ABSTRACT

A la hora de pronosticar el crecimiento del throughput y el uso de recursos, a menudo se tiende a extrapolar los percentiles superiores de las distribuciones de los datos. Este tipo de aproximación funciona correctamente en situaciones lejanas al punto de saturación del sistema. Sin embargo, cuando la utilización de los recursos se acerca a la capacidad máxima del sistema, los percentiles superiores tienden a nivelarse ("flattopping" o "clipping"), subestimando el crecimiento real y dando lugar a predicciones y asignación de recursos ineficientes. En el siguiente informe se explica este suceso, así como se propone una nueva aproximación con el objetivo de realizar predicciones útiles en base a históricos de datos y la detección de si un recurso se está aproximando a su punto de saturación.

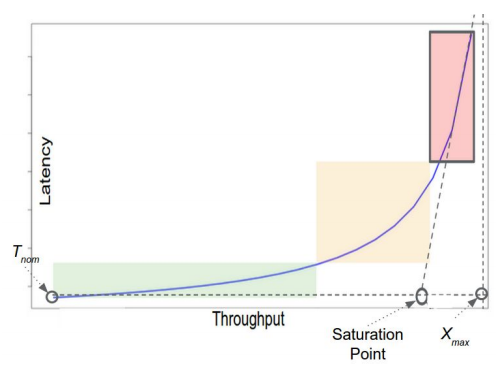
WORKLOAD

La carga de trabajo de un recurso IT, se define normalmente en términos como el número de comandos, peticiones, trabajos… que son o bien procesados o mantenidos en una cola de llegada. La ley de Little’s, conecta el workload, tasa de llegada, y el tiempo de servicio en una simple ecuación. W=X\*T donde X=Tasa de llegada, W=Workload y T=tiempo de servicio.

Las variables X y T, describen dos características del sistema, X caracteriza la demanda, mientras la latencia T caracteriza el tiempo de respuesta del sistema.

A medida que recogemos datos de rendimiento y latencia en el tiempo, obtenemos dos series temporales de mediciones

X(t) y T(t), que juntos definen una serie temporal de carga de trabajo W(t). Con una baja tasa de llegada la dependencia de T(t) de X(t) puede ser tratada como insignificante. Pero cuando el recurso se acerca a la saturación, observamos un punto de saturación en la Curva Operativa del Receptor (ROC).



En el punto donde termina la zona verde y comienza la zona amarilla la tasa de llegada y el tiempo de respuesta se vuelven significativamente independientes. Este tipo de comportamiento con un punto de saturación implican varias consideraciones prácticas. Una de ellas es que, teniendo en cuenta los tiempos necesarios para asignar los recursos o instalar las infraestructuras necesarias para aumentar la capacidad del sistema, imponen la necesidad de predecir el comportamiento en un momento muy lejano en el futuro. La segunda es que a menudo se buscan puntos de utilización por debajo del de saturación, teniendo que comparar el comportamiento normal del recurso (sin saturación) con situaciones de la zona amarilla y roja de la imagen (cerca de saturación o saturadas).

ENFOQUE ESTÁNDAR

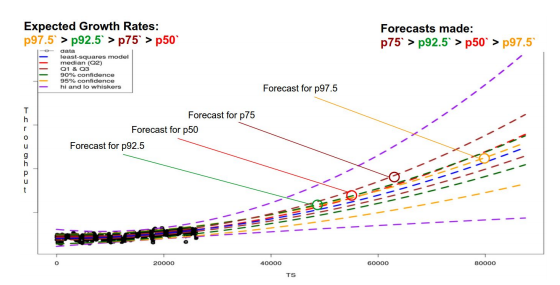
El enfoque estándar consta en la utilización de un histórico de datos, coger el 95th percentil y calcular W95(i) [[1]](#footnote-1)de la serie, seguidamente, predecir el W95(i)[[2]](#footnote-2). Por último, identificar el punto donde W95(i) ≥ N, con N valor del umbral máximo.

En el enfoque estándar la latencia no cambiará a mayor throughput, lo que implica que la trayectoria del throughput será una buena aproximación a la trayectoria del workload, y el pronóstico del workload se definirá por el throughput: W(t) = X(t) \* T. Además, las capacidades de los recursos de TI se miden normalmente en unidades de throughput (número de transacciones por segundo; bits por segundo; operaciones enteras por segundo; etc.), lo que hace que conveniente para medir el workload como la tasa de servicio de las unidades de trabajo que llegan.

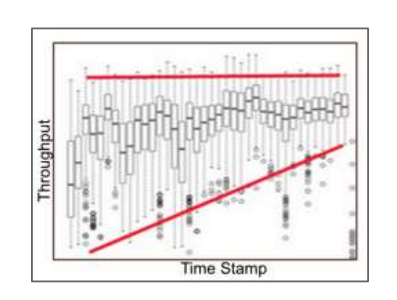
Este enfoque no tiene en cuenta el punto de saturación del workload descrito en el apartado anterior.

El enfoque estándar funciona correctamente cuando la trayectoria de los percentiles se comporta como “esperamos”, es decir W95(i)>W50(i)>W10(i) y W95(i)>W50(i)>W10(i) o X95(i)>X50(i)>X10(i) y X95(i)>X50(i)>X10(i)

Pero esto no siempre es así, en la siguiente ilustración se puede apreciar un ejemplo en el que las predicciones a futuro de los diferentes percentiles tienen un comportamiento distinto al esperado.

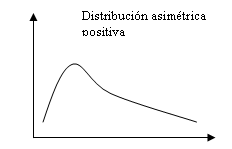
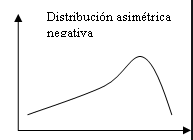
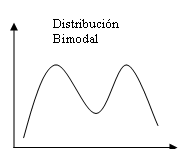


A medida que aumenta el throughput, la latencia solo puede aumentar, mientras que a medida que aumenta la latencia, el throughput solo puede disminuir. Como conclusión, al aumentar el throughput en un sistema cerrado, sus percentiles superiores deben crecer a un ritmo más lento, cerca del punto de saturación, que los percentiles más bajos; de ahí los patrones observados en la figura.

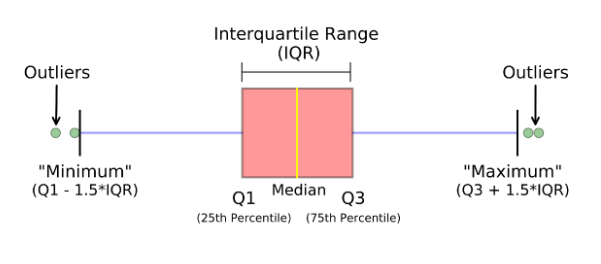


DETECCION DE CONGERSTION

El throughput no sigue una distribución normal. En un sistema sin congestión, su distribución es generalmente asimétricamente positiva. En situaciones de congestión la distribuciçón tiende a convertirse bimodal o incluso asimetrica negativa. El grado de asimetria de una distribución se puede calcular en función de los cuantiles de la distribución.



El método IQR Tukey genera intervalos de confianza para la detección de valores atípicos, el método hace uso del primer cuantil Q1 (25%), la mediana Q2 y el tercer cuantil Q3 (75%) para calcular los outlier superiores e inferiores. El uso del rango intercuartílico para la detección de outliers en lugar del percentil 95, como se hacía en la aproximación estándar, nos da resultados más fieles en situaciones con una distribución no normal, sobre todo en distribuciones multimodales en las que el percentil 95 dejaría fuera a una gran cantidad de datos no atípicos en la distribución.



PREDECIR EL CRECIMIENTO

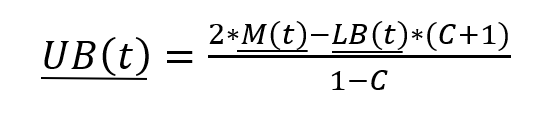
El cálculo de la asimetría de la distribución se puede obtener mediante los cuantiles Q1, Q2 y Q3 anteriormente descritos mediante la formula

Si el throughput está creciendo, podemos utilizar un tiempo pasado, cuando el sistema no estaba congestionado, para calcular la asimetría de la distribución y usarla para estimar su crecimiento. La asimetría es una propiedad que es distinta de los demás momentos (media, varianza, y curtosis). Se puede decir que es una propiedad de la distribución y se conservará a menos que el sistema se vea congestionado.

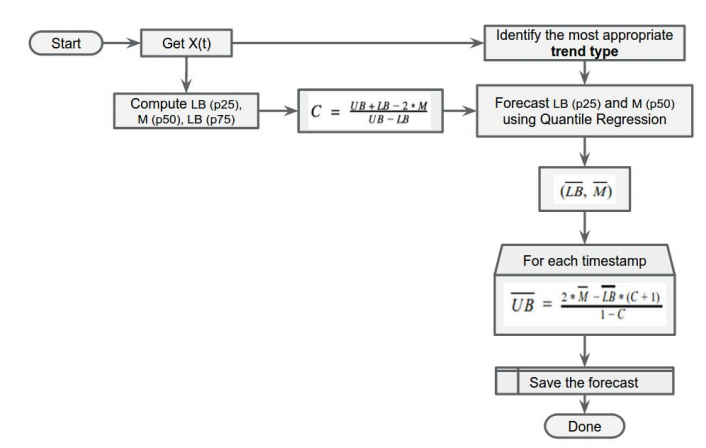
Al tratarse de una serie temporal podemos utilizar intervalos de tiempo para calcular la asimetría histórica de la serie temporal (C) mediante la fórmula:

Es natural suponer que la asimetría medida variará de un intervalo de tiempo a otro por lo que suponemos además que la asimetría, o al menos su mediana, puede tratarse como estacionaria por ser el throughput estacionario. La estacionariedad se perderá durante la transición hacia y desde el estado restringido; sin embargo, tales transiciones tienden a ocurrir indetectablemente rápido en los períodos de datos típicamente usados en los pronósticos (horas o días en transición vs. meses o años de datos históricos).

Por cada punto en el futuro t del horizonte de predicción, usamos regresión por cuantiles para calcular UB(t) (percentil 75 en función del parámetro C y los percentiles 25 y 50).



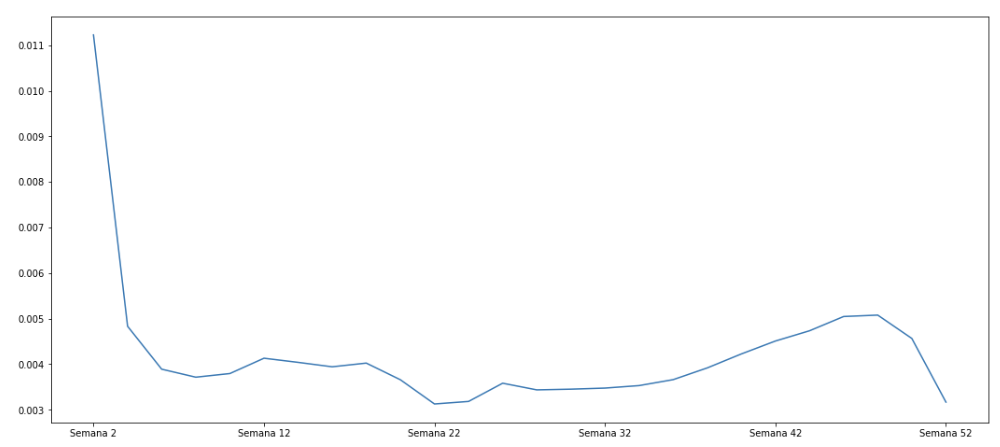
Con este valor de UB(t) y el método IQR Tukey podemos calcular los outlier a futuro que utilizar en la previsión en lugar del percentil 95.

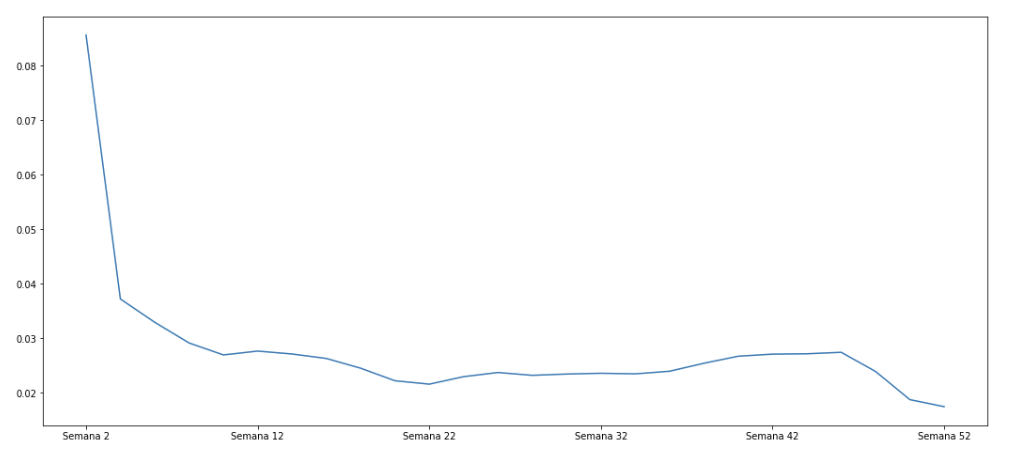
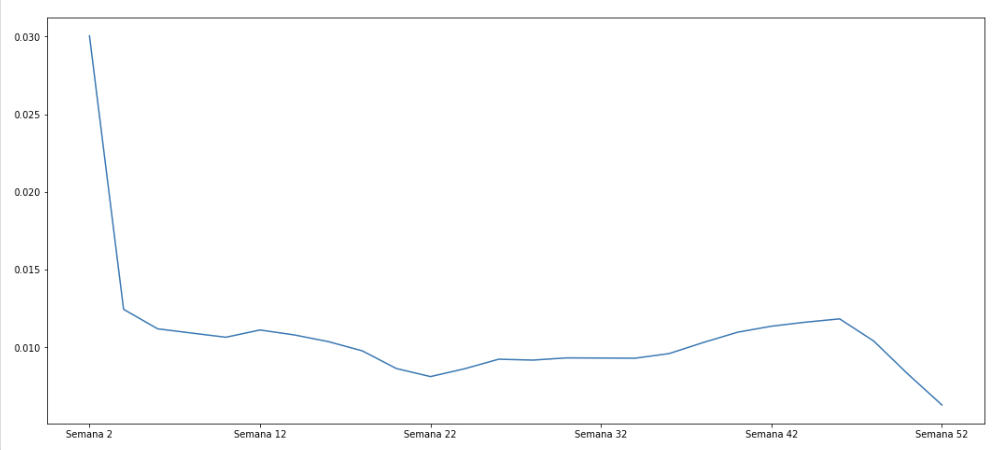


PRUEBAS Y RESULTADOS

El tamaño del histórico de datos disponible afecta en gran medida en el cálculo de las regresiones de los distintos percentiles en sus predicciones, por ejemplo, cogiendo un histórico de una semana, la disminución del tráfico en el fin de semana daría una previsión decreciente del tráfico, dando lugar a pendientes negativas en los distintos percentiles. Por otro lado, el uso de un histórico de datos excesivamente grande puede conllevar el suavizado de las pendientes en situaciones en las que el tráfico varíe en gran medida, por el peso de los datos anteriores.

Para evitar estas situaciones se ha calculado los errores obtenidos sobre datos reales (calculando sus percentiles reales por días) frente a los obtenidos mediante el forecast de la regresión por cuantiles[[3]](#footnote-3).





Los algoritmos AIC y BIC se utilizan para comparar la complejidad y los resultados de dos modelos distintos y decidir sobre eso si merece la pena un algoritmo más complejo que uno simple. En nuestro caso habría que decidir una métrica relativa al método (MASE MAE etc ... ) pero estamos ante el problema de que no es un forecast como tal, más bien una predicción de tendencias. O bien usar una métrica relativa a la aplicación, es decir, si el modelo satisface para lo que ha sido diseñado, con lo que no quedaría otra que implementarlo y hacer pruebas

1. Donde i = intervalo de tiempo sobre el que el percentil es calculado [↑](#footnote-ref-1)
2. Los elementos subrayados indican en el futuro. Es decir, en este caso indica la previsión a futuro del percentil 95. [↑](#footnote-ref-2)
3. Los percentiles diarios seguirán un comportamiento parecido al del tráfico, mientras que el forecast realizado se aproxima a una línea de tendencia. [↑](#footnote-ref-3)