UNIVERSIDAD DE TALCA | CHILE
SISTEMA DE BIBLIOTECAS

CLASIFICACIÓN ONLINE USANDO BOSQUES ALEATORIOS

JAVIER IGNACIO GONZÁLEZ PICHUANTE
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

El problema de clasificación consiste en identificar la categoría a la cual pertenece una instancia no vista con anterioridad a partir de un conjunto de instancias cuya categoría respectiva se conoce a priori. La mayoría de estos clasificadores consideran que la distribución estadística de los datos se mantiene constante a través del tiempo y que se tiene acceso a las instancias en todo momento. Sin embargo, en la actualidad ha surgido una nueva forma de obtener observaciones, conocido como flujo de datos. Este flujo se caracteriza por presentar nuevos datos a una alta tasa de velocidad y en grandes cantidades. Esta situación dificulta el normal funcionamiento de un clasificador estándar, por lo tanto es necesario trabajar este flujo de datos bajo un acercamiento online.

En este trabajo presentamos dos algoritmos de clasificación online específicamente diseñado para el caso en que la distribución de datos es dinámica. En primer lugar presentamos el algoritmo llamado Online Naive Bayes Classifier (ONBC) el cual corresponde a un clasificador estadístico basado en el teorema de Bayes que realiza supuestos respecto a la independencia de los atributos. Este clasificador genera dinámicamente un modelo de predicción apoyándose de histogramas en línea se incorpora la identificación automáticamente de cambios en la distribución de datos dentro del flujo de datos. El segundo clasificador llamado Online Naive Bayes Forest (ONBF) toma como base el algoritmo ONBC y los principios de Random Forest para crear un bosque de ONBCs. Este bosque realiza predicciones independientes entre sí para luego agruparlas y establecer una predicción final. Estas propiedades han sido confirmadas experimentalmente sobre numerosos conjuntos de datos pertenecientes a distintos dominios y cuyos resultados se resumen en la presente memoria. Palabras claves: Online Naive Bayes Classifier, Aprendizaje en línea, Cambio de concepto, Histograma Dinámico, Clasificación de Patrones

UNIVERSIDAD DE TALCA | CHILE
SISTEMA DE BIBLIOTECAS

ABSTRACT

The classification problem is to identify the category to which an instance previously unseen belongs from a set of instances whose respective category is known a priori. Most of these classifiers consider the statistical distribution of the data remains constant over time and that you have access to the instances at all times. Nowadays, there is a new way to obtain observations, known as the data stream. This strem is characterized by new data arriving at a high speed rate and in large quantities. This impedes the normal functioning of a standard classifier. Therefore we need to take an online approach in order to work on this data flow.

In this document we present two online classification algorithms specifically designed for the case where the data distribution is dynamic. We first present the algorithm called Online Naive Bayes Classifier (ONBC) which corresponds to a statistical classifier based on Bayes' theorem that makes assumptions about the independence of the attributes. This classifier dynamically generates a prediction model using online histograms and incorporates the automatic detection of changes in the distribution of data within the data stream. The second algorithm is called Online Naive Bayes Forest (ONBF) and its based on the algorithm ONBC and also takes the Random Forest principles in order to create a forest of ONBCs. This forest makes independent predictions among the elements within and then these are grouped and a final prediction is made. These properties have been confirmed experimentally on artificial data sets and whose results are summarized presently. ords: Histograms, Pattern Classification.