Generación Automática de Configuraciones Visuales

Victor Manuel Cardentey Fundora A.UNO@LAB.MATCOM.UH.CU
Grupo C511

Karla Olivera Hernández A.DOS@LAB.MATCOM.UH.CU
Grupo C511

Amanda González Borrell A.TRES@LAB.MATCOM.UH.CU
Grupo C511

Tutor(es):
Lic. Daniel Alejandro Valdés Pérez
Lic. Ernesto Estevanell

Resumen

La generación automática de visualizaciones sobre un conjunto de datos se puede dividir en dos procesos: determinar una consulta de interés para el usuario y generar la configuración gráfica para visualizar los resultados de la consulta. En particular la selección de configuraciones gráficas es un problema que presenta dificultades para llegar a consenso entre expertos del dominio y los sistemas tradicionales que brindan solución a este problema utilizan enfoques basados en reglas. En años recientes se ha planteado la posibilidad de aplicar técnicas de *Machine Learning* ampliamente utilizadas en sistemas de recomendación tradicionales a la recomendación de configuraciones gráficas. La propuesta de este trabajo consiste en utilizar y comparar distintos modelos de *Machine Learning* en la tarea de selección de configuraciones gráficas.

1. Introducción

2. Desarrollo

2.1 Selección de *features* de columnas

Medidas de dimensión

1. **Longitud**: La cantidad de elementos de una columna, esta medida obtuvo un alto índice de relevancia lo cual parece respaldar ciertas heurísticas utilizadas de forma común por analistas como pueden ser "no tener demasiadas barras en los gráficos" o "no tener demasiadas porciones en un gráfico de pastel" debido a que dificultan la correcta observación de los datos.

Medidas de tipo

1. **Tipo general**: El tipo general se refiere a la clasificación de la variable estadística pudiendo ser categórica (C), cuantitativa (Q) o temporal (T), esta clasificación se apoya en heurísticas comunes como utilizar variables temporales y categóricas en el eje x.

2. **Tipo Específico**: Se refiere al tipo de dato utilizado para representar la variable pudiendo ser una cadena de texto (*string*), un valor booleano (*boolean*), un entero (*integer*), un decimal (*decimal*) o una fecha (*datetime*).

Medidas de valores

Estas medidas se encargan de describir características de los datos de la columna y debido a que existen distintos tipos de variables estas medidas son dependientes del tipo.

1. Estadísticas [Q,T]:

   a) **Coeficiente de variación**: El coeficiente de variación expresa la razón entre la desviación típica y la media aritmética:

   $$CV = \frac{s}{\overline{x}}$$

   Tiene la ventaja de ser una medida de variabilidad relativa permitiendo poder comparar columnas con diferentes rangos de valores y unidades de medidas [1], además su interpretación como forma de evaluar la homogeneidad/heterogeneidad de un conjunto de datos permite discretizar esta variable.

CV	Interpretación
$CV \geq 0.26$	Muy heterogéneo
$0.16 \leq CV < 0.26$	Heterogéneo
$0.11 \leq CV < 0.16$	Homogéneo
$0 \leq CV < 0.11$	Muy homogéneo

   b) **Coeficiente de dispersión cuartil**: Este se define como:

   $$\frac{Q_3 - Q_1}{Q_3 + Q_1}$$

donde Q_1 y Q_3 son el primer y tercer cuartil respectivamente. Esta medida permite comparar los rangos de distintos conjuntos de datos aunque también es importante notar que es sensible a la presencia de valores extremos.

2. Distribución [Q]:

   a) **Entropía**: Se refiere a la definición de *entropía* en el campo de teoría de la información. Esta medida fue utilizada en [2] para establecer un *ranking* entre histogramas de acuerdo a la uniformidad de la distribución utilizando la siguiente definición. Sea un histograma de k intervalos entonces la entropía del histograma h es

   $$H(h) = -\sum_{i=1}^{k} p_i \log_2 p_i$$

   donde p_i es la probabilidad de que un elemento pertenezca al i-ésimo intervalo. Un alto valor de entropía se asocia a que los elementos pertenecen a una distribución uniforme y que el histograma tiende a ser plano.

   b) **Gini**: El coeficiente de Gini es una medida de dispersión definida como la media de las diferencias absolutas entre todos los posibles pares de individuos de una población para una medida dada.

   $$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \overline{x}}$$

   Donde n es la cantidad de medidas y $\overline{x}$ es la media aritmética. El valor mínimo es 0 cuando todas las medidas son iguales, esto puede ser utilizado para medir la homogeneidad/heterogeneidad de los datos.

   c) **Skewness**: Esta medida describe la asimetría de la distribución de acuerdo a la media, indicando la dirección y la magnitud relativa de la desviación de la distribución tomando como referencia una distribución normal. Es el tercer momento estándar definido como:

   $$\tilde{\mu}_3 = E\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$

   d) **Curtosis**: Esta medida permite describir el comportamiento de los datos en la cola de la distribución, esto permite obtener una medida de que tan susceptible es la distribución a la aparición de valores extremales. Es el cuarto momento estándar definido como:

   $$\tilde{\mu}_4 = E\left[\left(\frac{X - \mu}{\sigma}\right)^4\right]$$

   Además existe la definición alternativa de exceso de curtosis la cual permite discretizar esta variable, esta se define como $\tilde{\mu}'_4 = \tilde{\mu}_4 - 3$ y define las clases siguientes:

$\tilde{\mu}'_4$	Clase
$\tilde{\mu}'_4 = 0$	Mesokurtic
$\tilde{\mu}'_4 > 0$	Leptokurtic
$\tilde{\mu}'_4 < 0$	Platykurtic

   e) **Normalidad**:

   f) **Momentos de orden superior**:

3. Conclusiones

4. Recomendaciones

Referencias

[1] Prem S Mann. *Introductory statistics*, page 98. John Wiley & Sons, 2007.

[2] Jinwook Seo and Ben Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *IEEE Symposium on Information Visualization*, pages 65–72. IEEE, 2004.