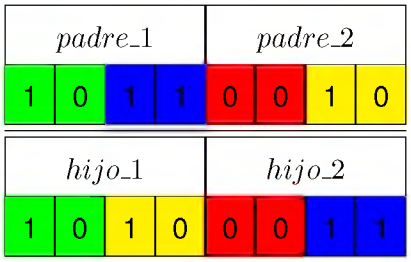
4 Eval es una función que evalúa a cada individuo de la población; dicha evaluación se hace sumando todas las características de cada individuo y se guardan en el vector F^t. Esta evaluación es conocida como la *función objetivo* y su labor será realizada de acuerdo a los criterios que se soliciten en el problema.

5 Elite es una funión que obtiene el mejor valor de F^t y lo guarda en f^t_{best} y tomando en cuenta la posición donde fue hallado el f^t_{best} se toma al individuo localizado en esa posición y se guarda en el vector X^t_{best}. Donde X^t_{best} tiene una dimensión de $nVar$.

6 Select es una función que ordena de mayor a menor los valores de función objetivo que contiene F^t, selecciona la mejor mitad del total de estos valores y obtiene la posición que tiene

el individuo dentro de la problación; S^t guarda en orden aleatorio las posiciones antes obtenidas, repitiendo cada valor de posición dos veces. Las dimensiones de S^t son las mismas que las de F^t; es decir $nPob$ es la dimensión de S^t.

7 Crossover es una función que hace la cruza entre padres para generar hijos, que más tarde conformarán una nueva población. Para elegir a los padres que generarán al nuevo par de hijos se toman de S^t las posiciones de par en par y se combina la primera parte de características del $padre\_1$ con la segunda parte de características del $padre\_2$ para generar al $hijo\_1$; para el $hijo\_2$ se combina la primera parte de características del $padre\_2$ con la segunda parte de características del $padre\_1$. El punto de partida en el que se combina el $padre\_1$ con el $padre\_2$ y viceversa es seleccionado de manera aleatoria. Ejemplo:



Cuando se ha finalizado la cruza, este par de hijos se insertan a la población sustituyendo a los padres.

8 Mutation es una función que cambia el valor de las características de los individuos; en este caso si es una población con valores binarios, se cambia 0 por 1 y viceversa; tomando en cuenta que $pmut$ sea mayor a un número generado aleatoriamente; donde $pmut = 1/nVar$.

9 Replacement es una funcion que toma el X_{best}; o sea el individuo que tuvo el mejor valor de función objetivo en toda la población y lo inserta en la población sustituyendo al primer individuo.

La condición de paro[**3**] de este algoritmo es que mientras $count < 5$ continuará iterando. Esta condición toma en cuenta la variación de la población; y si ésta no varía considerablemente ya no es necesario seguir produciendo nuevas generaciónes, ya que la función objetivo que se obtenga de ese punto en adelante no tendrá cambio significativo.

En la figura 5.1 se muestra gráficamente la el desarrollo del algoritmo genético con $nPob = 200$ y $nVar = 30$, maximizando la función objetivo.

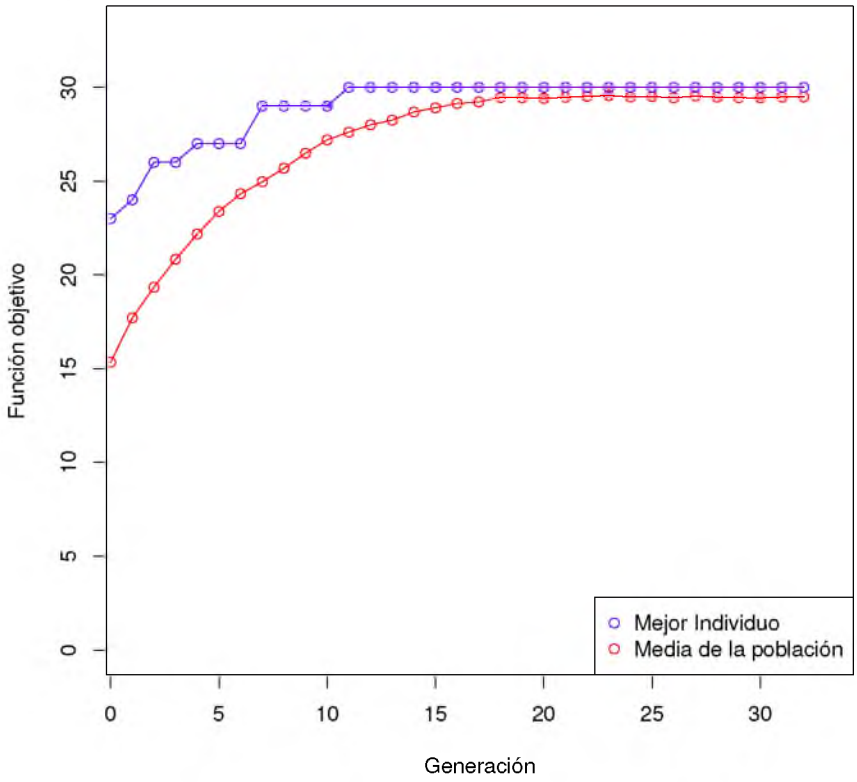


Figura 5.1: Evolución del algorítmo genético.

Capítulo 6

Algoritmo genético para la selección de características

La selección de características se hace mediante la implementación del algoritmo genético, utilizando como función objetivo la precisión obtenida de los clasificadores sencillos descritos en el capítulo 4.

En la sección 6.1 se describe la metodología implementada en cualquiera de los clasificadores sencillos que se utilizan en este trabajo y en las sección 6.2 se muestra la metodología implementada del algoritmo genético con los clasificadores sencillos.

6.1. Metodología para la obtención de la medida de calidad de clasificadores sencillos

La siguiente metodología que se presenta corresponde a la utilizada en la implementación de los clasificadores. Se describe el proceso a seguir desde que se inicia la lectura de datos hasta que se obtiene la medida de calidad que indica la eficiencia del clasificador.

6.1.1. Leer datos

Previo a la lectura, la base de datos que será objeto de prueba se encuentra dividida físicamente en dos partes; clase A y clase B. Teniendo esto, la lectura de datos se hace en dos pasos: 1) Leer clase A y 2) Leer clase B.

6.1.2. Preparar datos

Consiste en generar aleatoriamente los conjuntos de prueba y entrenamiento.

6.1.3. Clasificar

El clasificador seleccionado recibe los datos que fueron preparados previamente. Los clasificadores reciben parámetros distintos para realizar las clasificaciones. En las tablas 6.1, 6.2 y 6.3 se muestran los parámetros que recibe cada función de clasificación.

Parámetros de entrada del clasificador LDA:

double LDA(double CEntrenamientoCA, int nDatosCECA, double** CEntrenamientoCB, int nDatosCECB, double** CPruebaCA_CB, int nDatosCPCA_CB, int nCaracteristicas)**

Entrada	
double** CEntrenamientoCA	Matriz del conjunto de entrenamiento de la clase A
int nDatosCECA	Número de datos del conjunto de entrenamiento de la clase A
double** CEntrenamientoCB	Matriz del conjunto de entrenamiento de la clase B
int nDatosCECB	Número de datos del conjunto de entrenamiento de la clase B
double** CPruebaCA_CB	Matriz del conjunto de prueba
int nDatosCPCA_CB	Número de datos del conjunto de prueba
int nCaracteristicas	Número de características
Salida	
Precisión de la clasificación tipo double	
Al finalizar la clasificación crea un archivo en el que se almacenan las clasificaciones	

Tabla 6.1: Entrada y salida del clasificador LDA.

Parámetros de entrada del clasificador k-vecino más cercano:

double k_VMC(double CEntrenamientoCA, int nDatosCECA, double** CEntrenamientoCB, int nDatosCECB, double** CPruebaCA_CB, int nDatosCPCA_CB, int nCaracteristicas, int k)**

Entrada	
double** CEntrenamientoCA	Matriz del conjunto de entrenamiento de la clase A
int nDatosCECA	Número de datos del conjunto de entrenamiento de la clase A
double** CEntrenamientoCB	Matriz del conjunto de entrenamiento de la clase B
int nDatosCECB	Número de datos del conjunto de entrenamiento de la clase B
double** CPruebaCB_CB	Matriz del conjunto de prueba
int nDatosCPCB_CB	Número de datos del conjunto de prueba
int nCaracteristicas	Número de características
int k	El vecino del que se tomará la clase
Salida	
Precisión de la clasificación tipo double	
Al finalizar la clasificación crea un archivo en el que se almacenan las clasificaciones.	

Tabla 6.2: Entrada y salida del clasificador k-vecino más cercano.

Parámetros de entrada del clasificador promedio de los k-vecinos más cercanos:

double Promediok_VMC(double CEntrenamientoCA, int nDatosCECA , double** CEntrenamientoCB, int nDatosCECB, double** CPruebaCA_CB, int nDatosCPCA_CB, int nCaracteristicas, int k)**

Entrada	
double** CEntrenamientoCA	Matriz del conjunto de entrenamiento de la clase A
int nDatosCECA	Número de datos del conjunto de entrenamiento de la clase A
double** CEntrenamientoCB	Matriz del conjunto de entrenamiento de la clase B
int nDatosCECB	Número de datos del conjunto de entrenamiento de la clase B
double** CPruebaCA_CB	Matriz del conjunto de prueba
intn nDatosCPCA_CB	Número de datos del conjunto de prueba
int nCaracteristicas	Número de características
int k	Cantidad de vecinos de los que se obtendrá la clase promedio
Salida	
Precisión de la clasificación tipo double	
Al finalizar la clasificación crea un archivo en el que se almacenan las clasificaciones.	

Tabla 6.3: Entrada y salida del clasificador promedio de los k-vecinos más cercanos.

El archivo de clasificación creado por los clasificadores tiene el siguiente formato:

Dato	Clase de origen	Clase de clasificación
1	1	1
2	1	1
3	1	-1
4	1	1
⋮	⋮	⋮
nDatosCPCA_CB-3	-1	-1
nDatosCPCA_CB-2	-1	-1
nDatosCPCA_CB-1	-1	1
nDatosCPCA_CB	-1	-1

Tabla 6.4: Formato del archivo de clasificación.

La clase A siendo representada por 1 y la clase B por -1.

6.1.4. Obtener medidas de calidad

Se lee el archivo de clasificación y para cada dato se verifica si la clase de clasificación coincide con la clase de origen. Si la clase de origen es A y fue clasificado como A, entonces se declara TP; si la clase de origen es A y fue clasificado como B, entonces se declara FN; si la clase de origen es B y fue clasificado como A, entonces se declara FP, si la clase de origen es B y fue clasificado como B, entonces se declara TN.

De acuerdo a las cantidades de TP, FP, TN y FN se obtienen la sensibilidad, especificidad y precisión siendo ésta última de mayor importancia porque informa la calidad del clasificador.

6.2. Metodología del algoritmo genético para la selección de características

La siguiente metodología que se presenta corresponde a la utilizada en la implementación del algoritmo genético con clasificadores sencillos. Se describe el proceso a seguir desde que se inicia

la lectura de datos hasta que se obtiene el vector de características que maximizan la precisión.

6.2.1. Leer datos

Previo a la lectura, la base de datos que será objeto de prueba se encuentra dividida físicamente en dos partes; clase A y clase B. Teniendo esto, la lectura de datos se hace en dos pasos: 1) Leer clase A, 2) Leer clase B.

6.2.2. Preparar datos

Consiste en generar aleatoriamente los conjuntos de prueba y entrenamiento.

6.2.3. Definir clasificador

Se determina el clasificador que será utilizado para obtener el valor de función objetivo. Los parámetros de entrada del algoritmo genético dependerán del clasificador seleccionado, tomando en cuenta lo que se muestra en las tablas 6.1, 6.2 y 6.3.

6.2.4. Seleccionar características con el algoritmo genético

El algoritmo recibe la base de datos, las medidas de la población, y los parámetros adicionales para los clasificadores. El algoritmo genético implementa parte de la metodología de los clasificadores sencillos, dentro del cálculo de la función objetivo se utiliza la clasificación de la sección 6.1.3 y la obtención de la medida de calidad de la sección 6.1.4.

Los parámetros mostrados a continuación son los que recibe la función del algoritmo genético con el LDA y con los clasificadores del método del k-vecino más cercano.

Parámetros de entrada del algoritmo genético con el LDA:

int *algoritmo_genetico(int poblacion, int nPob, int nVar, double** CEntrenamientoCA, int**

nDatosCECA, double CEntrenamientoCB,int nDatosCECB, double** CPruebaCA_CB, int nDatosCPCA_CB, int nDatosCPCA, int nDatosCPCB)**

Entrada	
int** poblacion	Matriz de población binaria, generada aleatoriamente de dimensiones $nPob \times Nvar$
int nPob	Representa el número de individuos de la población, para este trabajo se utilizó $nPob = 200$ porque dentro de esa cantidad es posible hallar varias soluciones candidatas
int nVar	Representa el número de características que tiene cada base de datos, es decir, nVar tuvo dos valores distintos, $nVar = 9$ y $nVar = 100$
double** CEntrenamientoCA	Matriz del conjunto de entrenamiento de la clase A
int nDatosCECA	Número de datos del conjunto de entrenamiento de la clase A
double** CEntrenamientoCB	Matriz del conjunto de entrenamiento de la clase B
int nDatosCECB	Número de datos del conjunto de entrenamiento de la clase B
double** CPruebaCA_CB	Matriz del conjunto de prueba
int nDatosCPCA_CB	Número de datos del conjunto de prueba
int nDatosCPCA	Número de datos pertenecientes al conjunto de prueba que corresponden a la clase A
int nDatosCPCB	Número de datos pertenecientes al conjunto de prueba que corresponden a la clase B
Salida	
Vector de características que optimizan la precisión	

Tabla 6.5: Entrada y salida del algoritmo genético con el LDA.

Parámetros de entrada del algoritmo genético con cualquiera de las dos variaciones del método del k-vecino más cercano utilizadas en este trabajo:

int *algoritmo_genetico(int poblacion, int nPob, int nVar, double** CEntrenamientoCA, int nDatosCECA, double** CEntrenamientoCB,int nDatosCECB, double** CPruebaCA_CB, int nDatosCPCA_CB, int nDatosCPCA, int nDatosCPCB, int k)**

Entrada	
int** poblacion	Matriz de población binaria, generada aleatoriamente de dimensiones $nPob \times Nvar$
int nPob	Representa el número de individuos de la población, para este trabajo se utilizó $nPob = 200$ porque dentro de esa cantidad es posible hallar varias soluciones candidatas
int nVar	Representa el número de características que tiene cada base de datos, es decir, nVar tuvo dos valores distintos, $nVar = 9$ y $nVar = 100$
double** CEntrenamientoCA	Matriz del conjunto de entrenamiento de la clase A
int nDatosCECA	Número de datos del conjunto de entrenamiento de la clase A
double** CEntrenamientoCB	Matriz del conjunto de entrenamiento de la clase B
int nDatosCECB	Número de datos del conjunto de entrenamiento de la clase B
double** CPruebaCA_CB	Matriz del conjunto de prueba
int nDatosCPCA_CB	Número de datos del conjunto de prueba
int nDatosCPCA	Número de datos pertenecientes al conjunto de prueba que corresponden a la clase A
int nDatosCPCB	Número de datos pertenecientes al conjunto de prueba que corresponden a la clase B
int k	Para definir los k-vecinos más cercanos
Salida	
Vector de características que optimizan la precisión	

Tabla 6.6: Entrada y salida del algoritmo genético con el método del k-vecino más cercano.

Capítulo 7

Experimentos y resultados

Los resultados de cada prueba se muestran en dos tablas en las que se agregó el total de las características de la BD Diagnósticos y las palabras de la BD Libros. Para cada prueba la forma en la que se representan los resultados obtenidos es la siguiente: con $\surd$ se indica que la característica o palabra fue seleccionada y con χ que la característica o palabra no fue seleccionada.

En las tablas de resultados de la BD Libros se muestran en negritas las palabras que se encuentran en la figura 7.1. Las palabras de la figura fueron seleccionadas en el trabajo de José Nilo G. Binongo [6] y son utilizadas como referencia para ser comparadas con las obtenidas en las pruebas realizadas en este trabajo. Cuando en las tablas de la BD Libros aparece una palabra en negritas y también con $\surd$, quiere decir que esta palabra fue seleccionada tanto en este trabajo como en el de José Nilo G. Binongo.

En las siguientes secciones de este capítulo se muestran los resultados obtenidos al realizar cada una de las pruebas descritas a continuación. Son 3 pruebas distintas, en cada prueba se evalúa el algoritmo genético con cada uno de los 3 clasificadores; en total son 9 evaluaciones por cada base de datos.

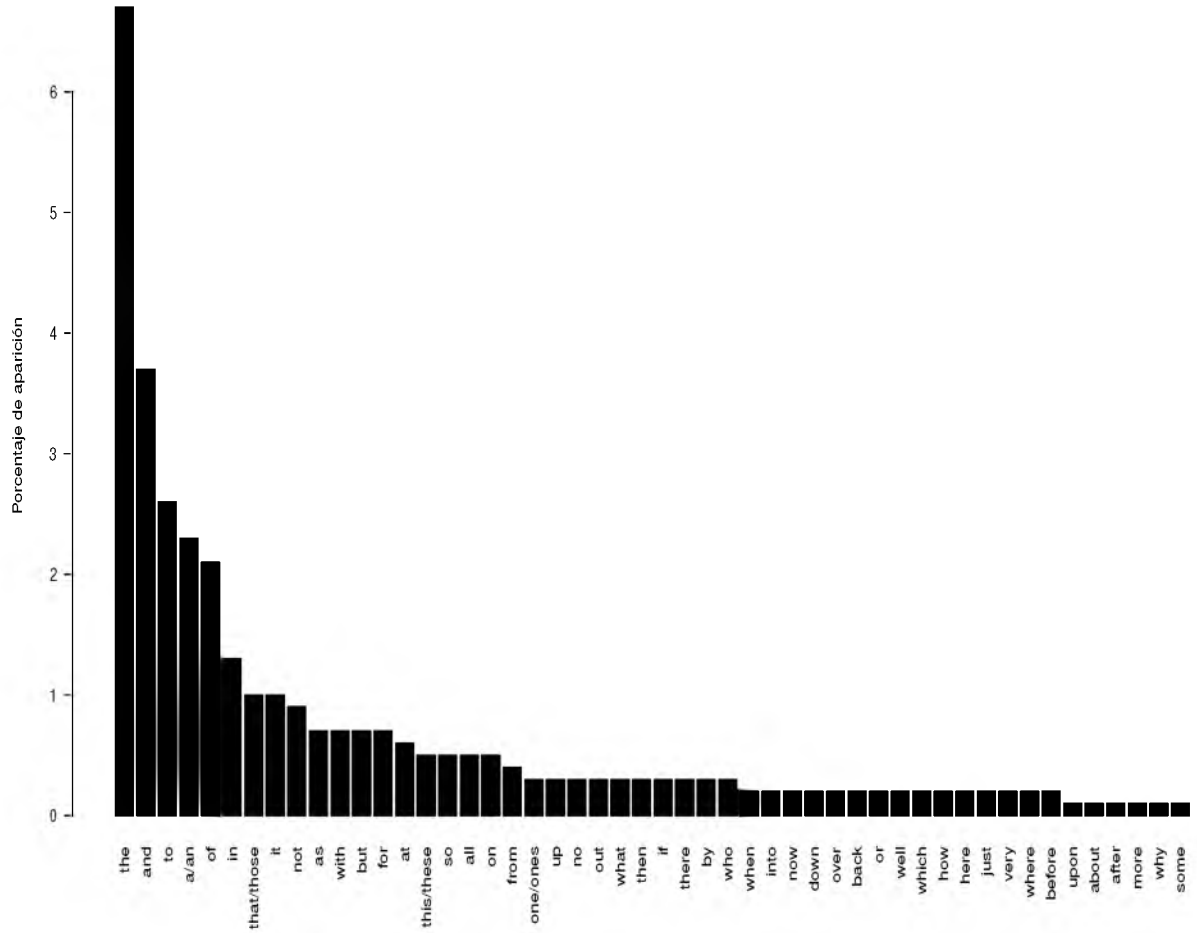


Figura 7.1: 50 palabras obtenidas del artículo de José Nilo G. Binongo.

Las pruebas realizadas son las siguientes:

Prueba 1 Consiste en la ejecución del algoritmo genético una vez, con conjuntos iniciales de prueba y entrenamiento. Seleccionar conjuntos iniciales quiere decir que los conjuntos de prueba y entrenamiento se seleccionaron antes de mandar a llamar a la función del algoritmo genético. El objetivo de esta prueba es verificar si el algoritmo genético es capaz de encontrar un subconjunto de características con el que se logre una precisión alta.

Prueba 2 Consiste en la ejecución del algoritmo genético 30 veces, con conjuntos distintos de prueba y entrenamiento para cada individuo. Cada individuo de la población del algoritmo genético calcula su función objetivo con un conjunto distinto de prueba y de entrenamiento. El objetivo de esta prueba es reducir la posible aparición de un sesgo que trae consigo la utilización del algoritmo genético; es posible que el sesgo aparezca ya que este algoritmo tiene una base estocástica, esto porque las soluciones que toma son puestas al azar y no siempre es posible llegar a la óptima en una sóla ejecución.

Prueba 3 Consiste en la ejecución del algoritmo genético una vez, promediando la función objetivo obtenida 15 veces con conjuntos distintos de prueba y entrenamiento para cada individuo. El objetivo de esta prueba es observar el funcionamiento promedio del algoritmo genético.

La selección del 70 % de los datos para el conjunto de entrenamiento y del 30 % de los datos para el conjunto de prueba se realiza en varias ocasiones. El motivo es que existen datos que van fuera de lo común, que pueden salir un poco del contexto en comparación con los demás datos y que en cierto punto hacen la diferencia al realizar las pruebas; estos datos son conocidos como datos atípicos. La forma en la que se reduce la diferencia que marcan estos datos es tomarlos de manera aleatoria, algunas veces como parte del conjunto de entrenamiento y otras como parte del conjunto de prueba.

7.1. Prueba 1, algoritmo genético con conjuntos iniciales de prueba y entrenamiento.

En esta sección se muestran los resultados que fueron obtenidos con los clasificadores llevando a cabo la prueba 1.

7.1.1. Resultados del LDA para la selección de características en la prueba 1

Los resultados obtenidos en la prueba 1 utilizando el algoritmo genético con el LDA son los que se muestran en la sección 7.1.1.1 y la sección 7.1.1.2.

7.1.1.1. LDA aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 1

La mejor precisión obtenida es 0.9512195 utilizando 4 características de 9, las cuales se muestran en la tabla 7.1.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
χ	$\surd$	χ	$\surd$	χ	$\surd$	χ	χ	$\surd$

Tabla 7.1: Características seleccionadas en la prueba 1. LDA con BD diagnósticos.

7.1.1.2. LDA aplicado a la base de datos de Libros en la prueba 1

La mejor precisión obtenida es 1.0 utilizando 40 palabras de 100, las cuales se muestran en la tabla 7.2, de estas, 10 se encuentran en la figura 7.1.

The √	**and** χ	his χ	was χ	you χ	**that** √	**with** χ
had √	**for** √	they χ	**but** χ	her √	**all** √	**Not** χ
this √	she √	said χ	**from** χ	were χ	him √	them χ
have χ	little χ	**one** χ	**out** √	are √	Dorothy χ	Their √
into χ	**When** χ	**Then** χ	could √	**who** χ	**down** χ	King χ
will χ	would χ	**over** χ	your χ	**There** √	**back** χ	**now** χ
what χ	been √	like √	head √	Can √	Ozma √	see √
which √	asked χ	time √	man √	**upon** χ	**about** χ	did χ
after χ	**before** χ	way χ	**more** χ	Wizard χ	**just** χ	magic √
very χ	great √	Scarecrow √	old √	made √	good χ	long √
girl χ	**some** χ	boy √	**where** χ	himself χ	must χ	through √
any √	other χ	than χ	City χ	came χ	**How** √	off χ
know √	eyes √	our χ	**here** √	never χ	only χ	While √
get χ	away √	began χ	around √	first χ	its χ	**it** χ
come √	Peter χ					

Tabla 7.2: Palabras seleccionadas en la prueba 1. LDA con BD Libros.

7.1.2. Resultados del k-vecino más cercano para la selección de características en la prueba 1

Los resultados obtenidos en la prueba 1 utilizando el algoritmo genético con el k-vecino más cercano son los que se muestran en la sección 7.1.2.1 y la sección 7.1.2.2.

7.1.2.1. K-vecino más cercano aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 1

La mejor precisión obtenida es 0.9804878 utilizando 6 características de 9, las cuales se muestran en la tabla 7.3.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
√	χ	χ	√	χ	√	√	√	√

Tabla 7.3: Características seleccionadas en la prueba 1. K-vecino más cercano con BD diagnósticos.

7.1.2.2. K-vecino más cercano aplicado a la base de datos de Libros en la prueba 1

La mejor precisión obtenida es 1.0 utilizando 47 palabras de 100, las cuales se muestran en la tabla 7.4, de estas, 17 se encuentran en la figura 7.1.

The $\surd$	**and** $\surd$	his $\surd$	was $\surd$	you $\surd$	**that** χ	**with** $\surd$
had $\surd$	**for** χ	they χ	**but** χ	her χ	**all** $\surd$	**Not** χ
this χ	she χ	said $\surd$	**from** χ	were χ	him χ	them $\surd$
have χ	little χ	**one** $\surd$	**out** $\surd$	are $\surd$	Dorothy χ	Their $\surd$
into $\surd$	**When** $\surd$	**Then** χ	could $\surd$	**who** $\surd$	**down** $\surd$	King χ
will χ	would χ	**over** χ	your $\surd$	**There** χ	**back** χ	**now** χ
what χ	been $\surd$	like χ	head χ	Can $\surd$	Ozma $\surd$	see χ
which χ	asked χ	time $\surd$	man χ	**upon** χ	**about** χ	did $\surd$
after $\surd$	**before** $\surd$	way χ	**more** $\surd$	Wizard χ	**just** $\surd$	magic χ
very $\surd$	great χ	Scarecrow χ	old χ	made χ	good χ	long $\surd$
girl χ	**some** χ	boy $\surd$	**where** χ	himself χ	must χ	through χ
any $\surd$	other $\surd$	than χ	City $\surd$	came $\surd$	**How** $\surd$	off χ
know $\surd$	eyes χ	our $\surd$	**here** $\surd$	never $\surd$	only χ	While χ
get $\surd$	away $\surd$	began χ	around $\surd$	first $\surd$	its χ	**it** $\surd$
come χ	Peter $\surd$					

Tabla 7.4: Palabras seleccionadas en la prueba 1. K-vecino más cercano con BD Libros.

7.1.3. Resultados del promedio de los k-vecinos más cercanos para la selección de características en la prueba 1

Los resultados obtenidos en la prueba 1 utilizando el algoritmo genético con el promedio de los k-vecinos más cercanos son los que se muestran en la sección 7.1.3.1 y la sección 7.1.3.2.

7.1.3.1. Promedio de los k-vecinos más cercanos aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 1

La mejor precisión obtenida es 0.9804878 utilizando 5 características de 9, las cuales se muestran en la tabla 7.5.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
$\surd$	χ	$\surd$	χ	χ	$\surd$	$\surd$	$\surd$	χ

Tabla 7.5: Características seleccionadas en la prueba 1. Promedio de los k-vecinos más cercanos con BD diagnósticos.

7.1.3.2. Promedio de los k-vecinos más cercanos aplicado a la base de datos de Libros en la prueba 1

La mejor precisión obtenida es 1.0 utilizando 54 palabras de 100, las cuales se muestran en la tabla 7.6, de estas, 21 se encuentran en la figura 7.1.

The $\surd$	**and** $\surd$	his $\surd$	was $\surd$	you χ	**that** χ	**with** $\surd$
had χ	**for** χ	they χ	**but** $\surd$	her χ	**all** $\surd$	**Not** χ
this $\surd$	she χ	said $\surd$	**from** χ	were χ	him χ	them $\surd$
have χ	little $\surd$	**one** $\surd$	**out** χ	are $\surd$	Dorothy χ	Their $\surd$
into χ	**When** $\surd$	**Then** $\surd$	could χ	**who** $\surd$	**down** χ	King $\surd$
will $\surd$	would χ	**over** $\surd$	your $\surd$	**There** χ	**back** $\surd$	**now** χ
what χ	been χ	like $\surd$	head χ	Can χ	Ozma χ	see χ
which $\surd$	asked χ	time $\surd$	man $\surd$	**upon** $\surd$	**about** χ	did $\surd$
after $\surd$	**before** $\surd$	way χ	**more** $\surd$	Wizard χ	**just** $\surd$	magic $\surd$
very $\surd$	great $\surd$	Scarecrow $\surd$	old χ	made $\surd$	good χ	long $\surd$
girl $\surd$	**some** $\surd$	boy χ	**where** χ	himself $\surd$	must $\surd$	through χ
any χ	other $\surd$	than χ	City $\surd$	came χ	**How** $\surd$	off χ
know $\surd$	eyes χ	our $\surd$	**here** χ	never $\surd$	only χ	While χ
get $\surd$	away $\surd$	began $\surd$	around $\surd$	first $\surd$	its χ	**it** $\surd$
come χ	Peter χ					

Tabla 7.6: Palabras seleccionadas en la prueba 1. Promedio de los k-vecinos más cercanos con BD Libros.

7.2. Prueba 2, algoritmo genético ejecutado 30 veces con conjuntos distintos de prueba y entrenamiento para cada individuo.

En esta sección se muestran los resultados que fueron obtenidos con los clasificadores llevando a cabo la prueba 2.

7.2.1. Resultados del LDA para la selección de características en la prueba 2

Los resultados obtenidos en la prueba 2 utilizando el algoritmo genético con el LDA son los que se muestran en la sección 7.2.1.1 y la sección 7.2.1.2.

7.2.1.1. LDA aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 2

La mejor precisión obtenida es 0.9414634 utilizando 5 características de 9, las cuales se muestran en la tabla 7.7.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
χ	$\surd$	χ	$\surd$	χ	$\surd$	$\surd$	$\surd$	χ

Tabla 7.7: Características seleccionadas en la prueba 2. LDA con BD diagnósticos.

7.2.1.2. LDA aplicado a la base de datos de Libros en la prueba 2

La mejor precisión obtenida es 1.0 utilizando 55 palabras de 100, las cuales se muestran en la tabla 7.8, de estas, 18 se encuentran en la figura 7.1.

The χ	**and** χ	his $\surd$	was $\surd$	you $\surd$	**that** χ	**with** χ
had $\surd$	**for** χ	they χ	**but** χ	her χ	**all** $\surd$	**Not** $\surd$
this χ	she χ	said χ	**from** $\surd$	were χ	him $\surd$	them χ
have $\surd$	little $\surd$	**one** χ	**out** χ	are $\surd$	Dorothy $\surd$	Their $\surd$
into χ	**When** $\surd$	**Then** χ	could χ	**who** $\surd$	**down** $\surd$	King χ
will $\surd$	would $\surd$	**over** χ	your χ	**There** $\surd$	**back** χ	**now** χ
what $\surd$	been $\surd$	like $\surd$	head χ	Can $\surd$	Ozma $\surd$	see $\surd$
which $\surd$	asked χ	time $\surd$	man $\surd$	**upon** $\surd$	**about** $\surd$	did $\surd$
after $\surd$	**before** χ	way $\surd$	**more** $\surd$	Wizard $\surd$	**just** $\surd$	magic $\surd$
very $\surd$	great χ	Scarecrow $\surd$	old χ	made $\surd$	good $\surd$	long χ
girl χ	**some** $\surd$	boy χ	**where** $\surd$	himself $\surd$	must $\surd$	through $\surd$
any χ	other χ	than $\surd$	City χ	came χ	**How** χ	off $\surd$
know $\surd$	eyes χ	our χ	**here** χ	never χ	only $\surd$	While $\surd$
get χ	away χ	began $\surd$	around $\surd$	first χ	its χ	**it** $\surd$
come $\surd$	Peter χ					

Tabla 7.8: Palabras seleccionadas en la prueba 2. LDA con BD Libros.

7.2.2. Resultados del k-vecino más cercano para la selección de características en la prueba 2

Los resultados obtenidos en la prueba 2 utilizando el algoritmo genético con el k-vecino más cercano son los que se muestran en la sección 7.2.2.1 y la sección 7.2.2.2.

7.2.2.1. K-vecino más cercano aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 2

La mejor precisión obtenida es 0.9853659 utilizando 5 características de 9, las cuales se muestran en la tabla 7.9.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
χ	χ	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$	χ	χ

Tabla 7.9: Características seleccionadas en la prueba 2. K-vecino más cercano con BD diagnósticos.

7.2.2.2. K-vecino más cercano aplicado a la base de datos de Libros en la prueba 2

La mejor precisión obtenida es 1.0 utilizando 39 palabras de 100, las cuales se muestran en la tabla 7.10, de estas, 14 se encuentran en la figura 7.1.

The $\surd$	**and** $\surd$	his χ	was $\surd$	you $\surd$	**that** χ	**with** χ
had χ	**for** χ	they $\surd$	**but** $\surd$	her χ	**all** $\surd$	**Not** χ
this χ	she $\surd$	said $\surd$	**from** $\surd$	were $\surd$	him χ	them $\surd$
have χ	little $\surd$	**one** $\surd$	**out** χ	are χ	Dorothy χ	Their $\surd$
into $\surd$	**When** χ	**Then** χ	could χ	**who** $\surd$	**down** χ	King χ
will $\surd$	would $\surd$	**over** χ	your $\surd$	**There** $\surd$	**back** $\surd$	**now** χ
what χ	been χ	like χ	head χ	Can χ	Ozma $\surd$	see χ
which $\surd$	asked $\surd$	time $\surd$	man χ	**upon** χ	**about** χ	did χ
after χ	**before** χ	way $\surd$	**more** χ	Wizard χ	**just** $\surd$	magic χ
very χ	great $\surd$	Scarecrow χ	old $\surd$	made χ	good χ	long χ
girl χ	**some** χ	boy χ	**where** χ	himself χ	must χ	through χ
any χ	other $\surd$	than χ	City χ	came χ	**How** $\surd$	off χ
know χ	eyes $\surd$	our χ	**here** $\surd$	never χ	only $\surd$	While $\surd$
get χ	away χ	began $\surd$	around χ	first $\surd$	its χ	**it** $\surd$
come χ	Peter χ					

Tabla 7.10: Palabras seleccionadas en la prueba 2. K-vecino más cercano con BD Libros.

7.2.3. Resultados del promedio de los k-vecinos más cercanos para la selección de características en la prueba 2

Los resultados obtenidos en la prueba 2 utilizando el algoritmo genético con el promedio de los k-vecinos más cercanos son los que se muestran en la sección 7.2.3.1 y la sección 7.2.3.2.

7.2.3.1. Promedio de los k-vecinos más cercanos aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 2

La mejor precisión obtenida es 0.9707317 utilizando 6 características de 9, las cuales se muestran en la tabla 7.11.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
$\surd$	$\surd$	χ	$\surd$	$\surd$	$\surd$	$\surd$	χ	χ

Tabla 7.11: Características seleccionadas en la prueba 2. Promedio de los k-vecinos más cercanos con BD diagnósticos.

7.2.3.2. Promedio de los k-vecinos más cercanos aplicado a la base de datos de Libros en la prueba 2

La mejor precisión obtenida es 1.0 utilizando 46 palabras de 100, las cuales se muestran en la tabla 7.12, de estas, 17 se encuentran en la figura 7.1.

The χ	**and** χ	his $\surd$	was $\surd$	you χ	**that** χ	**with** $\surd$
had χ	**for** $\surd$	they $\surd$	**but** $\surd$	her χ	**all** χ	**Not** χ
this χ	she χ	said $\surd$	**from** $\surd$	were $\surd$	him χ	them $\surd$
have $\surd$	little $\surd$	**one** χ	**out** χ	are χ	Dorothy $\surd$	Their $\surd$
into χ	**When** χ	**Then** χ	could χ	**who** $\surd$	**down** $\surd$	King $\surd$
will $\surd$	would $\surd$	**over** χ	your $\surd$	**There** $\surd$	**back** $\surd$	**now** $\surd$
what $\surd$	been $\surd$	like $\surd$	head $\surd$	Can χ	Ozma χ	see χ
which $\surd$	asked χ	time χ	man χ	**upon** $\surd$	**about** χ	did $\surd$
after $\surd$	**before** $\surd$	way χ	**more** $\surd$	Wizard χ	**just** χ	magic $\surd$
very $\surd$	great χ	Scarecrow χ	old $\surd$	made χ	good χ	long χ
girl χ	**some** χ	boy $\surd$	**where** χ	himself $\surd$	must χ	through $\surd$
any $\surd$	other χ	than χ	City χ	came χ	**How** χ	off $\surd$
know $\surd$	eyes χ	our χ	**here** χ	never $\surd$	only χ	While $\surd$
get $\surd$	away χ	began χ	around χ	first χ	its χ	**it** $\surd$
come χ	Peter χ					

Tabla 7.12: Palabras seleccionadas en la prueba 2. Promedio de los k-vecinos con BD Libros.

7.3. Prueba 3, algoritmo genético promediando la función objetivo obtenida 15 veces con conjuntos distintos de prueba y entrenamiento para cada individuo.

En esta sección se muestran los resultados que fueron obtenidos con los clasificadores llevando a cabo la prueba 3.

7.3.1. Resultados del LDA para la selección de características en la prueba 3

Los resultados obtenidos en la prueba 3 utilizando el algoritmo genético con el LDA son los que se muestran en la sección 7.3.1.1 y la sección 7.3.1.2.

7.3.1.1. LDA aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 3

La mejor precisión obtenida es 0.9447154 utilizando 3 características de 9, las cuales se muestran en la tabla 7.13.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
χ	$\surd$	χ	$\surd$	χ	$\surd$	χ	χ	χ

Tabla 7.13: Características seleccionadas en la prueba 3. LDA con BD diagnósticos.

7.3.1.2. LDA aplicado a la base de datos de Libros en la prueba 3

La mejor precisión obtenida es 0.9944444 utilizando 45 palabras de 100, las cuales se muestran en la tabla 7.14, de estas, 11 se encuentran en la figura 7.1.

The √	**and** χ	his χ	was √	you √	**that** χ	**with** χ
had √	**for** χ	they χ	**but** χ	her χ	**all** χ	**Not** χ
this √	she χ	said χ	**from** √	were √	him χ	them χ
have √	little χ	**one** χ	**out** √	are √	Dorothy χ	Their χ
into √	**When** χ	**Then** √	could χ	**who** χ	**down** χ	King χ
will √	would χ	**over** χ	your √	**There** χ	**back** √	**now** χ
what χ	been χ	like √	head χ	Can √	Ozma √	see √
which √	asked √	time χ	man χ	**upon** χ	**about** χ	did √
after χ	**before** √	way √	**more** √	Wizard χ	**just** √	magic √
very χ	great √	Scarecrow √	old √	made χ	good √	long χ
girl √	**some** χ	boy √	**where** χ	himself √	must χ	through √
any χ	other χ	than √	City √	came √	**How** √	off χ
know √	eyes χ	our √	**here** √	never √	only √	While χ
get χ	away χ	began χ	around χ	first χ	its √	**it** χ
come χ	Peter χ					

Tabla 7.14: Palabras seleccionadas en la prueba 3. LDA con BD Libros.

7.3.2. Resultados del k-vecino más cercano para la selección de características en la prueba 3

Los resultados obtenidos en la prueba 3 utilizando el algoritmo genético con el k-vecino más cercano son los que se muestran en la sección 7.3.2.1 y la sección 7.3.2.2.

7.3.2.1. K-vecino más cercano aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 3

La mejor precisión obtenida es 0.9691057 utilizando 8 características de 9, las cuales se muestran en la tabla 7.15.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
√	√	√	√	χ	√	√	√	√

Tabla 7.15: Características seleccionadas en la prueba 3. K-vecino más cercano con BD diagnósticos.

7.3.2.2. K-vecino más cercano aplicado a la base de datos de Libros en la prueba 3

La mejor precisión obtenida es 0.9988889 utilizando 52 palabras de 100, las cuales se muestran en la tabla 7.16, de estas, 21 se encuentran en la figura 7.1.

The $\surd$	**and** $\surd$	his $\surd$	was $\surd$	you χ	**that** $\surd$	**with** χ
had $\surd$	**for** χ	they χ	**but** χ	her χ	**all** χ	**Not** $\surd$
this χ	she χ	said $\surd$	**from** χ	were $\surd$	him $\surd$	them $\surd$
have χ	little χ	**one** $\surd$	**out** $\surd$	are $\surd$	Dorothy χ	Their χ
into χ	**When** $\surd$	**Then** χ	could $\surd$	**who** χ	**down** $\surd$	King $\surd$
will χ	would χ	**over** $\surd$	your χ	**There** χ	**back** $\surd$	**now** χ
what $\surd$	been $\surd$	like $\surd$	head $\surd$	Can χ	Ozma χ	see $\surd$
which $\surd$	asked $\surd$	time χ	man χ	**upon** $\surd$	**about** $\surd$	did $\surd$
after $\surd$	**before** $\surd$	way $\surd$	**more** $\surd$	Wizard $\surd$	**just** $\surd$	magic χ
very $\surd$	great χ	Scarecrow $\surd$	old χ	made $\surd$	good $\surd$	long χ
girl χ	**some** χ	boy $\surd$	**where** $\surd$	himself χ	must $\surd$	through $\surd$
any χ	other χ	than $\surd$	City $\surd$	came χ	**How** χ	off χ
know $\surd$	eyes $\surd$	our χ	**here** χ	never $\surd$	only χ	While χ
get χ	away $\surd$	began χ	around χ	first χ	its χ	**it** $\surd$
come $\surd$	Peter χ					

Tabla 7.16: Palabras seleccionadas en la prueba 3. K-vecino más cercano con BD Libros.

7.3.3. Resultados del promedio de los k-vecinos más cercanos para la selección de características en la prueba 3

Los resultados obtenidos en la prueba 3 utilizando el algoritmo genético con el promedio de los k-vecinos más cercanos son los que se muestran en la sección 7.3.3.1 y la sección 7.3.3.2.

7.3.3.1. Promedio de los k-vecinos más cercanos aplicado a la base de datos de cáncer de mama de Wisconsin en la prueba 3

La mejor precisión obtenida es 0.9762602 utilizando 6 características de 9, las cuales se muestran en la tabla 7.17.

Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Ashesion	Single Epithelial Cell Size	Bar Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses
$\surd$	χ	$\surd$	χ	$\surd$	$\surd$	$\surd$	$\surd$	χ

Tabla 7.17: Características seleccionadas en la prueba 3. Promedio de los k-vecinos más cercanos con BD diagnósticos.

7.3.3.2. Promedio de los k-vecinos más cercanos aplicado a la base de datos de Libros en la prueba 3

La mejor precisión obtenida es 1.0 utilizando 55 palabras de 100, las cuales se muestran en la tabla 7.18, de estas, 19 se encuentran en la gráfica 7.1.

The $\surd$	**and** χ	his $\surd$	was χ	you $\surd$	**that** $\surd$	**with** $\surd$
had χ	**for** χ	they $\surd$	**but** χ	her χ	**all** χ	**Not** $\surd$
this $\surd$	she χ	said $\surd$	**from** χ	were $\surd$	him χ	them χ
have $\surd$	little χ	**one** $\surd$	**out** χ	are χ	Dorothy $\surd$	Their $\surd$
into χ	**When** $\surd$	**Then** χ	could $\surd$	**who** χ	**down** $\surd$	King $\surd$
will $\surd$	would χ	**over** $\surd$	your $\surd$	**There** $\surd$	**back** χ	**now** $\surd$
what $\surd$	been χ	like χ	head χ	Can χ	Ozma $\surd$	see $\surd$
which $\surd$	asked $\surd$	time $\surd$	man $\surd$	**upon** $\surd$	**about** χ	did $\surd$
after $\surd$	**before** χ	way χ	**more** $\surd$	Wizard $\surd$	**just** χ	magic χ
very $\surd$	great $\surd$	Scarecrow χ	old $\surd$	made χ	good $\surd$	long χ
girl $\surd$	**some** $\surd$	boy χ	**where** χ	himself $\surd$	must χ	through $\surd$
any $\surd$	other χ	than χ	City $\surd$	came $\surd$	**How** $\surd$	off χ
know $\surd$	eyes χ	our $\surd$	**here** χ	never χ	only $\surd$	While χ
get $\surd$	away χ	began χ	around $\surd$	first $\surd$	its χ	**it** χ
come $\surd$	Peter $\surd$					

Tabla 7.18: Palabras seleccionadas en la prueba 3. Promedio de los k-vecinos más cercanos con BD Libros.

7.4. Resumen de resultados

Con la BD Diagnósticos, de acuerdo a los resultados obtenidos representados en la figura 7.2 se observa que las características Bar Nuclei, Marginal Ashesion y Bland Chromatin son las que se seleccionan en la mayoría de las pruebas realizadas. La posible razón de esto es que estas características tienen mayor impacto en las clasificaciones con alta precisión y las que se seleccionan en menos pruebas pueden no tener tanto impacto para generar una precisión alta.

Utilizando las 9 características se obtuvo una precisión=0.9121951 con el LDA, 0.9121951 con la metodología Boosting, 0.9658541 con el k-vecino más cercano y 0.9512200 con el promedio de los k-vecinos más cercanos. Haciendo la selección de características la mejor precisión obtenida es 0.9853659 con 5 características, utilizando como evaluador al clasificador k-vecino más cercano en la prueba 2, obteniendo las siguientes características: Uniformity of Cell Shape, Marginal Ashesion, Single Epithelial Cell Size, Bar Nuclei y Bland Chromatin; 3 de estas 5 características fueron seleccionadas en la mayoría de las pruebas, observar la figura 7.2.

En los resultados de las pruebas hechas con la BD Libros se observa que las palabras que fueron seleccionadas en la mayoría de las pruebas son: **the**, **which**, **after**, **very**, know, his, said, like, time, was, **more**, are, Ozma, **just**, **how**, their, did, through, **it**. De estas palabras, 8 de ellas se encuentran en negritas, esto significa que estas 8 palabras se hallan en la figura 7.1, es decir, también fueron seleccionadas en el trabajo hecho por José Nilo G. Binongo [6].

Puede ser que el motivo de que estas palabras sean seleccionadas en la mayoría de las pruebas y que algunas de ellas también lo hallan sido en un trabajo aparte es que posiblemente aporten la información más relevante al momento de clasificar.

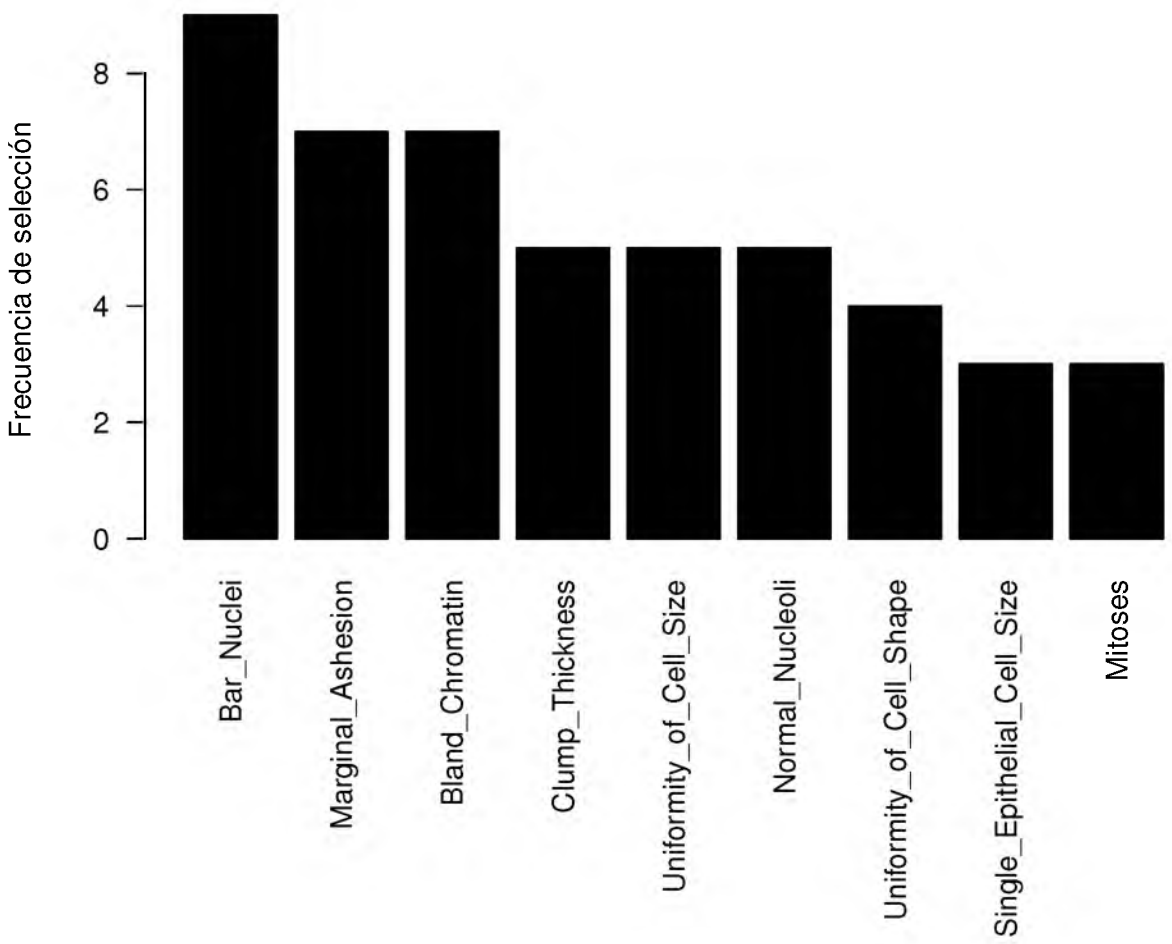


Figura 7.2: Frecuencia de selección de las características de la BD Diagnósticos.

Utilizando las 100 palabras se obtuvo una precisión=0.9166667 con el LDA, 1.0 con la metodología Boosting, 0.9333333 con el k-vecino más cercano, 1.0 con el promedio de los k-vecinos más cercanos. La mejor precisión obtenida con la selección de características fue 1.0, obtenida en 7 de las 9 pruebas realizadas, viendo esto se tomó en consideración la cantidad de palabras seleccionadas, por lo que se determinó que el menor subconjunto de características sería considerado como el mejor. Las palabras de este subconjunto elegido son 39, obtenidas en la prueba 2 con el k-vecino más cercano: **the**, into, will, **which**, **and**, she, little, would, asked, great, other, eyes, they, said, **one**, time, began, was, **but**, **from**, your, old, **here**, you, were, **who**, there, first, **all**, **back**, Ozma, **just**, **how**, only, them, **it**, while y their. De las cuales 10 aparecen entre las palabras que fueron seleccionadas en la mayoría de las pruebas; 13 de estas 39 aparecen en la gráfica 7.1. 5 de las 39 palabras fueron seleccionadas en la mayoría de las pruebas y también fueron seleccionadas en el trabajo de José Nilo G. Binongo: The, which, just, how, it.

En el trabajo de José Nilo G. Binongo las palabras fueron seleccionadas con la intervención de un *experto humano*, introduciendo un posible sesgo ocasionado porque el humano en base a su experiencia pudo haber seleccionado algunas palabras muy comunes o que el creyó que estaban bien; en cambio el algoritmo genético realizó esta selección tomando en cuenta las palabras que proporcionaron una precisión alta, es posible ver que algunas palabras si coinciden en los dos trabajos.

Como dato adicional de los resultados de esta base de datos se obtuvo que en todas las pruebas realizadas el libro de procedencia desconocida pertenece al autor R. Plumly Thompson.

Capítulo 8

Conclusiones y Trabajo a futuro

8.1. Conclusiones

Con los resultados vistos se deduce que el clasificador que funciona mejor para este trabajo es el método del vecino más cercano, pero eso no quiere decir que el LDA no funciona bien, es importante resaltar que una precision mayor a 0.9 es algo bueno pero es posible mejorarla.

Con la metodología boosting fue posible mejorar la precisión de clasificación con la BD Libros con las 100 palabras, pero con la BD Diagnósticos con las 9 características solo se logra igualar a la precisión del mejor clasificador LDA. La metodología boosting implementada en este trabajo es sencilla, ya que los pesos requeridos en una metodología más compleja necesitan de más cálculos. Incluir las técnicas usuales de boosting al algoritmo genético puede ocasionar que el algoritmo sea más costoso.

Gracias a las pruebas realizadas con la base de datos de libros fue posible comprobar que los resultados obtenidos van de acuerdo a lo esperado, se obtienen precisiones de clasificación altas y algunas de las palabras que fueron seleccionadas en el trabajo de José Nilo G. Binongo [6] también lo fueron en este trabajo.

Lo destacable de este trabajo es que se demuestra que con distintos subconjuntos de características seleccionadas sin la intervención de un *experto* se puede obtener una precisión alta, sin contar con el sesgo que se ocasionaría si en la selección de características interviniera un *experto humano*.

8.2. Trabajo a futuro

Existen más bases de datos que contienen casos de pacientes de cáncer de mama como lo es la base de datos CaMa [27], a la cual se podría aplicar la metodología desarrollada en este trabajo y así determinar las características más relevantes de ésta.

El algoritmo genético tiene como una característica particular que es altamente paralelizable por lo que a futuro se podría realizar la paralelización de este utilizando *Message Passing Interface* (MPI).

En el apendice 9 se muestran los pasos a seguir para la realización de un clúster que puede ser utilizado para correr el algoritmo genético paralelizado, con 1 maestro y 3 esclavos, utilizando placas Raspberry pi.

Capítulo 9

Apéndice

En este apendice se describe la implementación de un clúster con 4 nodos, 1 maestro y 3 esclavos, en el que se utilizan placas Raspberry pi. Éste sería de utilidad para correr el algoritmo genético cuando sea paralelizado.

Primero se definirá lo que es un clúster y después una breve descripción del procedimiento seguido para su implementación.

¿Que es un clúster?

Un clúster es un cúmulo, granja o clúster de computadoras, se puede definir como un sistema de procesamiento paralelo o distribuido, que consta de un conjunto de computadoras independientes, interconectadas entre sí, de tal manera que funcionan como una sola computadora.

A cada uno de los elementos del clúster se le conoce como nodo. Estos son aparatos o torres que pueden tener uno o varios procesadores, memoria RAM, interfaces de red, dispositivos de entrada y salida, y sistema operativo. [21]

9.1. ¿Que es una Raspberry Pi?

Una Raspberry Pi es una placa computadora de bajo costo, se podría decir que es un ordenador de tamaño reducido, del orden de una tarjeta de crédito, desarrollado en el Reino Unido por la Fundación Raspberry PI (Universidad de Cambridge) en 2011, con el objetivo de estimular la enseñanza de la informática en las escuelas, aunque no empezó su comercialización hasta el año 2012.

El concepto es el de un ordenador desnudo de todos los accesorios que se pueden eliminar sin que afecte al funcionamiento básico. Está formada por una placa que soporta varios componentes necesarios en un ordenador común y es capaz de comportarse como tal. [22]

La Raspberry Pi guarda en su interior un importante poder de cómputo en un tamaño muy reducido.

El diseño de la Raspberry Pi incluye:
Un Chipset Broadcom BCM2835, que contiene un procesador central (CPU) ARM1176JZF-S a 700 MHz (el firmware incluye unos modos Turbo para que el usuario pueda hacerle overclock de hasta 1 GHz sin perder la garantía), un procesador gráfico (GPU) VideoCore IV, un módulo de 512 MB de memoria RAM (aunque originalmente al ser lanzado eran 256 MB), un conector de RJ45 conectado a un integrado lan9512 -jzx de SMSC que nos proporciona conectividad a 10/100 Mbps, 2 buses USB 2.0., una Salida analógica de audio estéreo por Jack de 3.5 mm, salida digital de video + audio HDMI, salida analógica de video RCA, pines de entrada y salida de proposito general, conector de alimentación microUSB y un lector de tarjetas SD.

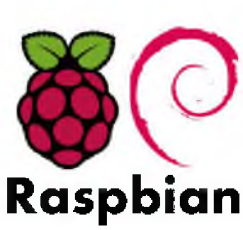
9.2. Materiales para elaborar un clúster con 4 Raspberry Pi:

4 Raspberry Pi (c) 2011,12. (Modelo B), 4 MicroSD de al menos 4GB con adaptador, 4 cables ethernet, 4 cargadores de 5v, 1 o más teclados USB, 1 o más cables HDMI, 1 monitor/pantalla, 1 switch y por útimo, si se desea un mouse USB.

9.3. Instalación de Raspbian (Debian Wheezy)

Para la instalacíon del SO en las placas se tomó en cuenta el siguiente proceso:

1. Descargar la imagen del sistema operativo: https://www.raspberrypi.org/downloads/raspbian/



2. Una vez descargado el ZIP, hay que descomprimirlo para obtener la imagen del sistema.
   Para instalar en la tarjeta microSD se hace lo siguiente:
   Desde una computadora independiente, con SO Linux.

   a. Antes de introducir la tarjeta se listan los dispositivos con el comando que se muestra debajo y se introduce la tarjeta y se vuelven a listar los dispositivos:

      ```
      df -h
      ```

   b. Habrá una línea nueva que empezará con la ruta del nuevo dispositivo:

      ```
      /dev/sdc1
      ```

   c. Ahora se va a desmontar la unidad para poder modificarla:

      ```
      umount /dev/sdc1
      ```

d. Copiar la imagen del sistema operativo en la tarjeta microSD. El comando es el siguiente:

```
dd if=/ruta/descarga/zip/2015-01-31-wheezy-raspbian.img of=/dev/sdc bs=1M
```

9.4. Instalación de Open-MPI 1.3

* Para la instalación de OpenMPI se introducen los siguientes comandos:

```
sudo apt-get install openmpi-bin
sudo apt-get install libopenmpi-dbg
sudo apt-get install libopenmpi-dev
```

* Para la correr OpenMPI.

  Entrar al directorio en el que se encuentra el archivo.c

  1. Compilar:

     ```
     mpicc testMPI.c -o testMPI
     ```

  2. Ejecutar:

     ```
     mpirun -np 4 -machinefile listanodos.txt testMPI.exe
     ```

     Donde *np* es el número de procesos.

     En *listanodos.txt* están las ip's de los nodos que pertenecen al clúster o los nodos que se han elegido para que ese programa se ejecute.

Nota: Las instalaciones se hacen en cada uno de los nodos del clúster.

Bibliografía

[1] Estadísticas a propósito del Día Internacional contra el cáncer de mama. Instituto Nacional de Estadística y Geografía, Aguascalientes, Ags. Octubre, 2013.

[2] Historia natural del cáncer de mama, Novoa, A., Pliego, M., Malagón, B. y Bustillos, R., Ginecología y Obstetricia de Mxico, 74, pp. 115-120, 2006.

[3] Perfil epidemiológico de los Tumores Malignos en México. Junio 2011. Secretaría de salud, subsecretaría de prevención y promoción de la salud dirección general de epidemiología. ISBN 978-607-460-236-4 http://www.epidemiologia.salud.gob.mx/

[4] Breast cancer in Mexico: a growing challenge to health and the health system. THE LANCET Oncology. Volume 13, Issue 8, Paginas e335 - e343, Agosto 2012, Yanin Chávarri-Guerra, Cynthia Villarreal-Garza, Pedro E R Liedke, Felicia Knaul, Alejandro Mohar, Dianne M Finkelstein, Paul E Goss.

[5] https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[6] Who Wrote the 15th Book of Oz? An Application of Multivariate Analysis to Authorship Attribution. Jos Nilo G. Binongo

[7] Minería de datos aplicada a la detección de Cáncer de Mama,Eugenio Hernández Martínez, Rodrigo Lorente Sanjurjo, Universidad Carlos III de Madrid.

[8] Estudio comparativo de técnicas de selección de características para la clasificación de lesiones de mama en ultrasonografía, Cristhian Muñoz Meza, Laboratorio de Tecnologías de Información, CINVESTAV-Tamaulipas

[9] Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks, Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau and Bart De Moor, Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium.

[10] Estudio comparativo de descriptores de textura para el desarrollo de un método computacional de segmentación automática de lesiones de mama en ultrasonografías, Refugio Ivan Rivera Islas, Laboratorio de Tecnologías de Información, CINVESTAV-Tamaulipas.

[11] Métodos de Clasificación -Alvaro Montenegro y Campo Elías Pardo

[12] Clasificación Supervisada basada en Redes Bayesianas. Aplicación en Biología Computacional, Víctor Robles Porcada, Universidad Politécnica de Madrid, Facultad de informática.

[13] Las máquinas de soporte vectorial(SVMs), Gustavo A. Betancourt,Grupo de Instrumentación y Control Facultad de Ingeniería Eléctrica Universidad Tecnológica.

[14] Induction of Decision Trees, J.R, Quinlan, Centre for Advanced Computing Sciences, New South Wales Institute of Technology, Sydney 2007, Australia.

[15] Computer Science Division, 387 Soda Hall, NIR FRIEDMAN, University of California, Berkeley, CA 94720., DAN GEIGER, Computer Science Department, Technion, Haifa, Israel, 32000., MOISES GOLDSZMIDT, SRI International, 333 Ravenswood Ave., Menlo Park, CA 94025., Bayesian Network Classifiers.

[16] Análisis de componentes principales, José Luis Vicente Villardón, Departamento de Estadística.

[17] Linear Discriminant Analysis, A Brief tutorial, S. Balakrishnama, A. Ganapathiraju

[18] http://www.ecured.cu_Distancia_euclídea

[19] Predicción de cáncer de mamas utilizando BI-RADS y un clasificador binario basado en redes neuronales artificiales, Alexander Hoyo, Dpto. Tecnología Industrial, Universidad Simón Bolívar Caracas, Venezuela.

[20] Selección de Características usando Algoritmos Genéticos para Clasificación de Vinos Chilenos, S.A. Salah, M.A. Duarte-Mermoud, N.H. Beltrán, M.A. Bustos, A.I. Pea-Neira, E.A. Loyola, and J.W. Jalocha,Depto. de Ing. Elctrica, Universidad de Chile

[21] Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS Implementations, Wen Zhu, Nancy Zeng, Ning Wang, KYL consulting services, Inc, Fort Washington, PA -Octagon Research Solutions, Wayne, PA

[22] Alfonso Mateos Andaluz, Algoritmos Evolutivos y Algoritmos Genéticos, Inteligencia en Redes de Comunicaciones Ingeniería de Telecomunicación,

[23] Algoritmos Genéticos y sus Aplicaciones, Carlos A. Coello Coello.

[24] Selección de características para clasificadores neuronales, Pablo Estévez Valencia, Instituto de Ingenieros de Chile, Diciembre 1999.

[25] http://www.revista.unam.mx/vol.4/num2/art3/cluster.htm. 2016.

[26] http://histinf.blogs.upv.es/2013/12/18/raspberry-pi/. 2016.

[27] N. Hevia-Montiel, Eduardo Sánchez-Soto y C.A. Gónzalez-Díaz. Early Breast Cancer Detection by Magnetic Induction Spectroscopy. Transactions of Japanese Society for Medical

and Biological Engineering Vol. 51(2013) No. Supplement p. R-196. ISSN Online: 1881-4379, Print: 1347-443X