Szebenyi Zoltán, IV. évfolyam, programtervező matematikus

Szegedi Tudományegyetem

Konzulens: Csirik János
Egyetemi tanár

Stringek osztályozása mediánok segítségével

A mesterséges intelligencia és a gépi látás területének fontos témája a stringek osztályozása. String alatt itt érthetünk bármilyen, véges abc feletti, tetszőleges hosszú jelsorozatot. Stringekkel nagyon sokféle dolgot reprezentálhatunk. Ilyenek lehetnek például tárgyak képeinek körvonalai, kromoszómák szerkezete, vagy épp kézzel írott számjegyek. Ebből is látható, hogy a témának rengeteg gyakorlati alkalmazása létezik.

Jelen dolgozatban a stringek osztályozásának a mediánok segítségével való megközelítését vizsgálom. A stringekre adott egy távolságfüggvény, például a Levenshtein-távolság, ami annak költségére ad egy mérőszámot, hogy az egyik stringet a másikba vigyük adott költségű beszúrás, törlés és csere operációk sorozatával.

Ezek után egy adott string osztály mediánját definiálhatjuk pl. egy olyan stringként, melynek az osztály elemeitől számított távolság-összege minimális. Egyéb definíciók is léteznek, pl. mikor a többi stringtől való maximális távolság minimális, illetve mikor a távolságnégyzetek összege minimális – ez utóbbira szokták a középpont (center) szót is használni.

A módszer előnye, hogy ha már ismerjük az egyes osztályok mediánját, egy újonnan érkező string osztályozása már nagyon alacsony műveletigénnyel elvégezhető, hiszen csak össze kell hasonlítani az egyes osztályok mediánjaival, és amelyiktől legkisebb a távolsága, az ahhoz tartozó osztályba besorolhatjuk.

Tehát ha viszonylag sok időnk van a tanulásra, de aztán nagyon gyorsan kell majd az érkező stringek osztályozását elvégezni, ez a megközelítés nagyon indokoltnak bizonyulhat.

Természetesen a legoptimálisabb medián megtalálása egy adott osztályra nagy problémák esetén rendkívül költséges lehet, és nem kifizetődő. Ehelyett használhatunk közelítő algoritmusokat, illetve olyan algoritmusokat, melyek csak az osztály elemei közül választják ki a legoptimálisabb jelöltet.

Jelen dolgozatban ilyen tanuló algoritmusokat vizsgálok, többféle változatban, többféle adathalmazra. Az algoritmusokat a futásidő és a kapott mediánok segítségével történő osztályozás hatékonyságának szempontjából vizsgálom. Összehasonlítom ezek tulajdonságait, előnyeit és hátrányait, és megkeresem legjobb paraméterezéseiket az egyes adathalmazokra. További hatékonyság-javító lehetőségeket vetek fel és tesztelek.