ÁRBOLES AUTO-ORGANIZATIVOS PARA LA CLASIFICACIÓN DE DATOS ATÍPICOS

JAVIER IGNACIO AROS MENDOZA
INGENIERO CIVIL EN COMPUTACIÓN

RESUMEN

La agrupación y clasificación son problemas estudiados en Machine Learning. La agrupación corresponde al problema de identificar dentro de un conjunto de datos las clases involucradas, para aglutinar estas instancias no existe información de las clases y en ocasiones se desconoce la cantidad de categorías posible. Esto consiste en asignar elementos similares a un grupo y a la vez asignar a grupos distintos instancias distintas. La clasificación corresponde a determinar mediante una clase conocida, una nueva instancia sin clasificar, determinando si esta última corresponde a la clase ya mencionada o es un elemento extraño, esto se realiza utilizando información estadística de los datos. Con este trabajo se presenta el desarrollo e implementación de Complete k-ary Tree SOM (CKTSOM), un algoritmo con la capacidad de agrupar y clasificar, esté utiliza una estructura de árbol capaz de aprender la distribución de un conjunto de datos de entrada. La estructura del árbol siempre es completa, esto significa que todos los nodos hojas tienen la misma altura. Estos nodos se pueden asociar a neuronas por su capacidad de aprender de la información presentada. En este algoritmo las neuronas compiten para poder representar un dato de entrada, pero hay que destacar un cambio en el paradigma, donde solo competirán las neuronas hojas. Esto es un cambio drástico ya que comúnmente toda la estructura es utilizada para representar la instancia. La búsqueda generada con este cambio de paradigma produce una búsqueda logarítmica que nos entrega un resultado aproximado, este es analizado y verificado para comprender su calidad. Por último se presenta la implementación del algoritmo para clasificar distintos conjuntos de datos que varían en cantidad de instancias y dimensiones. Esta implementación fue utilizada para comparar el algoritmo implementado con otros clasificadores. Para comparar el desempeño se utiliza la medida AUC que nos proporciona el rendimiento del algoritmo, un AUC menor a 0:5 indica que el clasificador es peor que una elección aleatoria.

ABSTRACT

Grouping and classification are problems studied in Machine Learning. The grouping corresponds to the problem of identifying within a set of data the classes involved, to agglutinate these instances there is no information of the classes and sometimes the number of possible categories is unknown. This consists of assigning similar elements to a group and at the same time assigning different instances to different groups. The classification corresponds to determine by means of a known class, a new instance without classifying, determining if this last one corresponds to the class already mentioned or it is a strange element, this is made using statistical information of the data. This work presents the development and implementation of Complete k-ary Tree SOM (CKTSOM), an algorithm with the ability to group and classify, using a tree structure capable of learning the distribution of a set of input data. The tree structure is always complete, this means that all leaf nodes have the same height. These nodes can be associated with neurons because of their ability to learn from the information presented. In this algorithm neurons compete to be able to represent an input data, but it is necessary to highlight a change in the paradigm, where only leaf neurons will compete. This is a drastic change since commonly the whole structure is used to represent the instance. The search generated with this paradigm shift produces a logarithmic search that gives us an approximate result, this is analyzed and verified to understand its quality. Finally we present the implementation of the algorithm to classify different sets of data that vary in number of instances and dimensions. This implementation was used to compare the implemented algorithm with other classifiers. To compare the performance we use the AUC measure that provides the performance of the algorithm, an AUC less than 0:5 indicates that the classifier is worse than a random choice.