UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

CLASIFICACIÓN FOTOMÉTRICA DE SUPERNOVAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

IGNACIO CANO DELGADO

PROFESOR GUÍA:
PABLO ESTÉVEZ VALENCIA

MIEMBROS DE LA COMISIÓN:
PABLO HUIJSE HEISE
GIULIANO PIGNATA

SANTIAGO DE CHILE
2016

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO
POR: IGNACIO CANO DELGADO
FECHA: 2016
PROF. GUÍA: SR. PABLO ESTÉVEZ VALENCIA

CLASIFICACIÓN FOTOMÉTRICA DE SUPERNOVAS

Entre los desafíos mas importantes para la cosmología actual se encuentran la expansión y composición del universo. Una de las herramientas mas útiles para la investigación en estos campos son las supernovas de tipo Ia, eventos de gran liberacion energetica que siguen al colapso de una estrella en un sistema binario, esto debido a que las caracteristicas de esta explosión permiten calcular distancias en base a su corrimiento al rojo. El problema es que su identificación y clasificación es un problema no trivial, y el método clasico, el espectroscópico, resulta incapaz de adaptarse al rápido aumento en la información disponible, proveniente de sondeos de última generación. Por lo que resulta de gran importancia encontrar la forma de aprovechar al máximo la información fotométrica, mucho mas abundante que la espectroscópica. El objetivo de esta memoria es diseñar una metodología para la clasificación de supernovas tipo Ia, que entregue resultados competitivos con los presentes en la literatura, esto mediante la aplicación de correntropía mutua ranurada, una medida discreta de disimilitud, sobre la información fotométrica de estas. Se generan matrices de disimilitud para cada uno de los filtros disponibles (*griz*) y se prueban diferentes métodos para la combinación de esta información. Se explora el efecto de añadir la información de corrimiento al rojo fotométrico (*photo-z*) y la forma de introducirla al proceso de clasificación. La clasificación es realizada utilizando tres implementaciones diferentes de algoritmos de vecinos cercanos (*K-nearest neighbours*, *weighted K-nearest neighbours*, y *distance-weighted K-nearest neighbours*). La base de datos utilizada corresponde a la versión corregida de un set de supernovas simuladas creada con motivo del *Supernova Photometric Classification Challenge* (SNPCC), que contiene la información fotométrica de cerca de 21000 supernovas. El entrenamiento se realiza utilizando un conjunto de 1100 objetos cuya clase ha sido espectroscópicamente confirmada, este subconjunto intenta simular las condiciones de captura esperables (e.g. distribución no representativa de clases, preferencia por objetos mas brillantes) y por lo tanto se ha decidido mantenerlo. Tambien se exploran los resultados obtenidos al utilizar una versión de este conjunto modificada para tener una distribución mas representativa, tanto en terminos de clases como de corrimiento al rojo. Se obtiene pureza = 0.556(0.824), eficiencia = 0.567(0.307), y FoM = 0.167(0.187) utilizando el conjunto espectroscópicamente confirmado (en su versión modificada) para el entrenamiento.

Agradecimientos

A toda mi familia que cada dia esta más unida.
A ambos Pablos por soportarme más de lo que tenia derecho a pedirles.
A los que están y los que estuvieron, los vishawty, los canelocos, loscas, lascas y la scorpion.
"Powered@NLHPC: Esta investigación fue parcialmente apoyada por la infraestructura de supercómputo del NLHPC (ECM-02)"

Tabla de contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

Una supernova es un evento cósmico que sigue a la explosión de una estrella masiva (con una masa superior a ocho masas solares), su magnitud es tal que puede llegar a opacar la luminosidad de la galaxia que la aloja ([2, p. 473]). Un tipo en particular de supernovas, denominadas Ia, resulta de gran importancia para la astronomía, esto debido a que sus características de luminosidad permiten su uso como candelas estandarizables, utilizadas para estimar la distancia a la tierra en función de su *redshift*, esto es, el corrimiento al rojo observado en su espectro, indicador de la velocidad a la que el objeto se aleja del punto de observación. La generación de un método para identificar fácilmente las supernovas de esta clase resulta crítica para resolver las grandes interrogantes de la cosmología moderna, específicamente la diferencia entre las mediciones y las predicciones realizadas para la velocidad de expansión del universo, y la naturaleza de la energía oscura.

El método utilizado normalmente para la clasificación de supernovas es el espectroscópico, consistente en el análisis del espectro del objeto en cuestión, pero este proceso es lento y costoso, en términos de tiempo de observación necesario, y resulta incapaz de adaptarse al alto influjo de información proveniente de sondeos de última generación, tales como el DES (*Dark Energy Survey* [3]) y futuros como el LSST (*Large Synoptic Survey Telescope* [4]). Debido a esto, esfuerzos actuales se han centrado en utilizar de manera eficiente la información fotométrica, mucho más fácil de recolectar ([5], [6], [7] [8]). La fotometría radica en el análisis de las curvas de luz de los objetos estelares, estas curvas consisten en series de tiempo de flujo de energia lumínica medida en un número variable de filtros diferentes. Existen varios problemas inherentes al método de recopilación de información; datos faltantes, tiempo de observación limitado, artefactos provenientes de las condiciones atmosféricas, o el ciclo natural día-noche.

La presente memoria de título tiene por objetivo la clasificación de supernovas mediante el análisis de sus curvas de luz, utilizando herramientas propias de la inteligencia computacional. En particular se propone la generación de una matriz de disimilitud mediante la aplicación de una versión discreta de la correntropía cruzada, una medida de la similitud entre dos curvas de luz, y su posterior utilización para la clasificación binaria entre supernovas de tipo Ia y no-Ia empleando algoritmos de vecinos cercanos.

1.1. Estructura de la memoria

La presente memoria consta de cinco capítulos; el primer capítulo, que está conformado por una breve introducción al tema, la estructura del trabajo, y los objetivos generales y específicos; el segundo, en el cual se presentan los conceptos básicos de astronomía necesarios para situarse en el contexto del trabajo, la revisión bibliográfica de aquellos trabajos con una relación directa con el presente, una revisión de los métodos provenientes del área de inteligencia computacional que han sido utilizados para enfrentar este problema en el pasado, y los que se ocupan en este trabajo; el tercero, en el cual se presenta la metodología diseñada para resolver el problema; el cuarto, en el que se presentan los resultados obtenidos; y el último, en el que se entregan las conclusiones del trabajo así como las posibles mejoras que se podrían aplicar al método.

1.2. Objetivos generales

El objetivo general de esta memoria es diseñar una metodología de clasificación fotométrica de supernovas utilizando diferentes técnicas de inteligencia computacional, en particular correntropía cruzada, que entregue resultados comparables a los presentes en la bibliografía al ser aplicado bajo las condiciones de observación e informacion espectroscópica disponible presentadas en el *Supernova Photometric Classification Challenge* ([5]).

1.3. Objetivos específicos

- Comparar la correntropía cruzada como medida de similitud entre curvas de luz con una técnica similar de correlación cruzada.
- Implementar una aplicación discreta del método con el objetivo de evitar interpolaciones globales.
- Comparar el desempeño de diferentes técnicas para la combinación de la información multicanal.
- Medir el impacto de la adición de información de corrimiento al rojo de galaxia huésped (*photo-z*).
- Comparar los resultados obtenidos con aquellos presentados en la bibliografía.

Capítulo 2

Antecedentes

En el presente capítulo se describen los antecedentes necesarios para la contextualización de este trabajo de título. En primer lugar se presentan resumidamente los conceptos básicos de astronomía necesarios para situarse en el contexto de este trabajo. Luego se examinan los trabajos más relevantes en cuanto al estado del arte. Finalmente se revisan las técnicas provenientes del área de inteligencia computacional usadas en los trabajos presentados anteriormente, así como las utilizadas en esta memoria de título.

2.1. Conceptos básicos de astronomía

2.1.1. Supernovas

Una supernova corresponde a un evento cósmico que sigue a la explosión de una estrella masiva (con una masa superior a ocho masas solares), en el cual una gran cantidad de energía es liberada en un corto periodo de tiempo, alcanzando su peak de energía liberada cerca de los 30 días, decayendo casi completamente entre los 100 días, en el caso de las supernovas tipo II, y los 300 días, en el caso de las supernovas tipo Ia. La magnitud es tal, que la luz liberada puede llegar a opacar la luminosidad de la galaxia anfitriona, sin embargo esta luz representa menos de un uno por ciento de la energía total liberada ([2, p. 473]).

Las causas que detonan el evento pueden ser varias. La clasificación clásica las divide en dos tipos generales, en base al tipo de líneas observadas en su espectro([9]). En la figura 2.1 se muestra el criterio de clasificación en cuanto a los elementos observados en su espectro.

Tipo I

Los espectros de las supernovas de esta clase no presentan líneas de hidrógeno, se subclasifican en tres categorías; Ia, Ib, e Ic.

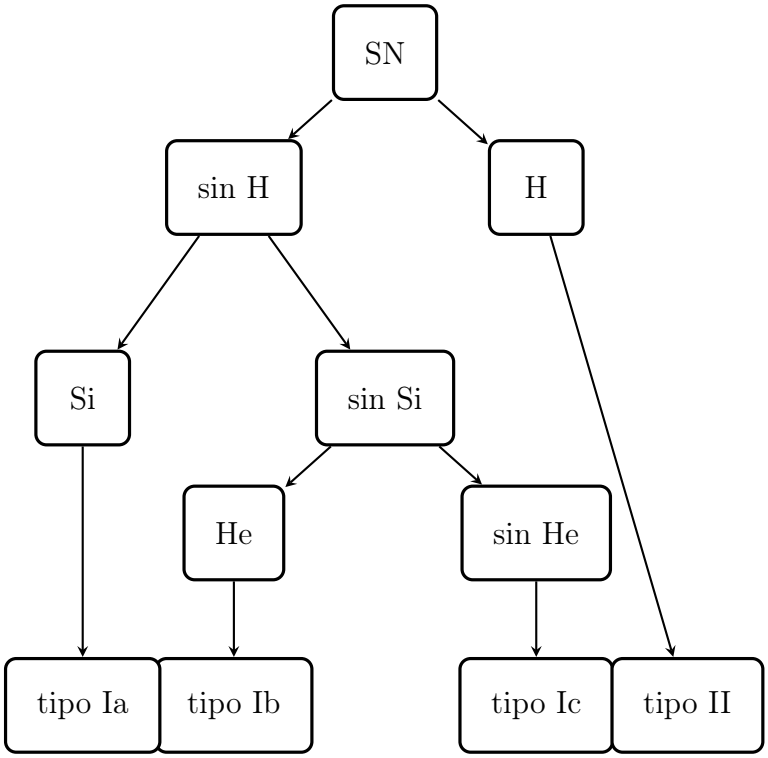

Figura 2.1: Clasificación espectroscópica de supernovas

Los espectros de las supernovas de tipo Ia presentan líneas de silicio y no de helio. Son formadas producto del rápido traspaso de materia entre una enana blanca y otra estrella, comunmente una gigante roja, en un sistema estelar binario. Mientras la enana blanca aumenta su masa, disminuye su tamaño y aumenta su temperatura. Cuando la temperatura en su centro alcanza los 10 billones K la estrella comienza a quemar su nucleo de carbón, provocando una reacción termonuclear que finaliza con la explosión de la enana. Se muestra una curva de luz típica de una supernova Ia en la figura 2.2 ([2, p. 513]).

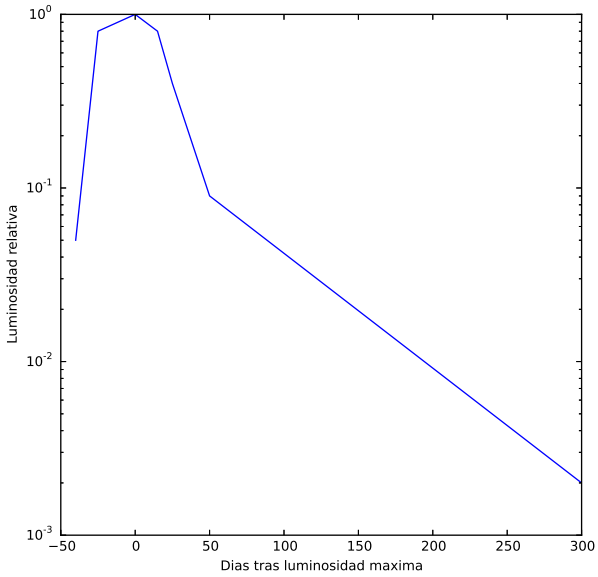

Figura 2.2: Curva típica de supernova tipo Ia

Las supernovas de tipo Ia son particularmente importantes para la cosmología debido a su uso como candelas estandarizables, utilizadas para estimar la distancia a la tierra como función de su corrimiento al rojo ([2, p. 629]). Debido a esto resultan herramientas útiles para estudiar la expansión del universo así como para encontrar límites para ciertas variables cosmológicas. Las supernovas de tipo Ib e Ic no son utilizadas directamente en cosmología pero son una fuente potencial de contaminación al generar catálogos de supernovas Ia y poseen cierto interés en el área de la astrofísica, ambas carecen de líneas de silicio en su espectro, y en las de tipo Ib se pueden observar líneas de helio ([9]).

Tipo II

Las supernovas de este tipo corresponden a explosiones iniciadas en el centro de estrellas masivas que han consumido todo su combustible nuclear. Sus espectros presentan líneas de hidrógeno. Se presenta una curva de luz típica para esta clase en la figura 2.3.

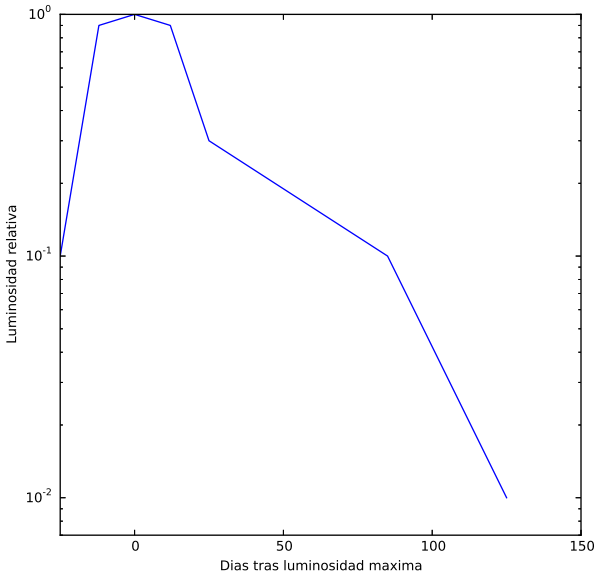

Figura 2.3: Curva típica tipo II

De manera similar a las supernovas de tipo Ia, las supernovas de tipo II pueden ser utilizadas como estimadores de distancia, pero con menor precisión y para distancias menores. En el terreno de la astrofísica, son mejor entendidas que las supernovas de tipo I. Presentan menores distorsiones ambientales debido a que solo se encuentran en galaxias tardías. Pueden ser utilizadas para complementar los análisis de supernovas Ia([2, p. 474]).

2.1.2. Expansión del universo

El modelo cosmológico actual posee diferentes parámetros que indicarían la forma y expansión del universo dependiendo de los elementos que lo conforman. Diferentes estudios([10])([11]) indican una expansión cosmológica acelerada, a partir de resultados basados en muestras de supernovas de tipo Ia con alto corrimiento al rojo ([12])([13])([14]).

Las recientes estimaciones de los parámetros se han apoyado principalmente en supernovas Ia clasificadas espectroscópicamente. Cerca de un noventa por ciento de las supernovas descubiertas no poseen un espectro indexado, por lo que los métodos de clasificación fotométrica son cada vez mas necesarios.

Los resultados se interpretan como evidencia de la existencia de un componente exótico denominado energía oscura. Esta representa el mayor desafío actual para la cosmología y la física teórica. Múltiples estudios se han realizado para mejorar nuestro entendimiento acerca de su naturaleza, así como para encontrar mejores limites para sus características.

2.1.3. Corrimiento al rojo (*redshift*) y distancias en astronomia

Al capturar la luz proveniente de un objeto astronómico lejano se puede observar una variación entre la longitud de onda emitida y la recibida. Esta distorsión, que se debe al efecto Doppler, depende de si el objeto que emite la luz se acerca o aleja del punto de observación.

Si el objeto se aleja, la longitud de onda recibida es mayor, a esto se le llama *redshift* o corrimiento al rojo, debido a que la luz visible con mayor longitud de onda es más roja. Esto sumado a ciertas características (e.g. líneas espectrales) esperadas o detectadas nos puede entregar una indicación de la velocidad a la que se está alejando de nosotros. En el caso de las supernovas, es posible calcular la distancia a la tierra como función del corrimiento al rojo.

Un caso particular de este fenómeno que es referenciado en este trabajo es el llamado *photo-z*, esto es el corrimiento al rojo correspondiente a la galaxia huésped de una supernova, calculado utilizando métodos puramente fotométricos. Uno de los objetivos de este trabajo es estudiar el impacto de añadir esta información, sea como parámetro de ajuste o como característica adicional durante el proceso de aprendizaje, a los algoritmos de clasificación utilizados.

2.1.4. Espectroscopía

La espectroscopía consiste en la obtención y estudio del espectro de objetos astronómicos. La luz proveniente de un objeto es capturada y posteriormente dispersada, comúnmente mediante el uso de un prisma, de acuerdo a la longitud de onda. Luego su intensidad es medida en regiones estrechas del espectro con el uso de un detector. Esto permite observar la forma, posición, e intensidad de las líneas de emisión y absorción de diferentes elementos y moléculas. El análisis espectroscópico es utilizado para descubrir propiedades de estos objetos astronómicos, tales como composición química, movimiento, y campos magnéticos([2, p. 114]).

La información espectroscópica es más escasa, lenta, y costosa (en términos de tiempos de observación necesarios) de generar que la información fotométrica, por lo que resulta importante encontrar métodos que utilicen eficientemente esta última.

2.1.5. Fotometría

La fotometría, en el contexto de la astronomía, consiste en el monitoréo de las variaciones de intensidad en regiones específicas del espectro lumínico, que incluye la radiación visible tanto como la no visible, delimitadas mediante el uso de filtros([15, pp. 1-6]).

Curvas de luz

Los datos observados corresponden a mediciones de flujo de energía lumínica en puntos de tiempo irregularmente espaciados en un número de filtros diferentes, generalmente entre dos y diez, dependiendo del sistema utilizado, además de sus medidas de error asociadas. De interés particular para este trabajo es el sistema de filtros *griz*, que consta de cuatro filtros con un rango de captura situado entre el verde y el infrarojo cercano, esto es, entre 500[nm] y 900[nm] de longitud de onda aproximadamente ([16]).

Podemos definir los datos para el objeto i como $X_{\mathrm{i}} = \{t^b_{\mathrm{i}k}, f^b_{\mathrm{i}k}, \sigma^b_{\mathrm{i}k}\}$

donde $k = 1, ..., p^b_{\mathrm{i}}$
b indexa el filtro al que corresponde la medición
p^b_{i} corresponde al número de mediciones de flujo en el filtro b para el objeto i
$t^b_{\mathrm{i}k}$ es la grilla de tiempo
$f^b_{\mathrm{i}k}$ la medición de flujo
$\sigma^b_{\mathrm{i}k}$ es el error en la medición

El estudio de las curvas de luz tiene diferentes problemas inherentes a las herramientas utilizadas para la captura de información(e.g. ciclo día-noche, condiciones atmosféricas). Las mediciones son comúnmente ruidosas, no uniformemente muestreadas en el tiempo, e incompletas. Además resulta difícil distinguir entre supernovas de tipo I con diferentes subclases.

2.1.6. Sondeos astronómicos

Una de las razones por las que este trabajo y otros similares resultan de interés es el gran aumento en la información disponible, así como los recursos limitados para analizar esta información. Los siguientes corresponden a sondeos, recientes o futuros, que se beneficiarían de nuevas técnicas de análisis fotométrico para su estudio:

- *Supernova Legacy Survey* (SNLS)([17]).
- *Sloan Digital Sky Survey II* (SDSS-II)([18]).
- *Dark Energy Survey* (DES)([3]).
- *Panoramic Survey Telescope and Rapid Response System* (Pan-STARRS)([19]).
- *Large Synoptic Survey Telescope* (LSST)([4]).

2.2. Estado del arte

2.2.1. SNPhotCC (Concurso 2010)

El año 2010 Richard Kessler et al. publicaron una muestra ciega de supernovas simuladas, de características similares a las esperadas en un sondeo real, específicamente a las condiciones de observación del DES([5]). El concurso, dirigido a científicos, consistía en la clasificación de las muestras en los tipos de supernova correspondientes y se dividía en cuatro desafíos independientes: con/sin photo-z; y curva completa/seis primeras mediciones. Los objetivos de la competencia consistieron en:

- Encontrar las fortalezas y debilidades relativas de diferentes técnicas de clasificación
- Utilizar los resultados para mejorar los algoritmos de clasificación
- Entender que conjuntos de muestras espectroscópicamente conformadas son necesarios para el correcto entrenamiento de estos algoritmos.

Las simulaciones fueron realizadas en los filtros *griz* del DES y basadas en modelos determinados empíricamente a partir de datos. Tomando en cuenta variaciones esperables correspondientes a factores tales como el ciclo día-noche y la transparencia atmosférica, así como datos incompletos.

Además se proporcionó un subconjunto de la muestra previamente clasificado espectroscópicamente, para que los algoritmos pudieran ajustarse a un set de entrenamiento realista. La muestra fue enriquecida mediante la contribución de curvas no-Ia espectroscópicamente confirmadas por parte de CSP, SNLS, y SDSS-II. En total se dispuso de 1256 muestras confirmadas, correspondiendo a un siete por ciento de los datos.

Dos criterios de selección fueron aplicados a las muestras. En primer lugar, cada objeto debía poseer al menos una observación en cualquier filtro con razón señal a ruído sobre 5. Segundo, debían existir al menos 5 observaciones posteriores a la explosión. La aplicación de estos criterios resulta en $1{,}8E4$ muestras disponibles.

Se utilizó una figura de mérito(FoM) para la evaluación de los resultados. Para el caso de clasificación binaria se propuso el uso de eficiencia y pureza. Las fórmulas para estas medidas son las presentadas en las ecuaciones siguientes.

Eficiencia (también llamada sensibilidad, razón de verdaderos positivos, o recall):

$$\mathrm{e}ff_{Ia} = \frac{N_{Ia}^{true}}{N_{Ia}^{total}} \tag{2.1}$$

Pureza (también llamada precisión o valor predictivo positivo):

$$pur_{Ia} = \frac{N_{Ia}^{true}}{N_{Ia}^{true} + N_{Ia}^{false}} \tag{2.2}$$

FoM (similar a F-score con la inclusión de un factor de penalización):

$$FoM_{Ia} = \frac{1}{N_{Ia}^{total}} \frac{(N_{Ia}^{true})^2}{N_{Ia}^{true} + W N_{Ia}^{false}} \tag{2.3}$$

donde N_{Ia}^{true} corresponde al número de supernovas correctamente clasificadas como Ia.
N_{Ia}^{false} al número de supernovas incorrectamente clasificadas como Ia.
N_{Ia}^{total} al número de supernovas de tipo Ia presentes en la muestra
y W a un factor de penalización que controla el costo relativo de los falsos positivos sobre los falsos negativos, con un valor de 3 para efectos de la competición.

2.2.2. Resultados del desafío (2010)

Tres errores principales fueron detectados y solucionados durante el plazo de la competición, resultando en una reducción de hasta un 10 por ciento en el tamaño de la muestra([1]). Un número de errores fueron detectados posteriormente a la entrega.

En cuanto a los resultados del concurso, se pueden identificar cuatro estrategias generales:

- Ajustar cada curva a un modelo de supernova Ia, cortando según criterio de mínimo χ cuadrado.
- Utilizar métodos probabilísticos bayesianos para determinar el tipo de supernova mas probable, contrastando modelos tanto de Ia como de no-Ia. El número de modelos Ia/no-Ia utilizados varia según el participante.
- Parametrizar un diagrama de Hubble mediante el uso de la supernovas Ia espectroscópicamente confirmadas, luego clasificar como Ia aquellas que yacen cerca del diagrama esperado. El diagrama fue generado mediante un polinomio de alto orden en un caso, y mediante estimación de densidad de kernel en otro.
- Ajustar cada curva de luz a una función paramétrica (e.g. spline), posteriormente utilizar los parámetros ajustados para realizar inferencias estadísticas.

Diferentes estrategias de clasificación entregaron resultados similares, para todos el rendimiento es altamente superior en el conjunto espectroscópicamente confirmado que en el completo, indicando un sobreajuste a estos datos. En la figura 2.4 se entrega un sumario de los equipos participantes en la competencia, y un resumen de su estrategia para enfrentarla.

El mejor resultado presenta una eficiencia de 0.96 y una pureza de 0.79 para la clase Ia en el conjunto espectroscópicamente confirmado. Sin embargo tres de las cuatro estrategias presentan resultados comparables y ninguna fue notablemente superior.

Dos problemas principales fueron identificados en las implementaciones. En primer lugar, el subconjunto de muestras espectroscópicamente confirmadas fue generalmente tratado como un subconjunto aleatorio, cuando en realidad presenta una sobrerepresentación de supernovas de tipo Ia. Segundo, muchas de las entradas no utilizaron toda la información disponible o añadieron ruido a la simulación.

Participants	Abbreviation[a]	Classified +Z[b]/noZ[c]	SN z_{ph}[d]	CPU[e]	Description (strategy class[f])
P. Belov and S. Glazov	Belov & Glazov	yes/no	no	90	light curve χ^2 test against Nugent templates (2)
S. Gonzalez	Gonzalez	yes/yes	no	120	cuts on SiFTO fit χ^2 and fit parameters (1)
J. Richards, Homrighausen, C. Schafer, P. Freeman	InCA[g]	no/yes	no	1	Spline fit & nonlinear dimensionality reduction (4)
J. Newling, M. Varuguese, B. Bassett, R. Hlozek, D. Parkinson, M. Smith, H. Campbell, M. Hilton, H. Lampeitl, M. Kunz, P. Patel (JEDI group[h])	JEDI-KDE JEDI Boost JEDI-Hubble JEDI Combo	yes/yes yes/yes yes/no yes/no	no no no no	10 10 10 10	Kernel Density Evaluation with 21 params (4) Boosted decision trees (4) Hubble diagram KDE (3) Boosted decision trees + Hubble KDE (3+4)
S. Philip, V. Bhatnagar, A. Singhal, A. Rai, A. Mahabal, K. Indulekha	MGU+DU-1[1] MGU+DU-2	no/yes no/yes	no no	< 1 < 1	light curve slopes & Neural Network (2) light curve slopes & Random Forests (2)
H. Campbell, B. Nichol, H. Lampietl, M .Smith	Portsmouth χ^2 Portsmouth-Hubble	yes/no yes/no	no no	1 1	SALT2-χ^2_r & False Discovery Rate Statistic (1) Deviation from parametrized Hubble diagram (3)
D. Poznanski	Poz2007 RAW Poz2007 OPT	yes/no yes/no	yes yes	2 2	SN Automated Bayesian Classifier (SN–ABC) (2) SN–ABC with cuts to optimize $\mathcal{C}_{\mathrm{FoM-Ia}}$ (2).
S. Rodney	Rodney	yes/yes	yes	230	SN Ontology with Fuzzy Templates (2)
M. Sako	Sako	yes/yes	yes	120	χ^2 test against grid of Ia/II/Ibc templates (2)
S. Kuhlmann, R. Kessler	SNANA cuts	yes/yes	yes	2	Cut on MLCS fit probability, S/N & sampling (1)

Figura 2.4: Estrategias presentadas para la competición. Tomado de [1]

Los resultados presentados por los participantes en el desafío se presentan en las figuras 2.5 y 2.6 en términos de FoM contra corrimiento al rojo, las líneas sólidas corresponden a los resultados obtenidos al aplicar los métodos sobre el conjunto espectroscópicamente confirmado, mientras que las punteadas a los resultados sobre el conjunto completo. La primera figura corresponde al desafío que no incluía la información de redshift fotométrico, la segunda al que si lo incluía. Se puede observar que, salvo en un par de casos, los buenos resultados en el conjunto espectroscópicamente confirmado no se traducen en el conjunto completo. Asimismo se puede notar que la mayoría de los métodos entregan resultados pobres para bajo corrimiento al rojo, esto se puede explicar debido a que el conjunto espectroscópicamente confirmado posee pocos ejemplos de supernovas tipo Ia con bajo corrimiento al rojo.

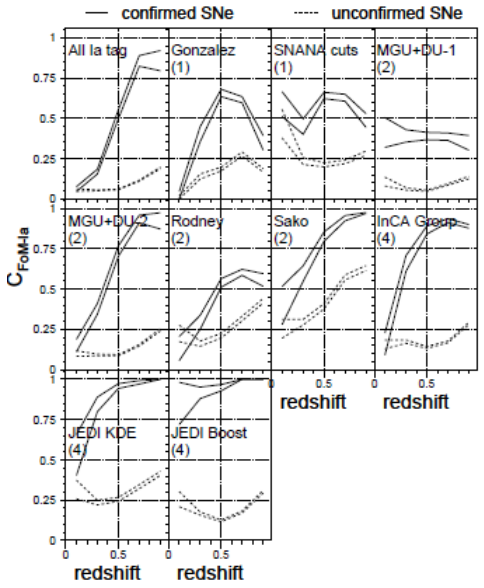

Figura 2.5: Resultados sin photo-z. Tomado de [1]

Por último, se liberó una muestra actualizada con algunas mejoras, los errores conocidos solucionados, y curvas de luz adicionales correspondientes a sondeos de LSST ([4]) y SDSS-II

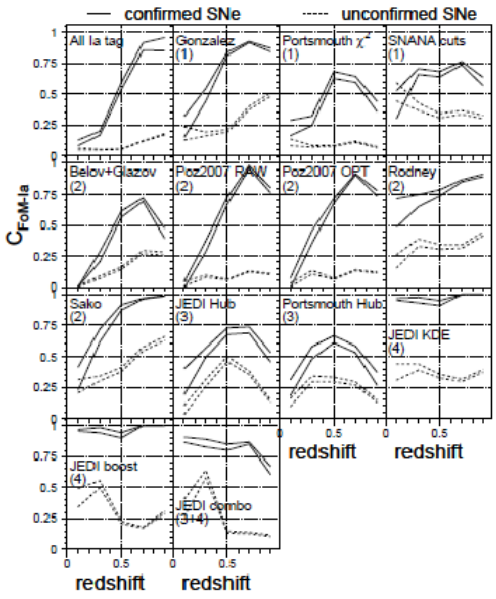

Figura 2.6: Resultados incluyendo photo-z. Tomado de [1]

(([18]). Esta muestra, comúnmente llamada SNPCC, será utilizada en estudios posteriores así como en este trabajo. Los resultados no son directamente comparables a los obtenidos en el concurso debido a las diferencias entre las bases de datos (e.g. la relación entre supernovas espectroscópicamente confirmadas y no confirmadas era de 1:13 en el desafío, y 1:18 en la muestra actualizada), por lo que se utilizarán como referencia los trabajos que utilizan la base de datos actualizada.

2.2.3. Publicaciones posteriores: Richards et al. (2012)

En este trabajo se utilizaron todas las curvas de luz para estimar una representación de baja dimensionalidad de cada supernova. Se usaron las muestras espectroscópicamente confirmadas para crear un modelo de clasificación([6]).

Los datos iniciales se encuentran en una grilla irregular de tiempo que difiere de supernova a supernova, y de filtro a filtro. Se utilizó un spline natural de regresión cúbica para transportar los datos a un grilla de tiempo uniforme. Un análisis de sensibilidad reveló que la densidad de la grilla no afecta notablemente los resultados.

Para alinear los datos en el eje temporal se definió el tiempo de punto cero de una curva de luz de supernova como el momento, en fecha Juliana, en el que se observa el flujo máximo en la banda r. Esto puede ocurrir al principio o al final del periodo de observación. Dependiendo del caso se utilizó correlación cruzada, o el descarte para las ocasiones en que el peak de una curva se encuentra al principio y el de la otra al final del periodo de observacion. Para aquellas curvas sin peaks en la banda r se utilizó un estimador mediante correlación cruzada con cada supernova que si posea un peak en la banda r.

Se utilizaron mapas de difusión para reducir la dimensionalidad. La distancia usada para

generar el mapa de difusión es llamada distancia cuadrática de banda x (donde x puede referirse a cualquiera de las bandas de *griz*), definida en la ecuación siguiente:

$$S_b(X_\mathrm{i}, X_j) = \frac{1}{t_u - t_\mathrm{i}} \sqrt{\sum_{k:t_u \leq k \leq t_\mathrm{i}} \frac{(\tilde{f}^b_{\mathrm{i}k} - \tilde{f}^b_{jk})^2}{(\tilde{\sigma}^b_{\mathrm{i}k})^2 + (\tilde{\sigma}^b_{jk})^2}} \tag{2.4}$$

donde k indexa la grilla de tiempo, usando bins de 1 día
$S_b(X_\mathrm{i}, X_j)$ corresponde a la distancia euclidiana ponderada entre curvas de luz, por bin de tiempo en que ocurre traslape
t_u y t_i corresponden a los límites inferior y superior de la superposición
$\tilde{f}^b_{\mathrm{i}k}$ y $\tilde{\sigma}^b_{\mathrm{i}k}$ corresponden, respectivamente, a las mediciones de flujo y error después del preprocesamiento, según la notación presentada en 2.1.5.

La distancia total entre dos curvas de luz resulta de la suma de esta distancia a lo largo de todas las bandas.

Finalmente se generaron *random forests* (ver 2.3.6), entrenándolos con la representación del mapa de difusión y las supernovas espectroscópicamente confirmadas.

Los criterios del desafío presentado en 2.2.1 para evaluar rendimiento fueron utilizados para ajustar los parámetros libres. Maximizando un estimador de FoM en una validación cruzada de 10 folds. Esto arrojó que los resultados son insensibles ante el valor de m (dimensiones del mapa de difusión) cuando este es suficientemente alto.

Se probaron dos métodos para incluir la información de photo-z: añadiéndolo a la medida de distancia; o entregándolo como característica a los árboles aleatorios. Se encontró que el primer método genera mejores resultados debido a que no se necesita asumir que la distribución de las SNs de entrenamiento es similar a la del conjunto de prueba.

Un análisis de los recursos computacionales utilizados señaló que el punto más costoso y variable es la generación de la matriz de correlación para estimar el punto cero de una curva sin peak.

Los resultados indicaron que el desempeño es altamente sensible ante el conjunto de entrenamiento utilizado. Obteniéndose los mejores resultados al utilizar conjuntos de entrenamiento construidos aplicando límites de magnitud, esto es, mediante la selección aleatoria de supernovas con brillo superior al límite predefinido.

Se reportó una pureza de 0.72/0.76 y una eficiencia de 0.65/0.74 (el segundo valor corresponde al resultado tras incluir la información de photo-z. Al entrenar utilizando el set espectroscópicamente confirmado se reportaron pureza y eficiencia de 0.5, y una FoM de 0.14.

2.2.4. Publicaciones posteriores: Karpenka et al. (2012)

El objetivo de este trabajo era generar una metodología simple y robusta para la clasificación de supernovas tipo Ia mediante la utilización de redes neuronales([7]).

Se probaron diferentes tamaños a utilizar para el conjunto de entrenamiento y no se realizaron cortes de selección de ningún tipo (a diferencia de otros trabajos referenciados en los cuales se seleccionaron datos por magnitud, corrimiento al rojo, o razón señal a ruido).

Los datos fueron ajustados a la función analítica parametrizada presentada en la ecuación siguiente:

$$f(t) = A[1 + B(t - t_1)^2]\frac{\mathrm{e}^{-(t-t_0/T_{fall})}}{1 + \mathrm{e}^{-(t-t_0/T_{\mathrm{rise}})}}. \tag{2.5}$$

Para cada supernova, t_0 corresponde a la primera medición en la banda r. La función no tiene una motivación física en particular pero es lo suficientemente general para adaptarse a virtualmente cualquier tipo de curva de luz de supernova, incluyendo aquellas con doble peak. Adicionalmente se realizaron ajustes para cada filtro de cada supernova, mediante una función de verosimilitud Gaussiana clásica como la mostrada en la ecuación siguiente:

$$L(\Theta) = \mathrm{e}xp[\frac{-1}{2}\chi^2(\Theta)], \tag{2.6}$$

donde $\Theta = \{A, B, t_1, t_0, T_{\mathrm{rise}}, T_{fall}\}$ corresponde al vector de los parámetros y

$$\chi^2(\Theta) = \sum_{k=1}^{n} \frac{[F_k - f(t_k; \Theta)]^2}{\sigma_k^2}, \tag{2.7}$$

donde n es el número de mediciones de flujo para la combinación supernova/filtro bajo consideración. El modelo es ajustado minimizando el error χ cuadrado.

Para entrenar la red neuronal se usó un vector de características consistente en:

- Los parámetros de la función ajustada.
- Sus incertidumbres.
- El número de mediciones del flujo.
- El valor de máxima verosimilitud del ajuste.
- Evidencia Bayesiana para el modelo.

Lo anterior resultó en un vector de características de 60 componentes (15 por cada filtro).

La red neuronal utilizada es una red perceptrón de dos capas, con 500 nodos en la capa oculta. Los tiempos de entrenamiento variaron entre 2 y 52 minutos dependiendo del conjunto utilizado. Se ocupó un método de optimización de segundo orden basado en el algoritmo del

gradiente conjugado. La salida de la red indica la probabilidad de que la supernova sea Ia. La adición de photo-z e incerteza aumentaron los resultados en 0.05-0.1 para las muestras seleccionadas aleatoriamente.

Para el caso de entrenamiento utilizando el conjunto espectroscópicamente confirmado se reportó pureza = 0.32, eficiencia = 0.94, y FoM = 0.12.

2.2.5. Publicaciones posteriores: Ishida y de Souza (2012)

El objetivo de este trabajo era diseñar un método que optimizara la pureza de la clasificación.([8]).

El eje temporal fue trasladado desde fecha Juliana a tiempo desde el brillo máximo en un filtro r, se probó que el filtro en particular a utilizar no influencia en demasía el resultado final. Se realizó una traslación hacia una grilla uniformemente espaciada en el tiempo, mediante un spline de regresión cúbico. Se normalizaron los valores de acuerdo al flujo máximo medido en todos los filtros para cada supernova en particular. Solo se consideraron aquellas supernovas que poseían suficientes observaciones tanto antes como después del punto cero en todas las bandas. No se realizó extrapolación en tiempo ni longitud de onda.

Kernel principal component analysis (KPCA, ver sección 2.3.3) fue aplicado para reducir la dimensionalidad utilizando solo las muestras espectroscópicamente confirmadas, el kernel ocupado corresponde al Gaussiano con parámetro σ libre. Se decidió utilizar solo dos coordenadas resultantes de este proceso. Para definir tanto las coordenadas a utilizar como el parámetro libre se realizó un proceso de validación cruzada LOO(*Leave One Out*) con las 5 componentes principales más altas y $\sigma \in \{0{,}1, 2\}$ con intervalos de 0.1.

Las muestras no confirmadas fueron proyectadas a este espacio para luego ser clasificadas utilizando *1-nearest neighbour* (ver 2.4.4) .

Los resultados son sensibles a la calidad de las muestras, representada por su razón señal a ruido (SNR, *Signal-to-Noise Ratio*), pero no tanto a la representatividad del conjunto de entrenamiento. Obteniéndose resultados vastamente superiores al seleccionar solo las muestras con SNR>5.

Para el caso de entrenamiento utilizando el conjunto espectroscópicamente confirmado se reportó pureza = 0.63, eficiencia = 0.71, y FoM = 0.26.

2.3. Técnicas utilizadas en la bibliografía

Una revisión más completa de métodos y técnicas utilizados en astronomía puede encontrarse en [20].

2.3.1. Splines cúbicos

Un spline es definido como una curva diferenciable definida en segmentos mediante polinomios. El término spline hace referencia a una amplia clase de funciones utilizadas en interpolación de datos y suavizado de curvas, aunque comúnmente se refiere a la versión unidimensional y polinomial de estas funciones. Localmente poseen una forma bastante simple pero son globalmente suaves y flexibles ([21]).

Los splines resultan particularmente útiles para estimar funciones suaves y continuas que poseen comportamiento complicado e incrementos rápidos. En los trabajos referenciados previamente es común el uso de splines cúbicos, esto es, conformados por polinomiales de tercer orden que pasan por un set de puntos de control.

2.3.2. Mapas de difusión

Los mapas de difusión consisten en un método de reducción no lineal de dimensionalidad. Se mantiene la "conectividad" de los datos en el contexto de un proceso ficticio de difusión. Se puede describir como la probabilidad de ir de una muestra a otra en un número dado de pasos.

El mapa de difusión se inicializa creando un grafo no dirigido con pesos en los datos fotométricos observados, donde cada dato X_{i} es un nodo en el grafo y los pesos entre pares de nodos son definidos por la siguiente ecuación:

$$w(X_{\mathrm{i}}, X_j) = exp(\frac{-s(X_{\mathrm{i}}, X_j)}{\varepsilon}, \tag{2.8}$$

donde ε es un parámetro ajustable y $s(\bullet, \bullet)$ es una medida de distancia (o similitud) definida por el usuario entre dos objetos (e.g. euclidiana). En esta construcción, la probabilidad de dar un paso de X_{i} a X_j en un proceso de difusión queda definida por la ecuación:

$$p_t(X_{\mathrm{i}}, X_j) = w(X_{\mathrm{i}}, X_j) / \sum_k w(X_{\mathrm{i}}, X_j). \tag{2.9}$$

La matriz P, de tamaño NxN, conserva los valores de probabilidades de paso entre los N datos. Por teoría de cadenas de Markov, para cada entero positivo t, el elemento $p_t(X_{\mathrm{i}}, X_j)$ de la matriz P^t entrega la probabilidad de ir desde X_{i} a X_t en t pasos.

El mapa de difusión a escala t queda definido por la fórmula siguiente:

$$\Psi^t : X \to [\lambda_1^t \Psi_1(X), \ldots, \lambda_m^t \Psi_m(X)], \tag{2.10}$$

donde Ψ_j y λ_j corresponden a los vectores y valores propios de P, respectivamente, en una descomposición espectral biortogonal y m es el número de coordenadas de mapa de difusión

seleccionadas para representar los datos. Las coordenadas se ordenan según los valores propios ($\lambda_1 > \lambda_2$), para que, análogamente a lo que ocurre en análisis de componentes principales, las m coordenadas superiores retengan la mayor cantidad de información posible acerca de P^t. Normalmente $m << N$ coordenadas son necesarias para capturar la mayor parte de la variabilidad del sistema.

La distancia Euclidiana entre dos puntos cualesquiera del espacio m-dimensional descrito por la ecuación 2.10 aproxima la distancia de difusión, una media de distancia que captura la geometría intrínseca de los datos al considerar simultáneamente todos los caminos posibles entre dos puntos en la caminata aleatoria de t-pasos de Markov construida.

Debido a que se promedia entre todos los caminos posibles, la distancia de difusión es robusta ante ruido aleatorio. Los parámetros ε, m y t se pueden ajustar para obtener mejores resultados.

2.3.3. Análisis de componentes principales (PCA, *Principal Component Analysis*)

PCA es el método más común de reducción de dimensionalidad. Los datos son proyectados a un hiperplano de menor dimensionalidad, conservando a la vez la mayor parte posible de la varianza([22, pp. 367-372]).

Cada componente principal puede ser vista como una combinación lineal de las variables originales, la reducción de dimensionalidad se realiza descartando aquellas componentes que conservan la menor cantidad de varianza del conjunto original. A mayor correlación de las características del conjunto inicial, menor cantidad de componentes principales son necesarias para representarlas adecuadamente. Mientras mayor es el valor propio de la componente, mayor es la varianza que representa.

Existen diferentes métodos tanto para obtener las componentes principales como para definir el número apropiado de componentes necesarias para representar la información original.

La teoría subyacente de PCA se basa en estadísticos de segundo orden (i.e. correlaciones) de los datos de entrada. Por lo que consiste en un método lineal de reducción de dimensionalidad. Asumir linealidad puede llevar a predicciones suboptimas debido a que se ignoran geometrías naturales y variaciones del sistema.

2.3.4. Kernel PCA

Generalización del método de PCA para adaptarse a problemas no-lineales ([22, pp. 401-405]). KPCA es robusto ante ruido aleatorio. $\Phi : R^{m_0} \rightarrow R^{m_1}$ denota el mapeo no-lineal entre el espacio de entrada, de dimensionalidad m_0, y el espacio de características, de dimensionalidad m_1. Correspondientemente $\Phi(x_\mathrm{i})$ denota la imagen de un vector de entrada x_i inducido al espacio de características. El método puede resumirse en los cinco pasos siguientes:

1. Se debe, en primer lugar, realizar un preprocesamiento tal que todos los vectores de características posean media cero en el conjunto de entrenamiento.
2. Dada la muestra de entrenamiento $\{x_\mathrm{i}\}_{\mathrm{i}=1}^{N}$, se debe computar el Gram $K = \{k(x_\mathrm{i}, x_j)\}$ de tamaño NxN donde $k(x_\mathrm{i}, x_j) = \Phi^t(x_\mathrm{i})\Phi^t(x_j), \mathrm{i}, j = 1, 2, \ldots, N$.
3. Resolver el problema de valores propios $K\alpha = \lambda\alpha$.
4. Normalizar los vectores propios calculados bajo el requerimiento $\alpha_r^T\alpha_r = \frac{1}{\lambda_r}, r = 1, 2, \ldots, l$. Donde λ_l corresponde al valor propio más pequeño diferente de cero, asumiendo orden decreciente.
5. Obtenemos las componentes para un punto x computando las proyecciones $a_k = \sum_{j=1}^{N} \alpha_{r,j}k(x_j, x)$ con $r = 1, 2, \ldots, l$

El kernel más comúnmente utilizado es el Gaussiano definido por la siguiente ecuación:

$$\kappa(x_\mathrm{i}, x_j) = \mathrm{e}xp[\frac{-\|x_\mathrm{i} - x_j\|^2}{2\sigma^2}], \tag{2.11}$$

donde σ es un parámetro libre de ajuste llamado tamaño o ancho de kernel.

Sin embargo la selección de un kernel óptimo continua siendo un problema abierto.

2.3.5. Comparación de plantillas(*Template matching*)

La comparación de planteillas es una técnica de clasificación altamente utilizada para resolver el problema de clasificación fotométrica.

Se construyen *templates* o plantillas usando conjuntos de datos bien estudiados, supernovas con alto SNR o modelos de distribución de energía espectral. Cada muestra observada es comparada a estos modelos y clasificada, usualmente evaluando máxima verosimilitud o máxima probabilidad posterior ([23], [24]).

Se asume que cada supernova puede ser aproximada por uno de los modelos, esto se vuelve problemático al analizar grandes bases de datos. Estos métodos requieren que todos los parámetros relevantes (como *redshift* y extinción) del modelo de la curva de luz sean simultáneamente ajustados o estimados, esto implica una alta carga computacional y errores catastróficos cuando los estimados son deficientes. No aprenden automáticamente mientras se reciben nuevos datos.

2.3.6. Árboles de clasificación

Un árbol de clasificación consiste en una serie de separaciones binarias recursivas del espacio de características. Cada separación es realizada en una coordenada, resultando en dos nodos, y es elegida de tal forma que produzca el mayor decremento de una función

previamente escogida que, dependiendo de la implementación particular, puede corresponder a un índice de Gini (como en el caso de CART), entropía (como en el caso de C4.5) u otros.

Tienden a poseer poco sesgo pero alta varianza además de ser muy sensibles a pequeños cambios en el conjunto de entrenamiento utilizado para generar el árbol. Para subsanar esta falencia se pueden generar múltiples árboles de clasificación no correlacionados, cada uno ajustado a diferentes subconjuntos del conjunto de entrenamiento, y promediar los estimados para obtener un predictor final, reduciendo la varianza del método. Esta técnica recibe el nombre de bosques aleatorios (*random forests*, [25]).

2.3.7. Redes neuronales

Las redes neuronales son un área de la inteligencia computacional motivada por la observación del funcionamiento del cerebro, máquina compleja, no lineal y capaz de computación paralela, especialmente eficiente en resolver muchos problemas de interés para el área, tal como reconocimiento de patrones o clasificación([22, pp. 1-15])([22, pp. 401-405]).

En términos generales, una red neuronal es una máquina diseñada para emular de manera simplificada la manera en que funciona el cerebro, aplicada a una tarea en particular, con la capacidad de aprender y adaptarse a nueva información y conformada por una red masiva e interconectada de células de computación simple llamada neuronas. La información aprendida es guardada en los pesos sinápticos, esto es, los enlaces entre las neuronas que conforman la red. Este mecanismo es explotado en el caso del aprendizaje supervisado mediante la entrega de un conjunto de entrenamiento junto con la salida deseada, la red se adapta a esta información hasta el punto en que las nuevas modificaciones son despreciables.

Las redes neuronales poseen la ventaja de ser capaces de enfrentarse a problemas no lineales; ser altamente adaptables; no paramétricas; robustas; y bastante rápidas luego del proceso de entrenamiento. Como desventajas podemos mencionar la posible lentitud del proceso de entrenamiento, así como la dificultad que presenta visualizar la información aprendida por la red (funciona como una caja negra).

La unidad básica de la red, la neurona, sigue el modelo presentado en la figura 2.7.Los tres elementos principales que podemos mencionar son:

- Las sinapsis, caracterizadas por los pesos w. La entrada conectada a la neurona k es multiplicada por el peso w_k. Estos pesos son modificados durante el entrenamiento para obtener la salida deseada.
- Una fase de suma de los pesos entrantes que combina toda la información que llega a la neurona. También se suma una nueva señal de entrada llamada bias que tiene la capacidad de modificar los umbrales de la función de activación, esta entrada también posee un peso propio modificado durante el aprendizaje.
- Una función de activación no-lineal que limita la amplitud de salida de la neurona (generalmente a valores entre [0,1] o [-1, 1]).

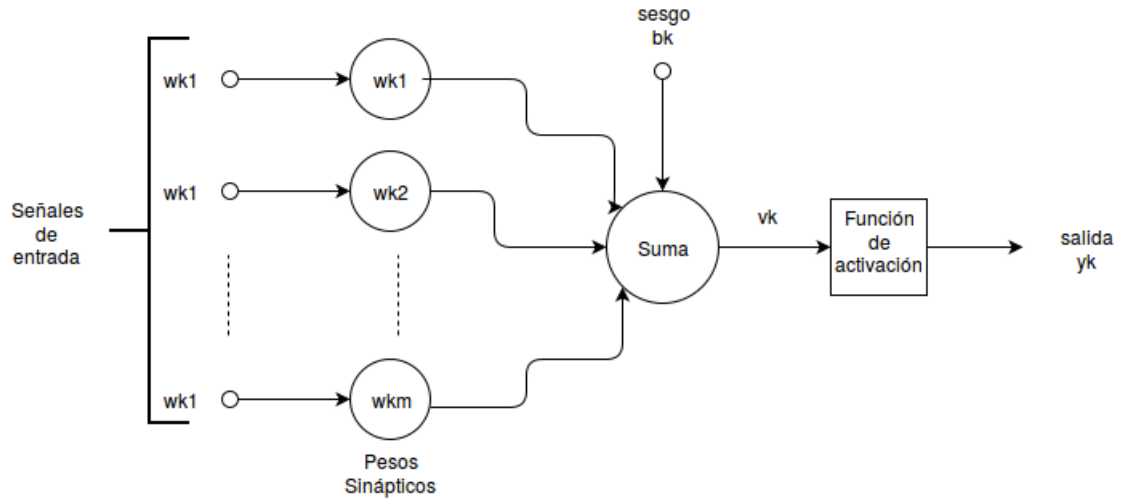

Figura 2.7: Modelo neurona

Una de las configuraciones más básicas y populares de redes neuronales es la llamada perceptrón multicapa (MLP, *Multilayer perceptron*). Esta se distingue por:

- Poseer una función de activación no lineal y diferenciable a la salida de cada neurona.
- Poseer una o más capas ocultas, que no interactúan directamente con las señales de entrada o salida.
- Exhibir un alto grado de conectividad, determinado por los pesos sinápticos de la red.

Las MLP son comúnmente entrenadas mediante un proceso de retropropagación del error, que se sustenta en dos pasos:

- Una señal de entrada se propaga hacia adelante en la red, neurona a neurona. Los pesos están fijos, los cambios están confinados a las funciones de activación de las unidades ocultas y las salidas de la red.
- Una señal de error que se origina en la salida de la red y se retro-propaga hacia atrás. La señal es producida mediante la comparación entre la salida de la red con la salida deseada. Los pesos sinápticos son modificados durante el entrenamiento para minimizar este error.

2.4. Técnicas utilizadas en el presente trabajo

2.4.1. Correntropía

La correntropía puede considerarse como una generalización de la función de correlación, que consiste en una medida de similitud en un espacio de alta dimensionalidad([26])([27]). A diferencia de la correlación, conserva estadísticos de alto orden, por lo que es capaz de utilizarse en procesos no lineales y no Gaussianos. Adicionalmente, hace uso de métodos de

kernel, convirtiéndola en una opción eficiente computacionalmente. La auto-correntropía es simétrica, semi-definida positiva y máxima para el retardo 0.

Para un proceso estocástico la función de correlación generalizada queda definida por:

$$V(s,t) = E(\kappa(x_s, x_t)), \tag{2.12}$$

Donde E corresponde a la esperanza y κ al kernel a utilizar. Al utilizar el kernel Gaussiano definido como:

$$G_\sigma(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{\|x\|^2}{2\sigma^2}), \tag{2.13}$$

y expandiendo la expresión a partir de su serie de Taylor se obtiene:

$$V(s,t) = \sum_{n=0}^{\infty} \frac{-1^n}{2^n \sigma^{2n} n!} E\|(x_s, x_t\|^{2n}. \tag{2.14}$$

Para n = 1 esta expresión es un estimador sesgado de la correlación convencional. El parámetro σ, correspondiente al ancho de banda del kernel Gaussiano, ajusta el énfasis que se les da a los estadísticos pares de orden mayor frente a los de segundo orden.

La correntropía ha sido utilizada anteriormente para encontrar el periodo de estrellas variables ([28]).

2.4.2. Correntropía ranurada

La correntropía cruzada es una medida de la similitud (mayor cuando las curvas son mas parecidas) entre dos conjuntos diferentes en un vecindario delimitado por el kernel utilizado. En su modalidad clásica, requiere que ambas curvas posean el mismo largo y espacio entre mediciones. En esta memoria se utiliza una versión discreta de la correntropía cruzada, inspirada en los trabajos [29], donde se presenta la técnica de ranurado para obtener correlaciones discretas, y [28], donde se adapta la técnica anterior para calcular autocorrentropía. El objetivo de la técnica de ranurado es permitir la compararación directa entre dos curvas sin tener que recurrir a métodos de interpolación globales. La correntropía cruzada ranurada se define como:

$$V[k\Delta\tau] = \frac{\sum_{\mathrm{i},j}^{N} G_\sigma(x_{1\mathrm{i}} - x_{2j}) * B_{k\Delta\tau}(t_{1\mathrm{i}}, t_{2j})}{\sum_{\mathrm{i},j}^{N} B_{k\Delta\tau}(t_{1\mathrm{i}}, t_{2j})}, \tag{2.15}$$

donde $k = 0, 1, 2, \ldots, [\tau_{max}/\Delta\tau]$, siendo [ ] la función de entero más cercano
$\Delta\tau$ corresponde al tamaño de la ranura

τ_{max} al retraso máximo considerado
$t_{1\mathrm{i}}$ y t_{1j} son los tiempos de las muestras $x_{1\mathrm{i}}$ y x_{2j} en sus respectivas curvas, trasladados con tal que t_{10} y t_{20} correspondan a los maximos en la banda r para cada curva
G_σ es el kernel Gaussiano definido en la sección 2.4.1
y $B_{k\Delta\tau}$ es calculado como sigue:

$$B_{k\Delta\tau}(t_{1\mathrm{i}}, t_{2j}) = \begin{cases} 1, & \mathrm{si}|(t_{1\mathrm{i}} - t_{2j} - k\Delta\tau)| < 0{,}2\Delta\tau \\ 0, & \text{de otra forma} \end{cases} \tag{2.16}$$

$\Delta\tau$ es obtenido mediante la multiplicación de dos componentes, $\Delta\tau = \mathrm{d}t * \mathrm{d}tx$. *dt* corresponde al tiempo promedio entre mediciones de las dos curvas a comparar. *dtx* es un parámetro de control que fija el tamaño de la ranura con respecto a *dt*.

2.4.3. Métrica inducida por correntropía (CIM, *correntropy induced metric*

La aplicacion de la correntropía cruzada induce una métrica en el espacio de las muestras. Esto mediante la función llamada CIM, o métrica inducida por correntropía, presentada en [30] y definida por:

$$CIM(X, Y) = \sqrt{\kappa(0) - V(X, Y)}, \tag{2.17}$$

donde X e Y son vectores del espacio de muestras (en este caso, curvas de luz),
$\kappa(0)$ es el kernel utilizado (comunmente Gaussiano) evaluado en 0,
y $V(\bullet, \bullet)$ es la función de correntropía cruzada.

La CIM es considerada una métrica debido a que cumple con las siguientes restricciones:

- Es no negativa. $CIM(X, Y) \geq 0$
- Cumple con la identidad de los indiscernibles. $CIM(X, Y) = 0 \Leftrightarrow X = Y$
- Es simétrica. $CIM(X, Y) = CIM(Y, X)$
- Cumple con la desigualdad triangular. $CIM(X, Z) \leqslant CIM(X, Y) + CIM(Y, Z)$

La prueba de estas propiedades puede encontrarse en [30], donde también se demuestra que la CIM es invariante ante la traslación al utilizarla junto a un kernel invariante(e.g. Gaussiano).

2.4.4. K vecinos más cercanos (KNN, *K-nearest neighbours*)

KNN es uno de los algoritmos de clasificación más simples, que entrega buenos resultados en espacios de baja dimensionalidad (≤ 10). Se inicializa con datos pre clasificados

correspondientes al conjunto de entrenamiento. Para cada nuevo dato a clasificar se computa la distancia (la métrica puede variar) a todos los puntos ya clasificados. Se seleccionan los elementos más cercanos, donde la clase de cada uno cuenta como un voto para clasificar la nueva muestra. 1NN es la versión más simple de este algoritmo donde cada muestra se clasifica dependiendo de la clase del elemento más cercano.

Los resultados obtenidos pueden ser altamente sensibles al parámetro K (que según la implementación puede indicar el numero de vecinos cercanos a considerar, o la distancia máxima del vecindario considerado), especialmente ante datos altamente ruidosos, presencia de outliers, o clases no balanceadas. A continuación se presentan dos implementaciones que intentan enfrentar esta debilidad mediante la adición de pesos al proceso de votación.

En [31] se presenta una implementación llamada *distance-weighted k-nearest neighbor* (WKNN) que introduce pesos al proceso de votación, estos pesos son dependientes tanto del parámetro k como de la distancia entre el elemento considerado y el elemento a clasificar. El peso asignado al elemento i queda definido por:

$$w_{\mathrm{i}} = \frac{\mathrm{d}(x, x_k) - \mathrm{d}(x, x_{\mathrm{i}})}{\mathrm{d}(x, x_k) - \mathrm{d}(x, x_1)}, \tag{2.18}$$

donde $\mathrm{d}(\bullet, \bullet)$ representa la métrica seleccionada. $x\mathrm{i}, \mathrm{i} \in [1, \ldots, k]$ representan los elementos ya clasificados y ordenados por distancia al nuevo elemento x, donde x_1 es el elemento mas cercano y x_n es el elemento más lejano considerado para la votación. Esto resulta en un peso igual a 1 para el elemento más cercano, 0 para el más lejano considerado, y una escala lineal para los elementos intermedios.

En [32] se introduce una actualización a WKNN llamada DWKNN que añade un nuevo peso que se multiplica al presentado en la implementación anterior resultando en el peso dual asignado al elemento i definido por:

$$w_{\mathrm{i}} = \frac{\mathrm{d}(x, x_k) - \mathrm{d}(x, x_{\mathrm{i}})}{\mathrm{d}(x, x_k) - \mathrm{d}(x, x_1)} \times \frac{\mathrm{d}(x, x_k) + \mathrm{d}(x, x_1)}{\mathrm{d}(x, x_k) + \mathrm{d}(x, x_{\mathrm{i}})}. \tag{2.19}$$

Esto resulta en un peso que decae más rápidamente conforme a la distancia.

Tanto WKNN como DWKNN son más estables ante variaciones del parámetro K que KNN, además de entregar mejores resultados de clasificación en múltiples escenarios.

Capítulo 3

Metodología e implementación

En este capítulo se presenta la metodología diseñada para enfrentar el problema de clasificación fotométrica de supernovas. En primer lugar se describen las características principales de la base de datos a utilizar. Luego se muestra la estructura general del método para posteriormente revisar con más detalle cada uno de los bloques que lo conforman.

3.1. Base de datos

La base de datos a utilizar contiene la información fotométrica de 21319 supernovas a lo largo de los cuatro canales correspondientes a un sistema de filtros *griz*, las longitudes de onda medias capturadas por cada filtro de un sistema de esta clase son presentadas en la tabla 3.1([16]). Los datos fueron generados mediante simulaciones que intentan emular de forma realista las condiciones de captura del DES por lo que incluyen variaciones y errores que se esperan encontrar en datos reales, tal como variaciones atmosféricas e incompletitud de datos. Por esta razón los datos poseen ciertas características que convierten a la clasificación en un problema no-trivial, estas son:

- Alto ruido
- Error heterocedástico
- Número variable de mediciones por supernova
- Tiempo variable entre mediciones
- Clases no balanceadas

Filtro	Longitud de onda media
g	512[nm]
r	668[nm]
i	792[nm]
z	912[nm]

Tabla 3.1: Sistema *griz*

Para objeto de la competición para la cual fue creada la base de datos ([5]) también se proporcionó un subconjunto de 1103 supernovas espectroscópicamente confirmadas, S de aquí en adelante. Este subconjunto posee una serie de características que lo convierten en una selección ineficiente como conjunto de entrenamiento (menor photo-z, mayor luminosidad, sobrerepresentación de supernovas tipo Ia), por lo que autores que han trabajado con el conjunto post concurso han hecho diferentes cortes de selección para compensar estos problemas, ya sea eliminando datos particularmente ruidosos, seleccionando las señales con magnitudes mayores, o equilibrando artificialmente las clases, lo que complica la comparación directa de los resultados. Para este trabajo se decidió seguir la metodología del concurso original, esto es, usar como conjunto de entrenamiento el conjunto espectroscópicamente confirmado, por motivos de comparación directa.

Adicionalmente se realizaron pruebas considerando un conjunto, SS, generado a partir de S que intenta suplir las carencias del original mediante la repetición de datos seleccionados aleatoriamente hasta alcanzar una proporción de aproximadamente 0.25/0.75 entre tipo Ia/no-Ia en once tramos de redshift. Este conjunto consta de cerca de 2000 objetos.

Estos dos conjuntos fueron utilizados como conjuntos de entrenamiento. Como conjunto de prueba se utilizó a C, que consiste en todos los datos no confirmados espectroscópicamente, es decir los correspondientes a 20216 supernovas.

Debido al imbalance en la distribución de las clases en los conjuntos C y S, se tomó la decisión de enfrentar el problema de clasificación binaria entre Ia no-Ia, aunque la metodología es fácilmente generalizable a clasificación multiclase.

En la tabla 3.2 se presenta un sumario de las características del conjunto C, la fórmula mediante la cual se calculó el flujo calibrado y su error corresponde a $FLUXCAL = 10^{[-0,4*mag+11]}$, más información al respecto puede encontrarse en [6].

Característica	Valor
Muestras promedio por curva de luz	93[muestras]
Tiempo promedio por curva de luz	121.541[dias]
Densidad de muestras promedio por curva de luz	0.846[muestras/día]
Flujo máximo promedio por curva de luz	62.311[FLUXCAL]
Error de flujo promedio por muestra	0.0002[FLUXCALERR]
Photo-z promedio	0.666

Tabla 3.2: Características conjunto C

En la tabla 3.3 se entrega el sumario de las características del conjunto S.

Característica	Valor
Muestras promedio por curva de luz	40[muestras]
Tiempo promedio por curva de luz	105.299[dias]
Densidad de muestras promedio por curva de luz	0.432[muestras/día]
Flujo máximo promedio por curva de luz	15.866[FLUXCAL]
Error de flujo promedio por muestra	0.011[FLUXCALERR]
Photo-z promedio	0.452

Tabla 3.3: Características conjunto S

La base de datos posee etiquetas para múltiples clases de supernovas, pero algunas están tan poco representadas que constituyen básicamente outliers. La distribución de clases Ia/no-Ia es presentada en las figuras 3.1, 3.2, y 3.3 para los conjuntos C, S, y SS respectivamente.

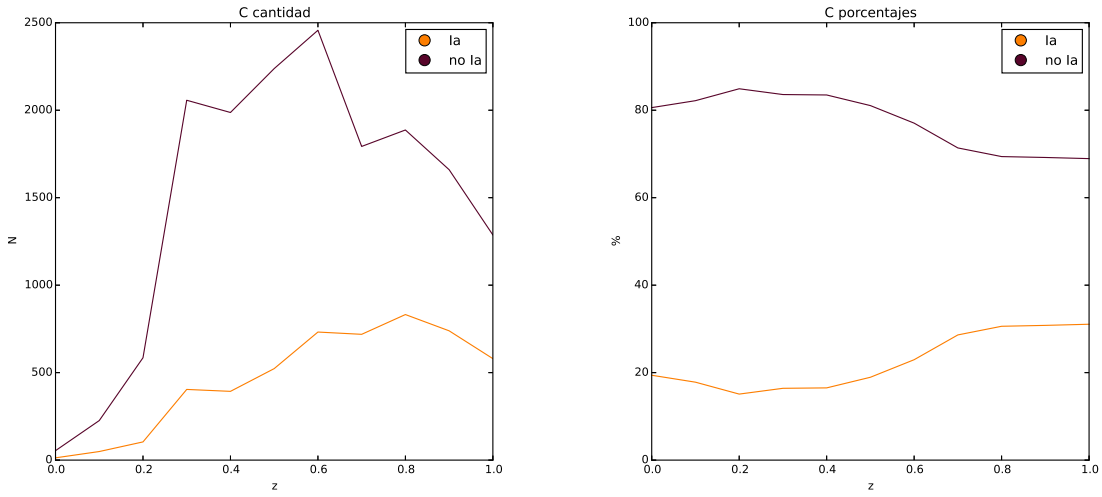

Figura 3.1: Distribución de clases en el conjunto C en función del corrimiento al rojo. A la izquierda en cantidad, a la derecha en porcentaje.

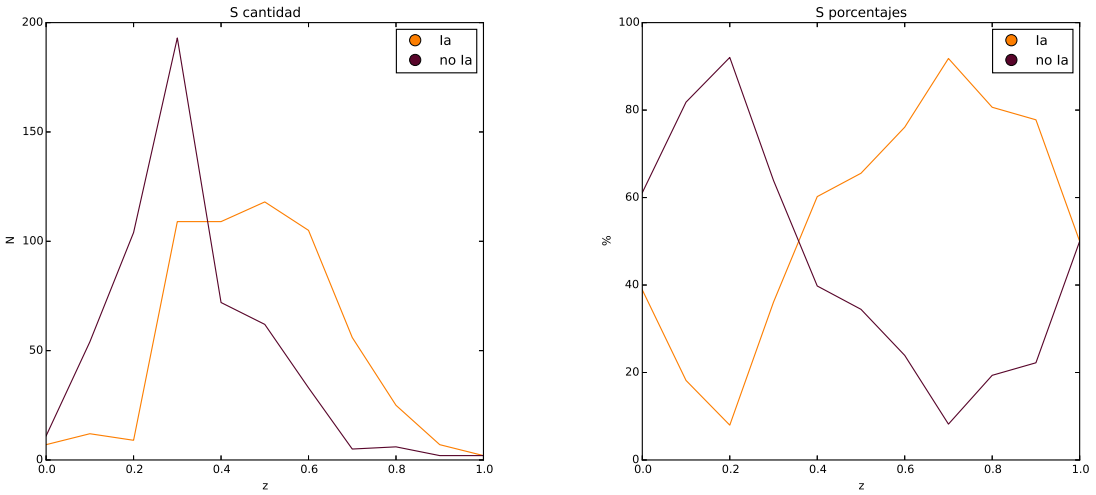

Figura 3.2: Distribución de clases en el conjunto S en función del corrimiento al rojo. A la izquierda en cantidad, a la derecha en porcentaje.

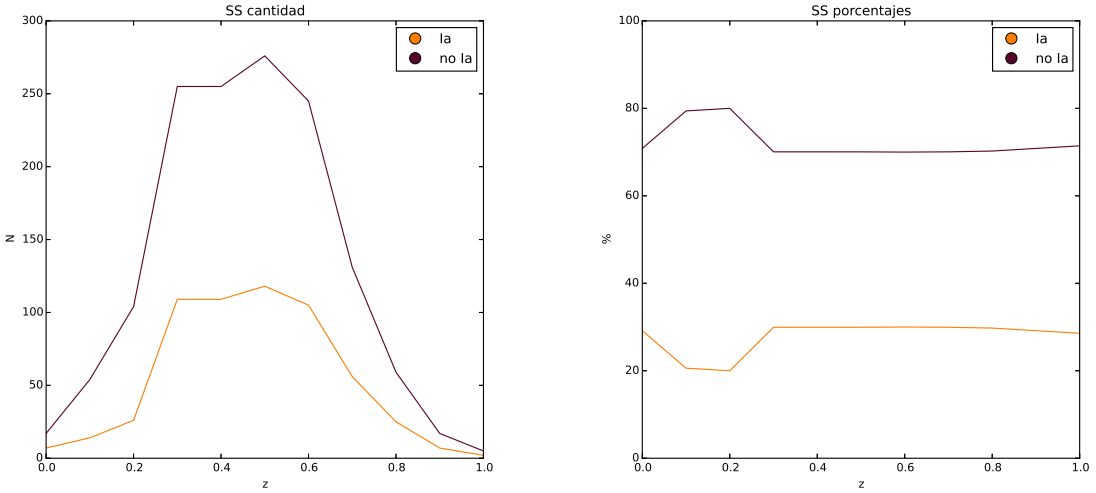

Figura 3.3: Distribución de clases en el conjunto SS en función del corrimiento al rojo. A la izquierda en cantidad, a la derecha en porcentaje.

3.2. Estructura metodología

En la figura 3.4 se presenta la estructura general de la metodología diseñada para resolver el problema de clasificación fotometrica. En primer lugar se realiza la extracción y selección de los conjuntos de datos a utilizar para el entrenamiento y prueba del método. Luego se preprocesan los datos, normalizándolos a media cero y desviación estandar 1. Los datos

normalizados son utilizados para computar la correntropía cruzada ranurada entre cada par de curvas de luz entrenamiento/prueba. Las cuatro matrices de correntropía obtenidas son transformadas a CIM y posteriormente combinadas para generar la matriz de disimilitud. Finalmente se realiza la clasificación usando alguno de los metodos presentados en el capítulo anterior.

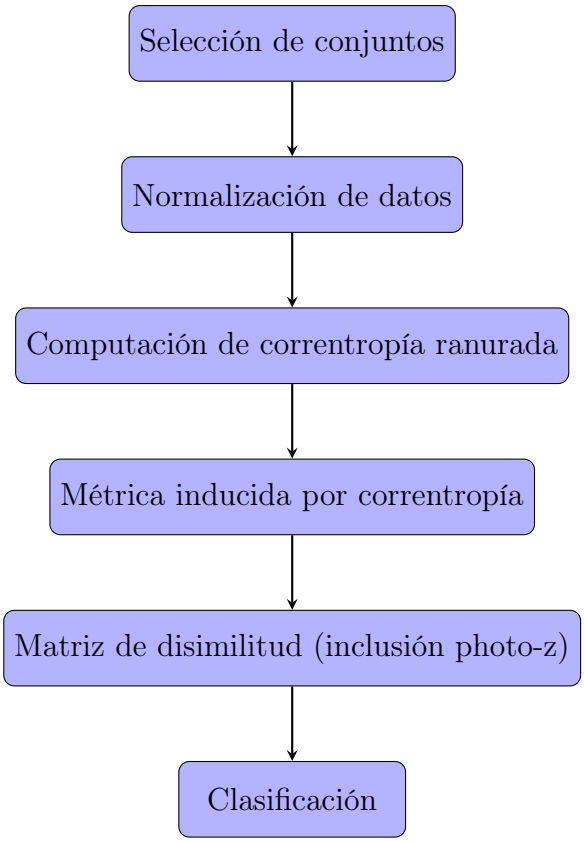

Figura 3.4: Resumen metodología

3.3. Selección de hiperparámetros

Existen dos hiperpárametros que deben ser fijados previamente al cálculo de la correntropía cruzada, estos son σ y *dtx*. σ controla el ancho del kernel Gaussiano utilizado para el cálculo de la correntropía. *dtx* es multiplicado por el tiempo entre mediciones promedio para las dos curvas de luz consideradas para fijar el ancho de la ventana utilizada durante el proceso de ranurado.

Ambos hiperparámetros fueron seleccionados mediante búsqueda de grilla (*grid search*) sobre subconjuntos de muestras provenientes del conjunto completo ($C \bigcup S$), seleccionadas aleatoriamente con repetición, con valores $\sigma \in [0{,}25\sigma_s, 0{,}5\sigma_s, 0{,}75\sigma_s, 1\sigma_s, 1{,}25\sigma_s, 1{,}5\sigma_s, 1{,}75\sigma_s]$, donde σ_s es calculado mediante la regla de Silverman [33], y d$tx \in [0{,}5, 0{,}6, 0{,}7, 1, 1{,}5, 2, 2{,}5, 3]$. Este proceso se repitio 8 veces, 4 con conjuntos de 500 muestras, y 4 con conjuntos de 1000 muestras, cada vez utilizando un subconjunto nuevo.

La correntropía ranurada fue calculada contra el conjunto S, y la medida de evaluación utilizada fue la distancia entre conjuntos menos la distancia intra conjunto.

3.4. Correntropía ranurada

En los trabajos referenciados, uno de los primeros pasos es generalmente un bloque de interpolación, comúnmente usando splines cúbicos, con el objeto de realizar extracción de características o comparación curva a curva. Para este trabajo se decidió usar una técnica de comparación discreta, esto para usar los datos reales. De la misma forma se adoptó una filosofía de preprocesamiento mínimo apoyándose en normalizaciones locales, es decir, cada curva fue normalizada independientemente a media cero y varianza uno.

El tiempo de computación necesario para una búsqueda completa de la correntropía máxima resulta excesivo por lo que para este trabajo se tomó la decisión heurística de utilizar el valor para un retardo 0 (cuando ambas matrices se encuentran alineadas en torno a los valores máximos en la banda R).

Se computa la correntropía discreta entre cada par de curvas correspondientes a supernovas de los conjuntos de entrenamiento y prueba para cada uno de los cuatro filtros, obteniéndose cuatro matrices de dimensiones n*m, donde n representa el número de supernovas en el conjunto de prueba, y m el número de supernovas en el conjunto de entrenamiento.

Adicionalmente se calculó la correlación cruzada ranurada como está presentada en [29], para comparar los resultados con los obtenidos mediante el cálculo de la correntropía.

3.5. Métrica inducida por correntropía

Los resultados obtenidos en el bloque anterior deben ser modificados con el objetivo de convertir una medida de similitud en una métrica utilizable en los siguientes pasos. Para esto se utiliza la CIM, presentada en la sección 2.4.3.

3.6. Matriz de disimilitud

Las cuatro matrices, cuyos valores han sido previamente convertidos mediante la CIM, son combinadas para generar un índice único para cada par de curvas. La forma de combinar las matrices no es única ni directamente discernible. Se debe tomar en cuenta que bajo condiciones reales no siempre se posee la información de todos los canales. Se tomó la decisión de efectuar la combinación de los canales mediante múltiples medidas con el objetivo de poder comparar sus desempeños.

La primera medida consiste en la promediación de las disimilitudes de cada canal.

$$D_1(X,Y) = (CIM_g(X,Y) + CIM_r(X,Y) + CIM_{\mathrm{i}}(X,Y) + CIM_z(X,Y))/4 \tag{3.1}$$

La segunda medida consiste en la máxima disimilitud entre los canales.

$$D_3(X,Y) = max(CIM_g(X,Y), CIM_r(X,Y), CIM_{\mathrm{i}}(X,Y), CIM_z(X,Y)) \tag{3.2}$$

La última medida utiliza la norma euclidiana de las disimilitudes de los cuatro canales.

$$D_2(X,Y) = \sqrt{(CIM_g(X,Y))^2 + (CIM_r(X,Y))^2 + (CIM_{\mathrm{i}}(X,Y))^2 + (CIM_z(X,Y))^2} \tag{3.3}$$

La combinación de los 4 canales resulta en una única matriz de dimensiones $m \times n$.

3.7. Clasificación

Para la clasificación se utilizó KNN, debido a su simplicidad y a que aprovecha la estructura generada por los bloques anteriores. También se probaron las implementaciones WKNN y DWKNN.

Pruebas iniciales indicaron que casi siempre el mejor resultado se obtenía para valores de $K \in [1, 20]$, por lo que la busqueda se restringió a valores de K menores a 50.

3.8. Inclusión de photo-z

Una de las principales problemáticas a la hora de comparar curvas de luz de objetos estelares radica en la traslación de sus frecuencias debido al efecto del corrimiento al rojo. Este problema es generalmente tratado mediante la aplicación de una *k-correction* ([34],[35]), procedimiento que induce una reconstrucción de los datos a su forma no trasladada, el problema es que esta corrección requiere de información contenida en múltiples canales no disponibles en la base de datos, por lo que se requiere otro método. Para abordar este problema se añadió una quinta matriz a combinar a la hora de generar la matriz de disimilitud. Esta está conformada por la diferencia entre photo-z de cada par de supernovas.

Esta información es combinada como un canal más en el primer y segundo caso, y como un multiplicador en el caso del máximo.

3.9. Evaluación de resultados

El desempeño de la diferentes implementaciones fue realizado utilizando las métricas presentadas en la sección 2.2.1, esto es pureza, eficiencia, y FoM, adicionalmente se incluye la información de razón de falsos positivos (*FPR, False positive ratio*).

3.10. Ejemplo ilustrativo

En la figura 3.5 se muestran las cuatro curvas de luz correspondientes a la supernova SN014719. Estas ejemplifican la morfología típica de las curvas de una supernova tipo Ia con bajo corrimiento al rojo.

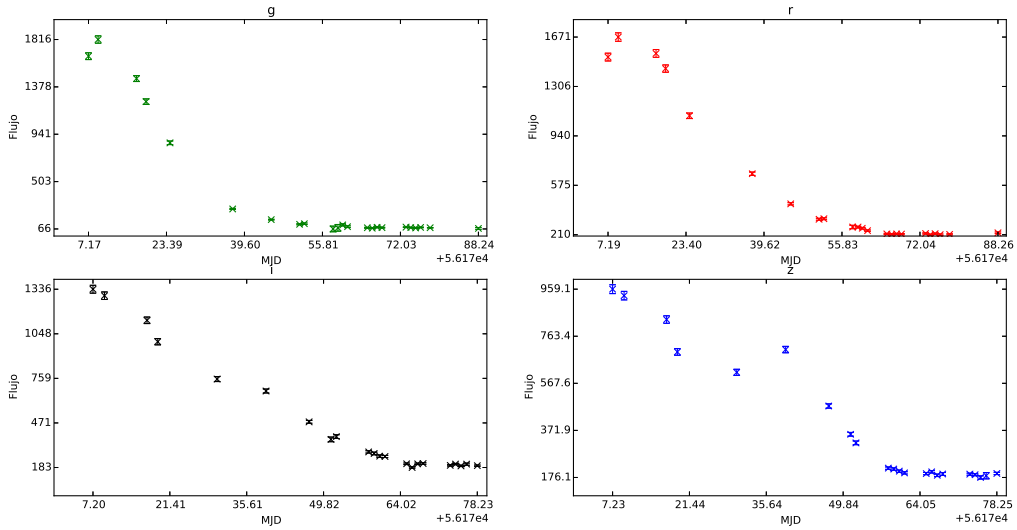

Figura 3.5: Curvas de luz SN014719

En las figuras 3.6, 3.7, y 3.8 se muestran tres curvas de luz normalizadas correspondientes al filtro r de las supernovas SN014719(Ia), SN020046(tipo Ia con mayor corrimiento al rojo), y SN000017(tipo II con mayor corrimiento al rojo) respectivamente.

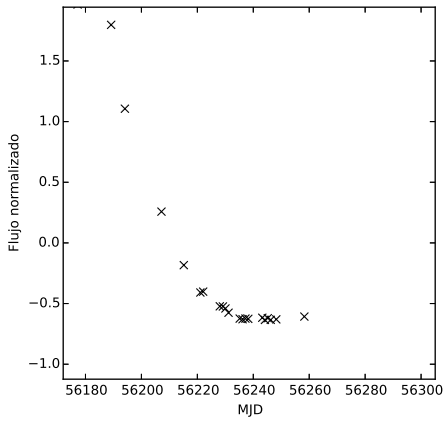

Figura 3.6: Curva de luz normalizada, filtro r, SN014719

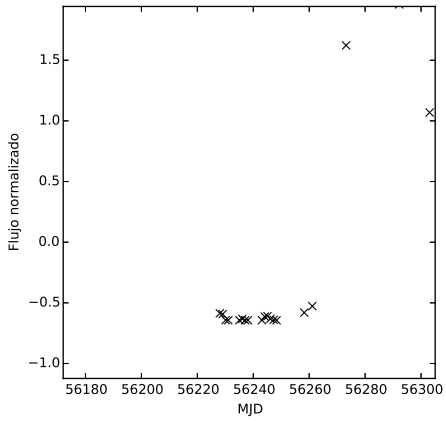

Figura 3.7: Curva de luz normalizada, filtro r, SN020046

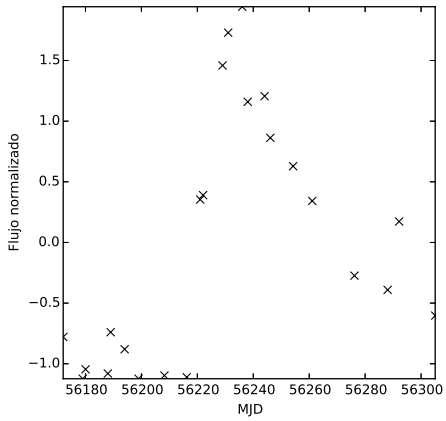

Figura 3.8: Curva de luz normalizada, filtro r, SN000017

En las figuras 3.9, 3.10, y 3.11 se muestra el resultado de la aplicación de correntropía ranurada en el canal r entre SN014719 y SN014719 , SN014719 y SN020046, y SN014719 y SN000017 respectivamente. Se puede observar que para este caso los valores para retraso 0 son consistentes con la selección del máximo, esto no es siempre cierto pero resulta una aproximación aceptable a cambio de una enorme disminución a la computación requerida.

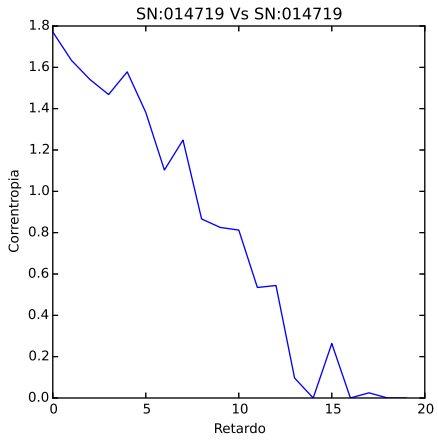

Figura 3.9: SN014719 vs SN014719, filtro r

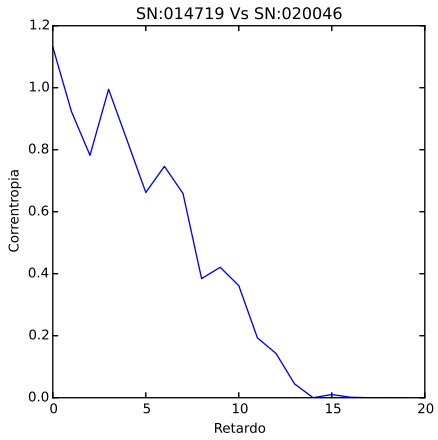

Figura 3.10: SN014719 vs SN020046, filtro r

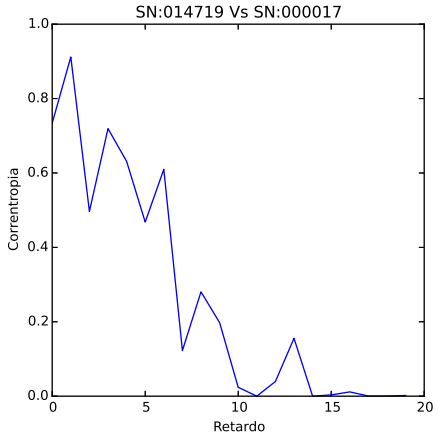

Figura 3.11: SN014719 vs SN000017, filtro r

3.11. Entorno Computacional

La implementación de la metodología presentada en este trabajo de memoria de titulo fue desarrollada en Python 2.7.9 y utiliza los siguientes paquetes:

- Ipython notebook: ambiente computacional interactivo que permite una mas cómoda aplicación, revisión y ejecución del código utilizado.
- mpld3: librería gráfica que facilita la visualización de datos en conjunción a *Ipython notebook*.
- numpy: paquete de computación científica que incluye herramientas útiles para el manejo de datos.
- pandas: conjunto de herramientas que incluye estructuras de datos útiles para el manejo y extraccion de la base de datos.
- os: módulo con funcionalidades propias del sistema operativo. Permite el fácil guardado y cargado de datos.
- matplotlib: librería gráfica especializada en visualización 2D.
- scipy: librería que agrupa herramientas científicas y matemáticas.
- pickle: módulo que permite la conversión de estructuras de python a flujo de bytes, con el objeto de guardado y cargado eficiente de los datos.
- collections: módulo que incluye estructuras de datos especializadas, en particular la estructura *counter* que es utilizada para las implementaciones de KNN.
- sklearn: librería de técnicas de aprendizaje de máquinas.
- copy: módulo que permite la copia de objetos de Python.

La computación de las matrices de correntropía fue realizada en el cluster Leftraru del NLHPC utilizando un entorno equivalente de Python. El cluster esta compuesto por 128 nodos delgados HP ProLiant SL230s Gen8 y 4 nodos gruesos HP ProLiant SL250s Gen8, cada uno conformado por 2 procesadores Intel Xeon E5-2660 de 10 cores. Adicionalmente

cuenta con 12 coprocesadores Xeon Phi 5110P de 240 cores cada uno. Utiliza Slurm como gestor de colas y Lustre como sistema de archivos distribuido. Cada cálculo en el cluster se realizó utilizando 20 cores.

Capítulo 4

Resultados

En esta sección se presentan los resultados obtenidos utilizando la metodología introducida en el capítulo anterior.

La búsqueda de grilla entregó como mejor resultado d$t = 0{,}7$ consistentemente para todos los conjuntos. En cuanto a σ los resultados presentaron un alta varianza, por lo que se tomó la decisión de generar matrices para los valores $0{,}5\sigma_s$, σ_s , y $1{,}5\sigma_s$.

4.1. Entrenamiento utilizando S

Se presentan los resultados de clasificación utilizando el conjunto espectroscópicamente confirmado (S) como conjunto de entrenamiento y el conjunto C como el conjunto de prueba. Se entregan los FoM máximos alcanzados entre los primeros 50 valores de K para cada configuración de métodos.En el anexo A se entregan los valores de los instrumentos de medición para el K en el que se alcanza FoM máximo.

En la tabla 4.1 se presentan los valores de FoM máximo utilizando la medida 1, esto es, el promedio de las disimilitudes, en la tabla 4.2 los valores utilizando la medida 2, el máximo de entre las disimilitudes, y en la tabla 4.3 los valores utilizando la medida 3, la norma euclidiana de las disimilitudes.

El mejor resultado, FoM = 0.167, es alcanzado para la medida 2, $\sigma = 0{,}5\sigma_s$, con similar valor para los tres algoritmos (diferencias del orden de 0.0001).

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.13	0.136	0.136
σ_s	0.123	0.154	0.157
$1{,}5\sigma_s$	0.132	0.114	0.118
Correlación	0.125	0.121	0.121

Tabla 4.1: FoM máximo alcanzado, medida 1(promedio), entrenamiento:S

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.167	0.167	0.167
σ_s	0.144	0.139	0.139
$1{,}5\sigma_s$	0.12	0.114	0.114
Correlación	0.112	0.106	0.106

Tabla 4.2: FoM máximo alcanzado, medida 2(máximo), entrenamiento:S

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.149	0.165	0.165
σ_s	0.127	0.154	0.157
$1{,}5\sigma_s$	0.13	0.131	0.136
Correlación	0.129	0.109	0.109

Tabla 4.3: FoM máximo alcanzado, medida 3(norma), entrenamiento:S

En promedio DWKNN entrega resultados levemente superiores a WKNN, superando ambos a KNN por un margen mas amplio.

Salvo en el caso de la medida 1, los mejores resultados se obtienen al calcular la correntropía utilizando $\sigma = 0{,}5\sigma_s$. Para los tres valores se obtiene un mejor desempeño que con la correlación.

En cuanto a las medidas utilizadas, la norma entrega mejores resultados en promedio. Siendo similares los resultados de las otras dos medidas.

4.2. Entrenamiento con *SS*

Se entregan los resultados de clasificación al utilizar el conjunto equilibrado (SS) para el entrenamiento y el conjunto C como conjunto de prueba. En las tablas 4.4, 4.5, y 4.6 se presentan los valores de FoM máximo utilizando las medidas 1, 2, y 3 respectivamente.

El mejor resultado, FoM = 0.187, es alcanzado para la medida 1, con $\sigma = 0{,}5\sigma_s$, con similar valor para WKNN y DWKNN, el resultado es superior al mejor alcanzado en el entrenamiento utilizando el conjunto S.

En promedio KNN entrega mejores resultados entre los tres algoritmos de clasificacion,

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.136	0.187	0.187
σ_s	0.124	0.083	0.083
$1{,}5\sigma_s$	0.149	0.037	0.037
Correlación	0.116	0.047	0.047

Tabla 4.4: FoM máximo alcanzado, medida 1(promedio), entrenamiento:SS

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.177	0.183	0.183
σ_s	0.133	0.171	0.171
$1{,}5\sigma_s$	0.09	0.101	0.101
Correlación	0.093	0.101	0.101

Tabla 4.5: FoM máximo alcanzado, medida 2(máximo), entrenamiento:SS

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.13	0.145	0.145
σ_s	0.117	0.071	0.071
$1{,}5\sigma_s$	0.169	0.022	0.022
Correlación	0.123	0.04	0.04

Tabla 4.6: FoM máximo alcanzado, medida 3(norma), entrenamiento:SS

sin embargo el desempeño de los tres algoritmos es peor que en el caso anterior.

$\sigma = 0{,}5\sigma_s$ es el único de los valores cuyo desempeño promedio aumenta. Tanto para $\sigma = \sigma_s$ y $\sigma = 1{,}5\sigma_s$, como para la correlación, los resultados son notablemente peores que en el caso anterior.

Similarmente los métodos de combinación de matrices entregan peores resultados en promedio, exceptuando a la medida 2.

4.3. Entrenamiento con S + photo-z

Se muestran los resultados del entrenamiento con el conjunto S, incluyendo información de photo-z en las tablas 4.7, 4.8, y 4.9, para las medidas 1, 2, y 3 respectivamente. Similarmente los métodos de combinación de matrices entregan peores resultados en promedio, exceptuando nuevamente a la medida 2.

El mejor resultado, FoM = 0.152, es alcanzado para la medida 1, con $\sigma = 0{,}5\sigma_s$, con similar valor para WKNN y DWKNN, el resultado es peor que los obtenidos en los casos anteriores.

En promedio KNN entrega mejores resultados entre los tres algoritmos de clasificación,

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.14	0.135	0.135
σ_s	0.139	0.134	0.134
$1{,}5\sigma_s$	0.139	0.132	0.132
Correlación	0.138	0.132	0.132

Tabla 4.7: FoM máximo alcanzado, medida 1(promedio), entrenamiento:$S{+}z$

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.152	0.098	0.097
σ_s	0.123	0.125	0.124
$1{,}5\sigma_s$	0.124	0.129	0.129
Correlación	0.115	0.037	0.037

Tabla 4.8: FoM máximo alcanzado, medida 2(máximo), entrenamiento:*S+z*

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.147	0.129	0.129
σ_s	0.133	0.129	0.128
$1{,}5\sigma_s$	0.133	0.132	0.132
Correlación	0.124	0.123	0.123

Tabla 4.9: FoM máximo alcanzado, medida 3(norma), entrenamiento:*S+z*

tanto WKNN como DWKNN entregan resultados bastante inferiores a los obtenidos antes de agregar la información de photo-z.

$\sigma = 1{,}5\sigma_s$ es el único de los valores cuyo desempeño promedio aumenta. Tanto para $\sigma = 0{,}5\sigma_s$ y $\sigma = \sigma_s$, como para la correlación, los resultados son notablemente peores que los obtenidos antes de la inclusión de la información de photo-z.

La medida 1 entrega resultados superiores a los obtenidos antes de añadir la información de photo-z, las otras dos medidas entregan resultados inferiores.

4.4. Entrenamiento con *SS* + photo-z

Se muestran los resultados de entrenar utilizando el conjunto *SS* como conjunto de entrenamiento, y el conjunto *C* como conjunto de prueba, añadiendo la información de photo-z. En las tablas 4.10, 4.11, y 4.12 se presentan los resultados para las medidas 1, 2, y 3 respectivamente.

El mejor resultado, FoM = 0.178, es alcanzado para la medida 3, con $\sigma = \sigma_s$, con similar valor para WKNN y DWKNN, el resultado es superior a los obtenidos al entrenar con los conjuntos *S* y *S+z*, pero inferior al obtenido al entrenar con el conjunto *SS*.

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.14	0.144	0.144
σ_s	0.155	0.149	0.148
$1{,}5\sigma_s$	0.153	0.154	0.154
Correlación	0.167	0.118	0.118

Tabla 4.10: FoM máximo alcanzado, medida 1(promedio), entrenamiento:*SS+z*

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.141	0.1	0.1
σ_s	0.149	0.162	0.162
$1{,}5\sigma_s$	0.123	0.155	0.154
Correlación	0.115	0.033	0.033

Tabla 4.11: FoM máximo alcanzado, medida 2(máximo), entrenamiento:*SS+z*

	KNN	WKNN	DWKNN
$0{,}5\sigma_s$	0.159	0.173	0.173
σ_s	0.172	0.178	0.178
$1{,}5\sigma_s$	0.163	0.128	0.128
Correlación	0.133	0.125	0.125

Tabla 4.12: FoM máximo alcanzado, medida 3(norma), entrenamiento:*SS+z*

En promedio KNN entrega los mejores resultados entre los tres algoritmos de clasificación. Los tres algoritmos entregan resultados superiores a los obtenidos al clasificar utilizando el conjunto *SS* sin información de photo-z.

$\sigma = 0{,}5\sigma_s$ entrega el peor desempeño promedio entre los tres valores. Para los valores $\sigma = \sigma_s$ y $\sigma = 1{,}5\sigma_s$, los resultados promedio son los mejores a lo largo de todos los conjuntos de entrenamiento.

La medida 1 y 3 entregan resultados superiores a los obtenidos con cualquier otro conjunto de entrenamiento.

4.5. Comparación con resultados de bibliografía

En la tabla 4.13 se muestra la comparación de los resultados obtenidos con aquellos presentados en la bibliografía. La entrada nombrada clasificación ingenua corresponde a los resultados de clasificación si todas las supernovas fueran clasificadas como tipo Ia. Se comparan dos resultados obtenidos en esta memoría, el primero corresponde al mejor resultado obtenido al entrenar utilizando el conjunto *S*, el segundo corresponde al mejor resultado obtenido al entrenar con cualquiera de los cuatro conjuntos. Los resultados de la bibliografía corresponden a los reportados en [6], [7], y [8] para el entrenamiento utilizando el conjunto espectroscópicamente confirmado (*S* en esta memoria).

	FoM	Pureza	Eficiencia
Clasificación ingenua	0.09	0.22	1
Esta memoria (S)	0.167	0.556	0.567
Esta memoria (Mejor)	0.187	0.824	0.307
Richards et al	0.14	0.5	0.5
Karpenka et al	0.12	0.32	0.94
Ishida & de Souza	0.26	0.63	0.71

Tabla 4.13: Comparación de resultados

Conclusión

A modo de conclusión no queda resuelto el problema de combinación de canales, siendo en promedio superiores por un leve margen las medidas 1 y 3, especialmente para la clasificación utilizando el conjunto S incluyendo información de photo-z como conjunto de entrenamiento.

Los resultados obtenidos son comparables a aquellos presentados en la bibliografía para los casos en que se entrenó utilizando conjuntos similares al set espectroscópicamente confirmado.

Aunque el método propuesto entrega resultados competitivos, el tiempo necesario para generar la matriz de disimilitud resulta excesivo (siete horas para generar la matriz de m*n, utilizando 20 cores de Leftraru), lo que dificulta una búsqueda más exhaustiva de los hiperparámetros óptimos. Existen implementaciones de KNN que agilizan este proceso cuya aplicación queda fuera de los alcances de esta memoria (e.g. *kd-trees*). Asi también es posible optimizar el cálculo de la correntropía mediante su implementación en otro lenguaje más básico que Python.

La adición de información de corrimiento al rojo fotométrico no induce una mejora en los resultados obtenidos para la clasificación utilizando el conjunto S, sin embargo esto si ocurre en el conjunto SS. Esto se puede explicar por la distribución de photo-z en el conjunto S lo que indica que la utilidad de esta información es altamente dependiente de la calidad del conjunto de entrenamiento.

Bibliografía

[1] R. Kessler, B. Basset, P. Belov, V. Bhatnagar, H. Campbell, et al. Results from the supernova photometric classification challenge. *PASP*, 122(898):1415, December 2010.

[2] John D. Fix. *Astronomy: journey to the cosmic frontier*. McGraw-Hill, New York, 5th edition, 2008.

[3] J.P. Bernstein, R. Kessler, S. Kuhlmann, and H. Spinka. Dark energy survey supernovae: Simulations and survey strategy. In *2008 Cosmology Proceedings of the 43rd Rencontres de Moirond*, pages 71–74, 2008.

[4] Z. Ivezic, J.A. Tyson, R. Allsman, J. Andrew, R. Angel, et al. Lsst: from science drivers to reference design and anticipated data products. Living Document, 2008. arXiv:0805.2366.

[5] R. Kessler, A. Conley, S. Jha, and S. Kuhlmann. Supernova photometric classification challenge, 2010. arXiv:1001.5210.

[6] J.W. Richards, D. Homrighausen, P.E. Freeman, C.M. Schafer, and D. Poznanski. Semi-supervised learning from photometric supernova classification. *MNRAS*, 419:1121, January 2012.

[7] N.V. Karpenka, F. Feroz, and M.P. Hobson. A simple and robust method for automated photometric classification of supernovae using neural networks. *MNRAS*, 429:1278, February 2012.

[8] E.E.O. Ishida and R.S. de Souza. Kernel pca for type ia supernovae photometric clasification. *MNRAS*, 430:509–532, March 2012.

[9] M. Turatto. *Classification of Supernovae*, pages 21–36. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[10] S. Perlmutter, G. Aldering, G. Goldhaber, P R.A. Knop, Nugent, et al. Measurements of ω and λ from 42 high-redshift supernovae. *ApJ*, 517:565–586, June 1999.

[11] A.G. Riess, B.P. Schmidt, A.V. Filippenko, P. Challis, A. Clocchiatti, et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *AJ*, 116(3):1009–1038, September 1998.

[12] R. Kessler, A.C. Becker, D. Cinabro, J. Vanderplass, J.A. Frieman, et al. First-year sloan digital sky survey-ii supernova results: Hubble diagram and cosmological parameters. *ApJS*, 185:32, November 2009.

[13] A. Conley, J. Guy, M. Sullivan, N. Regnault, P. Astier, et al. Supernova constraints and systematic uncertainties from the first three years of the supernova legacy survey. *ApJS*, 191:1, January 2011.

[14] S. Benitez-Herrera, F. Röpke, W. Hillebrandt, C. Mignone, M. Bartelmann, and J. Weller. Model-indipendent reconstruction of the expansion history of the universe from type ia supernovae. *MNRAS*, 419:513, January 2012.

[15] C. Sterken and J. Manfroid. *Astronomical photometry: a guide.* Springer, 1992.

[16] James Binney and Michael Merrifield. *Galactic Astronomy.* Princeton university press, Princeton, New Jersey, 1998.

[17] P. Astier, J. Guy, N. Regnault, R. Pain, E. Aubourg, et al. The supernova legacy survey: measurement of ω_M and ω_λ and w from the first year data set. *A&A*, 447:31–48, February 2006.

[18] D.G. York, J. Adelman, J.E. Anderson, S.F. Anderson, J. Annis, et al. The sloan digital sky survey: Technical summary. *AJ*, 120:1579, September 2000.

[19] Institute for Astronomy University of Hawai. pan-starrs, 2005.

[20] J.S. Bloom and J.W. Richards. *Data Mining and Machine Learning in Time-Domain Discovery Classification*, pages 89–112. CRC Press, 2012.

[21] R.H. Bartels, J.C. Beatty, and B.A. *An Introduction to Splines for Use in Computer Graphics and Geometric Modelling.* Morgan Kaufmann, San Francisco, 1st edition, 1998. from MathWorld–A wolfram Web Resource http://mathworld.wolfram.com/CubicSpline.html.

[22] Simon Haykin. *Neural Networks and Learning Machines.* Pearson, New Jersey, 3th edition, 2009.

[23] N.V. Kuznetsova and B.M. Conolly. A probabilistic approach to classifying supernovae using photometric information. *ApJ*, 659(1):530–540, April 2007.

[24] J.L. Tonry, B.P. Schmidt, B. Barris, P. Candia, et al. Cosmological results from high-z supernovae. *ApJ*, 594(1):1–25, September 2003.

[25] T.K. Ho. Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 278–282, 1995.

[26] A. Gunduz, A. Hedge, and J.C. Principe. Correntropy as a novel measure for nonlinearity tests. In *Proceedings of the 2006 International Joint Conference on Neural Networks*, 2006.

[27] I. Santamaría, P.P. Pokharel, and J.C. Príncipe. Generalized correlation function: Definition, properties, and application to blind equalization. *IEEE Transactions on Signal Processing*, 54(6), June 2006.

[28] P. Huijse, P. Estevez, P. Zegers, J.C. Principe, and P. Protopapas. Period estimation in astronomical time series using slotted correntropy. *IEEE Signal Processing Letters*, 18(6):371–374, June 2011.

[29] R.A. Edelson and J.H. Krolik. The discrete correlation function - a new method for analyzing unevenly sampled variability data. *ApJ*, 333:646–659, October 1988.

[30] W. Liu, P.P. Pokharel, and J.C. Príncipe. Correntropy: Properties and applications in non-gaussian signal processing. *IEEE Transactions on Signal Processing*, 55(11), November 2007.

[31] S.A. Dudani. The distance-weighted k-nearest neighbor rule. *IEEE Transactions on System, Man, and Cybernetics*, 6:325–327, November 1976.

[32] J. Gou, L. Du, Y. Zhang, and T. Xiong. A new distance-weighted k-nearest neighbor classifier. *Journal of Information & Computational Science*, 9(6):1429–1436, 2012.

[33] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[34] E. Hubble. Effects of red shifts on the distribution of nebulae. *ApJ*, 84:517–554, 1936.

[35] P Nugent, A. Kim, and S. Perlmutter. K corrections and extinction corrections for type ia supernovae. *Publications of the Astronomical Society of the Pacific*, 114(798):803–819, August 2002.

Apéndice A

Anexo: resultados completos

En este apéndice se presentan los resultados entregados por los instrumentos de médicion para el k donde se obtuvo FoM máximo para cada combinacion de conjuntos de entrenamiento, medidas y σ utilizados.

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.13	0.645	0.345	0.055	9
WKNN	0.136	0.674	0.333	0.046	3
DWKNN	0.136	0.678	0.331	0.045	4
Medida 2					
KNN	0.167	0.49	0.691	0.208	23
WKNN	0.167	0.556	0.567	0.131	15
DWKNN	0.167	0.556	0.566	0.131	15
Medida 3					
KNN	0.149	0.736	0.308	0.032	7
WKNN	0.165	0.755	0.326	0.031	3
DWKNN	0.165	0.754	0.326	0.031	3

Tabla A.1: $\sigma = 0{,}5\sigma_s$, entrenamiento S

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.123	0.712	0.273	0.032	4
WKNN	0.154	0.827	0.251	0.015	12
DWKNN	0.157	0.837	0.251	0.144	12
Medida 2					
KNN	0.144	0.49	0.593	0.178	17
WKNN	0.139	0.517	0.529	0.142	21
DWKNN	0.139	0.517	0.528	0.142	21
Medida 3					
KNN	0.127	0.676	0.309	0.43	9
WKNN	0.157	0.805	0.271	0.019	15
DWKNN	0.161	0.816	0.269	0.017	15

Tabla A.2: $\sigma = \sigma_s$, entrenamiento S

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.132	0.748	0.266	0.026	7
WKNN	0.114	0.71	0.253	0.03	15
DWKNN	0.118	0.724	0.252	0.028	15
Medida 2					
KNN	0.12	0.375	0.723	0.348	26
WKNN	0.114	0.357	0.657	0.3	41
DWKNN	0.114	0.386	0.655	0.3	41
Medida 3					
KNN	0.13	0.349	0.854	0.459	26
WKNN	0.131	0.75	0.261	0.025	18
DWKNN	0.136	0.766	0.26	0.023	18

Tabla A.3: $\sigma = 1{,}5\sigma_s$, entrenamiento S

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.125	0.338	0.859	0.485	28
WKNN	0.121	0.761	0.235	0.02	12
DWKNN	0.121	0.762	0.235	0.02	12
Medida 2					
KNN	0.112	0.356	0.723	0.377	26
WKNN	0.106	0.326	0.667	0.338	43
DWKNN	0.106	0.368	0.654	0.324	41
Medida 3					
KNN	0.129	0.791	0.231	0.017	4
WKNN	0.109	0.787	0.197	0.015	10
DWKNN	0.109	0.787	0.197	0.015	10

Tabla A.4: Correlación, entrenamiento S

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.136	0.725	0.291	0.032	4
WKNN	0.187	0.823	0.307	0.019	4
DWKNN	0.187	0.824	0.307	0.019	4
Medida 2					
KNN	0.177	0.736	0.368	0.038	2
WKNN	0.183	0.733	0.383	0.04	7
DWKNN	0.183	0.733	0.383	0.04	7
Medida 3					
KNN	0.13	0.719	0.283	0.032	4
WKNN	0.145	0.701	0.332	0.041	1
DWKNN	0.145	0.701	0.332	0.041	1

Tabla A.5: $\sigma = 0{,}5\sigma_s$, entrenamiento SS

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.124	0.713	0.274	0.032	32
WKNN	0.083	0.659	0.214	0.032	1
DWKNN	0.083	0.659	0.213	0.032	1
Medida 2					
KNN	0.133	0.722	0.267	0.032	2
WKNN	0.171	0.754	0.338	0.032	1
DWKNN	0.171	0.754	0.338	0.032	1
Medida 3					
KNN	0.117	0.705	0.264	0.032	32
WKNN	0.071	0.635	0.192	0.032	1
DWKNN	0.071	0.635	0.192	0.032	1

Tabla A.6: $\sigma = \sigma_s$, entrenamiento *SS*

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.149	0.736	0.309	0.032	29
WKNN	0.037	0.542	0.13	0.032	1
DWKNN	0.037	0.542	0.13	0.032	1
Medida 2					
KNN	0.09	0.674	0.228	0.032	2
WKNN	0.101	0.685	0.24	0.032	2
DWKNN	0.101	0.685	0.24	0.32	2
Medida 3					
KNN	0.169	0.753	0.336	0.032	28
WKNN	0.022	0.471	0.098	0.032	49
DWKNN	0.022	0.471	0.098	0.032	1

Tabla A.7: $\sigma = 1{,}5\sigma_s$, entrenamiento *SS*

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.116	0.549	0.403	0.095	33
WKNN	0.047	0.578	0.151	0.032	1
DWKNN	0.047	0.578	0.151	0.032	1
Medida 2					
KNN	0.093	0.674	0.228	0.032	2
WKNN	0.101	0.685	0.24	0.032	2
DWKNN	0.101	0.685	0.24	0.032	2
Medida 3					
KNN	0.123	0.7	0.28	0.03	26
WKNN	0.04	0.56	0.144	0.032	1
DWKNN	0.04	0.565	0.143	0.032	1

Tabla A.8: Correlación, entrenamiento *SS*

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.14	0.36	0.888	0.458	7
WKNN	0.135	0.341	0.921	0.514	17
DWKNN	0.135	0.341	0.921	0.514	17
Medida 2					
KNN	0.152	0.739	0.331	0.032	5
WKNN	0.098	0.441	0.472	0.173	5
DWKNN	0.097	0.441	0.468	0.171	5
Medida 3					
KNN	0.147	0.653	0.382	0.058	1
WKNN	0.129	0.362	0.813	0.414	27
DWKNN	0.129	0.365	0.805	0.404	25

Tabla A.9: $\sigma = 0{,}5\sigma_s$, entrenamiento $S{+}z$

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.139	0.352	0.904	0.48	7
WKNN	0.134	0.339	0.919	0.517	15
DWKNN	0.134	0.339	0.918	0.517	15
Medida 2					
KNN	0.123	0.384	0.717	0.331	6
WKNN	0.125	0.409	0.665	0.278	6
DWKNN	0.124	0.41	0.658	0.274	6
Medida 3					
KNN	0.133	0.347	0.882	0.479	14
WKNN	0.129	0.345	0.863	0.474	26
DWKNN	0.128	0.345	0.862	0.473	26

Tabla A.10: $\sigma = \sigma_s$, entrenamiento $S{+}z$

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.139	0.46	0.63	0.213	1
WKNN	0.132	0.356	0.852	0.445	4
DWKNN	0.133	0.356	0.852	0.445	4
Medida 2					
KNN	0.124	0.34	0.842	0.471	4
WKNN	0.129	0.444	0.615	0.222	8
DWKNN	0.129	0.445	0.61	0.219	8
Medida 3					
KNN	0.133	0.341	0.902	0.503	6
WKNN	0.132	0.343	0.888	0.49	16
DWKNN	0.132	0.343	0.888	0.49	16

Tabla A.11: $\sigma = 1{,}5\sigma_s$, entrenamiento $S{+}z$

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.138	0.36	0.873	0.448	5
WKNN	0.132	0.34	0.898	0.502	7
DWKNN	0.132	0.34	0.898	0.502	7
Medida 2					
KNN	0.115	0.289	0.968	0.688	4
WKNN	0.037	0.367	0.215	0.099	6
DWKNN	0.037	0.385	0.213	0.098	6
Medida 3					
KNN	0.124	0.321	0.913	0.557	9
WKNN	0.123	0.315	0.923	0.578	19
DWKNN	0.123	0.315	0.923	0.578	19

Tabla A.12: Correlación, entrenamiento $S{+}z$

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.14	0.729	0.297	0.032	15
WKNN	0.144	0.517	0.547	0.148	19
DWKNN	0.144	0.516	0.547	0.148	19
Medida 2					
KNN	0.141	0.73	0.298	0.032	4
WKNN	0.1	0.556	0.338	0.078	4
DWKNN	0.1	0.516	0.547	0.148	4
Medida 3					
KNN	0.159	0.745	0.322	0.032	6
WKNN	0.173	0.761	0.335	0.03	26
DWKNN	0.173	0.761	0.336	0.03	26

Tabla A.13: $\sigma = 0{,}5\sigma_s$, entrenamiento $SS{+}z$

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.155	0.597	0.468	0.091	3
WKNN	0.149	0.505	0.586	0.166	17
DWKNN	0.148	0.491	0.61	0.183	13
Medida 2					
KNN	0.149	0.737	0.309	0.032	4
WKNN	0.162	0.747	0.326	0.032	6
DWKNN	0.162	0.747	0.326	0.032	6
Medida 3					
KNN	0.172	0.671	0.426	0.06	1
WKNN	0.178	0.692	0.415	0.05	22
DWKNN	0.178	0.692	0.415	0.05	22

Tabla A.14: $\sigma = \sigma_s$, entrenamiento *SS+z*

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.153	0.716	0.335	0.038	25
WKNN	0.154	0.492	0.632	0.188	16
DWKNN	0.154	0.492	0.632	0.188	16
Medida 2					
KNN	0.123	0.717	0.272	0.032	3
WKNN	0.155	0.674	0.38	0.053	4
DWKNN	0.154	0.67	0.38	0.054	4
Medida 3					
KNN	0.163	0.667	0.408	0.059	7
WKNN	0.128	0.486	0.536	0.164	20
DWKNN	0.128	0.484	0.537	0.165	18

Tabla A.15: $\sigma = 1{,}5\sigma_s$, entrenamiento *SS+z*

	FoM	Pureza	Eficiencia	FPR	K
Medida 1					
KNN	0.167	0.722	0.329	0.04	19
WKNN	0.118	0.385	0.684	0.316	5
DWKNN	0.118	0.384	0.684	0.316	5
Medida 2					
KNN	0.115	0.351	0.753	0.402	2
WKNN	0.033	0.38	0.195	0.092	4
DWKNN	0.033	0.381	0.195	0.091	4
Medida 3					
KNN	0.133	0.561	0.447	0.101	7
WKNN	0.125	0.45	0.585	0.206	19
DWKNN	0.125	0.449	0.585	0.207	19

Tabla A.16: Correlación, entrenamiento *SS+z*