ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ
Факултет по Телекомуникации
Катедра Комуникационни мрежи

Маг. инж. Анкита Витхал Карале

ЕФЕКТИВЕН ПОДХОД ЗА ОТКРИВАНЕ НА НЕТИПИЧНИ СТОЙНОСТИ В ПОТОЧНИ ДАННИ С МЕТАЕВРИСТИЧНА ОПТИМИЗАЦИЯ

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен
"ДОКТОР"

Област: 5. Технически науки
Професионално направление: 5.3 Комуникационна и компютърна техника
Научна специалност: Комуникационни мрежи и системи

Научни ръководители: Проф. дн инж. Владимир Пулков
Проф. д-р инж. Милена Лазарова

СОФИЯ, 2021 г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Комуникационни мрежи“ към Факултет по Телекомуникации на ТУ–София на редовно заседание, проведено на 14.06.2021 г.

Публичната защита на дисертационния труд ще се състои на 19.10.2021 г. от 15,00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед № ОЖ-5.3-25 от 29.06.2021 г. на Ректора на ТУ-София в състав:

1. Проф. д-р Милена Лазарова–Мицева – председател
2. Проф. д-р Георги Илиев – научен секретар
3. Доц. д-р Станимир Садинов
4. Проф. д-р Олимпия Роева
5. Доц. д-р Филип Цветанов

Рецензенти:

1. Проф. д-р Георги Илиев
2. Проф. д-р Олимпия Роева

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет по Телекомуникации на ТУ–София, блок №1, кабинет № 1254.

Дисертантът е редовен докторант към катедра „Комуникационни мрежи“ на факултет по Телекомуникации. Изследванията по дисертационната разработка са направени от автора.

Автор: маг. инж. Анкита Карале
Заглавие: Ефективен подход за откриване на нетипични стойности в поточни данни с метаевристична оптимизация
Тираж: 30 броя
Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Откриването на нетипични стойности е важен изследователски проблем с приложение в множество различни област, изискващи обработване и извличане на данни. Задачата за откриване на нетипични стойности се състои в определяне на отклонения в очаквано или типично поведение на данните, дължащи се на аномалии, дефекти, грешки, повреди, шум. Приложните области включват откриване на необичайни събития при анализ на данни в сензорни мрежи и разпределени системи, финансови, метеорологични и екологични анализи, анализи при защита на данни и други. Съществуват множество методи и алгоритми за откриване на отклонения в статични набори от данни с краен брой стойности. Откриването на нетипични стойности при поточни данни, които имат динамичен характер, висока скорост на генериране и изискване обработването на данните да се извършва в реално време при ограничение на изчислителните ресурси и капацитета на използвана памет е актуален изследователски проблем, които е обект на голям брой научни изследвания за предлагане на ефективни методи и алгоритми.

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на научните изследвания в дисертационния труд е да се проектира и разработи ефективен алгоритъм за откриване на нетипични стойности при поточно предаване на данни в реално време при ограничения на използваната памет с висока точност и минимално време за изпълнение.

За постигане на поставената цел са дефинирани следните задачи:

1. Да се проучат различните техники за откриване на нетипични стойности и да се оцени тяхната ефективност.

2. Да се проектира и разработи метод за откриване на нетипични стойности, който може да оперира с поточни данни, може да се изпълнява при ограничения в използваната памет и при невъзможност всички генерирани в реално време стойности да се съхраняват, има висока точност на определяне на отклонения и аномалии в данните, изчислително ефективен е и осигурява минимално време за изпълнение.

3. Да се оцени производителността на проектирания и раработен метод по отношение изпозлвана памет, време за изпълнение и точност на определяне на отклонения и аномалии в поточните данни.

Научна новост

Предложен е нов подход за откриване на локални отклонения в поточни данни: Memory Efficient Outlier Detection (MEOD), който комбинира MiLOF и LOCI и използва техника за оптимизация с рояк частици и обобщаване на данните. MEOD се базира на оптимизация с рояк частици за определяне на оптимална стойност за радиуса r при алгоритъма LOCI, премахва зависимостта от броя на най-близките съседи при клъстеризация с алгоритъм k-NN и използва подход за определяне на стойност на фактора за нетипичност само за кандидат-стойностите, а не за целия набор от данни. С използване на предложения подход MEOD откриването на нетипични стойности в поточни данни е с по-добра точност, изисква по-малко изчислително време и използва по-малко памет в сравнение с алгоритъма MiLOF.

Предложен е нов подход за откриване на локални отклонения в поточни данни: Advanced Memory Efficient Outlier Detection (A-MEOD), който използва техника за оптимизация с рояк частици за определяне на оптимална стойност за радиуса r, премахва зависимостта от броя на най-близките съседи при клъстеризация с алгоритъм k-NN и подобрява изчислителната сложност на MEOD чрез определяне само на най-големите M отклонения при алгоритъма LOCI на базата на съотношението за локална плътност k/r. A-MEOD изисква по-малко памет и по-малко време за изпълнение в сравнение с MEOD и MiLOF при откриване на нетипични стойности в поточни данни.

Практическа приложимост

Предложените подходи за откриване на локални отклонения в поточни данни могат да бъдат изпозлвани за определяне на аномалии и отклонения в системи, при който данните се генерират динамично, изчисленията се извършват в реално време при ограничение на изчислителните ресурси и капацитета на използваната памет и се изисква висока точност при определянето на нетипични стойности.

Апробация

Резултатите от дисертационния труд са разглеждани, обсъждани и публикувани в:

1. Journal of Mobile Multimedia, River Publishers, Vol. 16, No. 3, 2020, индексиран в Scopus.
2. 43 международна научна конференция „Telecommunications and Signal Processing“ (TSP‘2020), Милано, Италия, 7 – 9 юли 2020, индексиран в Scopus и Web of Science.
3. MDPI Journal Symmetry, Vol. 13, No. 3, 2021, индексиран в Scopus и Web of Science, IF 2.645.
4. 44 международна научна конференция „Telecommunications and Signal Processing“ (TSP‘2021), 26–28 юли 2021, индексиран в Scopus и Web of Science.

Публикации

Основни постижения и резултати от дисертационния труд са публикувани в 4 научни статии, от които една самостоятелна и три като водещ съавтор. Статиите са публикувани в международни реферирани и индексирани издания.

Структура и обем на дисертационния труд

Дисертационният труд е в обем от 126 страници, като включва увод, 5 глави за решаване на формулираните основни задачи, списък на основните приноси, списък на публикациите по дисертацията и използвана литература. Цитирани са общо 184 литературни източници, като 181 са статии на латиница, а 3 са интернет адреси. Работата включва общо 30 фигури и 7 таблици. Номерата на фигурите и таблиците в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. МЕТОДИ ЗА ОТКРИВАНЕ НА НЕТИПИЧНИ СТОЙНОСТИ

В първа глава на дисертационния труд е направено проучване и сравнителен анализ на методите за откриване на нетипични стойности.

1.1. Видове нетипични стойности

Задачата за откриване на нетипични стойности, която се състои в определяне на отклонения в очаквано или типично поведение на данните, дължащи се на аномалии, дефекти, грешки, повреди, шум. Основните видове отклонения са: глобални (точкови), контекстуални (локални) и колективни отклонения.

1.2. Методи за откриване на нетипични стойности

Методите за откриване на нетипични стойности се категоризират в три класа в зависимост от стратегията за управление на процеса: управляеми (supervised), неуправляеми (unsupervised) и полу-управляеми (semi-supervised).

Управляемите стратегии за откриване на нетипични стойности използват набор от данни, който включва маркирани за типични и за нетипични стойности са създаване на модели за типични и за нетипични стойности. За откриване на нетипичи стойности в даден набор данни стойностите се сравняват с двата модела за да се определи принадлежност към един от двата класа. Недостатък е изискването за наличие на маркирани обучаващи данни, чието анотиране е трудоемко.

При методите за полу-управлеямо откриване на нетипични стойности се конструира модел за нормално поведение на обучаващо множество от типични данни, след което се определя вероятността тестов екземпляр да бъде генериран от създадения модел. Недостатък е изискването за наличие на набор от данни, който да обхваща всяко възможно отклонение, което може да се наблюдава в данните.

Неуправляемите стратегии за откриване на нетипични стойности не изискват примерни данни с информация за етикети на класове. Методите в тази категория откриват отклонения в немаркирани данни на базата на допускането, че повечето данни са типични, а всяка необичайна стойност се третира като отклонение. В зависимост от използвания подход неуправляемите методи за откриване нетипични стойности могат да бъдат категоризирани както следва (фиг. 1.5): методи базирани на плътност (density-based), методи базирани на статистически данни (statistical-based), методи базирани на разстояние (distance-based), методи базирани на клъстери (clustering-based), методи базирани на ансамбли (ensemble-based).

На базата на литературен обзор на съществуващите неуправляеми методи за откриване на нетипични стойности са определние предимствата и недостатъците на методите от всяка категория:

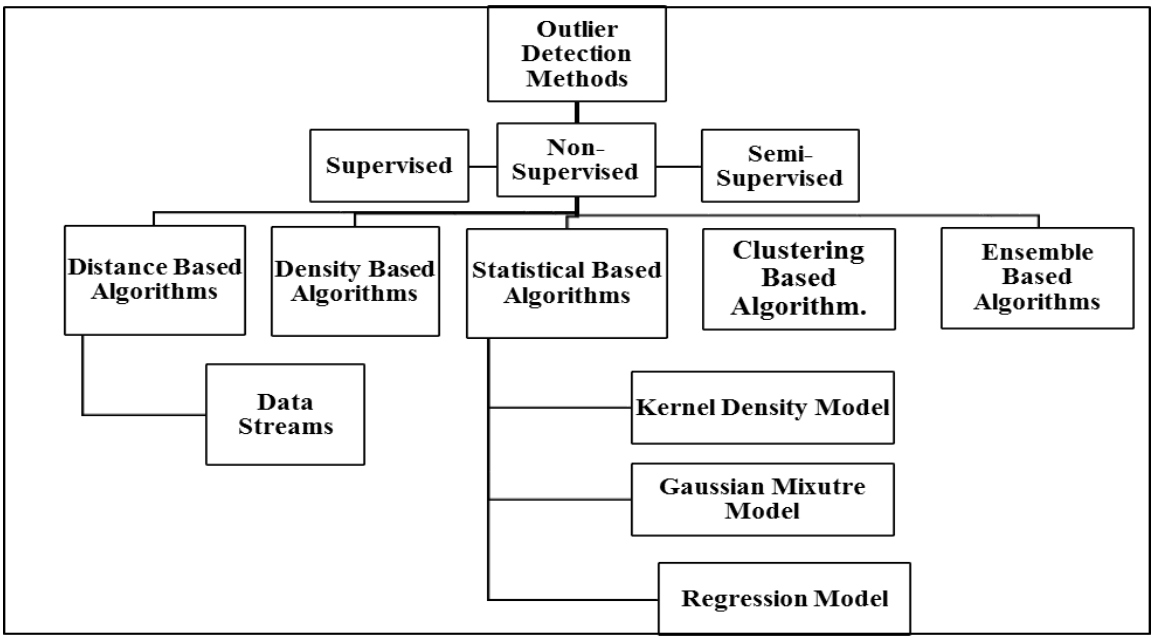
- Методи базирани на плътност

Предимства: Изискват малко априорна информация и използват малък брой параметри. Ефективни са при определяне на локални аномалии.

Недостатъци: Изчислително неефективни в сравнение със статистическите стратегии. Изискват предварително дефиниране на параметри. Не са приложими за поточни данни.

- Методи базирани на статистически данни

Предимства: Добро изчислително бързодействие след изграждане на модели. Подходящи за набори от данни с реални стойности.



Фигура 1.5: Методи за откриване на нетипични стойности

Недостатъци: Високи изчислителни разходи при многомерно пространство. Резултатите не са надеждни за реални сложни приложения.

- Методи базирани на разстояние

Предимства: Използват прости техники, не използват предположения. Изчислително ефективни и мащабируеми в сравнение със статистическите методи.

Недостатъци: Изчислителната сложност нараства при многомерни данни.

- Методи базирани на клъстери

Предимства: Не изискват априорна информация. приложими за различни типове данни. Добра изчислителна сложност. Адаптивност и приложимост за поточни данни.

Недостатъци: Не позволяват обратно проследяване за оценка на точността. Повечето техники изискват предварително определен брой клъстери, които може да не е известен предварително. Чувствителност към шумове. Изчислителната сложност нараства при многомерни данни.

- Методи базирани на ансамбли

Предимства: Подходящ за данни с голяма размерност. По-ефективнии от другите методи при наличие на шум. Подходящи за потончи данни.

Недостатъци: Липсват ефективни техники за оценка на характеристики. Неприложими за реални данни при малък размер на обучаващите примери.

1.3. Изводи от първа глава

Съществуват множество методи и алгоритми за откриване на отклонения в статични набори от данни с краен брой стойности. Откриването на нетипични стойности при поточни данни, които имат динамичен характер, висока скорост на генериране и изискване обработването на данните да се извършва в реално време при ограничение на изчислителните ресурси и капацитета на използвана памет е актуален изследователски проблем, които е обект на голям брой научни изследвания за предлагане на ефективни методи и алгоритми.

На базата на направеното обстойно проучване и сравнителен анализ на методите за откриване на нетипични стойности е формулирана целта на научните изследвания в дисертационния труд: да се проектира и разработи ефективен алгоритъм за откриване на нетипични стойности при поточно предаване на данни в реално време при ограничения на използваната памет с висока точност и минимално време за изпълнение.

За постигане на поставената цел са дефинирани следните задачи: да се проучат различните техники за откриване на нетипични стойности и да се оцени тяхната ефективност, да се проектира и разработи метод за откриване на нетипични стойности, който може да оперира с поточни данни, може да се изпълнява при ограничения в използваната памет и при невъзможност всички генерирани в реално време стойности да се съхраняват, има висока точност на определяне на отклонения и аномалии в данните, изчислително ефективен е и осигурява минимално време за изпълнени; да се оцени производителността на проектирания и раработен метод по отношение използвана памет, време за изпълнение и точност на определяне на отклонения и аномалии в поточните данни.

ГЛАВА 2. МЕТОДИ ЗА ОПТИМИЗАЦИЯ С ИНТЕЛИГЕНТНОСТ НА РОЯК ЗА ОТКРИВАНЕ НА НЕТИПИЧНИ СТОЙНОСТИ

Във втора глава на дисертационния труд е направено проучване и анализ на подходите за решаване на оптимизационни проблеми с използване на различни техники за оптимизация, в това число метаевристични методи за оптимизация с интелигентност на рояк.

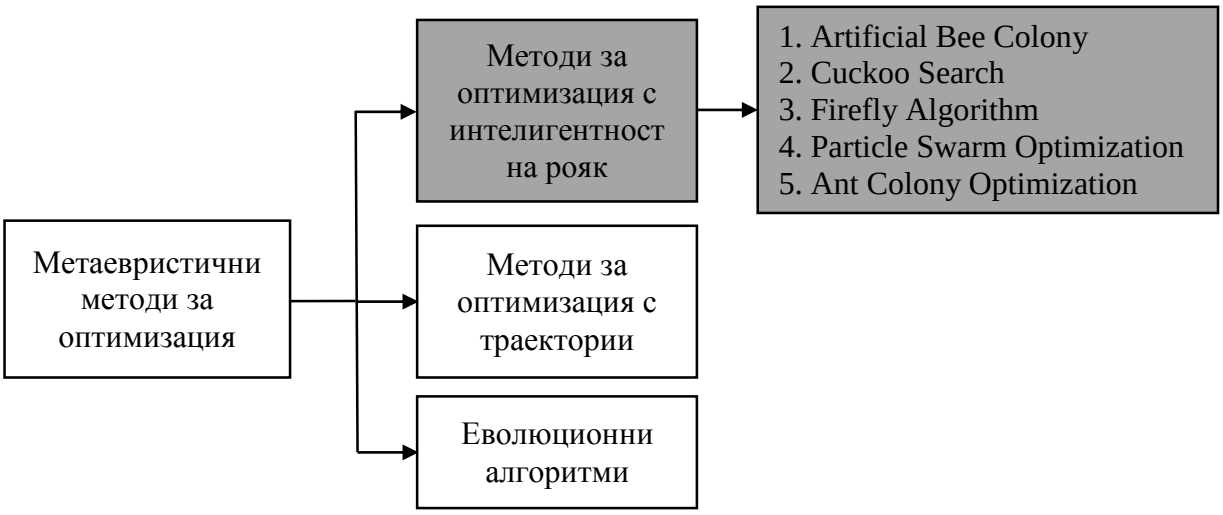
2.1. Оптимизационна задача и методи за оптимизация

Процесът на оптимизацация на даден проблем се дефинира като определяне на стойностите на множество от независими променливи, дефинирани в границите на пространството за търсене, които не нарушават предварително зададени ограничения, за които набор от целеви функции имат минимална или максимална стойност. За решаване на даден оптимизационен проблем се изисква дефиниране на граници на пространството на търсене, дефиниране на една или повече целеви функции, дефиниране на независими променливи, определяне на модел на оптимизационната задача, който дефинира връзката между независимите променливи и целевите функции.

Класическите методи за решаване на оптимизационни задачи се основават на диференциално смятане за намиране на оптимална стойност на целевите функции при следните предположения: целевите функции са диференцируеми по отношение на независимите променливи и производните им са непрекъснати функции.

Метаевристичните методи за решаване на оптимизационни задачи се използват за определяне на близко до оптималното решение. Метаевристичните методи не изискват изчисляване на производни на целевата функция. Качеството на полученото оптимално решение не се влияе от първоначални предположения, но изчислителната ефективност зависи от началния дизайн. Целевата функция и ограниченията не се влияят от непрекъснатост или диференцируемост на целевата функция. Сходимостта на процеа на търсене на оптимално решение не се влияе от пространството на търсене. Метаевристините методи са приложими при различен обхват на независимите променливи, както и при дискретни и непрекъснати оптимизационни задачи.

В зависимост от използвания подход за определяне на оптимално решение метаевристичните методи се класифицират като еволюционни алгоритми, методи базирани на траектории и методи базирани на интелигентност на рояк (фиг. 2.1).



Фигура 2.1: Метаевристични методи за оптимизация

2.2. Методи за оптимизация с интелигентност на рояк за откриване на нетипични стойности

Метаевристичните методи за оптимизация с интелигентност на рояк са инспирирани от интелигентното поведение на групи биологични системи (мравки, птици, пчели и други), които взаимодействат помежду си и със средата при отсъствие на глобално упраеление и координация.

Основните принципи на метафората за оптимизация с интелигентност на рояк са следните:

- Близост: Основните единици на рояка (агенти) реагират в зависимост от естествените колебания, активирани от комуникацията между тях и със средата.
- Качество: Роякът или агентът реагират в зависимост от фактори за качество като откриване на сигурност на местоположението им.
- Принцип на различен отговор: Агентите не се разполагат в ограничена зона, разпръскването им е планирано с цел всеки агент да бъде максимално обезпечен срещу естествените колебания.
- Принцип на стабилност: Агентът не променя поведението си при всяка промяна на средата.

Взаимодействието между агентите в рояка може да бъде пряко или косвено. Прякото взаимодействие в природата е под формата на звуково или визуално взаимодействие, докато при непряко взаимодействие контактът между агентите е чрез околната среда.

Метаевристични алгоритми за оптимизация с интелигентност на рояка са оптимизация на рояк частици (Particle Swarm Optimization, PSO), оптимизация с изкуствена пчелна колония (Artificial Bee Colony, ABC), оптимизация с колония мравки (Ant Colony Optimization, ACO), оптимизация с алгоритъм на светулката (Firefly Algorithm), оптимизация с алгоритъм на кукувицата (Cuckoo Search Algorithm).

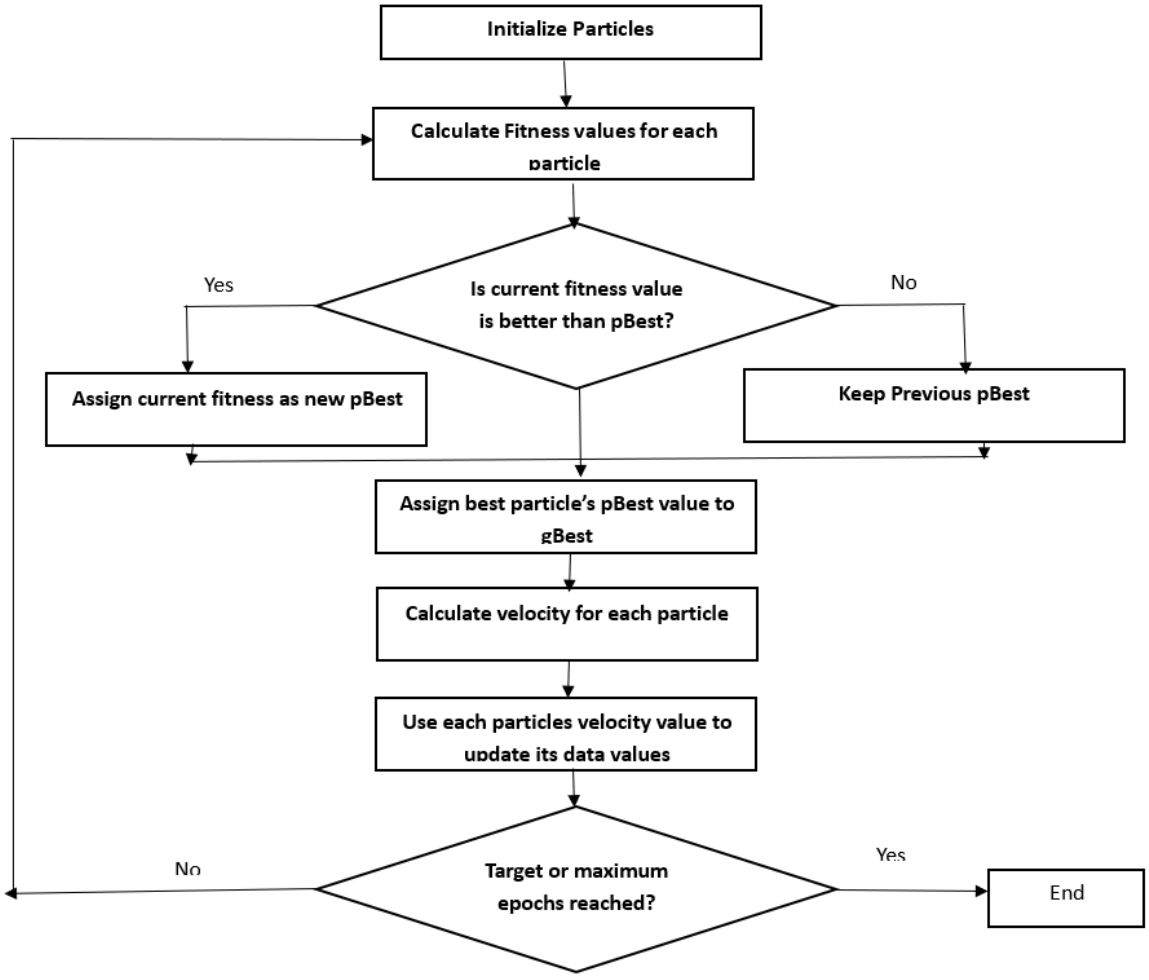
Оптимизацията на рояк частици (PSO) е метаевристичен алгоритъм с популации, инспириран от поведението на рояци птици или рибни пасажи. Колективното интелигентно поведение на рояка се базира на три основни правила, които могат да бъдат обобщени както следва:

1. Избягване на колизии: агентите променят позицията си в групада за избегнат колизия спазвайки безопасно разстояние помежду си.

2. Синхронизиране на скоростта: агентите регулират скоростта и посоката на движението си спрямо съседите за да се движат заедно в група.

3. Подреждане: агентите запазват позицията си спрямо съседите си за да останат близо до тях и да запазят подреждането на рояка.

Популацията (роякът) при алгоритъма с оптимизация на рояк частици се състои от колекция от частици (кандидат-решения) с произволно инициализирани начално положение и начална скорост на движение в дадено хиперпространство. Популацията еволюира във времеви итерации с прилагане на два оператора за актуализиране на положението и скоростта на всички частици. Операторите за актуализация използват данни за най-добро локално местоположение на всяка частица и текущо най-добро местоположение за всички частици в популацията. Алгоритъмът за оптимизация на рояк частици е представен на фиг. 2.5.



Фигура 2.4: Алгоритъм за оптимизация на рояк частици

Частица е индивид в рояка, която представя кандидат-решение на оптимизационния проблем. Всички частици в рояка определят позицията и скоростта си на движение с цел определяне на оптималното решение на проблема. Решението на оптимизационния проблем се определя итеративно като на всяка итерация положението и скоростта на движение на частицата се актуализира. Позицията на всяка частица определя място на частицата в пространството на търсене в N мерно пространство според разглеждания проблем, в което всеки набор от координати представлява решение на проблема. Фитнес функцията е целева функция, с която се определя качеството на дадена позиция в пространството за търсене като решение на оптимизационния проблем.

Най-доброто локално местоположение на дадена частица pbest определя най-доброто решение, което всяка частица е намерила в процеса на търсене и еволюция на рояка. На всяка итерация се сравнява стойността на фитнес функцията за текущата позиция с тази на pbest и ако текущата позиция има по-добра фитнес функция, то pbest се заменя с текущата позиция за дадената частица. Текущото най-добро местоположение за всички частици в популацията gbest е най-добрата позиция в целия рояк, към която всяка частица се стреми. Във всяка итерация на еволюцията на рояка фитнес функцията за текущата позиция за всички частици се сравнява с тази на gbest и ако за някоя частица текущата позиция има по-добра фитнес функция от gbest, то gbest се заменя с текущото положение на тази частица. Когнитивен фактор е коефициент за промяна на траекторията на движение на всяка частица в процеса на търсене така, че промяната на скоростта на частицата да промени позицията на частицата в посока към най-добрата ѝ позиция pbest. Социален фактор е коефициент, който променя скоростта на частиците така, че позицията на всяка частица да се променя в посока към най-добрата позиция за целия рояк gbest. Итеративното търсене на най-доброто решение на оптимизационния проблем продължава докато се изпълнят предварително дефинирани максимален брой итерации или до достигане на задоволителното решение с предварително дефинирана стойност на целевата функция.

Алгоритъмът за оптимизация на рояк частици има ниска изчислителна сложност, малък брой входни параметри и висока конвергенция към най-доброто решение, поради което се прилага за решаване на разнообразни оптимизационни задачи, в това число и при при откриване на нетипични стойности.

2.3. Мотивация за избор на оптимизация с рояк частици (PSO) за откриване на нетипични стойности

В резултат на направеното проучване и сравнителен анализ на метаевристичните методи за оптимизация с интелигентност на рояк могат да определят следните основни предимства при използването им в методи за откриване на нетипични стойности: мащабируемост, адаптивност, колективна устойчивост, простота на агентите.

При решаване на задачата за откриване на нетипични стойности метаевристични алгоритми с оптимизация на рояка се използват както на етапа на предварителна обработка, така и за селектиране на характеристики и в комбинация с методи базирани на клъстери.

В резултат на направеният сравнителен анализ на използването на метаевристични методи за оптимизация с интелигентност на рояк са определени следните предимства за избора на оптимизация с рояк частици (PSO) като оптимизационна техника при проектирането на ефективен подход за откриване на нетипични стойности за поточни данни при ограничения в използваната памет и изчислителните ресурси: изчислителна простота без използване на градиентен подход за оптимизация, малко алгоритмични параметри, които трябва да бъдат настроени, приложим за оптимизация в многомерни пространства, приложим при непълни данни. Оптимизацията с рояк частици е широко използвана метаевристична техника с приложение в редица различни области и е обект на активни научни изследвания с богата база от публични приложения.

2.4.Изводи от втора глава

Метаевристичните методи за оптимизация с интелигентност на рояк могат да бъдат използвани при проектиране на методи за откриване на нетипични стойности. Оптимизацията с рояк частици (PSO) е избрана като метаевристична техника, която да се използва при проектиране на ефективен подход за откриване на нетипични стойности за поточни данни при ограничения в използваната памет и изчислителните ресурси.

ГЛАВА 3. ЛОКАЛНИ МЕТОДИ ЗА ОТКРИВАНЕ НА НЕТИПИЧНИ СТОЙНОСТИ

В трета глава на дисертационния труд са представени методи за откриване на локални нетипични стойности, които могат да бъдат приложени за поточни данни и позволяват изчисления при ограничение на използваната памет.

От разгледаните в първа глава методи за откриване на нетипични стойности са избрани тези, базирани на разстояние, при които оценката за нетипичност (отклонение) се определя на базата на най-близки съседи в обработваните данни. Това прави методите базирани на разстояние приложими в случаи когато предварително не е известно разпределението на данните. Методите за откриване на нетипични стойности базирани на разстояние се категоризират като глобални и локални в зависимост от използвания подход за определяне на оценката за отклонение. Предимствата на подходите за локално откриване на нетипични стойности е способността им да откриват отклонения в набори от данни с нехомогенни плътности на разпределение.

3.1. Методи за откриване на локални нетипични стойности

- Метод базиран на локален фактор на нетипичност (Local Outlier Factor, LOF)

За определяне на нетипичните стойности се използва локален коефициент на отклонение, изчислен за всяка точка от данните p както следва:

$$LOF_k(p) = \frac{1}{k} \sum_{O \in N_{(p,k)}} \frac{lrd_k(o)}{lrd_k(p)}$$

където е $N_{(p,k)}$ е множеството от k най-близки съседа на p, а $lrdk(p)$ се изчислява чрез:

$$lrdk(p) = (\left(\frac{1}{k} \sum_{O \in N_{(p,k)}} reach\_dist_k\,(p,o) \right)^{-1}$$

$$reach\_dist(p,o) = max\{k\_distance(o), d(p,o)\}$$

С $d(p,o)$ е означено Евклидовото разсатояние между p и o, а с $k\_distance(o)$ – разстоянието между точка o и k-тия най-близък съсед.

Предимства: Добра точност и приложимост за определяне на нетипични стойности за набори от данни с нехомогенна плътност.

Недостатъци: Въпреки, че съществуват много различни версии и модификации на LOF, алгоритъмът не е приложим за определяне на локални отклонения в поточни данни.

- Инкрементален LOF метод (iLOF)

При метода LOF се изисква целия набор от данни за изчисляване на локалния фактор на отклонение за всяка точка от данните, докато при метода iLOF се използва инкрементален подход, който изисква само част от данните за актуализиране на LOF стойностите. За всяка входяща точка се определят k най-близки съседи, изчислява се локалния коефициент на отклонение на базата на локалните коефициенти на най-близките съседи и се актуализират локалните фактори на съседните точки ако е необходимо.

Предимства: Определянето на локалния фактор на отклонение е инкрементално и изчислително ефективно.

Недостатъци: За определянето на локалния фактор на отклонение се изисква всички данни да са налични и да се съхраняват в паметта, което прави алгоритъма неприложим за изчисления в реално време при поточни данни.

- Метод с ефективно използване на памет Memory Efficient iLOF (MiLOF)

Методът MiLOF е модификация на инкременталния метод с локален фактор на отклонение, приложим за поточни данни. При MiLOF се прилага обобщаване на фактора на отклонение за подмножества и натрупване на история за вече обработените поточни данни. По този начин се преодоляват ограничения за налична и използвана памет и се постига по-добра изчислителна ефективност.

Методът MiLOF включва няколко стъпки за локално откриване на отклонения при достигане на границите на паметта: стъпка на обобщаване, с-средна стъпка, стъпка на сливане, ревизирана стъпка на вмъкване. За всяка входна точка от данни стойността на локалния фактор на отклонение LOF се изчислява като се използва iLOF. След като се достигне ограничението на използваната памет се изчислява обобщение на фактора на отклонение за първите *b*/2 точки и те се изтриват от паметта. Обобщението за изтритите точки е под формата на набор от *c* прототипни вектора. Ако съществува набор от обобщени прототипи от предишен времеви прозорец, двата набора от прототипи се обединяват, с което се определя обобщение на минали точки от данни. Алгоритъмът продължава докато достигне края на потока от данни. Броят на съхранявани в паметта данни за откриване на отклонения е не повече от m = b + c.

Предимства: Добра скалируемост и ефективност по отношение на използвана памет

Недостатъци: Резултатите от прилагането на алгориртъма зависят от броя най-близки съседи, използвани за оценка на плътността.

- Метод с интеграл на локална корелация Local Correlation Integral (LOCI)

Методът LOCI използва коефициент MDEF (Multi-Granularity Deviation Factor), който позволява постигане на независимост при откриване на нетипични стойности от локални вариации на плътността в пространството на характеристиките, така че да бъдат определени както изолирани отклонения, така и отдалечени клъстери.

Коефициентът MDEF за точка p_i от входния поток данни оценява относителното отклонение на локалната плътност на съседство от средната плътност на локална околност с радиус r:

$$MDEF(\ p_i, r, \alpha\ ) = \frac{\mathrm{n'(p_i,r,\alpha) - n(p_i,\alpha r)}}{\mathrm{n'(p_i,\alpha,r)}} = 1 - \frac{\mathrm{n(p_i,\alpha r)}}{\mathrm{n'(p_i,\alpha,r)}}$$

където $n(p_i, r)$ е брой съседни точки за p_i в локална околност с радиус r, $\mathrm{n'}(pi, \mathrm{r}, \alpha)$ е средна стойност на $n(p, \alpha r)$ за множеството съседни точки на p_i в локална околност с радиус r:

$$\mathrm{n'}(pi, \mathrm{r}, \alpha) \equiv \frac{\sum_{p \varepsilon N(P_i, r)} n(p, \alpha \mathrm{r})}{n(p_i, r)}$$

С $\sigma_{n'}(pi, r, \alpha)$ е означено стандартното отклонение на $n(p, \alpha r)$ за множеството съседни точки на p_i в локална околност с радиус αr:

$$\sigma n'(pi, r, \alpha) \equiv \sqrt{\frac{\sum_{p \varepsilon N(P_i, r)} \left(n(p, \alpha r) - n'(p_i, r, \alpha)\right)^2}{n(p_i, r)}}$$

За точки, чиято плътност на съседство съвпада със средната плътност на локалната околност стойността на MDEF близка до 0, докато за нетипичните данни стойността на MDEF е висока. Методът LOCI приема, че точка е нетипична когато:

$$MDEF_{(pi, r, \alpha)} > k_\sigma \sigma_{MDEF}(pi, r, \alpha) \text{ за } k_\sigma = 3\sigma_{MDEF},$$

където $\sigma_{MDEF}(pi, r, \alpha)$ е нормализиран MDEF коефициент:

$$\sigma_{MDEF}\,(p_i\,,\,r,\,\alpha) = \frac{\sigma n'(p_i\,,r,\alpha)}{n'(p_i\,,r,\alpha)}$$

Предимства: Преодолява се зависимостта от входен параметър за броя на най-близките съседи.

Недостатъци: Няма механизъм за намиране на оптимална стойност на радиуса на локалната околност за определяне на отклонение. Стойността на радиуса може да бъде потребителски дефиниран входен параметър или да бъде определена с помощта на алгоритъм за оптимизация.

3.2. Изводи от трета глава

В резултат на направения сравнителен анализ на локалните методи за откриване на нетипични стойности могат да бъдат направени следните по-важни изводи:

- Методът LOF не позволява откриване на нетипични стойности в поточни данни;
- Методът iLOF изисква всички данни да се съхраняват в паметта и не е приложим за работа в реално време при ограничения на наличната памет;
- Методът MiLOF позволява откриване на нетипични стойности в поточни данни, но определянето на отклонения зависи от потребителски дефиниран входен параметър;
- Методът LOCI премахва зависимостта от броя на най-близките съседи, но не предлага механизъм за намиране на оптимална стойност на радиус на ликална околност.

Модификация на метода LOCI с използване на оптимизационен алгоритъм за определяне на радиус на локалната област на съседство може да осигури ефективен подход за откриване на нетипични стойности при поточни данни.

ГЛАВА 4. ЕФЕКТИВЕН ПОДХОД ЗА ОТКРИВАНЕ НА НЕТИПИЧНИ СТОЙНОСТИ ПРИ ПОТОЧНИ ДАННИ

Предложени са два нови подхода за откриване на нетипични стойности при поточни данни, които са приложими при ограничения за изпозлваната памет и позволяват работа в реално време: Memory Efficient Outlier Detection (MEOD) и Advanced Memory Efficient Outlier Detection (A-MEOD).

4.1. Подход MEOD

Подходът MEOD се основава методите LOCI и MiLOF и използва оптимизация с рояк частици (PSO). Методът LOCI се използва за откриване на отклонения на базата на плътността на съседни точки, които се определят с дефиниране на радиус на локална околност. Определянето на радиуса на локалната околност се разглежда като оптимизационен проблем, който се решава с оптимизация с рояк частици. Съгласно алгоритъма за оптимизация с рояк частици за всяка частица i позицията ѝ X_i в пространството на търсене и скоростта на движението ѝ V_i се актуализират на всяка итерация на алгоритъма съгласно следната зависимост:

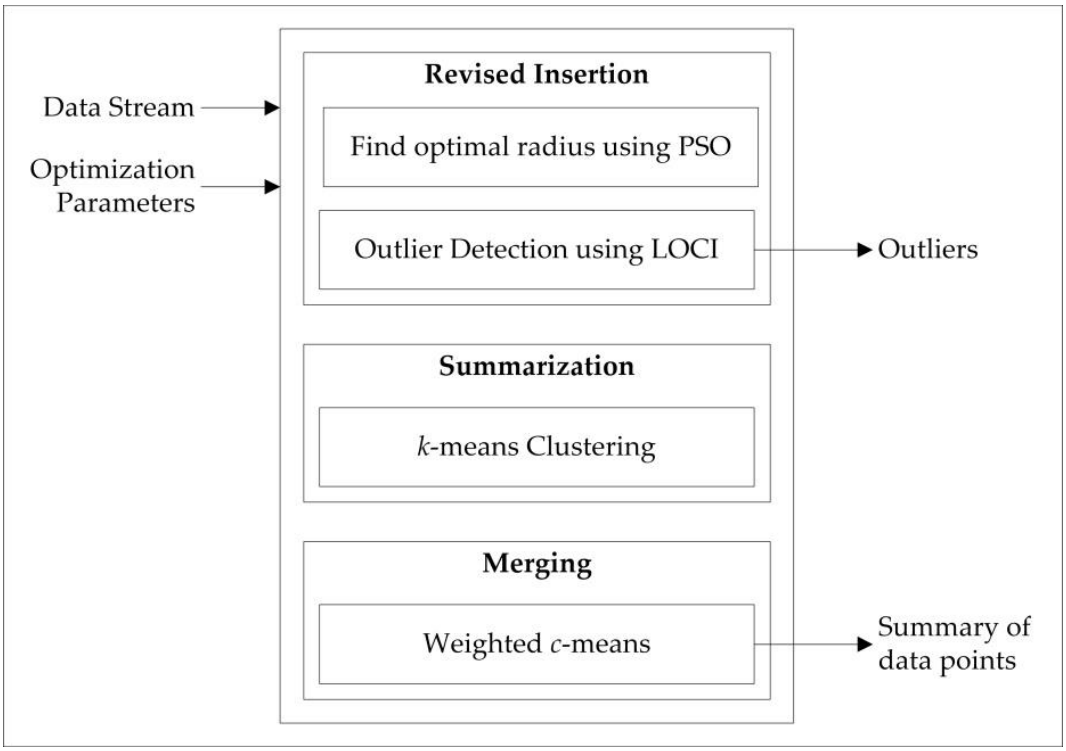
$$V_i = \chi X(V_i + \Theta_1 c_1 \left(X_i^{best} - X_i\right) + \Theta_2 c_2 \left(X_i^{gbest} - X_i\right))\text{-(4)}$$

където χ е коефициент на свиване (constriction factor), за който се препоръчва изпозлване на стойността 0.729, Θ_1 и Θ_2са коефициенти на ускорение (acceleration coefficients), c1 и c2 са произволни стойности в интервала [0, 1], X_i^{best} е най-добрата локална позиция на частица i, определена в процеса на търсене на оптимално решение, X_i^{gbest} е най-добрата глобална позиция, определена за всички частици в рояка.

В дисертационния труд се използва ринг-топология за обмен на данни между частиците, при която всяка частица е свързана с две съседни частици с цел избягване на локални оптимумални решения.

Предложеният подход MEOD за откриване на нетипични стойности в поточни данни използва оптимизация с рояк частици за определяне на оптимална стойност на радиуса r за изчисляване на MDEF за метода LOCI. Оптимизационната задача се свежда до определяне на точките, които имат минимално съотношение $\frac{k}{r}$. Фитнес функцията при прилагане на оптимизация с рояк частици се определя като $\frac{\alpha}{r\,X\,k} + \frac{k}{r} + \frac{k}{n-k}$, където α е константа, n е размерността на набора от данни, $\frac{\alpha}{r\,X\,k}$ е стойността за долна граница на стойността на r, а стойността $\frac{k}{n-k}$ определя горната граница на стойността на r. С използване на долна и горна граници се определя подходящата стойност на r: ако r има твърде малка стойност, то много малък брой съседни точки се използват при определяне на MDEF, а ако стойността на r е твърде висока, то се използват твърде много съседни точки, което влияе на точността на откриване на нетипични стойности.

Работният процес на техниката MEOD е представен на фиг. 4.1. Входните данни за прилагане на подхода MEOD са поточните данни от множеството, за което се определят нетипични стойности, както и предварително дефинирани параметри. Входните поточни данни се обработват с използване на плъзгащ се прозорец с определен размер. За потока данни b се идентифицират локални отклонения за всеки плъзгащ се прозорец. Изходните резултати са списък с определените нетипични стойности и обобщена информация за обработваните данни за избрания размер на плъзгащия се прозорец.



Фигура 4.1: Работен процес на подхода MEOD

За всеки плъзгащ се прозорец се обработват $\frac{b}{2}$ точки от потока данни с прилагане на следните три фази на обработване:

- Фаза „обобщение“ (Summarization)

При ограничения на използваната памет не е възможно да се съхраняват всички обработвани данни. Премахването на вече обработените данни от паметта не позволява

коректно определяне на нетипични стойности за следващите данни поради липса на история на данните, която се изисква за определяне на локални отклонения. Затова във фазата на обобщение се изчислява и съхранява обобщена оценка на предишните данни: за всеки плъзгащ се прозорец се обработват $\frac{b}{2}$ точки и се генерира обобщена информация с използване на клъстериране. Задава се голям брой клъстери и се определят центрове на клъстерите, които се запазват като обобщена информация за коефициент на отклонение MDEF и нормализиран коефициент на отклонение σ_{MDEF} при стойности на радиуса на локалната околност r. Останалите точки се премахват от паметта и се обработва следващ слот данни, определн от размера на плъзгащия се прозорец.

- Фаза „сливане“ (Merging)

Центровете на клъстерите, генерирани за данните от предишен плъзгащ се прозорец *i-1* и клъстерите, генерирани в текущия плъзгащ се прозорец *i* се обединяват в тази фаза и се запазва единствена стойност за клъстерните центрове с изпозлване на алгоритъм за клъстериране с претеглена средна стойност k-means.

- Фаза „вмъкване“ (Revised Insertion)

На тази фаза се определят нетипични стойности с изчислява на коефициента на отклонение MDEF с изпозлване на $\frac{b}{2}$ точки от текущия плъзгащ се прозорец и обобщените клъстерни центрове, запазени в паметта. Ако дадена точка се намира в локална околност с определния радиус за обобщените центрове на клустери, то тази точка не е нетипична стойност и за такива точки не се изисква изчисляване на коефициент на отклонение.

За определяне на нетипични стойности с предложения подход MEOD се използва метода LOCI. Алгоритъм 1 представя алгоритъм PSO-LOCI, който е модификация на метода LOCI с използване на оптимизация с рояк частици за определяне на оптимална стойност на радиуса r на локалната околност с фитнес функция, дефинирана съгласно метода LOCI. В стъпки 1 и 2 на алгоритъма частиците се инициализират със случайни стойности за позиция и скорост на движение. Процесът на оптимизация се изпълнява итеративно до достигане на максимален брой итерации. При итеративното изпълнение на стъпки 4 и 5 се определят k най-близки съседни частици и се изчислява стойността на фитнес функцията за всяка частица. С използване на тези стойности се актуализира глобалната най-добра позиция, локалната най-добра позиция и стойността на радиуса r в стъпки 7÷9. Позицията на всички частици в рояка се актуализира в стъпки 11÷13. Стойностите на коефициентите MDEF и σ_{MDEF} се изчисляват за всяка точка от набора данни с използване на определената оптимална стойност на радиуса r в стъпки 16÷18. Нетипичните стойности се определят на стъпка 19 чрез сравняване на стойностите на MDEF и σ_{MDEF} съгласно метода LOCI.

```
Алгоритъм 1: PSO-LOCI
```

```
Входни данни:  Number of points O1, O2,…, Ob
               Maximum Distance max_dist
               Number of particles P
Изходни данни: Outlier Points OP
               Radius r
```

```
Processing:
1. Initialize PSO parameters:
       X1_min = 0, X1_max = b, X2_min = 0, X2_max = max_dist
       V1_min = -10, V1_max = 10, V2_min= -1, V2_max = 1
```

```
2. For each particle P //Randomly initializes particles
       X1 = Random[X1_min, X1_max]
       X2 = Random[X2_min, X2_max]
       V1 = Random[V1_min, V1_max]
       V2 = Random[V2_min, V2_max]
   End For
3. For iteration = 0 to MaxIterations do
4.      For each particle P
5.         Find k for data point X1
6.         Calculate fitness value
7.         Assign r = X2
8.         Update local best values for particle X1_best, X2_best
9.         Update global best values for particle X1_best, X2_best
10.     End For
11.     For each particle
12.        Update the particle position by calculation the velocity
13.     End For
14. End For
15. r = X1_gbest
16. For each point O
17.     Calculate MDEF(oi, r, α) and σn'(oi, α, r)
18. End For
19. Find outlier points OP
```

Представяне на алгоритъма на предложения подход за определяне на нетипични стойности MEOD е показан в Алгоритъм 2. От входния поток данни O се използват b точки от данни и се определя отклонение за всяка от тях с алгоритъм PSO-LOCI (Алгоритъм 1). В стъпка 4 се определят обобщени данни за $\frac{b}{2}$ точки от данни и създават K клъстера с центрове Vi, като изчисленията се извършват за всеки плъзгащ се прозорец. Резултатите за текущия плъзгащ се прозорец се обединяват с тези за предишния плъзгащ се прозорец. При обединяването в стъпка 5 стойностите на MDEF (Vi, r, α) и σMDEF(Vi, r, α) се изчисляват за центровете Vi с използване на следните формули:

$$\mathrm{MDEF}(\mathrm{v_i}, \mathrm{r}, \alpha) = \frac{\sum_{p \in C_i} \mathrm{MDEF}(\mathrm{p_i}, \mathrm{r}, \alpha)}{|C_i|}$$

$$\sigma_{\mathrm{MDEF}}(\mathrm{v_i}, \mathrm{r}, \alpha) = \frac{\sum_{p \in C_i} \sigma_{\mathrm{MDEF}}(\mathrm{p_i}, \mathrm{r}, \alpha)}{|C_i|}$$

където $|C_i|$ е броя на точките в клъстер C_i. Стойностите на центровете са средно аритметичната стойност на MDEF и σ_{MDEF} за всички точки от клъстера. В стъпка 6 се запазват определените стойности на центровете на клъстерите и $\frac{b}{2}$ точки се премахват от паметта. В стъпка 7 се прилага алгоритъма за претеглени средни стойности c-means, след което центровете на клъстерите отново се актуализират със стойностите за коефициентите MDEF и σ_{MDEF} в стъпка 8. В стъпка 9 предишните стойности на центровете на клъстерите се изтриват от паметта и се актуализират обобщените данни.

```
Алгоритъм 2: MEOD
Входни данни:  Set of data points O (O1, O2, …, On)
               Window Size b
Изходни данни: Outlier Points OP
               Data summary (Z, W)
Processing:
1.   While O is not empty
2.      If number of data points = b then
3.         Find outlier using PSO-LOCI algorithm
4.         Apply K-means for b/2 data points and get cluster centers Vi
           and cluster member count Ni
5.         Compute MDEF(pi, r, α) and σMDEF(pi , r, α) for all centers
6.         Remove b/2 points from memory
7.         Apply weighted c-means for all cluster centers
           and get updated cluster centers Zi and cluster weights Wi
8.         Compute MDEF(pi , r, α) and σn'(pi, α, r) for all updated
           cluster centers Zi
9.         Remove Zi-1 and Wi-1 points from memory
10.    End If
11. End While
```

4.2. Подход A-MEOD

Подходът MEOD премахва зависимостта от броя на най-близките съседи, но увеличава изчислената сложност. Предложеният подход A-MEOD подобрява изчислителната ефективност на MEOD чрез намаляване на времето за изпълнение за намиране на оптимална стойност на радиуса на локалната околност на всяка итерация на MEOD.

Подходът A-MEOD не изчислява коенфициента MDEF за определяне на нетипични стойности, а изисква определяне на точки с ниска плътност на съседство като използва дефиниция на Knorr за определяне на отклонения, съгласно която отклоненията се определя с използване на функция за разстояние. Според дефиницията на Knorr точка O е отклонение ако (1–B) от точките в потока данни са на разстояние от O по-голямо от радиус r. Радиусът r е праг на разстоянието за определяне на нетипична стойност. При използване на дефиницията на Knorr не се изисква специфичен модел за разпределение на данните за откриване на отклонения и нетипичнии стойности могат да бъдат определени и за многомерни набори от данни. Параметрите r и B са входни потребителски дефинирани стойности.

Целта на оптимизационната задачата при подхода A-MEOD е да се минимизира стойността на k/r, където k е броя на съседните данни и се изменя с промяна на радиуса на локалната околност r. В предложения подход се използва стойността на радиуса r, определена чрез фитнес функцията за определяне на отклонения. На базата на изчислената стойност за r се изчислява съотношението k/r за другите точки и те се подреждат във възходящ ред за да се идентифицират най-големите k на брой нетипични стойности.

Подходът A-MEOD работи върху поточни данни. Стойността на радиуса на локалната околност r може да бъде променена за всеки плъзгащ се прозорец в зависимост от разпределението на данните, поради което при използването на отпимизация с рояк частици

се прилага следната модификация спрямо подхода MEOD. Инициализирането на параметърите за PSO се извършва на базата на данните от предишния плъзгащ се прозорец. За данните от първия плъзгащ прозорец позицията на частиците и скорост на движението им се инициализират със случани стойности и се изпълнява PSO са намиране на оптимално положение на частиците. Позицията на частиците се запазва заедно с обобщени данни за точките в плъзгащия се прозорец. В следващия плъзгащ се прозорец частиците се инициализират с предишната им позиция от предходния плъзгащ се прозорец, с което се намалява броя на итерациите за намиране на оптималната стойност на радиуса r за всеки плъзгащ се прозорец.

Използването на оптимизация с рояк частици в предложения подход A-MEOD е показано в Алгоритъм 3. В стъпки 1 до 5 се прави инициализация на позиции и скорост на частиците. В стъпки 6 до 16 позициите и скоростите на частиците се актуализират с използване на фитнес функцията за PSO. Стойността на радиуса r се актуализира на всяка итерация и оптималната стойност се определя на стъпка 18. Отношението k/r за всяка точка от набора данни се изчислява в стъпки 19 до 21. В стъпка 22 точките се сортират във възходящ ред по стойноста на k/r и първите m стойности с най-голяма стойност на отношението k/r се определят като нетипични стойности в резултат на работата на алгоритма.

```
Алгоритъм 3: PSO based outlier detection
```

```
Входни данни:  Number of points O1, O2, …, Ob
               Maximum Distance max_dist
               N: Number of particles P
               Particle Summary PS = {PS1, PS2,..PSN}
               Where PSi = {X1_prev, X2_prev, V1_prev, V2_prev}
Изходни данни: Top m Outlier Points OP
               Radius r
               Updated Particle Summary PS
```

```
Processing:
1. Initialize PSO parameters:
       X1_min = 0, X1_max = b, X2_min = 0, X2_max = max_dist
       V1_min = -10, V1_max = 10, V2_min = -1, V2_max = 1
2. For each particle P //Randomly initializes particles
3.     IF PS is NULL
           X1 = Random[X1_min, X1_max]
           X2 = Random[X2_min, X2_max]
           V1 = Random[V1_min, V1_max]
           V2 = Random[V2_min, V2_max]
4.     ELSE
           X1 = X1_prev
           X2 = X2_prev
           V1 = V1_prev
           V2 = V2_prev
5. End For
```

```
6. For iteration = 0 to MaxIterations do
7.     For each particle P
8.         Find k for data point X1
9.         Calculate fitness value
10.        Assign r = x2
11.        Update local best values for particle X1_best, X2_best
12.        Update global best values for particle X1_gbest, X2_gbest
13.    End For
14.    For each particle
15.        Update the particle position by calculation the velocity
16.    End For
17. End For
18. r = X1_gbest
19. For each point O
20.     Compute k/r for O
21. End For
22. Sort the points using k/r
23. Return top m outlier points OP
```

Представяне на алгоритъма на предложения подход за определяне на нетипични стойности A-MEOD е показан в Алгоритъм 4. На стъпка 2 се определят поточните данни за текущяи плъгащ се прозорец. Нетипичните стойности се определят с използване на Алгоритъм 3 в стъпка 3. В стъпки 4 до 6 се изпълнява фазата за обобщаване на данните като се прилага алгоритъм за клъстеризация K-means и средно аритметичната стойност k/r се присвоява на всеки център на клъстер. Средната стойност на отношението k/r за всички центрове v_i се изчислява като:

$$\mathrm{Avg}(\mathrm{v_i}\,, \mathrm{k/r}) = \frac{\sum_{p \in C_i} \mathrm{p}(\mathrm{k/r})}{|C_i|}$$

След генериране на обобщените данни за текуия плъзгащ се прозорец $\frac{b}{2}$точки се премахват от паметта. След това обединяването се изпълнява на стъпки 7 до 9. При обединяването се прилага алгоритъм за претеглена клъстеризация c-means и се обновява теглото за центъра на клъстера. Стъпки от 2 до 10 се изпълняват отново за всеки плъзгащ се прозорец.

Алгоритъм 4: A-MEOD

```
Входни данни:    Set of Data points (O1, O2, ..., On)
                 Window Size b
Изходни данни:   Outlier Points OP
                 Data summary (Z, W)
```

```
Processing:
1. While O is not empty
2.     If number of data points = b then
```

```
3.          Find Outlier using PSO based outlier detection algorithm
4.          Apply K-means for b/2 data points and get cluster centers Vi
            and cluster member count Ni
5.          Compute average k/r value for all centers
6.          Remove b/2 points from memory
7.          Apply weighted c-means for all cluster centers
            and get updated cluster centers Zi and cluster weights Wi
8.          Compute k/r value for all updated cluster centers Zi
9.          Remove Zi-1 and Wi-1 points from memory
10.     End If
11. End While
```

4.3. Изводи от четвърта глава

Въз основа на анализа на различни локални методи за откриване на нетипични стойности са предложени два подхода нови подхода за откриване на нетипични стойности при поточни данни, които са приложими при ограничения за изпозлваната памет и позволяват работа в реално време: Memory Efficient Outlier Detection (MEOD) и Advanced Memory Efficient Outlier Detection (A-MEOD). При двата подхода се използват методите MiLOF и LOCI и се прилага оптимизация с рояк частици. Предимствата на двата предложени подхода могат да бъдат обощени както следва:

- Предимства на подхода MEOD:

– подходът MEOD позволява откриване на локални нетипични стойности за поточни данни като използва метода LOCI;

– откриването на локални нетипични стойности позволява изпълнение при ограничени ресурси за наличната памет, използвайки механизъм за обобщаване на данните;

– прилага се оптимизация с рояк частици за определяне на оптимална стойност за радиуса r на използваната локална околност при изчисленията за откриване на нетипични стойности чрез метода LOCI;

– подходът използва определяне на коефициент на отклонение само за данните, които са потенциални нетипични стойностите, а не за целия набор от данни, с което се подобрява изчислителната ефективност спрямо други локални методи за откриване на нетипични стйности;

– откриването на локални нетипичнии стйности е с добра точност, сравнима с тази при изпозлване на метода MiLOF;

– предложеният подход MEOD е широко приложим за откриване на локални нетипични стойности за поточни данни;

- Предимства на подхода A-MEOD:

– подходът A-MEOD позволява откриване на локални нетипични стойности за поточни данни като използва отношението на локалната плътност k/r;

– откриването на локални нетипични стойности позволява изпълнение при ограничени ресурси за наличната памет, използвайки механизъм за обобщаване на данните;

- прилага се оптимизация с рояк частици за определяне на оптимална стойност за радиуса r на използваната локална околност при изчисленията за откриване на нетипични стойности чрез метода LOCI;
- при откриването на локални нетипични стойности се премахва зависимостта от стойността k на най-близките съседни данни от набора, която е потребителски дефиниран входен параметър, използван при други локални методи за откриване на нетипични стйности с влияние върху точността им;
- подходът A-MEOD е по-ефективен по отношение на изпозлваната памет в сравнение с MEOD и MiLOF;
- подходът A-MEOD е по-ефективен по отношение на времето за изчисления в сравнение с MEOD и MiLOF;
- при малка стойност на броя на определените нетипични стойности точността при използване на подхода A-MEOD откриване на локални нетипични стойности за поточни данни клони към 1.

ГЛАВА 5. ИМПЛЕМЕНТАЦИЯ НА ПРЕДЛОЖЕНИТЕ ПОДХОДИ И ЕКСПЕРИМЕНТАЛНИ РЕЗУЛТАТИ

5.1. Имплементация на предложените подходи

За експериментална оценка на двата предложени подхода за откриване на нетипични стойности е използвана имплементация на Python-3.8 с IDE Netbeans 11.2. Експерименталната оценка е направена на машина с CPU Intel i3, 4GB RAM, ОС Windows.

5.2. Използвани множества данни

За експерименталната оценка са използвани различни набори от данни от UCI и Kaggle (табл. 5.1). В допълнение на тези множества данни са използвани синтетични двумерни набори от данни, генерирани с използване на данни с Гаусово разпределение с различни размерности, дисперсия и ковариация, който са ръчно анотирани. В табл. 5.2 са представени изпозлваните параметри за оптимизацията с PSO.

За определяне на разстояние между две точки се използва Евклидово разстояние. Първоначално стойностите на атрибутите на набора от данни се нормализират в интервала [0, 1] както следва:

$$f_i = \frac{f_i - f_{i\min}}{f_{i\max} - f_{i\min}}$$

където f_i е атрибутна стойност, а $f_{i\max}$ и $f_{i\min}$ са минималната и максималната стойност за f_i.

Таблица 5.1. Набори от данни, използвани за експериментална оценка

No.	Набор от данни	Брой точки (n)	Размерност на данните (D)
1.	UCI Vowel (Vl)	1040	10
2.	UCI glass	214	10
3.	UCI Pendigit (Pt)	3 600	16
4.	IBRL	3 000	2
5.	Kaggle wine	177	2

Таблица 5.2. Параметри за оптимизацията с PSO

Параметър	Стойност
Брой частици	30
Максимален брой итерации	1000
Фактор на свиване	0.729
c1 и c2	2.02
Размер на прозореца	1000
Брой клъстери	50
Брой итерации за алгоритъм k-means	100
Брой итерации за алгоритъм c-means	10

5.3. Експериментални резултати

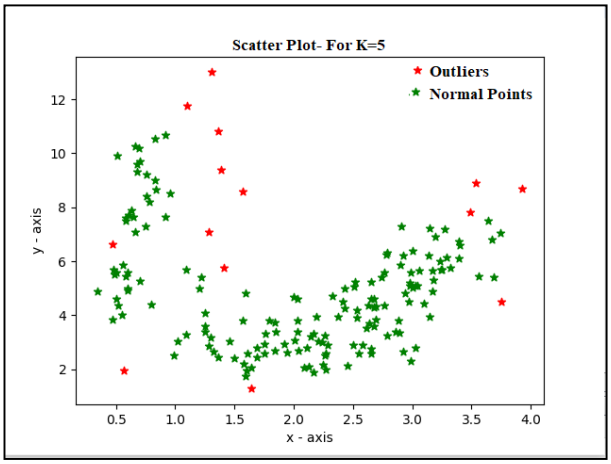
За оценяване на ефективността на предложените подходи MEOD и A-MEOD са проведени следните експерименти: оценка на точност при откриване на нетипични стойности, оценка на влиянието на параметъра K върху радиуса R, оценка на време за изпълнение и използвана памета, оценка на време за изпълнение при различен размер на плъзгащия се прозорец.

- Визуална оценка на точност при откриване на нетипични стойности

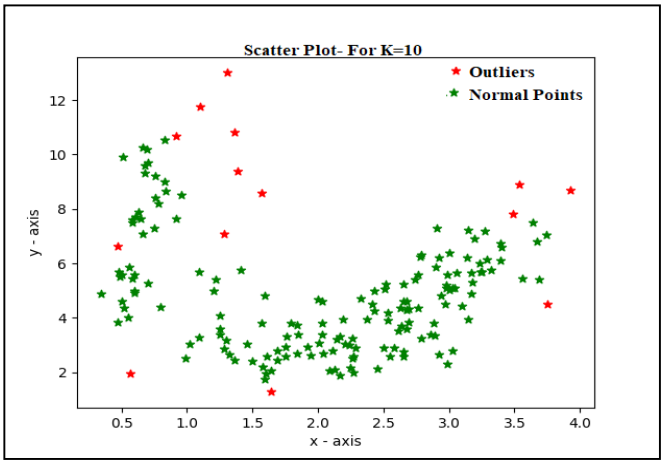
Точността на откриване на нетипични стойности с метода MiLOF зависи от стойността на параметъра K, използван за изчисляване на плътността на съседните точки. Резултатите за набора от данни Kaggle wine, получени за различни стойности на K са показани в scatter-plot диаграми на фиг. 5-1 ÷ 5-4: с увеличаване на стойността на K се откриват по-малко нетипични стойности. С използване на предложения подход MEOD се премахва зависимостта от стойността потребителски дефинирана стойност за K, тъй като се определя оптимална стойност на радуиуса на локалната околност (фиг. 5.5). Подобно на MEOD откриването на нетипични стойности при използване на подхода A-MEOD не зависи от стойността на K – на фиг. 5-6 са показани определените 20 за набора данни Kaggle wine.

- Оценка на точност при откриване на нетипични стойности

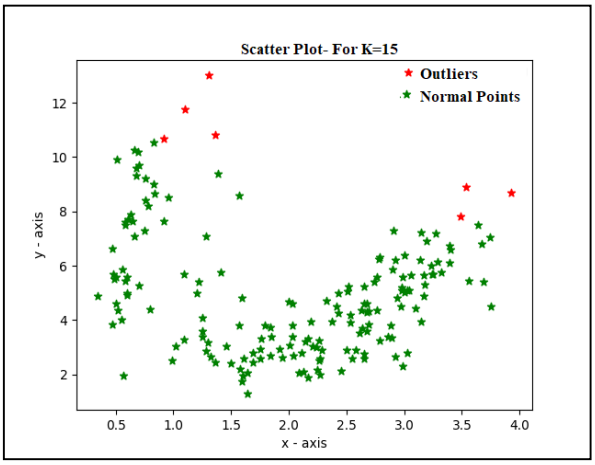
За оценка на точността при откриване на нетипични стойности с предложените подходи MEOD и A-MEOD са използвани три двумерни синтетични набора от данни. На фиг. 5.7 е показана точността за синтетичните набори от данни с използване на MEOD при различни стойности на параметъра K, а на фиг. 5.8 е дадена точността при използване на A-MEOD. MEOD има по-висока точност при откриване на нетипични стойности в сравнение с MiLOF за два от синтетичните 2D набори от данни и точността не се влияе от стойността на входен параметър. Точността на A-MEOD зависи от броя на най-добрите отклонения, като с увеличаването им точността намалява, а за малък брой точността на клони към 1.



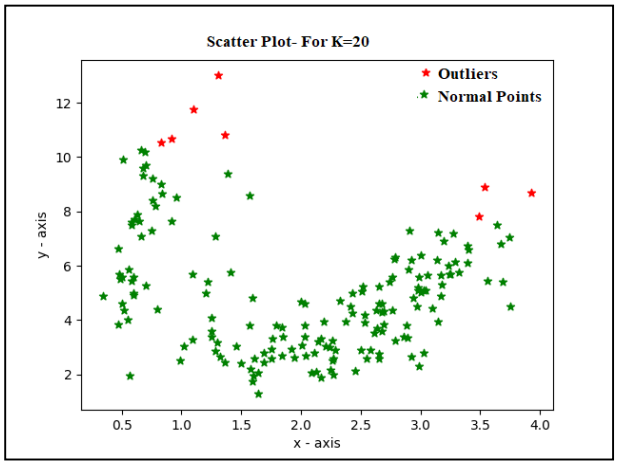
Фигура 5.1: Резултати с използване на MiLOF за K=5



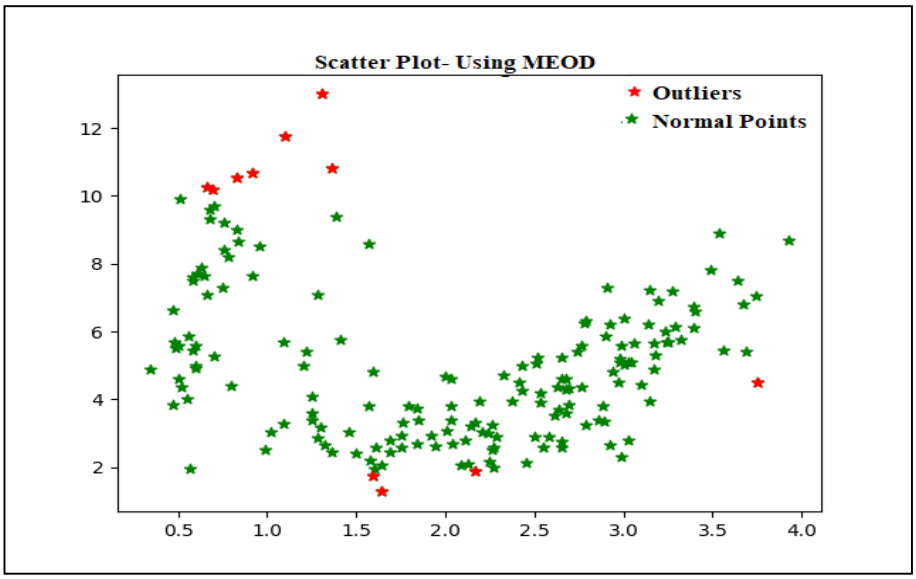
Фигура 5.2: Резултати с използване на MiLOF за K=10



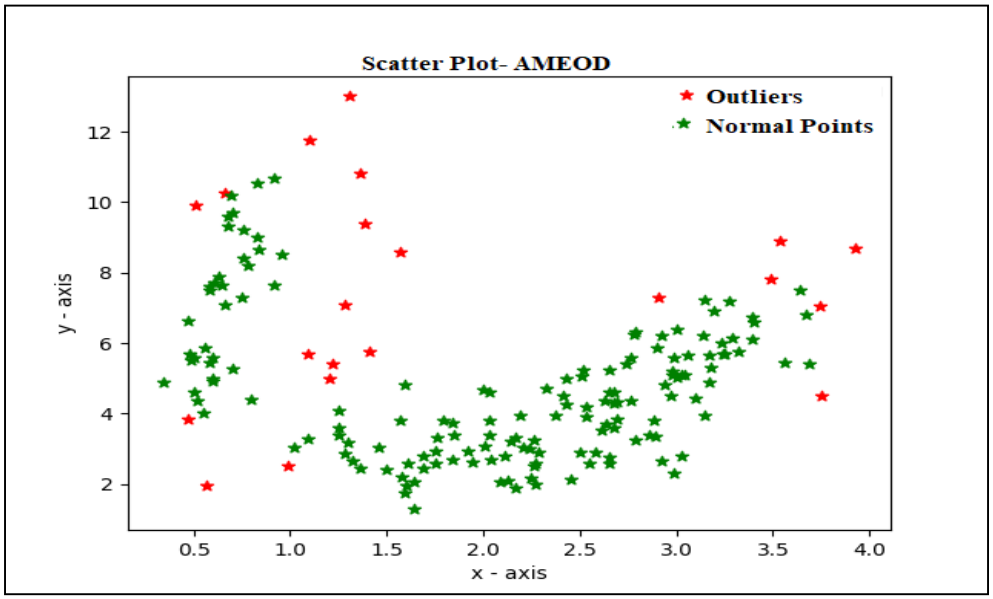
Фигура 5.3: Резултати с използване на MiLOF за K=15



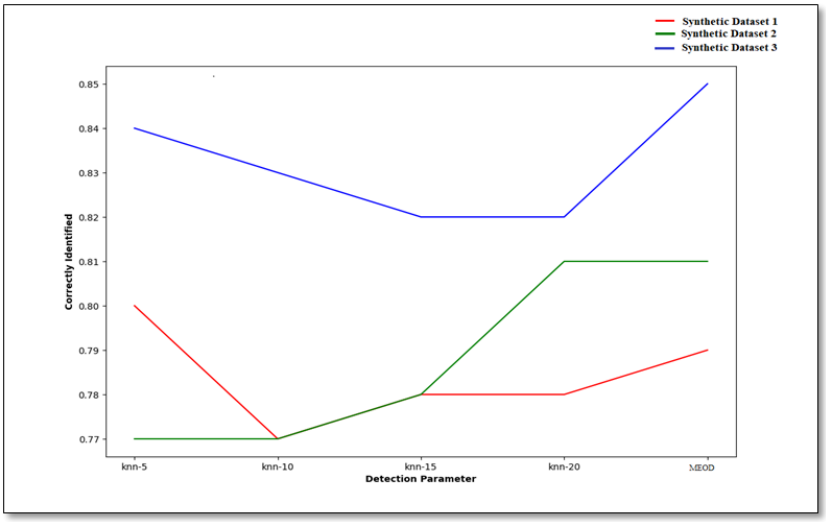
Фигура 5.4: Резултати с използване на MiLOF за K=20



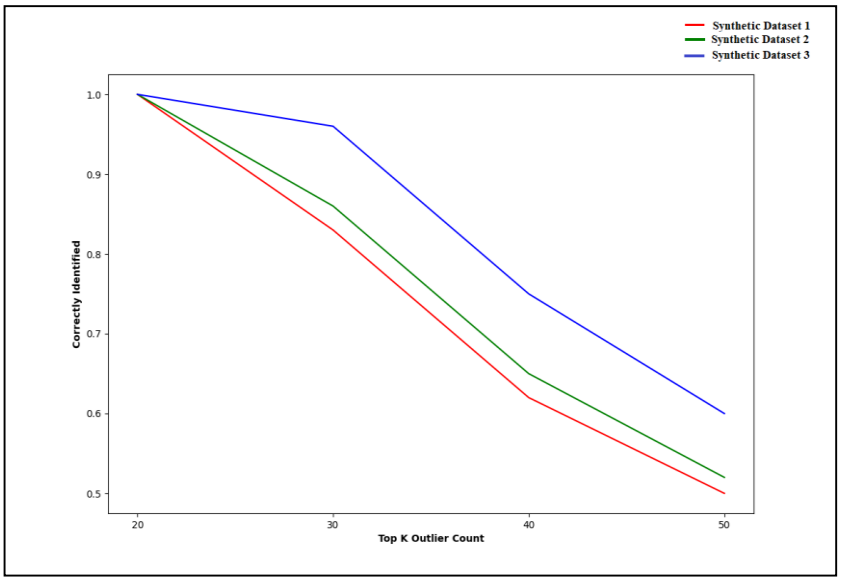
Фигура 5.5: Резултати с използване на MEOD за набора данни Kaggle wine



Фигура 5.6: Резултати с използване на A-MEOD за набора данни Kaggle wine



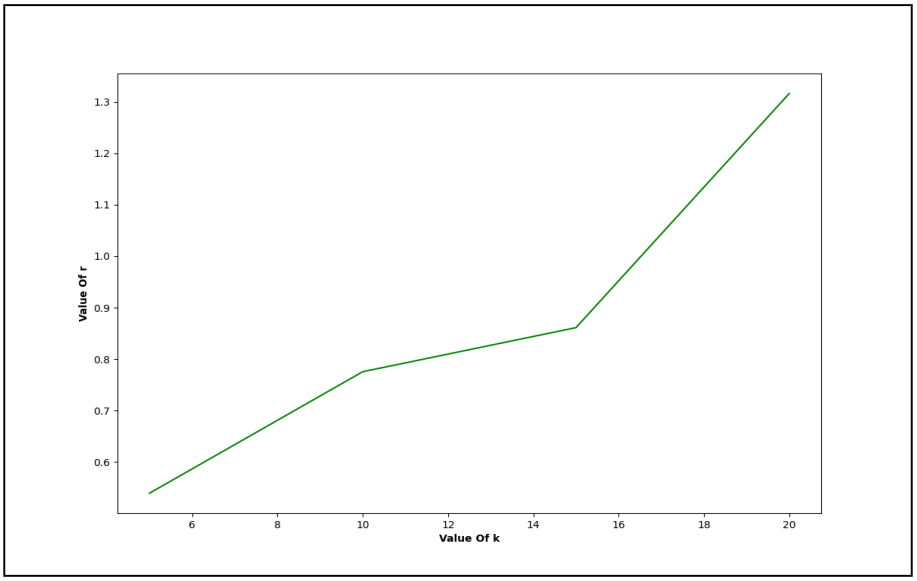
Фигура 5.7: Точност при използване на MEOD за синтетични набори от данни



Фигура 5.8: Точност при използване на A-MEOD за синтетични набори от данни

- Оценка на влиянието на параметър k върху радиуса r

Откриването на отклонения с подходите MEOD и A-MEOD не зависи от стойността на параметъра k. Направена е оценка на влиянието на стойността на k върху определената оптимална стойност на радиуса r на локалната околност. На фиг. 5.9 са показани резултати за определената оптимална стойност на R при различни стойности на k за набора данни Kaggle wine. Стойността на r се увеличава при увеличаване на броя най-близки съседи k, но от друга страна промяната на стойността на радиуса води до определяне на едни и същи отклонения, т.е. точността на MEOD и A-MEOD е независима от стойността на k.



Фигура 5.9: Стойност на радиуса r при различни стойности на k за набора от данни Kaggle wine

- Оценка на време за изпълнение на методите MiLOF, MEOD и A-MEOD

Времето, необходимо за определяне на нетипични стойностис използване на MiLOF, MEOD и A-MEOD е оценено за наборите данни UCI Vowel, UCI glass, UCI Pendigit, IBRL. На фиг 5.10 са показани резултати за времето за изпълнение при различните набори от данни за MiLOF, MEOD и A-MEOD. При MEOD се изисква повече изчислително време в сравнение с MiLOF, тъй като се определя оптимална стойност на радиуса r и след това се изчисляват MDEF и σ_{MDEF}. Този недостатък на MEOD се преодолява с подхода A-MEOD, при който времето за изчисления за оценка на MDEF и σ_{MDEF} намалява.

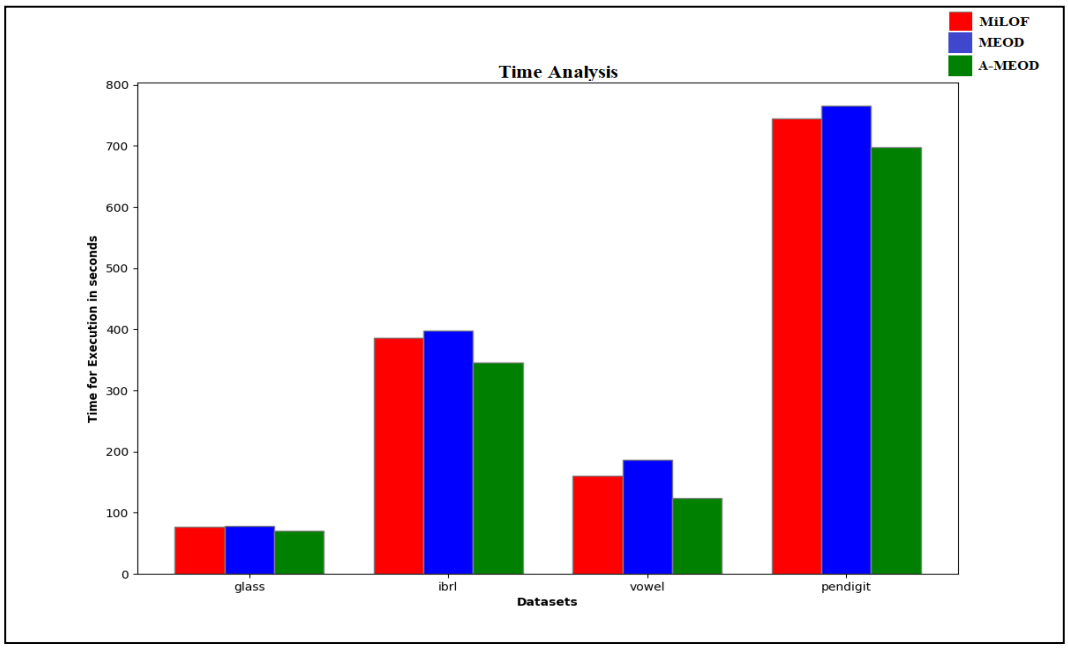
- Оценка на използвана памет на методите MiLOF, MEOD и A-MEOD

На фиг. 5.11 е показано сравнение на изискваната памет при определяне на нетипични стойности за за наборите данни UCI Vowel, UCI glass, UCI Pendigit, IBRL с използване на MiLOF, MEOD и A-MEOD. При подхода MEOD след намиране на оптималната стойност на радиуса r се съхраняват само данни за разстоянието до съседните точки в определената локална околност, а не за всички данни в набора. При подхода A-MEOD се определя единствено съотношението k/r за всяка точка, с което не само се намалява изчислителната сложност, но и изскваната памет.

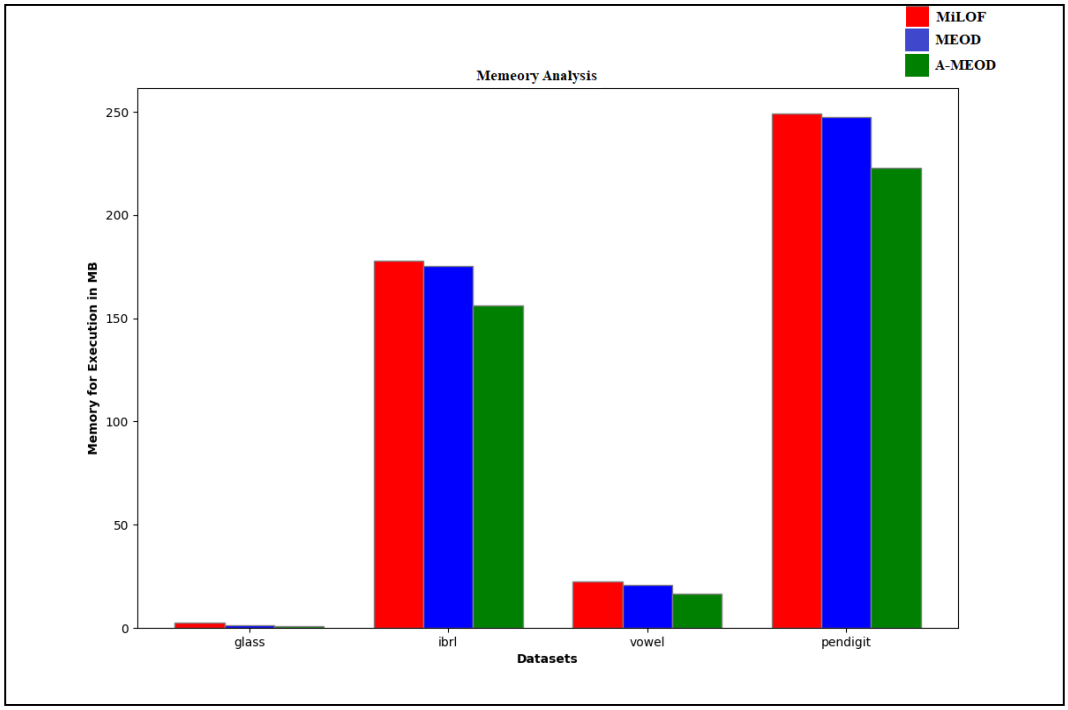
- Оценка на време за изпълнение при различен размер на плъзгащ се прозорец

Влиянието на размера на плъзгащия се прозорец върху времето за изчисления и необходимата памет при откриване на нетипични стойности с MiLOF, MEOD и A-MEOD е оценено за синтетичните набори данни. На фиг. 5.12 и фиг. 5.13 са показани резултати за времето за изпълнение и използваната памет при различни размери на плъзгащия се прозорец.

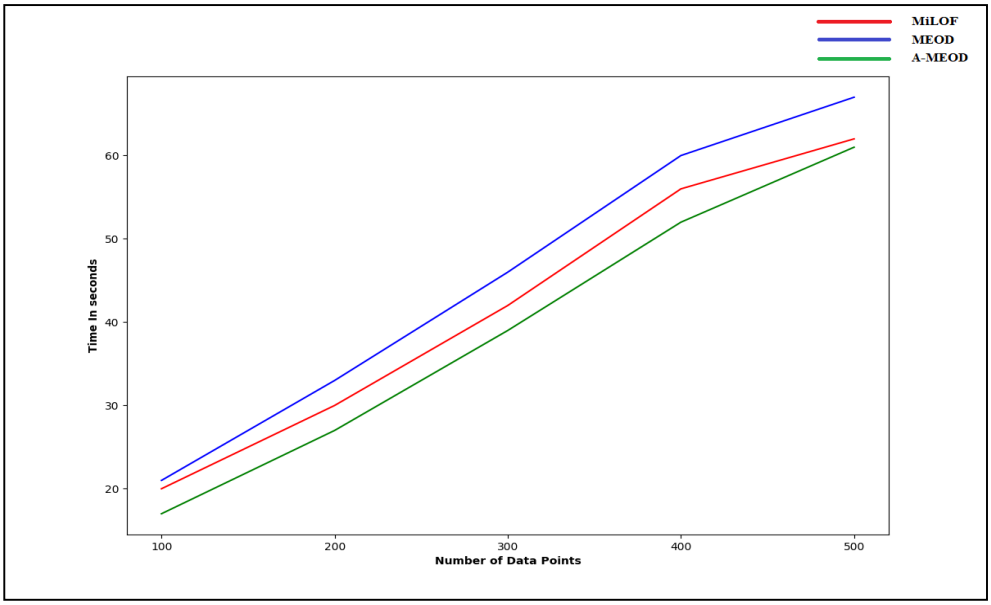
При увеличаване на размера на плъзгащия се прозорец се увеличава броя използвани точки за определяне на нетипични стойности, което води до увеличаване на времето за изпълнение и на използваната памет. При подхода MEOD се изисква повече време за изпълнение в сравнение с MiLOF и A-MEOD като времето нараства линейно при увеличаване на размера на плъзгащия се прозорец. Използваната памет при MEOD е по-малко в сравнение с MiLOF, а от своя страна A-MEOD е ефективен подход по отношение на изпозлваната памет в сравнение с MEOD.



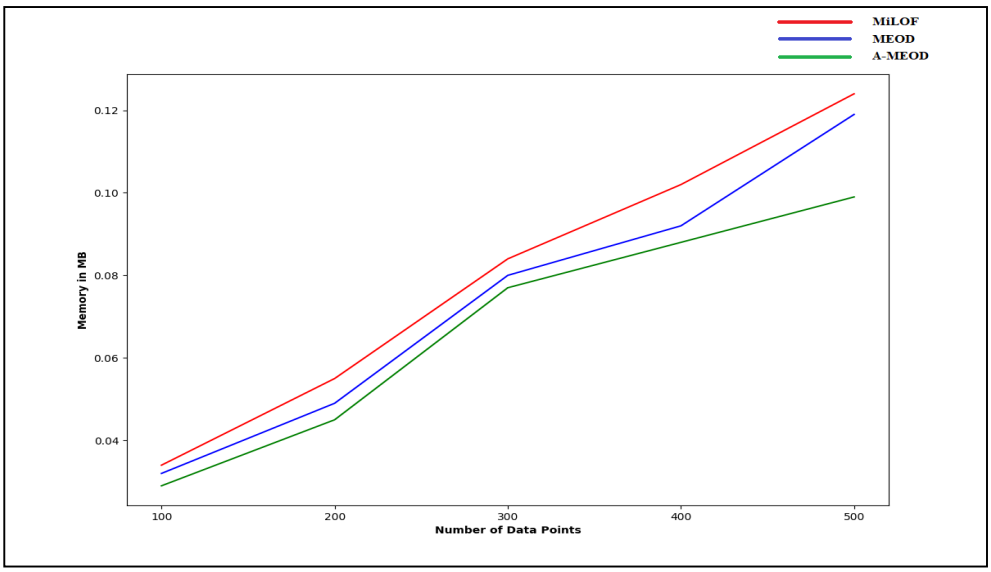
Фигура 5.10: Сравнение на време за изпълнение с използване на MiLOF, MEOD и AMEOD за набори от данни UCI



Фигура 5.11: Сравнение на необходима памет с използване на MiLOF, MEOD и AMEOD за набори от данни UCI



Фигура 5.12: Сравнение на време за изпълнение при използване на MiLOF, MEOD и A-MEOD за синтетични набори данни при различен размер на плъзгащия се прозорец



Фигура 5.13: Сравнение на използвана памет при MiLOF, MEOD и A-MEOD за синтетични набори данни при различен размер на плъзгащия се прозорец

5.4. Изводи от пета глава

Експерименталната оценка на ефективността на предложените подхдои MEOD и A-MEOD за откриване на нетипични стойности в поточни данни се базира на оценяване и анализ на няколко показателя за ефективност върху различни набори данни. На базата на направените експериментални изследвания за сравнение на производителността на методите MiLOF, MEOD и A-MEOD се потвърждават теоретичните твърдения за ефективността на предложените подходи MEOD и A-MEOD при определяне на отклонения в поточни данни. Използването MEOD и A-MEOD води до получаване на резултати с висока точност и устойчивост по отношение на входните множества данни и използваните параметри с ефективно използване на памет за съхраняване на неоходимите данни и по-добра изчислителна производителност в сравнение със съществуващите локални методи за откриване на нетипични стойности.

НАУЧНО-ПРИЛОЖНИ И ПРИЛОЖНИ ПРИНОСИ

Научно-приложни приноси:

- Предложен е нов подход за откриване на локални отклонения в поточни данни: Memory Efficient Outlier Detection (MEOD), който комбинира MiLOF и LOCI и използва техника за оптимизация с рояк частици и обобщаване на данните. MEOD се базира на оптимизация с рояк частици за определяне на оптимална стойност за радиуса r при алгоритъма LOCI, премахва зависимостта от броя на най-близките съседи при клъстеризация с алгоритъм k-NN и използва подход за определяне на стойност на фактора за нетипичност само за кандидат-стойностите, а не за целия набор от данни. С използване на предложения подход MEOD откриването на нетипични стойности в поточни данни е с по-добра точност, изисква по-малко изчислително време и използва по-малко памет в сравнение с алгоритъма MiLOF.

- Предложен е нов подход за откриване на локални отклонения в поточни данни: Advanced Memory Efficient Outlier Detection (A-MEOD), който използва техника за оптимизация с рояк частици за определяне на оптимална стойност за радиуса r, премахва зависимостта от броя на най-близките съседи при клъстеризация с алгоритъм k-NN и подобрява изчислителната сложност на MEOD чрез определяне само на най-големите M отклонения при алгоритъма LOCI на базата на съотношението за локална плътност k/r. A-MEOD изисква по-малко памет и по-малко време за изпълнение в сравнение с MEOD и MiLOF при откриване на нетипични стойности в поточни данни.

Приложни приноси:

- Направено е обстойно проучване на методите за откриване на нетипични стойности, като методите са структурирани и групирани в различни типове според използвания подход, определени са предимствата и недостатъците на всеки от подходите и е направено сравнение на ефективността им при откриване на нетипични стойности.

- Направено е проучване и анализ на подходите за решаване на оптимизационни проблеми с използване на различни техники за оптимизация, въз основа на което е избран подход за оптимизация с интелигентност на рояка като подходяща техника за оптимизация, която може да бъде използвана при откриване на нетипични стойности. Направено е проучване на съществуващи техники за оптимизация с интелигентност на рояка и е мотивиран избора на оптимизация с рояк частици (Particle Swarm Optimisation, PSO) като най-подходяща техника при решаване на задачата за откриване на нетипични стойности.

- Въз основа на анализ на различни техники за откриване на локални отклонения са избрани алгоритмите MiLOF и LOCI, които могат да бъдат приложени за поточни данни, тъй като отговарят на изискванията и преодоляват ограниченията на алгоритмите LOF и iLOF при работа с поточни данни. На тази база е предложено комбиниране на алгоритъма LOCI с техника за оптимизация с рояк частици за да постигне ефективно по отношение на използваната памет откриване на локални отклонения в поточни данни.

СПИСЪК НА ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИОННИЯ ТРУД

1. **Ankita Karale**, Outlier Detection Methods and the Challenges for their Implementation with Streaming Data, *Journal of Mobile Multimedia*, Vol. 16, No. 3, 2020, https://doi.org/10.13052/jmm1550-4646.1635 (индексирана в Scopus).

2. **Ankita Karale**, Milena Lazarova, Pavlina Koleva, Vladimir Poulkov, A Hybrid PSO-MiLOF Approach for Outlier Detection in Streaming Data, *Proc. of 43rd International Conference on Telecommunications and Signal Processing (TSP'2020)*, Milan, Italy, 7–9 July 2020, pp. 474–479, https://doi.org/10.1109/TSP49548.2020.9163430 (индексирана в Scopus и Web of Science).

3. **Ankita Karale**, Milena Lazarova, Pavlina Koleva, Vladimir Poulkov, MEOD: Memory-Efficient Outlier Detection on Streaming Data, *MDPI Journal Symmetry*, Vol. 13, No. 3: 458, https://doi.org/10.3390/sym13030458 (индексирана в Scopus и Web of Science, IF 2.645).

4. **Ankita Karale**, Milena Lazarova, Pavlina Koleva, Vladimir Poulkov, Advanced Memory Efficient Outlier Detection Approach for Streaming Data using Swarm Optimization, *Proc. of 44th International Conference on Telecommunications and Signal Processing (TSP'2021)*, 26 – 28 July 2021 [приета за публикуване, с индексиране в Scopus и Web of Science).

SUMMARY

Efficient Outlier Detection in Streaming Data Using Metaheuristics Optimization

Ankita Vitthal Karale

The detection of outliers is an important research problem with application in many different areas requiring data processing and extraction. The problem of outlier detection requires deviations to be determined in the expected or typical behaviour of the data due to anomalies, defects, errors, damage, noise, etc. The application areas of outlier detection problem include detection of unusual events in data analysis in sensor networks and distributed systems, financial, meteorological and environmental analysis, data protection analysis and many others. Many different approaches methods and algorithms exist for outlier detection in static data sets with a finite number of values. The outlier detection in streaming data with dynamic nature, high generation speed and requirement for real time processing with limited computational resources and memory capacity is a research problem of many scientific papers and efforts to suggest efficient approaches and methods.

Based on the state-of-the-art in the research approaches the objective of the thesis is to design and develop efficient approach for outlier detection which can work with real time streaming data in limited memory environment constraints and is able to provide reliable solution which with improved accuracy and minimize required execution time.

Two new approaches for outlier detection in streaming data are suggested: Memory Efficient Outlier Detection (MEOD) and Advanced Memory Efficient Outlier Detection (A-MEOD). MEOD approach combines MiLOF and LOCI methods, uses swarm intelligence optimization and data aggregation. Based on particle swarm optimization an optimal value for the radius of the local neighbourhood in LOCI algorithm is determined and thus the dependence on the number of nearest neighbours for clustering using k-nearest neighbour algorithm is eliminated. The outlier factor value for only candidate points rather than the whole dataset is applied to improve the efficiency of the algorithm. The proposed MEOD approach detects outliers over streaming data with good accuracy, computational time and memory requirements compared to MiLOF approach. A-MEOD approach uses swarm intelligence optimization to determine the optimal value for the radius of the local neighbourhood, eliminates the dependence on the number of nearest neighbours for clustering with the k- nearest neighbour algorithm and improves the computational complexity of MEOD approach by detecting only the top outliers using LOCI algorithm based on the k/r local density ratio. The proposed A-MEOD approach requires less memory and reduces the execution time compared to MiLOF and MEOD when applied for outlier detection over streaming data.

The proposed approaches for local outlier detection in streaming data can be used to determine anomalies and deviations in systems where data is dynamically generated, computations are performed in real time with limited computing resources and memory capacity and high accuracy in outlier detection is required.