

Ciencias de Datos: Preparación de los Datos para ser usados por Técnicas Inteligentes

Jose Aguilar

Contenido

¡Todos estamos generando datos de eventos!

tomando el tren

abastecimiento de combustible el
automóvil

comprando un café

ajustando la temperatura



poniendome una multa por exceso de
velocidad

enviando un correo electrónico

haciendo una cita

viendo esta clase

Mi teléfono inteligente

Más de 14 sensores



Acelerómetro

GPS

Sensor de luz ambiental

WIFI

Bluetooth

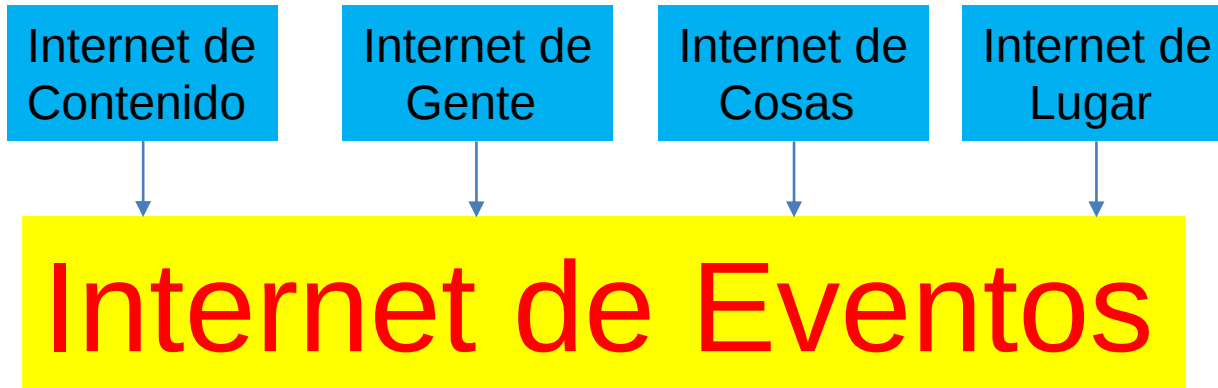
Sensor giroscópico

Sensor Proximidad

Pantalla táctil

Cámara(s)

Internet de eventos: 4 fuentes de datos de eventos



Eventos

Definición (Evento, atributo) Sea E el universo de eventos. Los eventos pueden caracterizarse por diversos atributos:

- $\#actividad(e)$ es la actividad asociada al evento e .
- $\#time(e)$ es la marca de tiempo del evento e .
- $\#resource(e)$ es el recurso asociado al evento e .
- $\#trans(e)$ es el tipo de transacción asociado al evento e ,

Definición Secuencias: son la forma más natural de presentar trazas en un registro de eventos. Operadores útiles en las secuencias:

- Una secuencia es una cadena de eventos de longitud dada ($|\sigma| = n$).
- Dos secuencias se pueden agregar $\sigma_1 \oplus \sigma_2$ dando como resultado una secuencia de longitud $|\sigma_1| + |\sigma_2|$.
- Una secuencia tiene una cabeza $hd_k(\sigma)$ que consiste de los primeros k elementos (si es posible). $hd_0(\sigma)$ es la secuencia vacía
- Una secuencia tiene una cola $tl_k(\sigma)$ que es la secuencia compuesta de los últimos k elementos. $tl_0(\sigma)$ es la secuencia vacía
- $\sigma \upharpoonright X$ es la proyección de σ en algún subconjunto $X \subseteq A$, (conjunto de datos)

Desafíos al extraer registros de eventos

Desafío 1: Correlación

Los eventos en un registro de eventos se agrupan por caso. Requiere la correlación de eventos, es decir, los eventos deben estar relacionados entre sí. Se deben recuperar los eventos dispersos en varias tablas o incluso en varios sistemas.

¿Cómo identificar eventos y sus casos correspondientes?

Desafío 2: marcas de tiempo

Los eventos deben ordenarse por caso. En principio, tal ordenamiento no requiere marcas de tiempo. Sin embargo, cuando se combinan datos de diferentes fuentes, uno normalmente necesita depender de las **marcas de tiempo para ordenar los eventos (en orden de aparición)**. Esto puede ser problemático debido a múltiples relojes y registros demorados, tiempos más largos de medida, etc. Como resultado, es imposible reconstruir el orden de los eventos en un día determinado.

- Una forma de abordar este problema es asumir solo un **orden parcial de eventos** (es decir, no un orden total) y posteriormente utilizar algoritmos de minería de procesos dedicados para esto.
- Otra forma es **"adivinar" el orden según el conocimiento del dominio o los patrones frecuentes a lo largo de los días.**

Desafíos al extraer registros de eventos

Desafío 3: instantáneas

Los casos pueden tener una vida útil que se extiende más allá del período registrado, por ejemplo, un caso se inició antes del comienzo del registro de eventos o aún se estaba ejecutando cuando se detuvo la grabación. Por lo tanto, es importante darse cuenta de que los registros de eventos normalmente solo proporcionan una instantánea de un proceso de ejecución más largo.

- Cuando la duración promedio grabación es corta en comparación con la duración del caso, es mejor **eliminar casos incompletos**.
- En muchos casos, las actividades iniciales y finales son conocidas, lo que facilita la filtración del registro de eventos: **simplemente elimine todos los casos con "cabeza" o "cola" faltante**.

Desafíos al extraer registros de eventos

Desafío 4: Alcance

El cuarto problema es el alcance del registro de eventos. Información de la empresa los sistemas pueden tener miles de tablas con datos relevantes para la empresa. **¿Cómo saber qué tablas incorporar?** El conocimiento del dominio es necesario para localizar los datos requeridos y para determinar su alcance.

Desafío 5: granularidad

En muchas aplicaciones, los eventos en el registro de eventos se encuentran en un nivel diferente de granularidad que las actividades relevantes para los usuarios finales. Algunos sistemas producen eventos de bajo nivel que son demasiado detallados. Afortunadamente, **hay varios enfoques para preprocesar registros de eventos de bajo nivel.**

Calidad de eventos

A nivel de entidad hay tres problemas potenciales:

- **Falta en el registro:** La entidad existe (o existió) en realidad, pero no fue registrada.

Por ejemplo, ocurrió un evento (por ejemplo, tomar una muestra de sangre) pero no fue capturado por el sistema de información.

- **Desaparecido en la realidad:** La entidad no existe y nunca existió en la realidad, pero fue grabada.

Por ejemplo, una cita médica programada nunca se realizó debido a una emergencia, pero de todos modos fue registrada por el sistema de información.

- **Oculto en el registro:** La entidad se registró y existe (o existió) en la realidad, pero está oculta en un conjunto de datos menos estructurado de mayor tamaño.

Por ejemplo, la misma entidad puede aparecer varias veces en el registro de eventos.

Calidad de eventos

Clasificación de los problemas relacionados con los atributos del evento.

- **Atributo perdido:** El atributo no se ha registrado para un evento en particular.

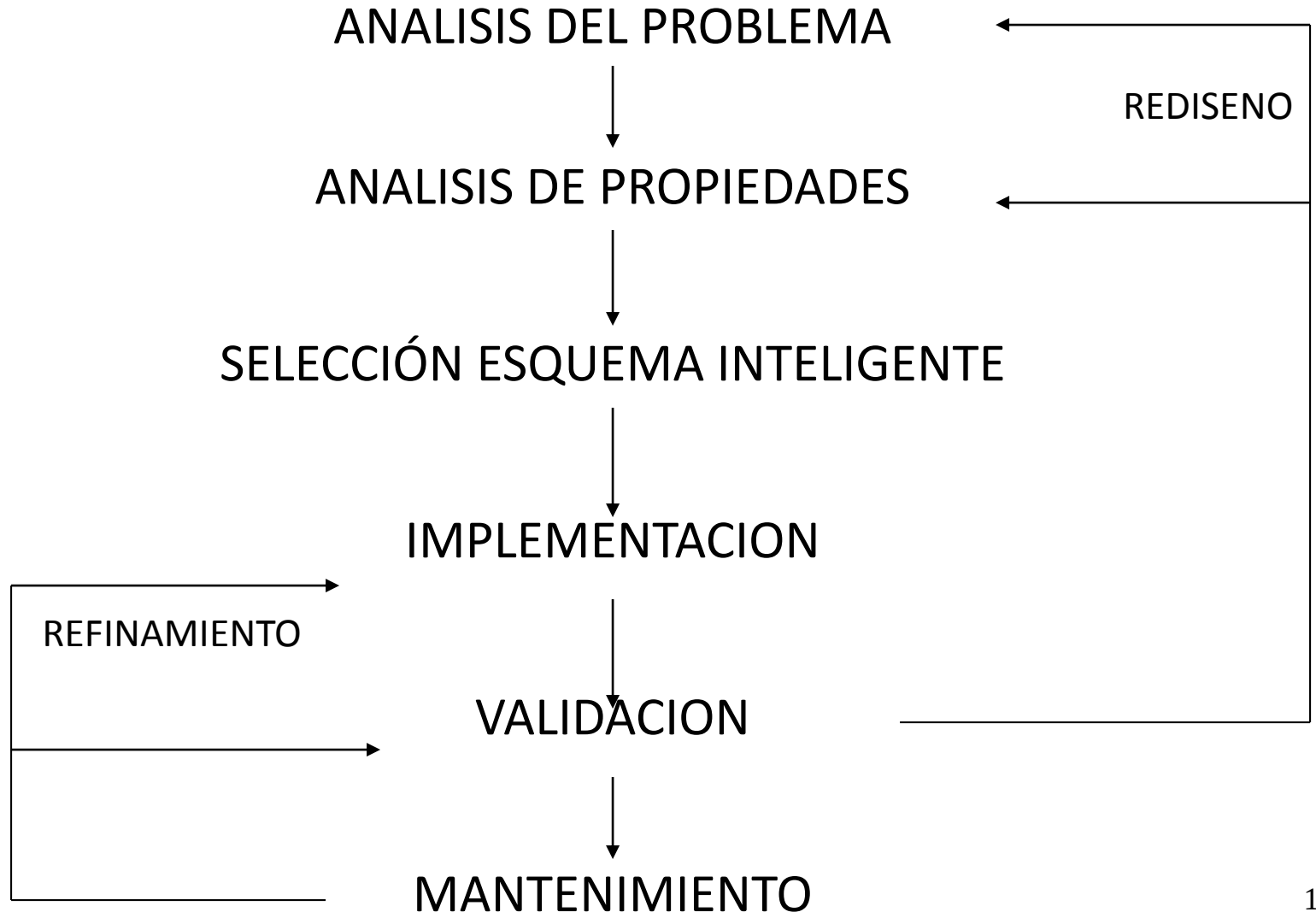
Por ejemplo, falta la marca de tiempo de un evento.

- **Atributo incorrecto:** El valor registrado del atributo de evento es incorrecto.

- **Atributo impreciso:** El valor del atributo de evento es demasiado impreciso.

Por ejemplo, el valor de una indicación de fecha y hora es demasiado grueso o la dirección es incompleta.

Metodología para Desarrollar Sistemas Inteligentes



Metodología para Desarrollar Sistemas Inteligentes

- **ANALISIS DEL PROBLEMA**
 - IDENTIFICAR LAS DIFERENTES TAREAS DEL PROBLEMA
 - IDENTIFICAR PROPIEDADES DEL PROBLEMA Y DE LAS TAREAS
- **ANALISIS DE PROPIEDADES**
 - QUE TECNICAS INTELIGENTES MEJOR SE ADAPTAN

Metodología para Desarrollar Sistemas Inteligentes

- **SELECCIÓN DE TECNICA**
 - SON NECESARIOS DIFERENTES TI?
 - EXISTE UNA TECNICA QUE CUBRA TODO?
 - SE SOLAPAN TAREAS?
 - ...
- **IMPLEMENTACION**
 - PARADIGMA DE PROGRAMACION (POO)
 - AMBIENTES Y HERRAMIENTAS

Metodología para Desarrollar Sistemas Inteligentes

- **VALIDACION**
 - PRUEBAS
 - VERIFICACION DE FUNCIONAMIENTO
- **MANTENIMIENTO**
 - COMPONENTES ADAPTATIVOS
 - AMBIENTES CAMBIANTES

Metodología para Desarrollar Sistemas Inteligentes

- **CADA TECNICA INTELIGENTE**
 - TIENE UNA PROPIEDAD COMPUTACIONAL
 - ES MAS SUSCEPTIBLE PARA CIERTOS PROBLEMAS
 - TIENE CIERTAS LIMITACIONES
- **PROBLEMAS DE ALGUNAS TECNICAS INTELIGENTES**
 - TIEMPO DE RESPUESTA
 - ADAPTATIVIDAD (TOLERANCIA A RUIDO)
 - EXPLICACION
 - ADQUISICION DE CONOCIMIENTO (APRENDER)
 - ESCALABILIDAD
 - REPRESENTATIVIDAD

Metodología para Desarrollar Sistemas Inteligentes: SISTEMAS HIBRIDOS INTELIGENTES

- RAZONES PARA HACER SHI
 - AUMENTAR CAPACIDADES DE LAS TECNICAS
 - MULTIPLICIDAD DE TAREAS
 - MULTIFUNCIONALIDAD
- ELEMENTOS A CONSIDERAR EN LA INTEGRACION
 - ADQUISICION DE CONOCIMIENTO
 - ROBUSTES
 - NIVEL DE RAZONAMIENTO
 - EXPLICACION

SISTEMAS HIBRIDOS INTELIGENTES

- **ADQUISICION DE CONOCIMIENTO**
 - ELITESCO
 - INTERPRETATIVO
 - REPRESENTATIVO
- **NIVEL DE RAZONAMIENTO**
 - TAREAS DE INFERENCIA=>COGNITIVAS
 - RECONOCIMIENTO=>ASOCIACIONES MENTALES
- **EXPLICACION**
 - ACEPTACION DE LA INFORMACION
 - ENTENDIMIENTO DEL PROCESO DE RAZONAMIENTO

SISTEMAS HIBRIDOS INTELIGENTES

TECNICA	ADQUIS. CONOCIM	ROBUST.	RAZON.	EXPLIC.
SISTEMAS EXPERTOS	BAJO	BAJO	ALTO	ALTO
LOGICA DIFUSA	BAJO	ALTO	ALTO	ALTO
RED NEURONAL	ALTO	ALTO	MEDIO	BAJO
COMPUTAC. EVOLUTIVA	ALTO	MEDIO	MEDIO	MEDIO
SISTEMA CLASIFIC.	MEDIO ALTO	ALTO	ALTO	

SISTEMAS HIBRIDOS INTELIGENTES

- **CLASIFICACION DE LOS SIH (SEGÚN FUNCIONALIDAD, ARQUITECTURA Y REQUER. COMUNIC)**
 - REEMPLAZO DE FUNCIONES
 - INTERCOMUNICACION
- **REEMPLAZO DE FUNCIONES**
 - AUMENTAR EFICIENCIA
 - AUMENTAR VELOCIDAD
- **INTERCOMUNICACION**
 - INDEPENDENCIA MODULAR
 - INTERCAMBIO DE INFORMACION
 - MECANISMO DE CONTROL

SISTEMAS HIBRIDOS INTELIGENTES

OTRA CLASIFICACION DE SIH

- FUSION
- TRANSFORMACION
- COMBINACION
- ASOCIACION

SISTEMAS HIBRIDOS INTELIGENTES

- **FUSION NEURODIFUSA**
 - DATOS DE ENTRADA DIFUSOS
 - OPTIMIZACION DE SISTEMAS DIFUSOS
 - IMPLEMENTACION DE OPERADORES DIFUSOS CON RN
- **TRANSFORMACION NEURODIFUSA**
 - APRENDER FUNCIONES DE MEMBRESIA
 - APRENDER REGLAS
 - APRENDER GRUPOS DIFUSOS

SISTEMAS HIBRIDOS INTELIGENTES

- **ENFOQUES DE COMBINACION**
 - NEURO-SIMBOLICO
 - GENETICO-SIMBOLICO
 - NEURO-DIFUSO
 - NEURO-DIFUSO BASADO EN CASOS
 - CONTROL NEURODIFUSO
 - CONTROL SIMBOLICO
 - CONNTROL NEURO-DIFUSO-GENETICO
- **SISTEMAS ASOCIATIVOS**
 - TAREAS
 - PROGRAMAS

Preparación de los Datos para ser usados por Técnicas Inteligentes

Pasos

- **Recopilación de datos**
- **Detección de características**
- **Extracción de características**
- **Reducción y selección de características**
- **Selección del algoritmo y construcción del modelo**

Preparación de los Datos para ser usados por Técnicas Inteligentes

Recopilación de datos:

Este paso tiene como objetivo recopilar información sobre un estudio de caso. Se establecen procedimientos de medición para capturar datos de sensores físicos o digitales. Dichos datos describen el estado actual o histórico de un banco de pruebas experimental definido.

Este aspecto es crucial debido a:

- Complejidad y escalabilidad del entorno (por ejemplo Internet),
- Dimensión y evolución constante de la carga de trabajo
- información privada y personal involucrada, entre otros.

Preparación de los Datos para ser usados por Técnicas Inteligentes

Recopilación de datos:

- Disponibilidad de datos
- Diversidad significativa en el costo de recopilar diferentes características
- Más, mejor (en términos de número de instancias, no necesariamente en términos de número de dimensiones / características)
- El aumento de los datos puede generar sobreajuste: ¡maneje con cuidado!
- Los datos etiquetados son los mejores. Si no está etiquetado: Podría configurar estudios/expertos para obtener datos etiquetados

Preparación de los Datos para ser usados por Técnicas Inteligentes

Recopilación de datos:

Al tratar de mejorar el rendimiento, es posible que necesite

- Más datos
- Mejores características de entrada
- Diferentes modelos de aprendizaje automático o hiperparámetros
- Etc.

Una forma de decidir si necesita más / mejores datos

- Compare su precisión en el entrenamiento y el conjunto de prueba
- **Si el entrenamiento incorrecto**, entonces probablemente necesite más y mejores datos, o podría necesitar un modelo de aprendizaje / hiperparámetros diferente.
- **Si la prueba establece una precisión mucho peor que la precisión del conjunto de entrenamiento**, la recopilación de más datos suele ser una buena dirección,

Preparación de los Datos para ser usados por Técnicas Inteligentes

Recopilación de datos:

Tendencias principales:

- **Adquisición de datos real:** se extrae directamente del sistema.
- **Generación de datos:** trata de imitar condiciones del reales similares del sistema bajo estudio, copiando el modelado de interacciones reales a través de scripts
- **Emulación sistema:** tiene como objetivo establecer escenarios tan cercanos como los reales, donde uno o más actores pueden emular intencionalmente las interacciones comunes del sistema

Preparación de los Datos para ser usados por Técnicas Inteligentes

Recopilación de datos:

Pasos para la medición de datos:

- Captura de eventos/flujos/paquetes,
- Medición y exportación de eventos/flujos/paquetes,
- Recopilación de datos.
- Asignación de etiqueta.

Recopilación de datos: Muestreo de datos

- Para reducir el número de instancias enviadas al algoritmo DM.
- Para apoyar la selección de solo aquellos casos en los cuales la respuesta es relativamente homogénea.
- Para ayudar con el balance de datos y
- Para evitar la ocurrencia de eventos raros.
- Para dividir un conjunto de datos

Recopilación de datos: Muestreo de datos

Formas de muestreo de datos

- Muestra aleatoria simple.
- Muestra equilibrada: la muestra se diseña de acuerdo con una variable objetivo y se fuerza a tener una cierta composición de acuerdo con un criterio predefinido.
- Muestra de conglomerados: se agrupan en G agrupamientos mutuamente inconexos,

Our Algorithm

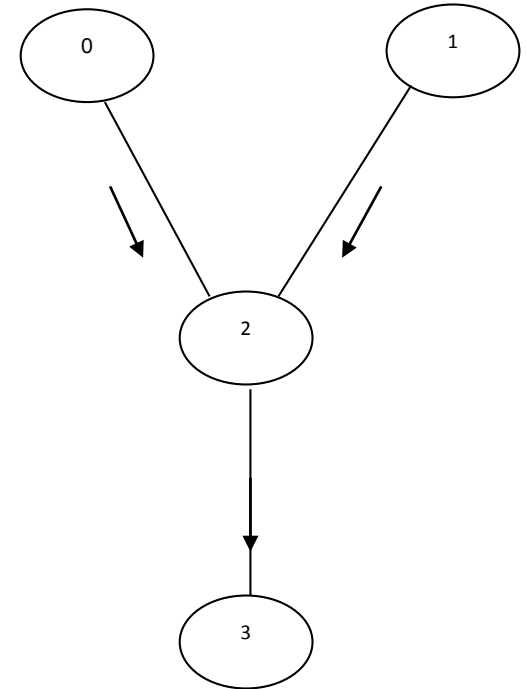
1. Preprocessing
2. Extracting descriptors
3. Specifying primitive patterns (classes)
4. Identifying Successions
5. Detecting events
6. Extracting sequences
7. Specifying temporal patterns (chronicles)
8. Updating the Chronicles Base

Phase	Description	Entrée	Traitement	Résultat	Outils
1) Prétraitement	Simulation des situations d'intérêt	Scénarios de simulation	Développement des scripts de création des scénarios	Fichier de traces *.tr	NS – 2, Langage OTCL

Simulations sur NS-2

I Hypothèses:

- Situation de perte par **congestion**
- Congestion provoqué par deux flux : UDP et TCP
 1. UDP/TCP
 2. UDP/UDP
 3. TCP/TCP
- Horloge synchronisée
- Etude d'un trafic régulier « Ping » : (ICMP)



II .Scénario des simulations:

- 3 congestions dans la même simulation
- Envoie d'un paquet ICMP chaque 0.1s

Feature Detection

- La detección de características es para identificar la presencia de un cierto tipo de característica u objeto en una imagen.
- La detección de características se logra generalmente mediante el estudio de las variaciones estadísticas de ciertas regiones y sus antecedentes para localizar actividades inusuales.
- Una vez que se ha detectado una característica interesante, la representación de esta característica se usará para compararla con todas las características posibles conocidas por el procesador.

Feature Detection

- Detección de rostros en una imagen llena de personas y otros objetos
- Detección de rasgos faciales como ojos, nariz y boca
- Detección de bordes, para que una característica se pueda extraer y comparar con otra

Feature Detection

- Todos los métodos dan una matriz de miles de columnas con una fila para cada imagen

	V_1	V_2	V_3	V_4	V_5	V_6							V_k
Image 1													
Image 2													
Image 3													
....													
....													
Image n													

- Un conjunto de datos complicado con muchas variables
- Difícil de analizar con estadísticas univariadas (por ejemplo, ANOVA)
- Debe usar técnicas multivariadas

Preparación de los Datos para ser usados por Técnicas Inteligentes

- **Feature Extraction:** el proceso de extracción de características (FE) es uno de los pasos más importantes porque permite construir las características que pueden proporcionar información sobre el estado del proceso.

procedimiento FE calcula diferentes métricas que reflejan propiedades específicas en los datos recopilados.

- El objetivo principal es obtener descriptores que caractericen mejor el problema.
- El proceso FE da como resultado una tabla estructurada formada por columnas de atributos y una columna opcional adicional con la etiqueta o clase.
- En caso de que el estado sea desconocido, las muestras no están etiquetadas.

Feature Extraction

- Conjunto de métodos y técnicas que buscan determinar la información relevante para caracterizar datos.
- Se busca generar un vector de características (feature vector) el cual contiene esas características relevantes extraídas.
- El vector de características resultante normalmente se obtiene mediante **reducción de dimensionalidad**, es decir, se busca mapear los datos originales que se encuentran representados con altas dimensionalidades a la mas baja posible con la menor perdida de información.

Feature Extraction

Clasificación de los métodos usados

- **Estadísticos**

- Se extraen propiedades estadísticas de los datos: como la distribución de su duración, tiempo que ocurre, (por ejemplo: en comunicaciones tiempo entre llegadas de paquetes, longitudes de paquetes, etc.)
- Presentan baja complejidad computacional y altas velocidades
- Son únicos para cierto tipo de datos.
- Medias, varianzas, moda, .. (basado en distribución normal de la variable)
- Cuidado con el **momento**

- **Geométricos, topológicos o de grafos**

- Buscan representar de forma gráfica información sobre los datos
- por ejemplo el contorno de un objeto o los nodos en una red.

- **Seriales**

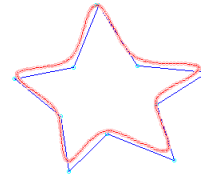
- Los datos de series de tiempo se pueden ver como una secuencia de eventos indexados en orden cronológico.
- El proceso de extracción normalmente se realiza en un tiempo discreto.
- Se usan técnicas como la transformada de Fourier o sus variaciones

Feature Extraction

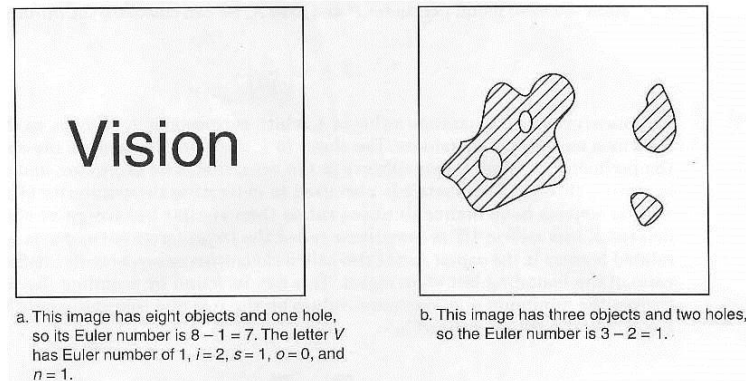
Clasificación de los métodos usados

Otras técnicas de Extracción de características en imágenes

- Características de la forma (silueta)
- Características del objeto binario: el área, el perímetro, el número de Euler, las proyecciones, etc.



$$A_i = \sum_{r=0}^{height-1} \sum_{c=0}^{width-1} I_i(r, c)$$



número de Euler

Feature Extraction

Clasificación de los métodos usados

Otras técnicas de Extracción de características en imágenes

- Características del histograma

The first-order histogram probability $P(g)$

$$P(g) = \frac{N(g)}{M}$$

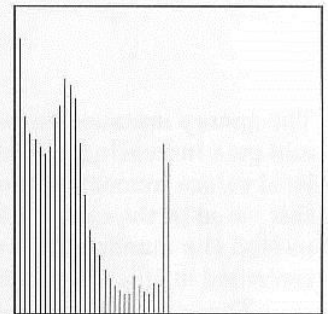
$P(g)$: probability of gray level g in image

$N(g)$: number of pixel with gray level g in image

M : total number of pixel in image



i. Dark image.



j. Histogram of (i) appears shifted to the left.

Las características basadas en la probabilidad de histograma de primer orden son Media, Desviación estándar, Energía, Entropía. $ENTROPY = -\sum_{g=0}^{L-1} P(g) \log_2 [P(g)]$

Feature Extraction

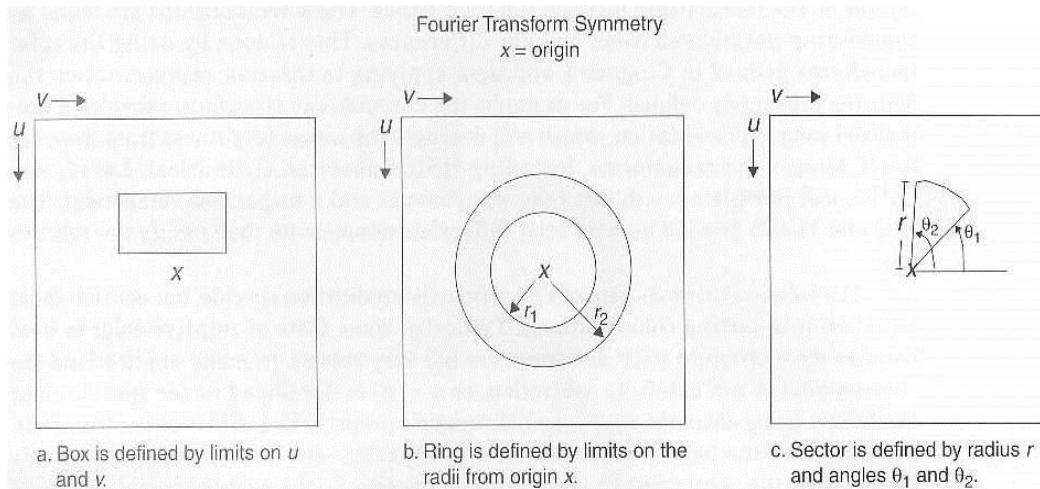
Clasificación de los métodos usados

Otras técnicas de Extracción de características en imágenes

- Características del color: iluminación, sensores, filtrado óptico y cualquier proceso fotográfico o de impresión.

- Características espectrales

$$\text{Spectral RegionPower} = \sum_{u \in \text{REGION}} \sum_{v \in \text{REGION}} |T(u, v)|^2$$



Phase	Description	Entrée	Traitement	Résultat	Outils
2) Extraction des descripteurs	Les attributs qui caractérisent un <u>individu</u> : Vecteur (délai, %perte, débit)	Fichier de traces NS *.tr	Développement de scripts pour calculer les critères à utiliser et les écrire dans un fichier	Fichier de traces raffinées *.dat	Compilateur Perl

Résultats des simulations et traitement

I .Fichiers de traces générés par NS:

[0]	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]
Action	Instant	Nsrc	Ndest	Type	Taille	Flag	IDflux	@src	@dest	N°Seq	IDpaquet	Estampille temporaire

- Traces collectés depuis tous les nœuds du réseau
- Informations sur tous les flux générés dans la simulation

II .Traitement:

1. Filtrage des informations concernant le flux ICMP
2. Filtrage côté récepteur puis côté émetteur
3. Calcul des descripteurs : ***Gigue / Délai de bout en bout / Débit / % de perte***

Résultats des simulations et traitement

III .Scripts de calcul

- Fenêtre glissante de 10 paquets
- Script côté récepteur (10 paquets reçus)
- Script côté émetteur (10 acquittements reçus)

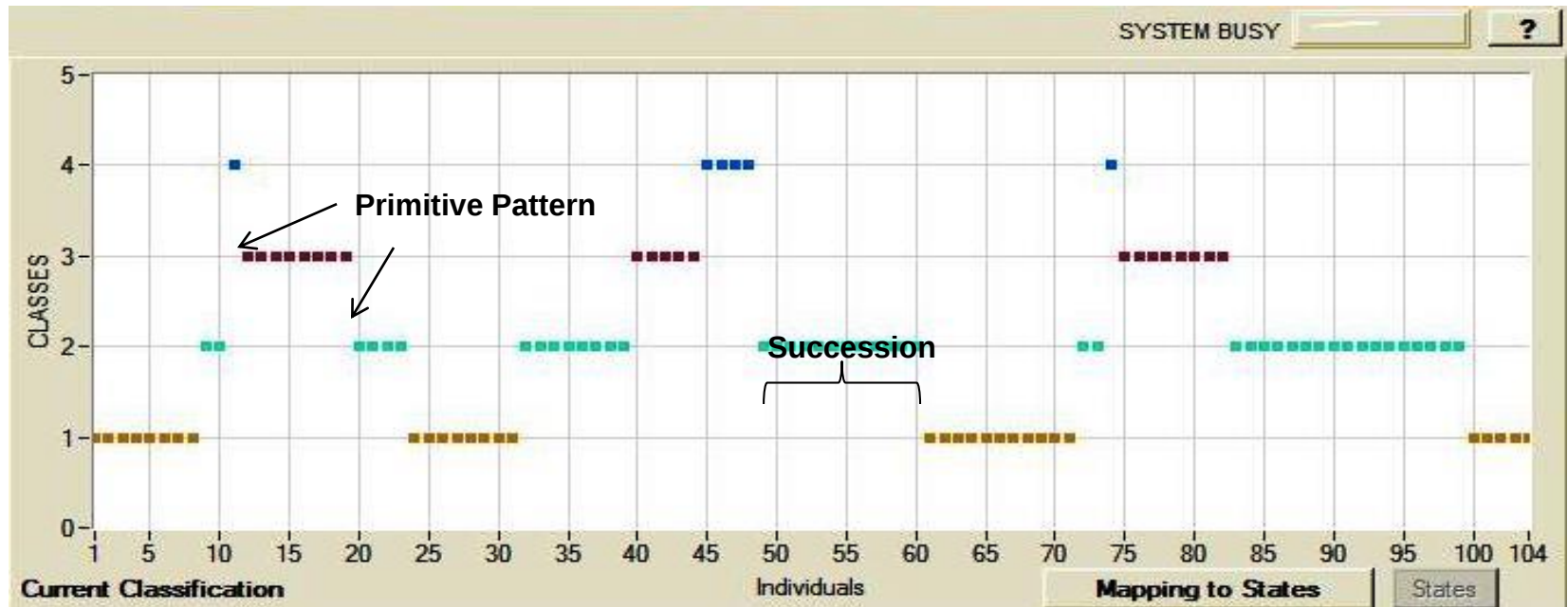
IV .Fichier résultant (entrée SALSA)

[0]	[1]
N° Individu	Instant

[0]	[1]	[2]	[3]
gigue	Délai	Débit	% perte

Phase	Description	Entrée	Traitement	Résultat	Outils
3) Spécification des motifs primitifs	• Vecteur qui décrit une unité de temps • Classe d'un individu	Fichier de données *.dat	Auto : Exploitation de la courbe résultante	Courbe de classification	SALSA
4) Détection des successions	Un intervalle de temps dans lequel tous les individus appartiennent à la même classe	Fichier de données *.dat	Auto: Exploitation de la courbe résultante	Courbe de classification	SALSA

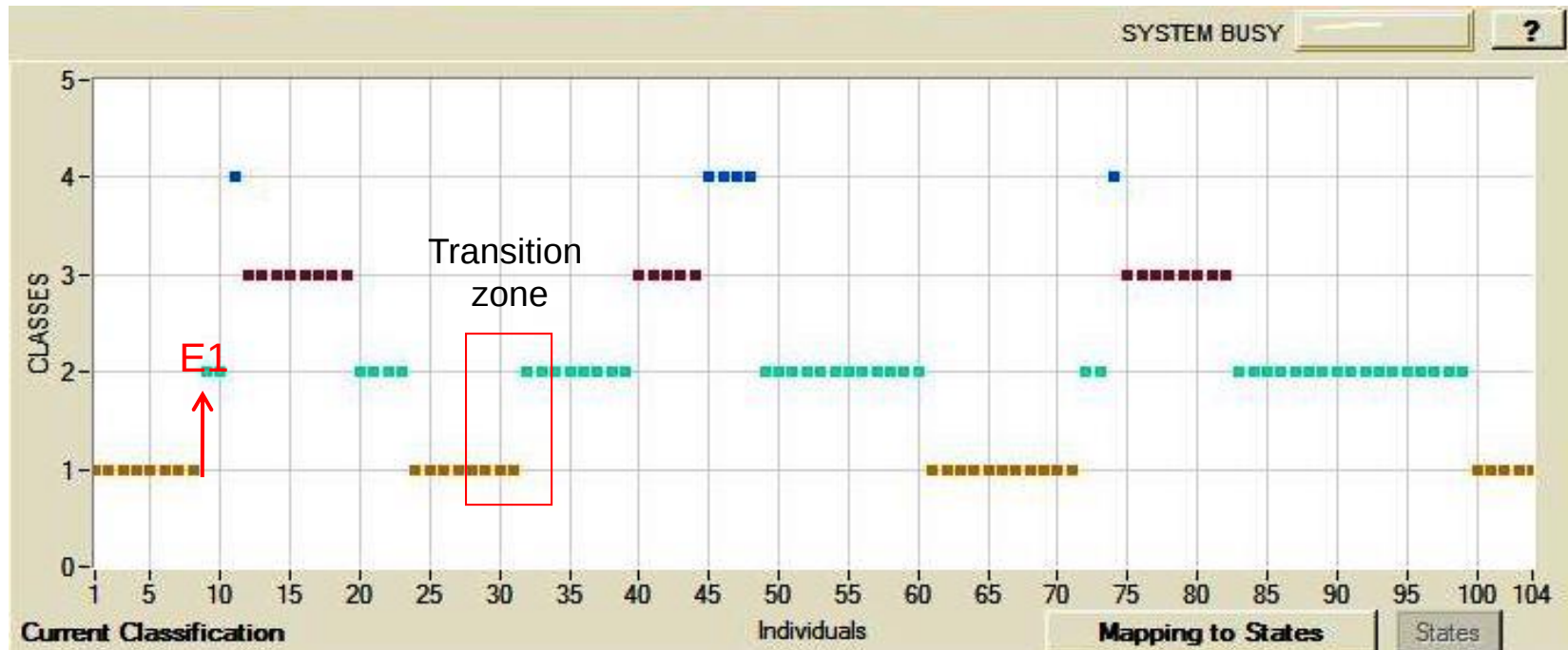
Specifying primitive patterns and Identifying Successions



- Based on the fuzzy clustering methodology LAMDA (Learning Algorithm for Multivariate Data Analysis) and its software tool called SALSA.
- Determination of the global membership degree of an individual to an existing class, considering all the contributions of each of its attributes (descriptors in our case), called the marginal adequacy degree (MAD).

Phase	Description	Entrée	Traitement	Résultat	Outils
5) Détection des événements	•Un changement de situation (transition d'une classe à une autre)	Fichier de données *.dat	Exploitation du « Membership Matrix »	Courbe de classification	SALSA
			Choix du degré d'appartenance pour déterminer le début et la fin de chaque succession		
			Choix de la date de l'événement		
6) Extraction des séquences	Occurrence de plusieurs événements	Fichier de données *.dat	Auto: Exploitation de la courbe résultante	Courbe de classification	SALSA

Detecting events and Extracting sequences



Event is determined from the transition zone and is described by an initial time and a final time

Extraction of sub-sequences leads on the notion of similarity between events.

Sequence 1 : E1, E2, E3, E4

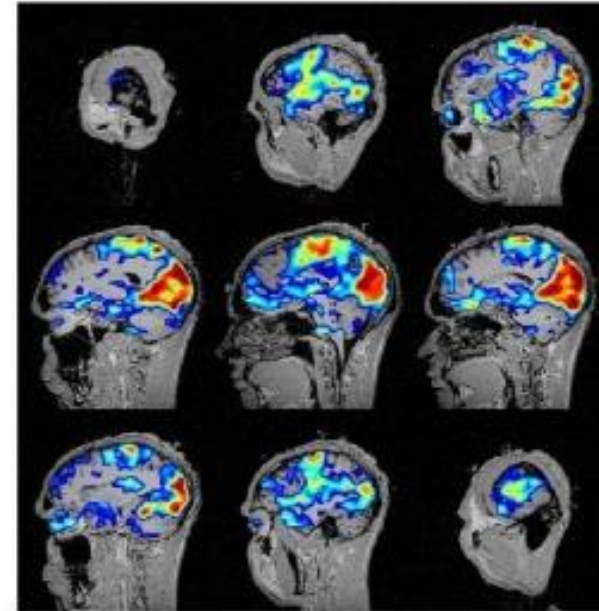
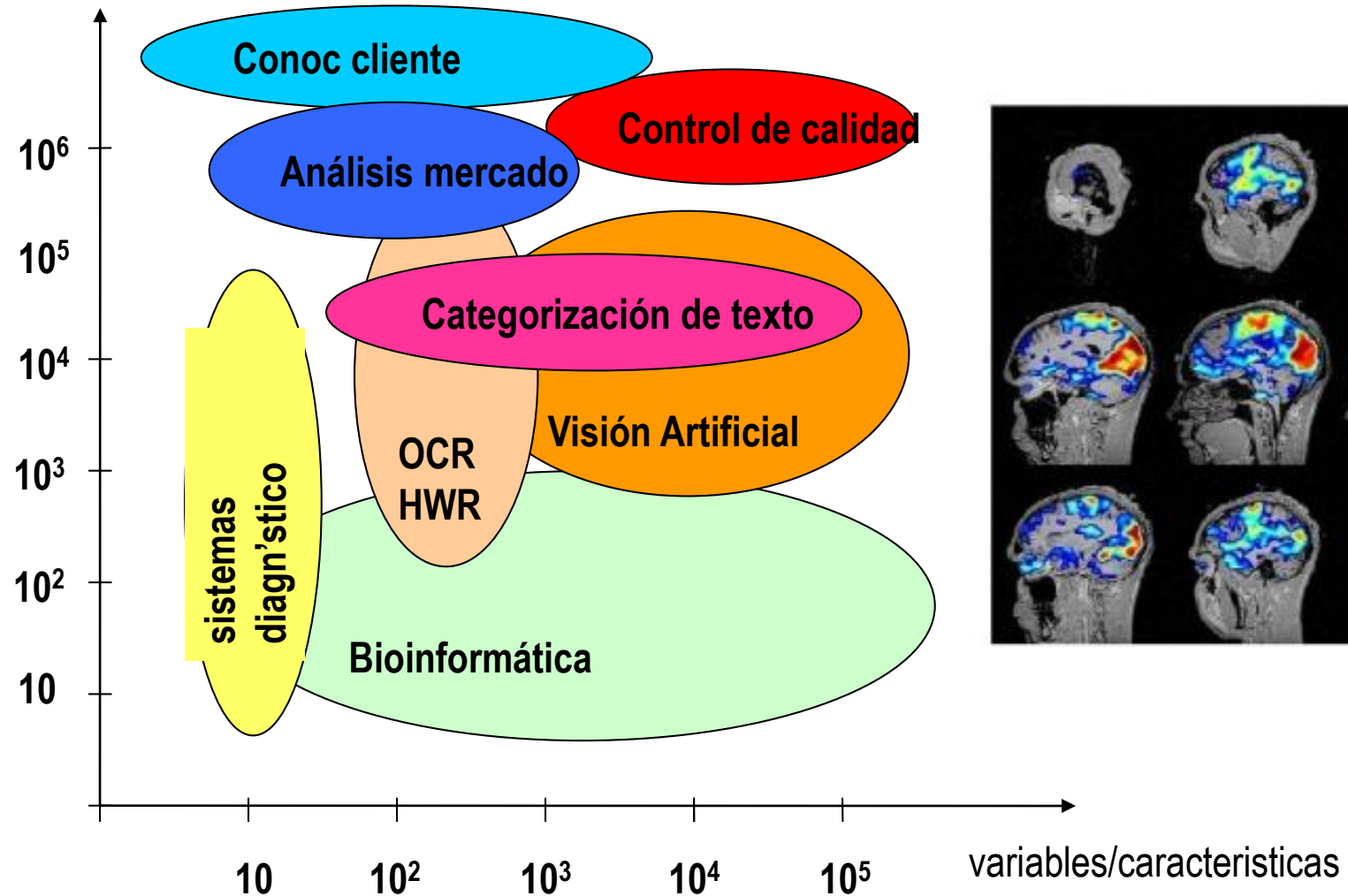
Sequence 2 : E1, E2, E5, E6, E4

Sequence 3 : E1, E7, E8, E3, E4

An event E_{ij} is similar to the succession of the two events E_{ik} E_{kj} .

Reducción y selección de características:

Ejemplos



Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características:

este es un paso opcional que permite seleccionar o reducir la cantidad de características extraídas.

- **La reducción de características** consiste en crear nuevos atributos utilizando los originales, mientras que
- **La selección de características** consiste en encontrar un conjunto bajo de atributos que describa mejor un proceso.
- Estos pasos apuntan a disminuir problemas como el consumo de tiempo, la maldición de la dimensionalidad y el aumento de la precisión, el desbalance de muestra por clases, entre otros.

Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características:

Las razones para realizar FS pueden incluir:

- Eliminar datos irrelevantes.
- Aumentar la precisión predictiva de los modelos aprendidos.
- Reducir el costo de los datos.
- Mejorar la eficiencia del aprendizaje, como reducir los requisitos de almacenamiento y el costo computacional.
- Reducir la complejidad de la descripción del modelo resultante, mejorando la comprensión de los datos y el modelo.

Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características:

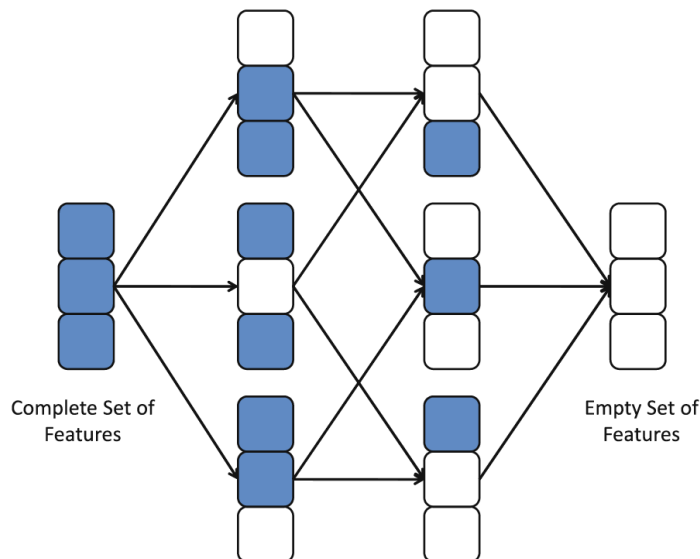
- . Buscar el mejor subconjunto de características.
- . Criterios para evaluar diferentes subconjuntos.
- . Principio para seleccionar, agregar, eliminar o cambiar nuevas características durante la búsqueda.

Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características:

Buscar el mejor subconjunto de características.

puede ser considerado como un problema de búsqueda



Ejemplos direcciones de búsqueda:

- **Generación progresiva secuencial (SFG):** comienza con un conjunto vacío de características S . Cuando comienza la búsqueda, las características se agregan a S de acuerdo con algún criterio que distinga la mejor característica de las demás.
- **Generación regresiva secuencial (SBG):** comienza con un conjunto completo de características y, iterativamente, se eliminan de a una por vez.

Selección de características:

- **Métodos**
 - **Univariante** (mira cada característica independientemente de los demás)
 - Coeficiente de correlación de Pearson, Puntuación F, Chi-cuadrado
 - **Multivariable** (considera todas las características simultáneamente)
 - Algoritmos de reducción de dimensionalidad
 - Clasificadores lineales como la máquina de vectores de soporte
 - Eliminación de características recursivas

Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características:

Criterios para evaluar diferentes subconjuntos.

- Medidas de información.

La información sirve para medir la incertidumbre del receptor cuando recibe un mensaje.

Entropía de Shannon:

$$-\sum_i P(c_i) \log_2 P(c_i).$$

- Medidas de distancia.
- Medidas de dependencia.
Pearson correlation coefficient
- Medidas de consistencia.
- Medidas de precisión

$$\rho(X, Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2]^{\frac{1}{2}}}$$

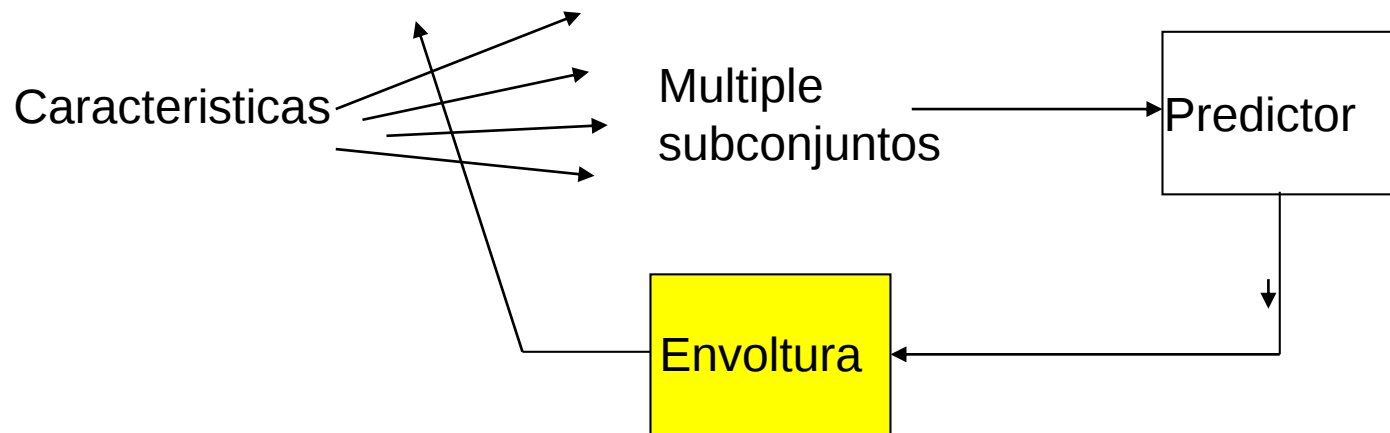
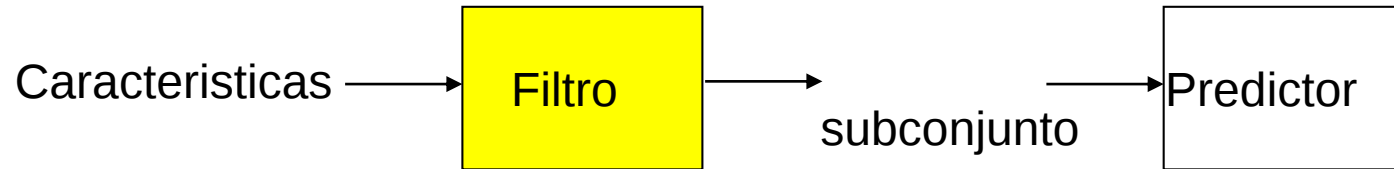
Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características:

Los Multivariables comúnmente se dividen en enfoques de filtro, envoltura e incrustados, que a su vez pueden desarrollarse mediante estrategias supervisadas y no supervisadas.

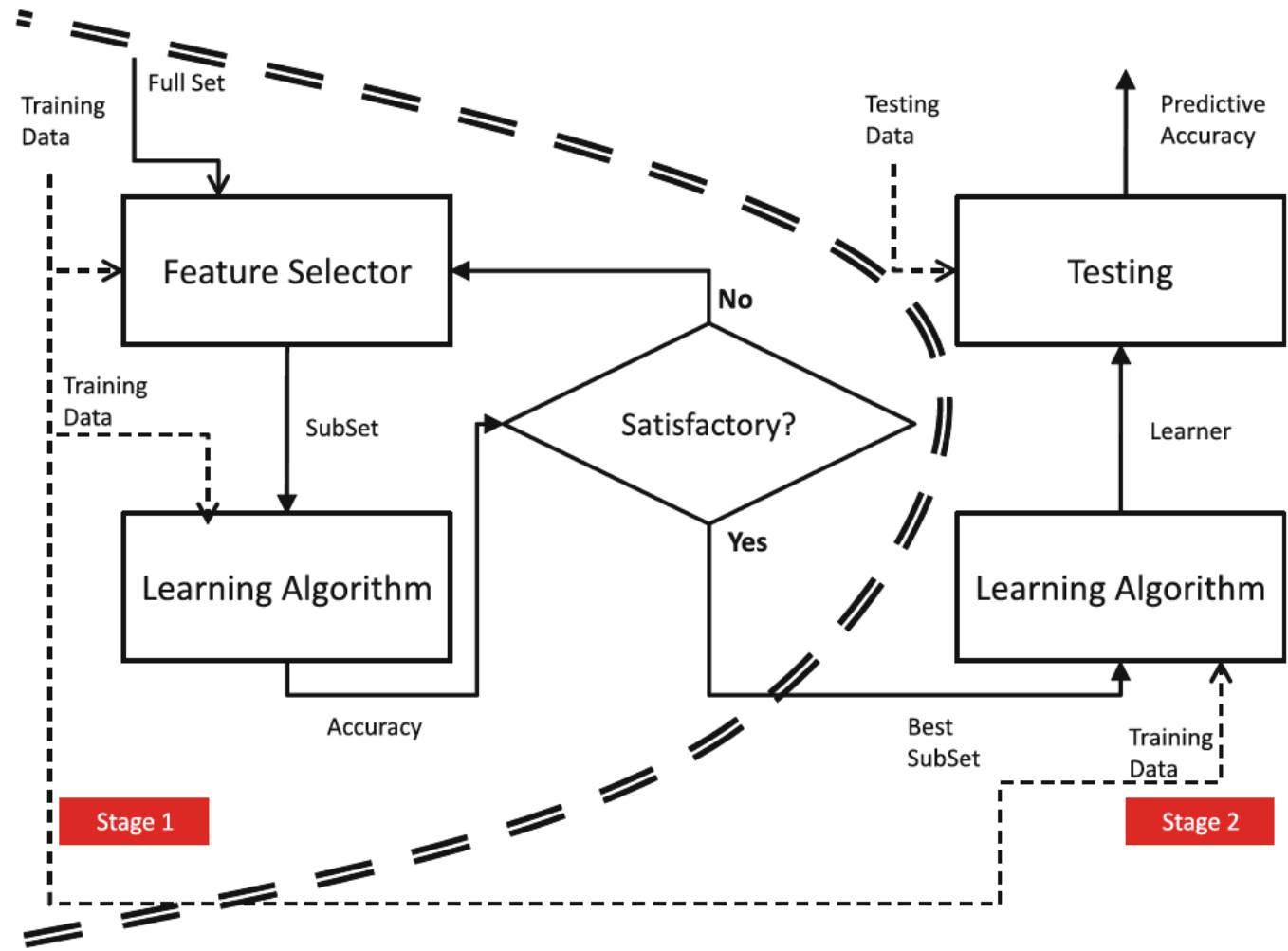
- **Los enfoques de filtro** dan un puntaje a cada característica usando una métrica, con el fin de ejecutar un proceso de clasificación, donde las características más relevantes cumplen con un umbral aceptado.
- **Los enfoques de envoltorio** pueden usar aprendizaje supervisado para buscar relaciones entre las características. Se define una función objetivo para determinar el efecto de diferentes conjuntos de características sobre la precisión del modelo.
- **El análisis de correlación** es un enfoque de filtro simple donde las relaciones entre las características se detectan a través de un coeficiente de correlación.

Filtros, Envolturas



Reducción y selección de características

- Envoltorio



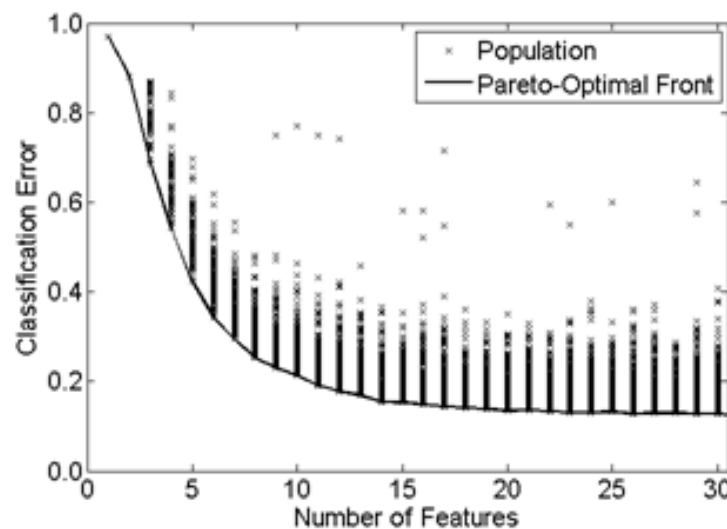
Preparación de los Datos para ser usados por Técnicas Inteligentes

Reducción y selección de características

- **En la estrategia supervisada**, el vector de características tiene una etiqueta de clase.
- **En la estrategia no supervisada**, el vector de características se usa para determinar las relaciones entre las características para descubrir su relevancia. Las estrategias no supervisadas se utilizan para respaldar el enfoque de filtro
- **Los algoritmos genéticos multiobjetivos (NSGA, etc.) y las estrategias de búsqueda secuencial** son ampliamente utilizados como métodos de envoltura.
- Hay métodos integrados de algoritmos ML que incluyen el proceso de selección de características en el diseño de su modelo, como los **modelos de regresión y modelos basados en árboles de decisión**.
- Otras técnicas utilizadas por el **enfoque de filtro** son **Gini Index, Information Gain, Laplacian Score y Sparsity Scores**, entre otros

Non-dominated Sorting Genetic Algorithm

- Use NSGA para buscar un conjunto óptimo de soluciones con dos objetivos:
 - Minimice la cantidad de características utilizadas en la clasificación.
 - Minimice el error de clasificación.

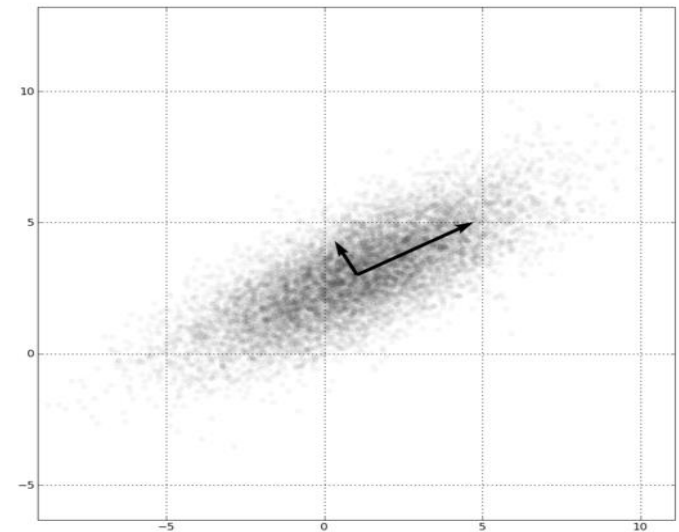


Reducción y selección de características: estadísticas multivariantes

- La estadística multivariada analiza todas las variables al mismo tiempo y encuentra patrones que describen la variabilidad
- Covarianza entre variables se toma en cuenta

- **Varios métodos estadísticos multivariantes**

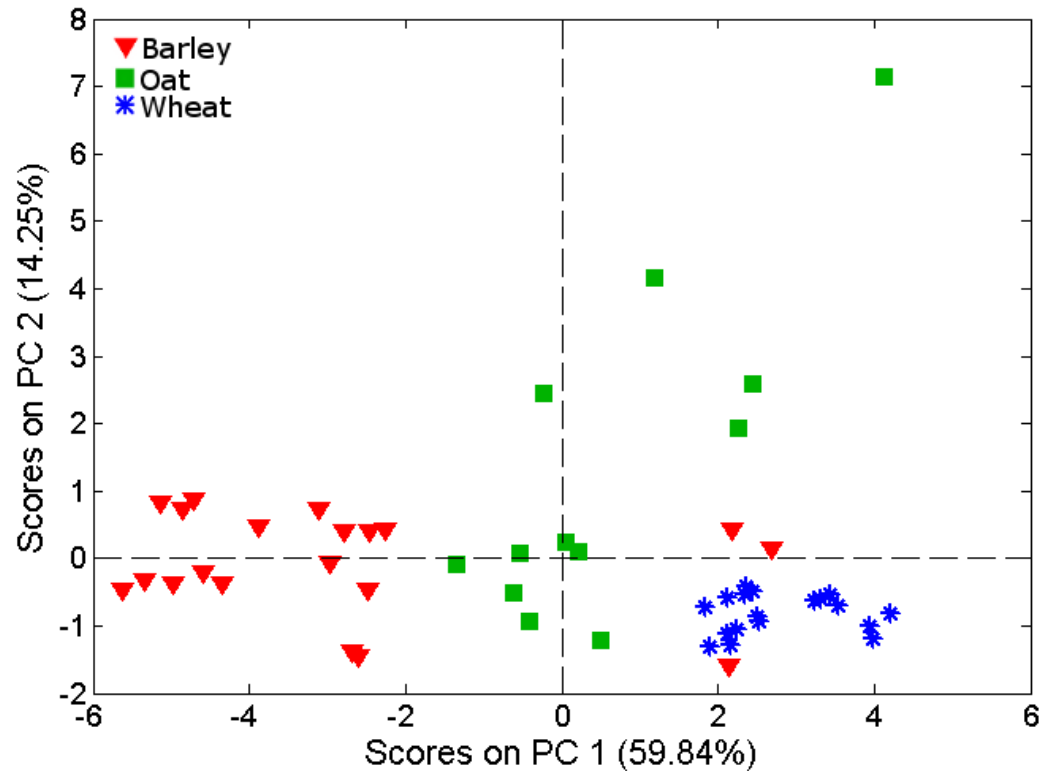
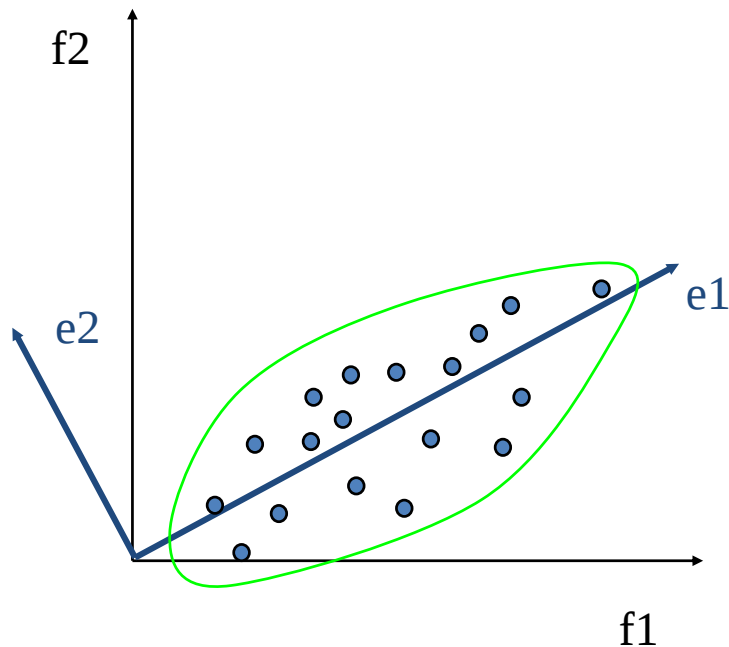
- Análisis de componentes principales (PCA)
- Análisis discriminante de mínimos cuadrados parciales (PLSDA)



Fra: http://en.wikipedia.org/wiki/Principal_component_analysis

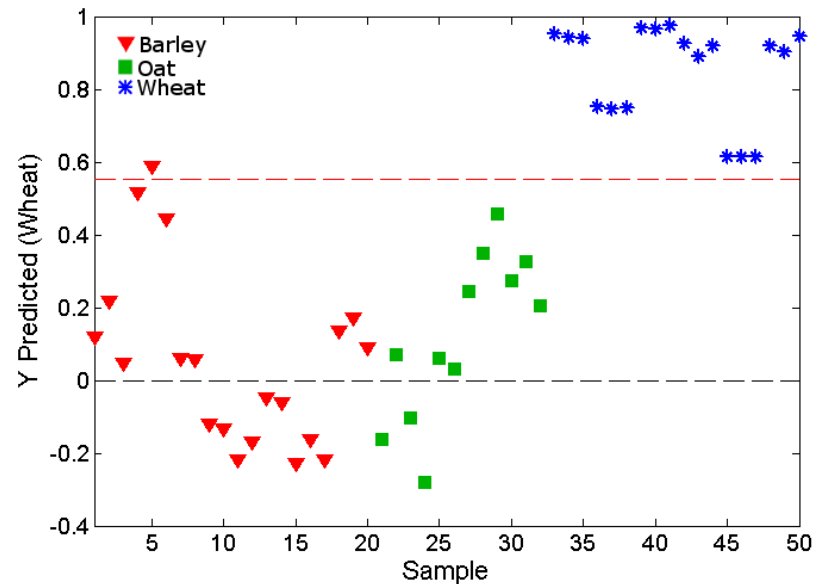
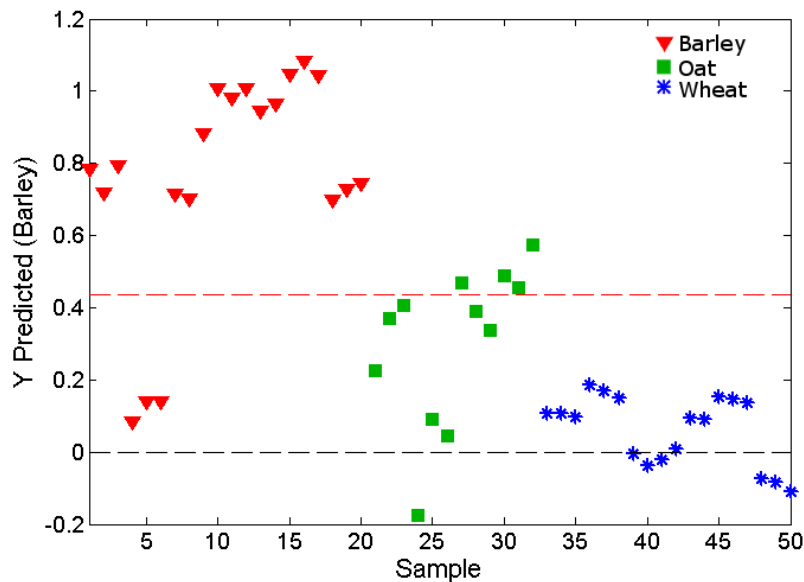
Análisis de componentes principales (PCA)

Intuición: encuentre el eje que muestre la mayor variación y proyecte todos los puntos en este eje



Análisis discriminante de mínimos cuadrados parciales (PLSDA)

- PLS es una técnica de regresión multivariante
- PLSDA permite la discriminación / clasificación de los datos



Preparación de los Datos para ser usados por Técnicas Inteligentes

Selección del algoritmo y construcción del modelo:

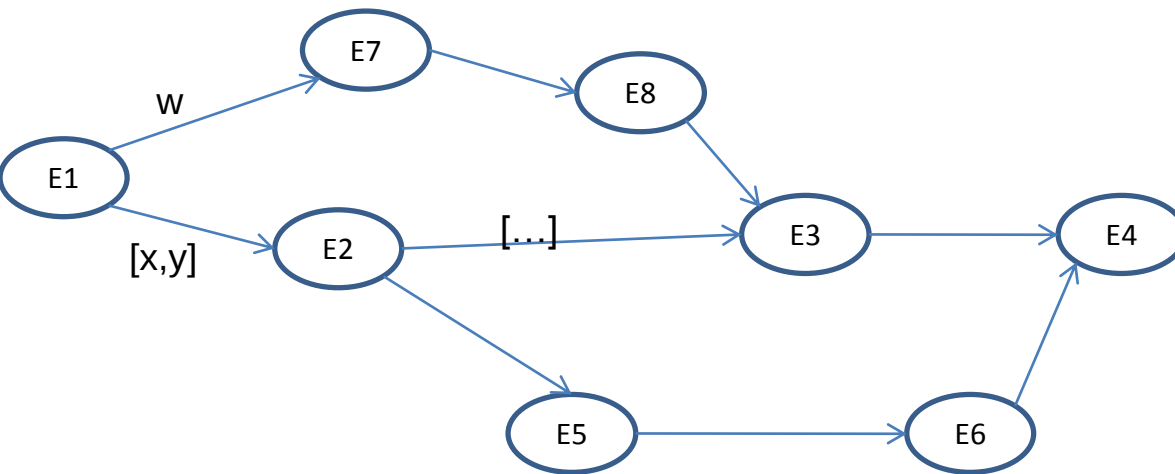
- El resultado de las fases previas conduce a un conjunto de datos que contiene información histórica sobre el estudio de caso, y también a la definición de la tarea y algoritmo ML.
- El conjunto de datos históricos es un recurso clave para construir modelos ML.
- La selección del algoritmo ML está relacionada con el problema a resolver o el tipo de conocimiento que los profesionales quieren descubrir.

Preparación de los Datos para ser usados por Técnicas Inteligentes

Selección del algoritmo y construcción del modelo:

- **La mayoría de los algoritmos de aprendizaje supervisado** ajustan los parámetros del modelo **minimizando el error entre la salida del modelo y la salida real esperada** para una entrada. Esto significa que los datos históricos deben estar etiquetados.
- **Los algoritmos no supervisados** intentan **encontrar relaciones entre las entradas sin conocimiento previo de las salidas**. Estas relaciones pueden ser similitudes, proximidades y relaciones estadísticas, entre otros.
- Como una consecuencia del aprendizaje, **los algoritmos supervisados se utilizan comúnmente para realizar tareas de clasificación**, mientras que los **no supervisados se utilizan para agrupar las entradas** con el fin de encontrar **comportamientos anómalos o similares entre ellos**.

Specifying temporal patterns (chronicles)



1. *Chronicle structure* based on four particular patterns:

- Sequence pattern
- Fork pattern
- Join pattern
- OR pattern

2. *Dynamical aspects*: the temporal constraints

➤ sequence 1:

E1	T= 0.89
E2	T= 1.22

Temporal Relationship between E1 and E2:

➤ sequence 2:

E1	T= 3.23	[0.33, 0.80]
E2	T= 4.03	