Random Forests

Leo Breiman

18 de Octubre de 2019

Keywords: classification, regression, ensemble

1 Idea 1

La idea que me surge a partir de este paper, es realizar un random forest, pero generando una muestra a partir de una simulación monte carlo y no usar el metodo bootstrap, esta idea utiliza el principio que presenta el paper de generar muestras a partir de una muestra dada, solo que en este caso yo generare una muestra a partir de la media y varianza de cada feature, la importancia de la idea radica en utilizar la tecnica de baggin pero a partide de generar features nuevas no reutilizando las ya observadas en la población. Y ademas se puede conseguir aumentar la cantidad de observaciones.

2 Idea 2

Esta idea consiste en implementar un random forest, utilizando el metodo boostrap percentil, esta idea parte de la misma base del paper que consiste en emsemble learning para decision tree, sim la importancia de la idea presentada radica en que se calcularan intervalos de confianza para las muestras bootstrap, lo cual servira como una métrica adicional que complementara el accuracy.

Top 10 algorithms in data mining

XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang

18 de Octubre de 2019

Keywords: data, mining, algorithm

1 Idea 1

Esta idea consiste en hacer una revisión literaria de todas las versiones y modificaciones del algoritmo de gradient descent, la idea utiliza el aporte del enfoque del paper que indica la revisíon de los algotirmos que han inlfuenciado data mining y consultar a personas referentes en el tema para que ellos puedan establecer el ranking de los algoritmos de gradient descent, asi como utilizar el principio de la validación del uso de los paper que explican el algotirmo segun la cantidad de citaciones. La importancia de la idea radica en que aparte de incluir el Stochastic Gradient Descent y Mini-Batch Gradient Descent, se debe incluir: Stochastic Gradient Descent con momentum, Nesterov accelerated gradient, Adagrad, Adadelta, Adam, AdaMax. Esto se podria incluir como una actividad del chapter de informs para Guatemala.

2 Idea 2

Esta ide hace una combinación con el paper de random forest, pues consiera el algoritmo Adaboost, y consiste en resolver un problema de clasificación implmentando un arbol de decisión, realizando Ensemble learning, con Adaboost y boosttraping, esta idea utilizara como base elgotirmot numero 7 explicacdo en el paper, el aporte de este sera realizar un analisis comprativo de los resultados de tomar dos caminos de muestreo.

The Google File System

Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung

18 de Octubre de 2019

Keywords: Fault tolerance, scalability, data storage, clustered storage

1 Idea 1

La idea consiste en proponer un analilsis de implementación de un data warehouse para bancos, estp parte del tema que presenta el paper en relación al almacenamiento de gran cantidad de datos, y como poder acceder a ellos de forma rapida. Este tema es relevante ya que según lo observado en el mercado, las empresas tienen un coceptualización diferenciada del concepto de datawarehouse, y esto permitira etandarizar este conocimiento.

2 Idea 2

La idea parte de hacer una revisión literaria de toda la literatura relacionada son sistemas de archivos distribuidos,esta idea presentaria el sistema de archivos distribuido do google, el Amazon Elastic File System, el sistema de archivos distriuido de Hadoop entre otros, la idea aportaria un analisis comparativo, ademas de ventajas y desventajas de las distintas implementaciones en diferentes empresas de tecnologia.