

DIPLOMAMUNKA KIVONAT

Metrikus terek beágyazása, alkalmazás klaszterezésre

Zsbán Ambrus

Témavezető: Lukács András, MTA Számítástechnikai és Automatizálási Kutatóintézet.

Belső témavezető: Rónyai Lajos, BME Matematikai Intézet, Algebra Tanszék.

Dolgozatomban először rövid áttekintést nyújtok a véges metrikus beágyazások elméletéről. Ez a módszer szélsőérték-feladatokra ad közelítő megoldást olyan módon, hogy egy metrikus teret egy másik, egyszerűbb szerkezetű metrikus térbe ágyaz be. Bemutatok néhány tételt, amelyek ilyen beágyazást konstruálnak; és egy lemmát, amelynek segítségével az ilyen beágyazásokat fel lehet használni.

Részletesen ismertetem a metrikus beágyazások elméletének egy alkalmazását: egy közelítő megoldást adó algoritmust egy konkrét klaszterezési feladatra, a min-sum k -clustering problémára. Ez az algoritmus egy általános metrikus teret először egy speciális formájú véletlen fa által meghatározott térbe képezi, majd ezen a metrikán rekurzív módon számítja ki egy jó közelítést.

A klaszterezést – vagyis adatok előre nem meghatározott, közeli pontokból álló csoportokba sorolását – az adatbányászatban többek között arra használják, hogy egy nagy méretű magas dimenziós adathalmaz viselkedését az egyes klaszterek néhány tulajdonságának vizsgálatával könnyebben megérthessük.

My thesis first gives a short review of the theory of finite metric embeddings. This method gives approximating solutions to optimization problems by embedding a metric space to another, simple metric space. I describe a few theorems that construct such embeddings, and a lemma allowing us to use these embeddings.

I comprehensively present an application of this theory, namely an algorithm that approximates a concrete clustering problem, min-sum k -clustering. This algorithm works by first transforming a general metric space to a space induced by a random tree of a certain special form, then gets a good approximation on this new metric with a recursive calculation.

Clustering, which means dividing points to groups of points close to each other without groups known in advance, is used in data mining for understanding the behaviour of large high-dimension data sets by examining few properties of each cluster.

Budapest, 2008. május 21.