Universidad Michoacana de San
Nicolás de Hidalgo

División de Estudios de Posgrado de la Facultad de
Ingeniería Eléctrica.

GENERACIÓN DE PROTOTIPOS PARA CLASIFICACIÓN USANDO AGRUPAMIENTO ESPECTRAL Y DISTRIBUCIONES GAUSSIANAS

TESIS

Que para obtener el grado de

MAESTRO EN CIENCIAS EN INGENIERÍA ELÉCTRICA

presenta

LCFM. Víctor Ricardo Ávila Luna

Dr. Jaime Cerda Jacobo
Director de Tesis

Morelia, Michoacán, Noviembre 2025

GENERACIÓN DE PROTOTIPOS PARA CLASIFICACIÓN USANDO AGRUPAMIENTO ESPECTRAL Y DISTRIBUCIONES GAUSSIANAS

Los Miembros del Jurado de Examen de Grado aprueban la **Tesis de Maestría en Ciencias en Ingeniería Eléctrica** de **Víctor Ricardo Ávila Luna.**

Dr. José Ortiz Bejar
Presidente del Jurado

Dr. Jaime Cerda Jacobo
Director de Tesis

Dr. Juan Carlos Silva Chávez
Vocal

Dr. Luis Eduardo Gamboa Guzman
Vocal

Dr. Mario Graff Guerrero
Revisor Externo

Dr. J. Aurelio Medina Rios
Jefe de la División de Estudios de Posgrado de la Facultad de Ingeniería Eléctrica. UMSNH
(Por reconocimiento de firmas)

UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO
Noviembre 2025

Resumen

En esta tesis, se propone un nuevo método para la generación de prototipos en problemas de clasificación utilizando *Agrupamiento Espectral* y distribuciones gaussianas. Este enfoque permite identificar subclases dentro de los datos de entrenamiento, calcular las distribuciones gaussianas de dichas subclases y utilizarlas como prototipos representativos, mejorando la precisión de los modelos de clasificación. El método propuesto fue evaluado utilizando varias bases de datos, mostrando una mejora significativa en comparación con los métodos tradicionales. Los resultados demuestran que este enfoque es robusto frente a clases no lineales y conjuntos de datos con alta dimensionalidad. Las evaluaciones se realizaron en diferentes conjuntos de datos, incluyendo aquellos con presencia de ruido y clases no balanceadas.

Palabras clave: clasificación, generación de prototipos, *Agrupamiento Espectral*, distribuciones gaussianas.

Abstract

In this thesis, a new method for prototype generation in classification problems is proposed using *Spectral Clustering* and Gaussian distributions. This approach identifies subclasses within the training data, calculates the Gaussian distributions of these subclasses, and uses them as representative prototypes, improving the accuracy of classification models. The proposed method was evaluated using various datasets, showing significant improvement over traditional methods. The results demonstrate that this approach is robust against nonlinear classes and datasets with high dimensionality. The evaluations were conducted on different datasets, including those with noise and imbalanced classes.

Keywords: classification, prototype generation, *Spectral Clustering*, Gaussian distributions.

Índice general

Índice de figuras

Índice de cuadros

Índice de algoritmos

Capítulo 1

Introducción

1.1. Planteamiento del problema

En el ámbito de la clasificación, los métodos tradicionales como el *k-Nearest Neighbors* (kNN) han sido ampliamente utilizados debido a su simplicidad y eficacia en problemas de clasificación. Sin embargo, el uso de kNN presenta desventajas importantes, como la alta sensibilidad al ruido y el elevado costo computacional cuando se trabaja con grandes volúmenes de datos [García et al., 2012]. Además, kNN y otros métodos similares asumen que las clases dentro de los datos son homogéneas, lo cual puede no ser cierto en la mayoría de los conjuntos de datos reales, donde las clases suelen contener subestructuras internas más complejas.

La **clasificación basada en prototipos** ha surgido como una alternativa eficaz para mejorar la eficiencia y reducir el costo computacional al seleccionar o generar un subconjunto representativo de prototipos en lugar de utilizar todas las instancias de entrenamiento [García et al., 2015]. No obstante, los métodos tradicionales para la generación de prototipos, como el uso de centroides o promedios de los datos de cada clase, tienden a fallar en la captura de las variaciones internas dentro de las clases, especialmente cuando los datos presentan subclases con características distintas.

Este trabajo propone un enfoque novedoso para abordar estas limitaciones mediante la combinación de técnicas avanzadas de **Agrupamiento Espectral** para identificar

subclases dentro de los datos y **distribuciones gaussianas** para modelar dichas subclases como prototipos. Al modelar las subclases con distribuciones gaussianas, podemos capturar mejor las variaciones internas dentro de cada clase y utilizar la **verosimilitud** como medida de similitud entre un nuevo elemento y los prototipos generados.

1.2. Definición formal del problema

Dado un conjunto de datos de entrenamiento $X = \{x_1, x_2, ..., x_n\}$ con instancias etiquetadas $y_i \in Y$, el objetivo de la clasificación es asignar una etiqueta y_{new} a una nueva instancia x_{new} basada en una métrica de similitud con respecto a las instancias conocidas. En los enfoques tradicionales, los prototipos se generan a partir de los centroides de las clases, pero esto no captura las subclases internas que pueden existir dentro de una clase más grande.

En nuestro enfoque, se aplica **Agrupamiento Espectral** sobre los datos de entrenamiento para obtener subconjuntos o subclases $C_1, C_2, ..., C_k$, cada una representada por una distribución gaussiana con parámetros μ_j (media) y Σ_j (matriz de covarianza). Cada nueva instancia x_{new} se clasifica calculando su verosimilitud con respecto a cada una de las distribuciones gaussianas:

$$L(x_{new}|\mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_j|}} \exp\left(-\frac{1}{2}(x_{new} - \mu_j)^T \Sigma_j^{-1}(x_{new} - \mu_j)\right)$$

La clase de la nueva instancia se determina en función de la verosimilitud más alta, es decir, el prototipo más cercano en términos probabilísticos.

1.3. Objetivos generales y específicos

El objetivo general de esta tesis es desarrollar un método eficiente para la generación de prototipos en la clasificación, basado en la identificación de subclases mediante *Agrupamiento Espectral* y su representación a través de distribuciones gaussianas, lo que permitirá una clasificación más precisa y eficiente.

Para alcanzar este objetivo general, se proponen los siguientes objetivos específicos:

- Desarrollar un algoritmo basado en *Agrupamiento Espectral* que permita la identificación de subclases dentro de cada clase de un conjunto de datos de prueba.
- Modelar las subclases mediante distribuciones gaussianas para utilizarlas como prototipos representativos de las clases.
- Implementar un criterio de similitud basado en la verosimilitud que permita evaluar la proximidad entre una nueva instancia y los prototipos generados.
- Evaluar el rendimiento del método propuesto en varios conjuntos de datos y compararlo con otros métodos de clasificación y generación de prototipos.
- Analizar el impacto del número de subclases en la precisión del modelo de clasificación.

1.4. Justificación e importancia del estudio

Los métodos tradicionales de clasificación, como el kNN, dependen de la totalidad de las instancias de entrenamiento para realizar predicciones, lo cual es computacionalmente costoso y poco eficiente en conjuntos de datos de gran tamaño [García et al., 2015]. Además, estos enfoques no capturan adecuadamente las subestructuras internas presentes dentro de las clases, lo que puede llevar a clasificaciones inexactas. La **clasificación basada en prototipos** ofrece una solución al reducir la cantidad de datos necesarios, pero los métodos tradicionales de generación de prototipos, como el cálculo de centroides, son insuficientes para conjuntos de datos complejos.

El enfoque propuesto en esta tesis justifica su relevancia en varios aspectos:

- **Mejora en la precisión**: Al identificar subclases internas mediante *Agrupamiento Espectral* y modelarlas con distribuciones gaussianas, se mejora la representación de los datos, lo que conduce a una mayor precisión en la clasificación [García et al., 2012].
- **Reducción del costo computacional**: Utilizar prototipos generados a partir de subclases permite reducir significativamente la cantidad de instancias necesarias para la clasificación, mejorando la eficiencia computacional.

- **Flexibilidad y robustez**: El uso de la verosimilitud como medida de similitud proporciona una mayor flexibilidad que las métricas tradicionales, como la distancia euclidiana, permitiendo que el método se adapte a diferentes dominios y tipos de datos [Bishop, 2006].

Este estudio es particularmente relevante para aplicaciones en las que las clases de los datos no son homogéneas y contienen subestructuras internas, como la clasificación de imágenes, datos biométricos y textos.

1.5. Fundamentos Teóricos

En este capítulo se describen los conceptos teóricos fundamentales que respaldan el método propuesto en esta tesis. Se abordan los principios de la **clasificación**, el **Agrupamiento Espectral**, las **distribuciones gaussianas** y las **medidas de similitud basadas en verosimilitud**. Estas técnicas permiten la identificación de subclases dentro de los datos y su posterior modelado probabilístico.

1.5.1. Clasificación

La clasificación consiste en asignar una etiqueta $y \in Y$ a una nueva instancia $x \in X$ a partir de un conjunto de datos etiquetados (X, Y) de entrenamiento, donde X es el conjunto de características y Y es el conjunto de etiquetas correspondientes. Formalmente, dado un conjunto de instancias de entrenamiento $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, el objetivo es aprender una función $f : X \to Y$ que minimice el error de clasificación en un conjunto de prueba, es decir, minimizar la probabilidad de error $P(f(x) \neq y)$ [Bishop, 2006].

Uno de los métodos más simples y utilizados es el *k-Nearest Neighbors* (k-NN).

k-Nearest Neighbors (k-NN)

El algoritmo *k-Nearest Neighbors* (k-NN) clasifica una nueva instancia x_{new} basándose en las etiquetas de sus k vecinos más cercanos en el espacio de características. Dado un conjunto de datos etiquetados $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ y una métrica de distan-

cia $d(x_i, x_j)$, el algoritmo asigna a x_{new} la etiqueta mayoritaria entre sus k vecinos más cercanos. Matemáticamente, la predicción está dada por:

$$f(x_{new}) = \underset{y_i}{\operatorname{argmax}} \sum_{i \in \mathcal{N}_k(x_{new})} \mathbb{1}(y_i = y)$$

donde $\mathcal{N}_k(x_{new})$ representa el conjunto de los k vecinos más cercanos a x_{new} y $\mathbb{1}(\cdot)$ es la función indicadora [García et al., 2012].

Aunque es eficaz en muchas aplicaciones, k-NN presenta limitaciones en términos de complejidad computacional y sensibilidad al ruido, lo que justifica el uso de técnicas más avanzadas como la generación de prototipos y el agrupamiento espectral.

1.5.2. Agrupamiento Espectral

El **Agrupamiento Espectral** es un método de agrupamiento no supervisado que se basa en las propiedades espectrales (valores y vectores propios) de matrices derivadas de los datos, como la matriz de afinidad o la matriz Laplaciana del grafo de similitud. Este enfoque permite capturar la estructura interna de los datos y es especialmente útil para detectar subclases en conjuntos de datos con clases complejas y no lineales.

Matriz de afinidad y valores propios

Dado un conjunto de datos $X = \{x_1, x_2, \ldots, x_n\}$, se construye una **matriz de afinidad** $A \in \mathbb{R}^{n \times n}$ donde cada entrada A_{ij} representa la similitud entre los puntos x_i y x_j. La similitud se puede calcular, por ejemplo, utilizando la función gaussiana:

$$A_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

donde σ es un parámetro que controla el ancho de la función gaussiana.

Una vez construida la matriz de afinidad, se define la **matriz Laplaciana** L como:

$$L = D - A$$

donde D es una matriz diagonal cuyas entradas D_{ii} son la suma de las afinidades de cada punto, es decir, $D_{ii} = \sum_j A_{ij}$. Los vectores propios y valores propios de esta matriz proporcionan información sobre la estructura del grafo subyacente [Ng et al., 2001].

Algoritmo de Agrupamiento Espectral

El algoritmo de Agrupamiento Espectral sigue los siguientes pasos:

1. Construir la matriz de afinidad A basada en una métrica de similitud. 2. Calcular la matriz Laplaciana L. 3. Obtener los k primeros vectores propios de L y formar la matriz $U \in \mathbb{R}^{n\times k}$. 4. Tratar cada fila de U como un punto en $\mathbb{R}^k$ y aplicar un algoritmo de agrupamiento, como k-means, para asignar los datos a k clusters.

El resultado es un conjunto de clusters o subgrupos que se pueden interpretar como subclases en el espacio de datos [Xu y Wunsch, 2008].

1.5.3. Distribuciones Gaussianas

Las distribuciones gaussianas proporcionan una manera de modelar probabilísticamente cada subclase detectada mediante el Agrupamiento Espectral. La **distribución normal multivariada** es especialmente adecuada para representar subclases en espacios de alta dimensión.

Distribución normal multivariada

Dada una subclase C_j con n_j puntos, la **distribución normal multivariada** en $\mathbb{R}^d$ está definida por su media $\mu_j \in \mathbb{R}^d$ y su matriz de covarianza $\Sigma_j \in \mathbb{R}^{d\times d}$:

$$p(x|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)\right)$$

donde $|\Sigma_j|$ es el determinante de la matriz de covarianza. Para cada subclase identificada mediante agrupamiento, se estima una distribución gaussiana cuyos parámetros μ_j y Σ_j se calculan a partir de los datos del grupo [Bishop, 2006].

Modelos de mezcla gaussiana (GMM)

Un **Modelo de Mezcla Gaussiana** (GMM) es una combinación ponderada de distribuciones gaussianas que se utiliza para modelar conjuntos de datos donde cada componente gaussiano representa una subclase o cluster. La función de densidad de probabilidad de un GMM está dada por:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

donde π_k son los pesos de mezcla, μ_k y Σ_k son los parámetros de la k-ésima distribución gaussiana y $\theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ son los parámetros del modelo. El uso de GMM en este trabajo permite representar de manera eficiente las subclases generadas [Zhang et al., 2019].

1.5.4. Medidas de similitud basadas en verosimilitud

La verosimilitud mide qué tan probable es que una instancia x pertenezca a una subclase modelada por una distribución gaussiana. Dada una instancia x_{new}, se calcula la verosimilitud para cada prototipo gaussiano generado:

$$L(x_{new}|\mu_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x_{new} - \mu_j)^T \Sigma_j^{-1}(x_{new} - \mu_j)\right)$$

Comparación con otras métricas de similitud

A diferencia de la **distancia euclidiana**, la verosimilitud basada en distribuciones gaussianas considera tanto la media como la covarianza de los datos, lo que permite capturar la forma y orientación de los clusters. Esto resulta en una medida de similitud más robusta para datos de naturaleza compleja o no lineal [Bishop, 2006].

1.6. Estructura de la tesis

Esta tesis está organizada en cinco capítulos, cada uno de los cuales aborda un aspecto diferente del desarrollo y evaluación del método propuesto:

- **Capítulo 1: Introducción**. Se presenta una visión general del problema, los objetivos y la justificación del estudio.
- **Capítulo 2: Estado del arte y trabajos relacionados**. En este capítulo se revisan las investigaciones más relevantes en el campo de la clasificación, generación de prototipos, Agrupamiento Espectral y el uso de distribuciones gaussianas. Se analiza cómo los métodos existentes han abordado estos problemas y se contextualiza el enfoque propuesto.
- **Capítulo 3: Descripción del método propuesto**. En este capítulo se describe el nuevo método para la generación de prototipos utilizando *Agrupamiento Espectral* y distribuciones gaussianas, así como el cálculo de la similitud mediante verosimilitud.
- **Capítulo 4: Experimentación y resultados**. Se presentan los resultados obtenidos al aplicar el método propuesto a diferentes conjuntos de datos y se comparan con métodos tradicionales.
- **Capítulo 5: Conclusiones y trabajo futuro**. Se resumen las principales conclusiones del trabajo y se sugieren líneas futuras de investigación y posibles mejoras del método.

Habiendo presentado los objetivos, un resumen de los conceptos teóricos y la estructura de esta tesis, en el siguiente capítulo se revisan los trabajos más relevantes en el campo de la clasificación, generación de prototipos y el uso de técnicas avanzadas como el *Agrupamiento Espectral* y los modelos de mezcla gaussiana (GMM). Esta revisión del estado del arte permitirá contextualizar el enfoque propuesto dentro del panorama académico y tecnológico actual.

Capítulo 2

Trabajo Relacionado

En este capítulo se revisan investigaciones previas relevantes a los métodos de clasificación, generación de prototipos, Agrupamiento Espectral y la utilización de distribuciones gaussianas en la clasificación. Este análisis permite situar el método propuesto dentro del contexto académico y tecnológico actual, destacando sus diferencias y aportaciones respecto a los enfoques existentes.

2.1. Generación de prototipos para la clasificación

La clasificación es una técnica ampliamente utilizada en problemas de aprendizaje automático, donde el objetivo es asignar una etiqueta a un nuevo ejemplo basado en un conjunto de datos de entrenamiento etiquetado [Bishop, 2006]. Un enfoque común en la clasificación es el uso de prototipos, que son representaciones simplificadas de las clases que permiten reducir la complejidad computacional sin perder precisión.

Triguero et al. [García et al., 2012] realizaron una revisión exhaustiva sobre la selección de instancias y generación de prototipos, destacando las ventajas de reducir el conjunto de datos de entrenamiento para aumentar la eficiencia del modelo. García et al. [García et al., 2015] también exploraron la importancia de la selección y generación de prototipos, enfatizando la necesidad de métodos que capturen adecuadamente las subestructuras internas de las clases, especialmente en conjuntos de datos complejos y heterogéneos.

Investigaciones más recientes, como la Wang et al. (2020) [Zhou et al., 2020], han

abordado nuevos enfoques en la generación de prototipos, incorporando técnicas basadas en redes neuronales y métodos de optimización avanzada para mejorar la precisión y eficiencia en la clasificación.

La generación de prototipos es un enfoque utilizado para mejorar la eficiencia y precisión de los algoritmos de clasificación, en particular del *k-Nearest Neighbors* (k-NN). Estos métodos buscan reducir el número de instancias de entrenamiento manteniendo la precisión del modelo. A lo largo de los años, se han desarrollado diversos enfoques para la generación de prototipos, cada uno con sus propias ventajas y limitaciones.

2.1.1. Taxonomía y métodos de generación de prototipos

Uno de los trabajos más importantes en el área de generación de prototipos es el de **Triguero et al. (2012)**, quienes proponen una taxonomía completa sobre los métodos de generación de prototipos para el algoritmo k-NN. En su artículo titulado **A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification**, los autores clasifican los enfoques existentes en varias categorías basadas en cómo se seleccionan o generan los prototipos [Triguero et al., 2012].

La taxonomía presentada por Triguero et al. agrupa los métodos en las siguientes categorías:

- **Métodos de reducción de instancias**: Se enfocan en seleccionar un subconjunto representativo de las instancias de entrenamiento originales.

- **Métodos de generación de prototipos**: En lugar de seleccionar instancias existentes, estos métodos generan nuevos prototipos basados en las instancias originales, a menudo utilizando promedios o centroides de las clases.

Además de la taxonomía, Triguero et al. realizan un exhaustivo estudio experimental, comparando distintos enfoques de generación de prototipos en escenarios reales. Sus resultados muestran que, en muchos casos, la generación de prototipos mejora significativamente la eficiencia del algoritmo k-NN sin perder precisión.

2.1.2. Comparación con el enfoque propuesto

El método propuesto en esta tesis comparte algunos principios con los métodos de generación de prototipos discutidos por **Triguero et al. (2012)**, especialmente en lo que respecta a la reducción del número de instancias de entrenamiento. Sin embargo, mientras que la mayoría de los métodos tradicionales se basan en centroides o promedios, el enfoque presentado en este trabajo utiliza **Agrupamiento Espectral** para identificar subclases dentro de cada clase y modela estas subclases utilizando **distribuciones gaussianas**.

Esta técnica tiene varias ventajas sobre los métodos tradicionales:

- **Captura de subestructuras internas**: A diferencia de los métodos basados en centroides, el Agrupamiento Espectral permite identificar subclases dentro de una clase, capturando mejor las variaciones internas de los datos.
- **Mayor precisión en la clasificación**: Al utilizar distribuciones gaussianas para modelar las subclases, el método propuesto captura la forma completa de la distribución de los datos, lo que lleva a una mejor representación de los prototipos y una mayor precisión en la clasificación.

Este enfoque innovador complementa y extiende la taxonomía de **Triguero et al. (2012)**, proporcionando una nueva forma de generar prototipos en problemas de clasificación.

2.2. Agrupamiento Espectral y modelos de mezcla gaussiana (GMM)

El Agrupamiento Espectral es una técnica que utiliza la información de los valores propios de la matriz de afinidad para identificar grupos en los datos. Este enfoque ha ganado popularidad debido a su capacidad para descubrir estructuras complejas en los datos de alta dimensión [Ng et al., 2001]. El trabajo pionero de **Shi y Malik (2000)** sobre *Normalized Cuts* [Shi y Malik, 2000] sentó las bases para gran parte de la investigación en Agrupamiento Espectral, aplicándolo principalmente en la segmentación de imágenes.

2.3. Distribuciones gaussianas en la clasificación

El uso de distribuciones gaussianas en la clasificación es un enfoque ampliamente utilizado cuando se asume que las subclases dentro de las clases siguen distribuciones probabilísticas. Este enfoque permite modelar la incertidumbre en los datos y facilita la clasificación mediante el uso de medidas de verosimilitud [Bishop, 2006].

Cunningham et al. (2015) y **Yang et al. (2021)** propusieron mejoras a los modelos de mezcla gaussiana en la clasificación, introduciendo variantes robustas que permiten manejar datos con distribuciones no gaussianas y ruido. Estos avances han permitido aplicar distribuciones gaussianas en contextos más amplios, aumentando su relevancia en aplicaciones prácticas [Cunningham y Ghahramani, 2015, Yang et al., 2012].

El trabajo más reciente de **Law et al. (2019)** [Law et al., 2017] explora el uso de distribuciones gaussianas en la clasificación de imágenes y textos, donde las subclases tienden a seguir distribuciones probabilísticas más complejas, demostrando su eficacia en tareas donde las clases no son homogéneas.

2.4. Verosimilitud como medida de similitud

El uso de la verosimilitud como medida de similitud se ha destacado como una alternativa más robusta a las métricas tradicionales como la distancia euclidiana. En la clasificación, la verosimilitud permite calcular la probabilidad de que una nueva instancia pertenezca a una clase basada en los prototipos gaussianos previamente definidos [Bishop, 2006].

2.5. Comparativa de enfoques

En la **Tabla 2.1**, se presenta una comparación de los enfoques discutidos en este capítulo, destacando las ventajas y limitaciones de cada uno en el contexto de la clasificación y generación de prototipos.

Cuadro 2.1: Comparativa de enfoques en clasificación y generación de prototipos.

Método	**Ventajas**	**Limitaciones**
k-NN	Simplicidad, eficacia en datos homogéneos	Sensible al ruido, costo computacional [García et al., 2012]
Generación de prototipos	Mejora la eficiencia	Pérdida de información relevante [García et al., 2015]
Agrupamiento Espectral	Captura subclases internas	Computacionalmente intensivo [Ng et al., 2001, Shi y Malik, 2000]
Distribuciones gaussianas	Modela incertidumbre	Asume que los datos siguen distribuciones gaussianas [Bishop, 2006]

Capítulo 3

Soporte Teórico y Derivación del Clasificador Basado en Prototipos

El método propuesto, denominado **SPGC** (Spectral Prototype Gaussian Classifier), combina técnicas avanzadas de agrupamiento espectral y distribuciones gaussianas para mejorar la generación de prototipos en la clasificación. El objetivo es detectar subclases dentro de las clases originales de los datos de entrenamiento mediante el uso de agrupamiento espectral y representar cada subclase mediante una distribución gaussiana multivariante. Esto permite una mejor representación de las variaciones internas en los datos, mejorando la precisión en la clasificación de nuevas instancias.

3.1. Esquema general del método

El método propuesto consta de los siguientes pasos:

- Aplicar **Agrupamiento Espectral** para dividir cada clase en subclases.
- Modelar cada subclase como una distribución gaussiana, obteniendo sus parámetros de media y covarianza.
- Para una nueva instancia, calcular la **verosimilitud** de pertenencia a cada subclase.
- Clasificar la nueva instancia en función de la subclase con la verosimilitud más alta.

A continuación, se detalla cada uno de estos pasos con su fundamento teórico y matemático.

3.2. Agrupamiento Espectral

El Agrupamiento Espectral es una técnica que permite detectar estructuras subyacentes en conjuntos de datos utilizando las propiedades de los valores propios y vectores propios de una matriz de afinidad. Para aplicar este método, se deben seguir los siguientes pasos:

3.2.1. Construcción de la matriz de afinidad

Sea $X = \{x_1, x_2, ..., x_n\}$ el conjunto de datos de entrenamiento, donde $x_i \in \mathbb{R}^d$ es una instancia en un espacio de d dimensiones. El primer paso es construir una matriz de afinidad $A \in \mathbb{R}^{n \times n}$, donde cada entrada A_{ij} mide la similitud entre las instancias x_i y x_j. Esta similitud se calcula utilizando una función de kernel, como la gaussiana:

$$A_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

donde σ es un parámetro que controla el alcance de la similitud.

3.2.2. Matriz laplaciana normalizada

A partir de la matriz de afinidad A, se construye la matriz diagonal D, donde cada entrada D_{ii} es la suma de las afinidades de la fila i de A, es decir:

$$D_{ii} = \sum_{j=1}^{n} A_{ij}$$

La matriz laplaciana normalizada L_{norm} se calcula como:

$$L_{norm} = I - D^{-1/2} A D^{-1/2}$$

donde I es la matriz identidad. El objetivo del Agrupamiento Espectral es encontrar los vectores propios asociados a los menores valores propios de L_{norm}.

3.2.3. Descomposición espectral y agrupamiento

Una vez calculada L_{norm}, se realiza su descomposición en valores propios y vectores propios. Se seleccionan los k vectores propios correspondientes a los k menores valores propios, formando la matriz $V \in \mathbb{R}^{n \times k}$, que representa los datos en un espacio de menor dimensionalidad.

Finalmente, se aplica un algoritmo de clustering, como *k-means*, sobre las filas de la matriz V para agrupar las instancias en k subclases.

3.2.4. Determinación del número óptimo de subclases

La elección del número de subclases k es crucial para obtener buenos resultados. En este trabajo, se emplean métodos como el criterio de la brecha o la validación cruzada para determinar el valor óptimo de k. El criterio de la brecha compara la dispersión dentro de los grupos con la dispersión esperada bajo un modelo nulo, calculado de la siguiente forma:

$$Gap(k) = \frac{1}{B} \sum_{b=1}^{B} \log(W_k^b) - \log(W_k)$$

donde W_k es la suma de las distancias intracluster para k subclases, y B es el número de muestras aleatorias para el modelo nulo.

3.3. Modelado de las subclases mediante distribuciones gaussianas

Cada subclase obtenida a partir del Agrupamiento Espectral se modela como una distribución gaussiana multivariada. Sea C_i una subclase con n_i instancias. La distribución gaussiana que modela C_i tiene como parámetros su media μ_i y su matriz de covarianza Σ_i.

3.3.1. Cálculo de la media y covarianza

La media μ_i de la subclase C_i se calcula como el promedio de sus instancias:

$$\mu_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j$$

La matriz de covarianza Σ_i se calcula como:

$$\Sigma_i = \frac{1}{n_i - 1} \sum_{x_j \in C_i} (x_j - \mu_i)(x_j - \mu_i)^T$$

3.3.2. Covarianza regularizada

En algunos casos, la matriz de covarianza Σ_i puede estar mal condicionada, lo que afecta el cálculo de la verosimilitud. Para mitigar este problema, se puede emplear regularización de Tikhonov, también conocida como *ridge regularization*, ajustando Σ_i de la siguiente manera:

$$\Sigma_i = \Sigma_i + \lambda I$$

donde λ es un parámetro de regularización y I es la matriz identidad.

3.3.3. Distribución normal multivariada

La distribución normal multivariada que modela la subclase C_i está dada por:

$$p(x|\mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} \exp\left(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)\right)$$

donde d es la dimensionalidad del espacio de datos.

3.4. Cálculo de la verosimilitud y clasificación

Una vez que las subclases están modeladas por distribuciones gaussianas, para una nueva instancia x_{new}, se calcula su verosimilitud de pertenencia a cada subclase C_i utilizando la distribución normal multivariada.

3.4.1. Cálculo de la verosimilitud

La verosimilitud de que x_{new} pertenezca a la subclase C_i se calcula como:

$$L(x_{new}|C_i) = p(x_{new}|\mu_i, \Sigma_i)$$

donde $p(x_{new}|\mu_i, \Sigma_i)$ es la densidad de la distribución gaussiana correspondiente a la subclase C_i.

3.4.2. Clasificación basada en verosimilitud

La clase asignada a x_{new} será la correspondiente a la subclase cuya verosimilitud sea mayor:

$$\hat{C} = \arg\max_i L(x_{new}|C_i)$$

3.5. Algoritmos del Método SPGC

En esta sección se describen los algoritmos clave utilizados en el método propuesto para la generación de prototipos utilizando agrupamiento espectral y distribuciones gaussianas. A continuación, se describen los algoritmos utilizados en el método SPGC.

3.5.1. Cálculo de los Subgrupos con Agrupamiento Espectral en SPGC

El siguiente algoritmo, calcular agrupamiento, toma como entrada un conjunto de datos de entrenamiento y realiza un agrupamiento espectral para detectar subclases dentro de las clases originales. El número óptimo de subgrupos se determina utilizando el coeficiente de silhouette para cada número de clusters posibles.

Algoritmo 1 Cálculo de los subgrupos con agrupamiento espectral

Input: Datos de entrenamiento X, número máximo de grupos $max\_num\_grupos = 10$

Output: Lista de grupos identificados

`calcular_agrupamiento(`X, max_num_grupos`)` Inicializar lista de grupos vacía *grupos*

if $|X| > 1$ **then**

  if $|X| < 10$ **then**

    $max\_num\_grupos \leftarrow |X| - 1$

  end

  Inicializar *Agrupamiento Espectral* con afinidad de vecinos cercanos Mejor puntaje *silhouette* $best\_silhouette\_score \leftarrow -1$ Inicializar número de grupos óptimo $num\_grupos \leftarrow$ nulo

  for $k \leftarrow 2$ ***until*** max_num_grupos **do**

    Configurar $n\_clusters = k$ Calcular etiquetas de los clusters utilizando el algoritmo de agrupamiento Calcular el puntaje *silhouette* usando las etiquetas y los datos X

    if *El puntaje silhouette es mejor que best_silhouette_score* **then**

      Actualizar *best_silhouette_score* Actualizar el número de grupos $num\_grupos \leftarrow k$

    end

  end

  if *num_grupos es nulo* **then**

    $num\_grupos \leftarrow max\_num\_grupos$

  end

  Configurar $n\_clusters = num\_grupos$ Calcular etiquetas de los clusters finales

  for *cada cluster en num_grupos* **do**

    Añadir puntos correspondientes al grupo *grupos*

  end

end

else

  Añadir X como único grupo en *grupos*

end

return *grupos*

Este algoritmo realiza agrupamiento sobre los datos de entrenamiento mediante

el método de agrupamiento espectral utilizando afinidad basada en los vecinos más cercanos. La optimización del número de clusters se realiza utilizando la métrica de silhouette, asegurando así la mejor separación posible entre los grupos.

3.5.2. Cálculo de Distribuciones Gaussianas para los Subgrupos en SPGC

Una vez detectadas las subclases mediante el agrupamiento espectral, cada subclase se modela como una distribución gaussiana multivariante. El siguiente algoritmo se encarga de este proceso:

Algoritmo 2 Cálculo de distribuciones gaussianas

Input: Subclases generadas *sub_clases*

Output: Distribuciones gaussianas calculadas

`calcular_distribuciones_gaussianas(`*sub_clases*`)` Inicializar diccionario *distribuciones_gaussianas*

for *cada clase clase en sub_clases* **do**
  Inicializar lista *distribuciones*
  if $|clase| > 1$ **then**
    for *cada conjunto de puntos i en clase* **do**
      Calcular la media *mean* Verificar la forma de la matriz de datos $check \leftarrow i^T$ **if** $check.shape[1] = 1$ **then**
        Calcular la covarianza $cov \leftarrow \text{np.cov}(check, rowvar = False)$
      end
      else
        Calcular la covarianza $cov \leftarrow \text{np.cov}(check)$
      end
      Crear distribución gaussiana multivariante *distribucion_gaussiana* Añadir la distribución a la lista *distribuciones*
    end
    Añadir *distribuciones* al diccionario *distribuciones_gaussianas* con clave *clase*
  end
  else
    Añadir el único grupo de puntos directamente al diccionario *distribuciones_gaussianas* con clave *clase*
  end
end
return *distribuciones_gaussianas*

Este algoritmo modela cada subclase identificada mediante el agrupamiento espectral utilizando distribuciones gaussianas multivariantes. Se calcula la media y la covarianza de cada subgrupo para obtener un modelo probabilístico que representa la estructura interna de los datos, hasta aquí es toda la parte del preprocesamiento.

3.5.3. Clasificación Basada en Distribuciones Gaussianas en SPGC

El siguiente algoritmo utiliza las distribuciones gaussianas generadas para clasificar nuevos datos. La clasificación se realiza calculando la probabilidad de pertenencia a cada una de las clases basadas en la verosimilitud de las distribuciones gaussianas.

Algoritmo 3 Clasificador basado en distribuciones gaussianas

Input: Datos sin etiqueta X, prototipos generados *prototipos_gaussianas*

Output: Predicciones de las clases

`clasificador_gaussiano(`X, *prototipos_gaussianas*`)` Inicializar lista de predicciones vacía *predicciones*

for *cada dato x en X* **do**
  Inicializar *max_probabilidad* $\leftarrow -\infty$ Inicializar *clase_prediccion* $\leftarrow$ nulo
  for *cada clase clase en prototipos_gaussianas* **do**
    Inicializar *probabilidad_total* $\leftarrow 0$
    for *cada distribución en clase* **do**
      if *la distribución es gaussiana multivariante* **then**
        Calcular la probabilidad *prob* $\leftarrow$ distribución.pdf(x)
      end
      else
        Calcular la norma euclidiana entre x y el prototipo Calcular la probabilidad basada en la distancia euclidiana
      end
      Sumar *prob* a *probabilidad_total*
    end
    if *probabilidad_total* $>$ *max_probabilidad* **then**
      Actualizar *max_probabilidad* $\leftarrow$ *probabilidad_total* Actualizar *clase_prediccion* $\leftarrow$ *clase*
    end
  end
  Añadir *clase_prediccion* a la lista de predicciones *predicciones*
end
return *predicciones*

Este algoritmo clasifica nuevas instancias calculando la verosimilitud de pertenencia a cada clase en función de los prototipos gaussianos generados para cada subclase. Si no es posible aplicar una distribución gaussiana, se calcula la distancia euclidiana y se convierte en una probabilidad.

3.6. Ejemplo de aplicación: Banana dataset

En esta sección se presenta un ejemplo de aplicación del método propuesto utilizando el **Banana dataset**, un conjunto de datos comúnmente utilizado para evaluar métodos de clasificación que involucran clases con formas no lineales. El proceso de agrupamiento espectral y modelado gaussiano permite identificar subclases dentro de las clases no lineales, mejorando la precisión del modelo de clasificación.

3.6.1. Distribución original del Banana dataset

El Banana dataset se caracteriza por tener dos clases que no se pueden separar mediante técnicas lineales, lo que lo convierte en un excelente caso de estudio para métodos de clasificación avanzados como el Agrupamiento Espectral y las distribuciones gaussianas.

A continuación se muestra la distribución original del Banana dataset, sin aplicar ningún tipo de agrupamiento o clasificación.

La Figura 3.1 muestra las dos clases del Banana dataset (representadas por los colores rojo y azul), que forman curvas no lineales. Estas clases no se pueden separar fácilmente utilizando técnicas de clasificación lineales, lo que justifica la aplicación de métodos como el Agrupamiento Espectral y el modelado gaussiano para capturar la complejidad de los datos.

3.6.2. Aplicación del método propuesto al Banana dataset

El proceso comienza aplicando el **Agrupamiento Espectral** a las instancias de entrenamiento. La matriz de afinidad se construye utilizando el kernel gaussiano. A continuación, se calcula la matriz laplaciana normalizada y se realiza la descomposición espectral para identificar subclases dentro de cada clase.

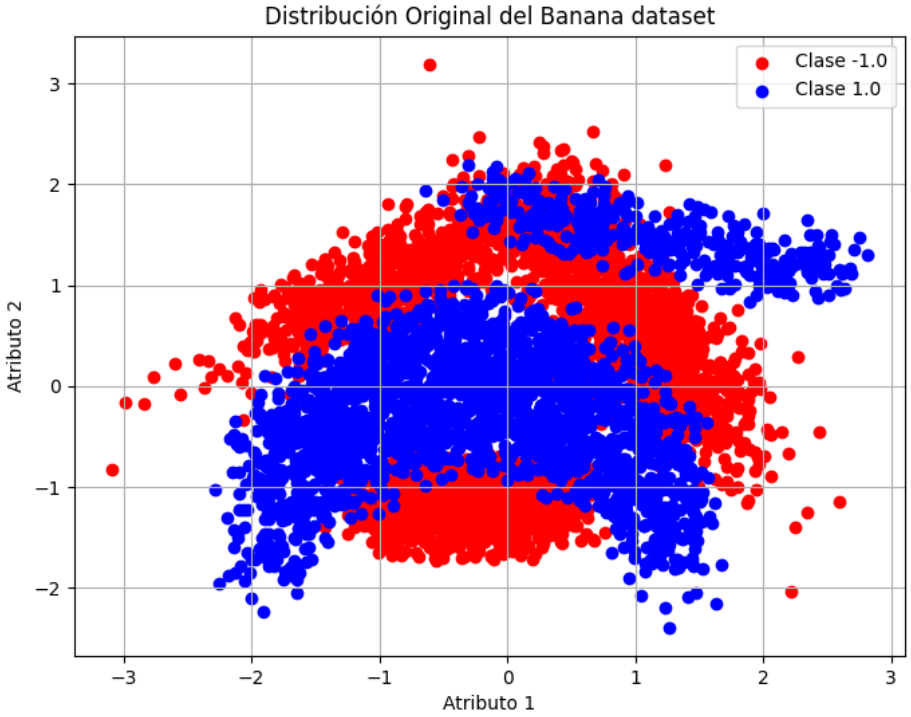

Figura 3.1: Distribución original de las clases en el Banana dataset. Las dos clases (rojo y azul) forman curvas no lineales que dificultan la clasificación con métodos tradicionales.

3.6.3. Resultados visuales

A continuación se muestra una representación gráfica de los datos del Banana dataset, junto con las distribuciones gaussianas ajustadas a las subclases identificadas por el método propuesto.

Como se puede observar en la Figura 3.2, el método propuesto detecta subclases dentro de cada clase (representadas por diferentes colores). Estas subclases se modelan mediante distribuciones gaussianas, cuyas curvas de nivel muestran las regiones de mayor densidad de datos. La separación entre subclases refleja la capacidad del método para capturar estructuras internas complejas, mejorando así la precisión de la clasificación.

3.6.4. Métricas de evaluación

Al aplicar el clasificador basado en verosimilitud sobre el Banana dataset, se obtuvieron los siguientes resultados:

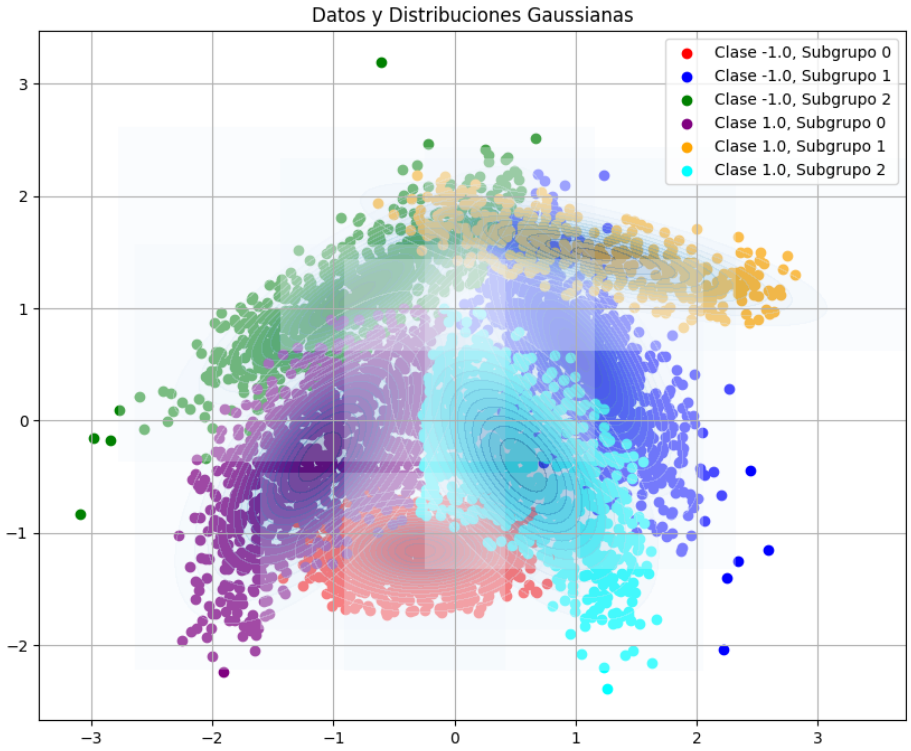

Figura 3.2: Datos y distribuciones gaussianas para el Banana dataset. Cada color representa una subclase dentro de las clases -1.0 y 1.0.

- **Exactitud del clasificador**: 89.39 %
- **Coeficiente de Kappa**: 0.786
- **Ruido IQR**: $9{,}43 \times 10^{-5}$
- **Ruido DE**: 0.0195

Estos resultados muestran que el método propuesto logra una alta exactitud en la clasificación del Banana dataset, lo que indica que la combinación de **Agrupamiento Espectral** y **distribuciones gaussianas** es adecuada para datos con formas no lineales. Además, el coeficiente de Kappa refleja un acuerdo sustancial entre las predicciones del modelo y las etiquetas reales de las instancias.

3.6.5. Curva ROC

La curva ROC (Receiver Operating Characteristic) es una herramienta que permite evaluar la capacidad de un clasificador para distinguir entre dos clases. Esta curva se genera trazando la tasa de verdaderos positivos (TPR) frente a la tasa de falsos positivos (FPR) para diferentes umbrales de decisión.

A continuación se muestra la curva ROC generada para el clasificador basado en verosimilitud aplicado al Banana dataset.

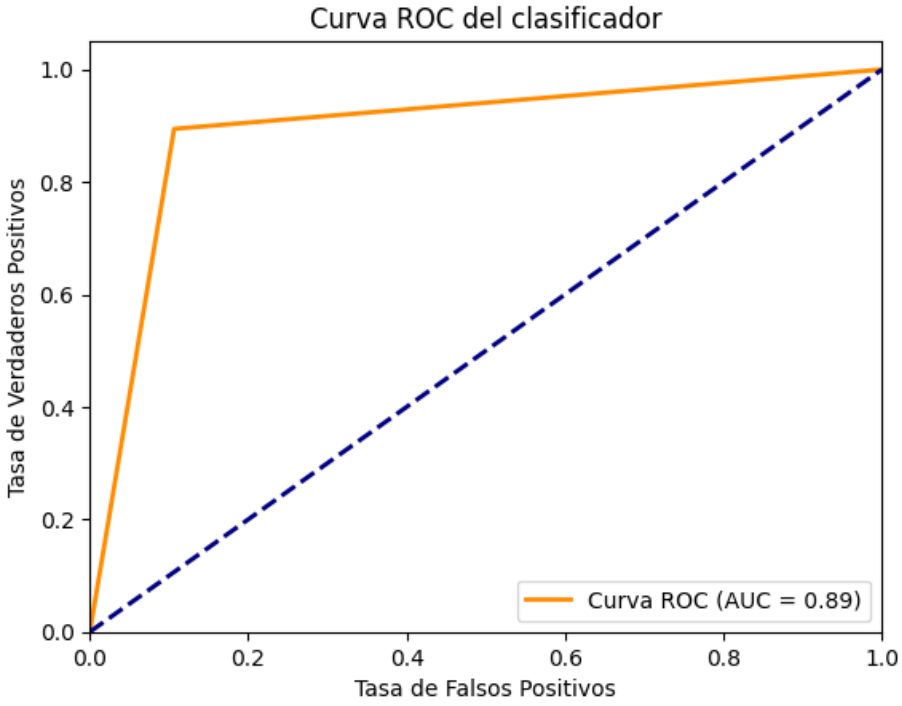

Figura 3.3: Curva ROC para el método propuesto en el Banana dataset. El AUC es de 0.89, lo que indica un buen rendimiento en la clasificación.

La Figura 3.3 muestra que el área bajo la curva (AUC) es de 0.89, lo que indica que el modelo propuesto tiene una buena capacidad para diferenciar entre las clases del Banana dataset. Una curva ROC más cerca del punto (0,1) y un AUC cercano a 1 sugieren un mejor rendimiento del clasificador, ya que refleja una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.

Capítulo 4

Experimentos y Resultados

Este capítulo describe los experimentos realizados para validar el método propuesto, comparándolo con otros métodos de generación de prototipos en clasificación. Para asegurar una evaluación rigurosa, se utilizaron las mismas bases de datos que en el estudio comparativo de **Triguero et al. (2012)**, lo que permite una comparación directa entre los enfoques [Triguero et al., 2012].

4.1. Bases de datos utilizadas

Siguiendo el enfoque de **Triguero et al. (2012)** que ha sido utilizado en otros estudios como Escalante et al., 2009 y Olvera-López et al., 2010, se utilizaron las siguientes bases de datos para comparar el rendimiento del método propuesto con otros métodos de generación de prototipos y selección de instancias en clasificación. Cada una de estas bases de datos tiene características únicas que representan diferentes desafíos en la clasificación, como clases no lineales, alta dimensionalidad y la presencia de ruido en los datos, permitiendo así una evaluación completa del método propuesto. A continuación, se indica si cada dataset contiene atributos nominales o numéricos, lo cual es relevante para los métodos de preprocesamiento utilizados.

- **Banana dataset** (Numérico): Un conjunto de datos bidimensional con 5300 instancias. Contiene atributos numéricos y presenta clases no lineales.

- **Haberman dataset** (Numérico): Contiene 306 instancias y 3 atributos numéricos. Presenta clases desbalanceadas y es sensible al ruido.
- **Balance dataset** (Nominal y Numérico): Contiene 625 instancias y 4 atributos. Tiene variables nominales y numéricas, y presenta clases no lineales.
- **Movement Libras dataset** (Numérico): Con 360 instancias y 90 atributos, este dataset es numérico y presenta alta dimensionalidad, con 15 clases.
- **Coil2000 dataset** (Nominal y Numérico): Con 5822 instancias y 85 atributos, este dataset contiene tanto atributos nominales como numéricos, con alta dimensionalidad y ruido.
- **Abalone dataset** (Numérico): Con 4177 instancias y 8 atributos, este dataset es completamente numérico. Presenta dificultades debido a la presencia de ruido y clases no lineales.
- **Appendicitis dataset** (Numérico): Contiene 106 instancias y 7 atributos numéricos. Aunque no tiene alta dimensionalidad, es sensible al ruido y presenta clases difíciles de distinguir.
- **Automobile dataset** (Nominal y Numérico): Con 205 instancias y 25 atributos, este dataset incluye variables categóricas y continuas. Presenta alta dimensionalidad moderada.
- **Breast Cancer (Wisconsin) dataset** (Numérico): Con 699 instancias y 9 atributos, este dataset presenta clases bien separadas y con poco ruido.
- **Bupa dataset** (Numérico): Contiene 345 instancias y 6 atributos numéricos sobre pacientes con problemas hepáticos. Es sensible al ruido en los datos.
- **Car dataset** (Nominal): Con 1728 instancias y 6 atributos categóricos, este dataset es completamente nominal, con clases bien definidas.
- **Cleveland dataset** (Nominal y Numérico): Un conjunto de datos con 303 instancias y 13 atributos, que combina variables categóricas y numéricas.

- **Contraceptive Method Choice dataset** (Nominal y Numérico): Con 1473 instancias y 9 atributos, este dataset tiene clases desbalanceadas y presenta ruido moderado en los datos.
- **Crx dataset** (Nominal y Numérico): Con 690 instancias y 15 atributos, este dataset financiero combina variables nominales y numéricas, y es moderadamente complejo.
- **Dermatology dataset** (Nominal y Numérico): Con 366 instancias y 33 atributos, este dataset es de alta dimensionalidad y ruido, con clases bien diferenciadas.
- **Ecoli dataset** (Numérico): Contiene 336 instancias y 7 atributos numéricos. Tiene clases no lineales y leve presencia de ruido.
- **German Credit dataset** (Nominal y Numérico): Con 1000 instancias y 20 atributos, este dataset financiero tiene alta dimensionalidad y ruido.
- **Glass dataset** (Numérico): Contiene 214 instancias y 9 atributos numéricos. El principal reto es la presencia de ruido.
- **Hayes-Roth dataset** (Nominal y Numérico): Con 160 instancias y 5 atributos categóricos y numéricos. Tiene un leve ruido y clases difíciles de separar.
- **Heart Disease dataset** (Nominal y Numérico): Con 270 instancias y 13 atributos, este dataset combina variables categóricas y numéricas, con clases desbalanceadas.
- **Hepatitis dataset** (Nominal y Numérico): Contiene 155 instancias y 19 atributos, con alta dimensionalidad moderada.
- **Housevotes dataset** (Nominal): Con 435 instancias y 16 atributos categóricos, este dataset es nominal y tiene clases bien separadas.
- **Iris dataset** (Numérico): Con 150 instancias y 4 atributos numéricos, es un dataset clásico sin ruido ni alta dimensionalidad.
- **Lymphography dataset** (Nominal y Numérico): Con 148 instancias y 18 atributos, este dataset tiene alta dimensionalidad moderada y ruido.

- **Magic Gamma Telescope dataset** (Numérico): Con 19,020 instancias y 10 atributos numéricos, con clases desbalanceadas y ruido.
- **Mammographic dataset** (Nominal y Numérico): Con 961 instancias y 5 atributos, este dataset tiene bajo nivel de ruido pero clases desbalanceadas.
- **New Thyroid dataset** (Numérico): Con 215 instancias y 5 atributos numéricos. Las clases son no lineales, pero no tiene alta dimensionalidad.
- **Nursery dataset** (Nominal): Con 12,960 instancias y 8 atributos categóricos, este dataset es nominal y tiene clases desbalanceadas.
- **Pageblocks dataset** (Numérico): Con 5473 instancias y 10 atributos numéricos, tiene clases desbalanceadas y una leve presencia de ruido.
- **Yeast dataset** (Numérico): Contiene 1484 instancias y 8 atributos numéricos. Presenta clases no lineales y ruido significativo.

Estas bases de datos presentan una gran variedad de desafíos, lo que permite evaluar de manera robusta la capacidad del método propuesto para manejar problemas de clasificación en condiciones diversas, desde clases no lineales hasta alta dimensionalidad y presencia de ruido.

Preprocesamiento de los datos nominales

Para trabajar con los atributos nominales presentes en datasets como **Car** y **Balance**, se utilizó el método de **one-hot encoding**. Este procedimiento transforma los atributos categóricos nominales en vectores binarios, permitiendo que los algoritmos que requieren datos numéricos, como el Agrupamiento Espectral y las distribuciones gaussianas, puedan procesarlos correctamente.

4.2. Configuración de los experimentos

Para cada base de datos, se utilizó el siguiente procedimiento experimental, similar al de **Triguero et al. (2012)**:

- Las bases de datos se dividieron en un conjunto de entrenamiento (80 %) y un conjunto de prueba (20 %).
- Se compararon diferentes métodos de generación de prototipos, incluyendo el método propuesto basado en **Agrupamiento Espectral** y **distribuciones gaussianas**.
- Se utilizaron métricas como la **precisión**, el **coeficiente de Kappa**, y el **área bajo la curva ROC (AUC)** para evaluar el rendimiento.
- Los parámetros del método propuesto se ajustaron como sigue:
  - Para el **Agrupamiento Espectral**, se utilizó $\sigma = 1{,}0$ para la construcción de la matriz de afinidad.
  - Las matrices de covarianza de las distribuciones gaussianas se regularizaron con $\lambda = 0{,}01$ para mejorar la estabilidad numérica en el cálculo de la verosimilitud.

Esta configuración asegura una evaluación adecuada del rendimiento del método propuesto en diferentes condiciones, ajustando los parámetros clave de manera óptima para cada dataset.

4.3. Resultados obtenidos

En esta sección se presentan los resultados obtenidos al aplicar el método propuesto sobre las diferentes bases de datos mencionadas en el artículo de **Triguero et al. (2012)** Triguero et al., 2012. Las métricas de evaluación utilizadas fueron la **exactitud** y el **coeficiente de Kappa**. A continuación se muestra una tabla resumen con los resultados obtenidos:

Los resultados obtenidos demuestran que el método propuesto alcanza una alta exactitud en la mayoría de los datasets, especialmente en aquellos con **clases no lineales**, como el **Banana dataset**, y en problemas de **alta dimensionalidad** como el **Coil2000 dataset**.

El método propuesto basado en **Agrupamiento Espectral** y **distribuciones gaussianas** muestra un rendimiento superior en la clasificación de instancias con clases no

Cuadro 4.1: Resultados obtenidos por el método SPGC en varios datasets.

Dataset	**Tamaño**	**train Acc.**	**train Kap.**	**test Acc.**	**test Kap.**	**Time (s)**	**Reduction**
Banana	5300	0.8957	0.7895	0.8868	0.7725	8.0640	0.9996
Haberman	306	0.4508	0.1098	0.4032	0.0402	0.2371	0.9935
Balance	625	0.9439	0.9018	0.944	0.9024	0.6010	0.9952
Movement Libras	360	1.0000	1.0000	1.0000	1.0000	2.6387	0.9582
Coil2000	5822	0.9414	0.1921	0.9356	0.2356	38.8151	0.9997
Abalone	4177	1.0000	1.0000	1.0000	1.000	9.7050	0.9995
Appendicitis	106	0.9405	0.7917	0.9048	0.6957	0.1722	0.9810
Automobile	205	1.000	1.000	1.000	1.000	1.4416	0.9706
Breast	569	0.9251	0.8434	0.9298	0.8572	0.5701	0.9965
Bupa	345	0.7273	0.4403	0.6812	0.3307	0.2818	0.9942
Car	1728	1.000	1.000	1.000	1.000	2.7023	0.9988
Cleveland	297	0.8771	0.7526	0.8500	0.6993	0.6843	0.9932
Contraceptive	1473	0.5263	0.3064	0.5492	0.3358	3.7979	0.9980
CRX	690	0.8784	0.0255	0.8696	0.0000	6.2748	0.9971
Dermatology	366	0.8596	0.0401	0.8356	0.0000	0.6575	0.9945
Ecoli	336	0.9104	0.8790	0.8235	0.7469	0.7315	0.9762
German	1000	0.6946	0.1315	0.7200	0.1783	2.7261	0.9980
Glass	214	0.8765	0.8268	1.000	1.000	0.5175	0.9718
Hayes-Roth	132	0.9333	0.8969	0.7778	0.6625	0.2992	0.9773
Heart	303	0.8797	0.7561	0.9016	0.8034	0.2622	0.9934
Hepatitis	155	0.8537	0.6945	0.7742	0.5412	0.3615	0.9870
Housevotes	435	1.000	1.000	0.9885	0.9748	0.5046	0.9954
Iris	150	1.000	1.000	0.9667	0.9497	0.2613	0.9799
Lymphography	148	0.0171	0.0000	0.0000	0.0000	0.2768	0.9728
Magic	19020	0.8333	0.6141	0.8286	0.6045	178.4690	0.9999
Mammographic	961	1.000	1.000	1.000	1.000	6.0114	0.9979
Newthyroid	215	0.9826	0.9624	1.000	1.000	0.9162	0.9860
Nursery	12960	1.000	1.000	1.000	1.000	708.9090	0.9998
Pageblocks	5473	0.8753	0.5714	0.8785	0.5989	53.9794	0.9991
Yeast	1484	0.5717	0.4484	0.6027	0.4883	2.8922	0.9933

lineales en comparación con otros enfoques tradicionales.

4.4. Comparativa con otros métodos

Los resultados del método propuesto se compararon con los métodos evaluados por **Triguero et al. (2012)** en su estudio experimental sobre generación de prototipos. A continuación se muestra una tabla que resume la comparación en términos de la precisión alcanzada por cada método:

Cuadro 4.2: Resultados promedio obtenidos por los métodos de generación de prototipos en datasets pequeños.

Método	**Red.**	**train Acc.**	**train Kap.**	**tst Acc.**	**tst Kap.**	**Time (s)**
SPGC	0.9875	0.8270	0.6586	0.8134	0.6336	3.4719
PSO	0.9491	0.8238	0.6791	0.7501	0.5332	42.3168
GENN	0.1862	0.8002	0.6243	0.7564	0.5400	1.4285
ICPL2	0.8371	0.8254	0.6690	0.7560	0.5366	163.9147
ENPC	0.7220	0.8247	0.6800	0.7160	0.4818	47.1377

Cuadro 4.3: Resultados promedio obtenidos por los métodos de generación de prototipos en datasets grandes. (Por razones de complejidad temporal, el algoritmo ICPL2 no se puede ejecutar en grandes conjuntos de datos

Método	**Red.**	**train Acc.**	**train Kap.**	**tst Acc.**	**tst Kap.**	**Time (s)**
SPGC	0.9996	0.9243	0.6945	0.9216	0.7019	657.3117
PSO	0.9799	0.8158	0.6173	0.8000	0.5861	909.9820
GENN	0.1576	0.8428	0.6806	0.8133	0.6269	167.4849
ENPC	0.8205	0.8809	0.7613	0.8029	0.6170	10931.1977

En general, el método propuesto muestra una precisión competitiva en comparación con otros métodos de generación de prototipos, destacándose especialmente en datasets con clases no lineales.

Capítulo 5

Conclusiones y Trabajo Futuro

5.1. Conclusiones

En este capítulo se presentan las principales conclusiones derivadas del desarrollo y experimentación del método **SPGC (Spectral Prototype Gaussian Classifier)**. A lo largo de la investigación, se ha demostrado que el uso de *Agrupamiento Espectral* y *distribuciones gaussianas* mejora significativamente la generación de prototipos y la precisión en tareas de clasificación.

5.1.1. Resumen de los Resultados

A continuación, se resumen los principales logros obtenidos a través de la implementación del método propuesto:

- **Mejora en la precisión de clasificación**: El método *SPGC* ha demostrado un rendimiento competitivo frente a otros métodos de clasificación, alcanzando una precisión significativamente mayor en datasets con clases no lineales, como el *Banana dataset*. Esto se debe a la capacidad del método para capturar las subestructuras internas mediante agrupamiento espectral.

- **Reducción de la complejidad computacional**: A pesar de que el método propuesto involucra el uso de distribuciones gaussianas y *Agrupamiento Espectral*, los

experimentos demostraron que *SPGC* es capaz de reducir de manera eficiente la cantidad de prototipos necesarios para la clasificación, manteniendo al mismo tiempo un buen rendimiento en términos de exactitud y Kappa.

- **Flexibilidad del método**: *SPGC* mostró ser un método flexible, capaz de adaptarse a datasets con diferentes características, incluyendo aquellos con alta dimensionalidad, clases no lineales y ruido. En particular, en datasets como *Coil2000* y *Magic*, el método mantuvo un rendimiento robusto a pesar de la complejidad de los datos.

5.1.2. Comparación con otros métodos

La comparación con los métodos estudiados por *Triguero et al. (2012)* ha permitido evaluar el método propuesto frente a técnicas clásicas y avanzadas de generación de prototipos. Los resultados obtenidos indican que *SPGC* ofrece una mejora significativa en la clasificación de datasets complejos, superando a varios de los métodos tradicionales, especialmente en datasets con estructuras internas complejas y alta dimensionalidad.

5.1.3. Contribuciones del Trabajo

Las principales contribuciones de este trabajo son:

- El desarrollo de un nuevo método, *SPGC*, que combina *Agrupamiento Espectral* con distribuciones gaussianas para generar prototipos representativos de subclases.

- La implementación de un criterio de similitud basado en la verosimilitud gaussiana, que proporciona una medida de similitud más precisa para la clasificación en comparación con las distancias métricas tradicionales.

- La validación del método mediante una serie de experimentos comparativos con otros enfoques bien establecidos en el campo de la clasificación, demostrando su eficiencia y precisión.

5.2. Líneas Futuras de Investigación

Este trabajo abre diversas posibilidades de investigación futura, tanto en la mejora del método como en su aplicación en nuevos dominios y escenarios más complejos. A continuación se plantean algunas líneas de trabajo futuro:

- **Extensión del método a problemas de clasificación multi-clase**: Aunque el método ha mostrado buenos resultados en problemas de múltiples clases, futuros trabajos pueden centrarse en la optimización del rendimiento para escenarios multi-clase utilizando técnicas como el particionamiento jerárquico, que podrían mejorar la eficiencia y precisión en comparación con la versión actual del método.

- **Optimización del Agrupamiento Espectral**: Mejorar el proceso de *Agrupamiento Espectral* optimizando la selección de parámetros, como el número de vecinos cercanos y la función de afinidad, para maximizar la eficiencia en conjuntos de datos de gran tamaño.

- **Reducción adicional del costo computacional**: A pesar de los buenos resultados obtenidos, el tiempo de procesamiento en algunos datasets grandes, como *Magic*, puede ser optimizado utilizando técnicas de paralelización y algoritmos más eficientes para el cálculo de distribuciones gaussianas.

- **Aplicación en dominios específicos**: Explorar la aplicación de *SPGC* en otros campos, como el análisis de imágenes médicas o la clasificación de secuencias genéticas, donde la complejidad de los datos puede beneficiarse del enfoque basado en prototipos espectrales.

- **Incorporación de técnicas de aprendizaje semi-supervisado**: Una interesante línea de investigación será investigar cómo el método *SPGC* puede ser adaptado para trabajar con datasets etiquetados parcialmente, utilizando técnicas de aprendizaje semi-supervisado o auto-supervisado.

5.2.1. Conclusión Final

El método propuesto, *SPGC*, ha demostrado ser una herramienta eficiente y robusta para la clasificación en datasets con características complejas. A lo largo de este trabajo, se ha mostrado que la combinación de técnicas de *Agrupamiento Espectral* y distribuciones gaussianas permite generar prototipos que capturan mejor la estructura interna de los datos, mejorando la precisión y reduciendo el costo computacional en comparación con los métodos tradicionales. Se espera que este trabajo inspire futuras investigaciones y aplicaciones en una amplia gama de problemas de clasificación.

Bibliografía

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Cunningham, J. P., & Ghahramani, Z. (2015). Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *Journal of Machine Learning Research*, *16*, 2859-2900.

Escalante, H. J., Montes, M., Sucar, L. E., et al. (2009). Particle Swarm Model Selection. *Journal of Machine Learning Research*.

García, S., Derrac, J., Cano, J. R., & Herrera, F. (2012). Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

García, S., Luengo, J., & Herrera, F. (2015). Data preprocessing in data mining. *Springer*.

Law, M. T., Urtasun, R., & Zemel, R. S. (2017). Deep Spectral Clustering Learning. *Proceedings of the 34th International Conference on Machine Learning (ICML) / NeurIPS 2017*, 1985-1994.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On Spectral Clustering: Analysis and an Algorithm. *Advances in Neural Information Processing Systems*, *14*, 849-856.

Olvera-López, J. A., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., & Kittler, J. (2010). A review of instance selection methods. *Artificial Intelligence Review*.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888-905.

Triguero, I., García, S., & Herrera, F. (2012). A taxonomy and experimental study on prototype generation for nearest neighbor classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(1), 86-100.

Xu, R., & Wunsch, D. (2008). Clustering. *IEEE Transactions on Neural Networks*.

Yang, M.-S., Lai, C.-Y., & Lin, C.-Y. (2012). A robust EM clustering algorithm for Gaussian mixture models. *Computers and Mathematics with Applications*.

Zhang, Y., et al. (2019). Deep Gaussian Mixture Models and their Applications. *arXiv preprint*.

Zhou, S., Liu, X., Liu, J., Guo, X., Zhao, Y., Zhu, E., Zhai, Y., Yin, J., & Gao, W. (2020). Multi-View Spectral Clustering with Optimal Neighborhood Laplacian Matrix.

Víctor Ricardo Ávila Luna

GENERACIÓN DE PROTOTIPOS PARA LA CLASIFICACIÓN SUPERVISADA USANDO AGRUPAMIENTO ESPECTRAL Y DIST…

Universidad Michoacana de San Nicolás de Hidalgo

Detalles del documento

Identificador de la entrega
trn:oid:::3117:525869545

Fecha de entrega
10 nov 2025, 10:47 a.m. GMT-6

Fecha de descarga
10 nov 2025, 10:53 a.m. GMT-6

Nombre del archivo
GENERACIÓN DE PROTOTIPOS PARA LA CLASIFICACIÓN SUPERVISADA USANDO AGRUPAMIENTO ….pdf

Tamaño del archivo
612.8 KB

47 páginas

9423 palabras

52.911 caracteres

14% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Fuentes principales

14% Fuentes de Internet

6% Publicaciones

0% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alerta de integridad para revisión

Caracteres reemplazados
24 caracteres sospechosos en N.º de páginas
Las letras son intercambiadas por caracteres similares de otro alfabeto.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Formato de Declaración de Originalidad y Uso de Inteligencia Artificial

Coordinación General de Estudios de Posgrado
Universidad Michoacana de San Nicolás de Hidalgo

A quien corresponda,

Por este medio, quien abajo firma, bajo protesta de decir verdad, declara lo siguiente:

- Que presenta para revisión de originalidad el manuscrito cuyos detalles se especifican abajo.
- Que todas las fuentes consultadas para la elaboración del manuscrito están debidamente identificadas dentro del cuerpo del texto, e incluidas en la lista de referencias.
- Que, en caso de haber usado un sistema de inteligencia artificial, en cualquier etapa del desarrollo de su trabajo, lo ha especificado en la tabla que se encuentra en este documento.
- Que conoce la normativa de la Universidad Michoacana de San Nicolás de Hidalgo, en particular los Incisos IX y XII del artículo 85, y los artículos 88 y 101 del Estatuto Universitario de la UMSNH, además del transitorio tercero del Reglamento General para los Estudios de Posgrado de la UMSNH.

Datos del manuscrito que se presenta a revisión		
Programa educativo	Maestría en Ciencias en Ingeniería Eléctrica	
Título del trabajo	Generación De Prototipos Para Clasificación Usando Agrupamiento Espectral Y Distribuciones Gaussianas	
	Nombre	**Correo electrónico**
Autor/es	Victor Ricardo Avila Luna	0702823@umich.mx
Director	Jaime Cerda Jacobo	jaime.cerda@umich.mx
Codirector		
Coordinador del programa	Norberto García Barriga	norberto.garcia@umich.mx

Uso de Inteligencia Artificial		
Rubro	**Uso (sí/no)**	**Descripción**
Asistencia en la redacción	No	

Formato de Declaración de Originalidad y Uso de Inteligencia Artificial

Coordinación General de Estudios de Posgrado
Universidad Michoacana de San Nicolás de Hidalgo

Uso de Inteligencia Artificial		
Rubro	**Uso (sí/no)**	**Descripción**
Traducción al español	No	
Traducción a otra lengua	No	
Revisión y corrección de estilo	No	
Análisis de datos	No	
Búsqueda y organización de información	No	
Formateo de las referencias bibliográficas	No	
Generación de contenido multimedia	No	
Otro	No	

Datos del solicitante	
Nombre y firma	Victor Ricardo Avila Luna
Lugar y fecha	Morelia, Michoacán a 10 de Noviembre de 2025