Estableciendo Objetivos de Minería de Datos

El primer paso en la minería de datos requiere que establezcas objetivos para el ejercicio. Obviamente, debes identificar las preguntas clave que necesitan ser respondidas. Sin embargo, más allá de identificar las preguntas clave, están las preocupaciones sobre los costos y beneficios del ejercicio. Además, debes determinar, de antemano, el nivel esperado de precisión y utilidad de los resultados obtenidos de la minería de datos. Si el dinero no fuera un problema, podrías destinar tantos fondos como fuera necesario para obtener las respuestas requeridas. Sin embargo, el intercambio costo-beneficio siempre es fundamental para determinar los objetivos y el alcance del ejercicio de minería de datos. El nivel de precisión esperado de los resultados también influye en los costos. Niveles altos de precisión en la minería de datos costarían más y viceversa. Además, más allá de cierto nivel de precisión, no obtienes mucho del ejercicio, dado el rendimiento decreciente. Así, los intercambios costo-beneficio para el nivel de precisión deseado son consideraciones importantes para los objetivos de minería de datos.

Selección de Datos

La salida de un ejercicio de minería de datos depende en gran medida de la calidad de los datos que se utilizan. A veces, los datos están fácilmente disponibles para su posterior procesamiento. Por ejemplo, los minoristas a menudo poseen grandes bases de datos de compras de clientes y demografía. Por otro lado, los datos pueden no estar fácilmente disponibles para la minería de datos. En tales casos, debes identificar otras fuentes de datos o incluso planificar nuevas iniciativas de recolección de datos, incluidas encuestas. El tipo de datos, su tamaño y la frecuencia de recolección tienen un impacto directo en el costo del ejercicio de minería de datos. Por lo tanto, identificar el tipo correcto de datos necesarios para la minería de datos que pueda responder a las preguntas a costos razonables es fundamental.

Preprocesamiento de Datos

El preprocesamiento de datos es un paso importante en la minería de datos. A menudo, los datos en bruto son desordenados, conteniendo datos erróneos o irrelevantes. Además, incluso con datos relevantes, a veces falta información. En la etapa de preprocesamiento, identificas los atributos irrelevantes de los datos y eliminas tales atributos de consideración futura. Al mismo tiempo, es necesario identificar los aspectos erróneos del conjunto de datos y marcarlos como tales. Por ejemplo, el error humano podría llevar a la fusión inadvertida o al análisis incorrecto de la información entre columnas. Los datos deben ser objeto de verificaciones para

garantizar su integridad. Por último, debes desarrollar un método formal para tratar los datos faltantes y determinar si los datos faltan de manera aleatoria o sistemática.

Si los datos faltaran de forma aleatoria, un conjunto simple de soluciones sería suficiente. Sin embargo, cuando los datos faltan de manera sistemática, debes determinar el impacto de los datos faltantes en los resultados. Por ejemplo, un subconjunto particular de individuos en un gran conjunto de datos puede haber rechazado divulgar su ingreso. Los hallazgos que dependen del ingreso de un individuo como entrada excluirían los detalles de aquellos individuos cuyo ingreso no fue reportado. Esto llevaría a sesgos sistemáticos en el análisis. Por lo tanto, debes considerar de antemano si las observaciones o variables que contienen datos faltantes deben ser excluidas de todo el análisis o de partes del mismo.

Transformación de Datos

Después de que se han retenido los atributos relevantes de los datos, el siguiente paso es determinar el formato apropiado en el que se deben almacenar los datos. Una consideración importante en la minería de datos es reducir el número de atributos necesarios para explicar los fenómenos. Esto puede requerir la transformación de datos. Los algoritmos de reducción de datos, como el Análisis de Componentes Principales (demostrado y explicado más adelante en el capítulo), pueden reducir el número de atributos sin una pérdida significativa de información. Además, puede ser necesario transformar variables para ayudar a explicar el fenómeno que se está estudiando. Por ejemplo, el ingreso de un individuo puede registrarse en el conjunto de datos como ingreso salarial; ingresos de otras fuentes, como propiedades de alquiler; pagos de apoyo del gobierno, y similares. Agregar los ingresos de todas las fuentes desarrollará un indicador representativo para el ingreso individual.

A menudo, es necesario transformar variables de un tipo a otro. Puede ser prudente transformar la variable continua de ingresos en una variable categórica donde cada registro en la base de datos se identifique como individuo de ingresos bajos, medianos y altos. Esto podría ayudar a capturar las no linealidades en los comportamientos subyacentes.

Almacenamiento de Datos

Los datos transformados deben almacenarse en un formato que facilite la minería de datos. Los datos deben almacenarse en un formato que otorgue privilegios de lectura/escritura inmediatos y sin restricciones al científico de datos. Durante la minería de datos, se crean nuevas variables, que se escriben de nuevo en la base de

datos original, por lo que el esquema de almacenamiento de datos debe facilitar la lectura y escritura eficientes en la base de datos. También es importante almacenar los datos en servidores o medios de almacenamiento que mantengan la seguridad de los datos y prevengan que el algoritmo de minería de datos busque innecesariamente fragmentos de datos dispersos en diferentes servidores o medios de almacenamiento. La seguridad y privacidad de los datos deben ser una preocupación principal al almacenar datos.

Minería de Datos

Después de que los datos son procesados, transformados y almacenados adecuadamente, son objeto de minería de datos. Este paso abarca métodos de análisis de datos, incluidos métodos paramétricos y no paramétricos, así como algoritmos de aprendizaje automático. Un buen punto de partida para la minería de datos es la visualización de datos. Las vistas multidimensionales de los datos utilizando las avanzadas capacidades de gráficos del software de minería de datos son muy útiles para desarrollar una comprensión preliminar de las tendencias ocultas en el conjunto de datos.

Evaluación de Resultados de Minería de Datos

Después de que se han extraído los resultados de la minería de datos, se realiza una evaluación formal de los resultados. La evaluación formal podría incluir probar las capacidades predictivas de los modelos en datos observados para ver cuán efectivos y eficientes han sido los algoritmos en reproducir datos. Esto se conoce como un “pronóstico en muestra”. Además, los resultados se comparten con las partes interesadas clave para obtener comentarios, que luego se incorporan en las iteraciones posteriores de la minería de datos para mejorar el proceso.

La minería de datos y la evaluación de los resultados se convierten en un proceso iterativo de tal manera que los analistas utilizan algoritmos mejores y mejorados para mejorar la calidad de los resultados generados a la luz de los comentarios recibidos de las partes interesadas clave.