ПОСЛІДОВНА ОН-ЛАЙН МОДЕРНІЗАЦІЯ МЕТОДУ КЛАСТЕРИЗАЦІЇ J- СЕРЕДНІХ

Бодянський Є.В., Патлань К.В

Харківський національний університет радіоелектроніки, Національний технічний університет «Харківський політехнічний інститут», м. Харків

На цей час методи інтелектуального аналізу даних та машинного навчання знаходять широке застосування у різноманітних галузях людської діяльності: техніці, економіці, маркетингових дослідженнях, медицині, біології, фармації, тощо. Завдання кластеризації тут займає особливе місце, оскільки вирішується на основі парадигми навчання без вчителя — самонавчання, тобто в умовах суттєвого дефіциту апріорної інформації відсутності апріорі розміченої навчальної вибірки.

Не дивлячись на існування великої кількості відомих методів та алгоритмів, найбільшого поширення набули алгоритми кластеризації засновані на використанні прототипів — центроїдів. Тут найпопулярнішим є метод К-середніх (K-means, HCM) та метод нечітких С-середніх (fuzzy C-means, FCM) у випадку перетинних класів-кластерів. Популярність цих підходів пояснюється, перш за все, простотою чисельної реалізації та некритичністю до рівня апріорної інформації. До недоліків слід віднести необхідності завдання кількості кластерів що формується. Ця проблема долається використанням методу Х-середніх (X-means). Більш серйозною проблемою є можливість «застигання» алгоритму в локальних мінімумах цільової функції, чого можна запобігти за допомогою методу J-середніх (J- means).

Сутність цього методу полягає в тому, що при попаданні у локальний екстремум, алгоритм реалізує інтенсивні випадкові стрибки (Jamps), що виводять процедуру з околу локального екстремум.

Всі ці методи призначені для роботи у пакетному режимі, тобто апріорі задана вибірка багаторазово опрацьовується у формі фіксованого масиву даних. Якщо ж дані надходять у послідовному он-лайн режимі, ці підходи є непрацездатними. В умовах, коли все більшого поширення набувають задачі, пов’язані з Big Data, викликає необхідність розробки нових підходів, пристосованих для нових умов.

Як відомо самоорганізовні мапи Т.Кохонена, що навчаються у послідовному режимі, можуть реалізувати HCM- кластеризацію на основі правила самонавчання «Переможець отримує все»(WTA) та FCM-кластеризацію на основі правила «Переможець отримує більше» (WTM), якщо функція сусідства обирається у вигляді кошіану.

Нескладно додати у процедуру налаштування системних ваг нейронної мережі Т.Кохонена, що є за штучно градієнтними алгоритмами оптимізації, випадкових збурень, що надає алгоритму самонавчання в цілому властивостей глобального випадкового пошуку. Таким чином реалізується но-лайн модифікація J-середніх, основною перевагою якої є простота чисельної реалізації та висока швидкість обробки даних, що надходять у послідовному режимі.