SEGUNDA DOCUMENTACIÓN PARCIAL (REV.2)

CIR-REQ-0011 STNG

Septiembre 2019

Árbol de decisión

El método original intenta abarcar una serie de restricciones predefinidas en un archivo Excel siguiendo un orden parcialmente respetado. Si algún resultado o conjunto de resultados viola alguna de las lógicas, se envía un mensaje y se modifican ciertos campos específicos en la vista del analista.

Un problema con este método consiste en la violación de lógicas por parte de resultados superfluos o de poco interés, provocando que el analista destine tiempo a añadir una lógica "normal" personalizada para que dicha combinación se ignore en un futuro escenario. Así también, el proceso manual de incorporación de nuevas lógicas involucra un retardo de tiempo significativo ya que se debe hacer vía solicitud a un encargado externo.

Cada lógica contiene evaluaciones basadas en umbrales como los mostrados en la Tabla 1.

Componente	Normal	Advertencia	Alarma
Al	< 20	20-40	> 40
Cr	< 10	10-20	> 20
Cu	< 50	50-100	> 100
Fe	< 200	200-300	> 300
Si	< 15	15-25	> 25

Table 1: Ejemplo de lógicas en árbol de decisión

Análisis Preliminar de la Estructura de los Datos

El análisis de la información entregada por medio de la base de datos es enfocado principalmente en la extracción de la información relacionada a los ensayos de tribología realizados por la empresa a equipos de diferentes clientes. Las variables de interés observadas a partir de las múltiples tablas presentes en la base de datos, y que se considera serán parte relevante de la alimentación de datos para el sistema de aprendizaje automático son las siguientes :

Datos propios de cada muestra-ensayo :

- **id_cliente** : Valor único correspondiente al cliente que solicita la muestra.
- **id_faena** : Valor único, propio de cada cliente, correspondiente a la faena desde donde proviene la muestra.
- **id_tipo_equipo** : Valor correspondiente al tipo de equipo desde donde proviene la muestra.
- **id_tipo_componente** : Valor correspondiente al tipo de componente desde donde proviene la muestra.
- **id_componente** : Valor único, propio de cada cliente y faena, correspondiente al componente desde donde proviene la muestra.

- **correlativo_muestra** : valor identificador de la muestra desde donde proviene el ensayo realizado.
- **id_ensayo** : Valor identificador del ensayo realizado.
- **valor** : valor correspondiente al ensayo de la muestra analizada.
- **id_protocolo** : Valor del protocolo usado para analisis del valor del ensayo determinado por el cliente.
- **m_fecha_muestreo_annio** : Fecha de muestreo correspondiente al *año*
- **m_fecha_muestreo_mes** : Fecha de muestreo correspondiente al *mes*
- **m_fecha_muestreo_dia** : Fecha de muestreo correspondiente al *día*
- **m_fecha_ingreso_annio** : Fecha de ingreso de la muestra correspondiente al *año*
- **m_fecha_ingreso_mes** : Fecha de ingreso de la muestra correspondiente al *mes*
- **m_fecha_ingreso_dia** : Fecha de ingreso de la muestra correspondiente al *día*

La Tabla 2 muestra el origen desde que tabla de la base de datos se obtiene el campo respectivo.

Dato	Tabla (en BD original en MySQL)
id_cliente	*trib_precalculo_reporte*
id_faena	*trib_precalculo_reporte*
id_tipo_equipo	*trib_precalculo_reporte*
id_tipo$_c$omponente	*trib_precalculo_reporte*
id_componente	*trib_muestra*
correlativo_muestra	*trib_muestra*
id_ensayo	*trib_resultado*
valor	*trib_resultado*
id_protocolo	*trib_resultado*

Table 2: Relación entre datos y su origen en BD

Datos propios de cada ensayo, obtenidos de la tabla *trib_ensayo* :

- **id_ensayo** : Identificador propio del ensayo realizado.
- **cp_3_tipo_protocolo** : Tipo de ensayo al que corresponde (metal o lubricante).
- **nombre** : Nombre del ensayo.

Datos propios de cada protocolo, obtenidos de la tabla *trib_protocolo* :

- **id_protocolo** : Identificador propio del protocolo realizado.
- **nombre** : Nombre del protocolo.

Datos propios de cada ensayo con su protocolo determinado según el cliente, obtenidos de la tabla *trib_ensayo_protocolo* :

- **id_protocolo** : Identificador del protocolo.
- **id_ensayo** : Identificador del ensayo.
- **lim_inf_marginal** : Límite inferior marginal del ensayo correspondiente al protocolo.
- **lim_sup_marginal** : Límite superior marginal del ensayo correspondiente al protocolo.
- **lim_inf_condenatorio** : Límite inferior condenatorio del ensayo correspondiente al protocolo.
- **lim_sup_condenatorio** : Límite superior condenatorio del ensayo correspondiente al protocolo.

Segmentación de Datos

Con el objetivo de analizar los datos y relacionar resultados de ensayos se comienza segmentando los datos que son posibles de correlacionar.

La primera variable identificada para la segmentación de los resultados de los ensayos es *id_faena*, debido a que los valores obtenidos para cada ensayo dependen directamente de las condiciones de trabajo propias de cada faena. Debido a que la variable *id_faena* es única (no compartida entre clientes), se identifica que el valor de la variable *id_cliente* se vuelve solamente informativa y no relevante para la toma de decisiones en base a los datos.

Posterior a la primera segmentación es posible segmentar nuevamente los resultados por otras tres variables :.

- "id_tipo_equipo"
- "id_tipo_componente"
- "id_componente"

Al hacer un primer análisis estadístico usando solo los datos provistos, se identifican los tipos de ensayo que presentan una mayor correlación entre sí, como se muestra en la Tabla 3 donde se listan los 5 pares de correlación más altos, y en la Figura 1 donde se ilustra un mapa de calor donde la intensidad del color se asocia con la cercanía a 1 y es directamente proporcional al grado de correlación, valor que naturalmente se observa máximo en la diagonal.

Ensayo 1	Ensayo 2	Correlación (%)
pH	vanadium	99.9965
chromium	pH	99.9964
acid_total_number	*basic_total_number*	98.8107
antifreeze	vanadium	97.9971
acid_total_number	*dilution_by_fuel*	97.9966

Table 3: Mayor correlación entre ensayos

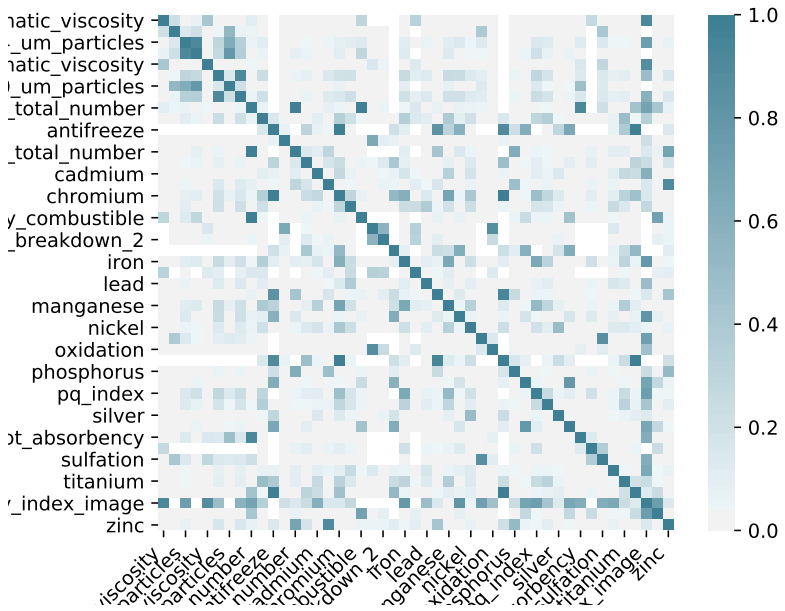

Figure 1: Mapa de calor con correlación de Pearson

PROPUESTA DE DESARROLLO DE INTELIGENCIA ARTIFICIAL

Con el propósito de entregar un resultado más completo al cliente que solicita el análisis de una muestra, es que se propone la incorporación de mecanismos de predicción de eventos futuros a partir del análisis de resultados acumulados de muestras históricas.

Luego de observar los datos disponibles presentes en la base de datos de tribología entregados, es que se generan las siguientes propuestas de posibles enfoques en los que se puede centrar el problema de aprendizaje automático.

Predicción de eventos y estados futuros de un componente

Por medio del uso de los resultados históricos acumulados de las muestras de un componente analizado es posible estimar, a partir de una nueva muestra, posibles eventos y estados futuros de dicho componente, permitiendo enriquecer el análisis de la nueva muestra con mecanismos de corrección temprana de fallos o estados no deseados.

El resultado de este análisis adicional de datos incorporará los siguientes campos:

- **alert_level_pred** : Predicción del nivel de alerta de una muestra futura (“Normal”, “Warning”, “Alert”)
- **id_state_pred** : Comentario (o descripción del “estado” futuro)
- **id_suggestion_pred** : Acción (o “sugerencia” sobre qué acción tomar para evitar dicho “estado” futuro)

La información adicional entregada al cliente permitirá evitar futuros problemas a presentarse en sus equipos, anticipándose mediante comentarios

de corrección preventiva en base a un mecanismo automático que apoye la toma de decisiones.

Con el objetivo de validar experimentalmente la calidad de los datos para este tipo de tarea, se prueba un clasificador tradicional mediante el algoritmo KNN (K-Nearest Neighbors). El objetivo planteado es predecir el nivel de alerta en un siguiente instante de tiempo, dada la información provista como entrada de la caracterización completa de ensayos realizados y su nivel de alerta correspondiente en instantes anteriores.

Se destaca que no se espera un desempeño cercano al óptimo para este algoritmo, por razones como:

- Existen alternativas que han probado tener mejor desempeño en tareas de clasificación, como las que se consideran como propuesta (LSTM por ejemplo, Long Short-Term Memory neural networks).
- Al haber cambios no informados en componentes de las muestras ensayadas, es posible que se produzca dualidad en la determinación del estado de alerta.
- Dado que cada cliente tiene protocolos distintos, se espera que este clasificador KNN no sea capaz de identificar la dualidad en nivel de alerta que podría tener para muestras similares.

Estimación de modificaciones no informadas de equipos o componentes

Un problema que se observa en el proceso actual es la dificultad del análisis de los resultados de las muestras históricas de un componente, debido a la no entrega completa de información por parte del cliente que solicita el análisis de una muestra.

En la mayoría de los casos, no es compartida la información relacionada con cambios en el estado físico de los equipos, como lo son cambios de componentes y relleno o renovación de lubricantes, afectando la predicción de la propuesta anterior y el análisis de la nueva muestra.

De presentarse una estimación positiva en la modificación del estado físico de un componente, se incorporará en la información incompleta entregada por el cliente facilitando el análisis de la nueva muestra y permitirá predecir correctamente posibles eventos y estados futuros.
El enriquecimiento de lo resultados para las muestras, producto de las propuestas que han sido expuestas anteriormente, únicamente conllevan un análisis computacional con un costo adicional asociado mínimo pero de gran valor agregado al servicio ofrecido al cliente. El costo monetario adicional asociado dependerá de los requerimientos de hardware computacional involucrados en la predicción.

Si los requerimientos para la predicción son posibles de solventar en máquinas computacionales locales, entonces no existirá costo adicional alguno. De no ser así, el costo del análisis se ve aumentado por los costos asociados al servicio AWS (Amazon Web Services) con características apropiadas para la predicción según el volumen de datos históricos en procesamiento.

Observaciones

Todo el código generado durante esta etapa se puede revisar en Github: `https://github.com/astng/module_ai` . Al momento de la fecha de entrega de este documento, se ha replicado el mecanismo de toma de decisiones original en forma de restful API como servicio HTTP, cuya documentación

de uso está disponible en una Wiki en https://github.com/astng/module_ai/wiki/Module-AI-API-Documentation