“Agrupación automática de descripciones de productos”

Loor Díaz Luis[1], Alvarado Juan[2]
Instituto de Ciencias Matemáticas (ICM)1
Escuela Superior Politécnica del Litroral(ESPOL)1
Campus Gustavo Galindo, Km 30.5 vía Perimetral1
Apartado 09-015863. Guayaquil-Ecuador1
lloordiaz@hotmail.com1
Ingeniero Electrico, Profesor de la ESPOL2
jao_ec@yahoo.es

Resumen

En este trabajo se realiza una aplicación acerca de la utilización del Data Mining. El propósito de esta tesis es automatizar el proceso de agrupamiento de las descripciones de productos de una gran cantidad de registros, para así tener grupos que puedan ser codificados posteriormente y así realizar inferencias sobre estos grupos que son una cantidad menor de datos, y que a la vez sea representativa.

Aquí se considera que, se toma una muestra de la población total (combinaciones de palabras), para poder realizar el análisis, ya que si tomamos la población total aunque serían más confiables los resultados no sería óptimo debido a que realizamos una comparación secuencial de cadenas de caracteres.

Cabe recalcar que esta tesis es una parte importante del proceso de extracción del conocimiento, porque permiten la agrupación de registros que antes se presentaban como separados (descripciones de productos) en cadenas de caracteres y esta agrupación nos permite un procesamiento de los datos de estos productos independientemente en que forma son presentadas estas descripciones

Este problema se lo conoce como Data Matching o De Duplex.

Palabras claves: *Agrupación, Data Mining, Distancias de edición.*

ABSTRACT

For this job an application is designed for the use of Data Mining. The purpose of this thesis is to automate the grouping process of the large number of entries with regards to Product Descriptions so that these groups can then be coded to generate inferences about the groups that have a lesser amount of entries and at the same time are representative.

In order to make a thorough analysis, we must consider that the sample be taken of the total population. Although it would be more reliable to consider the total population, the results would not be optimal because a sequential comparison of a chain of characters is required.

It is important to reiterate that this thesis is a very important part of the process of extraction as it will allow us to group the registries that were separated before (Product Descriptions) in chains of characters. This grouping will allow us to process the product’s data regardless of the manner in which these descriptions were presented.

This problem is also referred to as Data Matching or De Duplex.

1.0 INTRODUCCIÓN

En el campo del Reconocimiento de formas o patrones, las técnicas de clasificación basadas en distancia, necesitan de la obtención de prototipos adecuados para cada clase. Una de las posibilidades es usar la media de la clase (o el conjunto formado por la media de las diversas subclases que componen la clase) como prototipo de la misma.

Cuando se habla de espacios euclídeos (representación vectorial), hallar la media es un problema sencillo, pero no así si usamos la representación por cadenas de caracteres. En dicho caso, el problema de hallar la cadena media es duro ya que no existe el concepto de promedio de cadenas.

En este trabajo se propone usar la distancia de Levenshtein para obtener un valor por medio del cual mediremos la similaridad entre cadenas y así agrupar las descripciones, lo cual definiremos más adelante.

1.0 DATA MINING

Las técnicas del Data Mining son el resultado de un largo proceso de investigación. El Data Mining, es la exploración y análisis de grandes cantidades de datos para descubrir modelos significantes y reglas. También está dirigido a explicar o categorizar algún campo designado particular como detalles, descripciones, etc.

El Data Mining está principalmente dirigido a construir modelos entre los grandes grupos de datos o archivos. Un modelo Data Mining simplemente es un algoritmo o juego de reglas que conectan datos a un blanco particular o resultado.

Bajo las circunstancias correctas, un modelo puede producir una visión proporcionando una explicación de los resultados de un interés particular, como hacer un pedido o una orden de compra con un código que represente un determinado grupo de productos.

Data Mining está listo para su aplicación en la comunidad de negocios porque está soportado por tres tecnologías que ya están suficientemente maduras:

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de Data Mining.

2.0 EL ALCANCE DEL DATA MINING

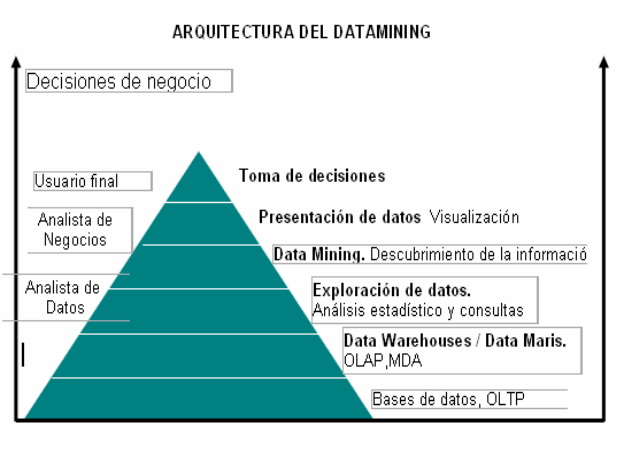
Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

- Predicción automatizada de tendencias y comportamientos.
- Data Mining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos. Un típico ejemplo de problema predecible es el marketing apuntado a objetivos (targeted marketing).
- Data Mining usa datos en mailing promocionales anteriores para identificar posibles objetivos para maximizar los resultados de la inversión en futuros mailing. Otros problemas predecibles incluyen pronósticos de problemas financieros futuros y otras formas de incumplimiento, e identificar segmentos de población que probablemente respondan similarmente a eventos dados.
- Descubrimiento automatizado de modelos previamente desconocidos.
- Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso.

2.1 ARQUITECTURA PARA DATA MINING

Para aplicar mejor estas técnicas avanzadas, éstas deben estar totalmente integradas con el data warehouse así como con herramientas flexibles e interactivas para el análisis de negocios.

Figura 1.1



Varias herramientas de Data Mining actualmente operan fuera del warehouse, requiriendo pasos extra para extraer, importar y analizar los datos. Además, cuando nuevos conceptos requieren implementación operacional, la integración con el warehouse simplifica la aplicación de los resultados desde Data Mining. El Data warehouse analítico resultante puede ser aplicado para mejorar procesos de negocios en toda la organización, en áreas tales como manejo de campañas promocionales, detección de fraudes, lanzamiento de nuevos productos, etc.

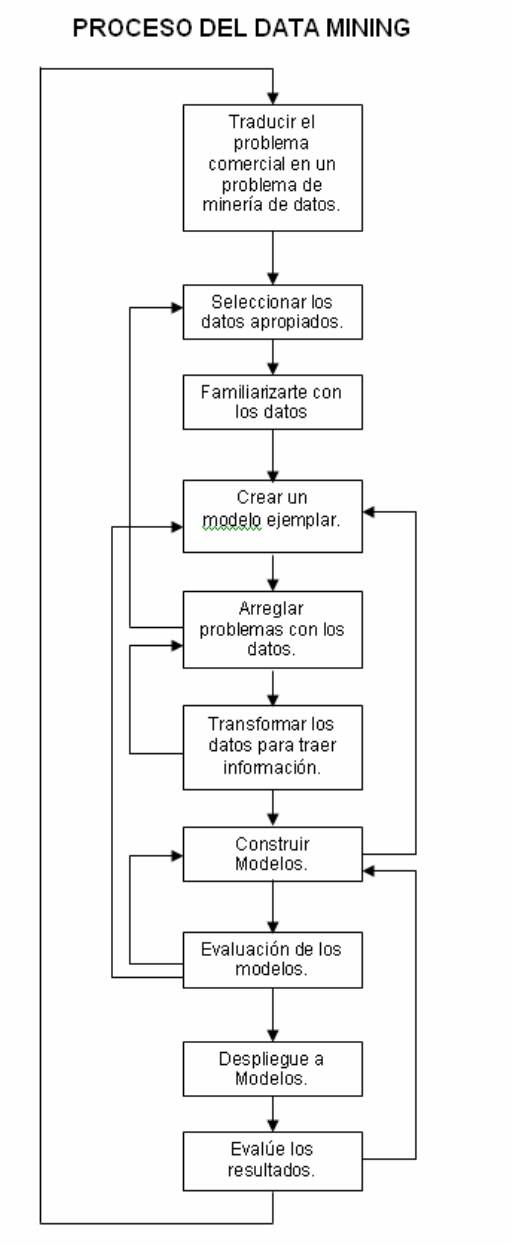
2.2 METODOLOGÍA DEL DATA MINING

La metodología del Data Mining tiene 11 pasos:

1.- Traducir el problema comercial o de negocios en un problema de minería de datos.
2.- Seleccionar los datos apropiados.
3.- Saber conseguir los datos.
4.- Crear un modelo ejemplar.
5.- Arreglar problemas con los datos.
6.- Transformar los datos para traer información.
7.- Modelos Robustos.
8.- Modelos de Evaluación.
9.- Despliegue a Modelos.
10.- Evalúe los resultados.
11.- Empiece de nuevo.

Como se muestra en la siguiente figura, el proceso del Data Mining es un conjunto de vueltas anidadas interactivas en lugar de una línea recta. Los pasos tienen un orden natural pero no es necesario seguirlo para terminar con un paso antes del siguiente.

Figura 1.2



Traducir el problema comercial en un problema de minería de datos.

Para transformar un problema de negocios en un problema de minería de datos, este debe ser reformulado como una de las seis tareas del data mining, tales como: clasificación, estimación, predicción, afinidad de agrupamiento, clustering, descripción y perfilamiento.

La tarea del modelamiento es encontrar reglas que expliquen los valores conocidos de las variables en blanco.

PASO 2: SELECCIONAR LOS DATOS APROPIADOS

La extracción del conocimiento requiere datos. Los datos requeridos estarían en un almacén corporativo de los datos, limpiados, disponible, históricamente exacto, y puestos al día con frecuencia. De hecho, se dispersa más a menudo en una variedad de sistemas operacionales en formatos incompatibles en la computadora, que funcionan diversos sistemas operativos, alcanzada a través de las herramientas de escritorio incompatibles.

Las fuentes de datos que son útiles y disponibles varían, por supuesto, de problema en problema y de industria en industria. Algunos ejemplos de datos útiles son:

- La garantía demanda datos (los campos incluyendo de fijo-formato y del texto libre).
- Expedientes de las cargas de tarjetas de crédito.
- Datos de puntos de venta (incluyendo códigos de anillo, cupones, descuentos aplicados)
- El seguro médico demanda de datos.
- Datos de registro del Web.
- Registros del uso del servidor del comercio electrónico.
- Expedientes de la respuesta del correo directo.
- Expedientes de centros de llamadas, incluyendo las notas escritas por los responsables del centro de llamadas.
- Expedientes del funcionamiento de la prensa.
- Expedientes de registro de motores de vehículos.
- Nivel de ruido en decibeles de micrófonos ubicados en comunidades cerca del aeropuerto.
- Expedientes de detalles de llamadas telefónicas.
- Datos de la respuesta del examen.
- Datos demográficos y de estilo de vida.
- Datos económicos.

Una vez que el problema de negocios ha sido formulado, es posible formar una lista deseada de datos que serían útiles tener. El primer paso es buscar los datos disponibles y crear una lista de candidatos a problemas de negocios.

PASO 3: FAMILIARIZARSE CON LOS DATOS

Los buenos mineros de datos se parecen confiar mucho en la intuición de alguna manera que puede conjeturar lo que pudo ser una buena variable derivada a intentar,

por ejemplo. La única manera de desarrollar la intuición para conocer lo que está pasando en un grupo de datos desconocido, es sumergirse en él. De manera que a la larga, se haga probable descubrir muchos problemas de la calidad de los datos, para responder cualquier pregunta que surja en determinado momento.

PASO 4: CREAR UN GRUPO DE DATOS

El sistema del modelo contiene todos los datos que son usados en el proceso del modelo. Algunos de los datos son usados para encontrar patrones y para verificar si el sistema del modelo de datos creado, es estable. Crear un sistema del modelo requiere datos que provienen de fuentes múltiples para el formulario de firmas del cliente y después preparar los datos para el análisis.

PASO 5: REPARAR PROBLEMAS CON LOS DATOS

Todos los datos son sucios. Todos los datos tienen problemas. Cuál es o no es un problema varía con la técnica de extracción del conocimiento. Para algunos, tal como árboles de decisión, los valores que faltan, y los afloramientos no causan demasiado apuro. Para otras, tales como redes neuronales, causan toda clase de apuro. Por esa razón, algunos de lo que tenemos que decir sobre problemas que fijan con datos se pueden encontrar en los capítulos en las técnicas donde causan la mayoría de la dificultad.
A continuación se nombran los problemas más comunes que necesitan ser reparados:

- Variables categóricas con demasiados valores.
- Variables numéricas con distribuciones sesgadas y outliers.
- Valores faltantes.
- Valores con los significados que cambian en un cierto plazo.
- Codificación contraria de los datos.

PASO 6: TRANSFORMAR LOS DATOS PARA TRAER INFORMACIÓN.

Una vez que los datos hayan sido revisados y los problemas importantes de los datos han sido reparados, los datos se deben preparar aún para el análisis. Esto implica el agregar campos derivados para traer la información a la superficie. Puede también implicar el quitar los outliers (puntuaciones extremas dentro una variable), variables numéricas, el agrupar clases para las variables categóricas, aplicando transformaciones tales como logaritmos, el dar vuelta de cuenta a proporciones, y los similares.

He aquí algunos ejemplos de transformaciones:

- Las tendencias de la captura.
- Crear cocientes y otras combinaciones de las variables.
- Conversión de cuentas a proporciones.

PASO 7: CONSTRUIR MODELOS

Los detalles de este paso varían de técnica en técnica y se describen en los capítulos dedicados a cada método de minería de datos. De modo general, éste es el paso donde la mayor parte del trabajo ocurre en crear un modelo. En la explotación minera dirigida a datos, el sistema de entrenamiento se utiliza para generar una explicación de la variable independiente o variable en blanco en términos de independiente o variables de entrada.
Esta explicación puede tomar la forma de una red neuronal, de un árbol de la decisión, de un gráfico del acoplamiento, o de una cierta otra representación de la relación entre la variable blanco y los otros campos en la base de datos. En la minería de datos indirecta, no hay variable blanco. El modelo encuentra relaciones entre los registros y los expresa como reglas de asociación o asignándolas a los clusters comunes.

La construcción de modelos es el único paso del proceso minería de datos que ha sido automatizado verdaderamente un moderno software de minería de datos. Por esa razón, relativamente toma poco tiempo un proyecto de minería de datos.

PASO 8: EVALUACIÓN DE LOS MODELOS

En este paso se determina si los modelos están trabajando o no. Un modelo determinado debe responder preguntas tales como:
¿Qué tan exacto es el modelo?
¿Cuan bien el modelo describe los datos observados?
¿Cuánta confianza se puede poner en las predicciones del modelo?
¿Cuán comprensible es el modelo?

Por supuesto, las respuestas a este tipo de preguntas corresponden al tipo del modelo que fue construido. Lo que refiere básicamente es a los méritos técnicos del modelo.

PASO 9: DESPLIEGUE DE LOS MODELOS

Desplegar un modelo significa el movimiento que hay de él, desde el ambiente de minería de datos al ambiente que llega. Este proceso puede ser fácil o difícil. En el peor de los casos, el modelo se desarrolla en un ambiente especial usando un software que no funciona en ninguna parte.
Para desplegar el modelo, un programador toma una descripción impresa del modelo y la recodifica en otro

lenguaje de programación de tal forma que pueda funcionar en la plataforma a la que llega.
Un problema más común es que el modelo utiliza las variables de entrada que no están en los datos originales. Esto no debe ser un problema puesto que las entradas se derivan por lo menos de campos que fueron extraídos originalmente del sistema del modelo.

Desafortunadamente, los mineros de datos no son siempre buenos al guardar un registro limpio, reutilizable de las transformaciones que se aplicaron a los datos.
Las cuentas representan a menudo una probabilidad que son valores típicamente numéricos entre 0 y 1, pero de ninguna manera tan necesariamente. Una cuenta pudo también ser una etiqueta de la clase proporcionada por un modelo de clustering, por ejemplo, o una etiqueta de la clase de una probabilidad.

PASO 10: DETERMINAR LOS RESULTADOS

Este paso nos ayuda a la toma de decisiones en la parte final del proceso del Data Mining. Una gráfica útil demostraría cuántos dólares se traen adentro para un gasto dado en la campaña de la comercialización. Después de todo, si desarrollar el modelo es muy costoso, un correo de la masa puede ser más rentable que el anterior nombrado.

¿Cuál es el coste fijo de creación de la campaña y el modelo que la ayuda?
¿Cuál es el coste por el recipiente de hacer la oferta?
¿Cuál es el coste por la respuesta de satisfacer la oferta?
¿Cuál es el valor de una respuesta positiva?

En fin, la medida que cuenta es la vuelta en la inversión. La elevación que mide un sistema ayuda en la prueba de elegir el modelo correcto. Pero, es muy importante medir estas cosas en el campo también. En una comercialización de la base de datos el uso, éste requiere siempre poner a grupos de control a un lado y cuidadosamente seguir la respuesta del cliente según varias cuentas del modelo.

3.0 DISTANCIA DE EDICIÓN

El problema de hallar la disimilitud entre dos cadenas de caracteres sobre un determinado alfabeto puede verse como una cantidad numérica que determine el "esfuerzo" necesario para convertir una cadena en otra.

En la Figura 1.3 podemos ver la representación gráfica habitual de las operaciones de edición, donde λ representa la cadena vacía (de manera que (λ;b) representa la inserción y (a;λ) representa el borrado).

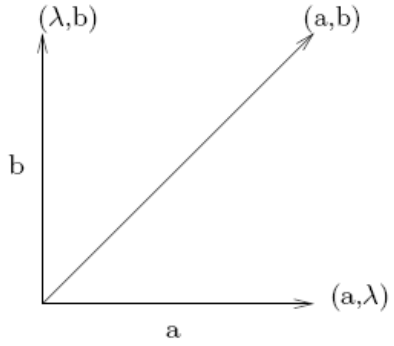


Figura 1.3

3.1 DISTANCIA DE LEVENSHTEIN

Sea F una función arbitraria de costo que asigne a cada operación de edición (ϕ , Ω) un número real no negativo, F(ϕ , Ω). Se puede extender F a la secuencia S definiendo:

$$F(S) = \sum_{j=1}^{n} F(S_j) \text{ si } n \geq \quad y \quad F(S) = 0 \text{ si } n = 0$$

Se denomina distancia de Levenshtein, DL(S,T), entre las cadenas S y T, a todas las secuencias de edición que transformen X en Y.

En este trabajo se ha seguido el criterio de que el costo de cualquier operación de edición se considera como unitario.

La mínima distancia de edición de dos cadenas s y t se define como el mínimo número de mutaciones requeridas para cambiar s a t, donde una mutación puede ser uno de los siguientes cambios:

- Cambiar una letra por otra
- Insertar una letra
- Eliminar una letra

3.2 IMPLEMENTACIÓN DEL ALGORITMO DE LEVENSHTEIN.

1.-Establecer una variable n que contenga el tamaño de la cadena de caracteres inicial s.
Establecer una variable m que contenga el tamaño de la cadena de caracteres final t.
Si n = 0, devuelva el valor de m y salga.
Si m = 0, devuelva el valor de n y salga.
Se construye una matriz d de m filas x n columnas.
2.-Inicializar la primera fila de 0…..n
Inicializar la primera columna de 0…..m
3.- Examine cada caracter de s (i de 1 a n).

4.- Examine cada caracter de t (j de 1 a m).
5.- Si s [i] es igual a t [j], el costo es 0.
Si s [i] no es igual a t [j], el costo es 1.
6.- Establecer la celda d [i , j] de la matriz igual al mínimo de:

a. La celda inmediatamente sobre 1: d [i-1, j] + 1.
b. La celda inmediatamente a la izquierda mas 1: d [i, j-1] + 1.
c. La celda diagonal sobre la izquierda más el costo: d[i-1, j-1] + cost.

7.- Después de que los pasos de iteración [3,4,5,6] son completados, la distancia es encontrada en la celda [n,m].

3.3 OBTENCIÓN DE LOS DATOS

Los datos que se tomaron para el desarrollo de esta tesis se asemeja mucho a lo que pueda existir en un servicio de Aduanas. Por lo tanto la mejor fuente considerada para la obtención de los mismos, fueron los listados de artículos que varias compañías ofrecen a través del Internet.

Se elaboro una lista con 1858 artículos de diferentes categorías de productos. Cabe recalcar que de este listado como en cualquier otro pueden existir datos repetidos. Por lo tanto, esto fue considerado en el desarrollo de la aplicación informática.

3.4 ANÁLISIS DE LOS DATOS

Luego de haber calculado el tamaño de muestra, utilizamos la distancia de levenshtein como herramienta, para obtener la mínima distancia de conversión (inserción, borrado o reemplazo) entre dos cadenas de caracteres. Para analizar los datos tomamos las cadenas de caracteres y las ubicamos en dos vectores, A y B. Figura 1.3

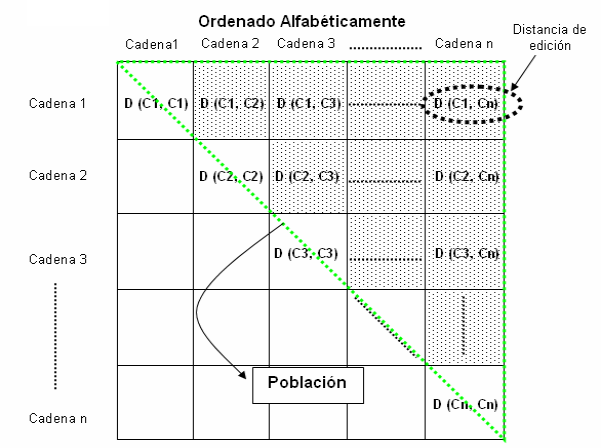


Figura 1.3

Entonces, realizamos una comparación secuencial para obtener las distancias que existen entre cada una de las cadenas para luego poder aplicar el criterio de agrupación en base al parámetro obtenido. El proceso de comparación secuencial se resume en una matriz en la cual sólo consideramos la triangular superior para evitar repeticiones y por lo tanto resultados erróneos.

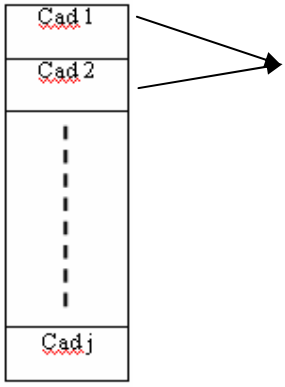
Como se observa en la figura 1.3, los cuadros denotados con una trama de puntos son las coordenadas de la matriz que consideramos para el análisis y cálculo de distancias secuenciales evitando así las repeticiones de los resultados.

3.5 CRITERIO DE AGRUPACIÓN

Dado una lista de cadenas de caracteres ordenadas alfabéticamente para realizar la agrupación se utilizara la siguiente regla:

Si d <= d* Los pares son iguales. (1)
Si d > d* Los pares no son iguales,

En donde d es la distancia entre las cadenas de caracteres, y d* es la distancia óptima. Llamamos distancia óptima a aquella distancia que nos genere el menor error de discriminación.



Si d (cad i, cad j) > d*, entonces se crea un grupo, G1

Se realiza nuevamente el calculo y la evaluación.

Si d(cad i, cad k) > d* , se crea otro grupo G2, caso contrario se queda en el grupo anterior G1 y así sucesivamente.

El error de discriminación de la regla (1) está dado por la siguiente fórmula:

Error de discriminación si d* = j

$$Error(d^* = j) = \sum_{i=1}^{j} 1 - \Pr ob(simi \setminus d) + \sum_{i=j+1}^{n} \Pr ob(simi \setminus d)$$

En donde es la probabilidad de que dos cadenas de caracteres sean similares a una determinada distancia y 1- es el complemento (lo contrario).

Error(d*=j) = Valor Esperado de Pares que no son similares y la distancia entre ellos es <= d* + Valor

Esperado de Pares que son iguales y la distancia entre ellos es > d*

	D(C1,C2) <= d*	D(C1,C2) > d*
C1 y C2 iguales	0	$\sum_{i=j+1}^{n} \Pr ob(simi/d)$
C1 y C2 no iguales	$\sum_{i=1}^{j} 1 - \Pr ob(simi/d)$	0

Por lo tanto, la distancia óptima esta dada por:

d* = min (error de discriminación)

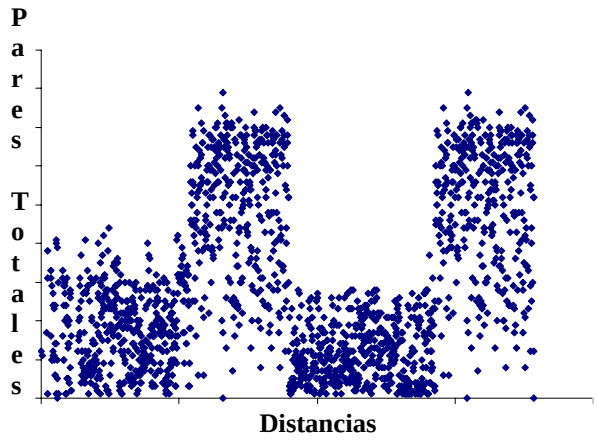
Tabla 1.1

dist	Probabilidad	Error
12	0.00840336	27.49
13	0.00840336	26.49
14	0.00840336	25.49
15	**0.00840336**	**24.49**
16	0.00840336	25.09
17	0.05042017	25.00
18	0.05882353	24.92
19	0.05042017	25.06
20	0.02521008	24.86
21	0.02521008	25.49
22	0.05882353	25.09
23	0.05882353	25.09
24	0.05882353	25.16
25	0.10084034	24.89
26	0.10084034	24.69
27	0.02521008	24.94
.		
78	0	26.51
79	0	27.51
80	0.00840336	26.51
81	0.00840336	25.51

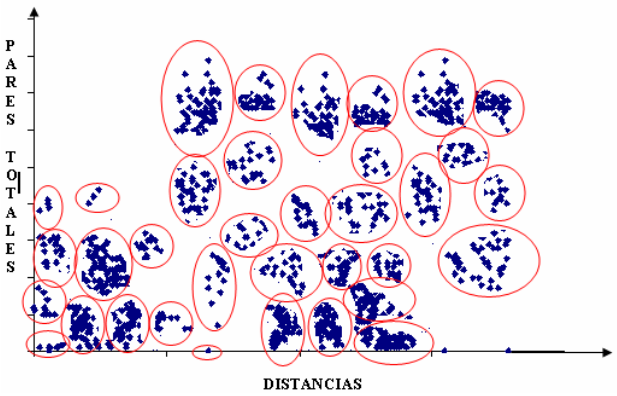
Como podemos observar en la tabla 1.1 encontramos una fila sombreada, la cual nos indica que genera el menor error de nuestro análisis y por lo tanto la distancia óptima (d*). Si observamos el gráfico encontramos que efectivamente la menor distancia de edición es 15 debido a que fue generada por el menor error de discriminación (24.49).

Aplicando esto en nuestra población total, porque recordemos que el análisis sobre la muestra es con el objetivo de proyectarlo sobre una cantidad de datos mucho mayor, en este caso nuestra población; nos da como resultado 202 grupos categóricos, en los cuales se encuentra distribuida la población total.

Para explicarlo gráficamente queremos llegar de una gran cantidad de registros, a una cantidad menor de grupos que representen dichos registros. Es decir, desde esto:



Hacia esto:



Grupos categóricos que representen a la población total, es decir, a los 1857 registros.

4.0 RECORD LINKAGE

El record linkage en una definición clara y concisa, es una fusión rápida y eficiente de múltiples ficheros con errores y con falta de datos.

Si partimos de un problema dado como identificar las distintas apariciones de coincidencias en una entidad en ficheros de registros y:

- Registros no identificados de manera global,
- Error en los datos.

La solución para nuestro problema es utilizar el record linkage. Las ventajas que este nos presenta es que no depende de un identificador global y es tolerante a errores en el registro e incompletitudes en los datos. La desventaja es que es un método estadístico e inexacto.

Otras de las desventajas que puede presentar el record linkage las detallamos a continuación:

Tratamiento previo de los datos (homogeneidad).
Formato
Errores de introducción.
Función de comparación de registros.
Determina si dos registros son similares.
Modelo probabilística.
Requiere un límite estadístico.
Realizar todas las comparaciones entre registros:
Coste cuadrático.
Por lo general inabordable.

Fases del proceso de record linkage

1.- Métodos de detección de similitudes (acercamiento de registros similares):

Blocking
Window

Esto implica la reducción del coste computacional del proceso y perdida de precisión en la detección de similitudes.

2.- Métodos de Análisis de las relaciones de similitud:

Clustering

Implica la inducción o discriminación de similitudes.

4.1 CARACTERÍSTICAS DEL RECORD LINKAGE

Dentro de las características del Record Linkage encontramos las siguientes:

- Precisión.
- Esfuerzo de revisión.
- Tiempo de ejecución.

La precisión

La perdida precisión para los casos correctos detectados se produce en la detección y en el análisis de similitudes.

Esfuerzo de revisión

El análisis de similitudes tiene un impacto significativo ya que existen casos en los cuales un par de registros pueden ser de la misma entidad y el sistema no los haya detectado y estaríamos perdiendo un integrando más de una determinada entidad.

Tiempo de Ejecución

Dentro del tiempo de ejecución se involucran múltiples ficheros y por lo tanto un número elevado de registros lo que conlleva a días para poder realizar un cruce de información para así verificar la efectividad del método.

DATOS BIBLIOGRAFICOS

1. http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/ MineriaDatosBressan.htm

2.-http://www.monografias.com/trabajos26/data-mining/data-mining.shtml

3.- http://www.monografias.com/trabajos/datamining/datamining.shtml

4.- **Data.Mining Techniques for Marketing Sales and Customer Support.(**2004),.2Ed. (Michael A. J. Berry & Gordon S. Linoff)

5**.- Data Mining Practical Machine Learning Tools and Techniques** (Ian H. Witten & Eibe Frank)

6.- **Técnicas de Muestreo Estadístico** – Teoría, práctica y aplicaciones informáticas. (César Pérez)

7.- **Estadística Matemática con Aplicaciones**. (Freud – Miller – Miller)