IEEE SUMMER SCHOOL ON "SYSTEMS, MAN AND CYBERNETICS"

Sponsored by: IEEE SMC Society
Organized by: IEEE CIS/SMC Chapters of Bulgaria and IEEE Young Professionals Bulgaria
Hosted by: Software University, Sofia

IEEE ЛЯТНО УЧИЛИЩЕ ПО "СИСТЕМИ, ЧОВЕК И КИБЕРНЕТИКА"

Спонсорирано от: IEEE SMC общество
Организирано от: IEEE CIS/SMC и IEEE Young Professionals в България
Домакинство: Софтуерен Университет, София

LECTURE 3

TITLE: HOW BIG IS TOO BIG? (MOSTLY) C-MEANS CLUSTERING IN BIG DATA

LECTURER: Prof. James Bezdek, IEEE Fellow, USA

ABSTRACT:

What is big data? Objectives of clustering in big data are acceleration for loadable data and feasibility for non-loadable data. History of c-means and least squares estimation. Acceleration methods for fuzzy c-means (FCM). Approximate clustering with FCM and Gaussian mixture models (EM) based on literal clustering of a sample followed by non-iterative extension. Incremental methods (spFCM and olFCM) that process data chunks sequentially. Extension of VAT to scalable VAT (sVAT) for arbitrarily large square data. sVAT marries single linkage (SL), resulting in two offspring: scalable SL and clusiVAT. Time and accuracy comparisons of clusiVAT to crisp versions of the three FCM models and CURE. Experiments on 48 synthetic data sets of Gaussian clusters, and three real world data sets (Iris, Forest and the KDD-99 cup). For example, clusiVAT recovers 97.06% of the crisp labels from the KDD-99 cup data, a set of 4, 292,637 vectors, each having 41 attributes, in 76 seconds.

ТЕМА: КОЛКО ГОЛЕМИ СА „ГОЛЕМИТЕ МАСИВИ ОТ ДАННИ“? C-MEANS КЛЪСТЕРИЗАЦИЯ НА „ГОЛЕМИ МАСИВИ ОТ ДАННИ“.

ЛЕКТОР: проф. Джеймс Бездек, IEEE Fellow, САЩ

РЕЗЮМЕ:

Какво представляват „Големите масиви от данни“? Целите на клъстеризация, приложена върху големи масиви от данни са:

- ускоряване на клъстеризацията за данни, които могат да се заредят в паметта;
- възможност за изпълнение на клъстеризация, работеща с данни, които немогат да се заредят в паметта.

Темата включва още история на c-Means и намиране на най-малките квадрати. Методи за ускоряване на размита c-Means клъстеризация (FCM). Апроксимираща клъстеризация с FCM и Гаусови смесени модели (EM), базирани на literal-клъстеризация на извадка, последвана от неитеративно разширение. Инкрементиращи методи (spFCM и olFCM), които обработват части от данните последователно. Разширение на VAT до скалируем VAT (sVAT) за произволно големи квадратни данни. sVAT свързва единична връзка (SL), от която произтичат 2 наследника: скалируем SL и clusiVAT. Сравнение на времето на изпълнение и точността на clusiVAT с чистите варианти на трите FCM модела и CURE. Експерименти върху 48 изкуствено-създадени множества от данни с Гаусови клъстери и три реални данни (Iris, Forest и KDD-99 cup). Например, clusiVAT покрива 97.06% от точните стойности на класификационната функция върху данните KDD-99 cup, множество от 4, 292,637 вектора, всеки от които има 41 атрибута, работейки 76 секунди.