KNOWLEDGE DISCOVERY TECHNIQUES TO IMPROVE THE SERVICES OF INTERNET ECONOMICS

Igor Ruiz Agúndez

Dirigida por Dr. Pablo García Bringas y
Dr. Yoseba Koldobika Peña Landaburu

Índice

- Conceptos básicos
- Estado de la discusión científica
- Modelo óptimo de agrupamiento
- Evaluación con un caso de uso real
- Conclusiones

Índice

- Conceptos básicos
- Estado de la discusión científica
- Modelo óptimo de agrupamiento
- Evaluación con un caso de uso real
- Conclusiones

Los servicios son un conjunto de actividades que buscan responder a las necesidades de un cliente

Nos referimos a los servicios ofrecidos en Internet

Los servicios son ofrecidos en un sistema económico de Internet

Los servicios se sustentan en los sistemas de soporte que se encargan de su producción, asignación y distribución

Los sistemas de soporte

- Gestionan la infraestructura
- Prestan el servicio
- Garantizan la disponibilidad y la calidad
- Ofrecen atención al cliente

Las ciencias económicas de Internet estudian los servicios que los usuarios consumen

El objetivo es gestionar los servicios de la manera más eficaz y eficiente posible

Para mejorar los servicios nos centramos en los datos que se generan en su uso

Analizamos estos datos con técnicas de extracción de conocimiento

Estas técnicas buscan patrones en grandes conjuntos de datos

De estos patrones puede extraerse conocimiento

Hipótesis

"Es posible desarrollar un método que extraiga conocimiento de los datos de los servicios de las ciencias económicas de Internet. Este conocimiento puede representarse como modelos que nos ayuden a mejorar los sistemas de soporte de estos servicios"

Índice

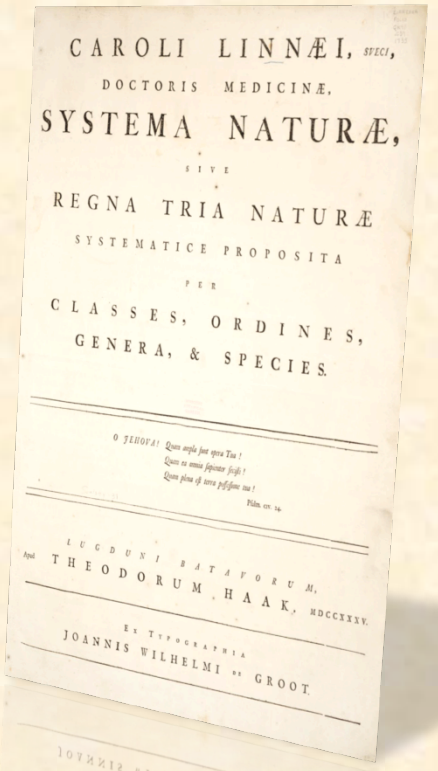
- Conceptos básicos
- Estado de la discusión científica
- Modelo óptimo de agrupamiento
- Evaluación con un caso de uso real
- Conclusiones

Actores del sistema

Taxonomía

- Ciencia de la clasificación
- Aplicada a múltiples campos
- Sirven como contenedores de información y permiten hacer predicciones

Ha sido utilizada para organizar el proceso de consumo de un servicio



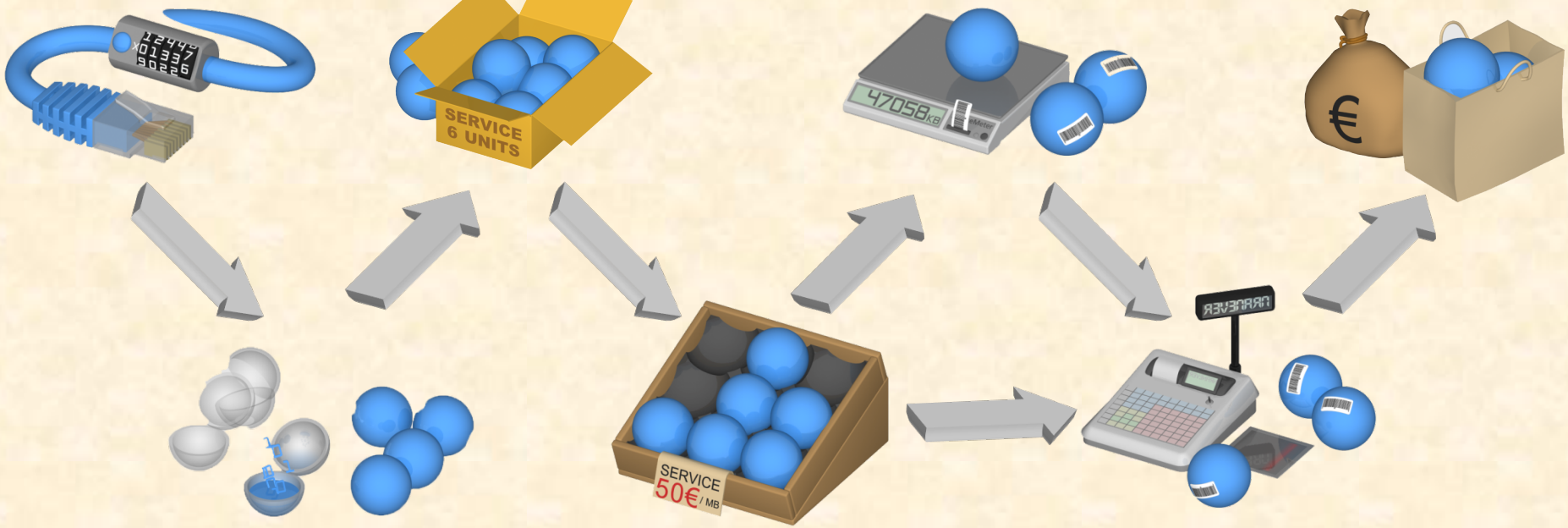
Proceso de consumo de un servicio

Medición

Tarificación

Cálculo de cargas

Compensación financiera

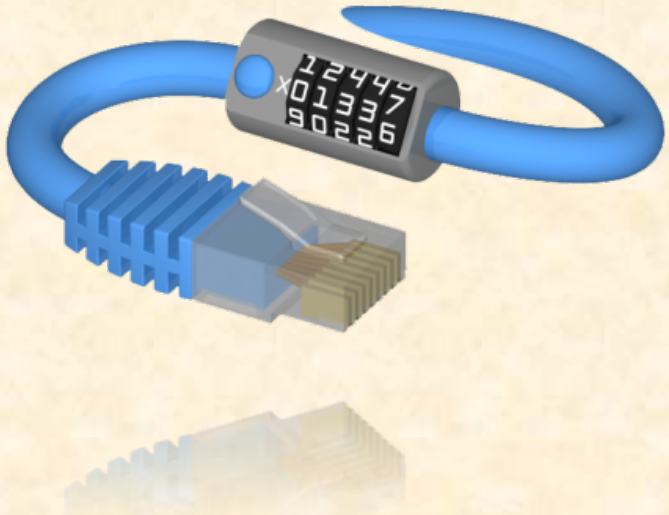


Mediación

Fijación de precio

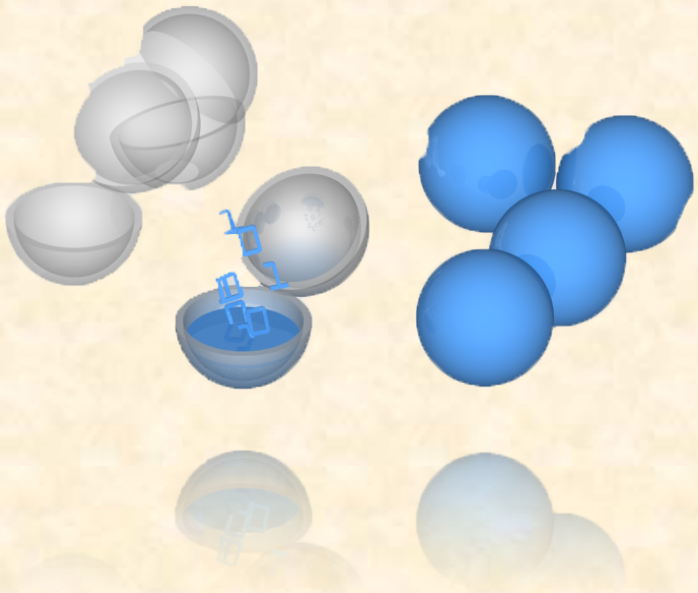
Facturación

Medición



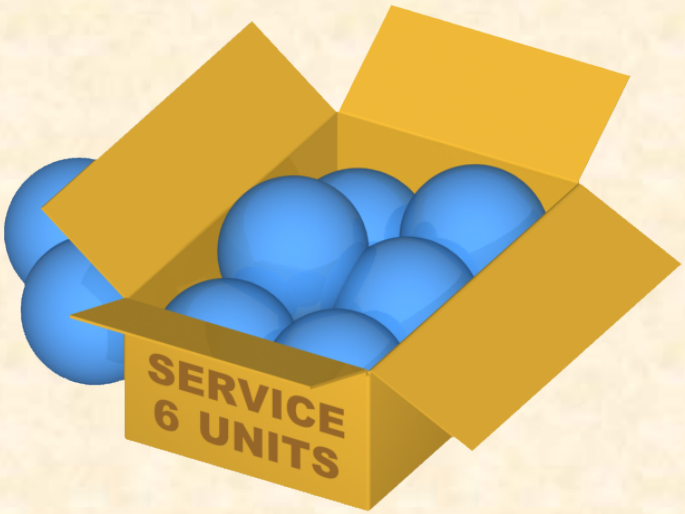
Recoge la información correspondiente al consumo de un recurso

Mediación



Transforma los datos en bruto en registros para su procesado

Tarificación

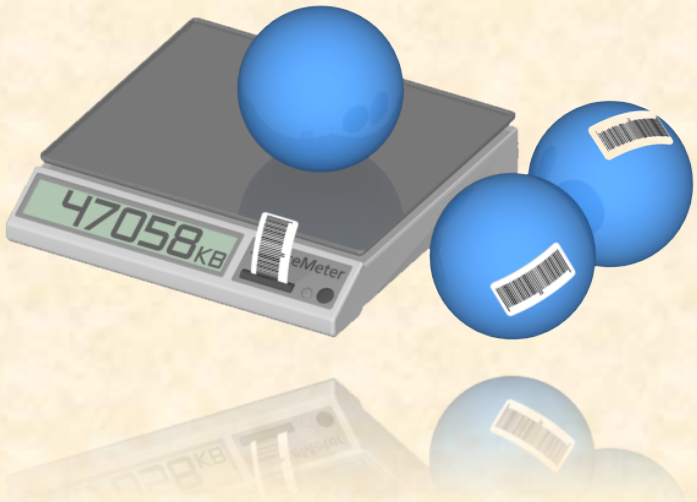


Prepara los registros de sesión de uso del servicio

Fijación de precio

Determina el precio que implica el uso de un recurso

Cálculo de cargas



Traduce en un precio monetario los registros de sesión

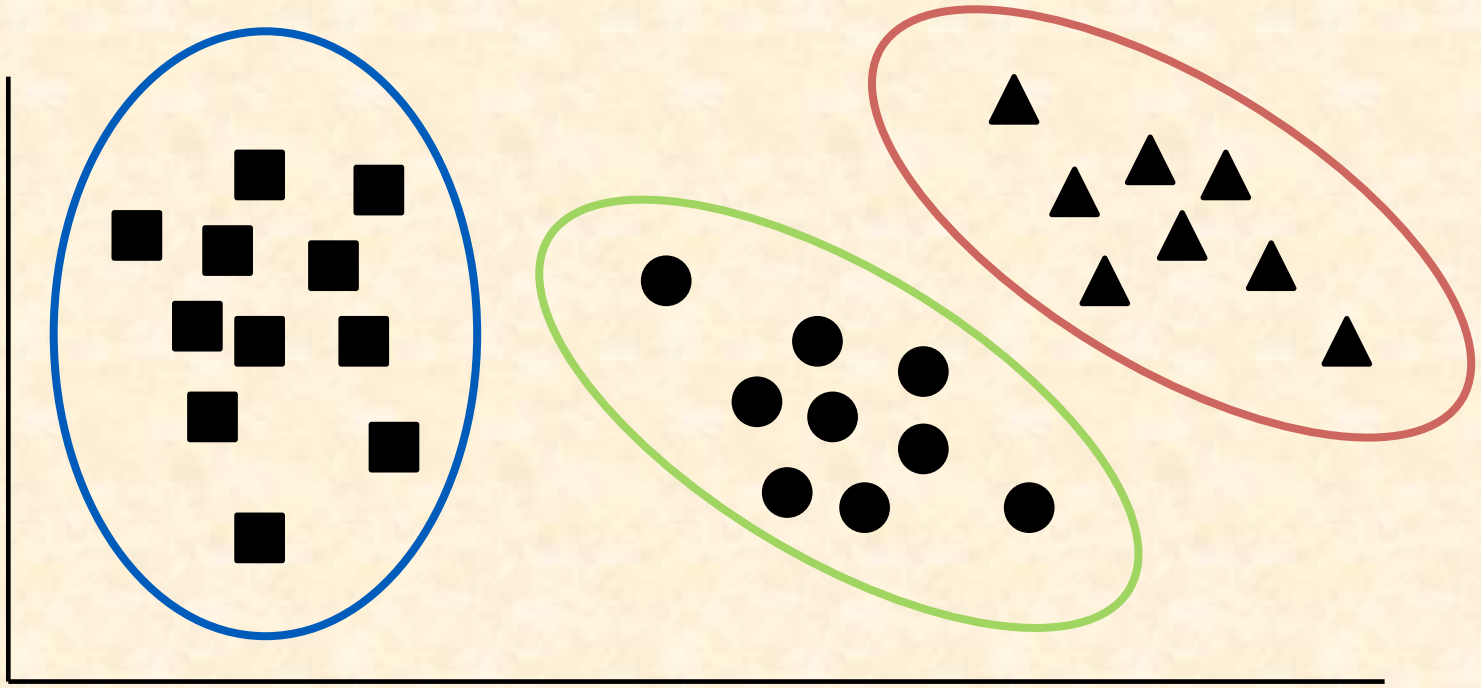
Facturación

Genera la factura final utilizando los registros de sesión y su precio monetario

Compensación financiera

Gestiona las transacciones económicas entre los participantes

Los algoritmos de agrupamiento son técnicas estadísticas de análisis de datos usadas en muchas áreas



Algoritmos de agrupamiento

Generan modelos que contienen:

- Distribución de las instancias en grupos
- Información generada por el algoritmo utilizado

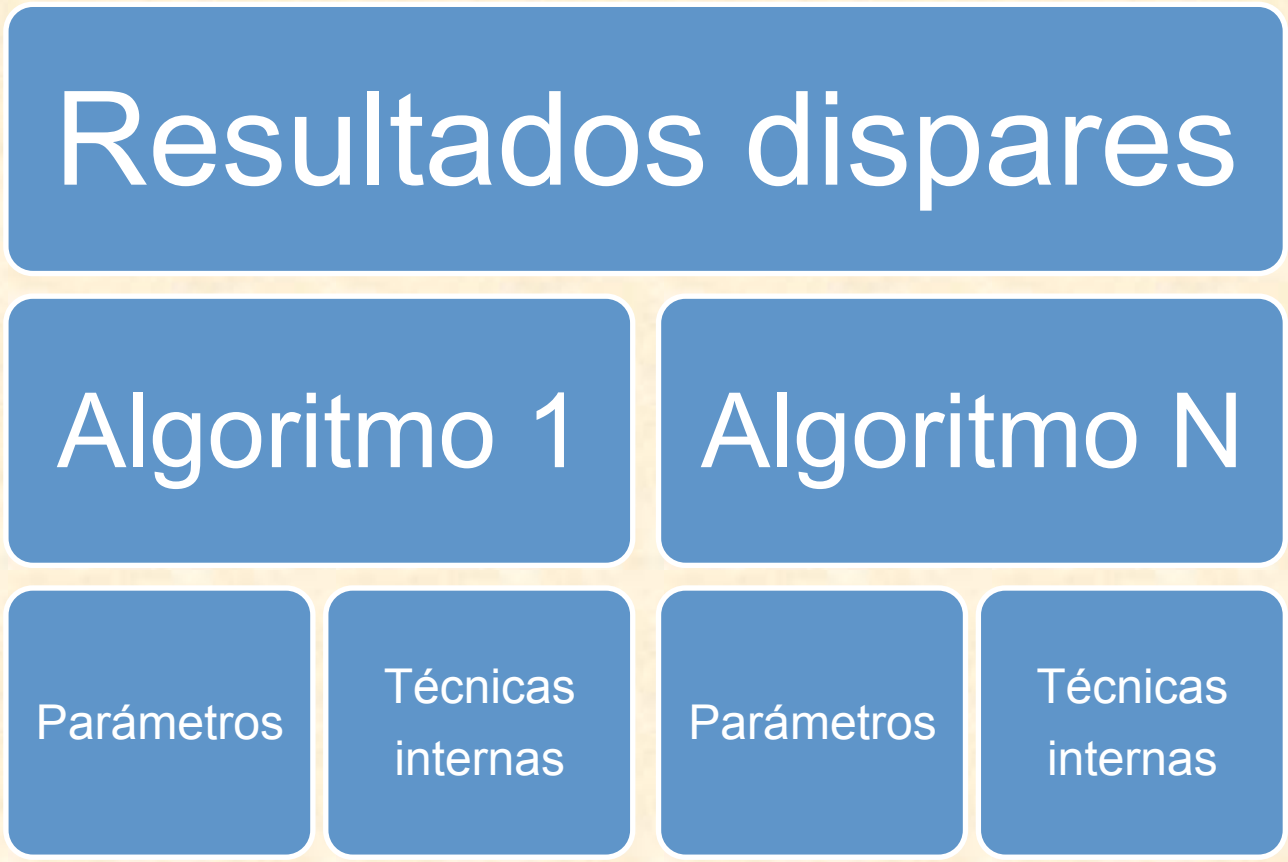
Existen multitud de enfoques y clasificaciones para organizar los diferentes algoritmos de agrupamiento

Cada algoritmo de agrupamiento tiene distintos:

- Parámetros de entrada-salida
- Técnicas internas de funcionamiento

Se generan resultados significativamente distintos

Necesitamos comparar



Es necesario validar

- ¿Cuál es el resultado correcto?
- ¿Cuál es el número correcto de grupos?
- ¿Son los grupos representativos?

Actualmente únicamente se puede determinar parcialmente si un resultado es correcto o no

Índice

- Conceptos básicos
- Estado de la discusión científica
- Modelo óptimo de agrupamiento
- Evaluación con un caso de uso real
- Conclusiones

Motivación

- Las metodologías tradicionales tienen ciertas deficiencias al validar y comparar modelos

- Elaboramos una metodología que extrae el modelo óptimo de agrupamiento analizando los datos del servicio

Este modelo puede utilizarse para mejorar los sistemas de soporte asociados al servicio

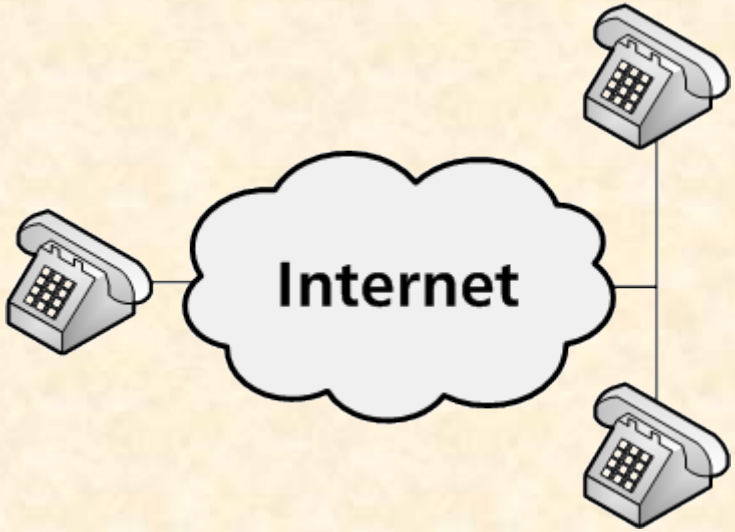
Metodología

1. • Agrupamiento
2. • Selección del modelo óptimo
3. • Operación con el modelo

Índice

- Conceptos básicos
- Estado de la discusión científica
- Modelo óptimo de agrupamiento
- Evaluación con un caso de uso real
- Conclusiones

Voz sobre IP



Permite realizar llamadas, teleconferencias, y otros usos

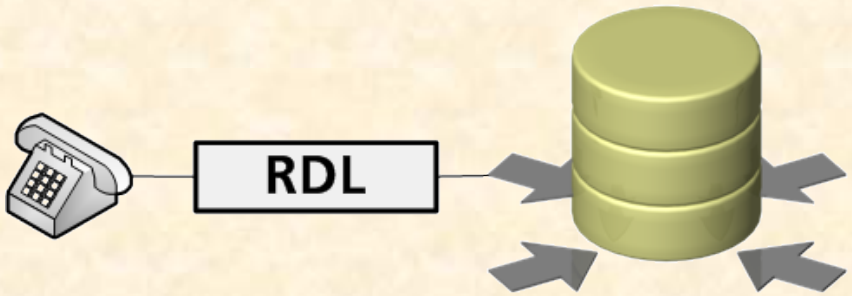
La aplicación de la metodología permite potencialmente mejorar múltiples aspectos de los sistemas de soporte de la Voz sobre IP

Podrían identificarse:

- Sumideros y nichos de llamadas
- Tendencias en las llamadas

Podría mejorarse:

- Las congestiones
- Estrategias de marketing
- Planificación de red



El conjunto de datos está formado por los Registros Detallados de Llamada (RDL) de los usuarios

Metodología

1. Agrupamiento
2. Selección del modelo óptimo
3. Operación con el modelo

Conjunto de datos

Atributos del conjunto de datos (I)

Atributo	Descripción
calldate	Cuando se registra un RDL
clid	Identificador del emisor en texto
src	Identificador del emisor en número
dst	Extensión de destino
dcontext	Contexto de destino
channel	Canal utilizado
dstchannel	Canal de destino
lastapp	Última aplicación interna utilizada

Atributos del conjunto de datos (II)

Atributo	Descripción
lastdata	Argumentos de la última aplicación interna utilizada
duration	Tiempo entre marcado y colgado
billsec	Tiempo entre descolgado y colgado
disposition	Resolución de la llamada
amaflags	Indicadores utilizados
accountcode	Número de tarificación
user field	Campo definido por el usuario

Los atributos tienen

- Un tipo de datos que define la naturaleza de los datos
- Una granularidad en la que pueden dividirse
- Unas relaciones entre atributos
- Una posible discretización
- Una relevancia en el problema

Conjunto de datos preprocesado (I)

Atributo	Descripción
year	Año en el que se registra un RDL
month	Mes en el que se registra un RDL
day	Día en el que se registra un RDL
hour	Hora en el que se registra un RDL
minute	Minuto en el que se registra un RDL
second	Segundo en el que se registra un RDL
src	Identificador del emisor en número
dst	Extensión de destino
dcontext	Contexto de destino

Conjunto de datos preprocesado (II)

Atributo	Descripción
channel	Canal utilizado
dstchannel	Canal de destino
lastapp	Última aplicación interna utilizada
lastdata	Argumentos de la última aplicación interna utilizada
duration	Tiempo entre marcado y colgado
billsec	Tiempo entre descolgado y colgado
answerTime	Creado con duration - billsec
disposition	Resolución de la llamada

Algoritmos de agrupamiento utilizados

Cobweb
DBSCAN
EM
FF
OPTICS
Simple K-Means

Cobweb

Realiza agrupamiento incremental jerárquico

Usa agrupamiento conceptual

Orientado al rendimiento

DBSCAN

Busca el número de grupos estimando la densidad de la distribución

Agrupa las instancias de acuerdo a sus relaciones

EM

Asocia una probabilidad de pertenencia de grupo a cada instancia

Trabaja de forma iterativa alternando dos pasos

FF

Comenzando en cualquier instancia mide la distancia más lejana con otra instancia.

Después la más lejana a ambas y así hasta analizar todas las instancias

OPTICS

Busca grupos utilizando la densidad del conjunto de datos

Procesa las instancias linealmente ordenando los datos espacialmente

Simple K-Means

Divide el conjunto de datos en N grupos

Las instancias son asignadas al grupo más cercano

Parámetros de configuración

Los valores de configuración seleccionados son los más habituales y se han elegido teniendo en cuenta las particularidades del problema

Cada combinación de:

- conjunto de datos
- algoritmo de agrupamiento
- parámetros de configuración

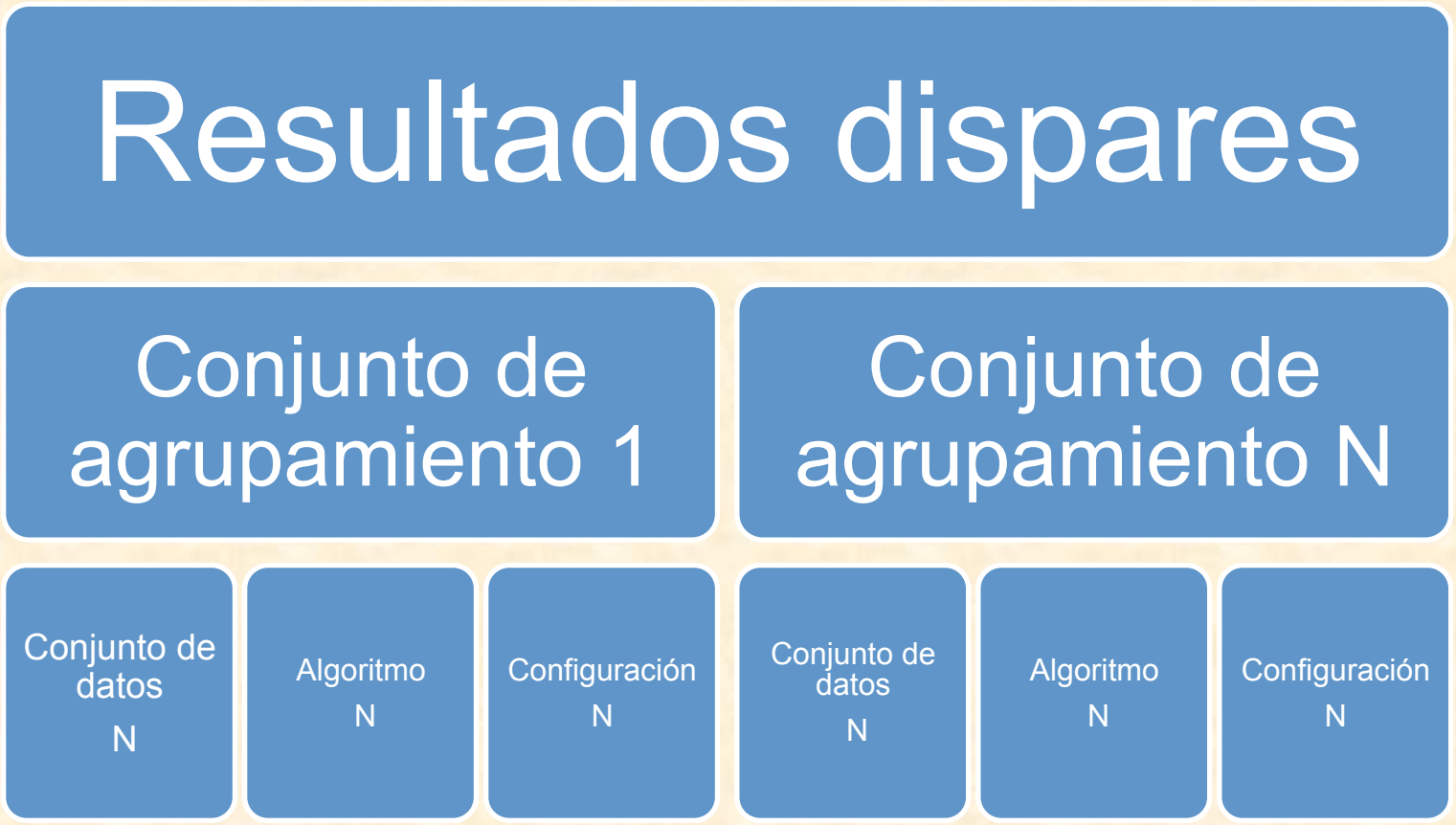
Genera un conjunto de agrupamiento o modelo

Conjunto de agrupamiento

Genera un conjunto de agrupamiento formado por:

- un modelo
- información de ejecución
- las asignaciones de grupo

Necesitamos comparar



Metodología

1. • Agrupamiento
2. • Selección del modelo óptimo
3. • Operación con el modelo

Atributo base y métrica

El único atributo común en los algoritmos utilizados es la asignación de grupo a cada una de las instancias

Este atributo será utilizado como base de la métrica

Restricciones

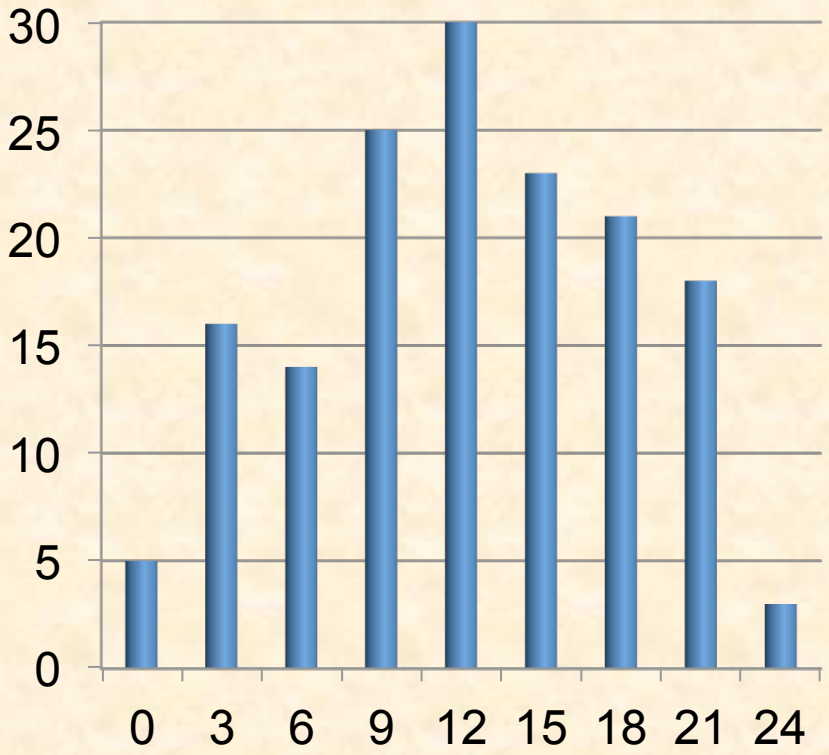
Las restricciones aplicadas purgan los resultados de los algoritmos Cobweb, DBScan y OPTICS por no ofrecer resultados relevantes

Criterio

El modelo óptimo debería tener:

- Un gran grupo representando el comportamiento normal
- Múltiples grupos más pequeños representando el comportamiento atípico

Curtosis

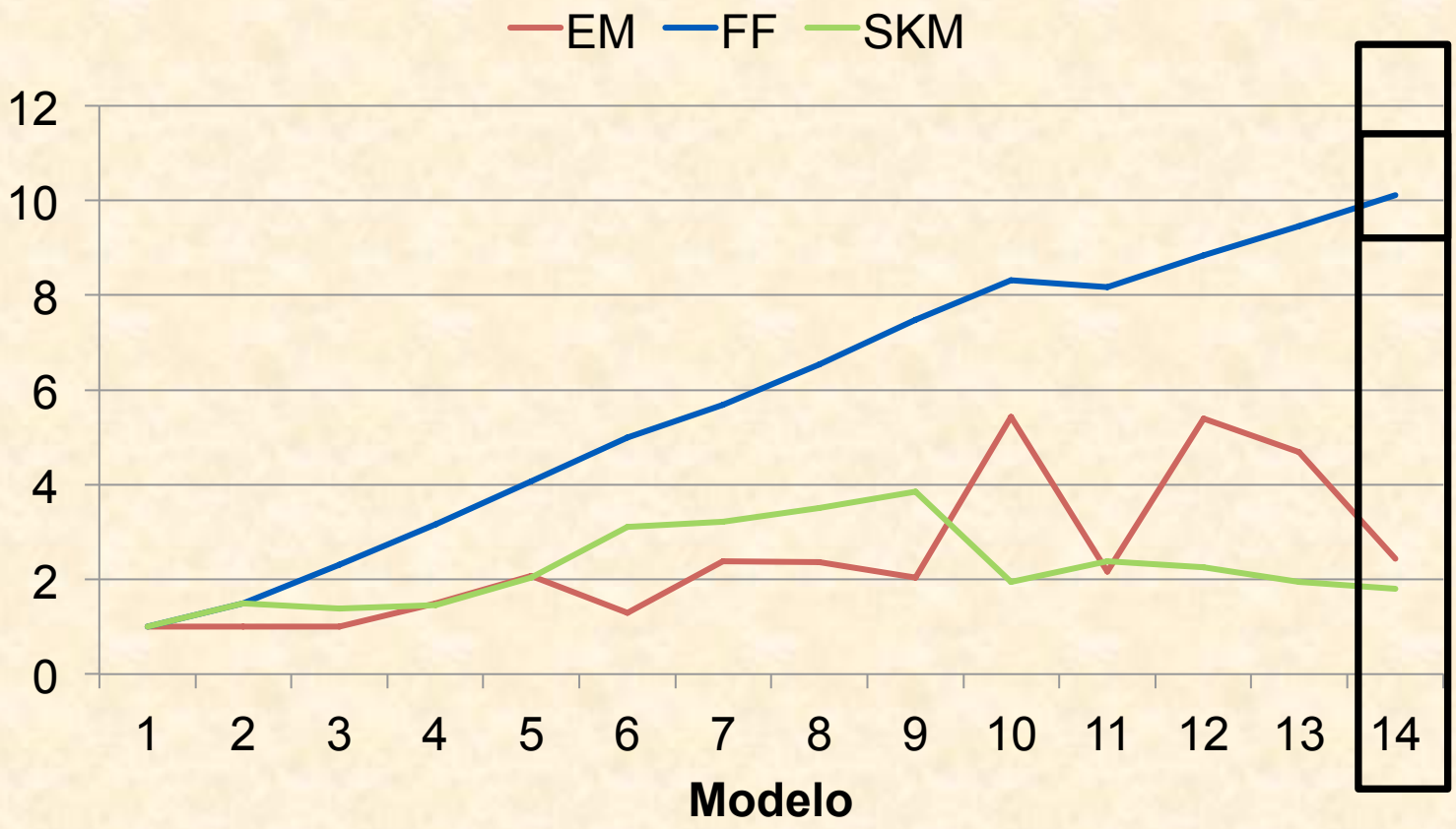


Mide la forma de una distribución estudiando la frecuencia alrededor de la media

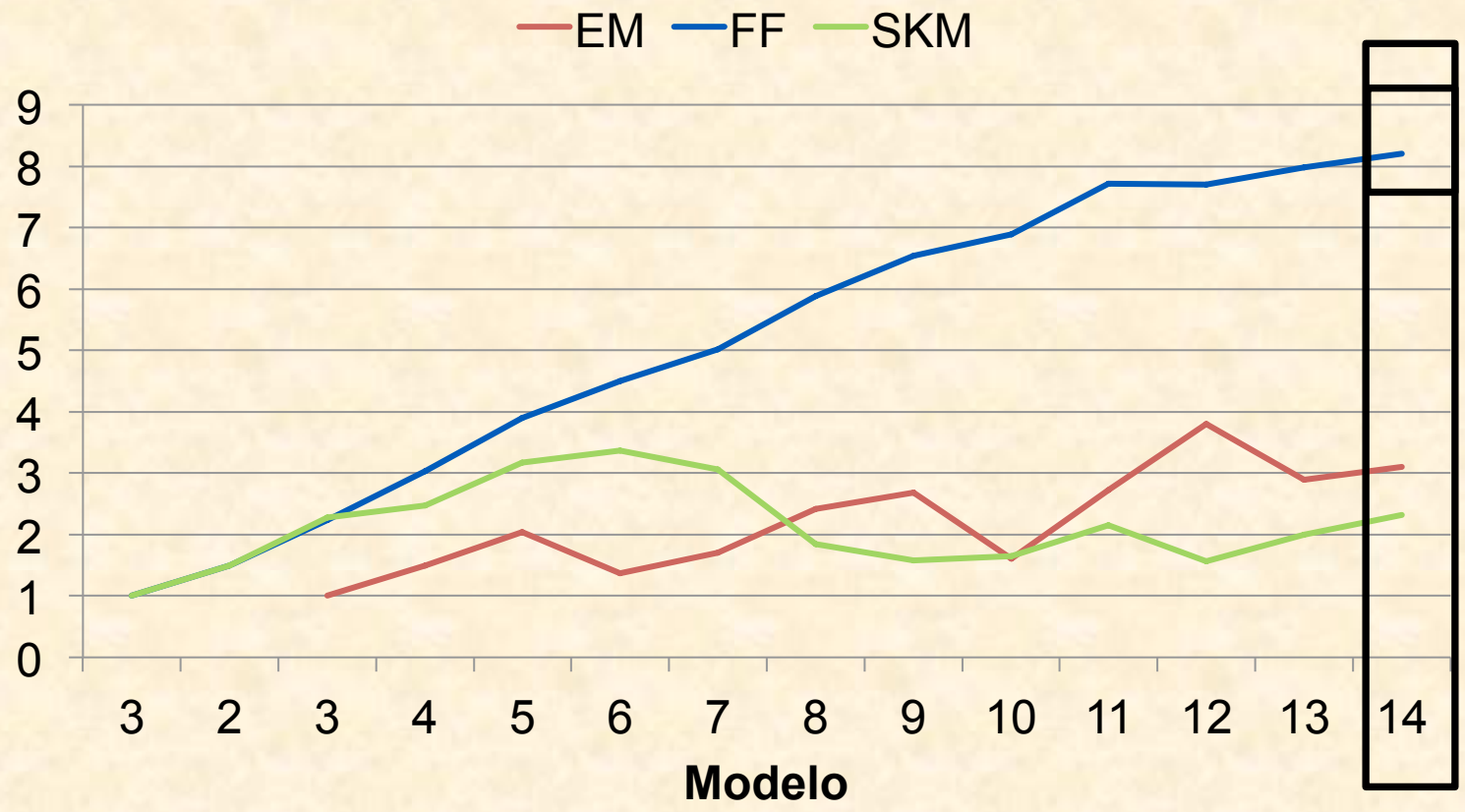
Curtosis

$$K(x) = \frac{\mu_4}{\sigma_4} - 3$$

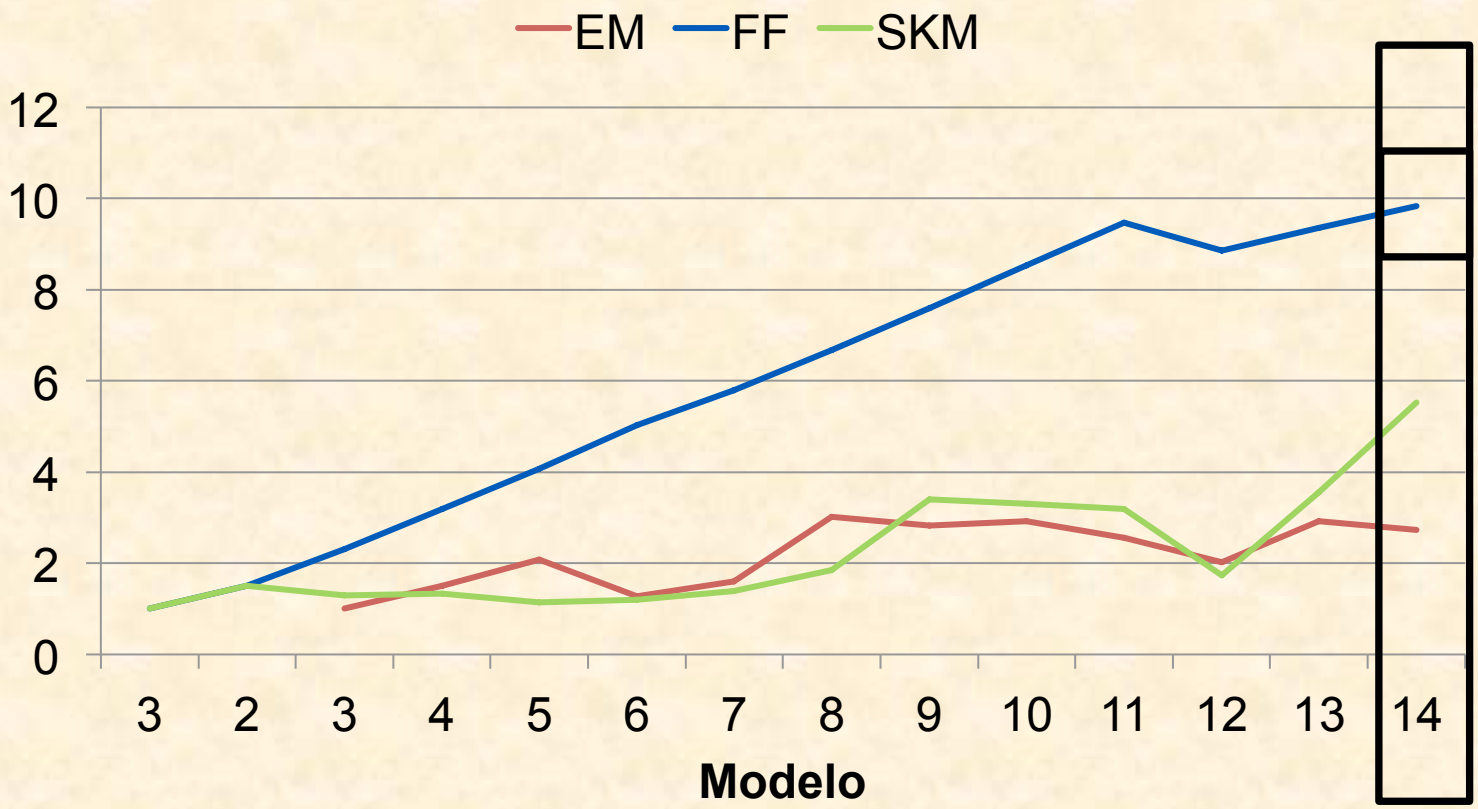
Curtosis en el conjunto de datos
1..31 Octubre



Curtosis en el conjunto de datos
1..7 Octubre



Curtosis en el conjunto de datos
1..7 Noviembre



El modelo óptimo ha sido el 14 FF

Las asignaciones de grupo son distintas

Instancias por grupo en cada conjunto de datos

Realizamos una validación para garantizar que las asignaciones de grupos a las instancias entre los distintos conjuntos de datos están dentro de un umbral aceptable

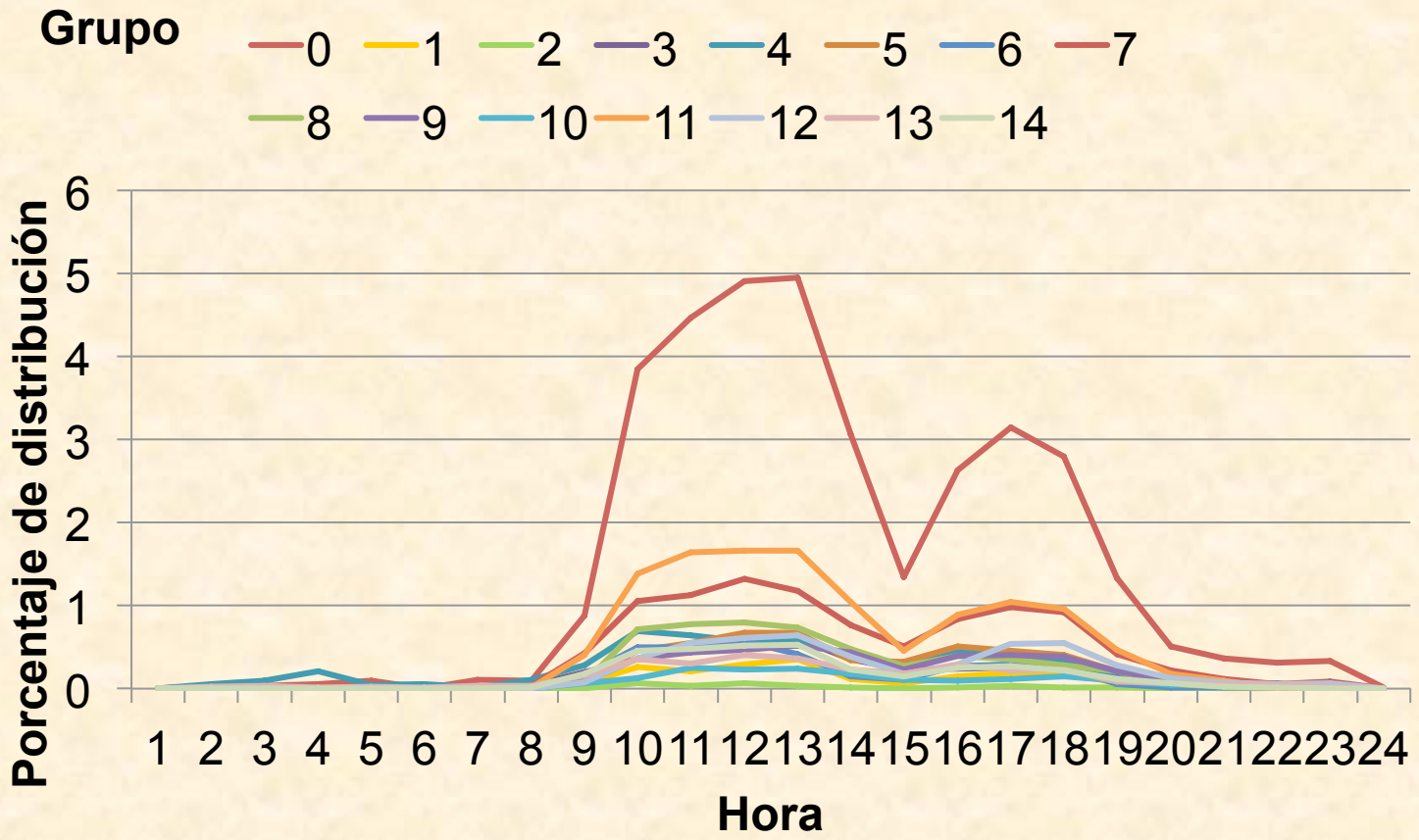
Metodología

1. • Agrupamiento
2. • Selección del modelo óptimo
3. • Operación con el modelo

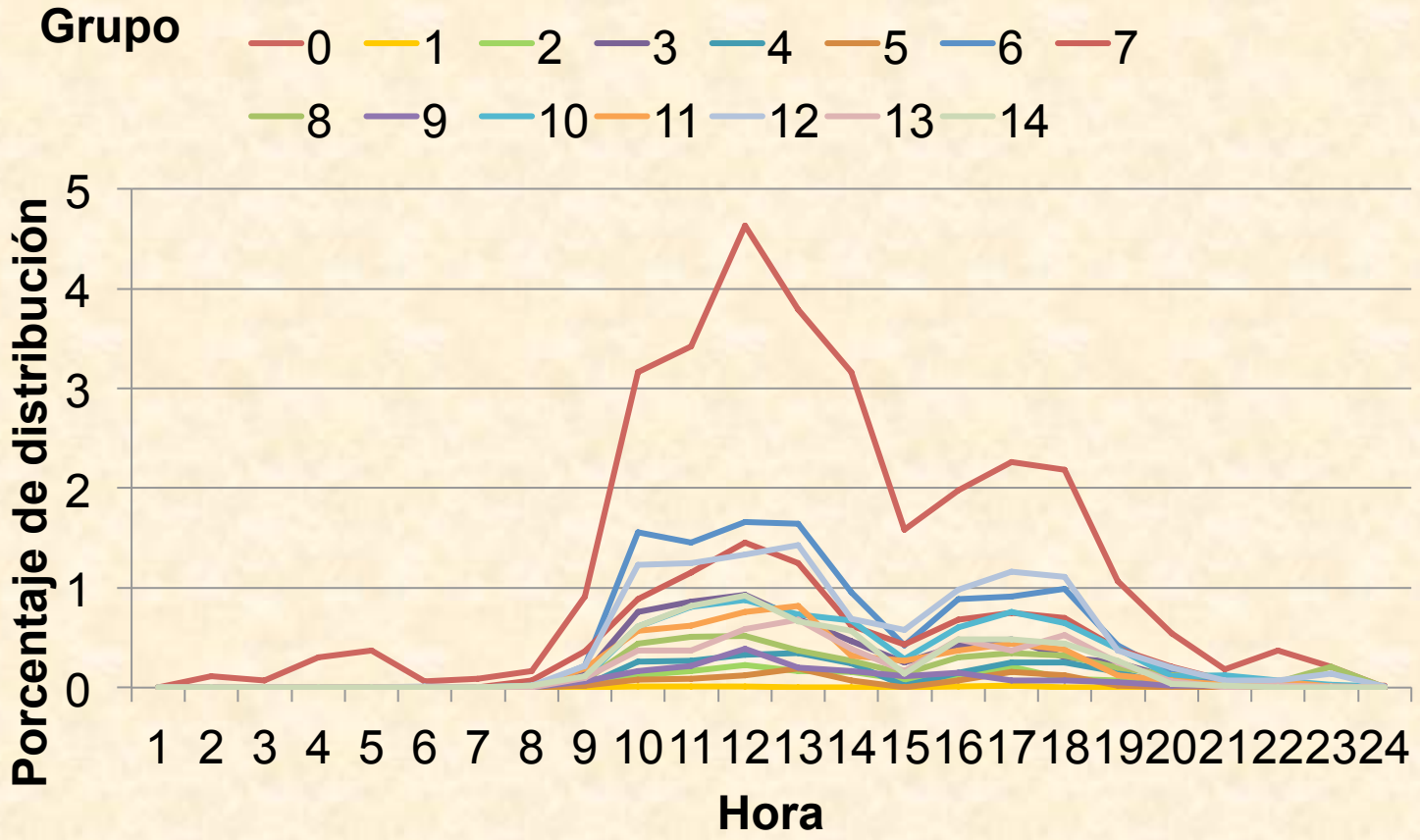
El modelo óptimo del problema corresponde al modelo 14 FF del primer conjunto de datos

Definimos el atributo hora para analizar el modelo en base a la operación de establecimiento de llamada

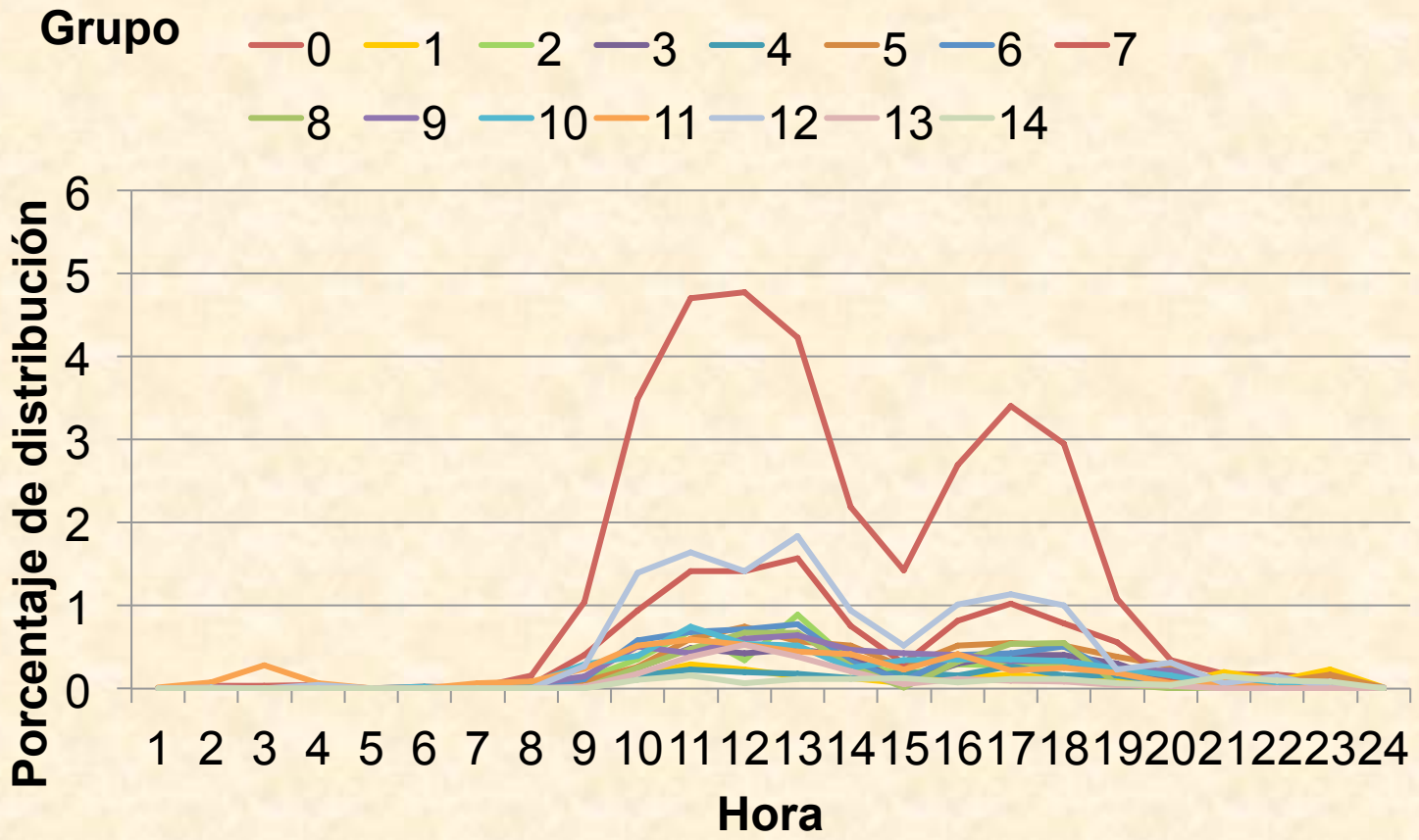
Instancias por grupo para el conjunto de datos 1..31 Octubre



Instancias por grupo para el conjunto de datos 1..7 Octubre

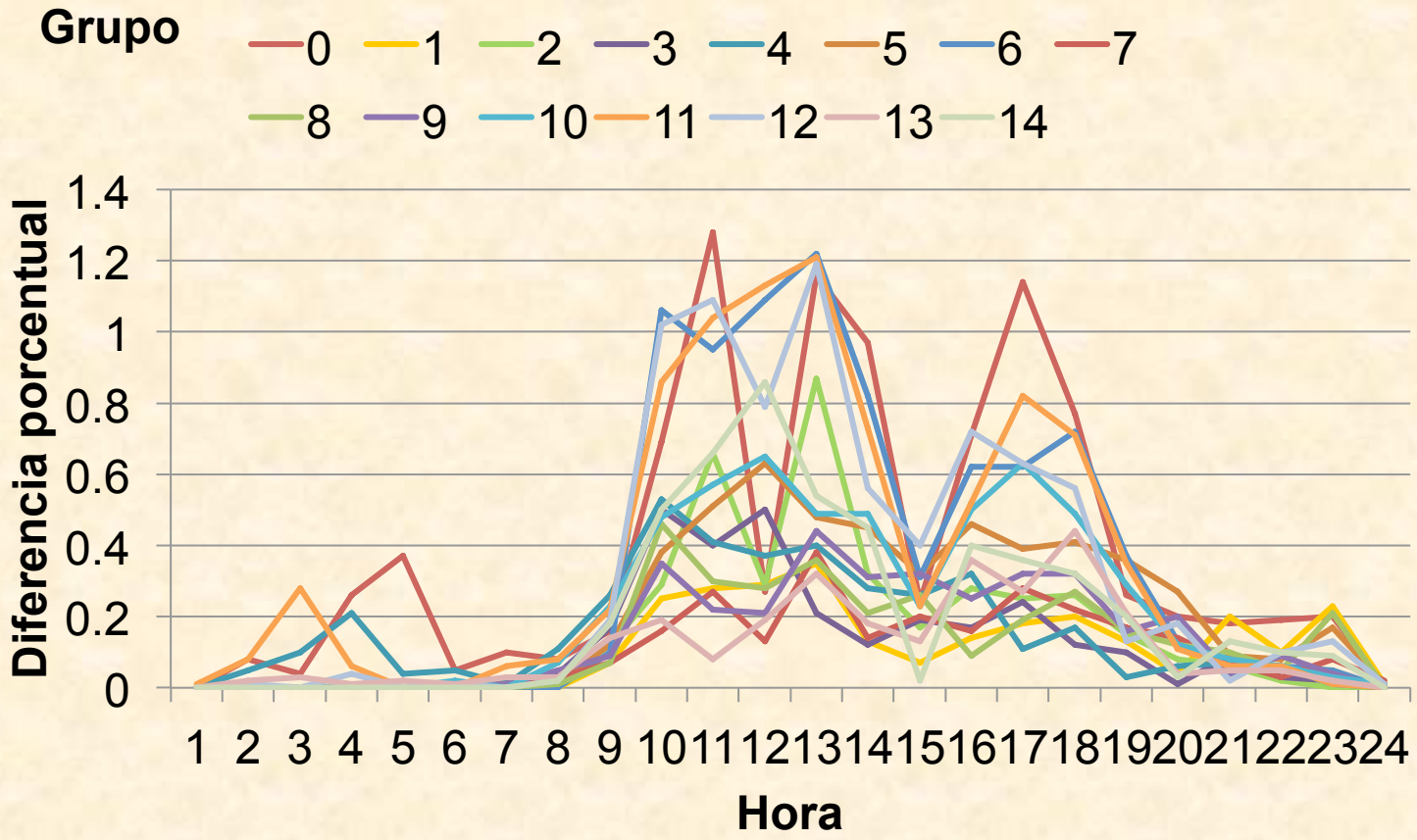


Instancias por grupo para el conjunto de datos 1..7 Noviembre



Unificamos el análisis de cada conjunto de datos obteniendo la diferencia porcentual de cada tupla(hora,grupo)

Diferencia porcentual de los modelos óptimos por grupo



Hemos logrado una compresión interna del modelo óptimo para este atributo

Podemos establecer qué grupo representa mejor cada valor del atributo hora

Atributo hora

Grupo	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Hora		7	21	12	0		3	2	8	23	6	5	1	4	14
				13	16			9	15			22	20	10	
				17	18			11							
				19											

- Grupos vacíos
- Grupos con una hora
- Grupos con múltiples horas

Se propone eliminar una línea primaria de la RTC

La correspondiente al grupo cinco por estar altamente infrautilizada

Supondría el ahorro del coste de esta línea

Podemos elegir esquemas de fijación de precio basados en el horario de oficina para determinados grupos

De esta forma podemos ajustar las necesidades específicas a la infraestructura utilizada

Si analizamos otros aspectos del problema obtendremos nuevas propuestas de mejora para los sistemas de soporte del servicio

Índice

- Conceptos básicos
- Estado de la discusión científica
- Modelo óptimo de agrupamiento
- Evaluación con un caso de uso real
- Conclusiones

Hipótesis

"Es posible desarrollar un método que extraiga conocimiento de los datos de los servicios de las ciencias económicas de Internet. Este conocimiento puede representarse como modelos que nos ayuden a mejorar los sistemas de soporte de estos servicios"

Contribuciones

- Presentamos un vocabulario común para las ciencias económicas de Internet
- Desarrollamos una metodología para representar el modelo óptimo de un servicio
- Diseñamos un conjunto de métricas, restricciones y criterios para seleccionar el mejor resultado posible
- Recomendamos unas mejoras para los sistemas de soporte de un servicio

6 Artículos en congresos internacionales

2 Capítulos de libros

1 Contribución a estándares

13 Publicaciones indirectamente relacionadas

Trabajo futuro

- Crear un banco de pruebas sobre el que simular el funcionamiento de nuevos servicios
- Buscar métricas que permitan la generalización de los problemas
- Usar la metodología desarrollada sobre otros servicios

KNOWLEDGE DISCOVERY TECHNIQUES TO IMPROVE THE SERVICES OF INTERNET ECONOMICS

Igor Ruiz Agúndez

Dirigida por Dr. Pablo García Bringas y
Dr. Yoseba Koldobika Peña Landaburu

Aplicación de la métrica a otros dominios

1. Agrupamiento
2. Selección del modelo óptimo
3. Operación con el modelo

Metodología

1. • Agrupamiento
2. • Selección del modelo óptimo
3. • Operación con el modelo

1

- Agrupamiento (entrada)

Conjunto de datos de un servicio transformados con un plan de acción
Algoritmos de agrupamiento
Parámetros de configuración

1

- Agrupamiento (salida)

La tripleta de datos genera un conjunto de agrupamiento formado por:

Un modelo de comportamiento

Información de los detalles de la ejecución

La asignación de grupo a cada instancia

2

- Selección del modelo óptimo (entrada)

El conjunto de agrupamiento
El atributo base indica como medir
La métrica cuantifica el atributo
Las restricciones indican requisitos
El criterio ordena los modelos

2

- Selección del modelo óptimo (salida)

El atributo, la métrica, las restricciones y el criterio elegidos seleccionan el modelo óptimo para el problema

3

- Operación con el modelo (entrada)

El modelo óptimo de comportamiento

El atributo determina que analizar

La operación indica que analizar

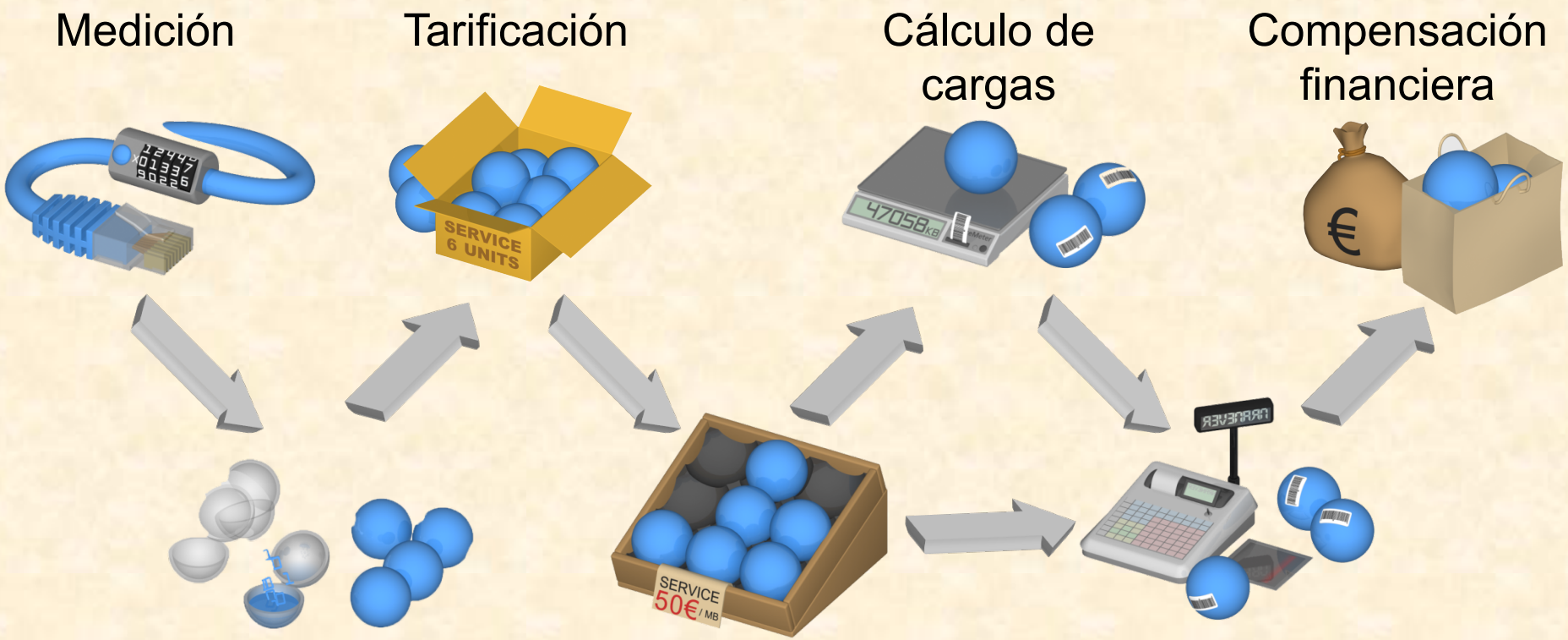
El umbral delimita el resultado

3

- Operación con el modelo (salida)

El modelo óptimo de comportamiento, el atributo, la operación y el umbral generan un conocimiento extrapolable

Proceso de consumo de un servicio



Métricas independientes

- Prueba error con distintas métricas y selección de las más generalizables
- Experimentación por aproximación, diseñando métricas ad hoc por áreas de conocimiento
- En el caso de aparecer nuevas técnicas disruptivas, incluirlas en la metodología
- Perseguir la métrica universal para conseguir aproximaciones realistas

Criterio

Recordemos que el modelo óptimo debería tener:

- Un gran grupo representando el comportamiento normal
- Múltiples grupos más pequeños representando el comportamiento atípico

Por ello hemos elegido el mayor valor de curtosis

Jerarquía de elementos

Restricciones

Las restricciones aplicadas purgan los resultados por no ofrecer resultados relevantes:

- Instancias sin un grupo
- Conjuntos de datos sin grupos
- Único grupo para todas las instancias

AIC y SRM

AIC: mide el ajuste de un modelo estadístico

SRM: balancea la complejidad de los modelos

Taxonomía

- Ciencia de la clasificación
- Aplicada a múltiples campos
- Sirven como contenedores de información y permiten hacer predicciones

Ha sido utilizada para organizar el proceso de consumo de un servicio

Curtosis

$$K(x) = \frac{\mu_4}{\sigma_4} - 3$$

Contribuciones

1

Presentamos un vocabulario común para las ciencias económicas de Internet

2

Desarrollamos una metodología para representar el modelo óptimo de un servicio

3

Diseñamos un conjunto de métricas, restricciones y criterios para seleccionar el mejor resultado posible

4

Recomendamos unas mejoras para los sistemas de soporte de un servicio

Otras decisiones de diseño en el caso de uso

Selección del modelo óptimo

- Atributo base y métrica
- Restricciones
- Criterio

Operación con el modelo

- Atributos