UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria Informàtica
Universitat Politècnica de València

Simulación de datos de sensores industriales

TRABAJO FIN DE GRADO

Grado en Ingeniería Informática

Autor: Pedro Henrique Mano Figueiredo Fernandes
Tutor: Francisco Sánchez Cid

Curso 2015-2016

Resumen

Los entornos industriales hacen uso de sensores para obtener mediciones de varios factores en su cadena de producción. En las fases iniciales, la cantidad de datos generados es pequeña, por lo que no es significativa para poder aplicar técnicas de Big Data. El proyecto descrito en este documento busca simular mayor cantidad de datos a través de los pocos datos generados por sensores industriales. La primera tarea en análisis de datos es la ingestión del dataset, que en este caso no tiene mayor dificultad técnica al tratarse de un dataset pequeño. Posteriormente se procede a la respectiva limpieza, para normalizar los datos y así poder aplicar cálculos precisos sobre los mismos.

Las técnicas para obtener conocimiento del dataset que se proponen son del ambito de Machine Learning. En este caso se tratan de técnicas de aprendizaje no supervisado (unsupervised learning), pues el dataset no contiene características con etiqueta. Se pretende obtener nuevo conocimiento de los datos al aplicar la reducción de dimensionalidad, a través de un método conocido por Principal Component Analysis o simplemente PCA. Este método transforma espacios multidimensionales en espacios ortogonales más reducidos, proyectando los datos sin perder información. Esto permite tener mayor claridad sobre los datos, mayor capacidad de visualización y reducción de ruido. Con los datos proyectados en el nuevo espacio ortogonal, el objetivo es generar nuevos datos y luego re-proyectar al espacio original, con todas las características.

La simulación de datos se hace a través de muestreo con distribución normal para cada componente principal. Se usan los restantes componentes para generar ruido. Como hay una característica de tipo fecha, tiene sentido usar técnicas basadas en séries temporales, para aprender la regularidad de generación de las muestras. Se propone la aplicación de modelos ARIMA para el aprendizaje de la série temporal.

Una vez generados los nuevos datos, se aplican técnicas estadísticas para evaluar su concordancia con los datos iniciales. El estadístico Hotelling T-squared ofrece un test de las médias multivariantes de diferentes poblaciones. Y el Box's M test evalua la homogeneidad de matrices de varianzas-covarianzas. De esta forma se comparan los dos espacios de datos, el original y el simulado.

Palabras clave: machine learning, big data, PCA

CAPÍTULO 1
Introducció

Este proyecto tiene como fuente de datos uno o varios sensores de un mecanizado de injección de plástico. Los sensores efectuan mediciones de varios factores con una regularidad temporal, generando así muestras de datos en determinados intervalos de tiempo. En este caso de estudo, la frecuencia de las mediciones no es constante y existen varias interrupciones, constituyendo un dataset con poco volumen. Las técnicas de Big Data, en particular Machine Learning, necesitan gran cantidad de datos para poder inducir modelos matemáticos que los expliquen. Cuanto más datos mejor, lo que garantiza un aprendizaje más robusto y fiable.

La fase inicial del análisis es la ingesta de los datos. El dataset proporcionado es de reducida dimensión, tiene tan solo 5MB, por lo que la tarea de ingestión no es intensiva. Sin embargo, hay que tratar los datos en bruto antes de empezar a usarlos. El fichero de datos es de texto, así que hay que hacer determinadas conversiones para poder usar tipos más específicos, como sean fechas y números. El primer problema tiene que ver con los formatos de fecha, que no son correctos para el locale España, lo que exige una adaptación de los algoritmos de parsing. La solución propuesta para simular datos empieza con la proyección de los datos del espacio multivariante original a un espacio ortogonal más reducido. Esta técnica de Machine Learning se conoce por PCA (Principal Component Analysis). PCA pertenece a la familia de aprendizaje no supervisado (unsupervised learning), donde no existen etiquetas o clases a predecir. Los datos de entrenamiento son simplemente observaciones, sin estar clasificados. Las técnicas de unsupervised learning se usan para descubrir grupos de similitud (clusters) en los datos; o para determinar la distribución de los datos en el espacio original; o también, que interesa a este estudio, para proyectar los datos en espacios de menos dimensiones.

Pocas dimensiones suelen ofrecer mejor insight sobre los datos. Como primera ventaja, permiten la aplicación de técnicas de visualización - resulta muy difícil visualizar datos en más de 3 dimensiones (incluso en 3 dimensiones puede ser difícil). Además, las componentes que no son principales suelen ser ruido o redundancia - con esta técnica se pueden atenuar los efectos de estos. La simulación de datos puede usar la ventaja de la eliminación de ruido y redundancia, generando así datos más relevantes.

PCA decompone el dataset en una série de componentes ortogonales que explican la variabilidad. Por ejemplo, si dos variables son directamente proporcionales, no hace falta tener en cuenta las dos para visualizar, porque conllevaría una complejidad inecesaria. PCA está muy relacionado con una técnica matemática llamada Singular Value Decomposition (SVD), tanto que muchas veces los nombres se usan intercambiados. De hecho, el algoritmo de PCA de scikit-learn usa la decomposición SVD de numpy. Sin embargo, el cálculo de SVD es muy intensivo para matrices grandes, porque calcula la matriz de varianzas-covarianzas para todos los componentes. Esto implica un gran consumo de

memoria y CPU. Como alternativa, se propone usar una técnica iterativa llamada NIPALS (Nonlinear Iterative Partial Least Squares), que usa el numero de componentes reducido.

El nuevo espacio está constituido por variables independientes y ortogonales. Para cada variable (componente), se calculan valores aleatorios con distribución normal. Estos son los scores de las componentes de los nuevos datos. Al proyectar la nueva matriz de componentes de vuelta al espacio original, se obtienen datos simulados para todas las características. Si la validación de esta técnica demuestra que es correcta, se puede usar para la simulación de nuevos datos. La validación de los datos generados pasa por la aplicación de estadísticos de comparación de las muestras originales y las nuevas. Las técnicas utilizadas son los tests de Hotelling T-squared y Box's M. En el caso de Hotelling T-squared (o T2) se prueba la igualdad en las médias de las poblaciones. Box's M test tiene como suposición la igualdad de matrices de varianzas-covarianzas de las poblaciones. Los resultados demuestran que es viable usar esta técnica de simulación de datos.

1.1 Motivación

TODO

1.2 Objectivos

TODO

1.3 Estructura de la memória

TODO

CAPÍTULO 2
Dataset

TODO

2.1 Limpieza

TODO

CAPÍTULO 3
Base teórica

TODO

CAPÍTULO 4
Desarrollo

TODO

CAPÍTULO 5

Evaluación de los datos simulados

TODO

CAPÍTULO 6
Conclusión

TODO

Bibliografía

[1] Jon Shlens. A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS. Derivation, Discussion and Singular Value Decomposition. Version 1, 25 March, 2003.

[2] Unknown Authors. The Truth about Principal Components and Factor Analysis. 28 September, 2009.

[3] Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006.

[4] Principal component analysis (PCA). Consultar http://scikit-learn.org/stable/modules/decomposition.html#principal-component-analysis-pca.

APÉNDICE A
Configuración del sistema

TODO

A.1 Fase de inicialización

TODO