DISEÑO DE UNA ESTRATEGIA DE RECONOCIMIENTO DE PATRONES EN UN ESCENARIO DE MÚLTIPLES EXPERTOS

IVÁN DARÍO GUSTIN SACANAMBUY

MAURICIO BOLAÑOS LEDEZMA

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE ELECTRÓNICA
SAN JUAN DE PASTO
2017

DISEÑO DE UNA ESTRATEGIA DE RECONOCIMIENTO DE PATRONES EN UN ESCENARIO DE MÚLTIPLES EXPERTOS

IVÁN DARÍO GUSTIN SACANAMBUY

MAURICIO BOLAÑOS LEDEZMA

TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE INGENIERO ELECTRÓNICO

DIRECTOR

PhD. DIEGO HERNÁN PELUFFO ORDÓÑEZ

INGENIERO ELECTRÓNICO

UNIVERSIDAD DE NARIÑO
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE ELECTRÓNICA
SAN JUAN DE PASTO
2017

NOTA DE RESPONSABILIDAD

“La Universidad de Nariño no se hace responsable por las opiniones o resultados obtenidos en el presente trabajo y para su publicación priman las normas sobre el derecho de autor.”

Acuerdo 1. Artículo 324. Octubre 11 de 1966, emanado del honorable Consejo Directivo de la Universidad de Nariño.

NOTA DE ACEPTACIÓN:

_____
_____
_____
_____
_____
_____

_____
Firma del presidente del jurado

_____
Firma del jurado

_____
Firma del jurado

San Juan de Pasto, 24 de febrero de 2017

DEDICATORIA

“A mi madre, que sin duda alguna es mi motor mi guía; la luz que me mantiene en el camino”

IVÁN DARÍO GUSTIN SACANAMBUY

“Es quizá mi sentimiento algo majestuoso, inefable. Son mis palabras llenas de amor a quien dejó la vida por verme aquí ya placido ya fuerte ya en pie. Creo saber Papá que lo que aquí está escrito refleja en gran parte el fruto de tu lucha incansable por la vida. Además de eso igual comprendí que la familia es lo único que se queda cuando todo se va. Para esas 3 mujeres: madre, hermana y sobrina mi vida de lo que queda de vida”.

MAURICIO BOLAÑOS LEDEZMA

AGRADECIMIENTO

“Agradezco a Dios por concederme la fuerza necesaria para seguir adelante, por ser mi guía en los momentos difíciles y por darme fortaleza para nunca desistir. Le doy gracias a mis padres, por todo el apoyo incondicional que siempre me brindaron a lo largo de este arduo camino, por todos los valores que me ha infundido y por su por su eterna paciencia y comprensión. A mis hermanos porque de alguna u otra manera siempre me acompañaron y me brindaron ánimos. Agradezco sentidamente al profesor Diego Peluffo por compartir sus conocimientos conmigo pero ante todo por brindarnos su amistad, gracias profe por mostrarnos el cautivador mundo de la investigación”. Les doy las gracias al grupo de estudio “multi-labeller” por su colaboración en las discusiones académicas que aportaron en el desarrollo de este trabajo. Por ultimo a todas aquellas personas que de alguna u otra manera colaboraron e hicieron esto posible”.

IVÁN DARÍO GUSTIN SACANAMBUY

“Agradecer es quizá ingrato, lo que se alcanza no solo depende de quién te da una mano sino también de quien te la niega porque aunque contradictorio eso también ayuda. Y es que no hay palabras, el mundo es grande pero el corazón pequeño y se remite a decir: gracias a Dios, madres, siempre tuve dos, y lo que ellas reflejan, amor de toda una familia. Profe Diego Peluffo, la grandeza está en la humildad, esa fue su me mejor enseñanza. David Imbajoa, Andres Arcineigas sin su ayuda no hubiese sido posible.”

MAURICIO BOLAÑOS LEDEZMA

CONTENIDO

Pág.

DEDICATORIA 5
AGRADECIMIENTO 6
CONTENIDO 7
LISTA DE FIGURAS 10
LISTA DE TABLAS 13
LISTA DE ANEXOS 14
RESUMEN 15
ABSTRACT 16
1. INTRODUCCIÓN 17
1.1. JUSTIFICACIÓN 18
1.2. CONTRIBUCIONES DE ESTA TESIS 19
1.3. ORGANIZACIÓN DEL DOCUMENTO 20
2. OBJETIVOS 21
2.1. OBJETIVO GENERAL 21
2.2. OBJETIVOS ESPECÍFICOS 21
3. MARCO TEÓRICO 22
3.1. INTELIGENCIA ARTIFICIAL 23
3.1.1 Minería de datos y aprendizaje automático. 24
3.1.2 Reconocimiento de patrones. 24
3.2. CLASIFICACIÓN 25
3.2.1 Clasificación no supervisada 25
3.2.2 Clasificación supervisada. 26
3.2.3 Tipología de algoritmos para clasificación supervisada 28
3.2.4 Mezcla de Clasificadores 30
3.3 ESCENARIOS DE MÚLTIPLES EXPERTOS 30

4 METODOLOGÍA 32
4.2 MÉTODOS DE CLASIFICACIÓN 33
4.2.1 Máquinas de soporte vectorial (SVM) 33
4.3 MÉTODOS CONVENCIONALES DE COMBINACIÓN PARA CLASIFICADORES 36
4.4 MODELO MATEMÁTICO PARA LA MEDIA ARITMÉTICA 36
4.5 MODELOS PARA EL CÁLCULO DE LOS FACTORES DE PONDERACIÓN 37
4.5.1. Modelo matemático propuesto para el cálculo de los factores de ponderación basado en matrices kernel 38
4.5.2 Agrupación basada en centroides aplicando optimización multi-criterio con algoritmos genéticos 41
4.6 DISEÑO DEL SIMULADOR PARA MÚLTIPLES EXPERTOS 45
4.7 MARCO EXPERIMENTAL 45
4.7.1. Base de datos Iris 45
4.7.2. Simulación de múltiples etiquetadores 46
4.7.3. Estudio comparativo de técnicas de clasificación 49
4.7.4. Estudio comparativo de técnicas de combinación para clasificadores 49
4.7.5. Estudio de desempeño del algoritmo de la media ponderada con los factores de ponderación generados por los métodos de matrices kernel y agrupación basada en centroides 50
4.8 MEDIDAS DE DESEMPEÑO 51
5 RESULTADOS Y DISCUSIÓN 53
5.1. PRUEBA DE LOS MÉTODOS DE CLASIFICACIÓN 53
5.2. PRUEBA DE MÉTODOS DE COMBINACIÓN DE CLASIFICADORES SVM 54
5.3. PRUEBA DE DESEMPEÑO DE LOS MÉTODOS PROPUESTOS BASADOS EN MATRICES KERNEL Y AGRUPACIÓN BASADA EN CENTROIDES 56

5.3.1. Método basado en matrices kernel 56

5.3.2. Método para el cálculo de η en la agrupación basada en centroides 59

5.4. SIMULADOR DE RECONOCIMIENTO DE PATRONES PARA MÚLTIPLES EXPERTOS 62

5.5. DISCUSIÓN 63

6 CONCLUSIONES Y TRABAJO FUTURO 64

6.1. CONCLUCIONES 64

6.2. TRABAJO FUTURO 65

RECOMENDACIONES 66

REFERENCIAS 67

ANEXOS 71

ANEXO 1. ARTICULO DE CONGRESO INTERNACIONAL CIARP 71

ANEXO 2. ARTICULO DE CONGRESO INTERNACIONAL INCISCOS 80

I. INTRODUCTION 80

II. RELATED WORKS AND BACKGROUND 81

III. PROPOSED MULTI-LABELER CLASSIFICATION APPROACH 81

A. Multi-labeler approach 81

B. Grouping based on centroids 82

C. Genetic algorithm for weights estimation 82

IV. RESULT AND DISCUSSION 83

V. CONCLUSION AND FUTURE WORK 85

ANEXO 3. MANUAL DE USUARIO 86

ANEXO 4. PAGINA WEB 96

ANEXO 5. ESTUDIO COMPARATIVO PARA LOS MÉTODOS DE MEZCLA UTILIZANDO TODOS LOS CLASIFICADORES CONVENCIONALES 97

LISTA DE FIGURAS

Figura 1. Clasificación de múltiples expertos frente a múltiples clases donde 1,2,3 y 4 son las fronteras de decisión generadas por los expertos y las figuras geométricas representan las múltiples clases. 18

Figura 2. Aplicaciones inmersas en el marco de la ciencia computacional donde se resaltan las herramientas de inteligencia artificial a partir del aprendizaje de máquina en conjunto con la minería de datos. 23

Figura 3. Proceso de agrupación con métodos de clasificación no supervisada en un espacio de características donde $\mathbf{x_1}$ y $\mathbf{x_2}$ representan dos características intrínsecas de los datos en un espacio bidimensional. Fuente [21]..... 26

Figura 4. Esquema de los procedimientos fundamentales que hacen parte de un algoritmo de clasificación supervisada: selección, entrenamiento y validación. Fuente: [20]. 27

Figura 5. Clasificación mediante algoritmos lineales en un espacio de características donde $\mathbf{x_1}$ y $\mathbf{x_2}$ representan dos características intrínsecas de los datos en un espacio bidimensional. Fuente [21] 27

Figura 6. Ejemplo de un híperplano generado por las SVM donde se distingue un esquema bi-clase; los VsA como vectores de soporte de la clase A y los VsB como vectores de soporte de la clase B. Además, se encuentra definido el margen máximo posible entre ellos y el híperplano generado por la función objetivo..... 29

Figura 7. Representación en un espacio de características de una posible combinación de clasificadores $\mathbf{F_n}$ a partir de clasificadores individuales $\mathbf{F_1}$ y $\mathbf{F_2}$. ... 31

Figura 8. Diagrama de flujo de una estrategia de clasificación supervisada para un escenario de múltiples expertos. 32

Figura 9. Híperplano de clasificación bi-clase con los parámetros representativos de una SVM, además, se representa los valores de holgura ξi que permiten obtener un clasificador de margen suave..... 35

Figura 10. Representación grafica de una función costo aplicada a la base de datos IRIS representada con sus tres clases en dos y tres dimensiones..... 37

Figura 11. Diagrama de bloques para el cálculo de los factores de ponderación mediante el método de agrupación basada en centroides. 41

Figura 12. Ejemplo de conjuntos etiquetados separados por clases para cada etiquetador con una tasa de error del [10 20 30 40 50] porciento, respectivamente. 42

Figura 13. Ejemplo de distribuciones de probabilidad separadas por clase para cada etiquetador..... 44

Figura 14. Base de datos IRIS representada con sus tres clases, donde 1 2 y 3 representan a cada una de estas (Versicolor, Virginica y Setosa). 46

Figura 15. Base de datos IRIS generada con diferentes porcentajes de error en sus etiquetas. 47

Figura 16. Vectores de etiquetas corruptos con 10% y 20% de error respectivamente. 48

Figura 17. Diagrama de cajas y bigotes para los métodos de clasificación estudiados, donde cada caja representa un método de clasificación (LDC, QDC, SVM, Fisher, Parzen); resultando que las SVM presentaron un menor error de estimación. 54

Figura 18. Diagrama de cajas y bigotes para los métodos de combinación ensayados con clasificadores SVM, donde cada caja representa el error promedio para cada método de combinación 55

Figura 19. Diagrama barras de la desviación estándar para el conjunto de resultados de cada método de combinación; resultando que el método 2 correspondiente a la media presenta un valor mas bajo comparado con los demas métodos..... 56

Figura 20. Diagrama de cajas y bigotes para la media ponderada con los factores η generados por el método de matrices kernel y los demás métodos de combinación de clasificadores; donde M.P.Kernel representa la media ponderada haciendo uso del método kernel..... 57

Figura 21. Diagrama de cajas y bigotes para la media ponderada haciendo uso de los factores η generados por el método de agrupación basada en centroides y los demás métodos de combinación de clasificadores. 60

Figura 22. Simulador de reconocimiento de patrones en un escenario con múltiples expertos..... 62

Figura 23. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador `LDC`, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error..... 97

Figura 24. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador QDC, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error..... 98

Figura 25. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador FISHER, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error. 99

Figura 26. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador PARZEN, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error. 100

LISTA DE TABLAS

Tabla 1. Error de clasificación calculado para los métodos de la media ponderada utilizando los factores η dados por el método de matrices kernel, la media y el voto mayoritario respectivamente; resultando que en todos los casos el método propuesto presenta siempre un mejor desempeño en la clasificación. 58

Tabla 2. Valores η calculados con el método de matrices kernel para cinco diferentes etiquetadores $\{y(1), \dots, y(5)\}$. 59

Tabla 3. Error de clasificación calculado para los métodos de la media ponderada utilizando los factores η dados por el método de agrupación basada en centroides, la media y el voto mayoritario respectivamente; resultando que en todos los casos el método propuesto presenta siempre un mejor desempeño en la clasificación.. 61

Tabla 4. Valores η calculados con el método de agrupación basada en centroides para cinco diferentes etiquetadores $\{y(1), \dots, y(5)\}$. 61

LISTA DE ANEXOS

ANEXO 1. ARTICULO DE CONGRESO INTERNACIONAL CIARP 71

ANEXO 2. ARTICULO DE CONGRESO INTERNACIONAL INCISCOS..... 80

ANEXO 3. MANUAL DE USUARIO..... 86

ANEXO 4. PAGINA WEB 96

ANEXO 5. ESTUDIO COMPARATIVO PARA LOS MÉTODOS DE MEZCLA UTILIZANDO TODOS LOS CLASIFICADORES CONVENCIONALES..... 97

RESUMEN

Actualmente, existe un avance computacional significativo en estrategias de aprendizaje automático para disciplinas donde se hace necesario la extracción de información en volúmenes de datos. Típicamente, aquellas estrategias de aprendizaje son entrenadas haciendo uso de conocimientos previos, establecidos por el criterio de un solo experto o profesional. Sin embargo, en ciertos contextos tales como el diagnóstico clínico, la evaluación académica de estudiantes y la valoración de alimentos, además de la cata de vinos, se hace necesario contar con múltiples criterios debido a que en la mayoría de casos la percepción subjetiva de un experto podría no ser suficiente para generar confiabilidad en los análisis. Si bien, el acceso a múltiples opiniones disminuye las ambigüedades en los resultados de análisis, los expertos pueden incurrir en errores en la asignación de etiquetas por diferentes motivos. Por esta razón se han desarrollado métodos capaces de clasificar datos con múltiples etiquetas asignando grados de confiabilidad a los etiquetadores en función de su desempeño.

El propósito de este proyecto de tesis es diseñar e implementar una estrategia de clasificación supervisada en un ambiente de múltiples expertos, partiendo del análisis de datos con múltiples etiquetas. Para este fin, se establece un estudio comparativo de clasificadores con el fin de identificar la factibilidad de su extensión a múltiples expertos, además, se propone un enfoque novedoso a partir de una mezcla ponderada de clasificadores donde los valores de ponderación son calculados por medio de matices kernel y agrupación de centroides con algoritmos genéticos, finalmente, se elabora un simulador para presentar de manera interactiva los resultados de clasificación. Los resultados demuestran que el enfoque propuesto tiene un mejor rendimiento que los métodos convencionales de clasificación.

ABSTRACT

Currently, there is a significant computational advance in machine learning strategies for disciplines where it is necessary to extract information in data volumes. Typically, those learning strategies are trained by an expert or professional, however, in some contexts such as clinical diagnosis, students' academic assessment, food evaluation and wine testing, it is necessary to consider multiple criteria due to, in most cases, subjective perception of an expert that could not be valid enough to achieve reliability of the analysis, although the access to multiple opinions diminishes ambiguous data outcomes, experts may incur in tag assignment errors because of different reasons. As a result of this situation, methods that classify data with multiple tags, assigning reliability of grades to taggers according to their performance, have been developed.

The purpose of this thesis project is to design and implement a controlled classification strategy in a multiple expert setting, starting from data analysis with multiple tags. To accomplish this, a comparative study of classifiers is set with the intention of identifying the extension to multiple expert's feasibility. Additionally, it is proposed a novel approach from a weighted mix of classifiers where the value of weighing is calculated through kernel matrixs and clustering of centroids whit genetic algorithms. Finally, a simulator is created to show interactively the outcomes of the classification. Results show that this approach has a better output than the conventional methods of classification.

1. INTRODUCCIÓN

En la actualidad, la cantidad considerable de datos que se generan a partir de múltiples sistemas y aplicaciones, ha desarrollado el interés por descifrar información importante que se encuentra almacenada en los datos, información que debe ser procesada a partir de diversas herramientas informáticas que permiten extraerla. La necesidad de distinguir patrones existentes entre los datos ha fomentado el crecimiento de enfoques de reconocimiento patrones de tal manera que sean más autónomos e inteligentes en la toma de decisiones. La capacidad de clasificar patrones con base en la experiencia, ha hecho de este tipo de herramientas un elemento fundamental en áreas como la minera de datos que requiere una demanda considerable de aplicaciones para la extracción de conocimiento dispuesto en bases de datos. Para definir y agrupar datos con patrones en común, se han diseñado algoritmos de clasificación supervisada y no supervisada que, dependiendo de las características de los datos a clasificar, cuentan con parámetros y un funcionamiento distinto según sea el contexto [1]. En general, no existe un criterio universal que permita identificar qué modelo de clasificación es el más oportuno para determinadas aplicaciones por lo que su elección sigue siendo un problema de investigación abierta [2].

En clasificación supervisada, los algoritmos por lo general se limitan en la mayoría de las investigaciones a clasificar bases de datos con una estructura en particular, donde las características o atributos de los datos son evaluados y etiquetados por un único experto o profesional, es decir, existe un único criterio de evaluación [3]. Sin embargo, para determinados contextos, es conveniente tener en cuenta una serie de opiniones en lugar de una sola opinión, esto es, contar con la evaluación de múltiples expertos como por ejemplo el diagnóstico de la patología de un paciente realizado por varios especialistas o, contar con un equipo de docentes para evaluar el rendimiento académico de un estudiante. Así mismo, es importante tener en cuenta que las apreciaciones de diferentes especialistas al momento de evaluar un evento están exentas a variaciones unas de las otras de acuerdo a la capacidad de percepción, experiencia en el tema y nivel de educación de los especialistas. De la misma manera, es difícil garantizar que un grupo de individuos emita un concepto similar de un fenómeno en particular, ya que la apreciación subjetiva de los expertos puede incurrir en errores [4]. Por tanto, alternativas emergentes de clasificación buscan reducir la influencia de los expertos con etiquetas equívocas de los datos, esto en función de factores de ponderación como método de penalización en una posible mezcla ponderada de clasificadores. [5]

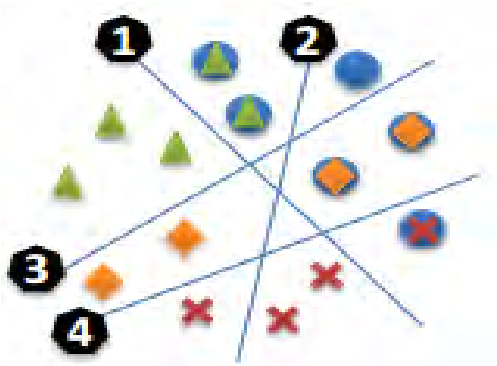


Figura 1. Clasificación de múltiples expertos frente a múltiples clases donde 1,2,3 y 4 son las fronteras de decisión generadas por los expertos y las figuras geométricas representan las múltiples clases. Fuente: [6]

En este trabajo de grado se presenta una estrategia de reconocimiento de patrones para la clasificación de datos en escenarios donde se tiene en cuenta el criterio de múltiples expertos generando múltiples etiquetas. En general, se contempla un enfoque de clasificación a partir de la mezcla ponderada de varios clasificadores entrenados, uno por conjunto de etiquetas. La estrategia en un comienzo introduce algoritmos matemáticos para la generación de factores de ponderación que determinan la importancia de los expertos al momento de clasificar, dichos algoritmos se basan en matrices Kernel y el agrupamiento basado centroides. Una vez que se han calculado, se realiza la mezcla ponderada de las funciones de costo generadas por los clasificadores entrenados. finalmente, se diseña un simulador de manera que el usuario pueda interactuar con los modelos que hacen parte de esta metodología de clasificación.

1.1. JUSTIFICACIÓN

La información se ha convertido en un elemento clave en los procesos organizacionales. En los últimos años, la tecnología ha tenido un crecimiento acelerado como herramienta útil y necesaria para facilitar dichos procesos y mejorar la productividad. La inteligencia se ha definido como la transformación de datos en conocimiento, a fin de sustentar la toma de decisiones desde el punto de vista estratégico y táctico en el momento y lugar oportuno y de generar una ventaja competitiva y de efectividad [7]. En la actualidad es evidente la necesidad de crear sistemas capaces de extraer “conocimiento” de bases de datos con el objetivo de encontrar y recuperar los datos más relevantes que se encuentran sumergidos en un conjunto de variables, y construir información que sea más comprensible e interpretable.

En diferentes entornos del reconocimiento de patrones, el origen para el análisis de datos, en muchos casos, es diverso y de considerable dificultad para la comprensión humana. Tal es el caso médico donde los estados clínicos del

paciente son diferentes y presentan una evaluación tediosa y monótona para el ser humano. No obstante, el caso médico no es el único en el que la evaluación humana se impone como la única forma de cuantificar los posibles estados de un fenómeno en particular. En general, todos aquellos eventos que son susceptibles de la evaluación sensorial de un ser humano son de difícil distinción para cualquier sistema automático. En otros escenarios, como por ejemplo los procesos de cata de alimentos y bebidas el resultado del etiquetado está ligado al carácter subjetivo que impone el factor sensorial de los evaluadores, por esta razón evidente se limita la obtención de un concepto objetivo. De nuevo es inevitable requerir el consentimiento de múltiples etiquetadores para conseguir una margen admisible de objetividad [4].

La finalidad de esta propuesta de investigación es realizar una estrategia para la clasificación de bases de datos enfocada hacia entornos de múltiples expertos que a su vez brinde resultados con una tolerancia aceptable, además, esta estrategia debe integrar un tipo de clasificador convencional eficiente para el proceso de clasificación. Finalmente se realizará un simulador en el cual se le brindará al usuario un instrumento que le permitirá tener acceso a información y conocimiento en clasificación de datos, por otra parte, el usuario dispondrá de datos e información de forma práctica y eficiente. Se debe agregar que se dispone de personal capacitado para ofrecer asesoría a lo largo del curso del proyecto debido a que se cuenta con expertos en la materia como el procesamiento de señales, reconocimiento de patrones y análisis de datos. Así mismo se tiene acceso a bases de datos y herramientas con los que se trabajará en las diferentes etapas de la metodología propuesta en el trabajo.

1.2. CONTRIBUCIONES DE ESTA TESIS

Generalmente, la clasificación de la información depende de la estructura de los datos, las características con que cuentan y las clases a los que son asignados, esto en función de la noción de un solo experto, lo que genera muchas veces ambigüedad en los resultados de clasificación, esto implica la apreciación de múltiples expertos para que la evaluación de los datos tenga una objetividad tolerable, aunque por lo general no basta con aumentar el número de expertos si no se evalúa su criterio. Se requiere entonces de un enfoque de clasificación que abarque la observación de múltiples expertos. Con el desarrollo de esta tesis se establecerá una estrategia de clasificación y análisis de datos con múltiples expertos, metodología que representa un aporte en el área de reconocimiento de patrones en términos de realizar una clasificación novedosa, esto partiendo de las estrategias preliminares de generar valores de ponderación posterior mezcla de clasificadores donde se obtienen buenos resultados de clasificación.

De igual manera, se genera un aporte al usuario donde se proporciona un simulador en el que los usuarios interactúan de manera directa con las bases de datos y su análisis. Para lograr eso, se da libre acceso a los parámetros

intrínsecos de los métodos de clasificación, así como también la elección en particular de las mezclas de los clasificadores y los enfoques a partir de los cuales se penaliza a los evaluadores.

1.3. ORGANIZACIÓN DEL DOCUMENTO

Este documento está compuesto por 7 secciones: Introducción, objetivos, marco teórico, metodología, resultados y discusión, conclusiones y recomendaciones, específicamente:

- En la sección 2, se presenta los objetivos que se plantearon como logros del desarrollo de esta investigación.

- En la sección 3, se presenta la revisión bibliográfica y un recorrido conceptual donde se aborda conceptos sobre diferentes temáticas fundamentales para el desarrollo de esta tesis tales como: Inteligencia artificial, aprendizaje automático, reconocimiento de patrones, clasificación de datos y el aprendizaje de múltiples expertos.

- En la sección 4, se describe las metodologías diseñadas y, adicionalmente, se describe las bases de datos empleadas.

- El resultado de los experimentos se discute en la sección 5.

- Finalmente, en la sección 6, se presenta las conclusiones de esta investigación, y se menciona el posible trabajo futuro.

2. OBJETIVOS

En esta sección se menciona los objetivos de esta tesis.

2.1. OBJETIVO GENERAL

Proponer un sistema de reconocimiento de patrones para realizar la clasificación supervisada de datos a partir de conocimiento a priori dado por múltiples expertos.

2.2. OBJETIVOS ESPECÍFICOS

- Desarrollar un estudio comparativo de técnicas de clasificación convencionales con el fin de identificar la más versátil e idónea para su aplicación en sistemas de múltiples expertos.

- Implementar una extensión del método de clasificación identificado, a un escenario de múltiples expertos.

- Diseñar una estrategia de clasificación supervisada de datos usando etiquetas de múltiples expertos.

3. MARCO TEÓRICO

En la actualidad, el incremento por el manejo de la información ha fomentado el interés, en particular, de campos innovadores en el tratamiento y análisis de sistemas de datos como la inteligencia artificial (IA). En este ámbito, el manejo de métodos computacionales modernos ha hecho que el procesamiento de información se realice de manera más eficiente, es por esto que las herramientas de IA se han consolidado en áreas de especial relevancia como la minería de datos o *data mining*. En efecto, existen cantidades enormes de datos donde encontramos codificada la información y donde se hace necesario tener estrategias que nos permitan extraer conocimiento a partir de sus fuentes [8]. Las bases de datos son fuente de una significativa y enorme cantidad de información potencialmente importante que no ha sido analizada. En consecuencia, surge la minería de datos como área de extracción de conocimiento a partir de herramientas de aprendizaje automático, entre ellas se encuentra algoritmos modernos que realizan un análisis minucioso de las regularidades de los datos en búsqueda de patrones y similitudes para finalmente establecer predicciones exactas sobre los datos futuros [1].

Las técnicas de reconocimiento de patrones hacen parte de esa amplia gama de estrategias de aprendizaje automático con las que se busca distinguir elementos similares o patrones reconocibles, en nuestro caso, en bases de datos. Los enfoques de estas estrategias se establecen de acuerdo a las características de los problemas de clasificación, entre los que se resalta los problemas de clasificación supervisada y de clasificación no supervisada. A su vez, se han generado herramientas en particular para cada caso, es decir, se han diseñado clasificadores con características que resuelvan las dificultades intrínsecas de estos problemas. Es así como se dispone de un sinnúmero de algoritmos en clasificación supervisada para el reconocimiento de patrones, entre los que encontramos: Clasificadores de estadísticos [10], clasificadores basados en densidad normal, clasificadores basados en estimación de Parzen [11], maquinas de soporte vectorial [12].

3.1. INTELIGENCIA ARTIFICIAL

La inteligencia artificial (IA) es una disciplina multifacética que incluye teorías computacionales con objetivos similares, todos encaminados a emular algunas facultades intelectuales propias del ser humano en sistemas inteligentes artificiales. Cuando de inteligencia humana se habla, es necesario tener en cuenta procesos relevantes relacionados con nuestros sentidos, percepciones sensoriales como la visión, audición y tacto como los procesos intuitivos de reconocimiento de patrones. Como resultado se tiene los procesos de tratamiento datos y aprendizaje e identificación de sistemas entre las aplicaciones más comunes de la IA (Ver Figura 2) [8]. En ese afán de establecer mecanismos de comportamiento y pensamiento con el fin que supone percibir, razonar, aprender, comunicarse y actuar de manera autónoma en entornos complejos, la IA surge como herramienta de comprensión del comportamiento de humanos animales u otros sistemas para desarrollar máquinas de aprendizaje que permitan, en torno a sus facultades, comportarse de igual o mejor manera que la naturaleza humana [9]. En [10] se define IA "como una rama de la informática dedicada a la creación artificial de conocimiento, es decir, una ciencia que tiene como aspiración el diseño y producción de sistemas computacionalmente inteligentes".

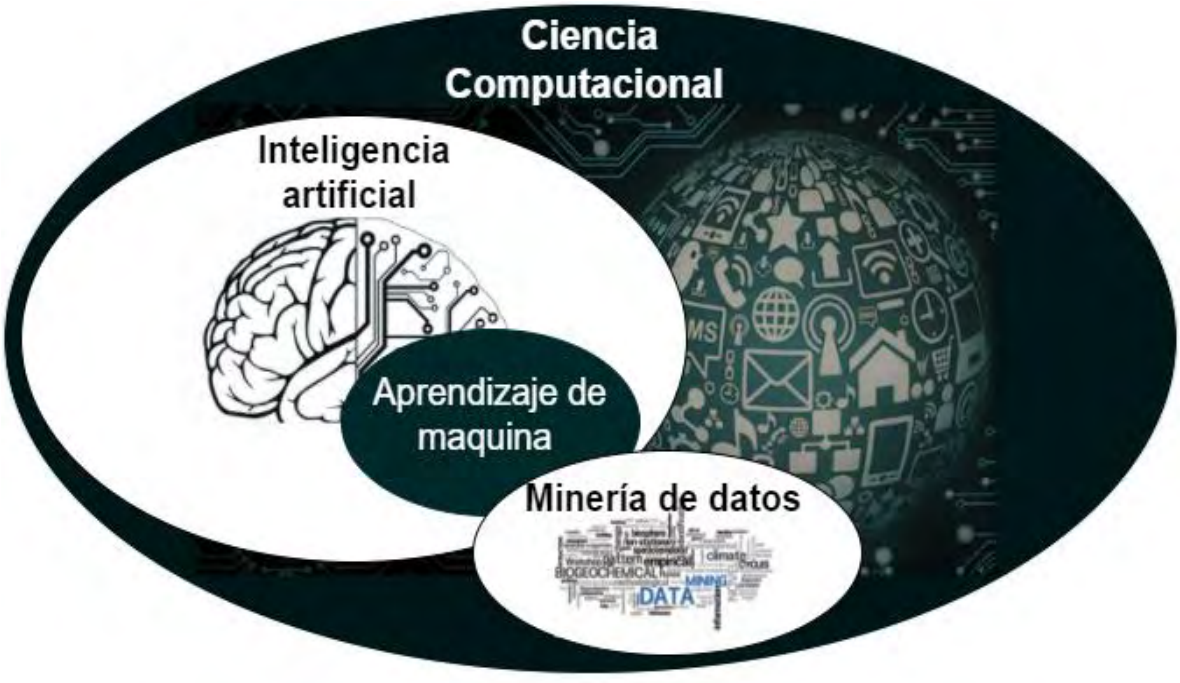


Figura 2. Aplicaciones inmersas en el marco de la ciencia computacional donde se resaltan las herramientas de inteligencia artificial a partir del aprendizaje de máquina en conjunto con la minería de datos.

3.1.1 Minería de datos y aprendizaje automático.

La minería de datos se define como un conjunto de técnicas que permiten explorar grandes volúmenes de información a través de procesos de análisis a bases de datos, esto en busca de patrones, tendencias, así como nuevas y significativas relaciones antes desconocidas. Las herramientas que proporciona la minería de datos tienen la capacidad de analizar grandes cantidades de datos con la finalidad de descubrir conocimiento almacenado en ellos [11, 12].

La minería de datos requiere acceso a los datos. Los datos pueden representarse como volúmenes de registros distribuidos de manera que se pueda diferenciar los registros que representan los vectores de características como los atributos del los datos y los registros de clases generados por un experto para etiquetar los datos [13].

El aprendizaje automático como herramienta de la minería de datos consiste en la programación de ordenadores y el manejo de elementos computacionales para el desarrollo de algoritmos capaces de generalizar comportamientos y reconocer patrones a partir de información suministrada en forma de bases de datos, con el fin de optimizar el rendimiento de determinados procesos utilizando criterios de los datos a partir de experiencias ya adquiridas. En otras palabras, el aprendizaje automático tiene la capacidad de predecir de manera eficiente y precisa el comportamiento futuro de los datos a partir de lo que ha ocurrido con ellos en el pasado. Las garantías de aprendizaje para un algoritmo están ligadas a la composición y complejidad de los datos que se va a analizar, de los múltiples atributos, de sus clases y del tamaño de la muestra de entrenamiento [14, 15].

3.1.2 Reconocimiento de patrones.

La definición elemental de reconocimiento de patrones se puede deducir a partir de una idea básica de igualdad o semejanza. Es posible relacionar a dos objetos en particular y reconocerlos con cierto grado de similitud ya que entre ellos comparten y son claros algunos atributos en común. Por lo general, cuando hablamos de similitud no siempre existe la comparación entre objetos ya que, en un sentido más abstracto, la hay entre un objeto y un concepto objetivo [16].

Se puede decir entonces que el reconocimiento de patrones es la disciplina científica relacionada con métodos para la descripción y clasificación de objetos [17]. De manera más general, como lo define el autor en [18] “El reconocimiento de patrones es la ciencia que se ocupa de los procesos sobre ingeniería, computación y matemáticas relacionados con objetos físicos y/o abstractos, con el propósito de extraer información que permita establecer propiedades de o entre conjuntos de dichos objetos, los cuales nos permitan interpretar el mundo que nos rodea”.

3.2. CLASIFICACIÓN

En reconocimiento de patrones existe la necesidad de cuestionarse acerca de lo que se va a reconocer y como se lo va a hacer. Por lo que existen las características del objeto o los a tributos de los datos como el patrón a reconocer y la clasificación como el medio para hacerlo. Se puede decir entonces que la clasificación es como lo define [8] “Una de las tareas de reconocimiento de patrones en la que queremos clasificar o agrupar a un individuo a partir de ciertas propiedades que lo caracterizan; entendiendo como individuo una entidad de cualquier tipo”. Los métodos de clasificación se han desarrollado en múltiples disciplinas donde se requiere establecer fronteras de decisión entre datos, la elección de un clasificador para cada caso depente de factores de las características de dichos datos que se encuentran en bases estructuradas de disciplinas como: medicina (diagnóstico de enfermedades), astronomía, ingeniería, control y robótica.

La construcción de un sistema de clasificación a partir de un conjunto de datos se puede definir a partir de dos conceptos fundamentales en función de la estructura de datos a clasificar como lo define [19]: “se puede tener un conjunto de observaciones con el objetivo de establecer la existencia de clases o grupos en los datos. O se puede saber que existen determinadas clases, y que el objetivo sea establecer una regla por la que se llegue a clasificar una nueva observación dentro de una de las clases existentes”. El primer caso se denomina clasificación no supervisada y el segundo clasificación supervisada.

3.2.1 Clasificación no supervisada

El Problema de clasificación no supervisada consiste en descubrir la estructura que se define en algún conjunto de datos, es decir, lo que se desea saber es si existen grupos en los datos y qué características hacen que los datos sean similares dentro del grupo y diferentes entre los grupos. Lo que diferencia este esquema de clasificación en reconocimiento de patrones es que no existen clases definidas con anterioridad [20]. En la Figura 3 se representa sobre un espacio de características, la metodología de agrupación en este tipo de clasificación.

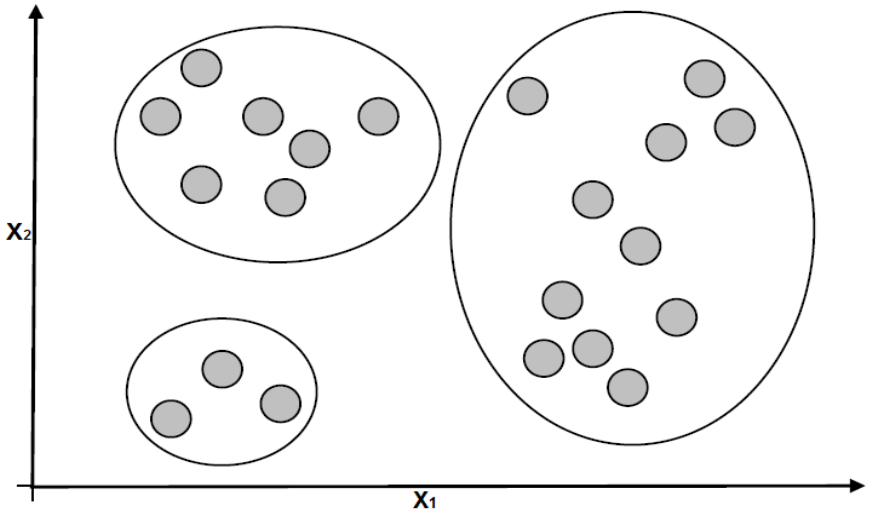


Figura 3. Proceso de agrupación con métodos de clasificación no supervisada en un espacio de características donde x_1 y x_2 representan dos características intrínsecas de los datos en un espacio bidimensional. Fuente [21]

3.2.2 Clasificación supervisada.

El problema de clasificación supervisada consta de un mecanismo que permite clasificar datos teniendo en cuenta una estructura en particular donde los atributos de los datos han sido etiquetados con anterioridad por un experto, es decir, han sido asignados a una clase. Los clasificadores a partir del conocimiento adquirido a través de los conjuntos de datos de entrenamiento generan criterios para discriminar entre las clases de datos predeterminadas. Finalmente si se quiere asignar un conjunto de datos nuevos a una clase, sólo se tiene que introducir sus características en el clasificador para que sean reconocidos [16].

La selección, el entrenamiento y la validación de un modelo clasificador constituyen el núcleo de reconocimiento de patrones. La selección tiene que ver con definir y seleccionar los datos relevantes que harán parte del proceso de clasificación ya que muchos de ellos son ruido. En entrenamiento, se introducen los datos en la herramienta de aprendizaje o clasificador para generar un modelo de acuerdo a la estructura de los datos. Finalmente en validación se introducen nuevos datos al modelo generado para que éste les asigne una etiqueta de clase [20]. Un esquema de este enfoque se puede ver en la Figura 4.

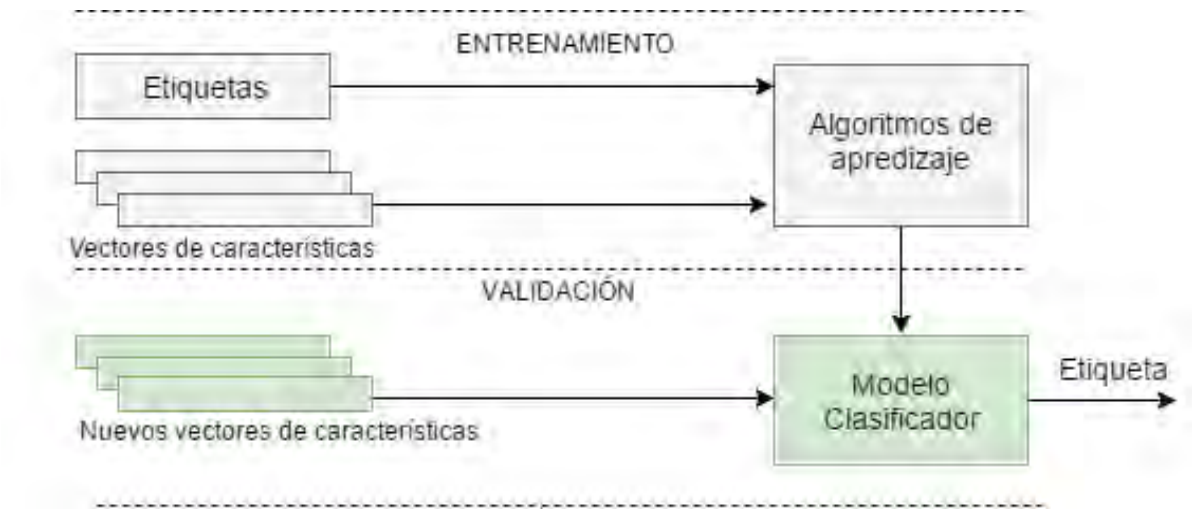


Figura 4. Esquema de los procedimientos fundamentales que hacen parte de un algoritmo de clasificación supervisada: selección, entrenamiento y validación. Fuente: [20].

En la Figura 5 se muestra sobre un espacio de características los resultados de un sistema de clasificación como el descrito anteriormente utilizando un clasificador lineal como algoritmo de aprendizaje en un proceso de clasificación supervisada.

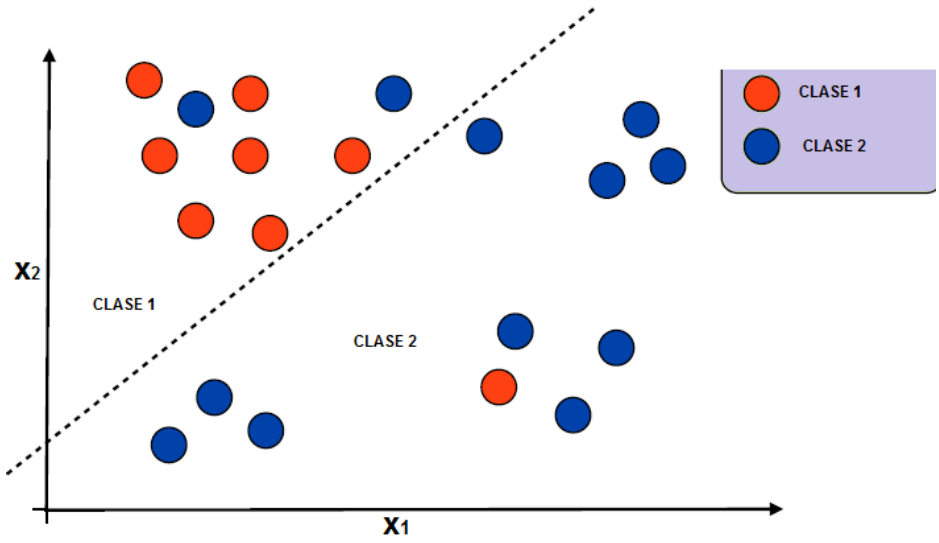


Figura 5. Clasificación mediante algoritmos lineales en un espacio de características donde $\mathbf{x_1}$ y $\mathbf{x_2}$ representan dos características intrínsecas de los datos en un espacio bidimensional. Fuente [21]

3.2.3 Tipología de algoritmos para clasificación supervisada

Para un conjunto de datos de entrenamiento que contiene diferentes clases, el objetivo es establecer límites de decisión en el espacio de características que permita separar los objetos que pertenecen a clases diferentes. Se puede clasificar los algoritmos en función de los métodos que se usan para adquirir los modelos o reglas de clasificación [8].

- **Clasificador discriminante lineal de Fisher**

  Clasificador que se caracteriza por ser uno de los procedimientos estadísticos mas utilizados por su facilidad y su vigencia desde tiempos atrás hasta la actualidad. Consiste básicamente en dividir el espacio muestral mediante una serie de lineas, planos en el caso de tres dimensiones y, en general, hiperplanos para muchas dimensiones. La línea que divide dos clases se traza de tal manera que corta por la mitad la línea que une los centros de la clases respectivas y su dirección se determina a partir de la forma de los grupos de los puntos [19].

- **Clasificadores basados en estimación de densidad Normal**

  Este tipo de clasificadores basados en estimaciones de densidad, se derivan de la base matemática de clasificación con mínimo de errores (Bayes) para clases normalmente distribuidas. En general, existen dos modelos de clasificación de interés: el clasificador discriminante lineal (LDC de sus siglas en ingles “Linear Discriminant Classifier”), para clases normalmente distribuidas con matrices de covarianza iguales, y el clasificador discriminante cuadrático (QDC de sus siglas en inglés “Quadratic Discriminant Classifier”), donde las clases son caracterizadas por matrices de covarianza diferentes. Estos clasificadores derivan su nombre del tipo de funciones discriminantes que utilizan. Son de funcionamiento razonablemente robusto, y generan buenos resultados al clasificar, incluso, clases que no tienen distribuciones normales [20].

- **Clasificador basado en estimación de Parzen**

  El diseño del clasificador basado en estimación de densidades de probabilidad de Parzen se realiza bajo la consideración de que se desconoce cualquier función de probabilidad antes establecida. Como objetivo principal, la estimación de Parzen se centra en la obtención de estimaciones de densidades de probabilidad condicional de las que se tiene poco conocimiento a priori al momento de realizarlas. Para obtener dichas estimaciones lo que se hace es una partición del espacio de

características de manera que se obtenga un numero finito de regiones denominadas cajas, finalmente se realiza un conteo de las muestras de cada caja siendo este número la estimación esperada [22].

- **Máquinas de soporte vectorial (SVM)**

Las maquinas de soporte vectorial (SVM por sus siglas en inglés “Support Vector Machine”) son una herramienta de aprendizaje supervisado utilizado en un principio para clasificación binaria, esto es, clasificación bi-clase. Básicamente se cuenta con un algoritmo de optimización que permite encontrar y seleccionar un híperplano de todos los posibles, con el máximo margen de separación entre las clases, de manera que haya una separación perfecta entre ellas. Esto sería para el caso donde las clases de los datos sean perfectamente diferenciables [23].

Sobre la Figura 6, se establece los conceptos propios de las SVM, se tienen dos clases representadas en figuras geométricas y definidas como clase A y clase B. para definir el híperplano se encuentran los vectores de soporte VsA y VsB que por lo general son los más cercanos a él, como el híperplano de margen máximo generado por las SVM para este caso [8].

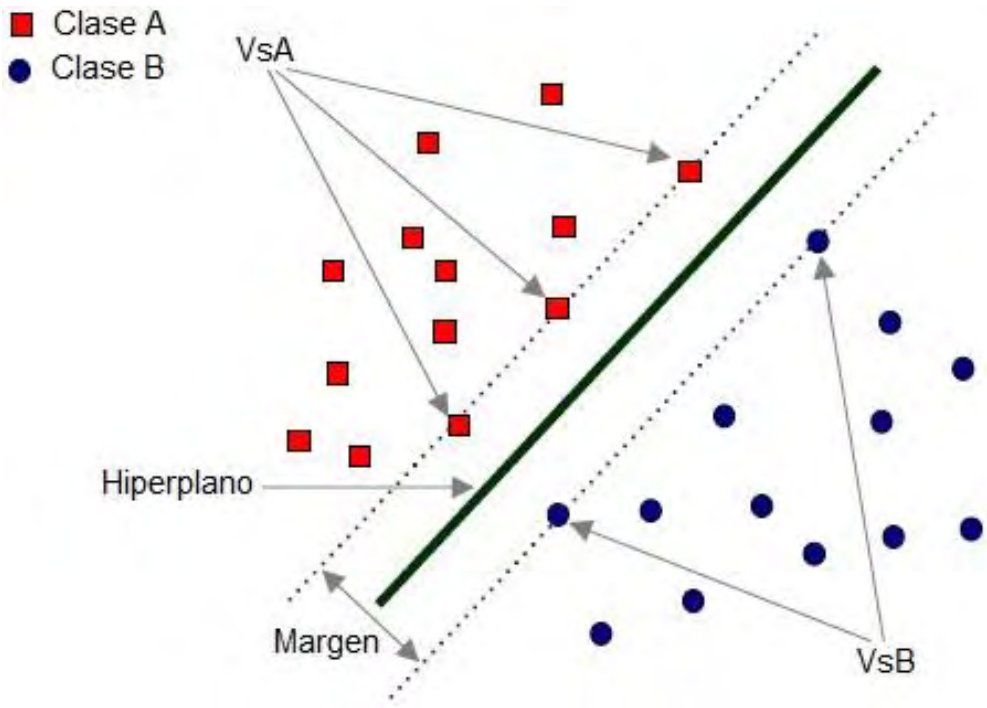


Figura 6. Ejemplo de un híperplano generado por las SVM donde se distingue un esquema bi-clase; los VsA como vectores de soporte de la

clase A y los VsB como vectores de soporte de la clase B. Además, se encuentra definido el margen máximo posible entre ellos y el híperplano generado por la función objetivo.

3.2.4 Mezcla de Clasificadores

De acuerdo con lo mencionado anteriormente, la selección de los clasificadores se definen en función de los requerimientos de clasificación. Si bien, estos métodos tienen un buen desempeño, el objetivo principal en un sistema de reconocimiento de patrones aplicado en bases de datos es alcanzar un margen bueno de clasificación. Una solución factible es combinar la información de múltiples clasificadores, lo que ha motivado el interés a probar con mezcla de clasificadores.

En un principio, la idea es desconfiar de una única solución. En su lugar, todos los diseños de clasificadores se utilizan para la toma de decisiones combinando sus opiniones individuales, para derivar en una decisión de consenso. En general, lo que se pretende, a partir de la base fundamental de una mezcla de clasificadores, es que se obténga una mayor precisión en la clasificación de la que se obtendría por cada uno de ellos. [19, 24].

Para realizar combinación de clasificadores existen varios enfoques, todos ellos con la finalidad de obtener buenos resultados en determinados escenarios. La regla del producto, la regla de la suma, la regla min, la regla Max, la regla de la mediana y el voto mayoritario y la media ponderada son algunos que se prueban en este trabajo [24].

3.3 ESCENARIOS DE MÚLTIPLES EXPERTOS

En la actualidad, no es nuevo encontrar datos asociados con un análisis generado por diferentes expertos o evaluadores, esto debido al incremento considerable de las bases de datos y las múltiples facetas en las que se encuentra predispuesta la información. En una de esas facetas se encuentran los escenarios de múltiples expertos que se caracterizan principalmente por disponer de una estructura de datos organizada de tal manera que se cuente con un conjunto de vectores de etiquetas generado por múltiples expertos.

Este tipo de análisis ha sido motivado en gran parte por la necesidad de encontrar una solución factible a problemas donde existen diferentes evaluaciones en base a un fenómeno en particular, algo que ocurre con frecuencia en los entornos médicos [25], donde los especialistas difieren en su análisis unos de otros por diferentes circunstancias: o los especialistas no cuentan con el conocimiento apropiado o los especialistas establecen sesgos en la evaluación por incurrir en

errores de percepción [25, 4]. De la misma manera, los contextos académicos se relacionan a los escenarios multietiquetador cuando se cuenta con la evaluación de múltiples maestros, donde se indaga acerca de la calidad de cada profesor independiente del algoritmo de aprendizaje [26].Otro escenario que se destaca al momento de compartir información es aquel donde se realiza un análisis a partir de las muestras de diferentes productos ya sea alimentos o en casos particulares como la catación de vinos [27].

Los trabajos de investigación en el tema se han manejado desde diferentes perspectivas, algunas de ellas pretenden definir qué tanto más discriminante es una maquina en relación con la capacidad de evaluación humana esto con el fin de cuantificar el error al clasificar automáticamente a partir de aprendizaje de máquina [28]. Otra de las perspectivas, se relaciona generalmente con metodologías de aprendizaje supervisado y técnicas donde se hace una combiación de información de clasificadores para obtener mejores resultados de clasificación, como se observa en la Figura 7.

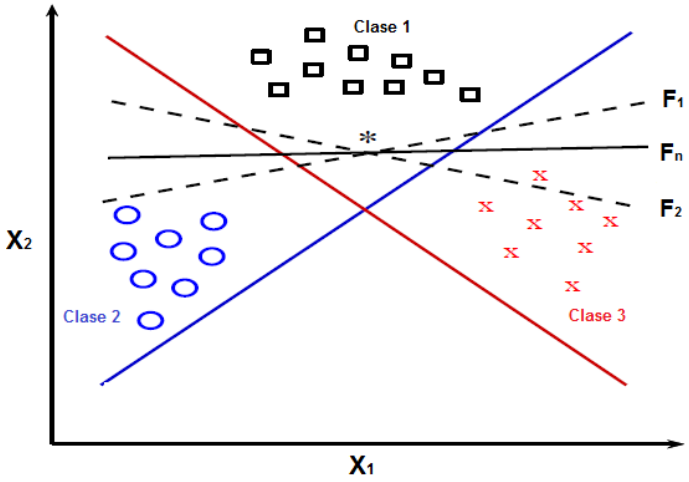


Figura 7. Representación en un espacio de características de una posible combinación de clasificadores **F_n** a partir de clasificadores individuales **F_1** y **F_2**.

Los algoritmos de clasificación basados en máquinas de soporte vectorial han mostrado ser una herramienta útil y poderosa al momento de identificar a los etiquetadores y establecer cuáles de ellos proporciona un mejor y más veraz análisis de los datos de manera que se puede lograr que su influencia sea mayor en el aprendizaje de máquina. La ventaja que ofrecen las SVMs son los vectores de soporte para la construcción del clasificador, lo que permite remover aquellos datos que no aporten a dicha construcción sin afectar o cambiar el aprendizaje de máquina [29]. En este ámbito, se propone también penalizar a los etiquetadores de manera que disminuye la influencia de los mismos al momento de entrenar los clasificadores [4, 5, 3].

4 METODOLOGÍA

La estrategia de clasificación se estructura en tres procedimientos fundamentales: estudio comparativo de métodos de clasificación para determinar cuál o cuáles clasificadores son los más pertinentes para el enfoque de múltiples expertos, implementación de algoritmos para estimar los valores de ponderación que cuantifican el desempeño de los expertos, implementación de una mezcla ponderada de clasificadores con los valores de ponderación ya obtenidos. Finalmente se realiza una interfaz a modo de simulador de manera que se pueda vislumbrar los resultados de clasificación.

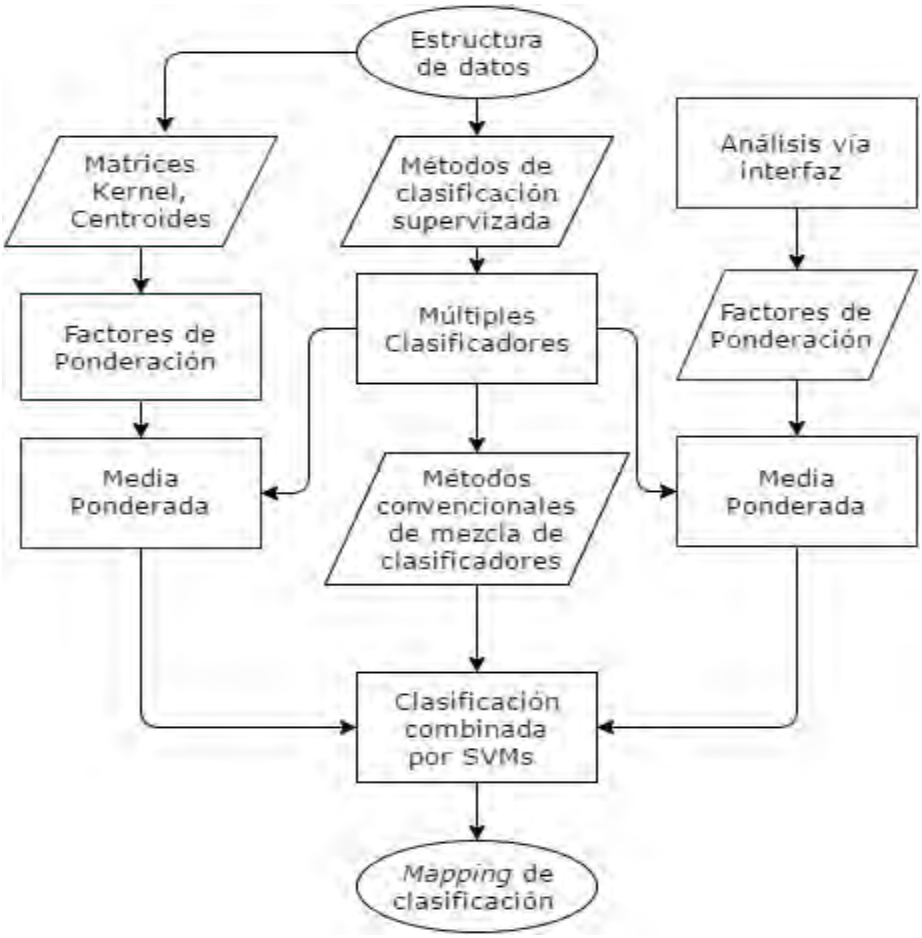


Figura 8. Diagrama de flujo de una estrategia de clasificación supervisada para un escenario de múltiples expertos.

4.2 MÉTODOS DE CLASIFICACIÓN

El objetivo final de diseñar diversos sistemas de clasificación es lograr conseguir un buen desempeño para la tarea que se está realizando. Este objetivo ha llevado tradicionalmente al desarrollo de diferentes esquemas de clasificación para cualquier problema a ser resuelto. Se debe considerar entonces los resultados de una evaluación experimental de los diferentes diseños para elegir uno de los clasificadores.

Para el desarrollo de esta tesis se ponen en consideración cuatro modelos de clasificación basados en métodos probabilísticos: clasificador discriminante lineal de Fisher [30], clasificador discriminante lineal [20], clasificador discriminante cuandrático [20], clasificador por estimación de parzen [31]. Además, se estudia el método basado en distancias de máquinas de soporte vectorial.

4.2.1 Máquinas de soporte vectorial (SVM)

Dada su versatilidad y excelente desempeño en varias aplicaciones, muchos enfoques de SVMs se han formulado para hacer frente a los problemas de múltiples etiquetas, en este sentido las SVM proporcionan una herramienta matemática sólida para clasificación en diferentes escenarios [32]. Es así como se ha seleccionado una serie de procedimientos que se hacen necesarios para el desarrollo de nuestro trabajo. En primer lugar se pondrá en consideración los clasificadores SVM para ejemplos bi-clase, para posteriormente examinar un enfoque para ejemplos de múltiples clases.

Para definir un clasificador bi-clase, se contempla una pareja ordenada $\{\boldsymbol{x}_i, \bar{\boldsymbol{y}}_i\}$ como la i-esima muestra donde $\boldsymbol{x}_i \in \mathbb{R}^{\boldsymbol{d}}$ y representa el vector de características con dimensión d, y $\bar{\boldsymbol{y}}_i$ es la etiqueta de clase correspondiente a la muestra i , tal que $\bar{\boldsymbol{y}}_i \in \{1, -1\}$. En términos matriciales, tendríamos $\boldsymbol{X} \in \mathbb{R}^{\boldsymbol{mxd}}$ y $\bar{\boldsymbol{y}}_i \in \mathbb{R}^{\boldsymbol{m}}$ donde d es el número de características y m es el número de muestras. Además, para la función objetivo que representa un híperplano en $\mathbb{R}^{\boldsymbol{d}}$, se asume un modelo de la forma:

$$e_i = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + b = \langle \boldsymbol{x}_i, \boldsymbol{w} \rangle + b, \qquad (1)$$

donde $\boldsymbol{w}$ es un vector d-dimensional, b una constante de sesgo y la notación $\langle \cdot, \cdot \rangle$. representa el producto interno euclidiano. Asumiendo como frontera de decisión el plano anterior para un clasificador bi-clase, y considerando que los puntos mayores a dicho híper-plano correspondan a una clase y los menores a otra clase, se establece el siguiente criterio de asignación de clase:

$\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + b > 0\,,\ \bar{\boldsymbol{y}}_i = 1$
$\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + b < 0\,,\ \bar{\boldsymbol{y}}_i = -1$

$\boldsymbol{w^T x_i} + b = 0$, $\boldsymbol{\bar{y}_i} = 0$, $\boldsymbol{x_i}$ pertenence a la frontera.

Por tanto, una función de decisión se puede escribir de la forma:

$$\boldsymbol{f(x)} = sign(e_i) = 0, \tag{2}$$

Además, para evitar que los puntos de los datos se encuentren en una región de ambigüedad o sobre la función de decisión, la distancia entre el híperplano y cualquier punto puede ser restringida, de manera que cualquiera de ellos cumpla con la condición $\boldsymbol{\bar{y}_i} e_i \geq 1, \ \forall i$, además la distancia entre los puntos de los datos $\boldsymbol{x_i}$ y el hiperplano e se puede calcular como $d(e, \boldsymbol{x_i}) = \boldsymbol{\bar{y}_i} e_i / ||\mathbf{w}||^2$ donde $||.||$ es la norma euclidiana. Por lo tanto dado que el límite de $d(e, \boldsymbol{x_i})$ es $1/||\mathbf{w}||^2$ se espera que $\boldsymbol{y}_i \simeq e_i$ además la función objetivo del clasificador a maximizar se puede escribir como $\max_{w} \boldsymbol{\bar{y}_i} e_i / ||\mathbf{w}||^2$, Asumiendo este numerador constante, la expresión puede reescribirse como un problema de minimización de la forma:

$$\min_{w} \frac{1}{2} ||\mathbf{w}||^2 \quad \text{s.a.} \quad \boldsymbol{\bar{y}_i} e_i = 1; \ \forall \boldsymbol{i} \tag{3}$$

Teniendo en cuenta que en la mayoría de casos las clases no son fácilmente separables, es necesario el uso de SVM de margen suave, donde se incluye la posibilidad de admitir errores en el entrenamiento del clasificador, esto con el fin de que la clasificación no sea tan exigente teniendo en cuenta un parámetro de holgura que permite constituir un margen más flexible al interior de los híperplanos de soporte como se observa en la Figura 9.

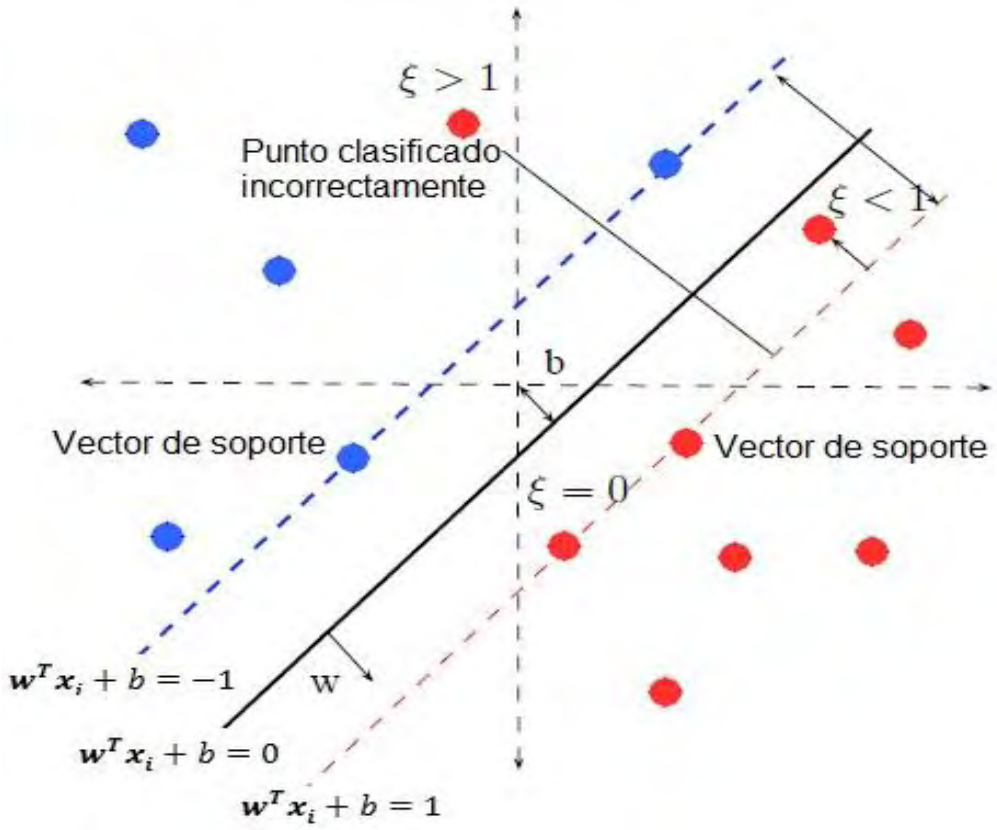


Figura 9. Híperplano de clasificación bi-clase con los parámetros representativos de una SVM, además, se representa los valores de holgura ξ_i que permiten obtener un clasificador de margen suave.

Teniendo en cuenta lo anterior se opta por una SVM de margen suave de la forma [5]:

$$\min_{w,\xi} f(w,\xi|\boldsymbol{\lambda}) = \min_{w,\xi} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{m}\sum_{i=1}^{m}\xi_i^2 ; \quad \text{s.a.} \quad \xi_i \geq 1 - \quad \boldsymbol{y}_i e_i, \tag{4}$$

donde λ es un parámetro de regularización y ξ_i es un término de holgura asociado a los puntos de los datos.

Lo que se ha hecho hasta el momento es examinar las bases para un clasificador biclase, sin embargo, en la mayoría de casos en general, los fenómenos a clasificar se dispone de múltiples clases con las que se entrena el clasificador. Con el fin de solucionar este tipo de problemas se ha desarrollado trabajos que extienden este enfoque bi-clase a múltiples clases, esto sera expicado brevemente mas adelante [3].

4.3 MÉTODOS CONVENCIONALES DE COMBINACIÓN PARA CLASIFICADORES

Dentro del ámbito del reconocimiento de patrones existe el problema en los escenarios donde se cuenta con múltiples etiquetadores, la razón de utilizar la clasificación múltiple es que al emplear un conjunto de clasificadores es posible obtener una mayor precisión que la que es capaz de lograr cualquiera de estos de una manera individual, además puesto que la mayoría de los métodos de clasificación tiene un punto fuerte o ventaja sobre algún grupo de clasificadores es necesaria unir las características más deseables de cada uno por medio de algún método de combinación [18].

Dicho lo anterior, el siguiente punto hace alusión a los métodos de combinación para clasificadores que se utilizaron en este trabajo de grado. Para esto se trabajó con diferentes enfoques de combinación para clasificadores, utilizando el Toolbox PRTools para MATLAB, entre los cuales se encuentran: la media aritmética, el máximo, el mínimo, la mediana, el producto y el voto mayoritario, detallados a fondo en [33, 24].

Con los métodos de combinación expuestos anteriormente se realizaron pruebas con los conjuntos de etiquetadores y los diferentes métodos de clasificación, obteniendo como resultado un nuevo clasificador combinado a partir de la información brindada por el grupo de clasificadores entrenados. Con este nuevo clasificador entrenado se procedió a realizar la clasificación del conjunto de datos original, observando el desempeño de cada método de combinación. Con los resultados de todos los métodos se aplicó algunas medidas de centralización con los que se siguió a realizar un análisis de estos.

4.4 MODELO MATEMÁTICO PARA LA MEDIA ARITMÉTICA

Como se mencionó anteriormente la combinación de clasificadores genera resultados aceptables en entornos donde se cuenta con varios clasificadores, de acuerdo con lo anterior se propone mejorar el enfoque de la media aritmética con la introducción de unos valores de ponderación que generarán un valor de relevancia para cada etiquetador dependiendo de cuan correctas sean sus etiquetas.

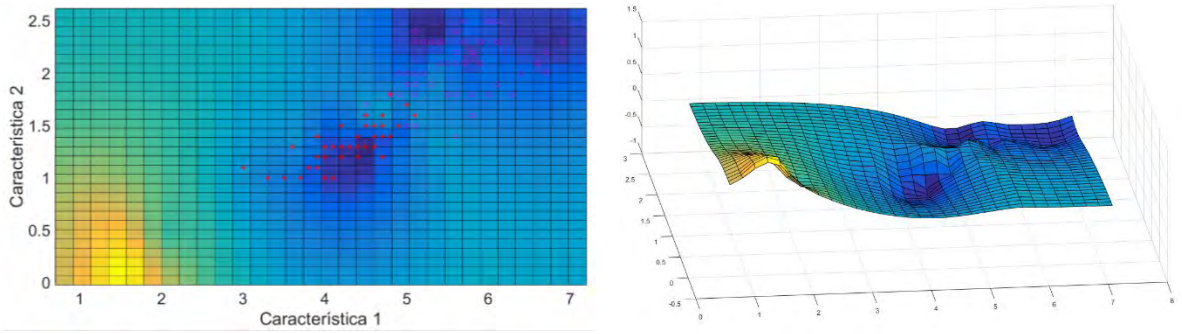
El siguiente punto trata del enfoque de combinación basado en la media ponderada, para esto es importante hablar de las medidas de tendencia central o medidas de centralización, su función es explicar cómo un conjunto de datos tiende a agruparse en las proximidades de su valor medio, un ejemplo de una medida de tendencia central es la media aritmética.

La media ponderada es considerada una medida de centralización que se construye asignándole a un dato un peso o ponderación. Enseguida se propuso realizar la extensión de este enfoque para desarrollar la combinación de clasificadores con lo que se formula un nuevo método de combinación basado en la media aritmética ponderada donde se asigna un factor de ponderación a cada clasificador.

Lo que se propone es desarrollar una combinación de clasificadores, donde se define $f^{(t)}(\boldsymbol{X})$ como la función coste entrenada usando las etiquetas dadas por el etiquetador t. Entonces, con el objetivo de aprovechar la información de todo el conjunto de etiquetadores, se propone un clasificador cuya función coste está dada por la siguiente combinación:

$$\bar{f}(\boldsymbol{X}) = \sum_{t=1}^{k} \eta_t f^{(t)}(\boldsymbol{X}), \tag{5}$$

donde η_t son los factores de ponderación que se definen en los siguientes apartados.



(a) Grafica de la función costo con la base de datos.

(b) Grafica de la función costo en 3 dimensiones.

Figura 10. Representación grafica de una función costo aplicada a la base de datos IRIS representada con sus tres clases en dos y tres dimensiones.

4.5 MODELOS PARA EL CÁLCULO DE LOS FACTORES DE PONDERACIÓN

El método de combinación basado en la media ponderada utiliza los factores de ponderación η mediante los cuales es posible inferir el grado de certeza que tiene un etiquetador. En este orden de ideas, es necesario idear un método que sea

capaz de calcular los factores η automáticamente, debido a que no se conoce cuál o cuáles etiquetadores son los que presentan un mejor conjunto de etiquetas y cuáles no.

En este trabajo se ponen en consideración dos métodos para el cálculo de los factores de ponderación: el primero por medio de un kernel supervisado, que se utiliza para construir matrices kernel (una por etiquetador), el segundo por medio del método de la agrupación basada en centroides. Cada uno de estos métodos es capaz de analizar la estructura general de los datos, generando de esta manera un valor η entre $[0,1]$ que indica cuán acertadas son sus etiquetas (cuanto mayor sea el valor de η, es más acertado).

4.5.1. Modelo matemático propuesto para el cálculo de los factores de ponderación basado en matrices kernel

Para la aplicación de un enfoque de matrices kernel, se debe hacer una extensión a datos con etiquetas multi-clase ya que se van a manejar bases de datos con más de dos clases, por lo tanto es necesario que el enfoque sea capaz de adecuarse a estos escenarios. En [3] se propone un método capaz de tratar con un número mayor de clases ℓ, donde $\ell \in \{1, \dots, c\}$ y c es el numero de clases a tomar. Entre los diferentes modelos tradicionales para múltiples clases se opta por trabajar con un enfoque basado en uno contra todos (one against all OaA). Este enfoque consiste en aplicar c numero de veces el enfoque bi-clase descrito anteriormente, la clase c es comparada con las restantes y se le da un valor de etiqueta positivo, mientras que para las otras clases el valor de etiqueta es negativo; de este modo se forma un vector de etiquetado binario individual para cada etiquetador t por cada clase $\boldsymbol{y}_{\ell i}^{(t)}$ asociado a la clase ℓ y es asumido como se muestra a continuación:

$$\boldsymbol{y}_{\ell i}^{(t)} = \begin{cases} 1 & \text{si } \mathbf{x_i} \text{ pertenece a la clase } \ell \\ -1 & \text{si no pertenece a las clase } \ell \end{cases},$$

de forma que se puede tener una matriz de etiquetas $\boldsymbol{Y}^{(\ell)}$:

$$\boldsymbol{Y}^{(\ell)} = \begin{pmatrix} \mathbf{1} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{1} & \mathbf{-1} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{1} & \mathbf{1} \\ \mathbf{-1} & \mathbf{1} & \mathbf{1} \\ \mathbf{-1} & \mathbf{-1} & \mathbf{1} \end{pmatrix}$$

Con esto, se plantea la ecuación kernel modificada por cada etiquetador, así:

$$\boldsymbol{G}_{ij}^{(t)} = \sum_{\ell=1}^{c} \boldsymbol{y}_{\ell i}^{(t)} \boldsymbol{y}'^{(t)}_{\ell j} \mathcal{K}(x_i, x_j), \tag{6}$$

donde $\mathcal{K}(x_i, x_j)$ respresenta la función kernel, que en esta tesis se considera de base radial (RBF kernel o kernel gaussiano), debido a que esta expresa una medida de semejanza entre los vectores x_i y x_j definido como:

$$\mathcal{K}(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}, \tag{7}$$

donde $\|x_i - x_j\|$ representa la disimilitud como una función de descomposición de la distancia entre los vectores. La norma elevada al cuadrado representa la distancia entre las muestras, es decir, cuando los vectores están muy cercanos entonces $\|x_i - x_j\|$ será menor, además el valor σ se despejo de la siguiente igualdad $\frac{1}{2\sigma^2} = 1$ [35], donde el valor que se determinó para σ es de 0.7, de esta manera, se limita a que la distancia calculada entre las muestras no se vea modificada por valores externos.

Para estimar los coeficientes η_t se propone usar una adaptación del enfoque de análisis de relevancia de variables aplicado a múltiples kernels abordados detalladamente [36, 37]. Se empieza definiendo una matriz $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{N^2 xk}$ donde en cada columna se encuentran la concatenación o vectorizacion de cada matriz kernel $\boldsymbol{G}^{(t)}$ para cada etiquetador. Ahora suponga que una representación de menor rango $\widehat{\boldsymbol{\mathcal{G}}} \in \mathbb{R}^{N^2 xk}$ es también conocida. Con respecto a cualquier matriz ortonormal $\boldsymbol{U} = [\boldsymbol{u}^{(1)} \dots \boldsymbol{u}^{(m)}] \in \mathbb{R}^{kxm}$, con $m < k$, se puede escribir la matriz de menor rango como:

$$\widehat{\boldsymbol{\mathcal{G}}} = \boldsymbol{\mathcal{G}}\boldsymbol{U}. \tag{8}$$

Siguiendo con el esquema de análisis de relevancia de variables, se propone el siguiente problema de optimización:

$$\min_{\boldsymbol{U}} \left\|\boldsymbol{\mathcal{G}} - \widehat{\boldsymbol{\mathcal{G}}}\right\|_F^2 \quad \text{s. a.} \quad \boldsymbol{U}^\intercal \boldsymbol{U} = \boldsymbol{I}_m, \tag{9}$$

donde $\|\cdot\|_F$ representa la norma de Frobenius e $\boldsymbol{I}_m$ denota una matriz identidad de dimensión m. El problema anterior tiene una versión dual que puede escribirse como un problema de maximización con respecto a la varianza de $\widehat{\boldsymbol{\mathcal{G}}}$ como se observa a continuación:

$$\max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^\intercal \boldsymbol{\mathcal{G}}\boldsymbol{\mathcal{G}}\boldsymbol{U}) \quad \text{s. a.} \quad \boldsymbol{U}^\intercal \boldsymbol{U} = \boldsymbol{I}_m. \tag{10}$$

Por último, los coeficientes η_t utilizados en la media ponderada son los valores de clasificación que cuantifican la cantidad que cada columna de la matriz $\boldsymbol{\mathcal{G}}$ (cada

matriz kernel) contribuye a minimizar la función dada en (9). Por otra parte, aplicando el enfoque de relevancia de variables, se calcula el vector de clasificación $\boldsymbol{\eta} = [\eta_1, \dots, \eta_k]$ usando la siguiente ecuación:

$$\boldsymbol{\eta} = \sum_{t=1}^{k} \lambda_t \boldsymbol{u}^{(t)} \circ \boldsymbol{u}^{(t)}, \tag{11}$$

donde λ_t y $\boldsymbol{u}^{(t)}$ son el *t*-ésimo autovalor y autovector (valor y vector propio) de $\boldsymbol{\mathcal{G}\mathcal{G}}$, respectivamente. El operador $\circ$ denota el producto (punto a punto) Hadamard. Con respecto a la formulación del problema, se puede garantizar que el valor de $\boldsymbol{\eta}$ siempre es positivo, por lo cual se puede utilizar directamente para llevar acabo la combinación lineal.

4.5.2 Agrupación basada en centroides aplicando optimización multi-criterio con algoritmos genéticos

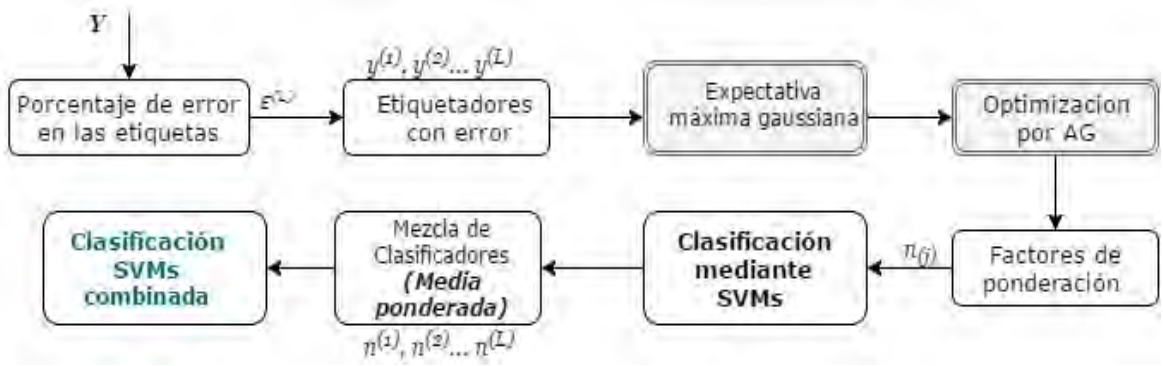


Figura 11. Diagrama de bloques para el cálculo de los factores de ponderación mediante el método de agrupación basada en centroides.

En esta tesis se propone emplear un método del clustering o agrupación basada en centroides y posteriormente realizar una optimización multi-criterio utilizando algoritmos genéticos con el objetivo de calcular los valores $\boldsymbol{\eta}$ para realizar la mezcla de clasificadores por medio del método de la media ponderada. Para llevarlo a cabo se efectuó el proceso descrito en la Figura 11, para lo cual se desarrolló un algoritmo capaz de identificar y penalizar los conjuntos etiquetados asignando un valor $\boldsymbol{\eta}$ que dependerá del porcentaje de error que estos contengan.

Este método consta de una etapa donde se separan los conjuntos etiquetados para cada clase, este proceso se realiza para cada conjunto etiquetado, luego se procede a generar una función de densidad de probabilidad para cada subconjunto, posteriormente los datos se ajustan a un modelo generando una ecuación para cada función de densidad de probabilidad y finalmente se aplicó una optimización multi-criterio haciendo uso del método Multiobjective Genetic Algorithm Options (gamultiobj) el cual hace parte del Global Optimization Toolbox, con el fin de generar los valores $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_k]$ para cada etiquetador.

Como se puede observar en la Figura 12, el proceso de separación de las clases de cada etiquetador se realizó con el fin de que estas fueran más diferenciables y de esta manera poder mirar el nivel de dispersión que existen entre cada clase. Es claro observar que a medida que el error de los subconjuntos aumenta, las muestras se comienzan a dispersar debido a que el error de los conjuntos va aumentando.

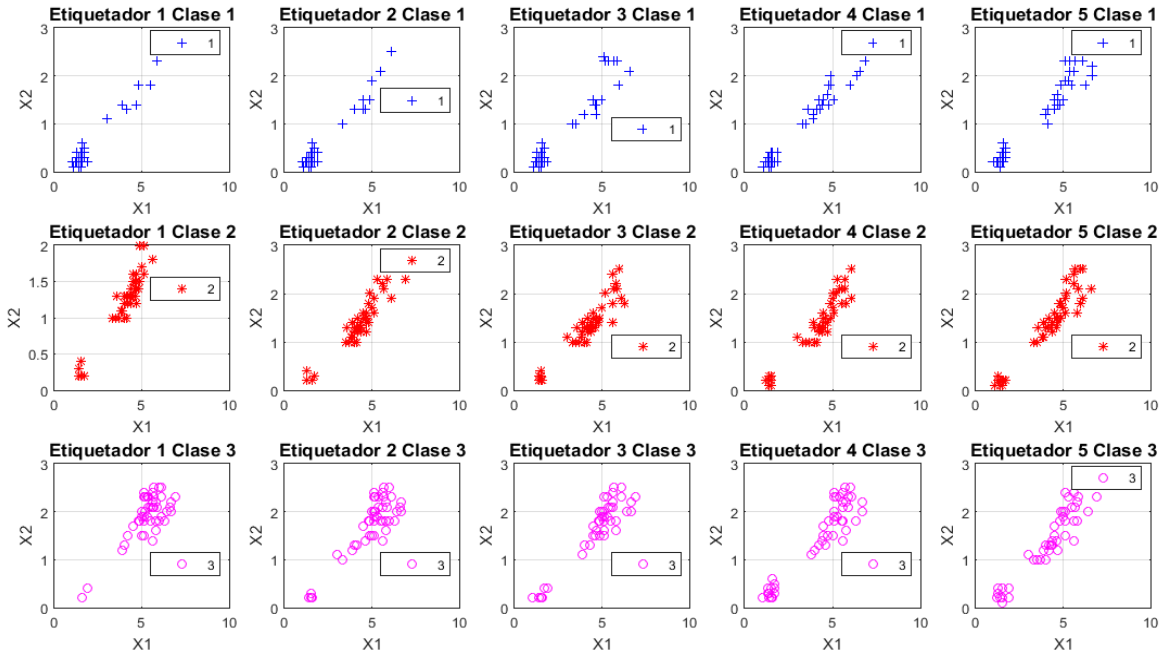


Figura 12. Ejemplo de conjuntos etiquetados separados por clases para cada etiquetador con una tasa de error del [10 20 30 40 50] porciento, respectivamente.

Una vez realizado el proceso de separación de las clases, se procede a aplicar una de las técnicas del clustering o agrupamiento basado en centroides (CBC). El propósito general del CBC consiste en poder realizar una minimización de una función objetivo. Este proceso de minimización establece que tan bueno es el resultado del agrupamiento y la solución de este proceso se logra de manera iterativa por medio de la generación de centroides que se actualizan. El resultado de cada iteración genera la asignación de los elementos al conjunto cuyo centroide se encuentre más próximo [38].

Una variante para realizar CBC se puede obtener analizando el grado de pertenencia de un elemento al grupo y cómo este elemento influye cada que un centroide es actualizado [39]. Esta variante llamada clustering o agrupamiento basado en la máxima esperanza gaussiana (GEMC), pertenece al grupo de métodos de clustering o agrupamiento basados en densidades (DBC) [40]. La función objetivo de este método es la combinación lineal de las distribuciones gaussianas centradas en los centroides de cada grupo como se muestra a continuación [41]:

$$GEMM_{log}(\mathbf{X},\mathbf{C}) = -\sum_{i=1}^{C} log\left(\sum_{j=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{q}_i)p(\boldsymbol{q}_j)\right), \tag{12}$$

Donde $p(\boldsymbol{x}_i|\boldsymbol{q}_i)$ es la probabilidad de $\boldsymbol{x}_i$ dado que es generado por una distribución gaussiana centrada en $\boldsymbol{q}_j$ y donde $p(\boldsymbol{q}_j)$ es la probabilidad a priori del grupo cuyo centroide en $\boldsymbol{q}_j$.

Las respectivas funciones membresía para cada uno de los elementos son:

$$m_{GEMM}(\boldsymbol{q}_j|\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i|\boldsymbol{q}_i)p(\boldsymbol{q}_j)}{p(\boldsymbol{x}_i)}, \tag{13}$$

en este caso la función membresía tiene un valor de probabilidad, de esta manera es posible utilizar la regla de Bayes para calcular su valor, considerando a $p(\boldsymbol{x}_i)$ como evidencia:

$$p(\boldsymbol{x}_i) = \sum_{j=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{q}_j)p(\boldsymbol{q}_j), \tag{14}$$

el factor $p(\boldsymbol{x}_i|\boldsymbol{q}_j)$ se puede obtener de la siguiente manera:

$$p(\boldsymbol{x}_i|\boldsymbol{q}_j) = f(\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}_j) = \frac{1}{\det(\boldsymbol{\Sigma}_j)^{\frac{1}{2}}}(2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\intercal}}, \tag{15}$$

donde $\boldsymbol{\mu}$ representa al centroide $(\mu = \boldsymbol{q}_j)$, d es la dimensión, $\boldsymbol{\Sigma}$ figura como la matriz de covarianza y $\det(\cdot)$ denota el determinante de su matriz argumento. La función objetivo a minimizar está dada por la siguiente ecuación:

$$\boldsymbol{F}_{log} = -\sum_{\boldsymbol{i=1}}^{C} \log(p(\boldsymbol{x}_i)). \tag{16}$$

Cada una de estas funciones tendrá una forma y comportamientos diferentes dado por la distribución y densidad de las clases, esto se muestra en la Figura 13. Después de realizar este proceso los datos se ajustan a un modelo para cada superficie dada por el método GEMC, encontrando una ecuación que en este caso representa una función de distribución gaussiana para cada clase dada por cada conjunto etiquetado.

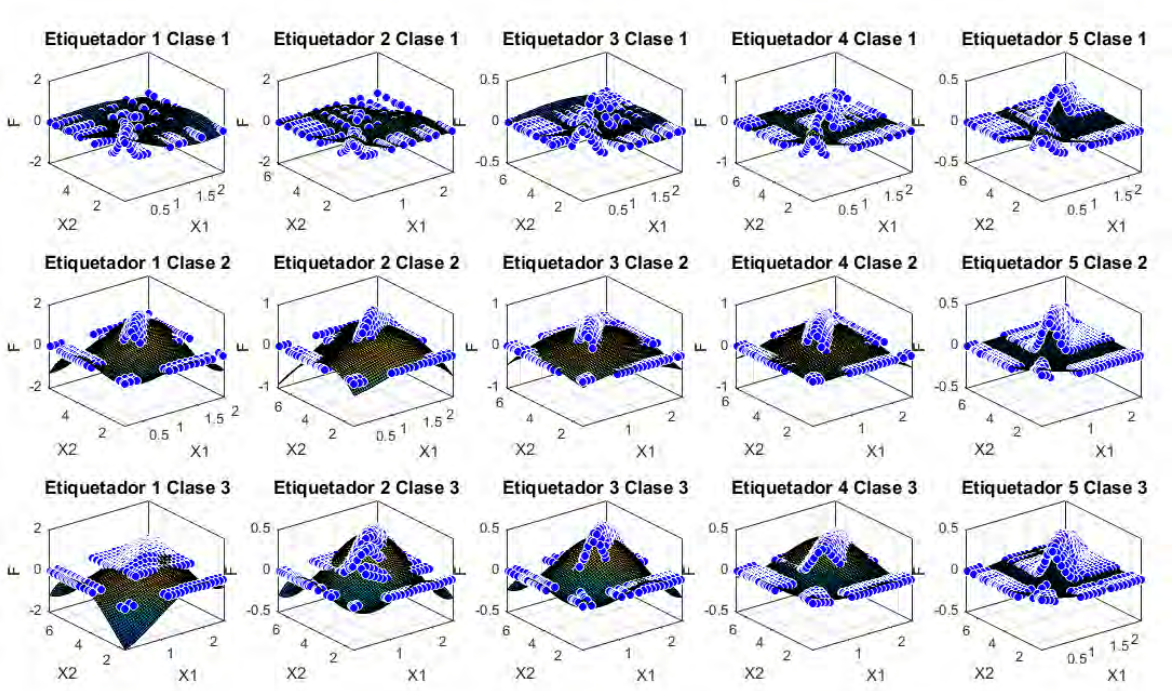


Figura 13. Ejemplo de distribuciones de probabilidad separadas por clase para cada etiquetador.

Luego se procede a generar una función objetivo que se optimizará. Finalmente, para la estimación de los factores de ponderación η se aplicó el método Multiobjective Genetic Algorithm Options (gamultiobj). Este es el encargado de resolver de manera iterativa los valores óptimos para la función multicriterio. Estos valores se promedian para cada etiquetador y este último valor representa en peso o ponderación η asociado a cada etiquetador. Con el objetivo de lograr mejores resultados, las funciones objetivo en 16, se encuentran sujetas a la siguiente condición:

$$\sum_{j=1}^{k} \eta_j = 1, \tag{17}$$

donde $\eta_j \in [0, 1]$.

4.6 DISEÑO DEL SIMULADOR PARA MÚLTIPLES EXPERTOS

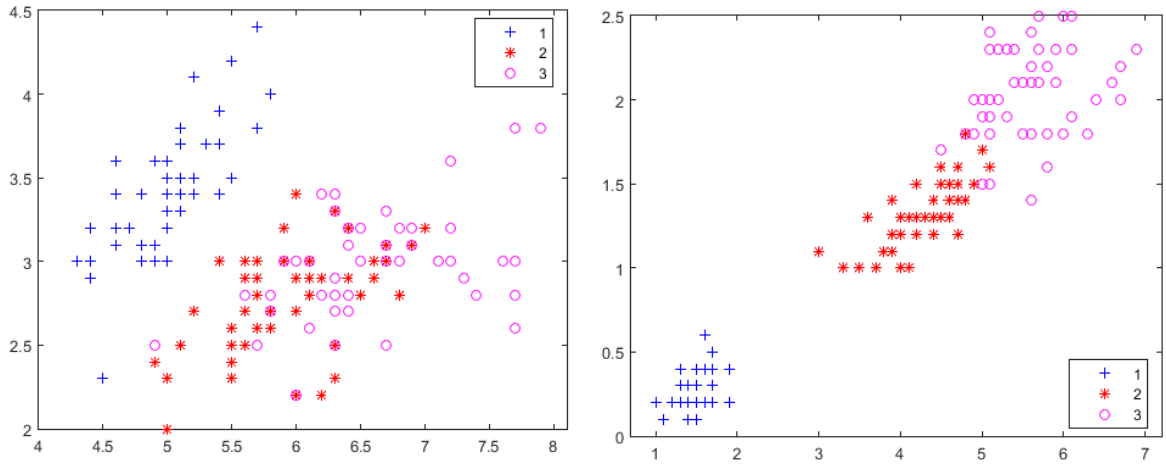
Con el fin de ejemplificar los procesos de clasificación y mezcla de clasificadores se realizo un simulador de tipo experimental, donde el usuario pueda afianzar el estudio de estos temas y aumentar la curva de aprendizaje mediante el descubrimiento de nuevos conocimientos. Este simulador fue pensado para el usuario común, por lo que se diseñó para que sea intuitivo y de uso interactivo. Esto se realizó con el objetivo del usuario sea capaz de realizar análisis de clasificación de datos y adentrarse en el escenario donde sea necesario simular un escenario de múltiples etiquetadores.

4.7 MARCO EXPERIMENTAL

Para realizar las pruebas y experimentos, se hará uso de la base de datos IRIS que se encuentran en el repositorio de *Machine Learning* (UCI) [42]. para efectuar la clasificación y también la combinación de clasificadores se utilizó las máquinas de soporte vectorial (SVM) empleando el toolbox de herramientas para reconocimiento de patrones PRTools para MATLAB [43]. Para el desarrollo del kernel gaussiano se tuvo en cuenta el análisis de relevancia variable para la extracción y selección de características y el cálculo de los factores de ponderación [36, 37, 44]. Para hacer la implementación del método de cálculo para los factores de ponderación haciendo uso de la agrupación basada en centroides se adaptó el modelo de agrupamiento de máxima esperanza gaussiana que es comúnmente utilizada en clustering [38, 45], así mismo se aplicó la optimización multi-criterio para generar los factores de ponderación [46]. Con estos métodos se genera un único clasificador que es el encargado de realizar la clasificación del conjunto de datos. Por otra parte, el simulador permite al usuario interactuar con las herramientas de clasificación, haciendo uso de todos los métodos anteriormente mencionados y haciendo más fácil la comprensión de las herramientas de clasificación.

4.7.1. Base de datos Iris

Esta base de datos fue creada por Sir Ronald Fisher y es una de las más usadas para realizar pruebas de clasificación de datos o análisis discriminante. En resumen, la base de datos IRIS relaciona la morfología de 3 variedades de flores de la especie iris, cuenta con 150 muestras de tres tipos diferentes: Versicolor, Virginica y Setosa. Por cada registro cuenta con cuatro características: ancho y largo del sépalo, y ancho y largo del pétalo. Por otra parte, esta base de datos se caracteriza por poseer dos clases traslapadas entre sí, así mismo contiene otra clase totalmente distinguible de las demás [4], a continuación se puede observar todas las características mencionadas (Figura 14).



(a) Ancho y largo del sépalo

(b) Ancho y largo del pétalo

Figura 14. Base de datos IRIS representada con sus tres clases, donde 1 2 y 3 representan a cada una de estas (Versicolor, Virginica y Setosa).

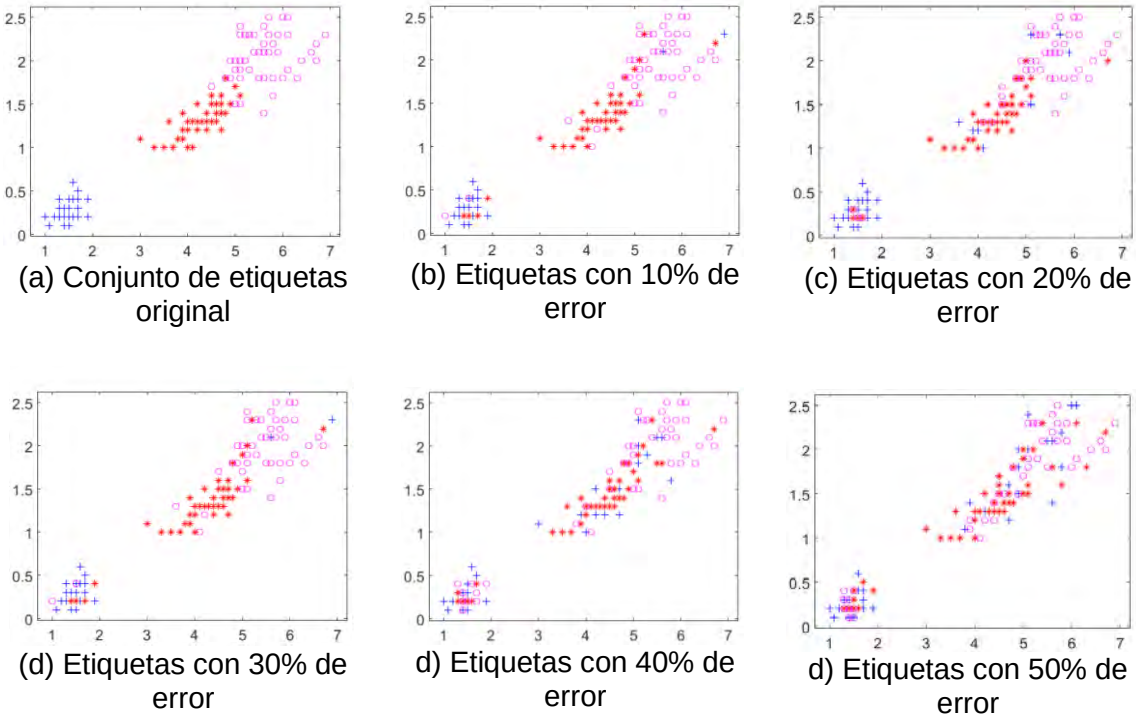
Para ejecutar las pruebas fue necesario emplear el conjunto de datos completo compuesto por las tres clases de flores Versicolor, Virginica y Setosa. El hecho de utilizar la base de datos iris para realizar las pruebas del estudio reside en que las etiquetas son acordes con los datos, por otra parte esta base de datos cuenta con la propiedad de separabilidad ya que sus clases no se encuentran totalmente traslapadas y son notoriamente diferenciables entre sí. De acuerdo con lo anterior, es factible aseverar que el conjunto de etiquetas presenta una alta tasa de confiabilidad, asegurando de esta manera que es apropiada para comprobar los resultados alcanzados aplicando los métodos propuestos [4].

4.7.2. Simulación de múltiples etiquetadores

En el proceso de desarrollo de proyectos de investigación es común utilizar bases de datos con el fin de extraer conocimiento haciendo uso de diferentes pruebas, métodos o herramientas, pero es habitual que existan inconvenientes en lo que se refiere a las bases de datos. Por ejemplo, cuando es necesario desarrollar modelos con múltiples bases de datos, por difícil acceso a estas o simplemente por costos no es posible acceder a ellas. En este orden de ideas se hace necesario una manera de simular bases de datos con el objetivo de aplicar los métodos de clasificación en un escenario de múltiples expertos.

Acorde con lo anterior se simuló diferentes conjuntos de etiquetas con la base de datos IRIS generando un escenario donde tenemos diferentes conjuntos

etiquetados provenientes de múltiples expertos, a cada uno de estos conjuntos se le asignó un porcentaje de error en sus etiquetas con el propósito de recrear un conjunto de etiquetadores corruptos, cabe señalar que el proceso de asignación de las etiquetas se hace de manera aleatoria. En la Figura 15, se puede examinar más a fondo como fueron asignados los porcentajes de error a los vectores de etiquetas corruptos y además como sus muestras a medida que el error en las etiquetas aumenta pierden la propiedad de separabilidad.



(a) Conjunto de etiquetas original

(b) Etiquetas con 10% de error

(c) Etiquetas con 20% de error

(d) Etiquetas con 30% de error

d) Etiquetas con 40% de error

d) Etiquetas con 50% de error

Figura 15. Base de datos IRIS generada con diferentes porcentajes de error en sus etiquetas.

El procedimiento que se efectuó para la mezcla se hizo de la siguiente manera: se generó un número que depende del número total de datos (N_d) y del porcentaje de error (per) que se desea, haciendo una regla de tres se puede obtener el conjunto de muestras que se permutaran (N_t) para efectuar el cambio en el conjunto total de datos, mediante la ecuación 18:

$$N_t = \frac{per * N_d}{100\%} \tag{18}$$

Por ejemplo si el per es del 10% aplicando la ecuación anterior se calcula el total de datos que se permutarán resultando que el valor N_t sería 15, debido a que el número total de datos de IRIS N_d es de 150. Esto se puede observar más

detenidamente en la Figura 16, donde en la primera columna están las 150 muestras representadas por escalas, además es posible apreciar que hay 15 escalas que no pertenecen a su respectiva clase por lo que se concluiría que en este caso estas serán las muestras permutadas. Con este valor se generan dos vectores aleatorios con números comprendidos entre [0-150], estos valores representan las posiciones de las muestras que se permutarán. Luego estas posiciones intercambian la muestra que corresponde al número aleatorio que está contenido en la primera posición del primer vector con la muestra correspondiente generada en el otro vector, finalmente el proceso es repetido hasta completar permutar las 15 posiciones con lo que se completa el proceso permutación.

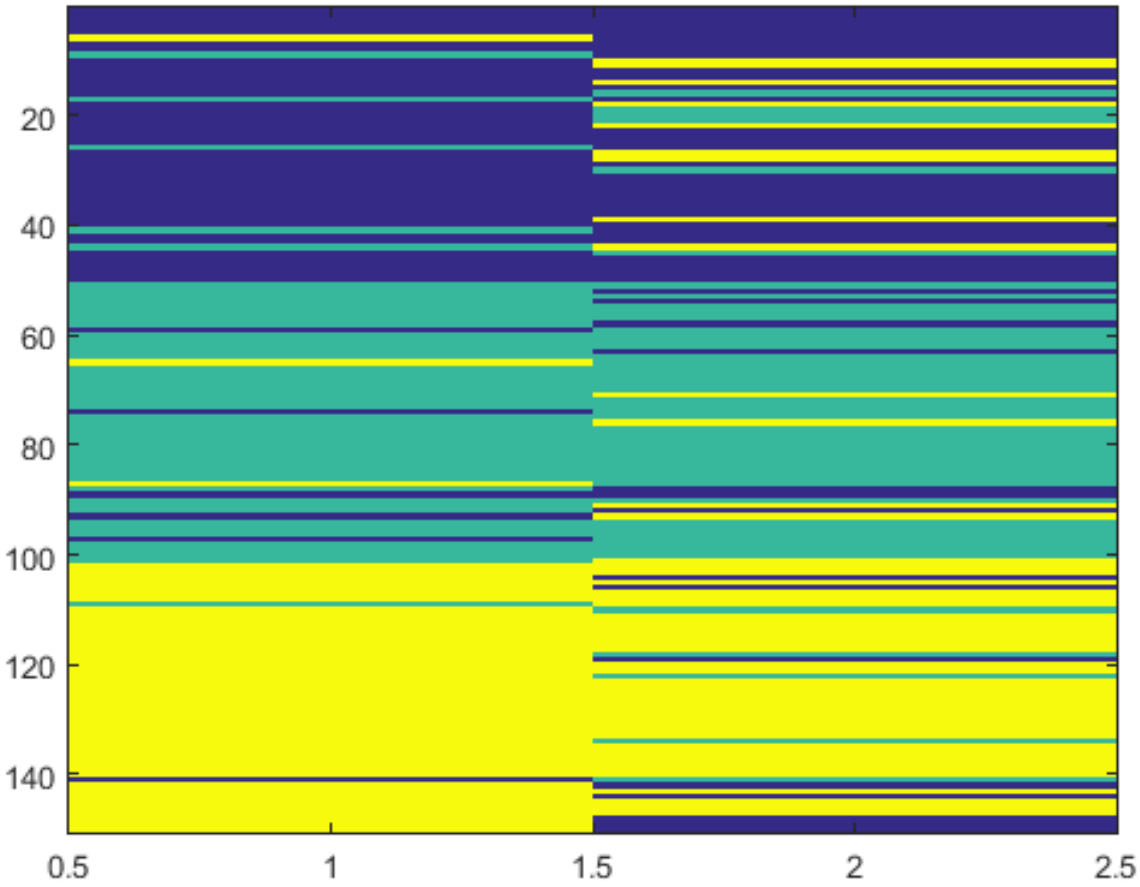


Figura 16. Vectores de etiquetas corruptos con 10% y 20% de error respectivamente.

4.7.3. Estudio comparativo de técnicas de clasificación

El método comparativo se presenta como una fase característica del ser humano, esta es aplicada en infinidad de tareas de una manera constante, por ejemplo en actividades en la que se requiere equiparar diferentes variables con el fin de constatar la veracidad o también el desempeño de una o más hipótesis. De acuerdo con lo anterior, en el presente trabajo se realiza un estudio comparativo de diferentes técnicas de clasificación convencional con el fin de contrastar su desempeño en términos de error de clasificación.

Para el estudio comparativo fue necesario abordar un conjunto de métodos de clasificación supervisada esto con el fin de identificar el que presentará un mejor desempeño en clasificar conjuntos de datos. Entre los métodos de clasificación estudiados que posteriormente fueron utilizados se puede mencionar los siguientes: máquinas de soporte vectorial, clasificador discriminante lineal, clasificador discriminante cuadratico, clasificador lineal discrimiante de fisher's y La optimización del clasificador de Parzen.

El procedimiento que se realizó con los clasificadores consta de realizar el procedimiento de clasificación con el conjunto de prueba perteneciente a cada uno de los conjuntos de etiquetas simulados con los diferentes porcentajes de error, de esta manera se obtiene el error de estimación de clasificación para cada conjunto. De igual modo este proceso se repite iterativamente durante 100 ciclos obteniendo como resultado un error de estimación para cada conjunto etiquetado. Luego de finalizar el proceso se conforma la matriz $\boldsymbol{E}_m \in \mathbb{R}^{100x5}$, donde m representa el número de métodos de clasificación que se utilizaron (en este caso fueron 5 métodos de clasificación).

A la matriz $\boldsymbol{E}_m$ se le aplico dos medidas estadísticas para cuantificar el desempeño de cada método de clasificación: el promedio y la desviación estándar. Estos se aplicaron obteniendo el promedio y la desviación por cada etiquetador resultando al final en un valor que representaría el desempeño de cada método de clasificación.

4.7.4. Estudio comparativo de técnicas de combinación para clasificadores

Es habitual que en los entornos donde se trabaja con diferentes clasificadores en ocasiones sea necesario extraer las mejores características o resultados de cada clasificador individual, para con esto proceder a acoplar todas estas propiedades en un solo clasificador para mejorar su desempeño, de esta manera los métodos de combinación se convierten en una herramienta fundamental en un entorno de múltiples expertos. En este apartado del trabajo de grado se presenta los métodos de combinación que fueron utilizados y como se realizaron las pruebas de

combinación de clasificadores por medio de los siguientes métodos: la media aritmética, el máximo, el mínimo, la mediana, el producto y el voto mayoritario.

Para el estudio comparativo se siguió con la misma metodología que se aplicó en la comparación de clasificadores, tomando como punto de partida que se tiene una matriz $\boldsymbol{C_m} \in \mathbb{R}^{100x5}$, la cual representa el conjunto de clasificadores entrenados que se obtuvo al realizar la clasificación con los conjuntos etiquetados con error. La combinación se realizó utilizando los vectores fila de la matriz $\boldsymbol{C_m}$, de esta manera se combinan los 5 clasificadores pertenecientes al vector fila 1, este procedimiento se repite iterativamente fila por fila hasta completar 100 iteraciones. Con el resultado adquirido después de aplicar el método de combinación se conforma un vector columna denotado como $\boldsymbol{cc_m} \in \mathbb{R}^{100}$, el cual está conformado por un clasificador que es el resultado de cada combinación del vector fila pertenecientes a la matriz $\boldsymbol{C_m}$.

Una vez se tiene al vector $\boldsymbol{cc_m}$ se procede a realiza nuevamente la clasificación del conjunto de datos original con el fin de obtener el error de estimación, este proceso se realiza de manera sucesiva durante 100 iteraciones, originando un vector columna denotado como $\boldsymbol{ec_m}$, el cual contiene los errores de estimación para cada clasificación. Posteriormente, al vector $\boldsymbol{ec_m}$ se le aplicaron la media arimética y la desviación estándar con lo que finalmente es posible medir el desempeño del método de clasificación y concluir cuál es el más eficaz.

4.7.5. Estudio de desempeño del algoritmo de la media ponderada con los factores de ponderación generados por los métodos de matrices kernel y agrupación basada en centroides

Para comprobar la efectividad y la estabilidad del método propuesto (media ponderada) haciendo uso de los factores de ponderación calculados por los métodos basados en matrices kernel y agrupación basada en centroides se realizó el mismo procedimiento general para calcular la efectividad en los métodos anteriores (sec. clasificación y combinación). Primero se genera una matriz $\boldsymbol{C} \in \mathbb{R}^{30x5}$, que representa un conjunto de clasificadores entrenados con los conjuntos etiquetados.

Luego se aplicaron los métodos de cálculo de los factores de ponderación (matrices kernel y agrupación basada en centroides) a cada conjunto etiquetado con los siguientes porcentajes de error [10 40 50 60 70] %, con el fin de generar los factores de ponderación $\boldsymbol{\eta} = [\eta_1, \dots, \eta_k]$. Una vez hecho este proceso se realiza la combinación aplicando el método basado en la media ponderada, donde cada valor de ponderación es asignado a cada clasificador ubicado en el subconjunto o vector fila contenido en la matriz $\boldsymbol{C}$, luego se efectúa la combinación de los clasificadores generando de esta manera un nuevo clasificador. Este proceso se repite durante 30 iteraciones donde finalmente se obtiene una estructura que

contendrá todos los clasificadores resultantes de la combinación con la media ponderada.

Finalmente con este grupo de clasificadores se procede a realizar la clasificación del conjunto de datos original, generando un nuevo conjunto de errores de estimación de clasificación. Con estos errores se puede aplicar las medidas estadísticas de desempeño con el fin de comparar los resultados adquiridos con los métodos de combinación convencionales.

También se realizaron pruebas con diferentes porcentajes de error en los conjuntos etiquetados con el fin de observar el comportamiento de los métodos en diferentes entornos, así de esta manera asegurar cuan correctos son los resultados obtenidos en los cálculos de los factores de ponderación.

4.8 MEDIDAS DE DESEMPEÑO

Para medir el desempeño de los métodos propuestos se usan las siguientes medidas convencionales tales como: el error estándar, el margen de error, la media estadística y la desviación estándar explicadas en [47]. Con estas dos últimas medidas se pretende medir el desempeño y la precisión de los métodos de clasificación y combinación.

La media estadística se define como el promedio aritmético de un conjunto de observaciones $A_1, A_2, \dots, A_n$ y se define como:

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} A_i \, . \tag{19}$$

La media es una medida de tendencia central apropiada que puede ser aplicada a diferentes conjuntos de datos con el fin de observar cual es la disposición de estos para agruparse alrededor del centro o también en algunos casos de ciertos valores numéricos [47].

La desviación estándar para un conjunto de observaciones $A_1, A_2, \dots, A_n$ puede calcularse como la raíz cuadrada positiva de la varianza y se define como:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} |A_i - \bar{\mu}|^2} \tag{20}$$

La desviación estándar se considera una medida de dispersión que tiene como funcion medir el nivel de agrupación de los datos estudiados con respecto al valor promedio. Entre menor sea el valor de la desviación estándar más agrupados

estarán los datos alrededor de la media, para el caso contrario, cuando el valor de la desviación estándar tiende a aumentar, mas dispersos se encontrarán los datos en relación a la media.

5 RESULTADOS Y DISCUSIÓN

A continuación se discuten los resultados obtenidos luego de efectuar los experimentos aplicando las metodologías descritas en 4 y según los métodos de clasificación y combinación descritos en 4.7.

5.1. PRUEBA DE LOS MÉTODOS DE CLASIFICACIÓN

En esta sección, se presenta las pruebas que se realizaron con los métodos de clasificación convencional, además los resultados concernientes con respecto a cuantificar el desempeño y precisión de los clasificadores.

En la Figura 17, se muestra un diagrama de cajas y bigotes, el cual contiene las cinco representaciones de los métodos de clasificación que se utilizaron, la línea roja representa la media del conjunto de datos representado por la matriz $\boldsymbol{E}_m$, los bordes de las cajas representan el percentil 25 y 75 respectivamente; el percentil es una medida de posición no central que muestra el valor de la variable por debajo de la cual se encuentra un porcentaje dado de muestras en un conjunto, además los bigotes se extienden hasta los datos más extremos o datos atípicos. En el eje (Y) se tiene al error de estimación de clasificación que se calculó para cada conjunto etiquetado en escala porcentual y finalmente en el eje (X) esta cada método de clasificación utilizado. Cabe anotar que de esta manera se seguirán presentando los demás resultados de comparación.

Ahora bien, en la Figura 17, es posible observar que los métodos de clasificación basados en LDC y Fisher son los que presentan en promedio un mayor error de estimación con respecto al conjunto de clasificadores. Por otra parte los métodos de clasificación QDC y Parzen presentan un menor error de estimación por conjunto etiquetado con respecto a LDC y Fisher, por último se tiene a las máquinas de soporte vectorial (SVM) las cuales presentan en promedio el menor error de estimación entre todos los métodos utilizados y por lo tanto presentan un mejor desempeño en la clasificación de los conjuntos de datos. Es por esta razón que se tomó la decisión de utilizarlas como el método elegido para proceder a realizar los demás análisis de clasificación y de combinación con las metodologías propuestas.

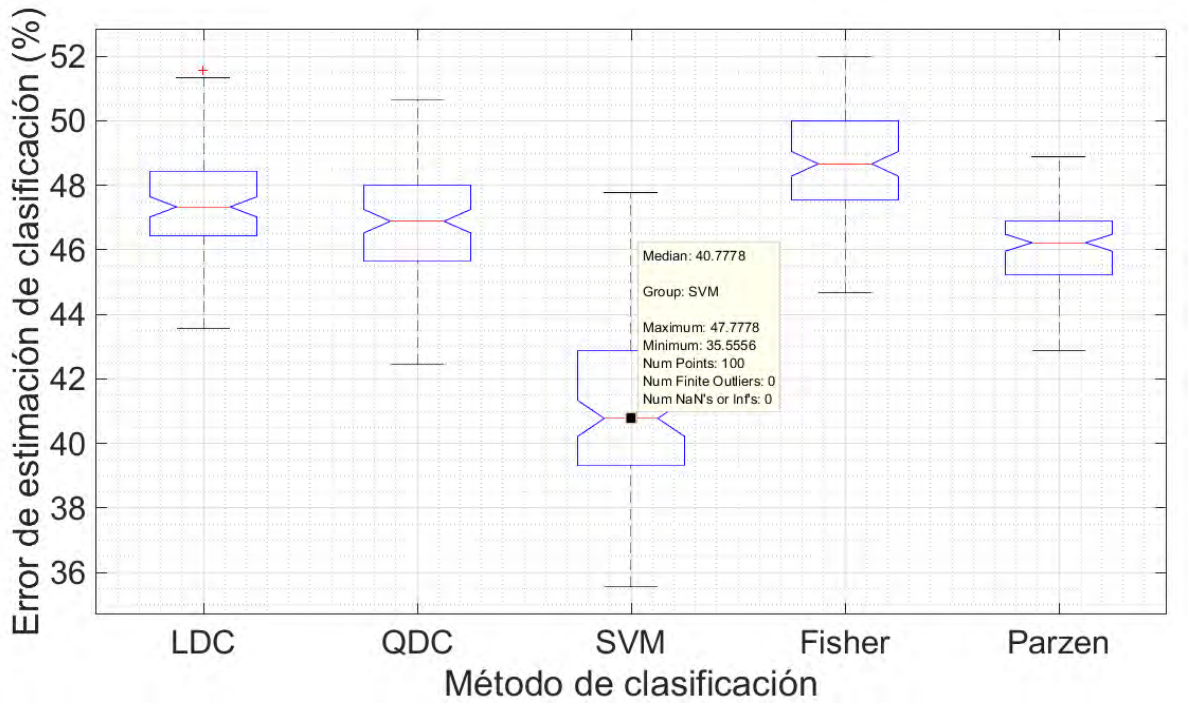


Figura 17. Diagrama de cajas y bigotes para los métodos de clasificación estudiados, donde cada caja representa un método de clasificación (LDC, QDC, SVM, Fisher, Parzen); resultando que las SVM presentaron un menor error de estimación.

5.2. PRUEBA DE MÉTODOS DE COMBINACIÓN DE CLASIFICADORES SVM

En esta prueba se realizó la combinación de varios clasificadores utilizando las máquinas de soporte vectorial (SVM). La mezcla se aplicó haciendo uso de los siguientes métodos: la media aritmética, el máximo, el mínimo, la mediana, el producto y el voto mayoritario.

En el diagrama de cajas y bigotes en la Figura 18, se encuentran plasmados los métodos de combinación que fueron aplicados. Como se puede ver, el mínimo es el método de combinación que presenta en promedio un mayor error de estimación de clasificación y es el que más datos atípicos presenta; por otra parte, el resto de métodos de combinación exhiben un comportamiento similar en términos de error de estimación. Haciendo un análisis más detallado es posible apreciar que el método basado en la media aritmética muestra el mejor desempeño en contraste con los otros métodos de combinación, ya que su promedio de error de estimación de clasificación es menor.

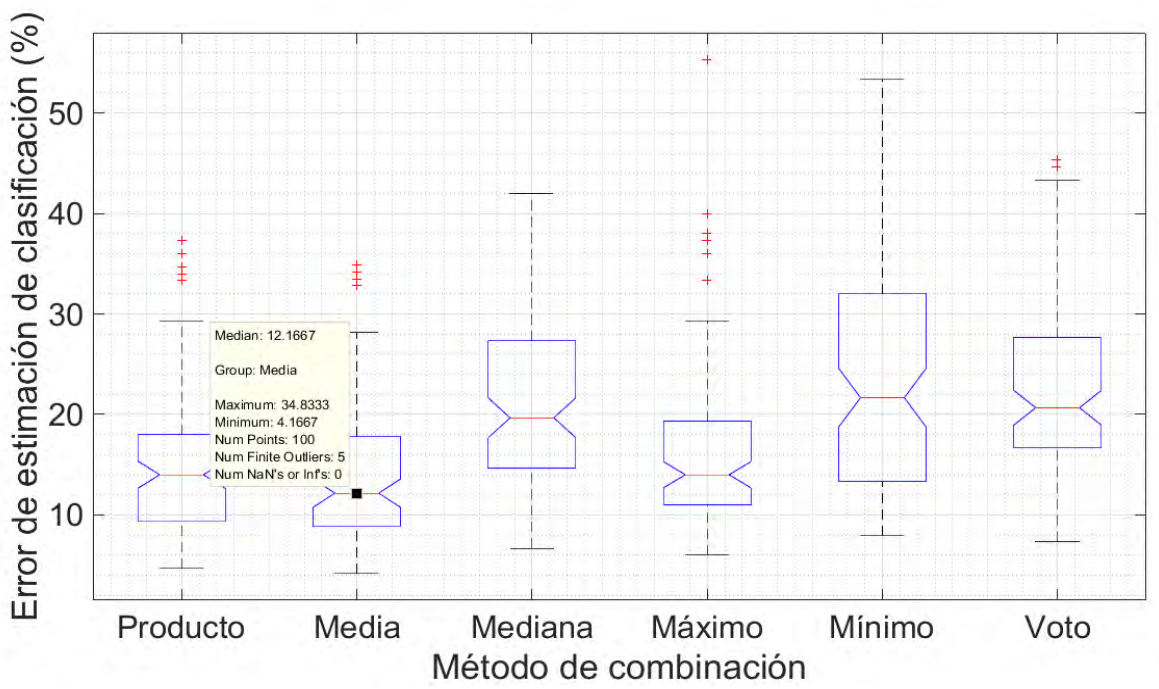


Figura 18. Diagrama de cajas y bigotes para los métodos de combinación ensayados con clasificadores SVM, donde cada caja representa el error promedio para cada método de combinación .

Otra medida de desempeño que se aplicó fue la desviación estándar (Figura 19) siendo esta una medida que representa el nivel de dispersión que presentan los datos con respecto a la media de los mismos, se puede decir que cuanto más pequeño sea el valor de la desviación más preciso será el método y menos dispersión se presentaran en sus resultados. Dando como resultado que el método de la media posee el valor más bajo para la desviación con respecto a los demás métodos de combinación. En este orden de ideas el método de la media tiende a presentar un mejor desempeño general frente a los demás métodos de combinación.

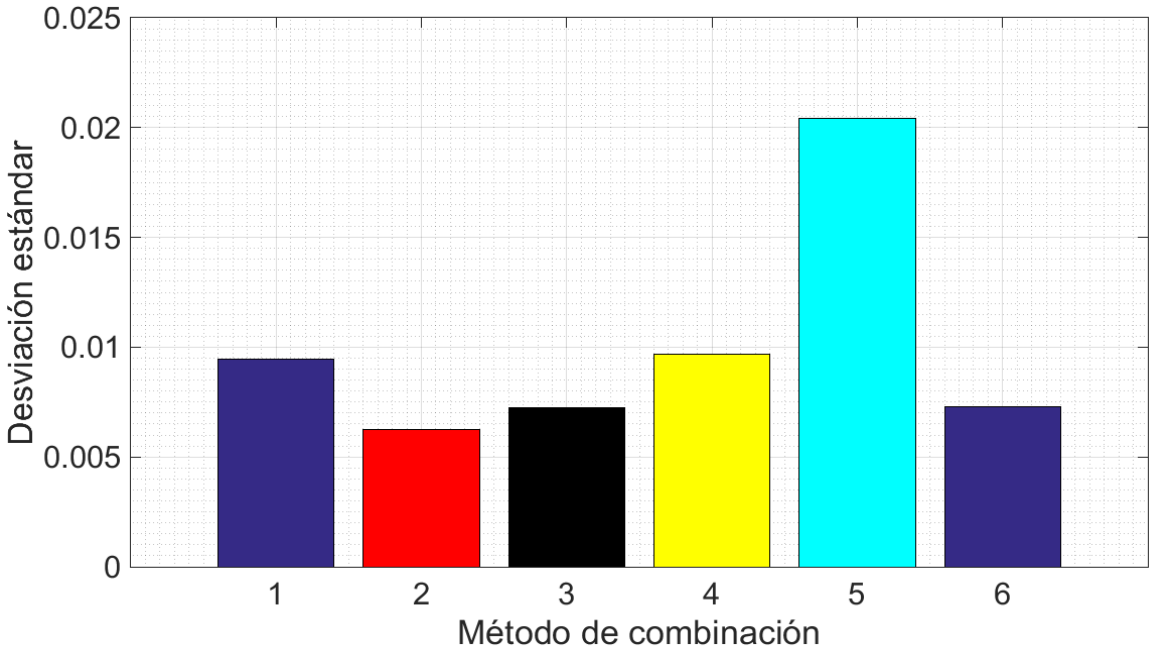


Figura 19. Diagrama barras de la desviación estándar para el conjunto de resultados de cada método de combinación; resultando que el método 2 correspondiente a la media presenta un valor mas bajo comparado con los demas métodos.

5.3. PRUEBA DE DESEMPEÑO DE LOS MÉTODOS PROPUESTOS BASADOS EN MATRICES KERNEL Y AGRUPACIÓN BASADA EN CENTROIDES

5.3.1. Método basado en matrices kernel

Para verificar el desempeño de los métodos para el cálculo de los factores de ponderación se realizó una prueba de precisión que consiste en aplicar reiteradas veces el método con los mismos porcentajes de error para esto fue necesario realizar 100 iteraciones.

La Figura 20, muestra todos los métodos de combinación probados y se incluye al método de combinación basado en la media ponderada, este hace uso de los factores de ponderación calculados por medio del enfoque de matrices kernel. Se puede apreciar que el método de la media ponderada presenta en promedio un error de estimación de clasificación considerablemente menor comparado con los resultados de los demás métodos en estudio.

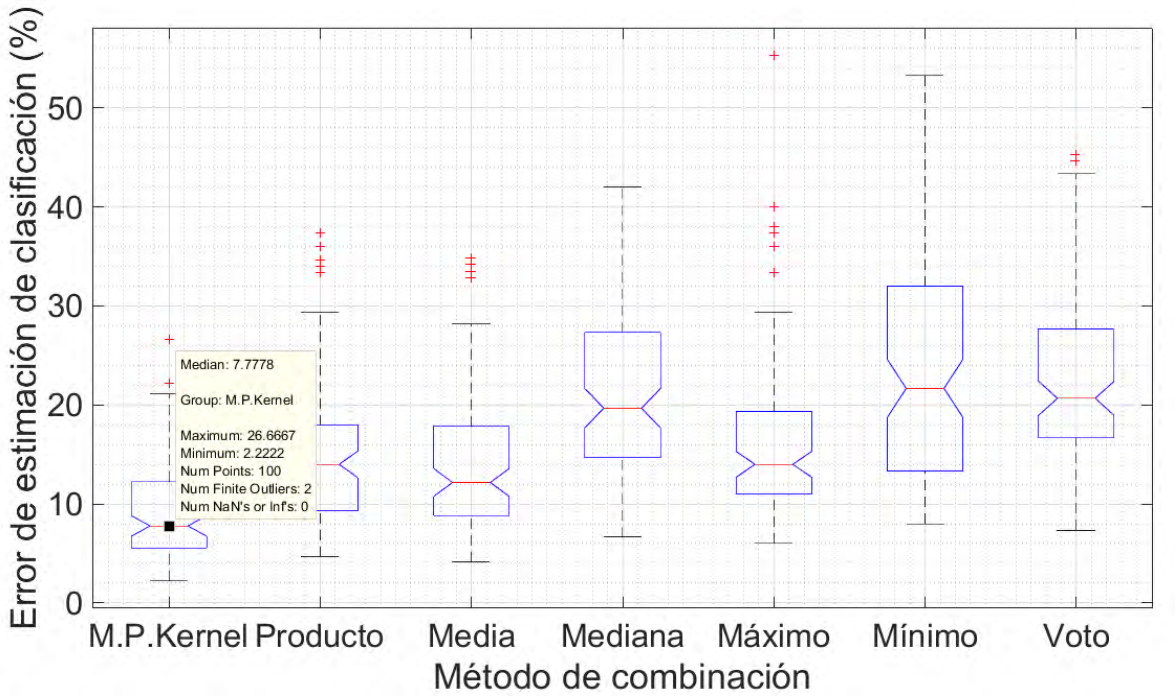


Figura 20. Diagrama de cajas y bigotes para la media ponderada con los factores η generados por el método de matrices kernel y los demás métodos de combinación de clasificadores; donde M.P.Kernel representa la media ponderada haciendo uso del método kernel.

Para probar la efectividad de este método se realizaron diferentes pruebas donde los conjuntos de etiquetas son contaminados con determinados porcentajes de error, los cuales están consignados en la Tabla 1. A continuación a cada conjunto se le realizaron 30 iteraciones y se aplicó el método de cálculo de los factores de ponderación basados en matrices kernel, de esta manera se obtiene los valores η expuestos en la Tabla 2, Con estos valores se procedió a realizar la combinación mediante el método propuesto (Media ponderada), luego los resultados fueron consignados en la Tabla 1, Además en esta se encuentran los errores de clasificación de la media y el voto mayoritario con el fin de comparar con los resultados del método propuesto (Media ponderada).

Experimento 1: En este experimento el primer anotador tiene el porcentaje de error más bajo en comparación con los demás al menos por un margen del 30%, En la Tabla 1, se puede observar que el valor η para el primer conjunto etiquetado es el mayor, por lo cual este tendrá mayor relevancia al momento de realizar la combinación.

Experimento 2: En este caso el etiquetador $y^{(1)}$ presenta la mayor tasa de error entre el grupo de etiquetadores. El resto tiene errores comprendidos entre 0 – 30 %. Observando la Tabla 2, es claro que el método le da menos peso al

etiquetador $y^{(1)}$, debido a que su porcentaje de error es mucho mayor en comparación con los demás.

Experimento 3: En este caso se prueba un escenario de error ascendente con un porcentaje de error iniciando en 30% y llegando al 90% en intervalos de 15%, los resultados presentados en la Tabla 2, para el experimento presentan un comportamiento descendente en los factores η, por esta razón al etiquetador $y^{(1)}$ el método le asigna un mayor peso, ya que es el que presenta la menor tasa de error entre en grupo de etiquetadores, el mismo proceso se efectúa de manera contraria a medida que esta tasa va aumentando para el resto de etiquetadores.

Experimento 4: En el caso 4 todos los etiquetadores presentan el mismo valor de error en las etiquetas, como se evidencian en la Tabla 2, los valores η tienden a permanecer en el mismo valor aunque presenten algunas variaciones mínimas.

Experimento 5: En este caso se presenta una situación similar al experimento 3 pero con un rango más amplio en el error, los resultados para este experimento tienden a presentar un comportamiento similar proporcionando más importancia al etiquetador $y^{(1)}$ y disminuyendo el valor de η a medida que la tasa de error aumenta.

$\boldsymbol{Experimentos}$	$\boldsymbol{y^{(1)}}$	$\boldsymbol{y^{(2)}}$	$\boldsymbol{y^{(3)}}$	$\boldsymbol{y^{(4)}}$	$\boldsymbol{y^{(5)}}$	$\boldsymbol{Método\ Prop}$	$\boldsymbol{Media}$	$\boldsymbol{Voto}$
1	20	50	60	55	65	$\mathbf{15.59 \pm 3.02}$	19.03 ± 3.47	26.77 ± 4.88
2	70	20	25	30	20	$\mathbf{4.01 \pm 0.82}$	4.77 ± 0.87	5.70 ± 1.04
3	30	45	60	75	90	$\mathbf{21.85 \pm 4.53}$	26.66 ± 4.86	39.03 ± 7.12
4	60	60	60	60	60	$\mathbf{20.74 \pm 3.78}$	20.88 ± 3.81	31.25 ± 5.70
5	20	40	60	80	100	$\mathbf{12.16 \pm 3.16}$	18.66 ± 4.17	23.83 ± 5.32

Tabla 1. Error de clasificación calculado para los métodos de la media ponderada utilizando los factores η dados por el método de matrices kernel, la media y el voto mayoritario respectivamente; resultando que en todos los casos el método propuesto presenta siempre un mejor desempeño en la clasificación.

$\%\boldsymbol{\eta}$	$\boldsymbol{y^{(1)}}$	$\boldsymbol{y^{(2)}}$	$\boldsymbol{y^{(3)}}$	$\boldsymbol{y^{(4)}}$	$\boldsymbol{y^{(5)}}$
$Experimento\ 1$	25.41 ± 0.41	19.20 ± 0.15	18.52 ± 0.15	18.41 ± 0.19	18.52 ± 0.15
$Experimento\ 2$	15.65 ± 0.19	21.70 ± 0.27	20.83 ± 0.20	19.46 ± 0.25	22.33 ± 0.39
$Experimento\ 3$	22.97 ± 0.32	20.52 ± 0.21	19.15 ± 0.13	18.63 ± 0.10	18.71 ± 0.14
$Experimento\ 4$	20.11 ± 0.18	20.09 ± 0.13	19.86 ± 0.09	19.90 ± 0.15	20.01 ± 0.15
$Experimento\ 5$	25.12 ± 0.37	20.83 ± 0.40	18.12 ± 0.10	17.91 ± 0.11	18.00 ± 0.12

Tabla 2. Valores η calculados con el método de matrices kernel para cinco diferentes etiquetadores $\{y^{(1)}, \ldots, y^{(5)}\}$.

5.3.2. Método para el cálculo de η en la agrupación basada en centroides

En la Figura 21 se representaron los valores promedios de error de estimación para los métodos de combinación probados anteriormente y también el del método propuesto (Media ponderada) haciendo uso de los factores η generados por el método de agrupación basada en centroides (Gauss), esto se realizó con el objetivo de probar la precisión del método. Para esto se realizaron 100 iteraciones con la misma tasa de error para los etiquetadores. Es claro notar que el método de la media ponderada haciendo uso de los factores η generados por este método muestran un mejor desempeño frente a los métodos de combinación convencional, presentando una reducción del error de estimación comparado con el método más cercano y además el del valor más bajo (Media) en términos de promedio de error de estimación.

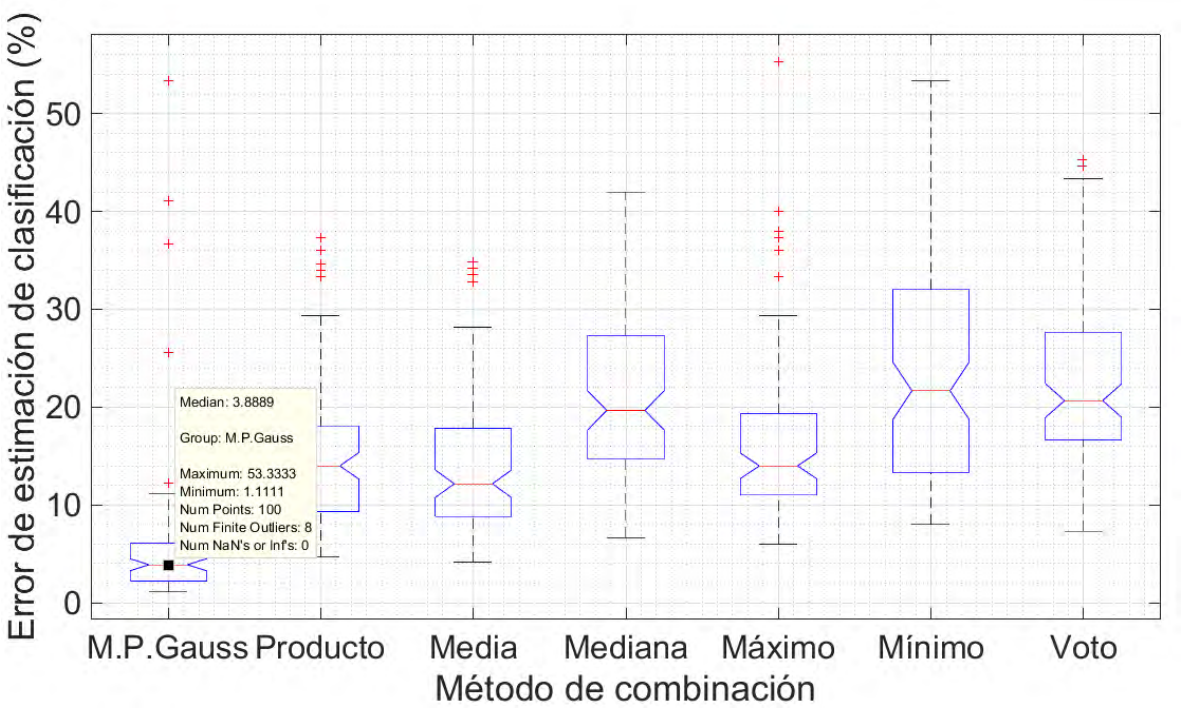


Figura 21. Diagrama de cajas y bigotes para la media ponderada haciendo uso de los factores η generados por el método de agrupación basada en centroides y los demás métodos de combinación de clasificadores.

Experimento 1: En este caso la mezcla se realizó con un porcentaje de error del 20% y se fue aumentado hasta 100% en intervalos de 20%. En la Tabla 3 se observa que los porcentajes de error corresponden con los valores η generados en la Tabla 4. Dando más peso al etiquetador $y^{(1)}$ y por ende mas importancia al momento de realizar la clasificación.

Experimento 2: En este caso el etiquetador $y^{(1)}$ tiene la tasa de error más baja y el resto de etiquetadores tiene un error comprendido mayor al 30% más que el primer etiquetador. Los resultados consignados en la Tabla 4, revelan que el método asigna un mayor valor η al etiquetador $y^{(1)}$ ya que este es el que menor tasa de error tiene asignado.

Experimento 3: En esta prueba se realizó el procedimiento contrario al proceso aplicado en el experimento 2, se le asignó un porcentaje de error alto al etiquetador $y^{(1)}$ y al resto se le asigno un porcentaje mucho menor en un rango entre el 20-30%. En la Tabla 4, se puede ver que el método ha asignado un valor de η bajo al etiquetador $y^{(1)}$ restándole importancia ya que presenta un nivel de error alto en sus etiquetas, también es claro que el valor η del resto de etiquetadores es mayor ya que sus tasas de error son claramente más bajas.

Experimento 4: En este caso se tiene un porcentaje de error ascendente empezando por el 30% y aumentado hasta el 90%, en intervalos de 15%, en este caso se espera una tasa de error alta, la cual ira aumentando dependiendo del error de cada etiquetador. Como se muestra en la Tabla 4, el desempeño del método es sobresaliente generando valores η de manera descendente asignando un mayor peso a $y^{(1)}$ y decreciendo a medida que el error aumenta.

Experimento 5: En este caso se evalúa una tasa de error igual para todos los etiquetadores en la Tabla 4, se puede ver que el valor η es aproximadamente el mismo para todos los etiquetadores.

$\boldsymbol{Experimentos}$	$\boldsymbol{y^{(1)}}$	$\boldsymbol{y^{(2)}}$	$\boldsymbol{y^{(3)}}$	$\boldsymbol{y^{(4)}}$	$\boldsymbol{y^{(5)}}$	$\boldsymbol{Método\ Prop}$	$\boldsymbol{Media}$	$\boldsymbol{Voto}$
1	20	40	60	80	100	$\mathbf{5.56 \pm 1.55}$	15.55 ± 2.02	23.72 ± 1.95
2	20	50	60	55	65	$\mathbf{2.42 \pm 1.59}$	10.11 ± 2.09	20.16 ± 2.02
3	70	20	25	30	20	$\mathbf{4.43 \pm 0.57}$	7.44 ± 0.72	8.83 ± 0.79
4	30	45	60	75	90	$\mathbf{2.22 \pm 1.75}$	23.11 ± 1.36	25.56 ± 1.85
5	60	60	60	60	60	$\mathbf{15.77 \pm 5.97}$	28.94 ± 2.86	23.94 ± 2.64

Tabla 3. Error de clasificación calculado para los métodos de la media ponderada utilizando los factores η dados por el método de agrupación basada en centroides, la media y el voto mayoritario respectivamente; resultando que en todos los casos el método propuesto presenta siempre un mejor desempeño en la clasificación.

$\%\boldsymbol{\eta}$	$\boldsymbol{y^{(1)}}$	$\boldsymbol{y^{(2)}}$	$\boldsymbol{y^{(3)}}$	$\boldsymbol{y^{(4)}}$	$\boldsymbol{y^{(5)}}$
$Experimento\ 1$	57.48 ± 6.7	26.72 ± 2.52	5.27 ± 2.33	5.99 ± 1.85	4.52 ± 1.73
$Experimento\ 2$	63.68 ± 7.05	13.56 ± 2.15	6.03 ± 1.38	7.13 ± 1.69	9.57 ± 3.56
$Experimento\ 3$	7.38 ± 1.67	26.90 ± 5.85	16.78 ± 3.44	16.19 ± 3.70	32.73 ± 4.43
$Experimento\ 4$	46.44 ± 8.37	28.36 ± 5.48	9.63 ± 1.56	8.07 ± 1.85	7.48 ± 4.50
$Experimento\ 5$	18.31 ± 6.34	17.58 ± 2.56	16.48 ± 2.95	23.72 ± 1.79	23.80 ± 3.74

Tabla 4. Valores η calculados con el método de agrupación basada en centroides para cinco diferentes etiquetadores $\{y^{(1)}, \ldots, y^{(5)}\}$.

5.4. SIMULADOR DE RECONOCIMIENTO DE PATRONES PARA MÚLTIPLES EXPERTOS

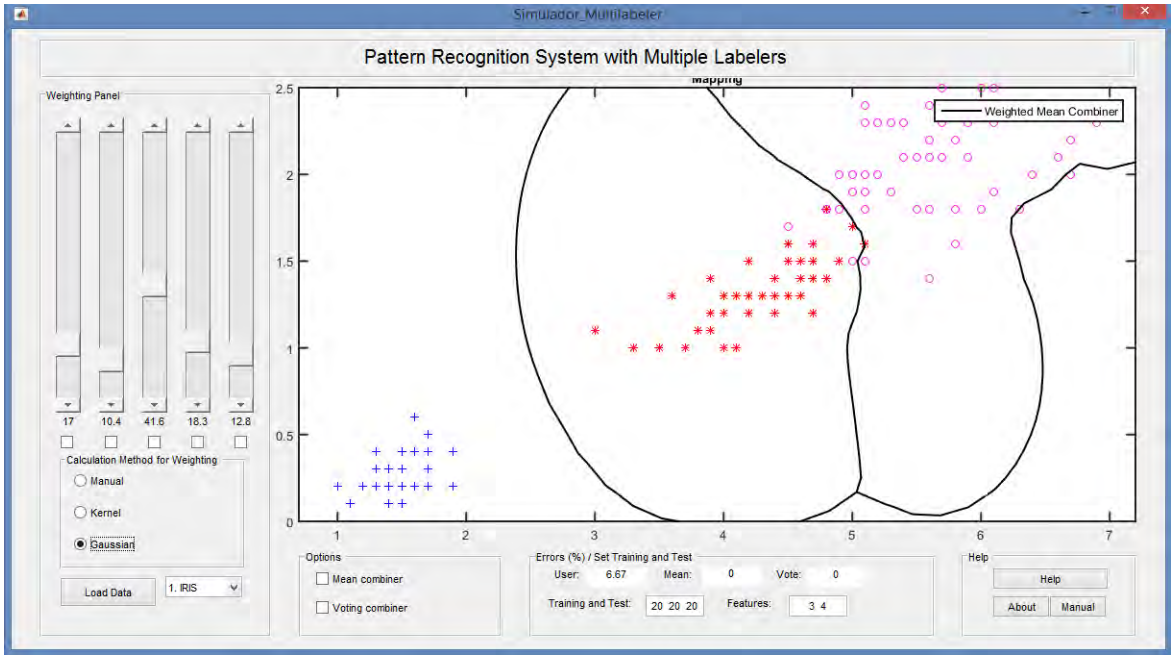


Figura 22. Simulador de reconocimiento de patrones en un escenario con múltiples expertos.

La Figura 22, muestra el simulador que fue diseñado con el objetivo de exhibir los resultados de clasificación dada por la combinación de múltiples etiquetadores. El simulador consta de tres modos Manual, Kernel y Gaussian, en el modo manual es posible elegir los pesos de cada etiquetador en el panel de ponderación. Así mismo es posible elegir entre diferentes conjunto de datos que vienen preseleccionados. Además tiene un apartado donde se puede seleccionar entre el método de combinación para la media (aritmética) y el voto mayoritario. Por otra parte se tiene un apartado donde es posible modificar los conjuntos de entrenamiento y prueba, así como también se puede seleccionar las características del conjunto de datos que se desea visualizar, el simulador cuenta con una sección de ayuda donde el usuario podrá encontrar el manual de usuario. Discusión

5.5. DISCUSIÓN

Es importante aclarar que el algoritmo de clustering en ningún momento modifica el número de grupos ni tampoco lo estima ya que éste se establece con anterioridad. Por esta razón el número de clústeres ya se encuentra preestablecido y se conserva en todo el proceso de agrupamiento.

6 CONCLUSIONES Y TRABAJO FUTURO

6.1. CONCLUCIONES

El modelo matemático propuesto basado en matrices kernel resulta ser un método novedoso para establecer factores de ponderación de manera que se pueda penalizar a los etiquetadores e inferir el grado de certeza que tiene cada uno de ellos al momento de evaluar los datos, esto, teniendo en cuenta la distribución natural de los datos en el espacio de características. Con lo anterior se garantiza que al realizar la mezcla de clasificadores se aproveche la información dada por todo el conjunto de clasificadores, así como también asegurar que esta información sea lo más acertada posible y que conduzca a resultados coherentes.

Las máquinas de soporte vectorial son una potente y eficiente herramienta matemática para clasificar datos por su excelente desempeño con respecto a los métodos de clasificación convencionales en escenarios donde existe un único conjunto etiquetado. Igualmente, en el presente trabajo, las SVM desempeñan un papel fundamental en el contexto requerido ya que el modelo matemático del que dispone permite establecer un método robusto de clasificación supervisada en entornos donde se cuenta con múltiples etiquetadores.

Los métodos de combinación de clasificadores se han establecido como una herramienta muy útil y eficiente a la hora de establecer un criterio de clasificación grupal en comparación a los métodos convencionales, es así como en este trabajo se evidenció la eficiencia de la mezcla de clasificadores haciendo uso del método de la media ponderada donde se relacionan las funciones de coste de los clasificadores y los factores de ponderación como medida de penalización a los etiquetadores. Este modelo de combinación le da mayor importancia a los etiquetadores más acertados minimizando el error de clasificación.

El simulador de datos propuesto como interfaz resulta ser un método innovador en el área de clasificación de datos, los elementos interactivos que se han predeterminado para el usuario permiten un manejo de los parámetros más importantes al llevar a cabo un proceso de clasificación en escenarios donde se cuenta con múltiples expertos. Desde la inclusión de múltiples bases de datos hasta el manejo de los parámetros intrínsecos de clasificación, los usuarios cuentan con múltiples herramientas para interactuar con los datos que desean analizar.

En general, el uso de la estrategia de múltiples expertos puede proporcionar un enfoque novedoso debido a la inclusión de los factores de ponderación ya que permiten penalizar a los malos etiquetadores generando una mejor clasificación de los datos. El método aquí propuesto permite hacer frente a los conjuntos de etiquetas con errores moderados, presentando un buen rendimiento en comparación con los métodos convencionales. Este enfoque ha sido bien visto por

investigadores en esta área como se puede evidenciar con la aprobación de este trabajo en el IberoAmerican Congress on Pattern Recognition (CIARP), el cual es un congreso con alta reputación en investigaciones de inteligencia artificial, dando cumplimiento con uno de los parámetros en los impactos esperados del trabajo.

6.2. TRABAJO FUTURO

En el proceso de realización de este trabajo se pudo evidenciar algunos enfoques hacia donde puede extenderse el desarrollo del presente trabajo y se listan a continuación:

- Se propone como trabajo futuro la implementación de clasificadores SVM combinados con técnicas Boosting con el fin de mejorar el desempeño en el tiempo de entrenamiento y clasificación.
- La realización de un nuevo método de mezcla de clasificadores aplicando el enfoque de otros tipos de media como por ejemplo la media geométrica ponderada.
- Se propone seguir explorando métodos para el cálculo de los factores de ponderación.

RECOMENDACIONES

En un escenario de múltiples expertos, es importante tener en cuenta que las bases de datos se disponen de tal manera que se cuente con múltiples etiquetas para los datos, estas a su vez pueden contar con valores equivocados propios de la observación subjetiva de cada etiquetador generando un sesgo al momento de clasificar. Por esta razón es necesario consolidar estrategias autónomas que permitan establecer que expertos tienen mayor validez en su criterio, ya que si bien es posible realizar un análisis manual de los datos, en algunos casos, sería un esfuerzo innecesario y poco eficiente teniendo en cuenta la complejidad de los mismos.

Existe la necesidad de seguir explorando y desarrollando nuevos métodos de mezcla de clasificadores que permitan generar mejores resultados que los del método de la media ponderada, ya que, si bien el método propuesto brinda buenos resultados, existen otras técnicas no convencionales para realizar la mezcla de las funciones costo que se podrían implementar, y quizá puedan generar resultados aceptables.

Es necesario tener en cuenta que el simulador se realizó pensando en el usuario común, por esta razón se hizo de una manera que sea interactivo y de fácil uso, pero es recomendable que el usuario cuente con unos conocimientos básicos en el campo de clasificación y el reconocimiento de patrones; esto con el fin de que pueda entender como funciona el simulador e interprete de mejor manera los resultados obtenidos.

REFERENCIAS

[1] I. H. Witten y E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann, 2005.

[2] J. A. C. Ochoa y J. F. M. Trinidad, «Reconocimiento de patrones. de patrones.,» *Komputer Sapiens,* vol. 2, 2011.

[3] D. H. Peluffo-Ordóñez, S. M. Rendón, J. D. Arias-Londoño y G. Castellanos-Domínguez, «A multi-class extension for multi-labeler support vector machines,» *In ESANN,* 2014.

[4] S. Murillo Rendón, «Metodología para el aprendizaje de máquina a partir de múltiples expertos en procesos de clasificación de bioseñales,» *Doctoral dissertation, Universidad Nacional de Colombia-Sede Manizales.*

[5] S. Murillo-Rendón, D. Peluffo-Ordónez, J. D. Arias-Londono y C. G. Castellanos-Domínguez, «Multi-labeler analysis for bi-class problems based on soft-margin support vector machines,» *In International Work-Conference on the Interplay Between Natural and Artificial Computation.Springer Berlin Heidelberg,* pp. 274-282, 2013.

[6] K. T. Lai, X. Y. Felix, M. S. Chen y S. F. Chang, «Video event detection by inferring temporal instance labels,» *In 2014 IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2251-2258, 2014.

[7] M. X. Dueñas-Reyes, «Minería de datos espaciales en búsqueda de la verdadera información,» *Ingeniería y universidad,* vol. 13, nº 1, pp. 137-156, 2009.

[8] R. Benítez, G. Escudero, S. Kanaan y D. M. Rodó, Inteligencia artificial avanzada, Editorial UOC, 2014.

[9] N. J. Nilsson, R. Marín Morales, J. T. Palma Méndez y E. Paniagua Aris, Inteligencia artificial: una nueva sintesis, vol. 15, 2001.

[10] L. Á. Munárriz, Fundamentos de inteligencia artificial, Editum , 1994.

[11] C. P. López, Minería de datos: técnicas y herramientas, Editorial Paraninfo, 2007.

[12] L. Rokach y O. Maimon, Data mining with decision trees: theory and applications, World scientific, 2014.

[13] R. J. Roiger, Data mining: A tutorial-based primer, CRC Press, 2017.

[14] R. S. Michalski, J. G. Carbonell, T. M. Mitchell y (Eds.)., Machine learning: An artificial intelligence approach, Springer Science & Business Media., 2013.

[15] E. Alpaydin, Introduction to machine learning, MIT press, 2014.

[16] W. L. Chao, « Introduction to pattern recognition,» *National Taiwan University, Taiwan,* pp. 1-31, 2009.

[17] J. M. DE SA, Pattern recognition: concepts, methods and applications, Springer Science & Business Media, 2012.

[18] J. Ruiz-Shulcloper, A. Guzmán Arenas y J. F. Martínez-Trinidad, «Enfoque Lógico Combinatorio al Reconocimiento de Patrones,» de *Selección de Variables y Clasificación Supervisada, Primera edición.*, Ed. IPN, 1999.

[19] E. Alfaro, Combinacion de clasificadores mediante el metodo boosting. Una aplicacion a la prediccion del fracaso empresarial en España, 2006.

[20] L. I. Kuncheva, Combining pattern classifiers: methods and algorithms, John Wiley & Sons, 2004.

[21] P. Lison, An introduction to machine learning, 2015.

[22] C. A. Trujillo Pulgarín, «Clasificación basada en la estimación de Parzen en espacios generalizados de disimilitudes,» *Tesis Doctoral. Universidad Nacional de Colombia-Sede Manizales.*

[23] D. Boswell, Introduction to Support Vector Machines, 2002.

[24] J. Kittler, «Combining classifiers,» *Proceedings - International Conference on Pattern Recognition,* vol. 2, pp. 897-901, 1996.

[25] Y. Yan, G. M. Fung, R. Rosales y J. G. Dy, «Active learning from crowds.,» *In Proceedings of the 28th international conference on machine learning (ICML-11),* pp. 1161-1168, 2011.

[26] O. Dekel, C. Gentile y K. Sridharan, «Selective sampling and active learning from single and multiple teachers,» *Journal of Machine Learning Research,*

vol. 13, pp. 2655-2697, 2012.

[27] A. London y T. Csendes, «HITS based network algorithm for evaluating the professional skills of wine tasters,» *In Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on,* pp. 197-200, 2013.

[28] P. Smyth, U. Fayyad, M. Burl, P. Perona y P. Baldi, «Inferring ground truth from subjective labelling of venus images.,» pp. 1085-1092, 1995.

[29] O. Dekel y O. Shamir, « Good learners for evil teachers,» *In Proceedings of the 26th annual international conference on machine learning ACM.,* pp. 233-240, 2009.

[30] R. O. Duda, P. E. Hart y D. G. Stork, Pattern classification, John Wiley & Sons, 2012.

[31] F. Van Der Heijden, R. Duin, D. De Ridder y D. M. Tax, Classification, parameter estimation and state estimation: an engineering approach using MATLAB, John Wiley & Sons, 2005.

[32] P. D. J. M. Ramírez, Máquinas de vectores soporte en entornos de supercomputación: aplicación a fusión nuclear, 2014.

[33] L. Xu, A. Krzyzak y C. Y. Suen, «Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recongnitio,» *Ieee Transactions On Systems, Man, And Cibernetics,* vol. 22, nº 3, p. 18, 1992.

[34] D. P. G. C. S Murillo, «Support Vector Machine-based approach for,» *European Symposium on Artificial Neural Networks, Computational Intelligence,* pp. 479-484, 2013.

[35] S. R. Geoffrey Hinton, «Stochastic Neighbor Embedding,» *Advances in neural information,* pp. 833-840, 2002.

[36] D. H. Peluffo, J. A. Lee, M. Verleysen, J. L. Rodriguez y G. C. Dominguez, «Unsupervised relevance analysis for feature extraction and selection,» *International conference on pattern recognition, applications and methods,* p. 6, 2013.

[37] D. H. P. Ordoñez, A. E. C. Ospina, J. C. A. Pérez y E. J. R. Fuelagan, «Multiple kernel learning for spectral dimensionality,» *Progress in pattern recognition, applications and methods Springer,* p. 8, 2015.

[38] D. H. P. Ordoñez, Estudio comparativo de metodos de agrupamiento no supervisado de latido de señales ECG, 2009.

[39] G. H. a. C. Elkan, «Alternatives to the k-means algorithm that find better clusterings.,» *CIKM '02: Proceedings of the eleventh international conference,* pp. 600-607, 2002.

[40] D. P. González, Algoritmos de Agrupamiento basados en densidad y Validacion de clusters, Castellón, 2010.

[41] A. P. Benavent, Clustering EBEM. Modelos de Mezclas Gausianas Basados En Maximizacion de Entropia, 2007.

[42] M. Lichman, «UCI Machine Learning repository,» 2013. [En línea]. Available: http://archive.ics.uci.edu/ml/.

[43] E. Pekalska y W. R. P, «Pattern Recongnition Tools,» [En línea]. Available: http://37steps.com/.

[44] R. Pant y B. Trafalis, «Svm classification of uncertain data using robust multi-kernel methods.,» *Springer,* p. 9, 2015.

[45] G. Hamerly y C. Elakn, «Alternatives to the k-means algorithm that find better clusterings,» *Proceedings of the eleventh international conference on information and knowledge management,* 2002.

[46] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, Wiley, 2001.

[47] G. C. Canavos, Probabilidad y estadistica, aplicaciones y metodos, Mexico: McGraw-Hill, 1991.

[48] C. W. Hsu y C. J. Lin, «A comparison of methods for multiclass support vector machines,» *IEEE transactions on Neural Networks,* vol. 13, nº 2, pp. 415-425., 2002.

[49] V. C. Raykar, S. Z. L. H. Yu, G. H. Valadez, C. Florin, L. Bogoni y L. Moy, « Learning from crowds,» *Journal of Machine Learning Research,* vol. 11, pp. 1297-1322, 2010.

ANEXOS

Esta sección ha sido destinada a los resultados tangibles logrados con el trabajo realizado en esta tesis. Estos anexos contienen una descripción más ampliada de los resultados mencionados en la sección 6 donde exponemos los detalles más relevantes.

ANEXO 1. ARTICULO DE CONGRESO INTERNACIONAL CIARP

Este anexo contiene uno de los artículos realizados, enviado y aceptado en el XXI Iberoamericano Congress on Pattern Recognition (CIARP), cabe resaltar que el artículo fue presentado en ponencia modalidad poster en la ciudad de Lima Perú, lugar donde se llevó a cabo el evento. Este articulo será publicado a través de Springer en la revista Lecture Notes in Computer Science

Multi-labeler classification using kernel representations and mixture of classifiers

D. E. Imbajoa-Ruiz[1], I. D. Gustin[1], M. Bolaños-Ledezma[1], A. F. Arciniegas-Mejía[1], F. A. Guasmayan-Guasmayan[1,2], M. J. Bravo-Montenegro[2], A. E. Castro-Ospina[3], and D. H. Peluffo-Ordóñez[1,4] ⋆

[1] Universidad de Nariño, Colombia,
[2] Universidad Mariana, Colombia,
[3] Research Center of the Instituto Tecnológico Metropolitano, Medellín, Colombia,
[4] Universidad Técnica del Norte, Ecuador.

Abstract. This work introduces a multi-labeler kernel a novel approach for data classification learning from multiple labelers. The learning process is done by training support-vector machine classifiers using the set of labelers (one labeler per classifier). The objective functions representing the boundary decision of each classifier are mixed by means of a linear combination. Followed from a variable relevance, the weighting factors are calculated regarding kernel matrices representing each labeler. To do so, a so-called supervised kernel function is also introduced, which is used to construct kernel matrices. Our multi-labeler method reach very good results being a suitable alternative to conventional approaches.

Keywords: Multi-labeler classification, supervised kernel, support vector machines.

1 Introduction

Typically,supervised pattern recognition systems are trained by using prior knowledge- expressed by labels- given by a single expert labeler or annotator. Nonetheless, for some applications is suitable to consider a set of labelers rather than only one. For instance, a set of specialists diagnosing a patient's pathology [1] or a teacher team assessing the academic performance of a student [2]. Despite it is necessary, considering multiple labelers makes the classification problem difficult since there is no a clearly, identified ground truth. To classify data within this scenarios, multi-labeler strategies have been proposed, which should be able to both compensating the influence of wrong labels regarding the assumed ground truth [3], as well as identifying the good and bad labelers [4, 5]. In this connection, support-vector-machines- (SVM-) based approaches have shown to be a suitable alternative [3, 6–8]. Most of the currently available methods make strong assumptions on the resultant labeling vector, introducing then naturally noise over the classification task [6].

In this work, we present a novel approach for data classification within a multi-labeler approach. The classification is done by making a mixture of classifiers trained by using each labeler. Therefore, there is no assumptions made on the estimation of ground truth. Our method just naturally obtains an adjusted objective function defining an improved boundary decision. The proposed approach start by introducing a so-called supervised kernel, which is used to construct kernel matrices -one per labeler. Obtained matrices are incorporated within a variable relevance procedure to estimate their corresponding weighting factors. Finally, a mixture of classifiers trained by the labelers using the cost functions and the aforementioned weighting factors. Particularly, the used mixture is a linear combination. Experiments are carried out using the well-known Iris databases. Labeling vectors are created introducing different percentage of noise into the beforehand ground truth. The simulated labellers are penalized. Five different experiments are performed with several iterations each, to prove stability of our approach. Our multi-labeler method reach very good results being an alternative to conventional approaches. The great advantage of our method is that classification is performed with no assumptions on the mixture of labeling vectors. Instead, we perform a mixture of classifiers.

This paper is organized as follows: Section 2 reviews some related works and outlines basics on support vector machines and their extensions to multi-labeler scenarios. Section 3 describes the operation of the proposed multi-labeler approach in three subsections. Section 4 gathers some results and discussion. Finally, conclusions and final remarks are drawn in Section 5.

2 Related works and background

Given its versatility and outstanding performance in several applications, many approaches to deal with multi-labeler problems are formulated within support-vector machines (SVM) frameworks. In a previous work [9], a bi-class multi-labeler classifier (BMLC) is introduced. It starts from the simplest formulation for a bi-class or binary SVM-based classifier. For further statements, let us define the ordered pair $\{x_i, y_i\}$ to denote the i-th sample or data point, where x_i is its d-dimensional feature vector and $y_i \in \{1, -1\}$ is its binary class label. If considering a data set of N samples, all feature vectors can be gathered into a $N\ x\ d$ data matrix $\boldsymbol{X}$ such that $\boldsymbol{X} = [\boldsymbol{x_1} \dots, \boldsymbol{x_N}]^T$ whereas labels into a labeling vector $\bar{\boldsymbol{y}} \in \mathbb{R}^m$. As well, consider k labelers or labeling vectors $\{\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(k)}\}$ That said, the labeling vector $\bar{\boldsymbol{y}}$ is a reference vector to be determined. There are some approaches to estimate it. For

instance, by calculating the simple average as done in [3]. To pose the classifier's objective function, we assume a latent variable model in the form: $e_i = \boldsymbol{w^T x_i} + b = \langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b$, where w is a d-dimensional vector, b is a bias term and notation $\langle \cdot, \cdot \rangle$ stands for Euclidean inner product. As can be readily noted, vector $[e_1 \dots, x_m]$ results from a linear mapping of elements of $\boldsymbol{X}$ which, from a geometrical point of view, is an hyperplane and can thus be seen as a projection vector. By design, if assuming $w \in \mathbb{R}^d$as an orthogonal vector to the hyperplane, projection vector can be used to encode the class assignment by a decision function in the form $sign(e_1)$. Alternatively, projection vector can be expressed in matrix terms as $e = \boldsymbol{Xw} + b1_m$being 1_m an m-dimensional all ones vector.

Moreover, in order to avoid that data points lie in an ambiguity region for the decision making, the distance between the hyperplane and any data point can be constrained to be at least 1 by fulfilling the condition: $\bar{y}_i e_i \geq 1, \ \forall i$. The distance between data point x_i and hyperplane e can be calculated as: $d(e, \boldsymbol{x_i}) = \bar{\boldsymbol{y}}_i e_i / ||\mathbf{w}||^2$, where $||.||$ denotes Euclidean norm. Therefore, since the upper boundary of $d(e, \boldsymbol{x_i})$ is $1/||\mathbf{w}||^2$, one expect that $\boldsymbol{y}_i \simeq e_i$Then, the classifier objective function to be maximized can be written as: $\max_{w} \bar{\boldsymbol{y}}_i e_i / ||\mathbf{w}||^2; \forall i$. Accordingly, for accounts of minimization, we can write the problem so: $\min_{w} \frac{1}{2}||\mathbf{w}||^2$s.a. $\bar{\boldsymbol{y}}_i e_i = 1; \ \forall \boldsymbol{i}$. Notice that previous formulation is attained under the *hard* assumption that $\bar{\boldsymbol{y}}_i = e_i$ and can then be named as hard-margin SVM. By relaxing it, and by adding slack terms, a soft-margin SVM (SM-SVM) can be written as:

$$\min_{w,\xi} f(w, \xi | \boldsymbol{\lambda}) = \min_{w,\xi} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i^2 \text{ ; s.a. } \xi_i \geq 1 - \ \bar{\boldsymbol{y}}_i e_i \qquad (1)$$

Where λ is a regularization parameter and . ξ_i is a slack term associated to data point i.

Binary approach: Aimed at designing a multi-label classifier, in [6, 9]the SM-SVM given in equation 1 is modified by adding penalty factors $\theta_{t_{t=1}}^{k}$ and computing $\bar{\boldsymbol{y}}_i$ as the average of the set of the labeling vectors. This factor is intended to make f increases when adding wrong labels otherwise f should not or insignificantly decrease. In other words, consider a set of k labelers or panelists who singly provide their corresponding labeling vectors. Then, the t-th panelist's quality is quantified by the penalty factor θ_t. Accordingly, incorporating the penalty factors θ, a new binary classification problem is introduced by modifying problem stated in 1, as detailed in [9]. The solution of this problem is conducted by a primal-dual formulation as explained.

Multi-class approach: Another work [10] naturally extends this approach to multiclass scenarios by using a one-against-all strategy. Basically, this approach consists of building a number of SVM models -one per class. Applying c times the BMLC approach, a multi-class approach is accomplished. In general, in case of using SVMapproaches, class c is compared with the remaining ones in such a way that it is matched with a positive label, meanwhile the others with a negative label [11]; so that a binary labeling vector per each single class is formed. Concretely, the labeling reference vector $\bar{y}^{(\ell)}$ associated to class ℓ is assumed as a binarized version of labeling vector, as explained in [10]. In this sense, the BMCL is generalized to deal with more that two clases. Consequently, the decision hyperplanes are given by $\{e_1^{\ell}, \dots, e_i^{\ell}\}$, where $e^{\ell} = \boldsymbol{X}w^{(\ell)} + b^{(\ell)}\mathbf{1_m}$.

3 Proposed multi-labeler classification approach

Briefly put, the proposed model works as follows: It involves mainly three stages. We first introduce a so-called supervised kernel, which is used to construct kernel matrices-one per labeler. Such matrices are incorporated within a variable relevance procedure to estimate their corresponding weighting factors. Finally, such weighting factors are used to performing the mixture of classifiers. Following are explained our approach's stages.

3.1 Modified supervised kernel

Both the binary and multi-class problems explained in Section 2 can be solved by a means of a primal-dual formulation as explained in [10]. Typically, the dual problem is in the form: $\alpha - (1/2)\alpha^{\intercal} G\alpha$ α where α is the Lagrangian multiplier vector and $\boldsymbol{G}$ is a symmetric and positive semi-definite matrix. As demonstrated in [12], vectors $\alpha^{(\ell)}$ (one per class) pointed out a feasible direction where classes are readily distinguished as the kernel captures the nature of data. Indeed, kernel matrices represent data by means of pairwise similarities between data points. Here, we propose to incorporate the supervised information within the design of a kernel similarly as done in kernelized SVM classifiers [13]. Specifically, we introduce the modified kernel as

$$\boldsymbol{G}_{ij}^{(t)} = \sum_{\ell=1}^{c} \boldsymbol{y}_{\ell i}^{(t)} \boldsymbol{y}_{\ell j}^{(t)} \mathcal{K}(x_i, x_j), \tag{2}$$

where $\mathcal{K}(x_i, x_j)$ is a kernel function and $\boldsymbol{y}_{\ell i}^{(t)}$ stands for the binary label assignment to sample i regarding class ℓ given by labeler t given by:

$$\boldsymbol{y}_{\ell i}^{(t)} = \begin{cases} 1 & if \text{ x}_\text{i} \text{ belongs to class } \ell \\ -1 & otherwise \end{cases}, \tag{3}$$

The kernel matrix $\boldsymbol{G}^{(t)}$ must be calculated for each labeler to account for multilabeler settings, i.e. $t \in \{1, \dots, k\}$.

3.2 Multi-labeler approach

Our approach may result appealing since it is easy to solve by means of a quadratic programming search, given the form of the dual formualtion. However, as BMLC, solution is highly dependent on the chosen reference vector ~~y~~ as well as a no new coordinate axis is provided since only one vector $\boldsymbol{\alpha}$ is yielded. Furthermore, to design a multi-labeler approach from this formulation, the quadratic problem should be solved k times(one per labeler). Instead, we propose to perform a mixture of classifiers. Let us define f t (X) the trained cost function by using the labels given by the labeler t. Then, in order to take advantage of the information of the whole set of labelers, we propose a classifier whose cost function is the following mixture:

$$\bar{f}(\boldsymbol{X}) = \sum_{t=1}^{k} \eta_t f^{(t)}(\boldsymbol{X}), \tag{4}$$

3.3 Estimation of weighting factors

Now, we are intended to estimate η_t. Here, we propose to estimate the coefficients by using an adapted version of the variable ranking approach proposed in [14]. Similarly as in [15], we start by define a matrix $\boldsymbol{\mathcal{G}} \in \mathbb{R}^{N^2 x k}$ holding the vectorization of the kernel matrices $\boldsymbol{G}^{(t)}$, as well as a lower-rank representation $\widehat{\boldsymbol{\mathcal{G}}} \in \mathbb{R}^{N^2 x k}$. Regarding any othonormal matrix $\boldsymbol{U} = \left[\boldsymbol{u}^{(1)} \ldots \boldsymbol{u}^{(m)}\right] \in \mathbb{R}^{k x m}$, with $m < k$, we can write the lower-rank matrix as $\widehat{\boldsymbol{\mathcal{G}}} = \boldsymbol{\mathcal{G}}\boldsymbol{U}$. Following the variable relevance scheme proposed in [14], we can pose the following optimization problem:

$$\min_{\boldsymbol{U}} \left\| \boldsymbol{\mathcal{G}} - \widehat{\boldsymbol{\mathcal{G}}} \right\|_F^2 \quad \text{s. a.} \quad \boldsymbol{U}^\intercal \boldsymbol{U} = \boldsymbol{I}_m \tag{5a}$$

where $\| \cdot \|_F$ stands for Frobenius norm. Previous problem has a dual version being a maximization problem regarding the variance of $\widehat{\boldsymbol{\mathcal{G}}}$ as:

$$\max_{\boldsymbol{U}} \operatorname{tr}(\boldsymbol{U}^\intercal \boldsymbol{\mathcal{G}}\boldsymbol{\mathcal{G}}\boldsymbol{U}) \quad \text{s. a.} \quad \boldsymbol{U}^\intercal \boldsymbol{U} = \boldsymbol{I}_m \tag{6a}$$

Finally, the coefficients η_t for mixture are the ranking values quantifying how much each column of matrix $\boldsymbol{\mathcal{G}}$ (each kernel) contributes to minimizing the cost function given in (5a). Again, applying the variable relevance approach presented in [14], we can calculate the ranking vector $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_k]$ using:

$$\boldsymbol{\eta} = \sum_{t=1}^{k} \lambda_t \boldsymbol{u}^{(t)} \circ \boldsymbol{u}^{(t)}, \tag{7}$$

being λ_t and $\boldsymbol{u}^{(t)}$ the t-th eigenvalue and eigenvector of $\boldsymbol{\mathcal{G}}\boldsymbol{\mathcal{G}}$, respectively. Operator $\circ$ denotes Hadamard (element-wise) product. Given the problem formulation, positiveness of η is guaranteed and then can be directly used to perform the linear combination.

4 Results and discussion

Database: For experiments, the well-known open Iris flower database, extracted from UCI repository [16] is considered. It contains three different types of flowers: Iris Setosa, Iris Versicolor and Iris Virginica, with 50 samples each. Four characteristics were recorded for each sample: width and length of sepal, and width and length of petal. Likewise, there are two overlapping classes and a linearly separable class. This database is used to built and simulate different labels from several experts in order to show the characteristics of the method and its effects. Before carrying out the classification procedures, data matrix is normalized so that its maximum value per column be 1.

Methods: As reference methods, we consider the average and the majority vote of the given labeling vectors.

Parameter settings: To perform our multi-labeler approach, we use kernel Gaussian kernel whose ij entry are given by exp(-0.5||xi - xj ||/σ2). Experimentally we set σ = 0.65.

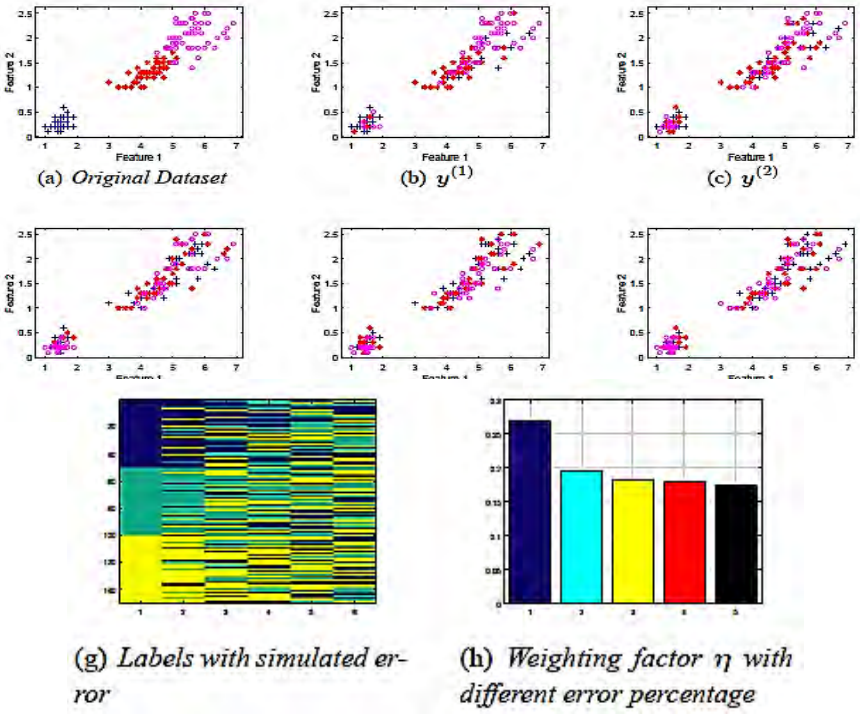
Performance measures: To quantify the performance of the considered multi-labeler approaches, conventional measures are used, such as: standard error, statistic mean, margin of error. Cohen's Kappa Index is measure also used in this work to eval-

uate the agreement relation between annotators. It is calculated considering the equallabeled individuals by the experts, where a total agreement equals a Kappa index of 100%, and no agreement at all, a Kappa index of 0%.

Experiments: To test the effectiveness of the proposed method, artificial annotators with different percentages of error in their labels are generated. In order to evaluate the stability of the presented approach, the procedure is iterated 30 times, and five cases are presented, with different induced error rates in the labellers. The labels' noise in labeling vectors are completely random, and the error rates in each annotator was chosen in several quantities and order to try the accuracy of the method. In Table 1, the assigned weights η are presented. These values were used in the experiments below, andare associated to the 'Proposed method' column in Table 2, where the overall results are presented.

The Kappa Index calculated and presented in Table 1 shows the agreement between labellers. In experiment 2, as all labellers have a low error rate, it is very likely that they have many labels in common. Thus, Kappa index is higher respect to the other cases. In the fourth experiment, unlike the second one, although the error rate is the same

Fig. 1. Generated labellers and data for Experiment No. 5



(a) *Original Dataset* (b) $y^{(1)}$ (c) $y^{(2)}$

(g) *Labels with simulated error* (h) *Weighting factor* η *with different error percentage*

Table 1. Weight η values

% η	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	Kappa Index
Experiment 1	25.41 ± 0.41	19.20 ± 0.15	18.52 ± 0.15	18.41 ± 0.19	18.19 ± 0.12	8.1 ± 1.2
Experiment 2	15.65 ± 0.19	21.70 ± 0.27	20.83 ± 0.20	19.46 ± 0.25	22.33 ± 0.39	21.0 ± 1.4
Experiment 3	22.97 ± 0.32	20.52 ± 0.21	19.15 ± 0.13	18.63 ± 0.10	18.71 ± 0.14	3.5 ± 0.8
Experiment 4	20.11 ± 0.18	20.09 ± 0.13	19.86 ± 0.09	19.90 ± 0.15	20.01 ± 0.15	2.0 ± 0.54
Experiment 5	25.12 ± 0.37	20.83 ± 0.40	18.12 ± 0.10	17.91 ± 0.11	18.00 ± 0.12	2.1 ± 0.92

Table 2. Performance results in terms of error percentage $\in$ of wrong classifications.

Experiment	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	Proposed method	Average	Majority vote
1	20	50	60	55	65	**15.59±3.02**	19.03±3.47	26.77±4.88
2	70	20	25	30	20	**4.01±0.82**	4.77±0.87	5.70±1.04
3	30	45	60	75	90	**21.85±4.53**	26.66±4.86	39.03±7.12
4	60	60	60	60	60	**20.74±3.78**	20.88±3.81	31.25±5.70
5	20	40	60	80	100	**12.16±3.16**	18.66±4.17	23.83±5.32

for all labellers, it is hardly expected that they share labels in common, so the Kappa index is very low. When the index presents a high value, it is expected that the standard deviation between the η values for a experiment should be low, and the annotators will be probably right, at least most of them, as shown in experiment 2. A low value of this index means that there is much disagreement between labellers, and that does not give much information about the accuracy of the method. This index should be taken into account only if presents high values.

Figure 1 depicts data generated used for the fifth experiment. Figure1(a) shows the original labels in the three classes. From Figure 1(b) to 1(f), the individual error corrupt data is shown for the five annotators. The misplaced labels can be noticed based on the colors of the clases. In Figure 1(g), contaminated labels are shown in a clearer way. Figure 1(h) shows the values of η for each annotator, representing the associated weight value. The classification accuracy in terms of percentage of wrong classifications is presented in Table 2. The error rate is decreased for all the cases using the method proposed. As the method assigns different weights to the annotators based on their certainty, the result error rate compared to the other methods is lower as the variation of the error in the labellers increases.

Experiment 1: Here, the first annotator $y^{(1)}$ has the lower error rate, unlike the others, surpassing the first one for at least 30%. In this case, it is observed that theweight η associated to that first annotator is higher, so his opinion will be more relevant in the mixture process. Thus, the performance of the mixture of classifiers will improve in the method proposed, as the worse annotators are not as considered as the first one.

Experiment 2: Unlike last experiment, the first annotator $y^{(1)}$ has now the higher error rate, with a 70%. The other labellers have a maximum of 30%. It is expected that the weight η of y(1) should be lower, so, his opinion will be proportionally ignored in the mixture process.

Experiment 3: An ascending error rate values is evaluated in this case, from 30% to 90%,15 by 15. As a general high error is presented among the annotators, it is expected that the final error is relatively high as well. Even so, the proposed method proves a better response to the noisy data.

Experiment 4: The case of same error rates in all labellers is assessed. The weight η is the same for each annotator, so the improvement in the results is slightly better.

Experiment 5: A similar case to the experiment 3, but in a wider range. The method proposed gives more importance in the mixture of classifiers to that labellers whose error rates are lower, that is $y^{(1)}$ and $y^{(2)}$, so the improvement respect to other methods is shown.

5 Conclusions and future work

Experimentally, we proved that the proposed approach is capable to quantify the confidence of a set of panelist taking into consideration the natural structure of data. Generally, the use of multi-labeler strategy may provide a better design and training of classifiers in comparison with one-labeler approaches. The here proposed method allows to deal with moderate noisy labels, with the capability to penalize labellers, keeping a good performance compared to conventional methods. For future work we are aiming to explore different alternatives that provide optimal procedures in the mixture of classifiers. Likewise, we are aiming to test other combination weighted methods to improve the classification, and make a sharper optimization for the kernel function for weight calculation, with a better and more strict penalizing for bad annotators.

References

1. Yan, Y., Fung, G.M., Rosales, R., Dy, J.G.: Active learning from crowds. In: Proceedings of the 28th international conference on machine learning (ICML-11). (2011) 1161–1168
2. Dekel, O., Gentile, C., Sridharan, K.: Selective sampling and active learning from single and multiple teachers. The Journal of Machine Learning Research **13**(1) (2012) 2655–2697
3. Dekel, O., Shamir, O.: Good learners for evil teachers. In: ICML. (2009) 30
4. Donmez, P., Carbonell, J.G., Schneider, J.: Efficiently learning the accuracy of labeling sources for selective sampling. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2009) 259–268
5. Wang, W., Zhou, Z.: Learnability of multi-instance multi-label learning. Chinese Science Bulletin **57**(19) (2012) 2488–2491
6. Murillo, S., Peluffo, D.H., Castellanos, G.: Support vector machine-based approach for multi-labelers problems. European Symposium on Artificial Neural Networks, Computational Inteligence and Machine Learning (2013)
7. Zhang, Y., Yeung, D.Y.: Multilabel relationship learning. ACM Transactions on Knowledge Discovery from Data (TKDD) **7**(2) (2013)
8. Cerri, R., de Carvalho, A.C.P., Freitas, A.A.: Adapting non-hierarchical multilabel classification methods for hierarchical multilabel classification. Intelligent Data Analysis **15**(6) (2011) 861–887
9. Murillo-Rendón, S., Peluffo-Ordóñez, D., Arias-Londoño, J.D., Castellanos-Domínguez, C.: Multi-labeler analysis for bi-class problems based on soft-margin support vector machines. In: Natural and Artificial Models in Computation and Biology. Springer (2013) 274–282
10. Peluffo-Ordóñez, D.H., Rendón, S.M., Arias-Londoño, J.D., Castellanos-Domínguez, G.: A multi-class extension for multi-labeler support vector machines. In: European Symposium on Artificial Neural Networks (ESANN). (2014)
11. Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. Neural Networks, IEEE Transactions on **13**(2) (2002) 415–425
12. Peluffo-Ordonez, D.H., Aldo Lee, J., Verleysen, M.: Generalized kernel framework for unsupervised spectral methods of dimensionality reduction. In: Computational Intelligence and Data Mining (CIDM), 2014 IEEE Symposium on, IEEE (2014) 171–177
13. Pant, R., Trafalis, T.B.: Svm classification of uncertain data using robust multi-kernel methods. In: Optimization, Control, and Applications in the Information Age. Springer (2015) 261–273
14. Peluffo, D.H., Lee, J.A., Verleysen, M., Rodríguez-Sotelo, J.L., Castellanos-Domínguez, G.: Unsupervised relevance analysis for feature extraction and selection: A distance-based approach for feature relevance. In: International conference on pattern recognition, applications and methods - ICPRAM 2014
15. Peluffo-Ordóñez, D.H., Castro-Ospina, A.E., Alvarado-Pérez, J.C., Revelo-Fuelagán, E.J.: Multiple kernel learning for spectral dimensionality reduction. In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. Springer (2015) 626–634
16. Lichman, M.: UCI machine learning repository (2013)

ANEXO 2. ARTICULO DE CONGRESO INTERNACIONAL INCISCOS

Este anexo contiene uno de los artículos realizados, enviado y aceptado en el congreso internacional INCISCOS llevado a cabo por la Universidad Tecnologica Equinoccial en la ciudad de Quito Ecuador. El artículo será publicado en un libro de actas con ISBN. aportando al avance de la informática y las telecomunicaciones.

Multi-Labeller classification Method based on Mixture of Classifiers and Genetic Algorithm Optimization

David E. Imbajoa R., Andrés F. Arciniegas M., Ivan D. Gustin, Mauricio Bolaños L., Dario F. Fajardo F.
Universidad de Nariño
Calle 18 No. 42
Pasto, Colombia
E-mail: deivy311@hotmail.com, arciandres@gmail.com, Ivanfly.92@gmail.com, arciandres@gmail.com, dario@udenar.edu.co

Fredy Guasmayán G., María J. Bravo
Universidad Mariana
—-
Pasto, Colombia
E-mail: ——

Diego H. Peluffo-Ordoñez˜
Universidad Técnica del Norte
Universidad Cooperativa de Colombia
—-
Pasto, Colombia
E-mail: dhpeluffoo@unal.edu.co

Abstract— This work presents a new method proposal applied to Multi-Labelers scenarios. This is a situation where labelling individuals in a set of data based on certain characteristics in the process of determining labels to individuals in a set of data based on certain characteristics. Our approach consists in processing a Support Vector Machine classifier to each labelers substantiated on his answers. We formulate a genetic algorithm optimization to obtain a set of weights according to their opinion, in order to penalize each panelist. Finally, their resulting mappings are mixed, and a final classifier is generated, showing to be better than majority vote. For experiments, the well-known Iris database is handled, with multiple simulated artificial labels. The proposed method reaches very good results compared to conventional multi-labeler methods, able to assess the concordance among panelists considering the structure data.

Abstract— Multi-labeller, multicriteria optimization, genetic algorithm, Gaussian distribution, support vector machine.

I. INTRODUCTION

Typically, the approaches of pattern recognition, based on supervised classification, require previous knowledge that takes into account a structure of labels, given by an unique expert or assessor. However, there are scenarios where information is assessed by multiple experts. That is to say, the characteristics of the data itself not only require supervision but also a range of opinions, so that its analysis obtains validity [1]. Some examples of these cases might be a group of specialists in the diagnosis of the pathology of a patient with specialized equipment [2] or the evaluation of the academic performance of a student [3]. In this type of scenario, where data is exposed to multiple interpretations, several factors that directly affect the adequate analysis of the information must be addressed. Among them, the inaccurate evaluation provided by the labelers that prevent a correct revision of the information. Therefore, it is necessary to find strategies that enable the reduction of the influence of the mistaken labels in relation to the real ones or the ground truth. The analysis of multiple experts focuses on the compensation of the negative effect of the mistaken labels. The mentioned compensation can improve the learning process in terms of factors of penalization or quantifying the efficiency of the evaluators [1] [4]. In particular, the support vector machines (SVMs) have shown to be a suitable alternative to approach this problem, mainly due to their versatility in regards to supervised classification [1]. In this paper, a new strategy for data classification contained inside an approach of multiple labelers is presented. The final classification is carried out using a variety of classifiers trained through the intervention of each labeler. Our method establishes a vector of decision variables that satisfies the restrictions and optimizes a function of vector whose elements represent the objective functions, as it generates values of decision in relation to the labelers.

The approach suggested initially provides a multi-objective criterion that, based on the functions generated by the labelled sets, estimating their respective optimal values with which the corresponding weighting values are generated. Finally, the combination of classifiers whose properties are established by means of the functions of cost and the factors of weighting aforementioned are carried out. The strategy proposed is evaluated on the database *IRIS* of the UCI learning machine. Label vectors are created entering in them, different noise percentages in relation to the vector designated as ground truth. N different experiments with m iterations were conducted to prove the stability of this approach, our multi-labeler method

accomplishes quite good results and stands for an efficient alternative in regards to conventional approaches.

The outline of this paper is as follows: Relevant related works are described in Section II. Section III explains the methods used in the proposed weighted Multi-labeler classification. Experimental results are shown in section IV. Finally, section V draws the conclusions and final remarks.

II. RELATED WORKS AND BACKGROUND

Many approaches to deal with multi-labeler problems are formulated within support-vector-machines (SVM) frame-works, due to its versatility and outstanding performance in several applications. For instance, a bi-class multi-labeler classifier (BMLC) is introduced in [5]. It starts from the simplest formulation for a bi-class or binary SVM-based classifier. Let us define the ordered pair $\{\boldsymbol{x}_i, \bar{y}_i\}$ to denote the i-th sample or data point, where $\boldsymbol{x}_i$its d-dimensional feature vector and $\bar{y}_i \in \{1, -1\}$, is its binary class label. All feature vectors can be gathered into a Nxd data matrix $\boldsymbol{X}$ such that $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N]^\intercal$, considering a data set of N samples, whereas labels into a labeling vector $\bar{\boldsymbol{y}} \in \mathbb{R}^m$. Consider k labelers or labeling vectors $\{\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(k)}\}$ as well. That said, there are some approaches to estimate the labeling vector $\bar{\boldsymbol{y}}$, which is a reference vector to be determined. By calculating the simple average as done in [6], for instance. We assume a latent variable model in the form: $e_i = \boldsymbol{w}^\intercal \boldsymbol{x}_i + b = < \boldsymbol{x}_i, \boldsymbol{w} > + b$ to pose the classifier's objective function, where $\boldsymbol{w}$ is a d-dimensional vector, b is a bias term and notation $<\cdot,\cdot>$ stands for Euclidean inner product. As can be readily noted, vector $\boldsymbol{e} = [e_i, \dots, e_m]$ results from a linear mapping of elements of $\boldsymbol{X}$, which is a hyperplane, from a geometrical point of view, and can thus be seen as a projection vector. By design, if assuming $\boldsymbol{w} \in \mathbb{R}^d$ as an orthogonal vector to the hyperplane, projection vector can be used to encode the class assignment by a decision function in the form $sign(e_i)$. Alternatively, projection vector can be expressed in matrix terms as $\boldsymbol{e} = \boldsymbol{X}\boldsymbol{w} + b\mathbf{1}_m$, being $\mathbf{1}_m$ an m-dimensional all ones vector.

In addition, the distance between the hyperplane and any data point can be constrained to be at least 1 by fulfilling the condition $\bar{y}_i e_i \geq 1$, $\forall_{\boldsymbol{i}}$, in order to avoid that data points lie in an ambiguity region for the decision making. The distance between hyperplane $\boldsymbol{e}$ and data point $\boldsymbol{x_i}$ can be calculated as: $d(\boldsymbol{e}, \boldsymbol{x}_i) = \bar{y}_i e_i / \|\boldsymbol{w}\|^2$, where $\|\cdot\|$ denotes Euclidean norm. Therefore, since the upper boundary of $d(\boldsymbol{e}, \boldsymbol{x}_i)$ is $1/\|\boldsymbol{w}\|^2$, one expect that $\bar{y}_i \cong e_i$. Then, the classifier objective function to be maximized can be written as: $max_{\boldsymbol{w}} \bar{y}_i e_i / \|\boldsymbol{w}\|^2; \forall_{\boldsymbol{i}}$. Consequentially, we can write the problem, for accounts of minimization, so: $min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2$, s. t. $\bar{y}_i e_i = 1$, $\forall_{\boldsymbol{i}}$. Notice that previous formulation is attained under the $hard$ assumption that $\bar{y}_i = e_i$, and can then be named as hard-margin SVM. By relaxing it, and by adding slack terms, a soft-margin SVM (SM-SVM) can be written as:

$$\min_{\boldsymbol{w},\xi} f(\boldsymbol{w}, \xi | \lambda) = \min_{\boldsymbol{w},\xi} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i^2 \quad s.t. \xi_i \geq 1 - \bar{y}_i e_i, \tag{1}$$

Where λ is a regularization parameter and ξ_i is a slack term associated to data point i.

Binary approach: Aimed at designing a multi-label classifier, in [4], [5], the SM-SVM given in equation 1 is modified by adding penalty factors $\theta_{j\,j=1}^{k}$ and computing $\bar{\boldsymbol{y}}$ as the average of the set of the labeling vectors. This factor is intended to make f increases when adding wrong labels otherwise f should not or insignificantly decrease. In other words, consider a set of k labelers or panelists who singly provide their corresponding labeling vectors. Then, the j-th panelist's quality is quantified by the penalty factor θ_j. Accordingly, with the penalty factors θ being included, a new binary classification problem is introduced by modifying problem stated in 1, as shown in [5]. As explained, the solution of this problem is accomplished by a primal-dual formulation.

Multi-class approach: Using a one-against all strategy, another work [7], naturally extends this approach to multiclass scenarios. This approach consists basically of building a number of SVM models -one per class. A multi-class approach is accomplished by applying c times the BMLC approach. In general, in case of using SVM-approaches, class c is compared with the remaining ones in such a way that it is matched with a positive label, meanwhile the others with a negative label [8]; so that a binary labeling vector per each single class is formed. Concretely, the labeling reference vector $\bar{\boldsymbol{y}}^{(\ell)}$ associated to class ℓ is assumed as a binarized version of labeling vector, as explained in [7]. In this sense, the BMCL is generalized to deal with more than two clases. Consequently, the decision hyperplanes are given by $\{\boldsymbol{e}_1^{(\ell)}, \dots, \boldsymbol{e}_i^{(\ell)}\}$, where $\boldsymbol{e}^{(\ell)} = \boldsymbol{X}\boldsymbol{w}^{(\ell)} + b^{(\ell)}\mathbf{1}_m$.

III. PROPOSED MULTI-LABELER CLASSIFICATION APPROACH

Unsupervised analysis covers all methods denominated as discriminative, which does not require a priori knowledge of the classes for classification. They usually require only one initialization parameter, like the number of resulting groups or any other indication about the initial partition. Then, the unsupervised analysis task is grouping homogeneous patterns without any information about the nature of classes present in the dataset. For this reason, the analysis does not generate unsupervised automatic classification, but generate s a homogeneous subset of data from some criterion based on distances, dissimilarities or statistical measures. Hence, the term of unsupervised classification refers to the grouping of data into subsets of similar elements and not some sort of automatic classification. There are several reasons why unsupervised methods are of special interest: converge quickly and they keep good performance if the characteristics change little over time, allowing categorizing items; they are useful when labeling a large set of samples is not feasible, among others. However, the solution generated by an unsupervised analysis system can be affected by factors such as inadequate initial parameters, which might generate a bad convergence, as explained in [9].

A. Multi-labeler approach

Our approach may result appealing since it is easy to solve by means of a quadratic programming search, given the form of the dual formulation. However, as BMLC, solution is highly dependent on the chosen reference vector $\bar{\boldsymbol{y}}$ as well as a no new coordinate axis is provided since only one vector α is yielded. Furthermore, to design a multi-labeler approach from this formulation, the quadratic problem should be solved k times (one per labeler). Instead, we propose to perform a mixture of classifiers. Let us define $f^{(j)}(\boldsymbol{X})$ the trained cost function by using the labels given by the labeler$j$. Then, in order to take advantage of

the information of the whole set of labelers, we propose a classifier whose cost function is the following mixture:

$$\bar{f}(\boldsymbol{X}) = \sum_{j=1}^{k} \eta_j f^{(j)}(\boldsymbol{X}) \tag{2}$$

Where η_t are the weighting factors to be defined.

B. Grouping based on centroids

The general idea of grouping based on centroids, is to minimize or maximize an objective function, which defines how good the solution pooling is. To achieve this, we use a method based on Gaussian Expectation Maximization, commonly used in clustering applications [10]. A generalized way to perform this grouping may be obtained by studying the proportion or degree of belonging of an element to a group, and the influence of each element in the centroid's updating. And the resulting partition for each iteration corresponds to the allocation of the subset elements whose centroid is nearest. Variants of these algorithms consist on changes of the objective function and therefore the update function centroids.

1) Gaussian Expectation maximization Mixture: (GEMM)

It is part of clustering methods based on probability density (DBC) and its objective function is the linear combination of Gaussian distributions centered in the centroids of each group, as follows:

$$GEMM_{log}(\mathbf{X}, \mathbf{C}) = -\sum_{i=1}^{C} log\left(\sum_{j=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{q}_i)p(\boldsymbol{q}_j)\right), \tag{3}$$

Where $p(\boldsymbol{x}_i|\boldsymbol{q}_i)$ is the probability of $\boldsymbol{x}_i$ since it is generated by a Gaussian distribution centered in $\boldsymbol{q}_j$, $p(\boldsymbol{q}_j)$ is the probability a priori of the group whose centroid is $\boldsymbol{q}_j$.

Another alternative to compute the objective function is by using an exponential operator:

$$GEMM_{log}(\mathbf{X}, \mathbf{C}) = -\sum_{i=1}^{C} exp\left(\sum_{j=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{q}_i)p(\boldsymbol{q}_j)\right), \tag{4}$$

When the estimated set of these probabilities present an elevated dispersion, the method with the logarithm function is used, as it segments the extraction of classification results, through the estimation of each distribution peak, maximizing the plausibility of each annotator; When the dispersion is low, the method with the exponential function should be used, which is a more aggressive method, as it segments the evaluation results in a more specific region. The minus sign is fixed in order to set a minimization operation with the objective function. The respective membership functions of each element are:

$$m_{GEMM}(\boldsymbol{q}_j|\boldsymbol{x}_i) = \frac{p(\boldsymbol{x}_i|\boldsymbol{q}_i)p(\boldsymbol{q}_j)}{p(\boldsymbol{x}_i)}, \tag{5}$$

Notice that the membership function is a probability value, thus Bayes' rule can be used to calculate its value, considering $p(\boldsymbol{x}_i)$ as evidence:

$$p(\boldsymbol{x}_i) = \sum_{j=1}^{k} p(\boldsymbol{x}_i|\boldsymbol{q}_j)p(\boldsymbol{q}_j) \tag{6}$$

$p(\boldsymbol{x}_i|\boldsymbol{q}_j)$ factor can be obtained easily with:

$$p(\boldsymbol{x}_i|\boldsymbol{q}_j) = \frac{1}{\det(\Sigma_j)^{\frac{1}{2}}}(2\pi)^{-\frac{d}{2}} e^{-\frac{1}{2}(\boldsymbol{x}_i-\mu)\Sigma_j^{-1}(\boldsymbol{x}_i-\mu)^{\intercal}}, \tag{7}$$

Where μ is the centroid ($\mu = \boldsymbol{q}_j$), d is dimension, Σ represent the covariance and $\det(\cdot)$ denotes the matrix determinant argument.

Objective functions to be minimized are given by:

$$\boldsymbol{F}_{exp} = -\sum_{i=1}^{C} \exp\big(p(\boldsymbol{x}_i)\big), \tag{8}$$

$$\boldsymbol{F}_{log} = -\sum_{i=1}^{C} \log\big(p(\boldsymbol{x}_i)\big), \tag{9}$$

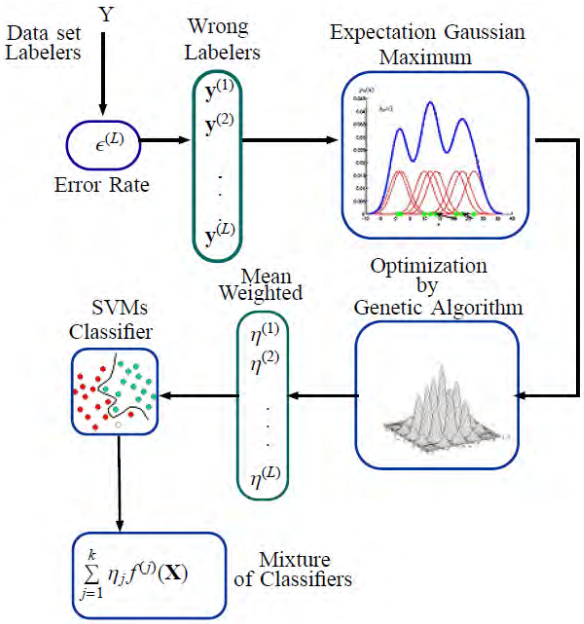


Figure 1: Process Diagram of Multi-labeler approach

C. Genetic algorithm for weights estimation

For estimation of weights for pattern classification, a genetic algorithm is chosen. It solves iteratively the optimum values for the weights in a multi-criteria objective function, depicted in 6, which consists in maximizing the sum of the Gaussian distribution for each labeler in every object to be classified, and also taking into account the probability a priori of each Gaussian distribution. For this purpose, the Pareto optimization method is used [11]. A Gaussian distribution function is generated for each class, labelled by each expert. And for each resulting classification, an objective

function is generated to be optimized. In Figure 1, the process diagram of the proposed approach is depicted. To achieve better results, according to the dispersion of each classifier's outcome, the objective functions in 8 and 9, are subject to:

$$\sum_{j=1}^{k} \eta_j = 1, \tag{10}$$

Where $\eta_j \in [0, 1]$

IV. RESULT AND DISCUSSION

Database: Open Iris flower database, extracted from UCI repository [12] is considered for experiments. Three different types of flowers are contained, with fifty samples each: Versicolor, Virginica and Setosa. For each sample, four characteristics were registered: width and length of petal and sepal. Moreover, there is

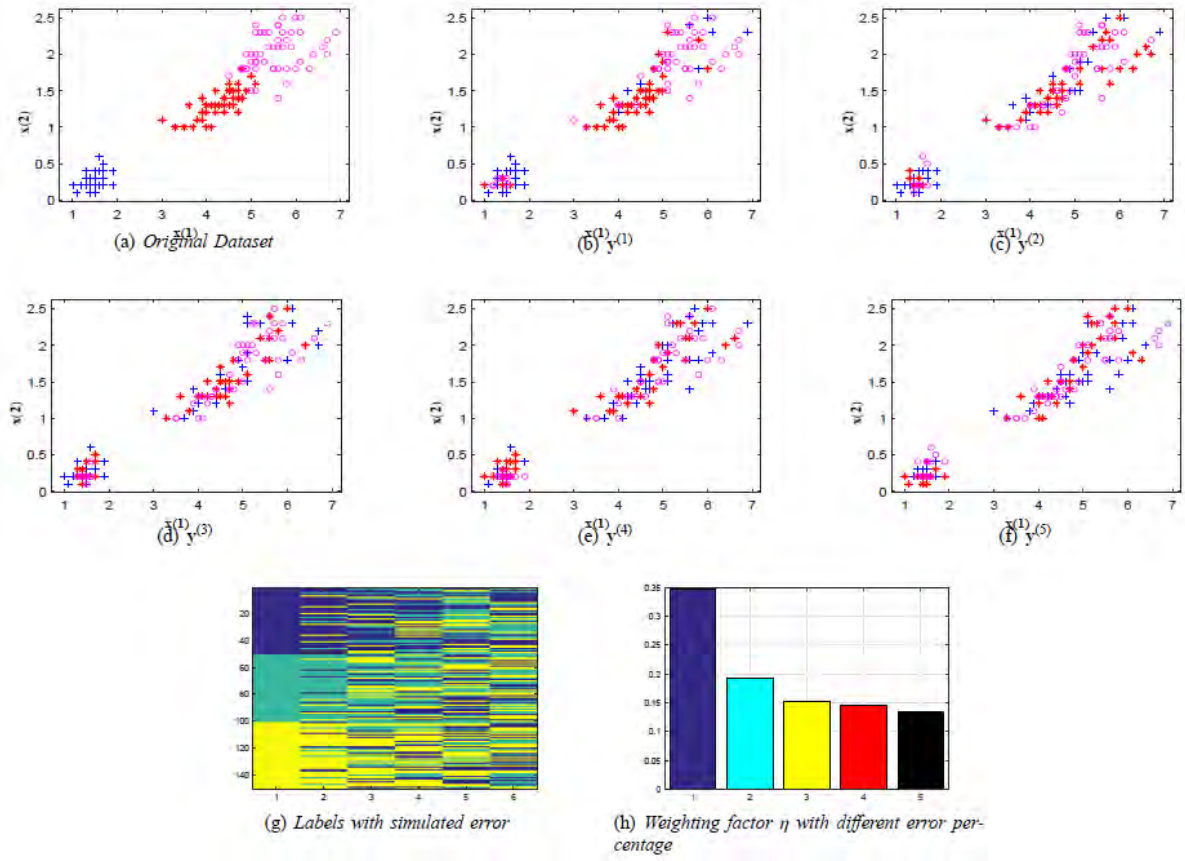


Figure 2: Generated Data and artificial Labelers for Experiment No. 5. Feature 1 is Petal length, and Feature 2 is Petal width. Scatter plots are displayed for *Ground Truth* and the five labelers.

A linearly separable class, and the two classes left are overlapped. The results presented here take into account petal length and width only. Different labels from several annotators are simulated and built using this database, in order to show the method effects and characteristics. Data matrix is normalized so that its maximum value per column be 1, before carrying out the classification procedures.

Methods: As reference methods, we consider the average and the majority vote of the given labeling vectors.

Parameter settings: To perform our multi-labeler approach, we use a multi-objective optimization with a genetic algorithm, with a random weights restricted between 0 and 1, as shown in equations (10)(11); a size of 20 individuals, and stopping criterion given by a tolerance of 0.01% error; tournament selection and 5% mutation rate.

Performance measures: Conventional measures are used to quantify the performance of the considered multi-labeler approaches, such as: standard error, statistic mean and margin of error. Cohen's Kappa Index is also used in this work to evaluate the agreement relation between annotators. It is calculated considering the equal labeled individuals by the experts, where a total agreement equals a Kappa index of 100%, and no agreement at all, a Kappa index of 0%.

Experiments: Simulated annotators with different percentages of error in their labels are generated to evaluate the efficiency of the method. The process is iterated thirty time s in order to reach the stability of the approach. Five cases are depicted, with different

induced error rates in the annotator. Noise of data in labeling vectors are completely random, and, in order to try the accuracy of the method, error rates in each annotator were chosen in several different quantities. The assigned weights η are presented in Table II. These values were used in the experiments below, and are associated to the 'Proposed method' column in Table I, where the general results are depicted.

The agreement between labelers in Table II are shown by Kappa Index calculations. In experiment 4, although the error rate is the same for all labelers, it is hardly expected that they share labels in common, so the Kappa index is very low. In experiment 2, as all labelers have a low error rate unlike fourth one, it is very likely that they have many choices in common. Thus, Kappa index is higher respect to the other cases. When the index presents a high value, it is expected that the standard deviation between the η values for an experiment is low, and the annotators will be probably right, at least most of them, as shown in experiment 2. Nevertheless, a low value of this index implicates a higher level of disagreement between labelers, and that does not give much information about the accuracy of the method. This index should be taken into account only if presents high values.

Table I: Performance results in terms of error percentage ∈ of wrong classifications.

Experiment	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	Method(log)	Method(exp)	Average	Majority vote
1	10	20	30	40	50	**2.22±0.36**	**3.27±1.05**	6.50±0.64	6.05±0.56
2	20	40	60	80	100	**5.56±1.55**	**8.67±2.20**	15.55±2.02	23.72±1.95
3	20	50	60	55	65	**2.42±1.59**	**9.02±2.25**	10.11±2.09	20.16±2.02
4	70	20	25	30	20	**4.33±0.57**	**5.44±1.81**	7.44±0.72	8.83±0.79
5	30	45	60	75	90	**2.22±1.75**	**18.88±4.55**	23.11±1.36	25.56±1.85
6	60	60	60	60	60	**15.77±5.97**	**11.57±4.1**	28.94±2.86	23.94±2.64
7	5	5	5	5	5	**3.51±0.37**	**2.18±0.9**	5.6±0.58	6.66±2.64

Table II: Weight η values (exp)

% η	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$	Kappa Index
Experiment 1	65.04 ± 4.23	16.25 ± 1.35	9.14 ± 2.05	5.83 ± 1.54	3.72 ± 1.13	2.4 ± 0.82
Experiment 2	60.54 ± 5.59	16.67 ± 4.27	13.10 ± 7.52	5.38 ± 1.68	4.29 ± 1.21	2.1 ± 0.92
Experiment 3	60.52 ± 7.21	16.29± 7.19	8.3 ± 1.16	10.36 ± 3.49	4.51 ± 0.86	8.1 ± 1.2
Experiment 4	3.07 ± 0.72	27.62 ± 3.85	15.77 ± 3.09	17.3 ± 4.52	36.34 ± 5.96	21.0 ± 1.4
Experiment 5	48.25 ± 6.35	13.42 ± 2.57	14.43 ± 2.97	7.99 ± 1.79	14.72± 3.75	3.5 ± 0.8
Experiment 6	17.61 ± 3.13	21.13 ± 2.74	2.17 ± 1.86	24.85 ± 4.96	24.22 ± 6.77	2.0 ± 0.54
Experiment 7	16.48± 2.57	19.09± 3.04	14.82 ± 2.99	22.02 ± 5.27	27.57 ± 5.27	2.0 ± 0.54

Table III: Weight η values (log)

% η	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	$y^{(4)}$	$y^{(5)}$
Experiment 1	57.76 ± 7.02	21.58 ± 5.78	7.36± 1.5	4.85 ± 1.22	8.42 ± 2.81
Experiment 2	57.48 ± 6.7	26.72 ± 2.52	5.27 ± 2.33	5.99 ± 1.85	4.52 ± 1.73
Experiment 3	63.68 ± 7.05	13.56± 2.15	6.03± 1.38	7.13± 1.69	9.57± 3.56
Experiment 4	7.38 ± 1.67	26.9± 5.85	16.78± 3.44	16.19± 3.70	32.73 ± 4.43
Experiment 5	46.44 ± 8.37	28.36 ± 5.48	8.63 ± 1.56	7.07± 1.85	9.48± 4.5
Experiment 6	18.31 ± 6.34	17.58± 2.56	16.48± 2.95	23.72 ± 1.79	23.8 ± 3.74
Experiment 7	15.7 ± 3.94	18.03 ± 5.15	15.19± 4.14	31.39 ± 9.77	19.66 ± 5.99

Figure 2 depicts generated data used for the second experiment. Figure 2(a) shows the original labels in the three classes. From Figure 2(b) to 2(f), the corrupt individual error data is shown for the five annotators. The misplaced labels can be noticed based on the colors of the classes. In Figure 2(g), contaminated labels are shown in a clearer way. Figure 2(h) shows the values of _ for each annotator, representing the associated weight value. The classification accuracy in terms of percentage of wrong classifications is presented in Table I. The error rate is decreased for all cases using the proposed method. As the approach assigns different weights to the annotators based on their certainty, the result error rate, compared with the other methods, is lower as the variation of the error in the labelers increases.

Experiment 1: In this case, the classifiers mixture with error percentage in the range of 10 percent in upward way until 50 percent is shown. In Tables II and III, weights for each labeler seen in Experiment 1 are corresponding with the error percentage shown in the Table I.

Experiment 2: This is a similar case to the experiment 1, but in a wider range. The proposed method gives more importance in the mixture of classifiers to that labelers whose error rates are lower. These are $\boldsymbol{y}^{(1)}$ and $\boldsymbol{y}^{(2)}$, so the improvement respect to other methods is shown.

Experiment 3: The first annotator $\boldsymbol{y}^{(1)}$ has the lower error rate in this case, unlike the others, surpassing the first one for at least 30%. In this case, it is observed that the weight η associated to that first annotator is higher, so his opinion will be more relevant in the mixture process. Thus, the performance of themixture of classifiers will improve in the proposed method as the worse annotators are not as considered as the first one.

Experiment 4: Unlike last experiment, the first annotator $\boldsymbol{y}^{(1)}$ has now the higher error rate, with a 70%. The other labelers have a maximum of 30%. It is expected that the weight η of $\boldsymbol{y}^{(1)}$ is lower, so his opinion will be proportionally ignored in the mixture process.

Experiment 5: Ascending error rate values are evaluated in this case, from 30% to 90%, 15 by 15. As a general high error is presented among the annotators, it is expected that the final error is relatively high as well. An outstanding performance of the proposed method is evidenced, where a total error avoidance is accomplished.

Experiment 6: The case of same error rates in all labelers is assessed. The weight η is the same for each annotator, so the improvement in the results is slightly better.

Experiment 7: This is a similar case to the last one, but with a lower general error rate. The method with the exponential function, shown in equation (4), presents better results than the rest of the methods, as expected.

An accomplishment of this work is the clear recognition of the best labelers, with only the natural structure of data. A general improvement is evidenced in the quality measures in every experiment performed, compared with the other conventional methods. We highlight the relevance of assigning weight to the opinions of the experts.

V. CONCLUSION AND FUTURE WORK

We proved experimentally that the proposed approach is capable of quantifying the confidence of a set of reliable labels, taking into account the given information by a group of experts and the variation in the natural structure of the data. In general, the use of this multi-labeler strategy provides a significant improvement in the classifiers design in comparison to the single-labeler approach. In addition, the proposed method has the capability of reducing the influence of wrong labelers establishing penalties and punishing to these bad experts, keeping a good performance in comparison with conventional methods. For future work, we are aiming to explore different alternatives for optimization procedures, to find more suitable penalty values that allow to identify bad annotators in a clearer way, and reduce their relevance. We are aiming also to explore different data sets and multi-labeler cases to apply and improve the algorithm.

REFERENCES

[1] O. Dekel and O. Shamir, "Good learners for evil teachers," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 233–240.

[2] Y. Yan, G. M. Fung, R. Rosales, and J. G. Dy, "Active learni ng from crowds," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 1161–1168.

[3] O. Dekel, C. Gentile, and K. Sridharan, "Selective sampli ng and active learning from single and multiple teachers," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2655–2697, 2012.

[4] S. Murillo, D. Peluffo, and G. Castellanos, "Support vec tor machine-based approach for multi-labelers problems," in *European Symposium on Artificial Neural Networks, Computational Inteligence a nd Machine Learning*, 2013.

[5] S. Murillo-Rendon,´ D. Peluffo-Ordonez,´ J. D. Arias-Londono, and C. G. Castellanos-Dom´ınguez, "Multi-labeler analysis for bi-class problems based on soft-margin support vector machines," in *Natural and Artificial Models in Computation and Biology*. Springer, 2013, pp. 274–282.

[6] O. Dekel, "From online to batch learning with cutoff-ave raging," in *Advances in Neural Information Processing Systems*, 2009, pp. 377–384.

[7] D. H. Peluffo-Ordo´nez,˜ S. M. Rendon,´ J. D. Arias-Londono,˜ and G. Castellanos-Dom´ınguez, "A multi-class extension for multi-labeler support vector machines." in *ESANN*, 2014.

[8] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multicl ass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[9] D. H. Peluffo Ordonez, J. L. Rodriguez Sotelo, and C. G. Castel-lanos Dominguez, "Estudio comparativo de m etodos´ de seleccion´ de caracter´ısticas de inferencia supervisada y no supervisada," 2009.

[10] G. Hamerly and C. Elkan, "Alternatives to the k-means algo rithm that find better clusterings," in *Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 600–607.

[11] K. Deb, *Multi-objective optimization using evolutionary algorithms*. John Wiley & Sons, 2001, vol. 16.

[12] M. Lichman, "UCI machine learning repository," 2013. [On line]. Available: http://archive.ics.uci.edu/ml

ANEXO 3. MANUAL DE USUARIO

Manual de usuario
Simulador de Reconocimiento de Patrones con múltiples expertos

DEPARTAMENTO DE INGENIERIA ELECTRONICA UNIVERSIDAD DE NARIÑO

https://sites.google.com/site/multipleexpertssimulator/
AÑO 2016

Manual de usuario

1. Introducción

El presente escrito está dirigido a otorgar las pautas de funcionamiento y operación del simulador de reconocimiento de patrones con múltiples expertos. Este simulador permite interactuar con un escenario de múltiples expertos y además realizar la clasificación de diferentes bases de datos.

La interfaz fue diseñada pensando en un escenario donde se emulan 5 diferentes conjuntos de etiquetas con el conjunto originalmente seleccionado, de esta manera se tiene un entorno de múltiples etiquetadores donde se realiza la clasificación para cada conjunto por medio de las máquinas de soporte vectorial (SVM) generando diferentes mappings por cada conjunto de etiquetadores.

Para realizar la mezcla de los mappings generados por las máquinas de soporte vectorial se puede optar por tres diferentes métodos de mezcla: Modo manual, modo Kernel y la aproximación basada en centroides. Además cuenta con la opción de comparar con otros métodos de mezcla convencionales como la media aritmética y el voto mayoritario.

2. Funcionamiento

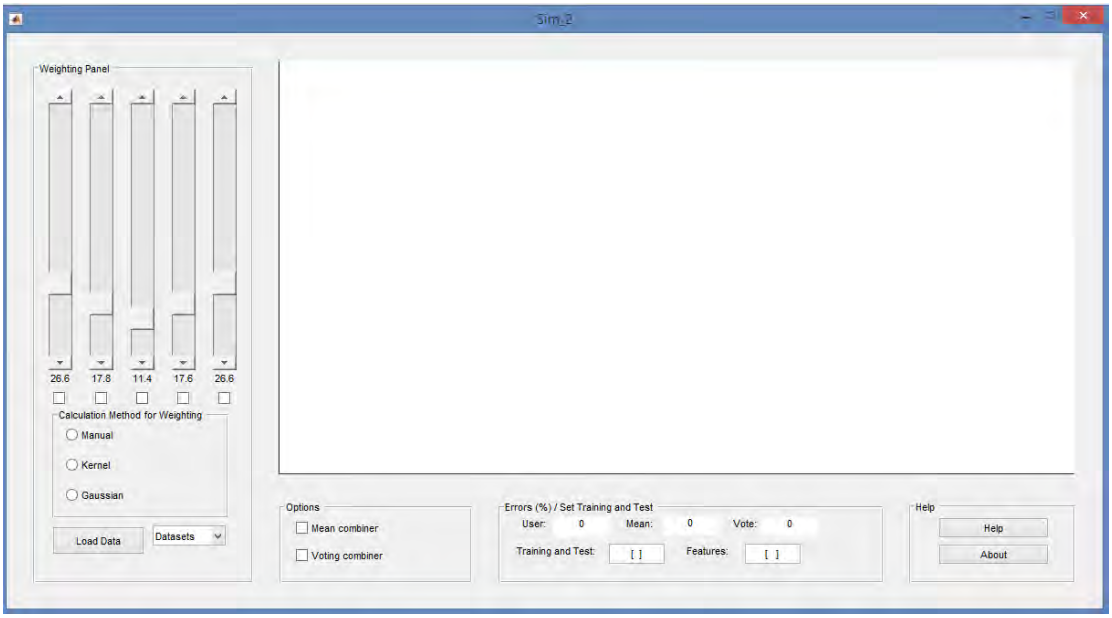


Figura 1. Simulador de múltiples expertos

❖ Selección de la base de datos:

Se puede seleccionar entre diferentes bases de datos predeterminadas en la sección datasets.

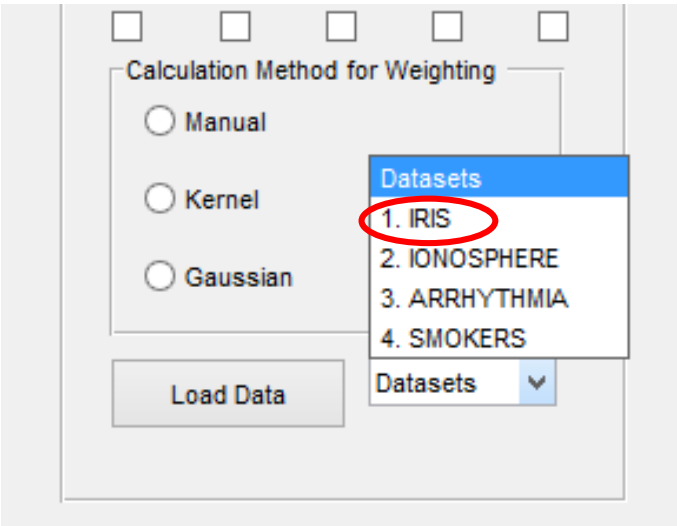


Figura 2. Sección Datasets

Una vez que haya seleccionado una base de datos aparecerá un cuadro de información donde se muestra algunas características sobre la base de datos seleccionada como por ejemplo el número de clases, los atributos y el número de elementos.

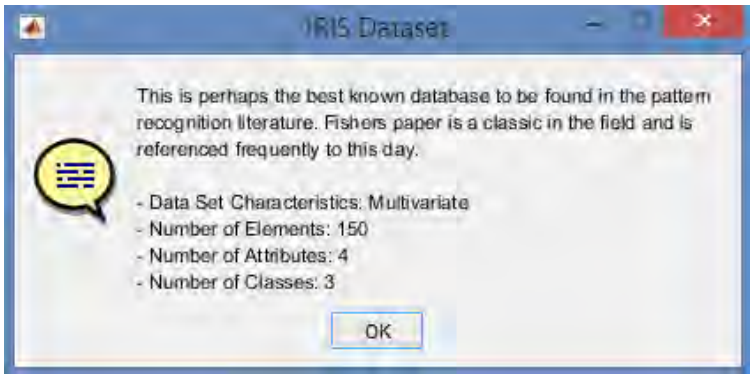


Figura 3. Cuadro de información

En la figura 4 se observa el cuadro de Error/Set training and test, aquí se tiene la opción para escoger los conjuntos de entrenamiento y prueba para los clasificadores. En este caso esta predeterminado en [20 20 20] pero el usuario puede escoger a su gusto, así mismo es posible modificar el campo de atributos, en este caso la base de datos IRIS cuenta con 4 características para este ejemplo se utilizara el atributo 3 y 4.

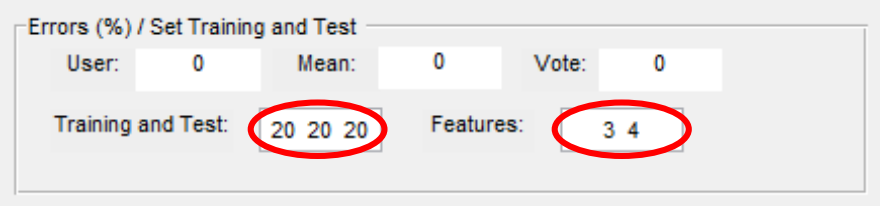


Figura 4. Cuadro de Error/Set training

Luego se procederá a presionar el botón LOAD DATA, con esto se cargara la base de datos y además será visible en pantalla como en la siguiente imagen. En este caso seleccionaremos la base de datos IRIS (usted puede seleccionar cualquier otra) en caso de el usuario requiera cambiar de base de datos en cualquier momento deberá repetir el proceso anterior.

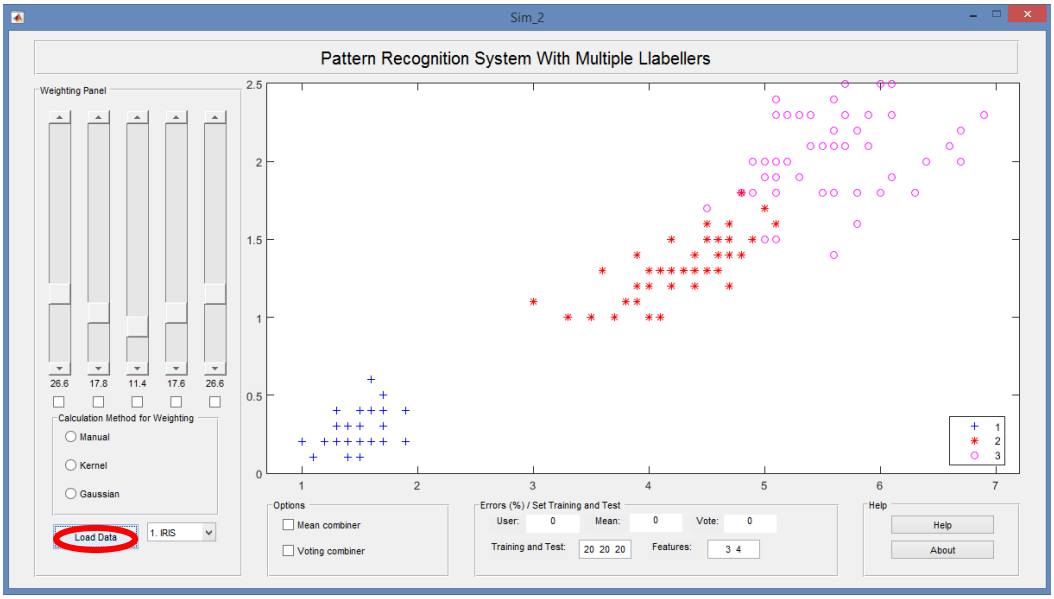


Figura 5. Simulador con la base de datos IRIS cargada

❖ Método de mezcla para clasificadores:

Para realizar la mezcla del conjunto de etiquetadores es posible optar por tres métodos: el método manual, el método basado en matrices kernel (Kernel) y la aproximación mediante centroides (Gaussian).

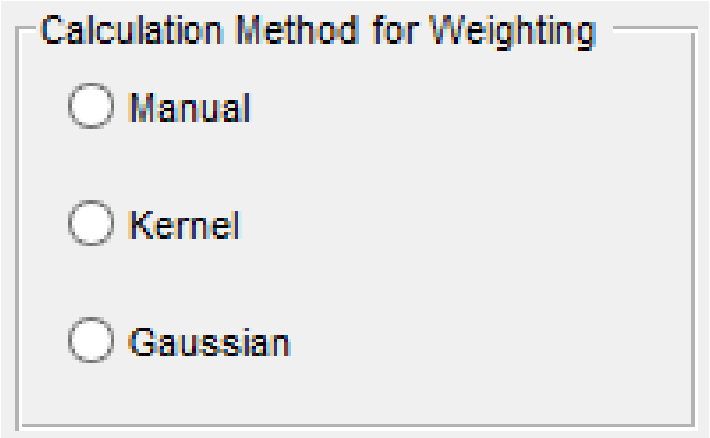


Figura 6. Cuadro de opciones para el método de selección de ponderación

- **Modo manual:** en el modo manual se tienen diferentes opciones que se pueden configurar como desee el usuario.

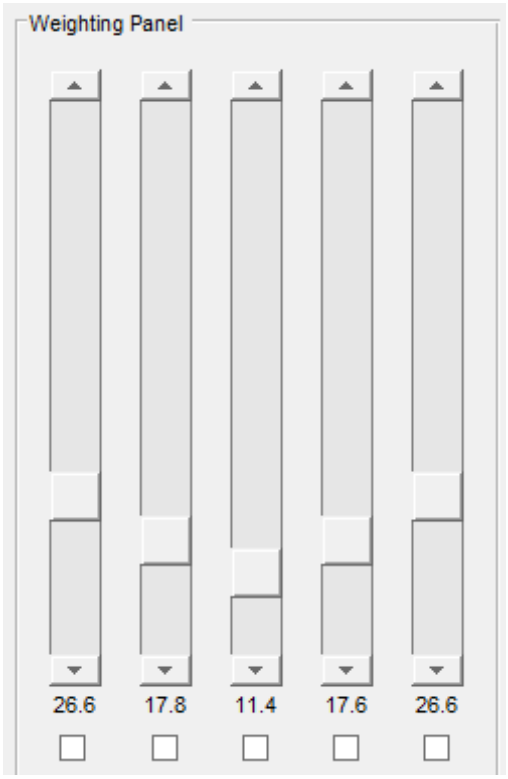


Figura 7. Panel de ponderación

Observemos el panel de ponderación, se tienen cinco sliders donde cada uno representa el peso que se le otorgara a cada etiquetador, entre más alto sea el valor que se le dé a un etiquetador mayor peso tendrá este al momento de realizarse la mezcla.

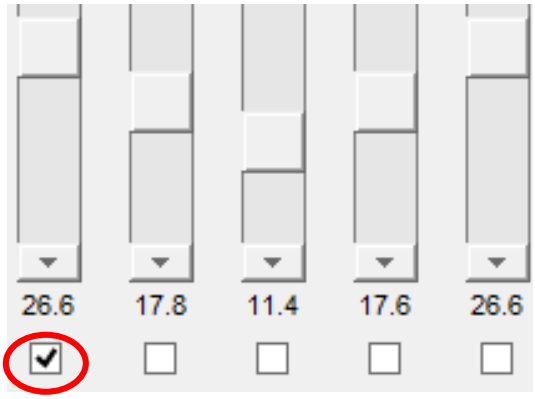


Figura 8. Panel de bloqueo de sliders

Como se puede observar en la imagen anterior en el círculo rojo se tiene la opción de bloquear un slider, con esto solo podrá cambiar el slider que se encuentre seleccionado. De esta manera podrá aumentar o disminuir su valor de ponderación, Así mismo los sliders restantes cambiaran su valor aumentando o disminuyendo respectivamente en la misma proporción.

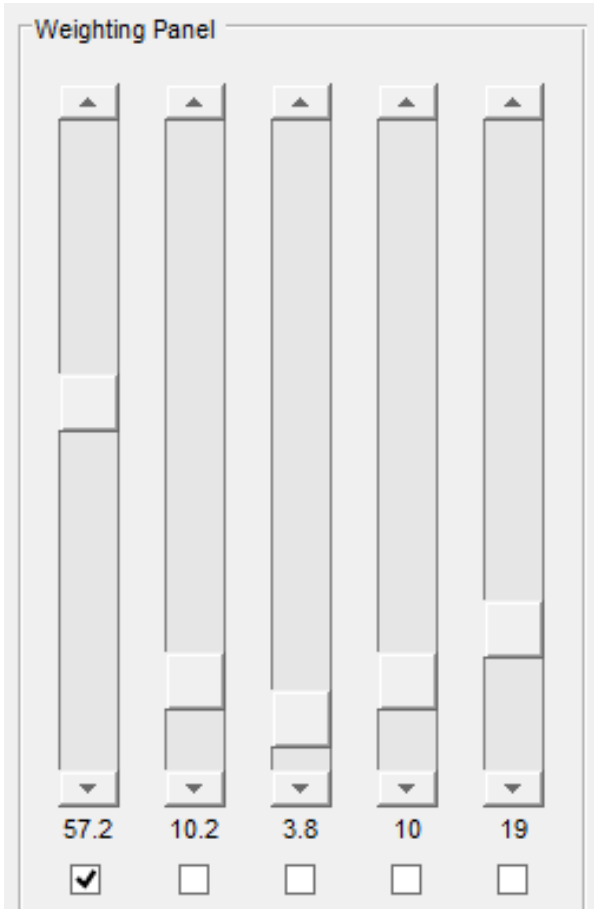


Figura 9. Activando el panel de bloqueo en el primer slider

Después de seleccionar los parámetros para la mezcla se debe entrar en el modo manual seleccionando la opción manual, luego se trazara un mapeo con los parámetros configurados como vemos a continuación:

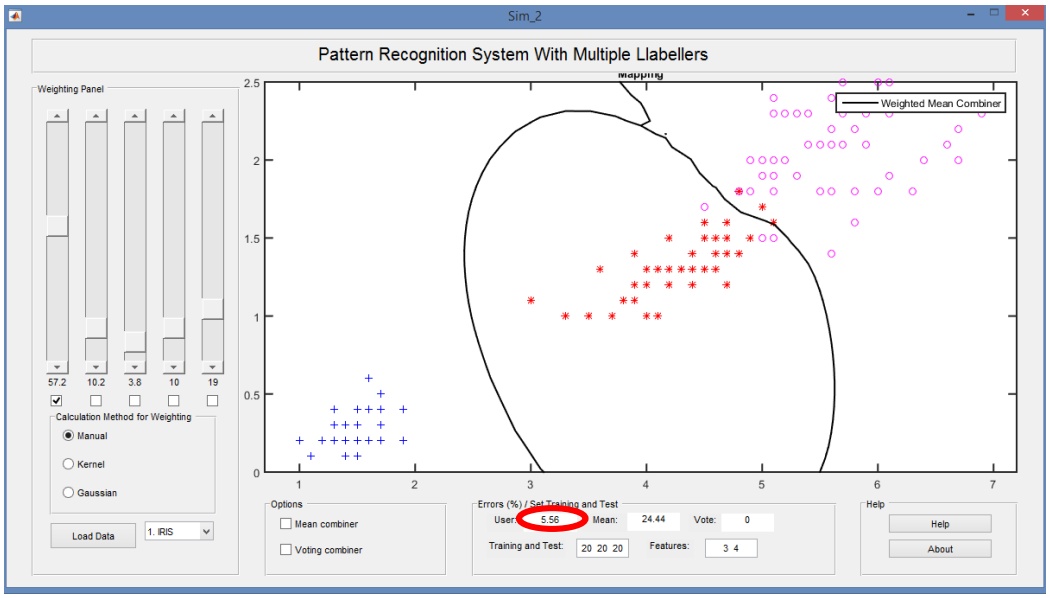


Figura 10. Mapping de la mezcla utilizando el modo manual

En el círculo rojo se observa el error porcentual que se obtiene al realizar la clasificación.

- **Modo kernel:** en el modo kernel la interfaz calcula automáticamente los pesos de ponderación para realizar la mezcla de clasificadores. Para entrar en el modo kernel solo se debe presionar el botón **Kernel** y se trazara el mapeo correspondiente**.**

- **Modo gaussian:** en el modo gaussian la interfaz calcula automáticamente los pesos de ponderación para realizar la mezcla de clasificadores por medio del método basado en centroides. Para entrar en el modo kernel solo se debe presionar el botón **Kernel** y se trazara el mapeo correspondiente**.**

- **Opción media aritmética y voto mayoritario:** el simulador cuenta con dos opciones de mezcla de clasificadores convencionales: la media aritmética y el voto mayoritario. Para acceder a ellas es necesario haber seleccionado antes

algún método de selección para los factores de ponderación, ya que de otra manera no será posible realizar ninguna de las dos opciones anteriores.

Puede escoger las dos opciones al mismo tiempo o solo una, si el usuario desea volver al mapping original solo debe desmarcar la casilla de la opción que haya elegido ya sea la media aritmética o el voto mayoritario o las dos simultáneamente.

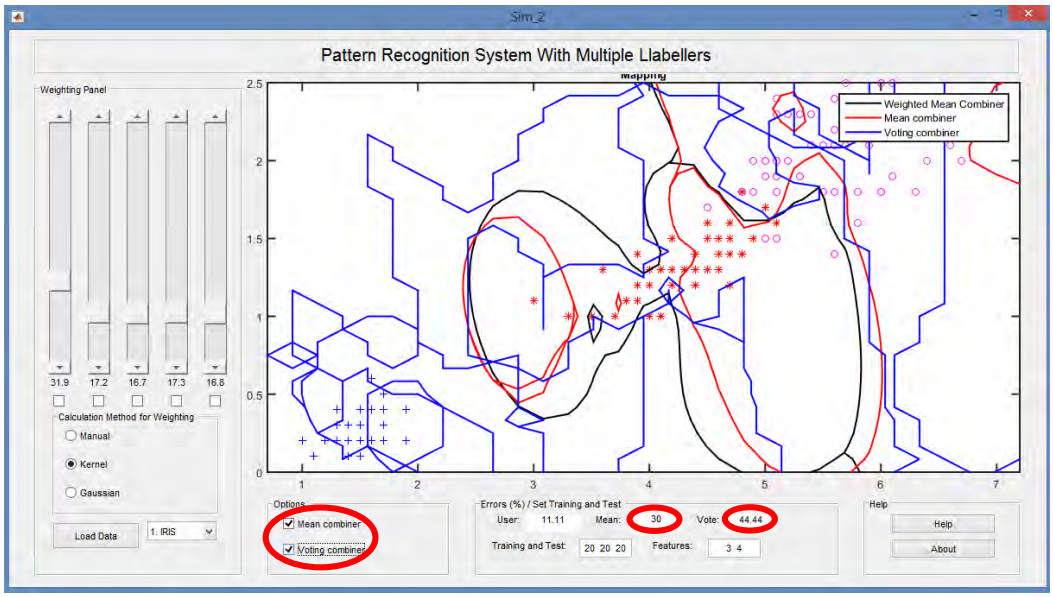


Figura 11. Mappings de las diferentes opciones de mezcla (media, voto y media ponderada)

En la imagen anterior en el primer círculo rojo se puede ver que se encuentran seleccionadas las dos casillas la media aritmética y el voto mayoritario además también se puede ver que esta activada la opción kernel, por lo que se trazan los tres mapeos. En el cuadro de errores se observa que en la casilla **mean** y **vote** están los errores en la clasificación de cada método respectivamente.

ANEXO 4. PAGINA WEB

En el desarrollo de este proyecto se propuso la implementación de una página web, con este fin se creó una página web en el sitio google sites en la cual se encuentra consignada la información relacionada con el simulador, algoritmos, datos personales, manual de usuario, artículos y un video tutorial entre otros. Se puede acceder a la página web por medio del siguiente vínculo:

https://sites.google.com/site/degreethesisdiegopeluffo/multi-labeler_simulator

ANEXO 5. ESTUDIO COMPARATIVO PARA LOS MÉTODOS DE MEZCLA UTILIZANDO TODOS LOS CLASIFICADORES CONVENCIONALES

En este anexo se presenta los resultados obtenidos con todos los metodos de mezcla de clasificacificadores que se utilizaron en el estudio comparativo para las SVM (Maquinas de Soporte Vectorial) aplicándolos a todos los clasificadores convencionales que se utilizaron en este trabajo listados a continuacion: Clasificador discriminante lineal (LDC), Clasificador discriminante cuadratico (QDC), Clasificador Lineal discriminante de Fisher's (FISHER), y La Optimización del Clasificador de Parzen (PARZEN).

El procedimiento que se realizó con los clasificadores consta de realizar el procedimiento de clasificación con el conjunto de prueba perteneciente a cada uno de los grupos de etiquetadores simulados con los diferentes porcentajes de error de esta manera se obtiene el error de estimación de clasificación. De igual modo este proceso se efectúa iterativamente durante 100 ciclos obteniendo como resultado que cada clasificador individual contenido en el grupo adquiera un error de estimación por cada conjunto etiquetado. Luego de finalizar el proceso se conforma la matriz $\boldsymbol{E}_m \in \mathbb{R}^{100x5}$, donde m representa el número de métodos de clasificación que se utilizaron. Finalmente este proceso se reitera en el mismo modo para cada método de clasificación.

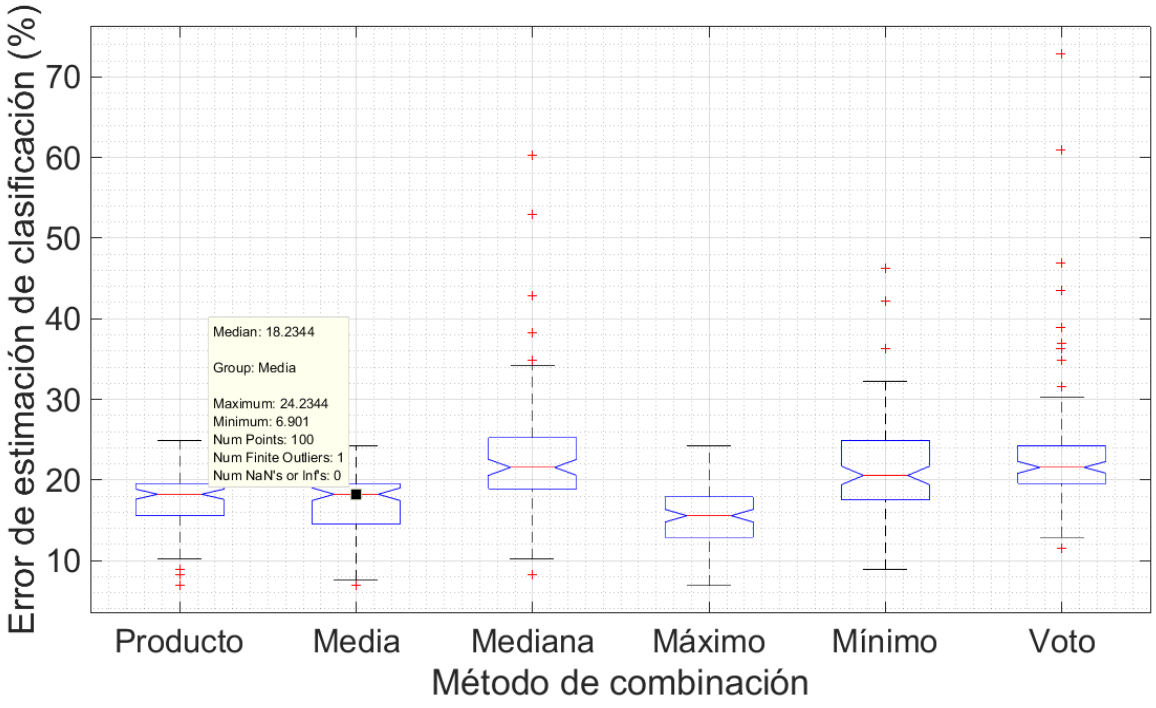


Figura 23. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador `LDC`, donde cada caja representa el error promedio

para cada metodo de combinacion; resultando que la media presenta una tasa menor de error.

La Figura 23 muestra los resultados de todos los métodos de combinación estudiados aplicados al clasificador `LDC`, se puede observar que la media presenta un menor error de estimación en promedio comparado con los demás métodos de combinación.

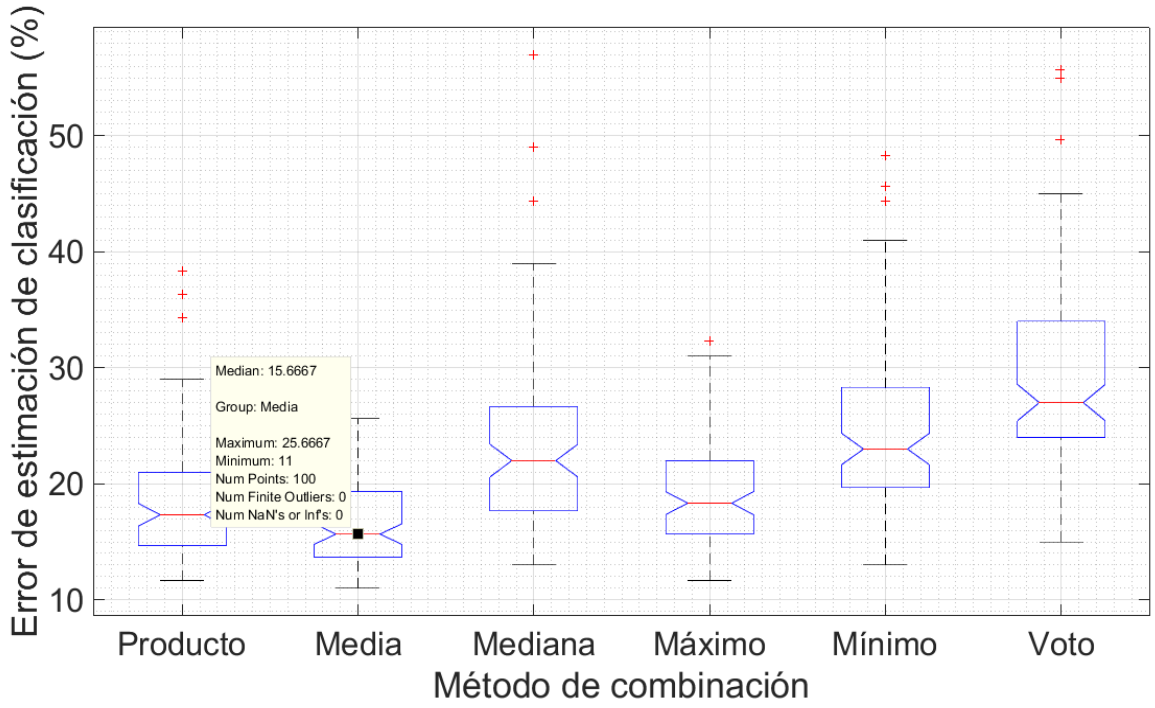


Figura 24. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador `QDC`, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error.

La Figura 24 muestra los resultados de todos los métodos de combinación estudiados aplicados al clasificador `QDC`, se puede observar que la media presenta un menor error de estimación en promedio comparado con los demás métodos de combinación.

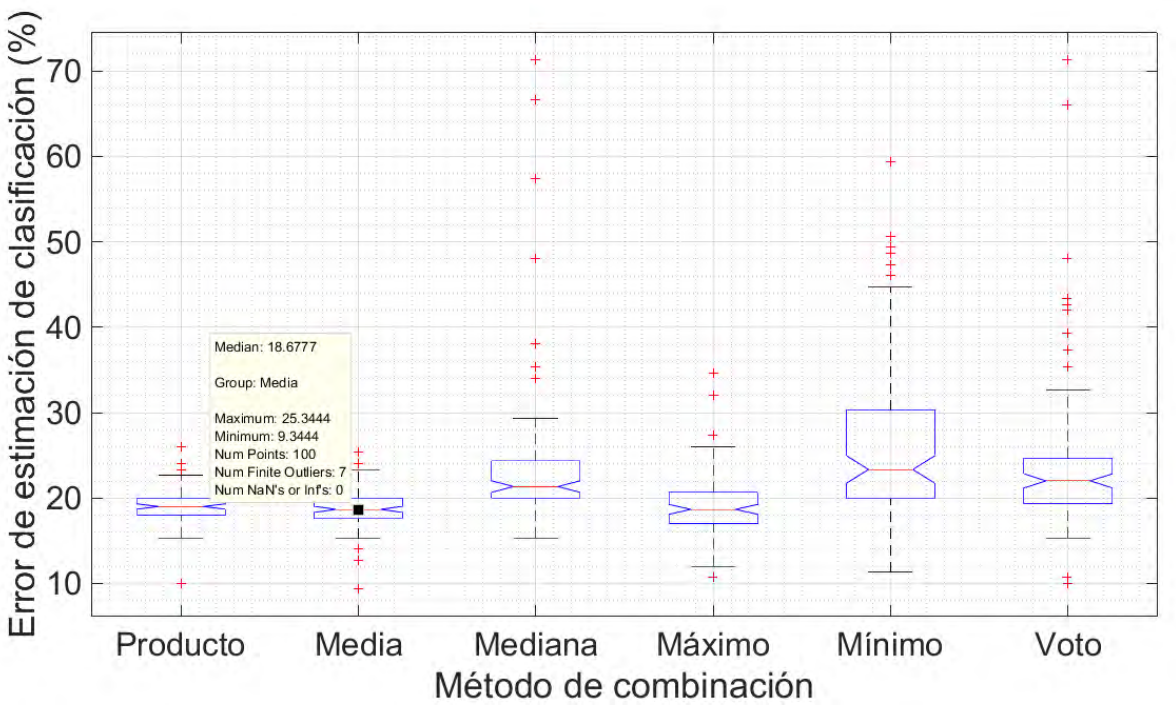


Figura 25. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador `FISHER`, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error.

La Figura 25 muestra los resultados de todos los métodos de combinación estudiados aplicados al clasificador `FISHER`, se puede observar que el máximo presenta un menor error de estimación en promedio comparado con los demás métodos de combinación, por otra parte, la media presenta un error de clasificación mayor pero en una proporción mínima y cabe recalcar que el tamaño de la caja (percentil 25 y 75) es menor.

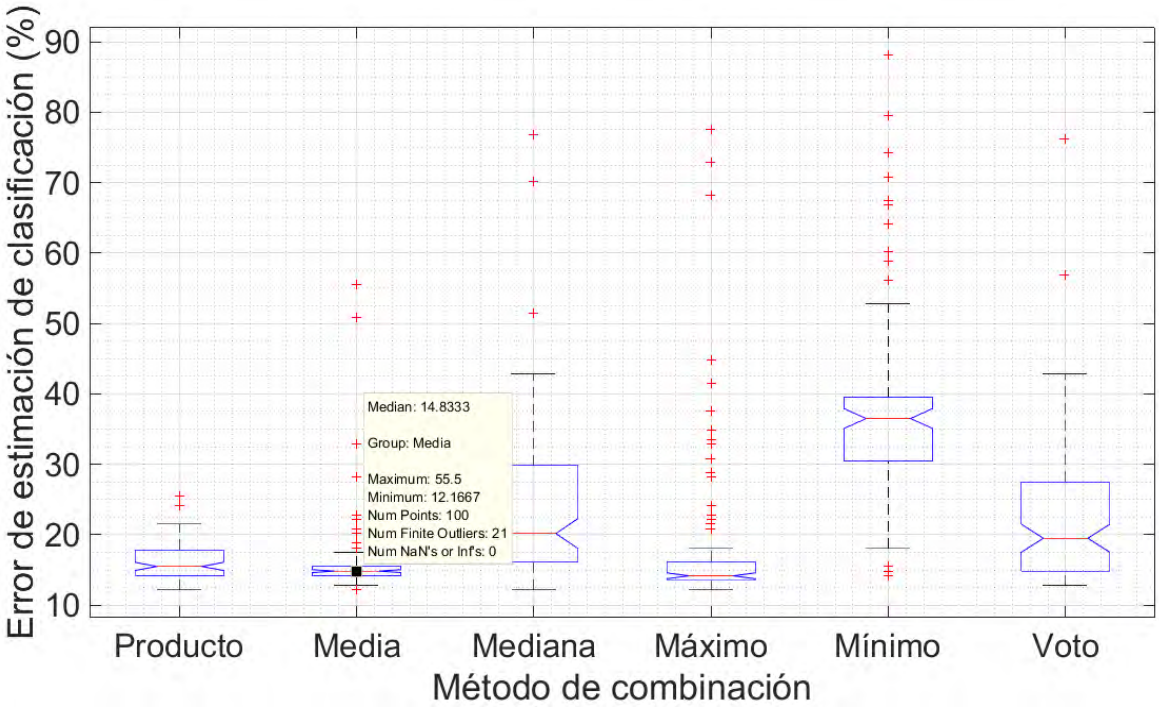


Figura 26. Diagrama de cajas y bigotes para los métodos de combinación ensayados para el clasificador `PARZEN`, donde cada caja representa el error promedio para cada metodo de combinacion; resultando que la media presenta una tasa menor de error.

La Figura 26 muestra los resultados de todos los métodos de combinación estudiados aplicados al clasificador `PARZEN`, se puede observar que el máximo presenta un menor error de estimación en promedio comparado con los demás métodos de combinación pero tiene una mayor proporción de datos atípicos (puntos rojos), por otra parte, la media presenta un error de clasificación mayor pero en una proporción mínima y cabe recalcar que el tamaño de la caja (percentil 25 y 75) es menor.