Algoritmos de minería de datos incluidos en SQL Server 2008

Los algoritmos que aquí se presentan son: Árboles de decisión de Microsoft, Bayes naive de Microsoft, Clústeres de Microsoft, Serie temporal de Microsoft, Reglas de asociación de Microsoft, Clústeres de secuencia de Microsoft, Regresión lineal de Microsoft, Red neuronal de Microsoft, Regresión logística de Microsoft.

1. Algoritmo de árboles de decisión de Microsoft [MIC2009a]

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporcionado por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos.

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Por ejemplo, en un escenario para predecir qué clientes van a adquirir probablemente una bicicleta, si nueve de diez clientes jóvenes compran una bicicleta, pero sólo lo hacen dos de diez clientes de edad mayor, el algoritmo infiere que la edad es un buen elemento de predicción en la compra de bicicletas. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto.

Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión.

Si se define más de una columna como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción.

Cómo funciona el algoritmo

El algoritmo de árboles de decisión de Microsoft genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

El algoritmo de árboles de decisión de Microsoft utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos de Analysis Services utilizan la selección de características para mejorar el rendimiento y la calidad del análisis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si se utilizan demasiados atributos de predicción o de entrada al diseñar un modelo de minería de datos, el modelo puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Entre los métodos que se usan para determinar si hay que dividir el árbol figuran métricas estándar del sector para la entropía y las redes Bayesianas.

Un problema común de los modelos de minería de datos es que el modelo se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está sobreajustado o sobreentrenado. Un modelo sobreajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobreajustar un conjunto de datos determinado, el algoritmo de árboles de decisión de Microsoft utiliza técnicas para controlar el crecimiento del árbol.

Predecir columnas discretas

La forma en que el algoritmo de árboles de decisión de Microsoft genera un árbol para una columna de predicción discreta puede mostrarse mediante un histograma. La Figura 1 muestra un histograma que traza una columna de predicción, Comprador, con una columna de entrada, Edad. El histograma muestra que la edad de una persona ayuda a distinguir si esa persona comprará una bicicleta.

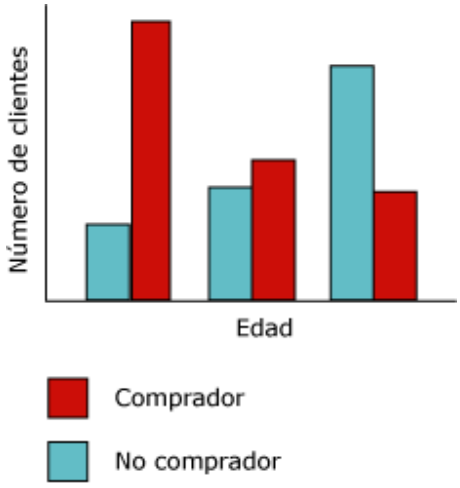


Figura 1: Histograma de una columna de predicción.

La correlación que aparece en la Figura 1 hará que el algoritmo de árboles de decisión de Microsoft cree un nuevo nodo en el modelo.

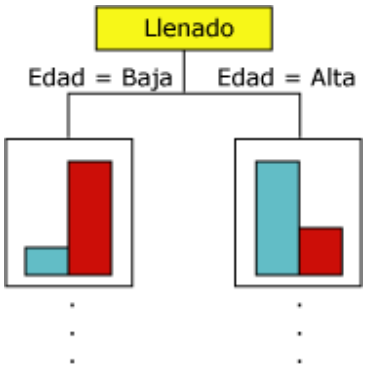


Figura 2: Llenado de un árbol de decisión.

A medida que el algoritmo agrega nuevos nodos a un modelo, se forma una estructura en árbol. El nodo superior del árbol describe el desglose de la columna de predicción para la población global de clientes. A medida que el modelo crece, el algoritmo considera todas las columnas.

Predecir columnas continuas

Cuando el algoritmo de árboles de decisión de Microsoft genera un árbol basándose en una columna de predicción continua, cada nodo contiene una fórmula de regresión. Se produce una división en un punto de no linealidad de la fórmula de regresión. Por ejemplo, considere la Figura 3.

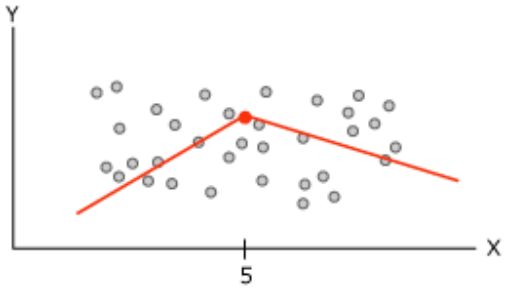


Figura 3: División en un punto de no linealidad de la fórmula de regresión.

La Figura 3 contiene los datos que pueden modelarse utilizando una sola línea o dos líneas conectadas. Sin embargo, una sola línea realizará un pobre trabajo en la representación de los datos. En su lugar, si se usan dos líneas, el modelo hará un mejor trabajo en la aproximación a los datos. El punto donde las dos líneas se unen es el punto de no linealidad y donde se dividiría un nodo de un modelo de árbol de decisión. Por ejemplo, el nodo que

corresponde al punto de no linealidad del gráfico anterior podría representarse mediante la Figura 4. Las dos ecuaciones representan las ecuaciones de regresión de las dos líneas.

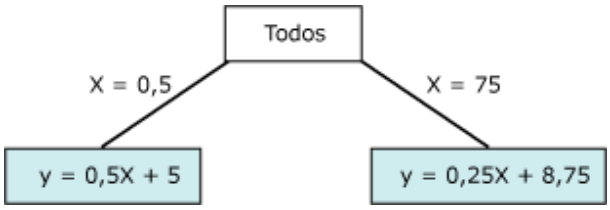


Figura 4: Representación de un punto de no linealidad.

Requisitos para un modelo de árboles de decisión

Una única columna key: Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Una columna de predicción: Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.

Columnas de entrada: Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

Ver un modelo de árboles de decisión

Para examinar el modelo, puede utilizar el Visor de árboles de Microsoft. Si un modelo genera varios árboles, puede seleccionar uno y el visor muestra un esquema de cómo se clasifican los casos para cada atributo de predicción. También puede ver la interacción de los árboles utilizando el visor de redes de dependencias.

Si desea obtener información más detallada sobre cualquier bifurcación o nodo del árbol, también puede examinar el modelo utilizando el Visor de árbol de contenido genérico de Microsoft. El contenido almacenado para el modelo incluye la distribución para todos los valores de cada nodo, las probabilidades en cada nivel del árbol y las fórmulas de regresión para los atributos continuos.

2. Algoritmo Bayes naive de Microsoft [MIC2009b]

El algoritmo Bayes naive de Microsoft es un algoritmo de clasificación que proporciona Microsoft SQL Server Analysis Services para el modelado de predicción. El nombre Bayes naive deriva del hecho de que el algoritmo usa el teorema de Bayes, pero no tiene en cuenta las dependencias que pudieran existir y, por consiguiente, se dice que sus suposiciones son ingenuas o "naive".

Desde el punto de vista computacional, el algoritmo es menos complejo que otros algoritmos de Microsoft y, por tanto, resulta útil para generar rápidamente modelos de minería de datos para descubrir relaciones entre columnas de entrada y columnas de predicción. Se puede utilizar este algoritmo para realizar exploraciones iniciales de datos y, más adelante, aplicar los resultados para crear modelos de minería de datos adicionales con otros algoritmos más complejos y precisos desde el punto de vista computacional.

Funcionamiento del algoritmo

El algoritmo Bayes naive de Microsoft calcula la probabilidad de cada estado de cada columna de entrada, dado cada posible estado de la columna de predicción. Puede utilizar el Visor Bayes naive de Microsoft en Business Intelligence Development Studio para consultar una representación visual del modo en que el algoritmo distribuye los estados, como se muestra en la Figura 5.

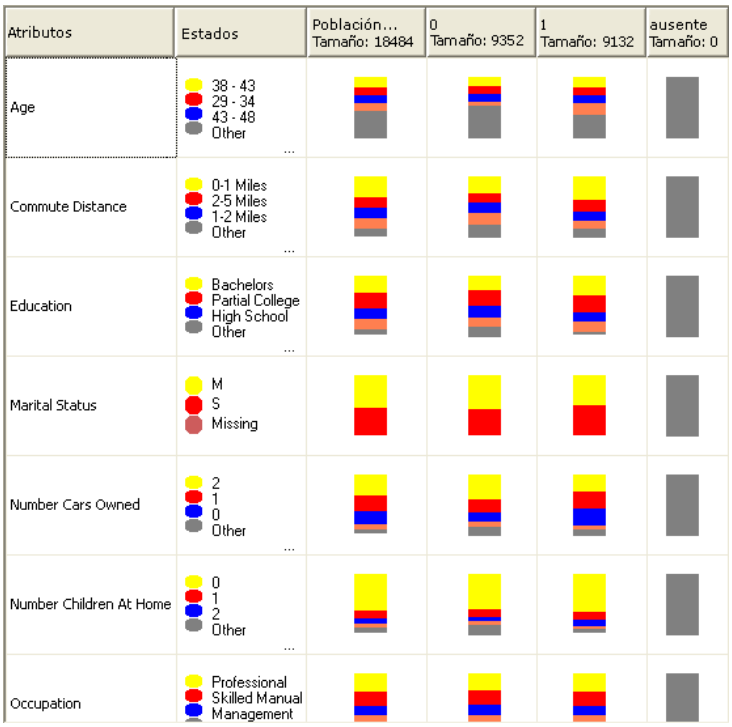


Figura 5: Columnas de entrada, dado cada probable estado de la columna de predicción.

El Visor Bayes naive de Microsoft muestra las columnas de entrada del conjunto de datos e indica cómo se distribuyen los estados de cada columna, dado cada estado de la columna de predicción. Puede usar esta vista para identificar las columnas de entrada que son importantes para diferenciar los distintos estados de la columna de predicción. Por ejemplo, en la columna Commute Distance (distancia que se ha de recorrer para llegar al trabajo), si el cliente tiene que desplazarse una distancia de dos a tres kilómetros, la probabilidad de que dicho cliente adquiera una bicicleta es de 0,387, mientras que la probabilidad de que no la adquiera es de 0,287. En este ejemplo, el algoritmo utiliza la información numérica derivada de un dato de cliente como la distancia entre el domicilio y el lugar de trabajo para predecir si un cliente compraría una bicicleta.

Requisitos para un modelo Bayes naive

Una columna de una sola clave: cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Columnas de entrada: en un modelo Bayes naive, todas las columnas deben ser discretas o de datos discretos. En un modelo Bayes naive, es importante asegurarse de que los atributos de entrada sean independientes unos de otros.

Al menos una columna de predicción: el atributo de predicción debe contener valores discretos o discretizados. Los valores de la columna de predicción se pueden tratar como entrada y, a menudo, se usan para buscar las relaciones entre las columnas.

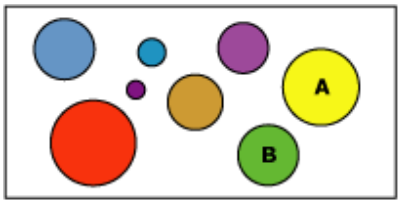
Ver el modelo

El Visor Bayes naive de Microsoft muestra cómo se relacionan los atributos de entrada con el atributo de predicción. El visor también proporciona un perfil detallado de cada clúster, una lista de los atributos que distinguen cada clúster de los demás, y las características del conjunto de datos de entrenamiento completo.

3. Algoritmo de clústeres de Microsoft [MIC2009c]

El algoritmo de clústeres de Microsoft es un algoritmo de segmentación suministrado por SQL Server 2008 Analysis Services (SSAS). El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones.

Los modelos de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual. Por ejemplo, puede discernir lógicamente que las personas que se desplazan a sus trabajos en bicicleta no viven, por lo general, a gran distancia de sus centros de trabajo. Sin embargo, el algoritmo puede encontrar otras características que no son evidentes acerca de los trabajadores que se desplazan en bicicleta. En la Figura 6, el clúster A representa los datos sobre las personas que suelen conducir hasta el trabajo, en tanto que el clúster B representa los datos sobre las personas que van hasta allí en bicicleta.



A = Trabajadores que conducen para ir al trabajo
B = Trabajadores que van en bicicleta al trabajo

Figura 6: Ejemplo de cluster.

El algoritmo de agrupación en clústeres se diferencia de otros algoritmos de minería de datos, como el algoritmo de árboles de decisión de Microsoft, en que no se tiene que designar una columna de predicción para generar un modelo de agrupación en clústeres. El algoritmo de agrupación en clústeres entrena el modelo de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

Cómo funciona el algoritmo

El algoritmo de agrupación en clústeres de Microsoft identifica primero las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas. Un gráfico de dispersión es una forma útil de representar visualmente el modo en que el algoritmo agrupa los datos, tal como se muestra en la Figura 7. El gráfico de dispersión representa todos los casos del conjunto de datos; cada caso es un punto del gráfico. Los clústeres agrupan los puntos del gráfico e ilustran las relaciones que identifica el algoritmo.

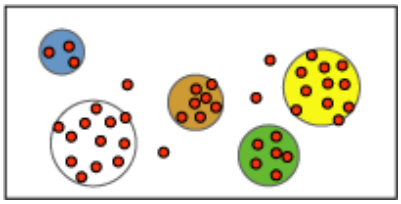


Figura 7: Gráfico de dispersión.

Después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos y, a continuación, intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los clústeres.

Se puede personalizar el funcionamiento del algoritmo seleccionando una técnica de agrupación en clústeres, limitando el número máximo de clústeres o cambiando la cantidad de soporte que se requiere para crear un clúster.

Requisitos para un modelo de agrupación en clústeres

Una única columna key: Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Columnas de entrada: Cada modelo debe tener al menos una columna de entrada que contenga los valores que se utilizan para generar los clústeres. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el modelo.

Una columna de predicción (opcional): El algoritmo no necesita una columna de predicción para generar el modelo, pero puede agregar una columna de predicción de casi cualquier tipo de datos. Los valores de la columna de predicción se pueden tratar como entradas del modelo de agrupación en clústeres, o se puede especificar que sólo se utilicen para las predicciones. Por ejemplo, si desea predecir los ingresos del cliente agrupando en clústeres de acuerdo con datos demográficos como la región o la edad, se deben especificar los ingresos como PredictOnly y agregar todas las demás columnas, como la región o la edad, como entradas.

Ver un modelo de agrupación en clústeres

Para explorar el modelo, puede utilizar el Visor de clústeres de Microsoft. Cuando se observa un modelo de agrupación en clústeres, Analysis Services presenta los clústeres en un diagrama que muestra las relaciones existentes entre ellos, además de un perfil detallado de cada clúster, una lista de los atributos que diferencian cada clúster de los demás, y las características de todo el conjunto de datos de entrenamiento.

4. Algoritmo de serie temporal de Microsoft [MIC2009d]

El algoritmo de serie temporal de Microsoft proporciona los algoritmos de regresión que se optimizan para la previsión en el tiempo de valores continuos tales como las ventas de productos. Mientras que otros algoritmos de Microsoft, como por ejemplo los árboles de decisión, requieren columnas adicionales de nueva información como entrada para predecir una tendencia, los modelos de serie temporal no las necesitan. Un modelo de serie temporal puede predecir tendencias basadas únicamente en el conjunto de datos original utilizado para

crear el modelo. Es posible también agregar nuevos datos al modelo al realizar una predicción e incorporar automáticamente los nuevos datos en el análisis de tendencias.

La Figura 8 muestra un modelo típico de previsión en el tiempo de las ventas de un producto en cuatro regiones de ventas diferentes. La línea de cada región consta de dos partes:

- La información histórica aparece a la izquierda de la línea vertical y representa los datos que el algoritmo utiliza para crear el modelo.
- La información de la predicción aparece a la derecha de la línea vertical y representa la previsión realizada por el modelo.

A la combinación de los datos de origen y los datos de la predicción se le denomina serie.

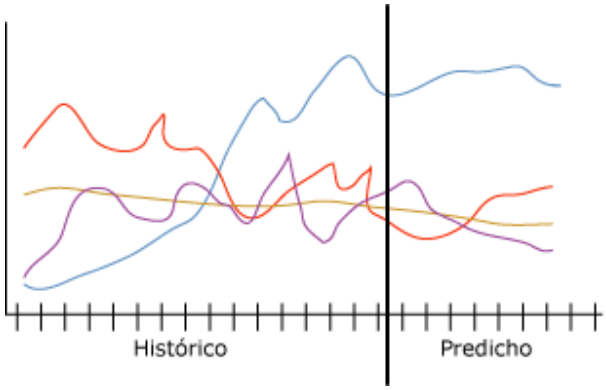


Figura 8: Modelo típico de previsión.

Una característica importante del algoritmo de serie temporal de Microsoft es su capacidad para llevar a cabo predicciones cruzadas. Si entrena el algoritmo con dos series independientes, pero relacionadas, puede utilizar el modelo generado para predecir el resultado de una serie basándose en el comportamiento de la otra. Por ejemplo, las ventas observadas de un producto pueden influir en las ventas previstas de otro producto. La predicción cruzada también es útil para crear un modelo general que se puede aplicar a múltiples series. Por ejemplo, las predicciones para una región determinada son inestables debido a que la serie no dispone de datos de buena calidad. Podría entrenar un modelo general sobre la media de las cuatro regiones y, a continuación, aplicar el modelo a las series individuales para crear predicciones más estables para cada región.

Cómo funciona el algoritmo

En SQL Server 2005, el algoritmo de serie temporal de Microsoft utilizaba un único algoritmo, ARTxp. El algoritmo ARTxp se optimizó para predicciones a corto plazo, y por consiguiente, predecía el siguiente valor probable en una serie. En SQL Server 2008, el algoritmo de serie temporal de Microsoft utiliza, además del algoritmo ARTxp, un segundo algoritmo, ARIMA. El algoritmo ARIMA está optimizado para la predicción a largo plazo.

De forma predeterminada, el algoritmo de serie temporal de Microsoft utiliza una mezcla de los dos algoritmos al analizar patrones y realizar predicciones. El algoritmo entrena dos modelos independientes sobre los mismos datos: uno de los modelos utiliza el algoritmo ARTxp y el otro modelo utiliza el algoritmo ARIMA. A continuación, el algoritmo combina los resultados de los dos modelos para obtener la mejor predicción sobre un número variable de intervalos de tiempo. Dado que ARTxp obtiene mejores resultados en las predicciones a corto plazo, se le da mayor importancia al principio de una serie de predicciones. Sin embargo, a medida que los intervalos de tiempo que se están prediciendo se adentran en el futuro, se va dando más importancia a ARIMA.

Es posible también controlar la mezcla de algoritmos para favorecer la predicción a corto o a largo plazo en las series temporales. En SQL Server 2008 Standard Edition, es posible especificar que el algoritmo de serie temporal de Microsoft use uno de los valores siguientes:

- Utilizar sólo ARTXP para la predicción a corto plazo.
- Utilizar sólo ARIMA para la predicción a largo plazo.
- Utilizar la mezcla predeterminada de los dos algoritmos.

En SQL Server 2008 Enterprise, es posible personalizar la manera en que el algoritmo de serie temporal de Microsoft combina los modelos para la predicción. Al utilizar un modelo mixto, el algoritmo de serie temporal de Microsoft combina los dos algoritmos de la manera siguiente:

- Sólo ARTXP se utiliza siempre para realizar el primer par de predicciones.
- Tras el primer par de predicciones, se utiliza una combinación de ARIMA y ARTxp.
- A medida que el número de pasos de la predicción aumenta, las predicciones se basan en mayor medida en ARIMA hasta que llega un momento en que ARTxp deja de utilizarse.

- Es posible controlar el punto de combinación, esto es, el ritmo al que el peso de ARTXP disminuye y el peso de ARIMA aumenta, mediante el parámetro PREDICTION_SMOOTHING.

Ambos algoritmos pueden detectar estacionalidad en los datos en varios niveles. Por ejemplo, sus datos podrían contener ciclos mensuales anidados en ciclos anuales. Para detectar estos ciclos estacionales, es posible proporcionar una sugerencia de periodicidad o bien especificar que el algoritmo deberá detectar automáticamente la periodicidad.

Datos requeridos para los modelos de serie temporal

Al preparar los datos para el entrenamiento de cualquier modelo de minería de datos, es preciso comprender los requisitos del modelo en particular así como la forma en que se utilizan los datos.

Cada modelo de previsión debe contener una serie de casos, que es la columna que especifica los intervalos de tiempo u otras series sobre las que se produce el cambio. Por ejemplo, los datos de la Figura 8 muestran las series correspondientes al historial y a la previsión de ventas de bicicletas para un período de varios meses. Para este modelo, cada región es una serie y la columna de fecha contiene la serie temporal, que también es la serie de casos. En otros modelos, la serie de escenarios puede ser un campo de texto o algún identificador tal como un id. de cliente o de transacción. Sin embargo, un modelo de serie temporal debe siempre utilizar una fecha, una hora o algún otro valor numérico único para su serie de escenarios.

Requisitos para un modelo de serie temporal

Una única columna Key Time: Cada modelo debe contener una columna numérica o de fecha que se utilizará como serie de casos y que define los intervalos de tiempo que utilizará el modelo. El tipo de datos para la columna de clave temporal puede ser un tipo de datos datetime o bien numérico. Sin embargo, la columna debe contener valores continuos y éstos deben ser únicos para cada serie. La serie de casos para un modelo de serie temporal no pueden estar almacenada en dos columnas como por ejemplo una columna Año y una columna Mes.

Una columna predecible: Cada modelo debe contener por lo menos una columna predecible alrededor de la que el algoritmo generará el modelo de serie temporal. El tipo de

datos de la columna predecible debe contener valores continuos. Por ejemplo, es posible predecir la manera en que los atributos numéricos tales como ingreso, ventas o temperatura, varían con el tiempo. Sin embargo, no es posible utilizar como columna predecible una columna que contenga valores discretos tales como el estado de las compras o el nivel de educación.

Una columna de clave de serie (opcional): Cada modelo puede tener una columna de clave adicional que contenga valores únicos que identifiquen a una serie. La columna de clave de serie opcional debe contener valores únicos. Por ejemplo, un solo modelo puede contener ventas de muchos modelos de producto, siempre y cuando haya un solo registro para cada nombre del producto para cada intervalo de tiempo.

Visualización de un modelo de serie temporal

Una vez entrenado el modelo, los resultados se encuentran almacenados como un conjunto de modelos, que se pueden explorar o utilizar para realizar predicciones.

Para explorar el modelo, se puede utilizar el Visor de series temporales. El visor incluye un gráfico que muestra las predicciones futuras y una vista de árbol de las estructuras periódicas en los datos.

5. Algoritmo de asociación de Microsoft [MIC2009e]

Este algoritmo de Microsoft es un algoritmo de asociación suministrado por Analysis Services, útil para los motores de recomendación. Un motor de recomendación recomienda productos a los clientes basándose en los elementos que ya han adquirido o en los que tienen interés.

Los modelos de asociación se generan basándose en conjuntos de datos que contienen identificadores para casos individuales y para los elementos que contienen los casos. Un grupo de elementos de un caso se denomina un conjunto de elementos. Un modelo de asociación se compone de una serie de conjuntos de elementos y de las reglas que describen cómo estos elementos se agrupan dentro de los casos. Las reglas que el algoritmo identifica pueden utilizarse para predecir las probables compras de un cliente en el futuro, basándose en los elementos existentes en la cesta de compra actual del cliente. La Figura 9 muestra una serie de reglas en un conjunto de elementos.

Regla
Road Bottle Cage = Existing, Cycling Cap = Existing -> Water Bottle = Existing
Mountain-200 = Existing, Mountain Tire Tube = Existing -> HL Mountain Tire = Existing
Mountain-200 = Existing, Water Bottle = Existing -> Mountain Bottle Cage = Existing
Touring-1000 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Road-750 = Existing, Water Bottle = Existing -> Road Bottle Cage = Existing
Touring Tire = Existing, Sport-100 = Existing -> Touring Tire Tube = Existing

Figura 9: Reglas derivadas de un conjunto de elementos.

Como muestra la Figura 9, el algoritmo de asociación de Microsoft puede encontrar potencialmente muchas reglas dentro de un conjunto de datos. El algoritmo usa dos parámetros, soporte y probabilidad, para describir los conjuntos de elementos y las reglas que genera. Por ejemplo, si X e Y representan dos elementos que pueden formar parte de la cesta de la compra, el parámetro de soporte es el número de casos del conjunto de datos que contienen la combinación de ambos elementos, X e Y. Mediante el uso del parámetro de soporte en combinación con los parámetros MINIMUM_SUPPORT y MAXIMUM_SUPPORT, definidos por el usuario, el algoritmo controla el número de conjuntos de elementos que se generan. El parámetro de probabilidad, también denominado parámetro de confianza, representa la fracción de casos del conjunto de datos que contiene X y que también contiene Y. Mediante el uso del parámetro de probabilidad en combinación con el parámetro MINIMUM_PROBABILITY, el algoritmo controla el número de reglas que se generan.

Cómo funciona el algoritmo

El algoritmo de asociación de Microsoft recorre un conjunto de datos para hallar elementos que aparezcan juntos en un caso. A continuación, agrupa en conjuntos de elementos todos los elementos asociados que aparecen, como mínimo, en el número de casos especificado en el parámetro MINIMUM_SUPPORT. Por ejemplo, un conjunto de elementos puede ser "Mountain 200=Existing, Sport 100=Existing", y puede tener un soporte de 710. El algoritmo generará reglas a partir de los conjuntos de elementos. Estas reglas se usan para predecir la presencia de un elemento en la base de datos, basándose en la presencia de otros elementos específicos que el algoritmo ha identificado como importantes. Por ejemplo, una regla puede ser "if Touring 1000=existing and Road bottle cage=existing, then Water bottle=existing", y puede tener una probabilidad de 0.812. En este ejemplo, el algoritmo

identifica que la presencia en la cesta del neumático Touring 1000 y del soporte de la botella de agua predice que probablemente la cesta de compra incluirá también una botella de agua.

Requisitos para los modelos de asociación

Una única columna key: Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Una única columna de predicción: Un modelo de asociación sólo puede tener una columna de predicción. Normalmente, se trata de la columna de clave de la tabla anidada, como el campo que contiene los productos que se han comprado. Los valores deben ser discretos o discretizados.

Columnas de entrada: Las columnas de entrada deben ser discretas. Los datos de entrada de un modelo de asociación suelen encontrarse en dos tablas. Por ejemplo, una tabla puede contener la información del cliente y la otra las compras de ese cliente. Es posible incluir estos datos en el modelo mediante el uso de una tabla anidada.

Ver un modelo de asociación

Para explorar el modelo, puede utilizar el Visor de asociación de Microsoft. Cuando se observa un modelo de asociación, Analysis Services presenta las correlaciones desde distintos ángulos para que se puedan comprender mejor las relaciones y las reglas halladas en los datos. El panel Conjunto de elementos del visor proporciona un análisis detallado de las combinaciones o los conjuntos de elementos más comunes. El panel Reglas presenta una lista de las reglas que se han generalizado a partir de los datos, agrega cálculos de probabilidad y otorga un rango a las reglas según su importancia relativa.

6. Algoritmo de agrupación en clústeres de secuencia de Microsoft [MIC2009f]

El algoritmo de clústeres de secuencia de Microsoft es un algoritmo de análisis de secuencias que proporciona Microsoft SQL Server Analysis Services. Puede utilizar este algoritmo para explorar los datos que contienen eventos que pueden vincularse mediante rutas o secuencias. El algoritmo encuentra las secuencias más comunes mediante la agrupación, o agrupación en clústeres, de las secuencias que son idénticas. Éstos son algunos ejemplos de secuencias:

- Los datos que describen las rutas de clicks que se crean cuando los usuarios navegan o examinan un sitio web.
- Los datos que describen el orden en el que un cliente agrega elementos en una cesta de compra de un comerciante en línea.

Este algoritmo es similar en muchas maneras al algoritmo de clústeres de Microsoft. Sin embargo, en lugar de encontrar clústeres de casos que contienen atributos similares, el algoritmo de clústeres de secuencia de Microsoft encuentra clústeres de casos que contienen rutas similares en una secuencia.

Cómo funciona el algoritmo

El algoritmo de clústeres de secuencias de Microsoft es un algoritmo híbrido que combina técnicas de agrupación en clústeres con el análisis de cadenas de Markov para identificar los clústeres y sus secuencias. Una de las marcas distintivas del algoritmo de clústeres de secuencias de Microsoft es que utiliza los datos de las secuencias. Estos datos suelen representar una serie de eventos o transiciones entre los estados de un conjunto de datos, como una serie de compras de productos o los clicks en web para un usuario determinado. El algoritmo examina todas las probabilidades de transición y mide las diferencias, o las distancias, entre todas las posibles secuencias del conjunto de datos con el fin de determinar qué secuencias es mejor utilizar como entradas para la agrupación en clústeres. Después de que el algoritmo ha creado la lista de secuencias candidatas, usa la información de las secuencias como entrada para el método EM (Expectation Maximization) de agrupación en clústeres.

Requisitos de un modelo de clústeres de secuencia

Una única columna key: un modelo de clústeres de secuencia requiere una clave que identifique los registros.

Una columna de secuencia: para los datos de la secuencia, el modelo debe tener una tabla anidada que contenga una columna de identificador de secuencia. El identificador de secuencia puede ser cualquier tipo de datos ordenable. Por ejemplo, puede usar el identificador de una página web, un número entero o una cadena de texto, con tal de que la columna identifique los eventos en una secuencia. Solo se admite un identificador de secuencia por cada secuencia y un tipo de secuencia en cada modelo.

Atributos opcionales no relacionados con la secuencia: el algoritmo admite la incorporación de otros atributos que no tengan que ver con las secuencias. Estos atributos pueden incluir las columnas anidadas.

Ver un modelo de clústeres de secuencia

El modelo de minería de datos que crea este algoritmo contiene descripciones de las secuencias más comunes en los datos. Para explorar el modelo, puede usar el Visor de clústeres de secuencia de Microsoft. Cuando se ve un modelo de clústeres de secuencia, Analysis Services muestra los clústeres que contienen varias transiciones. También pueden verse las estadísticas pertinentes.

Si desea obtener más detalles, puede examinar el modelo en el Visor de árbol de contenido genérico de Microsoft. El contenido almacenado para el modelo incluye la distribución para todos los valores de cada nodo, la probabilidad de cada clúster y detalles acerca de las transiciones.

7. Algoritmo de regresión lineal de Microsoft [MIC2009g]

El algoritmo de regresión lineal de Microsoft es una variación del algoritmo de árboles de decisión de Microsoft que ayuda a calcular una relación lineal entre una variable independiente y otra dependiente y, a continuación, utilizar esa relación para la predicción.

La relación toma la forma de una ecuación para la línea que mejor represente una serie de datos. Por ejemplo, la línea de la Figura 10 muestra la mejor representación lineal de los datos.

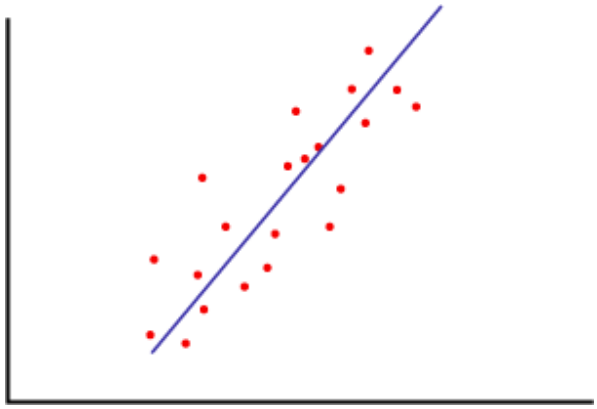


Figura 10: Línea de regresión.

Cada punto de datos del diagrama tiene un error asociado con su distancia con respecto a la línea de regresión.

Hay otros tipos de regresión que utilizan varias variables y también hay métodos no lineales de regresión. Sin embargo, la regresión lineal es un método útil y conocido para modelar una respuesta a un cambio de algún factor subyacente.

Aunque hay muchas maneras de calcular la regresión lineal que no requieren herramientas de minería de datos, la ventaja de utilizar el algoritmo de regresión lineal de Microsoft para esta tarea es que se calculan y se prueban automáticamente todas las posibles relaciones entre las variables. No tiene que seleccionar un método de cálculo, como por ejemplo para resolver los mínimos cuadrados. Sin embargo, la regresión lineal podría simplificar en exceso las relaciones en escenarios en los que varios factores afectan al resultado.

Cómo funciona el algoritmo

El algoritmo de regresión lineal de Microsoft es una variación del algoritmo de árboles de decisión de Microsoft. Al seleccionar el algoritmo de regresión lineal de Microsoft, se invoca un caso especial del algoritmo de árboles de decisión de Microsoft, con parámetros que restringen el comportamiento del algoritmo y requieren ciertos tipos de datos de entrada. Además, en un modelo de regresión lineal, el conjunto de datos completo se utiliza para calcular las relaciones en el paso inicial, mientras que en un modelo de árboles de decisión estándar los datos se dividen repetidamente en árboles o subconjuntos más pequeños.

Requisitos para los modelos de regresión lineal

Una única columna key: Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Una columna de predicción: Se requiere al menos una columna de predicción. Se pueden incluir varios atributos de predicción en un modelo, pero deben ser tipos de datos numéricos continuos. No se puede utilizar un tipo de datos de fecha y hora como atributo de predicción aunque el almacenamiento nativo para los datos sea numérico.

Columnas de entrada: Deben contener datos numéricos continuos y se les debe asignar el tipo de datos adecuado.

Ver un modelo de regresión lineal

Para examinar el modelo, puede utilizar el Visor de árboles de Microsoft. La estructura de árbol de un modelo de regresión lineal es muy simple, con toda la información sobre la ecuación de regresión contenida en un nodo único.

En un modelo de regresión lineal, el contenido incluye metadatos, la fórmula de regresión y estadísticas sobre la distribución de los valores de entrada.

8. Algoritmo de red neuronal de Microsoft [MIC2009h]

En SQL Server Analysis Services, el algoritmo de red neuronal de Microsoft combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente, puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada.

Los modelos de minería de datos construidos con el algoritmo de red neuronal de Microsoft pueden contener varias redes, en función del número de columnas que se utilizan para la entrada y la predicción, o sólo para la predicción. El número de redes que contiene un único modelo de minería de datos depende del número de estados que contienen las columnas de entrada y las columnas de predicción que utiliza el modelo.

El algoritmo de red neuronal de Microsoft es útil para analizar datos de entrada complejos, como los datos de un proceso comercial o de producción, o problemas empresariales para los que hay una cantidad importante de datos de entrenamiento disponibles pero en los que no es fácil derivar reglas mediante otros algoritmos.

Los casos sugeridos para utilizar el algoritmo de red neuronal de Microsoft son:

- Análisis de comercialización y promoción, como medir el éxito de una promoción por correo directo o una campaña publicitaria en la radio.
- Predecir los movimientos de las acciones, la fluctuación de la moneda u otra información financiera con gran número de cambios a partir de los datos históricos.
- Analizar los procesos industriales y de producción.
- Minería de texto.

- Cualquier modelo de predicción que analice relaciones complejas entre muchas entradas y relativamente pocas salidas.

Cómo funciona el algoritmo

El algoritmo de red neuronal de Microsoft crea una red formada por hasta tres niveles de neuronas. Estas capas son una capa de entrada, una capa oculta opcional y una capa de salida.

Nivel de entrada: las neuronas de entrada definen todos los valores de los atributos de entrada para el modelo de minería de datos, así como sus probabilidades.

Nivel oculto: las neuronas ocultas reciben entradas de las neuronas de entrada y proporcionan salidas a las neuronas de salida. El nivel oculto es donde se asignan pesos a las distintas probabilidades de las entradas. Un peso describe la relevancia o importancia de una entrada determinada para la neurona oculta. Cuanto mayor sea el peso asignado a una entrada, más importante será el valor de dicha entrada. Los pesos pueden ser negativos, lo que significa que la entrada puede desactivar, en lugar de activar, un resultado concreto.

Nivel de salida: las neuronas de salida representan valores de atributo de predicción para el modelo de minería de datos.

Requisitos para los modelos de red neuronal

El modelo de red neuronal debe contener una columna de clave, una o más columnas de entrada y una o más columnas de predicción.

Los modelos de minería de datos que usan el algoritmo de red neuronal de Microsoft están muy influenciados por los valores que se especifican en los parámetros disponibles para el algoritmo. Los parámetros definen cómo se muestrean los datos, cómo se distribuyen o cómo se espera que estén distribuidos en cada columna, y cuándo se invoca la selección de características para limitar los valores usados en el modelo final.

Ver un modelo de red neuronal

Para trabajar con los datos y ver cómo el modelo pone en correlación las entradas y salidas, puede usar el Visor de redes neuronales de Microsoft. Con este visor personalizado, puede filtrar los atributos de entrada y sus valores, y ver gráficamente cómo afectan a las

salidas. La información sobre herramientas del visor muestra la probabilidad y la mejora respecto al modelo predictivo asociados a cada par de valores de entrada y de salida.

La manera más fácil de explorar la estructura del modelo consiste en usar el Visor de árbol de contenido genérico de Microsoft. Este visor permite ver las entradas, las salidas y las redes creadas por el modelo, así como hacer clic en cualquier nodo para expandirlo y ver las estadísticas relacionadas con los niveles de entrada, los niveles de salida y los niveles ocultos de los nodos.

9. Algoritmo de regresión logística de Microsoft [MIC2009i]

El algoritmo de regresión logística de Microsoft es una variación del algoritmo de red neuronal de Microsoft. La regresión logística es una técnica estadística conocida que se usa para modelar los resultados binarios, como los resultados sí-no.

La regresión logística es muy flexible; puede tomar cualquier tipo de entrada y admite varias tareas analíticas diferentes:

- Usar datos demográficos para realizar predicciones sobre los resultados, como el riesgo de contraer una determinada enfermedad.
- Explorar y ponderar los factores que contribuyen a un resultado. Por ejemplo, buscar los factores que influyen en los clientes para volver a visitar un establecimiento.
- Clasificar los documentos, el correo electrónico u otros objetos que tengan muchos atributos.

Cómo funciona el algoritmo

La regresión logística es un método estadístico conocido que se usa para determinar la contribución de varios factores a un par de resultados. La implementación de Microsoft usa una red neuronal modificada para modelar las relaciones entre las entradas y los resultados. Se mide el efecto de cada entrada en el resultado y se ponderan las diversas entradas en el modelo acabado. El nombre regresión logística procede del hecho de que la curva de los datos se comprime mediante una transformación logística para minimizar el efecto de los valores extremos.

Requisitos para los modelos de regresión logística

Una columna clave: cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Columnas de entrada: cada modelo debe tener al menos una columna de entrada que contenga los valores que se utilizan como factores en el análisis. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el modelo.

Al menos una columna de predicción: el modelo debe contener al menos una columna de predicción de cualquier tipo de datos, incluidos datos numéricos continuos. Los valores de la columna de predicción también se pueden tratar como entradas del modelo, o se puede especificar que sólo se utilicen para las predicciones. No se admiten tablas anidadas en las columnas de predicción, pero se pueden usar como entradas.

Ver un modelo de regresión logística

Para explorar el modelo, puede usar el Visor de redes neuronales de Microsoft o el Visor de árbol de contenido genérico de Microsoft.

Cuando se ve el modelo con el Visor de redes neuronales de Microsoft, Analysis Services muestra los factores que contribuyen a un resultado determinado, clasificados por su importancia. Se puede elegir un atributo y los valores que se desea comparar.

Referencias

[MIC2009a] MICROSOFT CORPORATION. *"Algoritmo de árboles de decisión de Microsoft"*, Mayo 2009. [Fecha de Consulta: 06 de Julio de 2009]. Disponible en: http://technet.microsoft.com/es-es/library/ms175312.aspx

[MIC2009b] MICROSOFT CORPORATION. *"Algoritmo Bayes naive de Microsoft"*, Mayo 2009. [Fecha de Consulta: 06 de Julio de 2009]. Disponible en: http://technet.microsoft.com/es-es/library/ms174806.aspx

[MIC2009c] MICROSOFT CORPORATION. *"Algoritmo de clústeres de Microsoft"*, Mayo 2009. [Fecha de Consulta: 06 de Julio de 2009]. Disponible en: http://msdn.microsoft.com/es-es/library/ms174879.aspx

[MIC2009d] MICROSOFT CORPORATION. *"Algoritmo de serie temporal de Microsoft"*, Mayo 2009. [Fecha de Consulta: 07 de Julio de 2009]. Disponible en: http://msdn.microsoft.com/es-es/library/ms174923.aspx

[MIC2009e] MICROSOFT CORPORATION. *"Algoritmo de asociación de Microsoft"*, Mayo 2009. [Fecha de Consulta: 07 de Julio de 2009]. Disponible en: http://msdn.microsoft.com/es-es/library/ms174916.aspx

[MIC2009f] MICROSOFT CORPORATION. *"Algoritmo de agrupación en clústeres de secuencia de Microsoft"*, Mayo 2009. [Fecha de Consulta: 07 de Julio de 2009]. Disponible en:

http://msdn.microsoft.com/es-es/library/ms175462.aspx

[MIC2009g] MICROSOFT CORPORATION. *"Algoritmo de regresión lineal de Microsoft"*, Mayo 2009. [Fecha de Consulta: 07 de Julio de 2009]. Disponible en: http://msdn.microsoft.com/es-es/library/ms174824.aspx

[MIC2009h] MICROSOFT CORPORATION. *"Algoritmo de red neuronal de Microsoft"*, Mayo 2009. [Fecha de Consulta: 07 de Julio de 2009]. Disponible en: http://msdn.microsoft.com/es-es/library/ms174941.aspx

[MIC2009i] MICROSOFT CORPORATION. *"Algoritmo de regresión logística de Microsoft"*, Mayo 2009. [Fecha de Consulta: 07 de Julio de 2009]. Disponible en: http://msdn.microsoft.com/es-es/library/ms174828.aspx