# Data Base Analysis using a Compact Data Set

# 

**Main Ideas:**

* With the increasing amounts of CRM data companies are gathering data warehouses are growing. Characterizing the customers with so much data is costly and not always fruitful. This issues has been approached with parallel computation, optimization of clustering algorithms, distributed and grid computing. The use of statistical methods to reduce the search space and some data treatment can solve this issue in a more efficient way.
* The project used the data from a major multinational and required gathering data from multiple sources related to customer characteristics, history of products and services contracted and billing history. 118 vars \* 12 million customers.
* Methodology:
  + Data analysis - eliminate deficiencies or limitations, detect input errors, bad numerical types, outliers, variables with too little too many values, too much too little numerical precision.
  + Programming language selection - used SAS for convenience, Oracle as the dbms, developed an external utility in low level language to calculate entropy from a smaller subset of the data.
  + Categorical var encoding - clustering algorithms work with numerical values only, write a program that identifies categorical vars, determine number of values that each var can take, create pseudo binary variables (dummy vars can only take 0 or 1). It avoids certain spurious correlations.
  + Calculate smallest equivalent sample - reduce the number of rows and columns to obtain a minable view that has as much “information” (think shanon concept of information) as the original. Two reductions.
    - Vertical - reduce attributes - multivariate analysis is used, many techniques exist. Pearson’s correlation coefficient was used, a correlation matrix is constructed. Expert consultation concluded that all vars with more than 0.85 coeff could be assumed to be similar. 118 to 73 vars. It is not relevant which var is kept in the minable view.
    - Horizontal reduction - reduce tuples - sampling - use of statistical theory of communication, consider customer info as a message and each attribute as a symbol. Estimate entropy contained in each message. See shanon’s formula of entropy H(X) = .There is an approximation formula described in the paper. The idea is to select a sample that has an entropy equal to the population. Sample randomly from the population and put in a sample set, calculate sample entropy, when entropy between population and sample are almost the same (based on an epsilon threshold - 0.0001) stop increasing the sample set. Process is done for each attribute and calculate the M number of sample size for each one. The largest M for all attributes is selected and this is the sample size needed. To validate the sample, a monte carlo non parametric test was applied with a hypothesis test verifying that any sample of size M keeps the same entropy level for all attributes. To keep the relationships across the attributes in the sample a set of 36 mathematical models was applied across arbitrarily selected attributes and computation of the errors of each sample and compared the maximum error to the minimum error, when this difference is small enough across all samples of size M has been found then M is set.
    - Seasonality - time dependent variables required to remove seasonality through moving averages.
  + Clustering - the number of clusters to find was not given - N was determined using the expert’s input as to what made sense. Used a Fuzzy C Means algorithm and the “elbow criterion” to determine the optimal number of clusters. This is the point where the partition’s entropy and the partition’s coefficient (how compact the data set is) changes are simultaneous. Optimal was achieved at 6.
  + Use of a self organizing map to find the definitive center of each cluster.
  + Use of pca to provide meaning to each cluster
  + The results are being used by the company worldwide.

**Interesting concepts:**

Clustering - obtain groups where the elements of each group exhibit small differences and large dissimilarities across groups.

The article mentions several articles for efficient clustering including the divide and merge and snakes and sandwiches approaches.

Sampling - is the process o selecting certain subset of the population - it can be probabilistic (Random, systematic or stratified) or non-probabilistic (convenience, judgment or quota). It has been criticized due to the possibility of missing small clusters in the population. But small clusters might not be important from a business perspective.

Feature selection - reduce dimensionality of a data set by droping some attributes. Some techniques include multivariate regression analysis, principal component analysis, variance and covariance analysis, canonical correlation analysis.

# A search space reduction methodology for data mining in large databases.

**Main Ideas:**

CRM databases volumes require new approaches for large data set clustering and processing techniques. Direct clustering over a database of several Terrabytes with million of registers is costly and not always fruitful. This article presents a methodology using statistical methods to reduce the search space of the problem. The article explains similar concepts than the first Kuri article, clustering techniques and sampling. Here they used 400,000 customers and 415 variables in 9 data tables.

**Methodology:**

* Data preprocessing
  + Search for incomplete, inconsistent or missing data
  + Transform non-numeric data to numeric
  + Define a single table denormalized after a primary/foreign key analysis
* Search space reduction
  + Vertical reduction - Using pearson´s as with the previous article. At a level of 0.75 all correlated variables were dropped. Only 129 variables were kept.
  + Horizontal reduction:
    - Sampling of 20% of the original data was taken based on expert consultation. Random selection. Validate the sample:
      * Select n equally sized samples (5)
      * Select sets of m variables to perform goodness-of-fit, for example m=2 to prove that within each sample the behavior of these variables is statistically equivalent.
      * Search for the best regressive from 34 different models and selected the one with the highest pearson correlation factor.
      * The article explains how it defined a way to establish that the fits found in each couple of variables associations are not given by pure chance. Check the algorithm.
* Clustering - no apriori constraint for the number of clusters, some idea based on expert consultation.
  + Validate the expert N number of clusters with empirical proof.
  + Use of Fuzzy C Means and determine the optimal number of clusters based on the “Elbow Criterion”.
  + Use of a Self-organizing map to find the optimal segmentation.
  + Use of entropy calculation is possible in fuzzy clustering. Using the membership degree and the probability of its appearance you can calculate the expected value of membership for a given cluster. As N increases the Partition’s Entropy also increases, PE is maximal when each element is a cluster. We need another measure. Use of partition coefficient a measure of how compact the dataset is. The article explains the formulas to calculate PE and PC. The elbow criterion uses curvature to determine the number of clusters. In this case it was at N = 6.
  + There is a neural network being used and a Chi-squared criterion.
  + Validation of the reduced search space:
    - Compare model 1 the sample and model 2 the whole population keeping the vertical reduction strategy intact for both models.
    - COmpare by seeing the amount of registers in percentage in each cluster. DIfferences are quite small (below 3%). Several tests are made using model 1 and model 2 to the sample and population data and cardinality of clusters is kept with a variation of less than 3% in every case.
  + Cost analysis
    - Cost-benefit analysis of the reduction of the search space. As the complexity of the clustering algorithm increases the savings of using a space reduction strategy becomes better.
    - A reduction of 93% of the original dataset was achieved.

**Interesting Concepts:**

# **Prueba de Normalidad**

**Test de normalidad.**

**Es muy útil visualizar las distribuciones de valores, detectar posibles errores y realizar pruebas de normalidad de los datos antes de realizar un análisis más complejo. En estas prácticas se van a trabajar tanto test de normalidad gráficos y numéricos. Los test de normalidad gráficos más utilizados en geoquímica son el histograma (explicado anteriormente), el gráfico Q-Q, el grafico de cajas y bigotes y el grafico probabilístico. El gráfico Q-Q normal representa los datos de la variable frente a los datos esperados si la distribución fuera normal. Si los puntos están cerca de la diagonal podemos decir que la distribución es normal.**

# **Referencia anafórica**

**La referencia anafórica o *anáfora* es un mecanismo mediante el cual un elemento del textoremite a otro que ha aparecido anteriormente, denominado *antecedente*. Se establece, pues, una relación interpretativa entre dos unidades lingüísticas en la que la segunda unidad adquiere sentido por su relación con la que se ha mencionado antes.**

**El fenómeno de la anáfora se inscribe entre los procedimientos lingüísticos que otorgan cohesión a un texto, como la progresión temática y la conexión. El estudio de las relaciones anafóricas constituye uno de los grandes objetivos de la gramática del texto y su desarrollo en la lingüística textual.**

# **Muestra minima**

**1. El radio (r) de un intervalo de confianza de la predicción de la media poblacional, la Desviación estándar (DE) y el tamaño muestral (n) mantienen una relación que puede expresarse en una ecuación muy importante en Estadística:**

**2. El algoritmo de Levenberg-Marquardt: In** [**mathematics**](https://en.wikipedia.org/wiki/Mathematics) **and computing, the Levenberg–Marquardt algorithm (LMA), also known as the damped least-squares (DLS) method, is used to solve** [**non-linear least squares**](https://en.wikipedia.org/wiki/Non-linear_least_squares) **problems. These minimization problems arise especially in** [**least squares**](https://en.wikipedia.org/wiki/Least_squares)[**curve fitting**](https://en.wikipedia.org/wiki/Curve_fitting)**.**

**3. Algoritmo de Ascenso: En** [**ciencia de la computación**](https://es.wikipedia.org/wiki/Ciencia_de_la_computaci%C3%B3n)**, hill climbing (ascenso de colinas, en alguna literatura) es una técnica de** [**optimización matemática**](https://es.wikipedia.org/wiki/Optimizaci%C3%B3n_(matem%C3%A1tica)) **que pertenece a la familia de los algoritmos de**[**búsqueda local**](https://es.wikipedia.org/w/index.php?title=B%C3%BAsqueda_local(optimizaci%C3%B3n)&action=edit&redlink=1)**. Es un algoritmo iterativo que comienza con una solución arbitraria a un problema, luego intenta encontrar una mejor solución variando** [**incrementalmente**](https://es.wikipedia.org/w/index.php?title=Heur%C3%ADstica_de_b%C3%BAsqueda_incremental&action=edit&redlink=1) **un único elemento de la solución. Si el cambio produce una mejor solución, otro cambio incremental se le realiza a la nueva solución, repitiendo este proceso hasta que no se puedan encontrar mejoras.**

**Simultaneous reduction—horizontal and verti- cal—yields the smallest representation of the original data set.**

# **Compresión de Datos**

**Sin pérdida**

**- Huffman: El algoritmo de codificación/compresión Huffman se propuso en 1952 como una forma sencilla y óptima de mapear cada símbolo de un alfabeto con un código (codeword) de longitud óptima. para comprimir cada símbolo de la cadena, simplemente debemos usar el código que se ha calculado mediante Huffman. Para conseguir esta asignación óptima, los símbolos se representan con códigos cuya longitud es inversamente proporcional a la probabilidad del símbolo. De esta forma, los símbolos menos probables se representan con códigos más largos, y los más probable con códigos más cortos.**

**- Lempel-Ziv-Welch**

**Con pérdida**

# **Tabla desnormalizada: vista minable**

**en un contexto de BD relacionales, es una única tabla con todos los atributos relevantes para el proceso de MD. Ya no segues una forma normal.**

**Leer más:** [**http://www.monografias.com/trabajos81/mineria-datos/mineria-datos2.shtml#ixzz3mV7YeOt9**](http://www.monografias.com/trabajos81/mineria-datos/mineria-datos2.shtml#ixzz3mV7YeOt9)

# **Prueba de bondad de ajuste**

**En la construcción del modelo de simulación es importante decidir si un conjunto de datos se ajusta apropiadamente a una distribución específica de probabilidad. Al probar la *bondad del ajuste* de un conjunto de datos, se comparan las frecuencias observadas *FO* realmente en cada categoría o intervalo de clase con las frecuencias esperadas teóricamente *FE*.**

**Prueba Ji cuadrada**

**Prueba de Kolmogorov - Smirnov**

# **Logica difusa**

**La lógica difusa (también llamada lógica borrosa) se basa en lo relativo de lo observado como posición diferencial. Este tipo de lógica toma dos valores** [**aleatorios**](https://es.wikipedia.org/wiki/Aleatoriedad)**, pero contextualizados y referidos entre sí. Así, por ejemplo, una persona que mida dos metros es claramente una persona alta, si previamente se ha tomado el valor de persona baja y se ha establecido en un metro. Ambos valores están contextualizados a personas y referidos a una medida métrica lineal.**

# **Clave primaria vs clave foránea**

**La llave primaria se utiliza para distinguir un registro con el fin de que se pueda tener acceso a ellos, organizarlos y manipularlos. En el caso de un registro de un empleado, él numero de este representa un ejemplo de una llave primaria.**

**Una clave o llave foránea (o Foreing key FK) es una limitación referencial entre dos tablas. La clave foránea identifica una columna o grupo de columnas en una tabla (tabla hija o referendo) que se refiere a una columna o grupo de columnas en otra tabla (tabla padre o referenciada).**

# **Coeficiente de correlación de Pearson**

**El coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente.**

# **Fit:**

**It is the set of n model codes obtained from the regressive analysis of the couples of the n samples. #similar fits/#possible fits**

# **Maximality:**

**The largest number of similar models in a fit. For instance, the fit (2, 5, 6, 5, 5) has a maximality of 3.**

**Similar fits:**

**Fits with the same maximality.**

**Possible fits**

**The total number of fits is given by the number of combina- tions with repetition for d models in n samples.**

**Number of clusters**

**Elbow method**

**Another method looks at the percentage of variance explained as a function of the number of clusters: One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data.**

# **Rule of Thumb**

**One simple** [**rule of thumb**](https://en.wikipedia.org/wiki/Rule_of_thumb) **sets the number to**[**[1]**](https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set#cite_note-1)**:365**

**with *n* as the number of objects (data points).**

# **Information Criterion Approach**

**An alternative approach to choosing the number of clusters that makes limited parametric assumptions, can be rigorously theoretically motivated using ideas from the field of rate distortion theory, is both simple to understand and compute, and is highly effective on a wide range of problems. The procedure is based on “distortion” which is a measure of within cluster dispersion.**