

Университета в Констанц

Виолета Николаева Иванова-Ролинг

Безкомуникационни стратегии за Widening

Автореферат

Научен ръководител:

проф. Михаел Бертолд

Рецензенти:

проф. Михаел Бертолд

проф. Свен Косуб

Публикация на дисертацията:

URI <http://nbn-resolving.de/urn:nbn:de:bsz:352-2-kw8k2dxtevg26>

Защитата на дисертационния труд се е състояла на 05.10.2019 в Университета в Констанц, гр Констанц. Дисертацията е публикувана на уебсайта на университета URI: <http://nbn-resolving.de/urn:nbn:de:bsz:352-2-kw8k2dxtevg26>

Дисертационният труд съдържа 175 страници, в които 12 страници литература, включваща 155 заглавия.

Автор: Виолета Николаева Иванова-Ролинг

Заглавие: Стратегии за Widening без комуникация

Глава 1

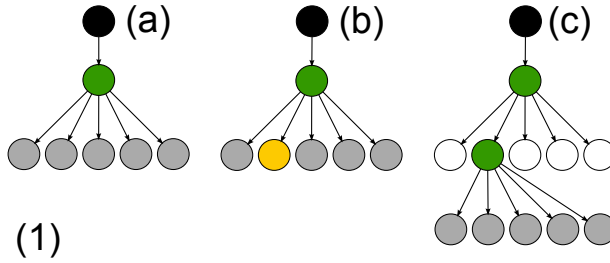
Увод и основни понятия.

Живеем в ерата на свръхдостъпните изчислителни ресурси, и като следствие интензивни изследвания в областта на паралелните алгоритми за Дейта айнинг и машинно обучение. По-голямата част от тези паралелизации подобряват ефективността (running time) на съществуващи вече алгоритми. Тази дисертация изследва друг аспект: вместо да се използват паралелни алгоритми за да се получи решение от същото качество, но по-бързо, ние целим да инвестираме паралелни ресурси по начин, който води до намирането на по-добър модел или решение, като същевременно запазим ефективността на оригиналния алгоритъм.

Тъй като пространството от потенциални решения при типичните проблеми на машинното обучение е огромно, то не може да бъде изследвано изцяло. Мнозинството алгоритми прилагат евристики, като например, грийди (алчно) търсене, с цел да се намери достатъчно добро решение в разумно време. Алгоритмите от тип Грийди и други евристики не гарантират намирането на оптималното решение в общия случай. Целта на изследванията в тази дисертация е да се разработят стратегии, целящи подобряване на резултата получен от стандартна грийди евристика или евристика от друг тип чрез инвестиране на паралелни изчислителни ресурси така, че пространството от потенциални решения да бъде по-добре и по-пълно изследвано. Подходът се базира на различни стратегии за паралелни търсения в пространството от решения и се нарича *Widening* на даден алгоритъм за дейта майнинг или машинно обучение.

Може да разглеждаме много от алгоритмите в дейта майнинг и машинно обучение като грийди търсене в пространството от потенциални решения. Грийди алгоритъмът избира на всяка стъпка локално най-добрия модел, докато се намери достатъчно добро решение, задоволяващо конкретен критерий.

Алчното (грийди) търсене може да бъде представено като итеративно



Фигура 1.1: Алчен търсещ алгоритъм. Вляво (a) е показан временният модел m в зелено. Опциите за подобрения $r(m)$ са показани в сиво. Операторът за селекция s избира жълтото подобрение (b) на следващото ниво търсенето продължава като този избор става временния модел (c).

прилагане на два оператора:

оператор за рафиниране r и оператор за селекция s .

Операторът за рафиниране при зададен модел m прави нови, по-специфични и потенциално по-добри модели, които ще наричаме подобрения (refinements) на модела m .

Операторът за селекция избира локално най-добрия модел от всички възможни подобрения на модела m . Нека \mathcal{M} е фамилия от модели и е дефиниционната област на двата оператора. Операторът за селекция е базиран на функция за оценка ψ , която оценява качеството на модела $m \in model\ family$:

$$\psi : \mathcal{M} \rightarrow \mathbb{R}.$$

Използвайки тези означения, ние можем да представим една итеративна стъпка на алчния алгоритъм по следния начин:

$$m' = s_{\text{best}}(r(m)),$$

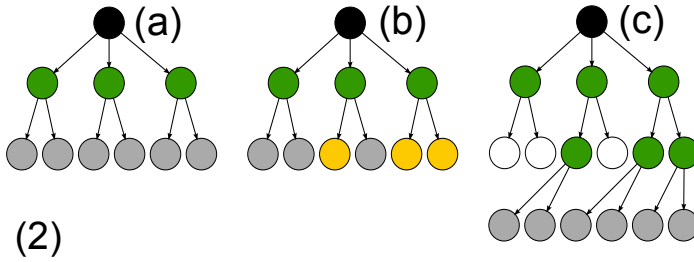
където

$$s_{\text{best}}(M) = \arg \max_{m'' \in M} \{\psi(m'')\},$$

Фигура 1.1 илюстрира алчен търсещ алгоритъм.

1.0.1 Widening на алчна евристика.

За да бъде подобрена работата на алчния алгоритъм, трябва да се вземе предвид неговото "ограничение": на всяка стъпка той избира локално най-доброто решение, което може да не ни отведе до глобалния оптимум. Заради това, вместо да избираме един модел, ние можем да инвестираме



Фигура 1.2: Widening. От множество от модели M (зелени кръгове), операторът за рафиниране създава няколко множества от модели (сиви) a set of models M (green circles) (a). Селекторът избира ново множество от модели The selection now picks a subset of the refined models (жълти кръгове в (б) и търсенето продължава (в)).

паралелни ресурси в това да разглеждаме няколко модела едновременно и да търсим в различни части на пространството едновременно. Това е и идеята на Widening на алгоритми.

Използвайки обозначенията, въведени по-горе, итерация на Widening може да бъде представена по следния начин:

$$M' = \{m'_1, \dots, m'_k\} = s_{\text{widened}} \left(\bigcup_{m \in M} r(m) \right).$$

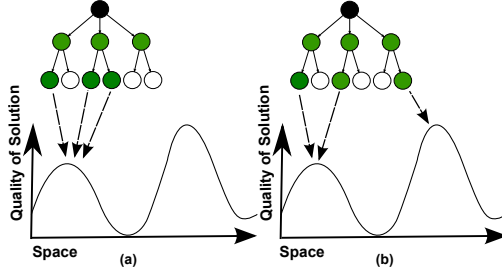
Ще наричаме параметъра k параметър на Widening. Фигура 1.2 демонстрира този процес. Лесна имплементация на идеята за Widening е $Top-k$ търсене в лъч. Вместо да се следва една алчна пътека, се изследва пътеката от k най-добри кандидати.

Всяка итерация на $Top-k$ може да бъде представена по следния начин:

$$\{m'_1, \dots, m'_k\} = s_{Top-k} \left(\bigcup_{i=1, \dots, k} s_{Top-k}(r(m_i)) \right)$$

където s_{Top-k} избира най-добрите k модела, използвайки оценката ψ .

В [1] показваме, че този пост подход вече води до подобряване на качеството на намереното решение. Въпреки това, той има два основни проблема. Единият е, че избраните модели могат да са подобни, тъй като сходните модели често имат сходна оценка и за да се избегне риска от конвергиране в конкретна част от пространството на търсене, се налага да се използва изискване за разнообразие на избраните модели, за да може да се осигури



Фигура 1.3: Widening на даден алчен алгоритъм може да доведе до локална експлоатация на пространството от решения, а не до неговото експлорационно изследване (а). Добавянето на изисквания за разнообразие окуражава по-широкото изследване на пространството от решения (б).

изследването на разнообразни части от пространството от решения. Фигура 1.3 илюстрира ползата и значението от изискването на разнообразие при търсенето на оптималния модел в пространството от решения.

Вторият проблем е свързан с нуждата от комуникация между паралелните агенти при този подход, което значително увеличава оперативното време. Целта на подхода Widening на алчен алгоритъм е да се подобри изследването на пространството от решения и, по този начин, качеството на открития модел, но без да се увеличава оперативното време. Поради това тази дисертация се занимава със стратегии за разнообразно търсене и изследване на пространството, но без нуждата от комуникация.

1.0.2 Разнообразно $Top - k$ Widening.

Прост метод да се добави разнообразие е да се използва прост праг на разнообразие θ , който да се удовлетворява от всички избрани модели, функция за разстояние между два модела, δ и модифициран селектор $s_{Top-k, \delta}$.

Операторът $s_{Top-k, \delta}$ итеративно избира най-добрите k подобрения, които задоволяват прага на разнообразие. Това може да бъде формализирано по следния начин:

- 1: $M_{\text{all}} = \cup_{i=1, \dots, k} r(m_i)$ create set of all possible refinements
- 2: $m_1 = \arg \max_{m \in M_{\text{all}}} \{\psi(m)\}$ pick the locally optimal model as first model
- 3: $M_1 = \{m_1\}$ add as initial model to solution
- 4: for $i = 2, \dots, k$: iteratively pick next, sufficiently diverse model:
- 5: $m_i = \arg \max_{m \in M_{\text{all}}} \{\psi(m) \mid \neg \exists m' \in M_{i-1} : \delta(m, m') < \theta\}$
- 6: $M_i = M_{i-1} \cup \{m_i\}$
- 7: endfor

8: return M_k

1.0.3 Комуникация между паралелните агенти и разнообразие на търсенето в пространството от решения.

Налагането на разнообразие (diversity) при избраните модели допълнително изисква комуникация между паралелните агенти, за да могат те да сравняват множеството от избраните на дадена стъпка модели не само по качество, но и по разнообразие. Това е противоречие с изискването към Widening на даден алгоритъм да има оперативно време максимално близко до това на оригиналния алгоритъм. За да постигнем това, ние целим Widening без комуникация между паралелните агенти. Постигане на разнообразие на изследването без постоянно сравняване на временните модели, избирани от паралелните агенти и без предварително знание за пространството от потенциални решения е трудна задача.

Това изисква механизъм за предварително разпределение на моделите, които да бъдат изследвани от всеки паралелен агент. Подобни стратегии за разнообразно паралелно изследване на пространството от решения е и основната задача в тази дисертация.

1.1 Основен принос на дисертацията и основни резултати

Описваме *идеален Widening* и теоретично го дефинираме като експлицитно разбиване с подобрени свойства на пространството от потенциални решения. Всеки дял от разбиването е се изследва от различен паралелен агент. Идеалния Widening е трудно постижим на практика за пространство на модели от общ вид. Заради това, ние разглеждаме различни методи на Widening, които да не налагат комуникация и са постижими на практика. Това най-често са методи базирани на имплицитно разделяне на пространството от решения, чрез използването на индивидуализирани модификации на оператора за избор на всеки паралелен агент. Друг потенциален проблем на Widening на алгоритъм, който трябва да бъде отчетен, е опасността от клонене към локален оптимум. Това е добре познат проблем при търсенето в лъч и се решава чрез налагане на разнообразие при решенията, избирани от всеки агент при търсенето. Постигането на разнообразие на изследването на пространството от решение без да има комуникация между паралелните агенти и без познание относно структурата на пространството от решения, не е тривиално. Първоначално ние използваме прост подход,

който преди началото на търсенето приписва индивидуални предпочитания относно модели за всеки паралелен агент.

Този подход зависи от параметри, свързани с пространството от решения, и не позволява структурирано изследване на пространството от решения, което е целта на Wideningто на даден алгоритъм. Тъй като нашата цел е структуриран подход, който да не изисква комуникация между агентите, ние изследваме друг подход—Widening на алгоритъм чрез околности, подход, който използва понятието съседство/ околност. Ние дефинираме подход, базиран на околности, който цели структурирано изследване на пространството от решения.

В дисертацията се демонстрират и доказват теоретични свойства на Widening чрез околности под конкретни допускани свойства на пространството от решения. Доказваме, че Widening чрез околности може да емулира *Top - k* Widening, без необходимостта от комуникации за достатъчно голямо k (брой на паралелни агенти). Widening чрез околност от подобни модели може да бъде използвано за изследване на обещаващи области от пространството от решения и за достатъчно голям брой k на паралелните ресурси то може да гарантира, че оптималното решение в конкретен интервал на пространството от решения може да бъде намерено. Може да се гарантира, че за конкретно k намереното най-добро решение от Wideningто на алгоритъма ще бъде най-много δ от най-доброто решение в изследвания интервал. Widening чрез околности от разнообразни модели подпомага да се преодолее потенциално клонене към локален оптимум, принуждавайки всеки паралелен агент да изследва достатъчно различно и обещаващо решение.

Ние изследваме структурата на полето от решения за прост тип алгоритми и доказваме, че той е Булева решетка. Използваме тези свойства, за да получим разбиване и всяко непресичащо се подмножество от разбиването да го разпределим на даден паралелен агент. Демонстрираме ползите на Wideningто и сравняваме всички тези стратегии за Widening, използвайки два алгоритъма, алчния алгоритъм за задачата за покритието на множества (*set cover problem* или *SCP*) и *CN2* алгоритъма за индукция на правила. Провеждаме експериментални изследвания на теоретически обоснованите стратегии и резултатите от тези изследвания подкрепят теоретически доказаните свойства на стратегиите за Widening.

Всички подходи за Widening показват подобрение на качеството на решението спрямо решението, намерено от алчния алгоритъм. Правилното използване на разнообразие води до подобряване на качеството на решението. Разглежданите методи без комуникация са сравними с тези, които използват комуникация, а имат много по-добро оперативно време. Добавянето на паралелни агенти при добре разработена стратегия компенсират липсата на комуникация. Увеличението на размера на околността не во-

ди до подобрене на резултата, ако няма и съответно увеличение на паралелните агенти. Използването на разнообразие при построяването на околностите от модели така, че околностите да съдържат обещаващи и разнообразни модели, води до подобрене на качеството на крайния резултат. Различните подходи за постигане на разнообразие постигат различни нива на разнообразие в крайния резултат. Взаимосиност от типа данни и структурата на пространството от решения, различни подходи за разнообразие са по-удачни от други. Wideningто с подобни околности е успешно в откриването на решения с подобни свойства и добро качество. Използването на околности от подобни в комбинация с околности от разнообразни модели помага за прилагане на двата основни принципа на търсенето – експлорация и експлоатация (exploration and exploitation). Използването на малки околности от подобни модели помага изследването в детайл на подобрени региони пространството от решения и се представя по-добре от големи околности от подобни модели. Размерът на околността доминира оперативното време при подходите за Widening, които използват околности. Това важи особено при Wideningто с разнообразни модели, където размерът на околността определя количеството сравнения и изчисления необходими да се построи околността. При Widening с околности от подобни, размерът на околността не е така определящ. За експлоатация чрез околности от подобни модели по-добре работи Widening с околност от подобни с малък размер и голям брой паралелни агенти. Прагът на разнообразие δ не оказва силно влияние върху оперативното време. Поради това е по-печелившо да се използва различен праг за разнообразие вместо непременно по-голяма околност. Липсата на комуникация може донякъде да бъде компенсирана чрез повече работа от паралелните агенти. За много големи околности подходът Widening чрез разнообразни околности може да има по-лошо оперативное време от *Top – k* Widening поради нуждата от това всеки паралелен агент да построи голяма част от пространството от решения. Но експерименталните резултати демонстрират, че дори за малък размер околности за конкретните данни, за които сме тествали методите, с малка околност се постигат добри резултати и голяма околност не е необходима. Въпреки, че този резултат е зависим от използваните данни, вместо увеличаване на околността може да се използва добре подбран достатъчно голям праг на разнообразие.

1.1.1 Публикации, свързани с дисертацията

Резултатите от дисертацията са публикувани в следните публикации. Widening е първо дефинирано в [1] като парадигма за интелигентно инвестиране на паралелни ресурси с цел да се подобри качеството на намереното решение. Ние мотивираме нуждата от Widening без комуникация и нуждата от разнообразие за подобреното изследване на пространството от решения

в [2]. В публикациите [3] и [4] се демонстрира подхода на разширение чрез околности за два алгоритъма – алчния алгоритъм за минимално покритие на множества и алгоритъма CN2 за Rule Induction.

Глава 2

Предходни научни изследвания в областта (Related Work).

В тази глава са описани различни научни публикации, свързани с темата на дисертацията. Публикациите в областта основно целят паралелизацията на алгоритми с цел подобряване на оперативното време на тези алгоритми. Дискутират се и паралелни или лесно-паралелизуеми алгоритми, които целят подобряването на намерения модел. Подчертават се разликите между вече съществуващите научни изследвания и приноса на тази дисертация.

Глава 3

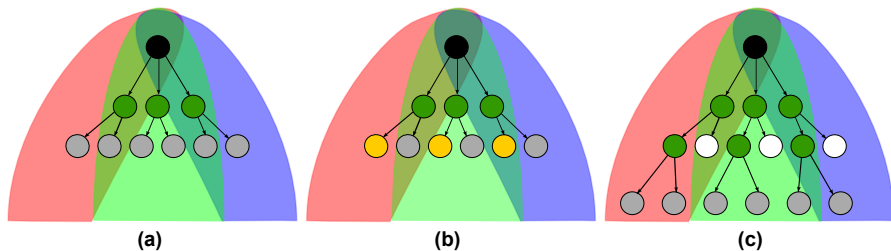
Идеален Widening

В тази глава се дискутират свойства, които идеален подход за Widening би трябвало да притежава и защо в практиката е трудно/невъзможно един подход за Widening да ги изпълни. Идеалният Widening би било разбиване на непресичащи се, затворени множества, всяко, от които, да бъде възложено за изследване на даден паралелен агент. Това е много трудно в общия случай, а понякога и невъзможно. В практиката, по-реалистичен сценарий е този, в който пространството от решения е разделено на подмножества, които частично се пресичат, но всеки модел е достъпен в поне едно множество. Това в много случаи вече е достатъчно. Кардиналностите на сеченията на множествата има директно влияние върху количеството излишно изследване на пространството, когато се използва този подход за приблизително разбиване на пространството при Widening на даден алгоритъм. Фигура 3.1 илюстрира този подход. Целта е пресичането да е в началото на търсенето, а в края всеки агент да достига до различно решение.

Най-различни критерии могат да бъдат приложени за да се оцени качеството и разнообразието на избраното множество от модели, в зависимост от целите. При дадена функция за оценка на качеството на модела, целта за множество от модели може да е да се максимизира най-високата за множеството оценка, средната оценка, сумата им или други свойства на множеството от модели. За разнообразието на множество от модели, човек може да разглежда различни метрики и да оцени колко разнородно е множеството от модели, например средната подобност по двойки, сумата на всички подобности по двойки, минимумът или максимумът, в зависимост от целите на оценката.

Най-общо:

Definition Нека $M \subset \mathcal{M}$ е подмножество на пространството от потенциални решения (тоест, множество от модели), и нека Ψ and Δ са функции,



Фигура 3.1: Widening базиран на приблизително разбиване на пространството от решения. Пространството от потенциални решения е разделено на пресичащи се подмножества, но може да се гарантира, че всеки модел е достъпен за даден агент в поне едно подмножество. Вместо да избира измежду временни решения, всеки агент избира най-доброто временно решение именно в своето подмножество от пространството от потенциални решения.

оценяващи качеството и съответно разнообразието на множество от модели k е параметърът на Widening, тогава функцията $s_{\text{topdiv}-k}$ се нарича **най-добър-разнороден- k оператор за избор** ако $\forall M' \subseteq M \wedge |M'| = k : \Psi(s_{\text{Topdiv}-k}(M)) \geq \Psi(M') \vee \Delta(s_{\text{Topdiv}-k}(M)) \geq \Delta(M')$.

Тоест, не съществува друго множество от модели с размер k , което да е с по-висока оценка за качество и същевременно да е по-разнообразно. За многокритерийните оптимизационни проблеми целият Парето фронт съдържа недоминирани решения.

Глава 4

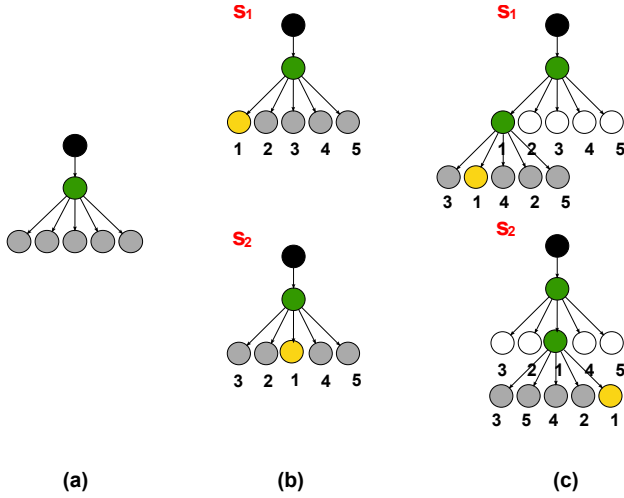
Widening чрез хаширане без нуждата от комуникация

Този раздел се базира на [2]. Ако не желаем да разчитаме на глобална комуникация между паралелните работници при избора на висококачествени и разнообразни модели, ние трябва да подсигурием всеки паралелен работник да има индивидуална стратегия за избор на (висококачествени и разнообразни) модели, за да може изследваното множество модели да е разнообразно. Ние можем да го постигнем чрез модификация на индивидуалните оператори за селекция преди началото на процедурата за търсене чрез което да постигнем промяна в поведението на всеки един паралелен агент.

Идеалното и приблизително идеалното разбиване са трудни за реализиране в общия случай без познаване на структурата на пространството от решения априори. Поради това ние се целим да постигнем имплицитно разделяне на пространството от потенциални решения.

4.0.1 Разнообразие без комуникация чрез индивидуални предпочитания (хаширане)

Ние искаме да постигнем имплицитно Widening с параметър на Wideningto k без комуникация между паралелните агенти, чрез индивидуализиране на оператора за селекция $s_i, i \in \{1, \dots, k\}$ на всеки агент. Това означава, че всеки оператор s_i предпочита различно множество от модели. Пространството от решения е неизвестно в началото на търсенето, поради това тези индивидуални предпочитания трябва да бъдат дефинирани имплицитно и преди търсенето. Поради това не е възможно да бъдат зададени преференции за цели модели в началото на търсенето, но могат да се ползват



Фигура 4.1: Widening без комуникация: две различни пътеки на селекция, генерирани от два различни оператора за селекция s_1 и s_2 .

фрагменти от модела или самите данни, тъй като те са известни в началото на търсенето. Най-лесният подход е да се назначат индивидуалните предпочитания директно по статичен начин.

Ще дефинираме този прост подход за Widening без комуникация както следва.

Нека за даден временен модел m ψ_i е оценъчната функция, която оценява рафинирания модел m' въз основа на оригиналната оценъчна функция (мярка ?) и на индивидуални за агента предпочитания π_i , където i е паралелният агент, $i = 1, \dots, k$. Тогава :

$$\psi_i(m') = \psi(m') + t * \pi_i(m').$$

Нашата цел е да имаме k различни и разнообразни редове на предпочитание preference orders π_1, \dots, π_k , така, че индивидуалните оператори за селекция са принудени да изследват различни и разнообразни пътеки в пространството от решения.

Definition При дадено множество от паралелни агенти $\{w_1, \dots, w_k\}$, Widening без комуникация с **разнообразие базирано на глобални предпочитания** е множеството от оператори за селекция $s_{hash} = \{s_1^{hash}, \dots, s_k^{hash}\}$, чиито индивидуализирани функции за качествена оценка $\{\psi_1, \dots, \psi_k\}$ възлагат на модели от \mathcal{M} различни предпочитания, зададени от функциите

$p_i : \mathcal{M} \rightarrow \mathbb{R}$ да имплементират множество от разнообразни частични наредби над \mathcal{M} $\{(\leq_i, \mathcal{M})\}$.

Балансът между разнородност и качество на изследвани модели могат да бъдат оптимизирани ако индивидуализираните функции за оценка $\{\psi_1, \dots, \psi_k\}$ комбинират резултатът, получен от оригиналната функция за оценка ψ и индивидуализираните предпочитания за моделите. Този прост подход води до изненадващо високо подобряване на качеството на крайния модел.

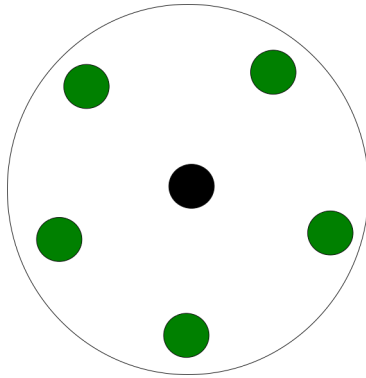
Глава 5

Widening на алчен алгоритъм чрез околности

Тук непременно да сложа дефиниция на Widening чрез околности. Wideningto на алчен алгоритъм чрез околности дефинира околности на локално оптимален модел от дадено множество на рафинирани модели с различни свойства и разделя тази околност между паралелните агенти. Три различни вида околности са изследвани: околности, базирани на оптималност, за които метриката е изчислена чрез model quality measure; околности, базирани на подобност, за които метриката се базира на предефинирано разстояние подобност; and diversity околности, които съдържат разнообразни и добре представящи се k модела. Различни подходи за имплементиране на diverse neighborhoods са разгледани – такива, базирани на прост праг от разнообразие, други използват сложни софистицирани стратегии за разнообразие, като примерно ниширане и, взаимствано от генетичните алгоритми. Други изследвани подходи за постигане на разнообразие включват

Definition При дадено метрично пространство (X, d) и естествено число $k > 0$, we дефинираме k -околност N_k на фрагмент x , като нареденото множество от k моделни фрагмента $N_k = (x, x_1, \dots, x_{k-1})$, за които $d(x, x_i) \leq d(x, x_{i+1})$, $i \in \{1, \dots, k-1\} \wedge d(x, x_j) \geq d(x, x_i) \forall x_i \in N_k, x_j \in X \setminus N_k$.

С други думи, k -околността на моделен фрагмент x е множеството, което съдържа x и $k-1$ моделови фрагмента, които са най-близко до x според метриката d .



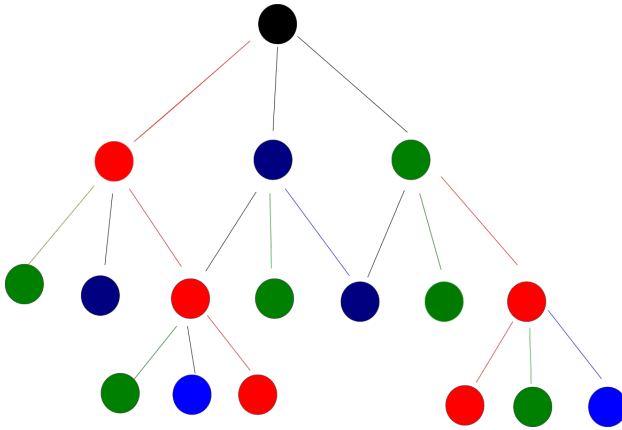
Фигура 5.1: Околност на метричното пространство от моделни фрагменти k -околност за $k = 6$.

5.1 Околности от оптимални модели

Wideningто чрез околности от оптимални модели е опит да се емулира *Top – k Widening* без да се налага комуникация между паралелните агенти. Вместо да се избира само най-добрия модел на всяка стъпка, както е при алчния алгоритъм, се строи околност от k най-добри модела, които се разпределят между паралелните агенти.

5.2 Околности от подобни модели

Околностите от подобни модели са k -околности, при които метриката d оценява подобност е метрика за подобност въз основа на подбрани важни характеристики на модела. Widening чрез околности от подобни модели изследва моделите, подобни на локално оптималния модел на всяка стъпка, тоест модели подобни на алчния избор. Wideningто чрез оптимални модели има различни приложения като се използва за експлоатация на част от пространството от потенциални решения, за която е известно априори, че съдържа добри решения. Това представлява интензивно търсене в малка част от пространството *exploitation*. Друго приложение на тези околности е *similarity search*, при който да се търсят подобни модели с конкретни характеристики, от които да се изберат най-добрите според оценката за качество. Подобни търсения на подобни намират широко приложение във фармацевтиката при лекарствен дизайн, например.



Фигура 5.2: Схематично представяне на Widening чрез локално разбиване на $M^r = r(m)$ използвайки k -околности. Различни съседи са възложени на различни паралелни агенти. Разбиването е представено с различно оцветяване на моделите и се проследява пътеката на всеки от паралелните агенти. Тук $k = 4$.

5.3 Разнообразие в околностите от модели

Widening чрез околности от оптимални модели не гарантира разнообразие на изследваните в пространството модели, тъй като сходни модели често имат сходна оценка на качество и съседи по оптималност (в околност от оптимални модели) често имат сходни характеристики. Чрез построяване на околности от *разнообразни* модели, които се състоят от качествени, но разнообразни решения (модели), паралелните агенти са принудени да изследват по-голяма част от пространството от потенциални решения. Целта на Wideningто чрез околности от разнообразни модели е чрез локално разнообразие на избраните модели на всяка стъпка, да се постигне глобално разнообразие на изследването на пространството от решения без да се налага комуникация между паралелните агенти.

Построяване на околности от разнообразни модели чрез методи за разнообразие от генетичните и еволюционни алгоритми. Разнообразието трябва да бъде експлицитно наложено, защото увеличаване на размера на околността не води непременно до разнообразие на изследваните модели. Например, при силно неравномерно разпределение на моделите в пространството от решения, увеличаване на околността няма непременно да доведе до изследване на разнообразни модели.

В областта на генетичните алгоритми клонене към локален оптимум

прекалено рано в търсенето е голям проблем, обикновено причинен от недостатъчно генетично разнообразие на поколенията. Този проблем е много добре изучен като за целта е създаден е клас от подходи, наречен “ниширане” (*niching*), който да поддържа разнообразието на поколенията при всяка стъпка. Този проблем, за клонене към локален минимум, е много близък до този, с който се сблъсква Wideningто чрез околности на даден алчен алгоритъм. Поради това тези подходи могат да бъдат използвани при Wideningто с околности от разнообразни модели за подбора на разнообразни модели при всяка стъпка. Използваме Fitness sharing и crowding, които са класически подходи в генетичните алгоритми. За подобряване на оперативното време използваме не само стандартните версии на тези подходи, но и по-бързи “приблизителни” версии на тези подходи deterministic crowding.

В нашия контекст на Widening на алчен алгоритъм чрез разнообразна околност ние разглеждаме временните модели при дадена стъпка на рафиниране като нови индивиди в популацията. Прилагаме методите за ниширане към множеството от рафинирани модели на всяка стъпка с цел да създадем околност от k разнообразни модела. $M = \{m_1, \dots, m_n\}$, а model quality measure ψ and a metric d , модифицирана оценка на качеството, базирана на ψ , която, обаче, включва поделената годност на моделите от една ниша, ψ_{fsh} е дефинирана като

$$\psi_{fsh}(m_i) = \frac{\psi}{\sum_{j=1}^n sh(d(m_i, m_j))},$$

$$sh(d) = \begin{cases} 1 - (\frac{d}{\sigma})^\alpha & \text{if } d < \sigma \\ 0 & \text{иначе.} \end{cases}$$

Където σ е прагът на разстоянието между два модела, под който те са в една и съща ниша, а параметърът α до каква степен стойността на оригиналната оценка за качество на модела (преди преизчисляването на годността на модела) контролира резултатното качество след прилагането на поделената годност the niche count of a given model. Този подход имплицитно налага разнообразие в околността от модели.

При Crowding методите, разнообразието се поддържа чрез заместването на прекалено подобни модели с нов модел като съществуват различни подходи . Ние използваме стандартен и детерминистичен Crowding.

5.3.1 Разнообразие чрез прост праг. Разнообразие чрез околности от подобни.

При multi-objective оптимизация, рядко има решение, което минимизира всички функции едновременно. Заради това се търсят Парето оптимал-

ни решения, които не могат да бъдат подобрявани за нито една objective без същевременно да се получи по-ниска стойност за поне една друга objective.

Definition За multi-objective оптимизационен проблем $\max(f_1(M), f_2(M))$, за решението M' се казва, че решението доминира M'' ако $f_i(M') > f_i(M'')$ и $f_j(M') \geq f_j(M'')$, където $(i, j) = (1, 2)$ или $(i, j) = (2, 1)$.

В нашия контекст, multi-objective оптимизация се занимава с разнообразие и оптималност на множеството от решения $\max(\Delta, \Psi)$.

За да построим околност от оптимални и разнородни модели за дадено множество от рафинирани модели, може да бъде използван прост праг за разнородност δ , базиран на метрика за подобност. Тогава, решенията, които ще бъдат избрани на всяка стъпка са множеството с най-висок Ψ от всички възможни множества от модели, които покриват изискването δ . Това множество от решения ще бъде недоминирано решение от Парето фронта за проблема за даденото множество от рафинирани модели. Вместо използването на фиксиран праг, може да се ползва гъвкав праг на разнообразие, който се адаптира на всяка стъпка, съобразно структурата на пространството от потенциални решения. Също така, при формирането на околност от разнообразни модели могат да бъдат използвани и околности от подобни модели, като се избират модели, които не принадлежат на едни и същи околности от подобни модели.

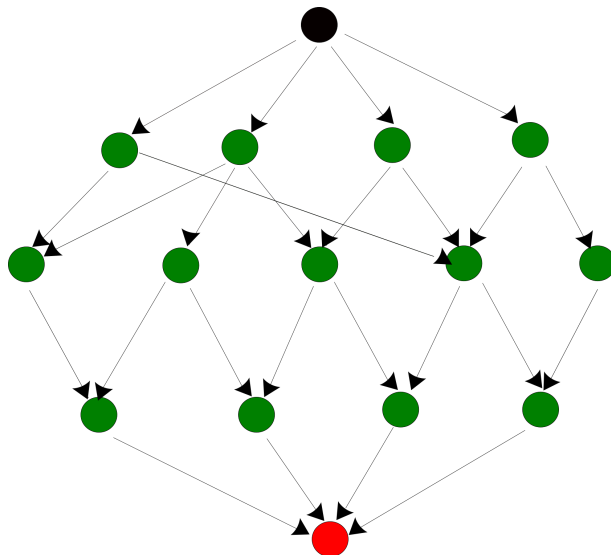
Глава 6

Локални подходи за Widening: теоретични свойства на Widening чрез околности за оператор за рафиниране от тип 1

В тази глава дефинираме вид оператор за рафиниране от тип 1, и за него доказваме конкретни теоретични свойства на Widening чрез околности. Основен интерес представляват следните въпроси: Колко паралелни агенти трябва да се гарантира, че Widening без комуникация с и без разнообразие с различни околности ще може да бъде сравнявано с *Top-k* Widening и *diverse Top-k* Widening, където се използва комуникация. Също така се интересуваме колко паралелни агенти са нужни при Widening с околности със сходство, за да гарантираме, че ще намерим оптималното решение. За тази цел доказваме конкретни граници на методите.

In this chapter we will refer to a direct refinement simply as a refinement. Let \mathcal{M} be a family of models, X be the set of model fragments in \mathcal{M} , r be a refinement operator over \mathcal{M} . We can use \mathcal{M} and the refinement operation $r(\cdot)$ to define a graph $G_{\mathcal{M}}(V, E)$, where V is the set of vertices, and E is the set of edges, defined as follows: $v \in V \Leftrightarrow v \in \mathcal{M}$ and $\forall m, m' \in \mathcal{M}, m' \in r(m) \exists e(m, m') \in E$.

Графът от рафинирани модели е показан във Фигура 6.1. Фигура ?? показва Widening чрез околности в графа $G_{\mathcal{M}}$. Свойствата на класа методи Widening чрез околности ще зависят от структурата на пространството от



Фигура 6.1: A general refinement graph $G_{\mathcal{M}}$ for \mathcal{M} .

решения.

Видно е, че $G_{\mathcal{M}}$ е DAG.

6.1 Графът на пространството от потенциални решения е решетка за семейство модели с оператор за рафиниране от тип 1.

Различни видове оператор за рафиниране могат да се дефинират за различните алгоритми, в зависимост от тяхната комплексност. Типът оператор дефинира структурата на пространството от решения.

Definition Нека \mathcal{M} е семейство от модели, X е множеството от моделни фрагменти в \mathcal{M} , r е операторът за рафиниране върху \mathcal{M} със следните свойства: само един моделов фрагмент се добавя при една операция за рафиниране и редът, в който се добавят фрагментите няма значение. По-конкретно, множеството от моделови фрагменти $\{x_1, \dots, x_l\}$ по единствен начин дефинира модел m и $\forall m' \in r(m) : m' \setminus m = x', x' \in X$. Ние ще наричаме подобен вид оператор за рафиниране *r оператор за рафиниране от тип 1*.

Lemma 6.1.1. *Нека \mathcal{M} е семейство от модели с оператор за рафиниране r от тип 1. Тогава, \mathcal{M}, \leq дефинира решетка, където \leq е частична наредба дефинирана от оператора r над \mathcal{M} .*

Lemma 6.1.2. *Решетката от фамилия от модели \mathcal{M} , $L_{\mathcal{M}}$ е дистрибутивна решетка.*

За оператор за рафиниране от тип 1, решетката е известна като power set lattice, за които се знае, че са булеви решетки.

6.2 Недостатъци на Widening чрез околности

Недостатъците на Widening чрез околности произлизат от това, че методът е локален и без комуникация. Липсата на комуникация между паралелните агенти ги лишава от получаване на информация от другите агенти. Локалните методи определят поведението на даден паралелен агент за всяко множество от рафинирани модели $M^r = r(m)$ индивидуално. Не се знае каква е релацията на моделите от различни множества от рафинирани модели, тъй като всеки модел може да е член на повече от една околност/множество от рафинирани модели. А по-точно, паралелните агенти могат да достигнат едно и също решение по различни пътеки. Целта е да се използва локално поведение на паралелния агент, зададено за всяко конкретно refinement set M_i^r , за да се постигне желано глобално поведение на паралелните агенти в пространството от решения.

6.3 Свойства на N_k^o

N_k^o Widening цели да емулира, но без да използва комуникация, подходът $Top - k$ Widening. За сравнение, $Top - k$ Widening е метод, който използва много комуникация между паралелните агенти. Важно е да се оцени експериментално и теоретично как тези два метода се представят един спрямо друг, като качество на намереното решение. За $k = 1$, и двата метода изследват пътеката на алчния алгоритъм и ще получат един и същи резултат. Ние ги сравняваме за $k > 1$.

Lemma 6.3.1. *Нека $m_i, m_j \in \mathcal{M}$ са два различни модела, тогава околността на тези модели може да има най-много един общ модел :*

$$|N_k^o(m_i) \cap N_k^o(m_j)| \leq 1.$$

Всъщност, те се пресичат тогава и само тогава, когато двата модела принадлежат на едно и също множество от рафинирания $m_i, m_j \in r(m)$.

Нека да погледнем изкуствено конструирания подход за Widening $FullTop-k$.

Definition Дадена функция за оценка на модела $\psi : \mathcal{M} \rightarrow \mathbb{R}$, и модели m_1, \dots, m_k функцията $s_{FullTop-k}$ се дефинира по следния начин:

$$s_{FullTop-k}(r(m_1, \dots, m_k)) := \bigcup_{i=1}^k s_{Top-k}(r(m_i))$$

Ние използваме $FullTop-k$ за да ограничим подпространствата на пространството от решения, изследвано от методите $Top-k$ and N_k^o и да ги сравним.

Lemma 6.3.2. 1. $Top-k(\mathcal{M}) \in FullTop-k(\mathcal{M})$.
2. $N_k^o(\mathcal{M}) \in FullTop-k(\mathcal{M})$.

The relationship between $Top-k$, N_k^o , and $FullTop-k$ is visualized in Figure 6.3.

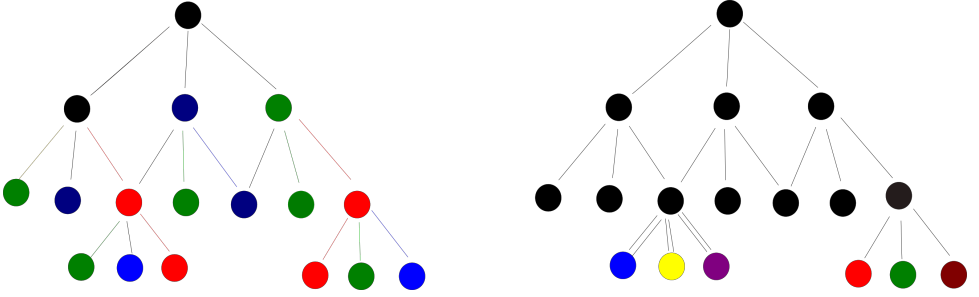
Definition Given a family of models \mathcal{M} with a refinement operator r of type 1, we define *randomized k -neighborhood Widening*, Nr_k^o , as optimality k -neighborhood Widening, where each member of a given k -neighborhood is selected by a parallel worker with equal probability $\frac{1}{k}$.

Lemma 6.3.3. При дадено $FullTop-k^l(\mathcal{M})$ с G_{FT-k} , за който вероятностното разпределение на всеки модел на ниво l , P^l е униформно, функцията за качество ψ , разстояние d . Ние налагаме намаляваща наредба, базирана на ψ , в $\{FullTop-k^l(\mathcal{M})\}$, $(\{FullTop-k^l(\mathcal{M})\})_{ord}$, така, че $\psi(m_j) \geq \psi(m_{j+1}), \forall j = 1, \dots, \max - 1, \max = |\{FullTop-k^l(\mathcal{M})\}|$. Това индуцира намаляваща наредба върху подмножеството на $(\{FullTop-k^l(\mathcal{M})\})_{ord}$, $(\{(Nr_k^o)^l(\mathcal{M})\})$. Тогава за всеки три последователни модела в нареденото множество $(\{(Nr_k^o)^l(\mathcal{M})\})$, $m_i, m_{i+1}, m_{i+2} \in (\{(Nr_k^o)^l(\mathcal{M})\}), i = 1, \dots, k-2$, on average, $d(m_i, m_{i+1}) = d(m_{i+1}, m_{i+2})$.

6.3.1 Графът, G_{FT-k} , генериран от $FullTop-k$.

Да разгледаме графът, който се състои от подпространството на модели, изследвани от $FullTop-k$ до рафинираща стъпка l .

Definition Нека G_{FT-k} е графът, генериран от $FullTop-k$, който изследва пространството от модели. Тогава множеството от върхове V се състои от множеството от модели, изследвани от $FullTop-k$ до ниво на рафиниране l . Множеството от ребра E представя връзката между два модела, чрез оператор за рафиниране, $e = e(m_i, m_j) \in E \iff m_j \in r(m_i)$.



Фигура 6.2: Widening чрез рандомизирани околности като случайни блуждания по графа от модели.

The graph G_{FT-k} is a subgraph of the search space graph $G_{\mathcal{M}}$.

Lemma 6.3.4. Nr_k^o Widening е еквивалентно на k независими случайни walk във G_{FT-k} .

Lemma 6.3.5. Нека X , където $|X| = n$, е множеството от фрагменти на модела за оператор за рафиниране от тип 1 r and let \mathcal{M} е фамилията от модели, дефинирани чрез r, X . Тогава графът G_{FT-k} има най-много $\min(k^l, \binom{n}{l})$ върха на ниво l .

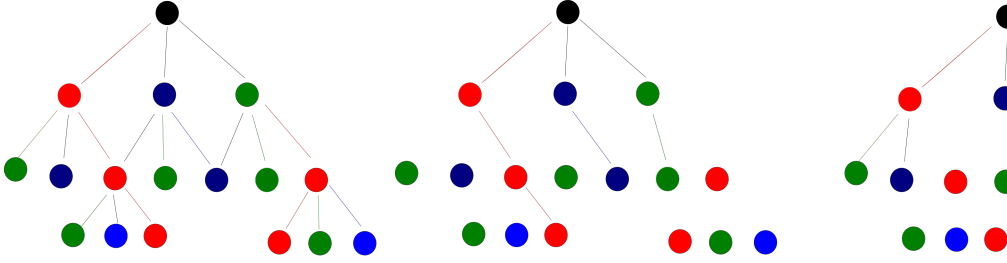
Нека P^l е вероятностно разпределение за вероятността всеки връх от ниво l да бъде открит от случайно блуждаене. На всяко ниво l , the probability p_i^l for reaching a node m_i^l depends on the number of in-degrees to m_i^l as well as the probability distribution P^{l-1} . Let T be the transition matrix associated with G_{FT-k} .

$$P^l = P^{l-1}T.$$

6.3.2 Униформено разпределение в G_{FT-k} .

Разпределението на ребра в G_{FT-k} определя вероятността да бъде достигнат всеки връх на графа/ всеки модел/.

Lemma 6.3.6. Ако P^l е униформено разпределение, решенията, намерени от Nr_k^o , $\{Nr_k^o(\mathcal{M})\} \in \{FullTop - k(\mathcal{M})\}$ са средно униформено разпределени сред решенията $\{FullTop - k(\mathcal{M})\}$. Тогава $\max_{\psi}(\{(Nr_k^o)^l(\mathcal{M})\})$ ще бъде средно най-много $\frac{k^{l-1}}{2}$ модела от $\max_{\psi}(\{FullTop - k^l(\mathcal{M})\})$.



Фигура 6.3: Изкуствената $FullTop-k$, която съдържа и ограничава и двата метода N_k^o and $Top-k$

6.3.3 Widening Graph with Strongly Non-Uniform Distribution

В най-лошия случай, $\max_{\psi}\{Top-\theta\} = \max_{\psi}\{FullTop-\theta\}$. Колко голямо трябва да е k , за да гарантира $\Psi(\{Top-\theta(\mathcal{M})^l\}) = \Psi(\{Nr_{\theta,k}^o(\mathcal{M})^l\})$?

Lemma 6.3.7. *Ние допусваме равномерно разпределение на ребрата на $G_{FullTop-\theta}$. За $k = \min(\theta^l, \binom{n}{l})$, $Nr_{\theta,k}^o$ изследва напълно моделите изследвани от $FullTop-\theta$ на стъпка l и гарантира $\Psi(Nr_{\theta,k}^o(\mathcal{M}))^l \geq \Psi(\{Top-\theta(\mathcal{M})^l\})$.*

6.3.4 $N_{\theta,k}^o$ със силно неуниформно разпределение

Силно неуниформно разпределение е много не non-uniform distribution is very неизгодно за N_k^o Widening метода в сравнение с $Top-k$ подхода.

Lemma 6.3.8. *Да допуснем, че $P^l(x)$ представя вероятността за всеки модел на ниво l да бъде достигнат от рендъм уок по $G_{FullTop-\theta}$. Тогава, за $k = \frac{1}{\min P^l(x)}$ на ниво l всеки модел ще бъде достигнат.*

6.3.5 Свойства на of θ, k -околности.

Ние намираме колко точно паралелни агента са нужни за $\max_{score}\{(N_{\theta,k}^o)^l(\mathcal{M})\}$ да бъде на фиксирана дистанция от най-добрия модел намерен от $Top-k$ метод $\max_{score}\{Top-\theta^l(\mathcal{M})\}$. За да намират и двата метода модели от едно и също качество, липсата на комуникация може да се компенсира чрез наличието на повече паралелни агенти.

Lemma 6.3.9. *При дадено униформно разпределение P^l , за $k = \theta^l/p$, най-доброто решение открито от $N_{\theta,k}^o$ е средно p модела разстояние от най-доброто решение от $FullTop-\theta$.*

Глава 7

Глобални Подходи: Разделяне на решетката на пространството от решения за оператори за рафиниране от тип 1 и Widening чрез глобално разнообразие.

В тази глава се дискутират методи за имплементиране на глобално разнообразие при изследването на пространството от решения. Изследва се структурата на решетката, образувана от пространството от решения и оператора на рафиниране от тип 1; ползите и недостатъците на глобалното разнообразие, както и как да се използва методът "Symmetric Chain Decomposition" за Widening.

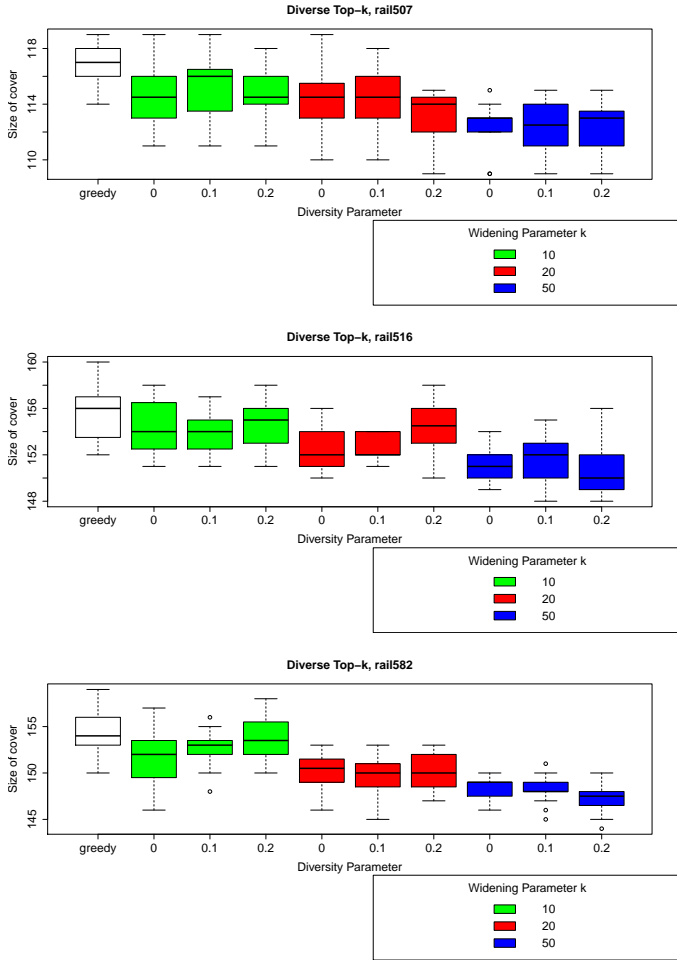
Глава 8

Widening на алчния алгоритъм за минимално покритие на множества

Тази глава дава конкретен пример за Widening подходите, като ги прилага към алчния алгоритъм за намиране на минималното покритие на множества (Set Cover problem). Подходите, представени в предходните глави, са приложени тук, имплементирани и експериментално оценени. Главата съдържа формализация на проблема за намиране на минимално покритие на множествата и представяне на алчния алгоритъм чрез дефинирания в първа глава формализъм за селекция и рафиниране. Представени са методи за Widening с и без комуникация чрез псевдокод.

8.0.1 *Top - k* и *Top - k* с разнообразие

С увеличението на параметъра k , качеството на намереното финално решение се повишава. Оптималните стойности на параметъра за разнообразие зависят от данните и k , ако стойността е подбрана правилно, води до повишаване на качеството на намереното решение. Високи стойности на параметъра за разнообразие в съчетание с малък брой паралелни работници води до влошен резултат поради рандомизиране на търсенето – търсенето не разглежда качеството на решения. Разнообразието, комбинирано с достатъчно голям брой паралелни работници води до подобряване на качеството на решения.



Фигура 8.1: Резултати от $Top - k$ Widening с и без разнообразие.

8.0.2 Резултати от Widening без комуникация, базирано на хаширане.

Променияйки параметъра t , можем да контролираме до каква степен всеки работник приоритизира качеството на решението спрямо разнообразието и доколко могат да се отклоняват от траекторията на работник, следващ чисто алчната стратегия. За стойности на параметъра $t > 1$, пътеките за селекция на паралелните работници включват и субоптимални

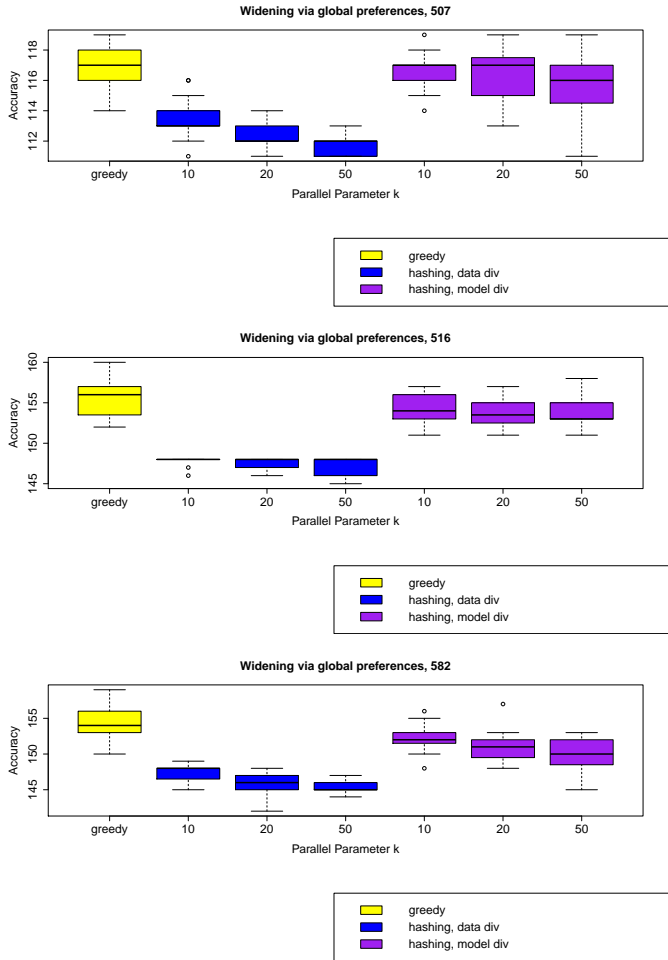
временни решения, стига те да допринасят за разнообразието на изследвани решения. Високи стойности на параметъра t ($\gg 1$) ще доведат до рандомизирано изследване на пространството от решения, в което качеството не се взема предвид при избиране на временните решения. Това ще доведе до загуба на интелекта на алчния алгоритъм. Фигура 8.2 показва резултатите от Widening без комуникация, използващо разнообразие, базирано на данни и разнообразие, основаващо се на модели. Фигура ?? сравнява Widening без комуникация с $Top-k$ Widening с разнообразие. Резултатите водят до две основни заключения: първо, очаквано, по-голяма широчина на търсенето (т.е. повече паралелни работници) подобрява качеството на финалното решение. Второ, налагането на разнообразие при избор на временните решения, изследвани от паралелните работници, подобрява допълнително качеството на откритото решение. Оптималната стойност на параметъра t зависи силно от данните и прецизното подбиране на стойност на този параметър води до значително подобрение в качеството на решение. За добре подбрано t и достатъчно паралелни работници, Widening без комуникация може да бъде сравнимо с $Top-k$ Widening с изискване за разнообразие, но без изключително високата цена на комуникацията.

8.0.3 Widening чрез оптимални θ , k -околности.

Вижда се, че при фиксиран размер на околността, колкото с повече паралелни агенти разполага методът, толкова по-добър е резултатът от разширеното търсене. При фиксиран брой паралелни агенти, увеличаването на размера на околността безконтролно ще доведе до случайно търсене в пространството, при което се губи интелигентността на метода. Прекалено малка околност пък води до изследване само на малка част от пространството на потенциални решения и до избор на подобни решения при всяка стъпка от алгоритъма. Оптималният размер на околността ще е различен за всяко множество от данни и ще зависи от пространството от решения. За добро и смислено изследване на пространството от решения по-добре е да се разчита експлицитно на изискване за разнородност при търсене, вместо на голям размер на околностите.

8.0.4 Widening чрез разнообразни околности

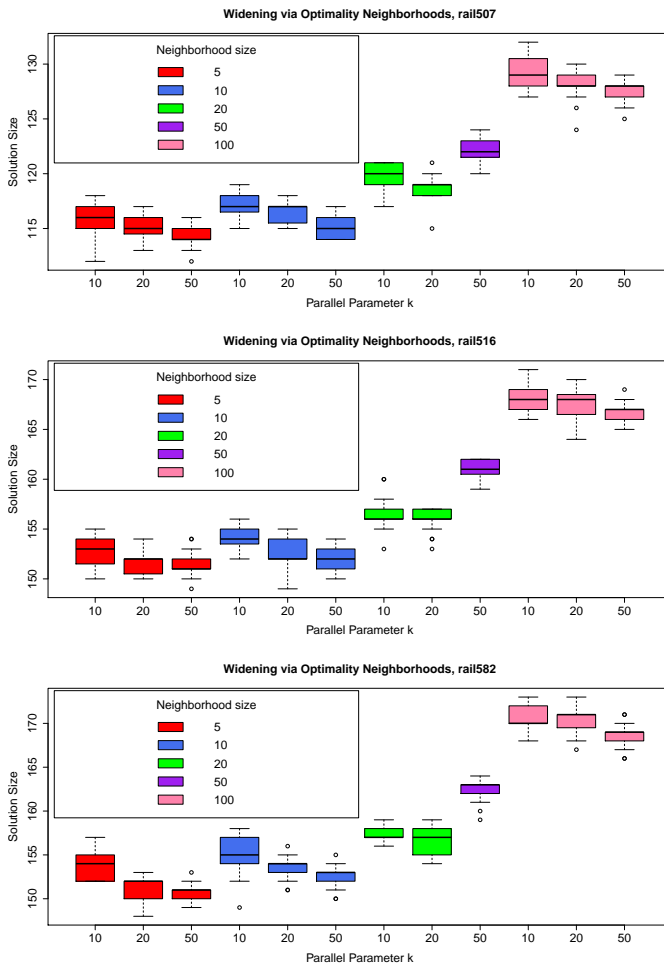
Целта на Wideningто не е рандомизирана експлорация на пространството от потенциални решения, но изследване на върховете от обещаващи и разнородни решения в това пространство. Заради това се налага експлицитно да наложим разнородност при избора на текущи модели при търсенето, вместо просто да използваме голям размер на околността. Резултатите, представени във фигура 8.5 показват решенията намерени от



Фигура 8.2: Резултати от Widening без комуникация, базирано на хаширане.

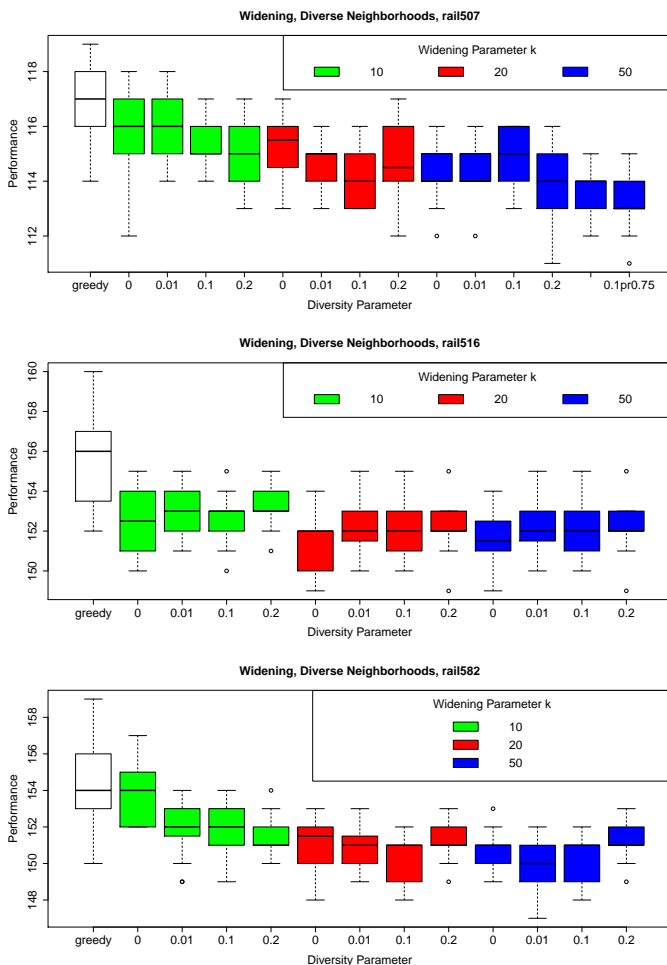
Wideningто чрез околности от разнородни модели за различни стойности на параметъра на разнородност δ .

Резултатите за Widening чрез околности от разнородни модели, където разнородността е получена чрез fitness sharing при различни стойности на параметрите за fitness sharing са показани на фигура 8.6. Очевидно подобрения са постигнати чрез използването на fitness sharing за разнородност.



Фигура 8.3: Widening чрез околности от оптимални модели, за различни стойности на параметрите за Widening θ, k (размер на околностите и брой на паралелните агенти).

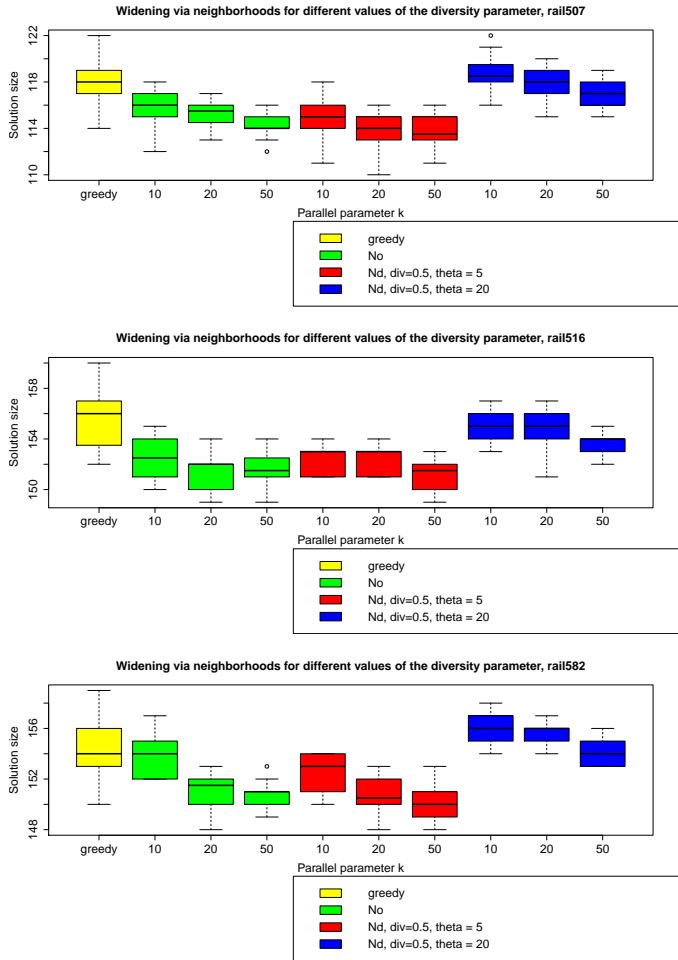
Трудно е да се изберат оптимални стойности на параметрите на поделяне на фитнеса α, σ за конкретни множества от данни. Тези стойности ще са различни за всяко конкретно множество от данни.



Фигура 8.4: Widening чрез разнородни околности за различни стойности на параметъра на разнородност δ , параметъра на Разширяване k и фиксиран размер на околността $\theta = 5$. За оценяване на разнородността на моделите е използвана мярката за дистанция на Jaccard..

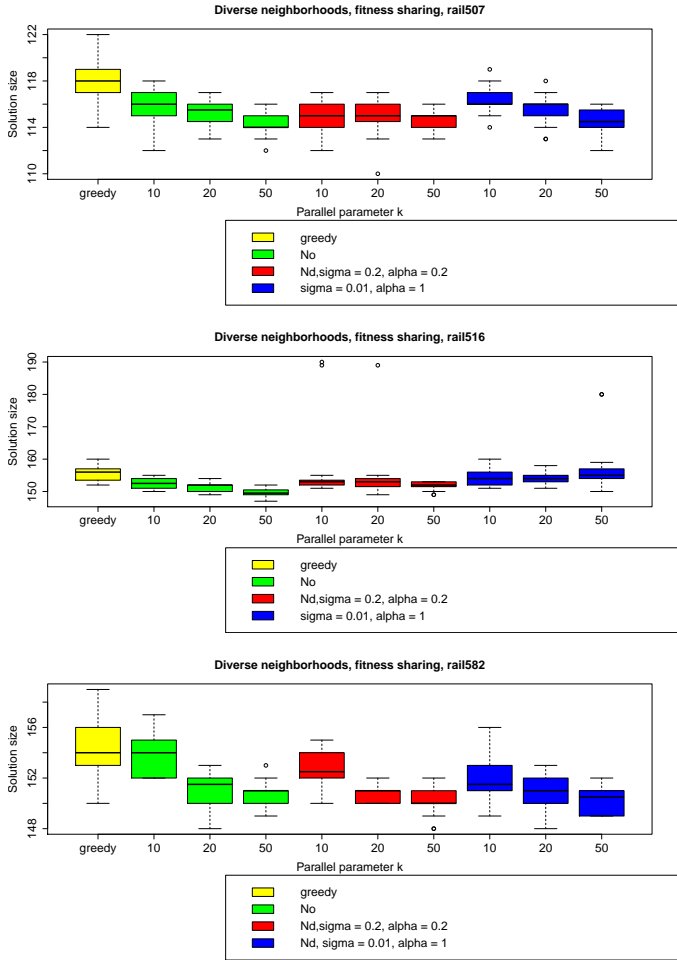
8.1 Анализ на оперативното време на различни Widening подходи с и без комуникация.

Тук ще сравним оперативното време за различни подходи на Разширяване с и без комуникация.



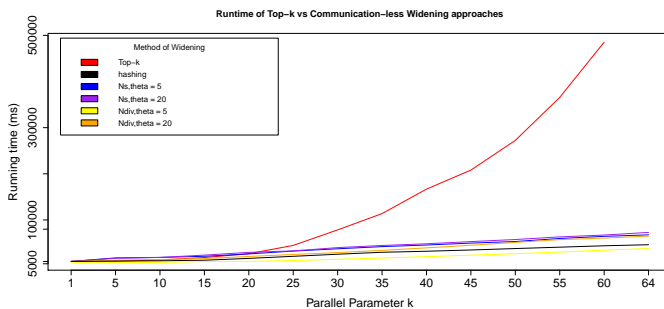
Фигура 8.5: Widening чрез разнородни околности за различни стойности на параметъра на разнородност δ , параметъра на Разширяване k и размера на околността θ . За оценяване на разнородността на моделите е използвана мярката за дистанция на Jaccard.

Експериментите по-горе оценяват основно различни свойства на решенията като качество на модела и подобност или разнородност на откритите решения. Обаче, една от задачите на Разширяването на даден алчен алгоритъм е и запазването на оперативното време да е същото (или кол-



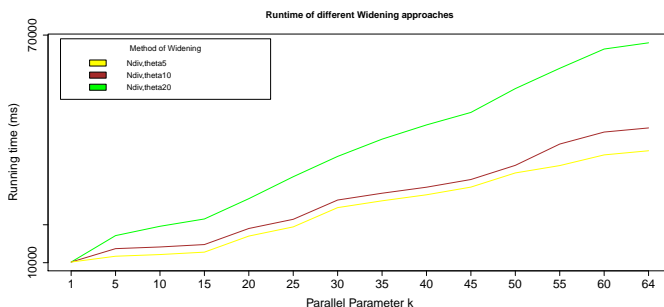
Фигура 8.6: Widening чрез околности на разнородни модели, където разнородността е постигната чрез fitness sharing. Оценен е ефектът на параметрите за Widening θ, k (размер на околност и брой на паралелни агенти при Widening) и ефекта на fitness sharing α, σ чрез задаване на различни стойности на тези параметри. Използвани са три различни множества от данни *rail507*, *rail516*, *rail582*.

кото се може по-близко до това на) като на оригиналния алчен алгоритъм. Изискването на честа комуникация между паралелните агенти за да бъ-

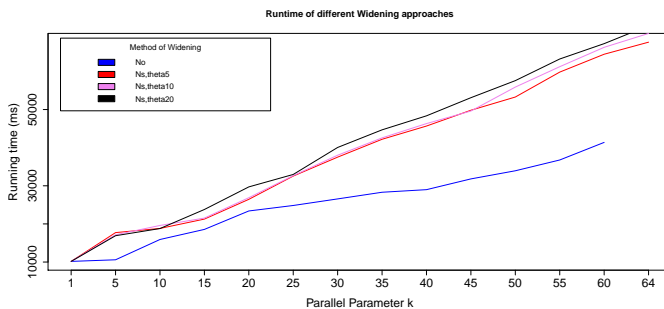


Фигура 8.7: Сравнение на оперативното време при различни стратегии за Widening на алчния алгоритъм за проблема на покритие на множество (SCP) с и без комуникация

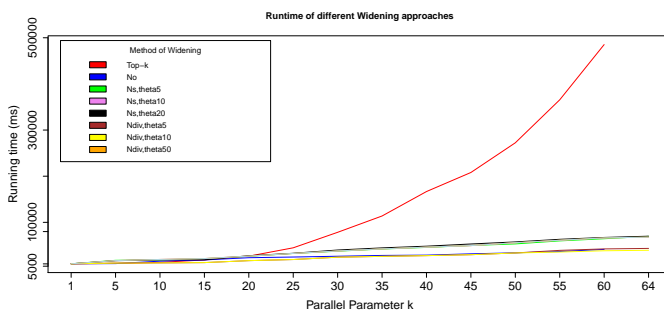
дат намерени най-добрите k решения е bottleneck за оперативното време, когато k расте. Поради това, ние се интересуваме от методи, при които не се изисква честа комуникация между паралелните агенти. Ще сравняваме стратегиите за Widening, които не използват комуникация с подхода *Top-k* Widening. Първо го контрастираме с метода на Widening, чрез хашване. Експериментите използват *rail507* на 64-ядрена машина и го повтаряме 10 пъти. Фигура 8.11 показва оперативното време на различните стратегии за Widening. .



Фигура 8.8: Сравнение на оперативното време на Widening чрез разнородните околности за различни стойности на параметъра θ , размер на околността, използвайки *rail507* data set. Размерът на околността е най-влиятелният фактор за оперативното време.



Фигура 8.9: Сравнение на оперативното време за Widening чрез околности от подобни модели за различни стойности на параметъра θ , размер на околността, използвайки *rail507*. Размерът на околността няма силно влияние върху оперативното време.



Фигура 8.10: Сравнение на оперативното време за различни методи на Widening с или без комуникация, използвайки *rail507* множество от данни.

Ние използваме метода *Top-k* Widening и го сравняваме с различните методи, използващи околности. Експериментите са изпълнени върху множеството от данни *rail507* на 64-ядрена машина и са повторени 10 пъти.

Както се очаква, the running time на различните подходи на Wideningто на алгоритмите се различава в зависимост от типа околност, която се използва. Widening, чрез околност от оптимални модели има изчислително време близко до константното, до оригиналното за алчния алгоритъм, независимо от броя на паралелните агенти. Изчислителното време на Wideningто чрез околности от подобни не се влияе силно от размера на околността, но

зависи от размера на данните, не зависи силно от броя на паралелните агенти. Това може да бъде видяно във Фигура 8.9, където за различни размери на околности изчислителните оперативното време за всеки от методите не се различава драстично. (Няма драстични разлики в оперативното време). Има два аспекта, които влияят на оперативното време на Widening чрез разнообразни околности. Единият е размерът на околността, а другият е праг на разнообразие. От фигура 8.8 се вижда, че размерът на околността е най-влиятелният фактор за оперативното време на Widening чрез разнообразни околности. Колкото по-голям е размерът на околността, толкова повече изчислително време е нужно за построяването на околността. Броят на паралелните агенти не влияе на изчислителното време, при наличие на нужните изчислителни ресурси. Размерът на threshold оказва влияние на running time, защото колкото по-голям е прагът, толкова повече сравнения са нужни на всяка стъпка.

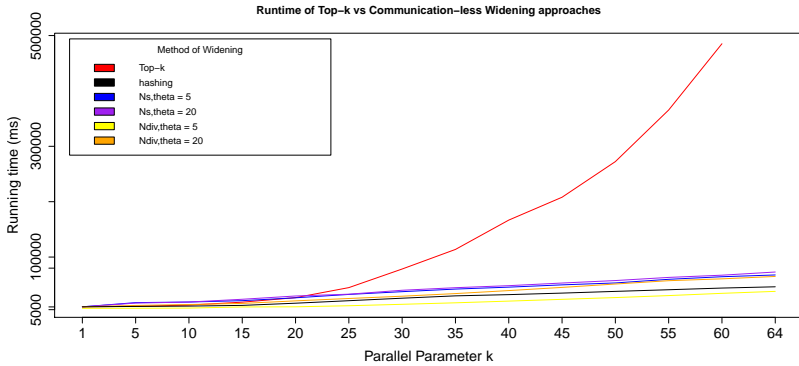
Очевидно, обаче, комуникацията между паралелните агенти остава основният проблем за оперативното време, показано във Figure 8.10. Експериментите показват, че методите, които използват комуникация имат значително по-лоши running time сравнение с методите, които не изискват комуникация между паралелните агенти.

От резултатите може да бъде видяно, че дори за малки по размер околности, резултатите (намерените модели) на Widening чрез разнородни околности са със сходно качество като тези, намерени чрез метода, използващ комуникация. Но времето на Wideningто чрез околности е много по-добро от това на метода, изискващ комуникация. Разнородността на околностите може да бъде увеличена, без да се увеличава размерът на околността и така да бъде избегнат негативния импакт /негативното влияние/ на големия размер на околността. Тъй като построяването на околностите отнема най-много време е най-изчислително интензивния аспект на Wideningто с околности, този етап може да бъде ускорен чрез различни видове preprocessing.

8.2 Анализ на оперативното време при Widening с и без комуникация.

В тази глава сравняваме оперативното време за различни подходи на Widening с и без комуникация.

Експериментите по-горе оценяват основно различни свойства на решенията като качество на модела и подобност или разнородност на откритите решения. Обаче, една от задачите на Widening на даден алчен алгоритъм е и запазването на оперативното време да е същото (или колкото се може по-близко до това на) като на оригиналния алчен алгоритъм. Изискването



Фигура 8.11: Сравнение на оперативното време при различни стратегии за Widening на алчния алгоритъм за проблема на покритие на множество (SCP) с и без комуникация

за честа комуникация между паралелните агенти за да бъдат намерени най-добрите k решения е основният фактор, влияещ на оперативното време, когато k расте. Поради това, ние се интересуваме от методи, при които не се изисква честа комуникация между паралелните агенти. Ще сравняваме стратегиите за Widening, които не използват комуникация с подхода *Top - k* Widening. Първо го контрастираме с метода на Widening, чрез хаширане. Експериментите използват *rail507* на 64-ядрена машина и го повтаряме 10 пъти. Фигура 8.11 показва оперативното време на различните стратегии за Widening.

Глава 9

Widening на алгоритъм за Rule Induction

В тази глава се дискутира проблемът за Rule Induction, класически в областта на машинното обучение. След това за конкретен алгоритъм за Rule Induction, CN2, се представят различни подходи за разширяване, както и имплементацията им. Експериментално се оценяват и сравняват подходите.

Библиография

- [1] Zaenal Akbar, Violeta N. Ivanova, and Michael R. Berthold. Parallel data mining revisited. Better, not faster. In *IDA*, 2012.
- [2] Violeta Ivanova and Michael R. Berthold. Diversity-driven widening. In *Proceedings of the 12th International Symposium on Intelligent Data Analysis(IDA 2013)*, 2013.
- [3] Violeta N. Ivanova-Rohling. Communication-less strategies for the widening of rule induction. In *Proceedings of the 19th International Conference on Computer Systems and Technologies*, CompSysTech'18, pages 33–37, New York, NY, USA, 2018. ACM.
- [4] Violeta N. Ivanova-Rohling. Neighborhood-based strategies for widening of the greedy algorithm of the set cover problem. In *Proceedings of the 19th International Conference on Computer Systems and Technologies*, CompSysTech'18, pages 27–32, New York, NY, USA, 2018. ACM.