Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs Rendszerek Tanszék

Többkerneles adat- és tudásfúziós módszerek a gyógyszerkutatásban

Tézisfüzet

Bolgár Bence

Témavezető:
Dr. Antal Péter

Budapest, 2017

1. Előzmények és célkitűzések

A heterogén adatok és tudás integrációjának problémáját számos megfogalmazásban jelent meg az elmúlt hatvan évben; megoldására több tudományterületen javasoltak elméleti és gyakorlati eszközöket, ideértve az irányításelméletet, adatbázisokat, gépi tanulást, mesterséges intelligeciát és számos egyéb diszciplínát. A több forrásból származó, heterogén tudást integráló, nagyléptékű tudásbázisok felhasználása különösen a mesterséges intelligencia területén vált hangsúlyossá, például válaszkereső, automatizált következtető, döntéstámogató és szakértő rendszerekhez kapcsolódóan.

1.1. Adat- és tudásfúzió

A tudásfúzió egyik központi kérdése a reprezentáció, azaz egy olyan „közös nyelv" megalkotása, amely kellő kifejezőerővel bír a különböző tudásformák formalizációjára, ám emellett hatékony tanulást és következtetést is lehetővé tesz. A korábban javasolt formális rendszerek elsősorban a komplex relációk reprezentációjának képességében, a következtető algoritmusok számítási hatékonyságában, valamint a hiányzás, ellentmondások és bizonytalanság kezelésében különböznek. Ilyen keret például a formális logika, amelyet széles körben, többek között orvosbiológiai vagy jogi tudásalapú szakértő rendszerekben használtak fel, az induktív oldalon pedig a bayesi statisztika, szemantikus integráció és hasonlóság-alapú megközelítések váltak elterjedtté.

Valószínűségi megközelítések. A bayesi statisztika a valószínűségszámítás nyelvét használja fel a bizonytalan háttértudás és megfigyelések axiomatikus alapokon nyugvó kombinációjára. A háttértudás *a priori* eloszlások formájában fogalmazható meg, a teljes modellt pedig az együttes eloszlás adja. A bizonytalanság direkt és axiomatikus kezelésén túl a bayesi modellek a következtetés számos formáját lehetővé teszik; használhatók deduktív, induktív és nem-monoton következtetésre is (azaz az új evidenciák érvényteleníthetnek korábbi konklúziókat). Bár az együttes eloszlásból tetszőleges célváltozó marginális eloszlása megkapható, így a bayesi modellekben történő következtetés is elméletileg triviális, a gyakorlatban ez rendszerint analitikusan nem kezelhető és közvetlenül nem kiszámítható integrálokhoz vezet.

Szemantikus gráfok. A szemantikus gráfok némiképp különböznek más tudásreprezentációs sémáktól, mivel nem rendelkeznek „alapértelme-

zett" következtető algoritmusokkal, azonban sokkal rugalmasabb módot nyújtanak komplex relációk formalizálására. Hagyományos gráfelméleti szempontból olyan gráfoknak tekinthetők, amelynek csúcsai és élei szemantikus információt hordoznak, azaz különböző típusokba tartozhatnak és attribútumokkal rendelkezhetnek. A gyakorlatban a csúcsok tetszőleges fogalmakat, az élek pedig közöttük létező kapcsolatokat reprezentálnak, amelyeket az adatbázis-séma definiál. A tudásmenedzsment területén általánosan használt a Resource Description Framework (RDF) keret, amely alapelemként alany–állítmány–tárgy hármasokat alkalmaz, az adódó gráfadatbázis pedig speciális lekérdező nyelvek segítségével használható.

Esetalapú következtetés hasonlóságok felhasználásával. A gépi tanulásban a valószínűségek mellett a hasonlóságok váltak széles körben alkalmazott egységesítő keretté, amely először a legközelebbi szomszéd (nearest neighbor) sémáknál jelent meg [DGL97], ma pedig a kernel módszerek alapját képezi. A hasonlóságok az adatfúziós kutatásokban is központi szerepre tettek szert Pavlidis és mtsai munkássága nyomán, amely a korai, késői és köztes integráció kategóriáit fektette le [Pav+02]. A korai integráció a vektoriális reprezentációk adatszintű, egyszerű konkatenáción alapuló fúzióját jelenti. A késői integráció döntés-szinten történik, azaz a fúzió a következtetés eredményeit kombinálja például szavazásos módszerek felhasználásával. A köztes integráció az adatoknak valamilyen átmeneti reprezentációit kombinálja, leggyakrabban páronkénti hasonlósági mátrixok (kernelek) súlyozott átlagával, vagy hálózatok integrációjával.

1.2. Nyitott kérdések a gyógyszerkutatásban

Az adat- és tudásfúziós technikák fontos alkalmazási területe a gyógyszerkutatás, ahol a kémiai szerkezetek, célpont-profilok, mellékhatások, gyógyszerhasználati adatok vagy génexpressziós mérések kulcsfontosságú információforrások lehetnek az egyes tulajdonságok predikciójának szempontjából; számítási modellekben történő együttes felhasználásuk azonban máig nyitott kérdés. A „gyógyszerfelfedezés" (*drug discovery*) kifejezés olyan jelölt molekulák felderítését takarja, amelyek potenciálisan törzskönyvezett gyógyszerekké válhatnak, de további kísérletek szükségesek tulajdonságaik feltárására. A gyógyszerfelfedezés hagyományosan szerencsés „rábukkanások" útján történt, míg a nagy áteresztőképességű szűrések (HTS) csak később, a '90-es évek vége felé épültek be a munkafolyamatba. Ez az experimentális megközelítés nagy méretű molekuláris könyvtárakat használ, ahol

a molekulák aktivitását biológiai szempontból releváns célpontokon, például receptorokon vagy enzimeken mérik. A legígéretesebb találatok (*hit*) megerősítés után ún. vezérmolekulává (*lead*) lépnek elő, majd a kívánt tulajdonságaik fokozását célzó szerkezeti optimalizáción esnek át (*lead optimization*). A folyamat végeredményeként a preklinikai gyógyszerfejlesztési folyamatba lépő jelölt molekuák állnak elő.

A bioinformatika más területeihez hasonlóan az új méréstechnikák hatalmas mennyiségű kísérleti adatot termeltek, amelynek menedzselése, integrációja, elemzése és prediktív modellek építése a kemoinformatika feladata. A technológiai fejlődés ellenére a HTS módszerek továbbra is költségesek, így szükségessé vált a gépi tanulás eszköztárát felhasználó komplementer *in silico* metodológiák fejlesztése.

A gyógyszerfejlesztés költségeinek emelkedésével és az évente engedélyezett új molekuláris entitások (NME) számának csökkenésével a gyógyszeripar alternatív stratégiákat fogalmazott meg [Tob09]. Ezek egyike az Open PHACTS projekt [Wil+12], amely vezető cégek és akadémiai intézmények együttműködésével valósult meg, és egy közös, nyilvános gyógyszerkutatási platform létrehozását célozta. Ez egy átfogó, számos orvosbiológiai adatbázist integráló szemantikus tudásbázist, nyílt forráskódú eszközöket, sőt, privát gyógyszeripari adatokat is tartalmaz, lehetővé téve a gyógyszerfelfedezési folyamat szűk keresztmetszeteinek megkerülését.

A gépi tanulási metodológiák különösen alkalmasak e hatalmas, heterogén adatkincs kezelésére. A fizikai tulajdonságok, interakciók vagy aktivitási értékek jóslására szolgáló prediktív modelleket már-már rutinszerűen alkalmazzák. A gyógyszer–célpont interakció-predikciós feladatot már a '90-es években megkísérelték neurális hálózatokkal [Köv+99], illetve kernel módszerekkel megoldani [Bur+01]. A hasonlóság-alapú technikákat a '90-es években virtuális screening kísérletekben használták fel [WBD98]; az ezredfordulón a molekuláris dokkolási szimulációk terjedtek el [Kit+04]; a 2000-es évektől a mátrixfaktorizációs módszereket adaptálták erre a célra [Yam+08]. Ahogy az adat- és tudásfúzió egyre nagyobb hangsúlyt kapott [WSN07; Agr+07], a több forrásból származó háttértudás beépítése vált a fő kutatási iránnyá, és vezetett a prediktív teljesítmény javulásához [Zhe+13; GKK13].

1.3. Célkitűzés

A kutatás célja adat- és tudásfúziót megvalósító gépi tanulási modellek kidolgozása volt gyógyszerkutatás támogatására. Az értekezés első része egy általános munkafolyamatot ad a gyógyszer-újrapozicionálás, vagyis már törzskönyvezett gyógyszerek új indikációkban történő alkalmazhatóságának jóslására, amelynek alapját az ℓ^p-regularizált többkerneles szupportvektor-gépek egyosztályos adaptációja adja. Megmutatjuk az algoritmus előnyeit és alkalmazhatóságának határait elméleti szempontból és numerikus kísérletek útján is.

A második rész a leíró és hálózat-jellegű tudás integrációját tárgyalja az orvosbiológiai tudásfúzióban. A javasolt többkerneles metrika-tanulási algoritmus képes hasonlósági mátrixokat és páronkénti ekvivalencia-relációkat befogadni. A feladat egy konvex optimalizációs problémára vezet, amelynek megoldására egy sztochasztikus gradiens-projekciós eljárást dolgozunk ki, amelynek bemutatjuk és kiértékeljük GPU-alapú implementációját is. A kvalitatív és kvantitatív kiértékelés során megmutatjuk, hogy az algoritmus prediktív teljesítmény szempontjából felveszi a versenyt a korábbi egyosztályos modellekkel, miközben egy konzisztens, többosztályos megoldást ad; teljesítménye egyúttal felülmúlja a korábbi metrika-tanulási algoritmusokét.

A harmadik rész a heterogén információforrások fúzióját tárgyalja nagy számú gyógyszer–célpont interakció együttes jóslásában. Kiterjesztjük a bayesi mátrixfaktorizációs módszertant entitásszintű háttértudás befogadására Gauss-folyamatok felhasználásával, valamint megoldást adunk az interakciószintű háttértudás integrációjára és nem véletlenszerű hiányzás kezelésére. Leírjuk továbbá egy korábbi logisztikus mátrix-faktorizációs módszer többkerneles, bayesi adaptációját, amely egyesíti a többkerneles tanulás, súlyozott megfigyelések, Laplace-típusú regularizáció előnyeit, valamint lehetővé teszi bináris gyógyszer-célpont interakciók valószínűségének explicit modellezését. A modellben való hatékony következtetéshez kidolgozunk egy variációs approximációs sémát, amelynek GPU-alapú implementációja a számítások jelentős gyorsulását eredményezi. A numerikus kiértékelés során megmutatjuk, hogy mindkét módszer prediktív teljesítménye meghaladja a korábbi módszerekét standard benchmark adathalmazokon, valamint lehetővé teszik gyógyszer-promiszkuitás jóslását.

2. Kutatási módszertan és új tudományos eredmények

Ebben a szakaszban ismertetjük az alkalmazott kutatási módszertant a mélyebb technikai részletek bemutatása nélkül. Az új tudományos eredményeket az alszakaszok végén, tézisek formájában is kimondjuk.

2.1. Prioritizálás többkerneles szupportvektor-gépekkel

Ebben az alszakaszban bemutatunk egy, az *in silico* gyógyszer-újrapozicionálást támogató kernel-alapú metodológiát, amely a gyógyszermolekulák különféle reprezentációit használja fel, valamint megoldást ad ezen heterogén adatforrások statisztikailag optimális kombinációjára, illetve a gyógyszerek kombinált hasonlóság-alapú rangsorolására (prioritizációjára). Megvizsgáljuk az eljárás előnyeit és korlátait mind elméleti, mind gyakorlati szempontból. A ligand-alapú virtuális szűrési módszerekkel analóg módon, adott indikációt a benne alkalmazott gyógyszerek halmazával karakterizálva a módszer kimeneteként adódó prioritizált listák felhasználhatók az indikációban alkalmazható további gyógyszerek jóslására.

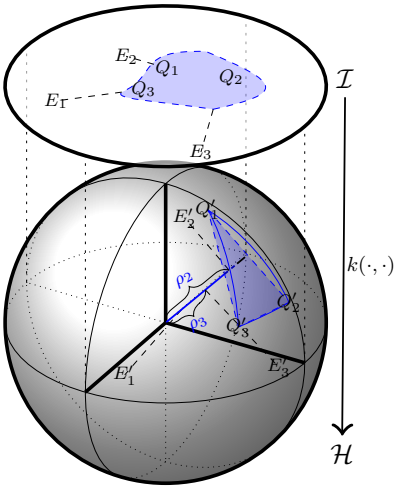
Az egyosztályos szupportvektor-gépeket (SVM) Schölkopf és mtsai javasolták olyan kis méretű régiók becslésére nagy dimenziójú vektorterekben, ahol a tanító minták (például adott indikációban alkalmazott gyógyszerek vektoriális reprezentációi) megtalálhatók [SS01]. Az algoritmus a kernel által definiált térben olyan, $\boldsymbol{w}$ által paraméterezett hipersíkot számít ki, amely a kérdéses $\phi(\boldsymbol{x}_i)$ reprezentációkat a legnagyobb margóval elválasztja az origótól (1. ábra). Ehhez a következő kvadratikus problémát oldja meg:

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi},\rho} \quad & \frac{1}{2}\left\|\boldsymbol{w}\right\|^2 - \rho + \frac{1}{\nu P}\sum_i \xi_i \\ s.t. \quad & \boldsymbol{w}^T\phi(\boldsymbol{x}_i) \geq \rho - \xi_i, \quad \boldsymbol{\xi} \geq 0, \end{aligned} \tag{1}$$

ahol a margót ρ jelöli, a modell komplexitását pedig ν szabályozza. A duál

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{K}\boldsymbol{\alpha} \\ & 0 \leq \boldsymbol{\alpha} \leq 1, \quad \mathbf{1}^T\boldsymbol{\alpha} = \nu P, \end{aligned}$$

ahol $\boldsymbol{K}$ a kernel mátrix, másképpen a páronkénti hasonlóságokat tartalmazó mátrix. Moreau és mtsai. egy adott entitás tanító halmazhoz vett általános



1. ábra. Prioritizálás egyosztályos SVM-mel. Konstans diagonálist feltételezve a kernel mátrixban, az entitások reprezentációi a $\mathcal{H}$ kernel térben egy hipergömb felszínén helyezkednek el. A tanító minták Q halmaza maximális margóval szeparált az origótól. A további minták E halmaza sorrendezhető a szeparáló hipersík normálvektorára vett projekcióik és az origó távolsága alapján, ami az ábrázolt esetben az E_2, E_3, E_1 sorrendet adja.

„hasonlóságának" mérésére a kapott hipersík normálvektorára vett projekció és az origó távolságát javasolta:

$$f(\boldsymbol{x}) = \frac{\sum_i \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha}}},$$

amely így az entitások prioritizációjára is felhasználható [MT12; Yu+10].

Felhasználva Vishwanathan eredményeit [Vis+10], az (1) egyenlet ℓ^p-regularizált többkerneles változatát a következőképpen adjuk meg:

$$\begin{aligned} \min_{\boldsymbol{w},\rho,\boldsymbol{\xi},\boldsymbol{m}} \quad & \frac{1}{2} \sum_k \frac{\|\boldsymbol{w}_k\|^2}{m_k} - \rho + \frac{1}{\nu P} \sum_i \xi_i + \frac{\lambda}{2} \|\boldsymbol{m}\|_p^2 \\ s.t. \quad & \sum_k \boldsymbol{w}_k^T \phi_k(\boldsymbol{x}_i) \geq \rho - \xi_i, \quad \boldsymbol{\xi} \geq 0, \quad \boldsymbol{m} \geq 0, \end{aligned}$$

amely a következő differenciálható duálra vezet:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & -\frac{1}{8\lambda}\left(\sum_k \left(\boldsymbol{\alpha}^T \boldsymbol{K}_k \boldsymbol{\alpha}\right)^q\right)^{\frac{2}{q}} \\ s.t. \quad & \mathbf{1}^T\boldsymbol{\alpha} = \nu P, \quad 0 \leq \boldsymbol{\alpha} \leq 1, \end{aligned}$$

ahol már több $\boldsymbol{K}_k$ kernelt kombinálunk.

Ez a sorrendezési stratégia több szempontból is nehézkes. Megmutatjuk, hogy az m_k kernel súlyok monoton növekvő függvényei a $\boldsymbol{K}_k$-beli értékeknek, ami lényegében a tanító halmaz különböző információforrásokban vett „önhasonlóságán" alapuló súlyozási sémához vezet. Másképpen, a súlyok a megfelelő kernel terekben nézett befoglaló gömbök sugara alapján alakulnak, amelyből az alábbi következtetéseket vonjuk le:

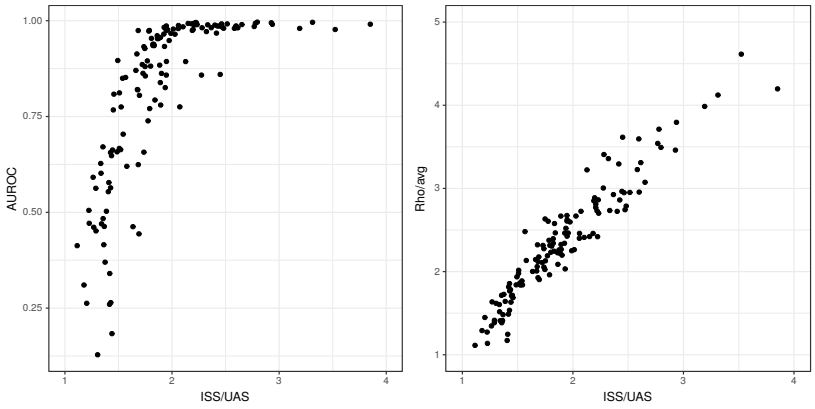
- Csupán a kernel mátrixok skálázásával is jelentős torzítás vihető be, amely egyúttal rámutat a normalizáció kulcsfontosságú szerepére.
- A korábbi vélekedésekkel ellentétben önmagában a többkerneles tanulási algoritmusokra hagyatkozni az optimális kombináció megtalálására nem megalapozott.

Szintén megmutatjuk, hogy a tanító halmaz heterogenitása erősen behatárolja a prediktív teljesítményt. Szemléletesen, egy túlzottan heterogén tanító halmaz elemeinek befoglaló hipergömbje a kernel térben nagy sugárral bír, így a további entitások nagy részét is befogja. Ekvivalens módon, a tanító minták eloszlásának becsült tartója nagy, a minták pedig ennek határán helyezkednek el. Ez lényegében megfordítja a sorrendezési stratégiát: a tanító halmaz a „hasonló" entitásokkal együtt a lista legvégére kerül. Ennek formális vizsgálatához először bizonyítjuk a következő állítást:

Állítás. *Az ℓ^p-regularizált egyosztályos SVM margója arányos a duál célfüggvénnyel, és a következő formában írható:*

$$\rho = \frac{1}{\nu P}\boldsymbol{\alpha}^T\bar{\boldsymbol{K}}\boldsymbol{\alpha} = \frac{2\lambda}{\nu P}\left\|\boldsymbol{m}\right\|_p^2 = -\frac{4}{\nu P}\mathcal{D}(\boldsymbol{\alpha}),$$

ahol $\bar{\boldsymbol{K}} = \sum_k m_k \boldsymbol{K}_k$.



2. ábra. Bal oldal: az ISS/UAS hányados előrejelzi a heterogén gyógyszercsoportok okozta alacsony AUROC értékeket. A kísérletben 1041 kémiai szerkezeti leírót és 135 gyógyszercsoportot használtunk. Jobb oldal: a valódi margó/átlagos távolság és az ISS/UAS heurisztika erős korrelációt mutat.

Ezen eredmény felhasználásával bevezetünk egy heurisztikát, amely a margó/átlagos távolság hányadost becsüli, és felhasználható heterogén tanító halmazok kiszűrésére, valamint a gyenge prediktív teljesítmény előrejelzésére.

Definíció. *Az ISS/UAS hányados*

$$\frac{ISS}{UAS} := \frac{\nu P}{T} \frac{\mathbf{1}^T \bar{\boldsymbol{K}} \mathbf{1}}{\mathbf{1}^T \bar{\boldsymbol{K}}' \mathbf{1}},$$

ahol $\bar{\boldsymbol{K}}'$ a kombinált kernel mátrix tanító mintáknak megfelelő oszlopait tartalmazza.

Megmutatjuk, hogy ez a hányados jól közelíti a valódi margó/távolság hányadost, valamint korrelál a prediktív teljesítményt jellemző AUROC mértékkel (2. ábra).

Végül bemutatunk egy, a gyógyszerkutatást támogató általános munkafolyamatot, ami magában foglalja a fejlesztett algoritmust és egy feldúsulási elemzési technikát, így lehetővé teszi új indikációk jóslását is gyógyszerek vagy gyógyszerhalmazok számára.

I.1. tézis. Kidolgoztuk az ℓ^p-regularizált többkerneles szupportvektor-gépek prioritizációs adaptációját, amely képes gyógyszer-újrapozicionálási jelölteket jósolni heterogén információforrások felhasználásával.

I.2. tézis. A módszer elméleti elemzésével és numerikus kiértékelésével megmutattuk annak előnyeit és határait, valamint kidolgoztunk egy heurisztikát, amely az alkalmazhatóságát mutatja.

A kutatás a Semmelweis Egyetem Szerves Vegytani Intézetével közösen folyt. A társszerzők Dr. Arany Ádám, Dr. Temesi Gergely, Dr. Balogh Balázs, Prof. Mátyus Péter és Dr. Antal Péter. A közös munka számos eredménye itt nem kerül bemutatásra; a disszertáció kizárólag a szerző saját hozzájárulását tartalmazza. A tézisek a következő publikációkon alapulnak: [1], [2], [3].

2.2. Kernelek és ekvivalencia-relációk fúziója

Az orvosbiológiai adatok heterogenitása gyakran kettős formában fejezhető ki. A korábbi megközelítések páronkénti hasonlóságokat használnak, amely egységesítő, de végső soron „deskriptív" jellegű, és rendszerint az entitások különböző leírásain ill. reprezentációin alapul (lásd az előző rész módszereit). A heterogenitás kezelésére újabb megközelítést jelent az ekvivalencia-relációk felhasználása, amely a rendszer- és hálózatbiológiai paradigmán keresztül újította meg a bioinformatikai kutatásokat. Az utóbbi szemléletmód inkább „relációs", semmint „deskriptív", mivel nem magukra az entitásokra, hanem ezek komplex interkciós mintázataira fókuszál.

Ez az irányvonal különösen erős az *in silico* gyógyszerkutatás területén. A többféle kémiai hasonlóság együttes alkalmazása már lassan két évtizede használt technika a virtuális szűrések területén [WBD98]. Az ekvivalencia-relációk természetes módon adódnak a legkülönfélébb esetekben, például gyógyszerkombinációk, interakciók vagy közös indikációk kutatásánál. Ezen két alapvetően különböző tudásforma kombinációja matematikailag megalapozott keretben nyílt kihívás maradt. Ebben az alszakaszban kidolgozunk egy távolságmetrika-tanulási (DML) keretrendszert, amely alkalmas ilyen kettős priorok befogadására.

A Mahalanobis-távolság bizonyos értelemben a legáltalánosabb távolságmérték vektortereken, és így a legtöbb DML módszer alapját is adja.

Definíció. *Legyen $\mathcal{X}$ az $\mathbb{R}^D$ vektortér. A Frobenius-szorzat segítségével az $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ minták Mahalanobis-távolsága*

$$\begin{aligned} d(\boldsymbol{x}_i, \boldsymbol{x}_j) &= \|\boldsymbol{L}\boldsymbol{x}_i - \boldsymbol{L}\boldsymbol{x}_j\| \\ &= \sqrt{\left\langle (\boldsymbol{x}_i - \boldsymbol{x}_j), \boldsymbol{W}(\boldsymbol{x}_i - \boldsymbol{x}_j) \right\rangle} \\ &= \sqrt{\left\langle \boldsymbol{W}, (\boldsymbol{x}_i - \boldsymbol{x}_j) \otimes (\boldsymbol{x}_i - \boldsymbol{x}_j) \right\rangle_F}. \end{aligned}$$

ahol $\boldsymbol{W} = \boldsymbol{L}^T\boldsymbol{L}$, azaz $\boldsymbol{W}$ pozitív definit.

A legtöbb felügyelt DML formalizációban a $\boldsymbol{W}$ Mahalanobis-mátrixot tanulják, a tanítóminták pedig *a priori* kényszerekként lépnek be, egyfajta „strukturált lekérdezést" alkotva.

- Páronkénti kényszerek: ekvivalencia-relációk (entitáspárokra definiált „kötődési" és „nem-kötődési" kényszerek),
- Triplet-alapú kényszerek: relatív távolságok (pl. $\boldsymbol{x}_i$ legyen közelebb $\boldsymbol{x}_j$-hez, mint $\boldsymbol{x}_k$-hoz).

A DML eljárások gyakran alkalmaznak regularizáló tagot a Mahalanobis-mátrixon, ami SVM-hez igen hasonló optimalizációs problémákra vezet. Utóbbiakkal ellentétben azonban itt a pozitív definit tulajdonságot is biztosítani kell, amely rendszerint a legnagyobb kihívás. Ismert megoldások például a pozitív szemidefinit kúpra való projekció minden iterációban (LMNN [WS09]), megfelelően megválasztott divergenciák alkalmazása a regularizáló tagban (ITML [Dav+07]), vagy egyszerűen a tulajdonság elhagyása [Wan+11].

Legyen $\left\{ \left(\{(\boldsymbol{x}_{ik}, \boldsymbol{z}_{ik})\}_{k=1}^{R}, y_i \right) \right\}_{i=1}^{P}$ a tanító halmaz, ahol k az információforrásokat, i pedig a tanító mintákat indexeli, $\boldsymbol{x}_{ik}, \boldsymbol{z}_{ik} \in \mathbb{R}^{D_k}$, $y_i \in \{-1, +1\}$. A cél, hogy amennyiben az $\boldsymbol{x}_i$, $\boldsymbol{z}_i$ entitások azonos osztályba tartoznak, ($y_i = 1$), távolságuk kicsinek, különböző entitások esetén pedig ($y_i = -1$) a távolságuk nagynak adódjon. Jelölje d_i az i. mintapár négyzetes távolságát, azaz

$$d_i = \sum_k \left\langle \boldsymbol{W}_k, (\boldsymbol{x}_{ik} - \boldsymbol{z}_{ik}) \otimes (\boldsymbol{x}_{ik} - \boldsymbol{z}_{ik}) \right\rangle_F,$$

ahol $\boldsymbol{W}_k$ a k. metrika mátrixa és $\langle \cdot, \cdot \rangle_F$ a Frobenius-szorzat.

Algoritmus A duál probléma megoldása projektált gradiens eljárással.

Initialize $d_i^0,\ \boldsymbol{\alpha} \leftarrow 0,\ \boldsymbol{m} \leftarrow 0,\ (\boldsymbol{Q}_k\boldsymbol{\alpha})_i \leftarrow 0,\ b \leftarrow 0.$
while $\boldsymbol{\alpha}$ not optimal **do**
  for all i **do**
    for all k **do**
      for all j **do in parallel**
        Compute $\boldsymbol{Q}_{k_{ij}} = y_i y_j \left\langle \tilde{\boldsymbol{x}}_{ik} - \tilde{\boldsymbol{z}}_{ik}, \tilde{\boldsymbol{x}}_{jk} - \tilde{\boldsymbol{z}}_{jk} \right\rangle^2.$
    Compute $A = \frac{1}{2\lambda} \sum_k \boldsymbol{Q}^2_{k_{ii}}.$
    Compute $B = \frac{3}{2\lambda} \sum_k \boldsymbol{Q}_{k_{ii}} (\boldsymbol{Q}_k\boldsymbol{\alpha})_i.$
    Compute $C = \sum_k m_k \boldsymbol{Q}_{k_{ii}} + \frac{1}{\lambda} \left((\boldsymbol{Q}_k\boldsymbol{\alpha})_i \right)^2.$
    Compute $D = \sum_k m_k \left(\boldsymbol{Q}_k\boldsymbol{\alpha}\right)_i + y_i(r - d_i^0) - \rho.$
    $\delta^* \leftarrow$ largest root of $A\delta^3 + B\delta^2 + C\delta + D$
      projected into the feasible region.
    $\alpha_i \leftarrow \alpha_i + \delta^*.$
    $b \leftarrow b + \delta^* y_i \left\| \boldsymbol{x}_{ik} - \boldsymbol{z}_{ik} \right\|^2$
    for all k **do**
      $m_k \leftarrow m_k + \frac{1}{2\lambda} \left(\boldsymbol{Q}_{k_{ii}} {\delta^*}^2 + 2(\boldsymbol{Q}_k\boldsymbol{\alpha})_i \delta^* \right).$
    for all k, j **do in parallel**
      $(\boldsymbol{Q}_k\boldsymbol{\alpha})_j \leftarrow (\boldsymbol{Q}_k\boldsymbol{\alpha})_j + \boldsymbol{Q}_{k_{ij}} \delta^*.$

A primál problémát a következőképpen adjuk meg:

$$\min_{\boldsymbol{W},\boldsymbol{m}} \quad \frac{1}{2} \sum_k \frac{\left\| \boldsymbol{W}_k - \boldsymbol{W}_k^0 \right\|_F^2}{m_k} + \frac{\lambda}{2} \left\| \boldsymbol{m} \right\|_p^2$$
$$s.t. \quad y_i(r - d_i) \geq \rho, \quad m_k > 0, \quad \boldsymbol{W}_k \succeq 0,$$

ahol $\boldsymbol{W}_k^0$ a k. *a priori* metrika mátrixa, m_k a megfelelő súly, λ egy trade-off paraméter és r, ρ pedig a hasonló és nem hasonló entitások távolságának arányát szabályozza. Az egyenlet azt az intuitív preferenciát tükrözi, hogy az *a priori* metrikákhoz (pl. egységmátrix) minél közelebb maradva a tanult metrika elégítse ki a definiált kényszereket, amelyek átfogalmazva így írhatók:

$$\begin{aligned} (d_i, y_i = +1) \quad &\Rightarrow \quad d_i \leq r - \rho, \\ (d_i, y_i = -1) \quad &\Rightarrow \quad d_i \geq r + \rho. \end{aligned}$$

A duál feladatot a következő formában kapjuk:

$$\max_{\boldsymbol{\alpha}} \quad -\frac{1}{2\lambda}\left(\sum_k \left(\frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{Q}_k \boldsymbol{\alpha}\right)^q\right)^{\frac{2}{q}} - \sum_i \alpha_i y_i \left(r - d_i^0\right) + \boldsymbol{\rho}^T \boldsymbol{\alpha}$$
$$s.t. \quad \boldsymbol{\alpha} \geq 0, \quad \sum_i \alpha_i y_i \left\|\boldsymbol{x}_{ik} - \boldsymbol{z}_{ik}\right\|^2 \leq 0,$$

ahol d_i^0 a prior metrikák szerinti távolság, $\boldsymbol{Q}_k$ pedig az előző alszakasz $\boldsymbol{K}_k$ mátrixaival analóg szerepet tölt be. A $p = 2$ regularizáló tag felhasználásával javasolunk egy párhuzamos projektált gradiens-alapú algoritmust, amely a fenti feladatot minden lépésben egy harmadfokú polinom megoldására vezeti vissza, a módosított duál változót pedig visszavetíti a megengedett tartományba.

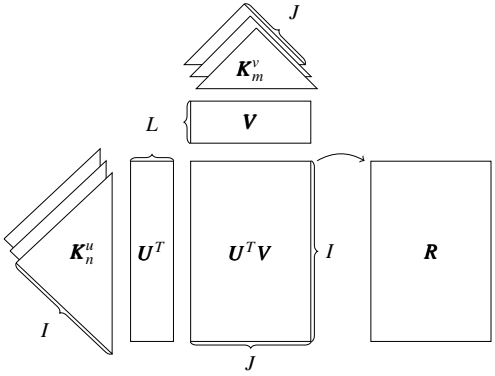
Megmutatjuk, hogy a konvex és differenciálható célfüggvény kiváló párhuzamosítási lehetőségeket rejt. 10^4 párból álló tanító halmazra a GPU-alapú implementáció két nagyságrendnyi gyorsulást ér el Radeon HD7970 típusú grafikus kártyán egy asztali Intel Core i5-2500K CPU-hoz képest. Demonstráljuk az algoritmus prediktív teljesítményét $100\times70\%-30\%$ keresztkiértékeléses keretben, a gyógyszer-újrapozicionálási feladatban, amelynek során gyógyszercsoportba tartozást jósolunk kémiai szerkezetek, mellékhatás-profilok és célpontprofilok felhasználásával. Annak ellenére, hogy az algoritmus a korábbinál jelentősen nehezebb többosztályos problémát old meg, prediktív teljesítménye felveszi a versenyt az előző szakasz egyosztályos módszerével és meghaladja más DML eszközök teljesítményét.

II.1. tézis. Kidolgoztunk egy többkerneles távolságmetrika-tanulási algoritmust, amely képes páronkénti ekvivalencia-relációk és hasonlósági mátrixok befogadására.

II.2. tézis. Kidolgoztunk egy hatékony, projektált gradiens-alapú stratégiát a célfüggvény optimalizálására, ennek GPU-alapú implementációjával együtt.

II.3. tézis. Megmutattuk, hogy a módszer prediktív teljesítménye eléri az egyosztályos modellekét és meghaladja más DML tecnikákét, miközben konzisztens, többosztályos megoldást ad.

A tézisek a következő publikáción alapulnak: [5].



3. ábra. Mátrixfaktorizáció háttérinformációval. Az $\boldsymbol{U}$ és $\boldsymbol{V}$ faktorok szorzata az $\boldsymbol{R}$ mátrix alacsony rangú approximációja. A tanulás célja a faktorok oszlopainak meghatározása a kernelek formájában megadott háttértudás figyelembe vételével.

2.3. Többkerneles bayesi mátrix-faktorizáció

Bár a HTS kísérletek időigényes és költséges volta nagyban behatárolja a gyógyszer–célpont interakciós (DTI) adatbázisok növekedését, ilyen korlátok sokkal kevésbé érvényesek a komplementer háttértudás forrásaira. A kémiai szerkezeti leírók, betegek által jelentett és hivatalos mellékhatásadatok, off-label gyógyszerhasználati adatok, protein-protein hálózatok és egyéb bio- illetve kemoinformatikai adatbázisok ma is hatalmas sebességgel növekednek. A gyógyszer-újrapozicionáláshoz hasonlóan a gyógyszer–célpont interakció-predikcióban is felhasználhatók ezen információforrások, ami az adat- és tudásfúziós megközelítéseket különösen fontossá teszi. Ebben az alszakaszban a DTI predikciós folyamatot támogató bayesi többkerneles fúziós eljárásokat dolgozunk ki.

A gépi tanulásban a mátrixfaktorizációs módszerek a diadikus adatok elemzésének egyik leghatékonyabb megközelítésévé váltak. A mátrixfaktorizáció célja egy $\boldsymbol{R} \in \mathbb{R}^{I \times J}$ interakciós mátrix alacsony rangú approximációja az $\boldsymbol{U} \in \mathbb{R}^{L \times I}$ és $\boldsymbol{V} \in \mathbb{R}^{L \times J}$ faktorok szorzataként:

$$\boldsymbol{R} \approx \boldsymbol{U}^T \boldsymbol{V},$$

ahol $L \ll \operatorname{rank}(\boldsymbol{R})$, amely a látens dimenziók számát jelöli. Egy szemléletes interpretáció szerint az $\boldsymbol{U}$ mátrix oszlopai az I darab „sor-entitás" (pl. gyógyszerek) alacsony dimenziós reprezentációi, míg $\boldsymbol{V}$ oszlopai a J darab „oszlop-entitás" (pl. célpontok) reprezentációi. Az $\boldsymbol{R}$ interakciós mátrix $I \times J$ értéket (pl. bioaktivitási értékeket) tartalmaz. A legtöbb kutatásban veszteségfüggvénynek a Frobenius-távolságot választják, azaz a célfüggvény

$$\min_{\boldsymbol{U},\boldsymbol{V}} \sum_{\{(i,j):\boldsymbol{R}_{ij} \text{ known}\rangle} (\boldsymbol{R}_{ij} - \boldsymbol{u}_i^T \boldsymbol{v}_j)^2,$$

ahol $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ az $\boldsymbol{U}$ mátrix $i.$ oszlopa és a $\boldsymbol{V}$ mátrix $j.$ oszlopa. Ha $\boldsymbol{R}$ nem tartalmaz csupa nulla sort és oszlopot, a hiányzó értékek a megfelelő látens reprezentációk szorzataként prediktálhatók:

$$\boldsymbol{R}_{ij} \sim \boldsymbol{u}_i^T \boldsymbol{v}_j.$$

Az alapmodell több szempontból is korlátozott. Egyrészt, mivel a faktorok tetszőlegesen naggyá válhatnak, könnyen felbukkanhat a túlilleszkedés problémája; másrészt valós feladatokban gyakran áll rendelkezésre a sor- és oszlop-entitásokra vonatkozó további háttérinformáció az interakciók mellett, amelyek célfüggvénybe való beépítése jelentősen javíthatná a teljesítményt. Mindkét nehézség kezelhető generatív modellek alkalmazásával, amelyek megfelelő paraméter-priorokon keresztül a háttérinformáció befogadására is lehetőséget nyújtanak (3 .ábra).

Az alapmodell bayesi verzióit először Salakhutdinov és mtsai javasolták, a PMF és BPMF algoritmusok azonban még nem használtak háttérinformációt [SM08b; SM08a]. Gönen és mtsai dolgozták ki a KBMF2MKL algoritmust, amely egy teljesen bayesi modellben kombinál több információforrást többkerneles tanulás segítségével [GKK13]. A Liu és mtsai által fejlesztett NRLMF algoritmus a teljes bayesi megközelítés helyett MAP approximációt használ, a háttérinformáció Laplace-típusú regularizációs tagon keresztüli beépítésével, valamint egy logisztikus likelihood-függvény alkalmazásával, amely a Bernoulli-eloszlás súlyozott verziójával képes nagyobb fontosságot rendelni ismert gyógyszer–célpont interakciókhoz [Liu+16]. A disszertációban bemutatott első modell ezen módszerek továbbfejlesztése. Megmutatjuk, hogy a Laplace-típusú regularizáció többkerneles verziója kompatibilis a logisztikus mátrixfaktorizáció teljesen bayesi megfogalmazásával, és hatékony variációs approximációs módszert adunk a következtetésre.

A hiperparamétereket elhagyva a bináris interakciók feltételes eloszlását a következőképpen definiáljuk:

$$p(\mathbf{R}|\mathbf{U},\mathbf{V}) \propto \prod_i \prod_j \left[\left(\sigma\left(\mathbf{u}_i^T \mathbf{v}_j\right)\right)^{c\mathbf{R}_{ij}} \left(1-\sigma\left(\mathbf{u}_i^T \mathbf{v}_j\right)\right)^{1-\mathbf{R}_{ij}} \right]^{\mathbf{m}_i^u \mathbf{m}_j^v},$$

ahol σ a logisztikus szigmoid függvény, $\boldsymbol{m}$ pedig az $\boldsymbol{R}$ mátrix üres sorait és oszlopait jelzi. A faktor-mátrix oszlopaira kombinált Laplace–ℓ^2 priort téve

$$p(\mathbf{U}|\alpha^u,\gamma^u,\mathbf{K}^u) \propto \prod_i \prod_k \exp\left\{ -\frac{1}{2} \sum_n \gamma_n^u \mathbf{K}_{n,ik}^u \left\|\mathbf{u}_i - \mathbf{u}_k\right\|^2 \right\} \times \prod_i \exp\left\{ -\frac{\alpha^u}{2} \left\|\mathbf{u}_i\right\|^2 \right\},$$

amely már több $\mathbf{K}^u$ kernelt tartalmaz. $\mathbf{V}$ prior eloszlása hasonlóképpen írható. A γ_n^u súlyok optimális értékének tanulásához Gamma priort választunk:

$$p(\gamma_n^u|a,b) = \frac{b^a (\gamma_n^u)^{a-1} e^{-b\gamma_n^u}}{\Gamma(a)}.$$

Célunk a

$$p(\mathbf{U},\mathbf{V},\gamma^u,\gamma^v|\mathbf{R},\mathbf{K}_n^u,a^u,b^u,\mathbf{K}_n^v,a^v,b^v,\alpha^u,\alpha^v,c) \tag{2}$$

poszterior variációs közelítése, amelyhez bevezetjük a

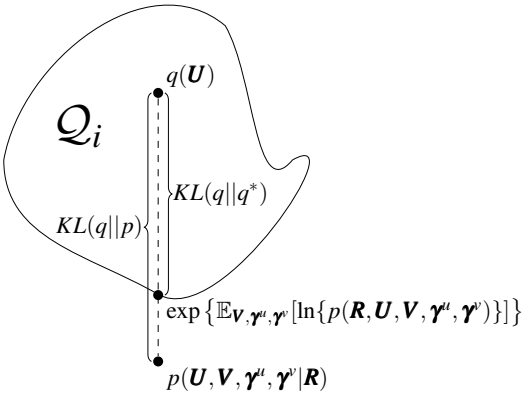
$$q(\mathbf{U},\mathbf{V},\gamma^u,\gamma^v) = q(\mathbf{U})q(\mathbf{V})q(\gamma^u)q(\gamma^v). \tag{3}$$

faktorizált variációs eloszlást. Mivel $\ln p(\boldsymbol{R})$ dekomponálható

$$\ln p(\mathbf{R}) = \mathcal{L}(q) + KL\left(q||p\right),$$

formában, a (2) and (3) kifejezések közötti Kullback–Leibler divergencia minializálása ekvivalens az $\mathcal{L}$ alsó korlát maximalizálásával (4. ábra). Az optimális eloszlás a következő nem-konjugált formában adódik:

$$\ln q^*(\mathbf{U}) = \mathbb{E}_{\mathbf{V},\gamma^u,\gamma^v}\left[\ln\left\{p(\mathbf{R},\mathbf{U},\mathbf{V},\gamma^u,\gamma^v)\right\}\right] + \text{const.}$$



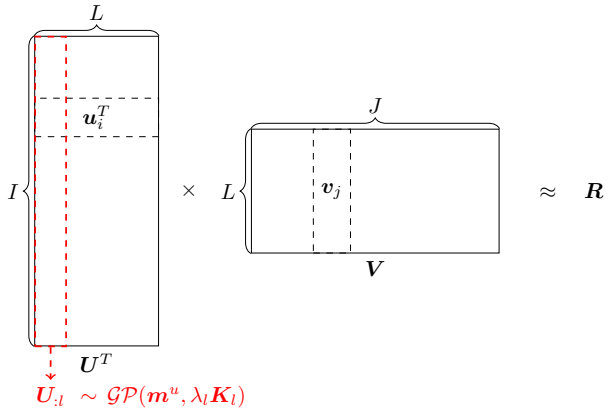
4. ábra. A (2) poszterior variációs közelítése a (3) eloszlással, az $\boldsymbol{U}$ faktormátrixra nézve.

A szimmetrizált logisztikus függvény Taylor-approximációja (Jaakkola korlát [JJ00; HHG14])

$$\sigma(z) \geq \tilde{\sigma}(z,\xi) = \sigma(\xi)\exp\left\{\frac{z-\xi}{2} - \frac{1}{2\xi}\left(\sigma(\xi)-\frac{1}{2}\right)\left(z^2-\xi^2\right)\right\},$$

amellyel alsó korlátot adunk $p(\mathbf{R}|\mathbf{U},\mathbf{V})$-re a ξ_{ij} lokális variációs paraméterek bevezetésével, tehát új $\tilde{\mathcal{L}}$ alsó korláthoz jutunk. Ez azonban már csak kvadratikus tagokat tartalmaz, azaz $q^*(\boldsymbol{U})$ normál eloszlásúnak adódik. A modell konjugálttá vált az összes változóra nézve, így a variációs faktorok a konjugált priorokra vonatkozó összefüggésekkel meghatározhatók.

A disszertációban bemutatott második modell valós-értékű interakciós pontszámok jóslását tűzi ki célul nem-véletlenszerű hiányzás mellett. Azon esetekben, amikor a megfigyelés ténye függ a megfigyelt változó értékétől, a véletlenszerű hiányzást (MAR) feltevő modellek pontatlan predikciókat adhatnak. Ilyen függésre lehet példa egy irreleváns gyógyszer–célpont interakció, amely nem kerül be a mérésbe, vagy egy mért erős interakció, amely nem kerül közlésre. A modell Gauss-folyamat priorokat alkalmaz több kernel beépítésére (Figure 5), valamint egy háttértudás-modellt interakció-szintű adatok befogadására.



5. ábra. Háttérinformáció beépítése a sorokon értelmezett Gauss-folyamat priorokkal, amely a $\boldsymbol{K}_l$ kernel mátrixok által diktált hasonlóságokat kényszeríti a látens reprezentációkra, azaz $\boldsymbol{U}$ és $\boldsymbol{V}$ oszlopaira.

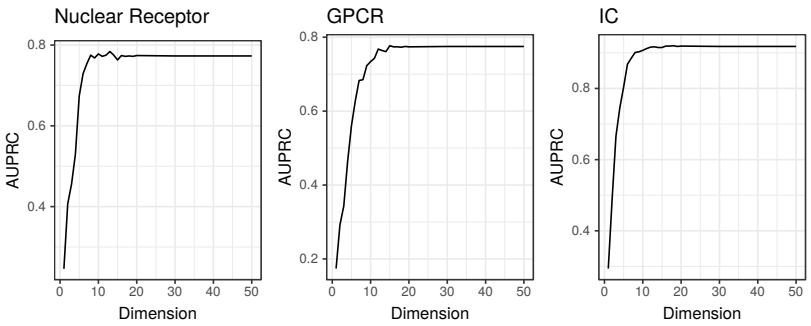
A hiányzó adat-modell *a priori* specifikált intervallumokat használ, amelyek az $\boldsymbol{R}_{ij}$ értékek megfigyelését jelző $\boldsymbol{X} \in \{0,1\}^{I\times J}$ bináris változó Bernoulli-eloszlását paraméterezik:

$$p(\boldsymbol{X}|\boldsymbol{R}, s_1, s_2, \mu) = \prod_i \prod_j f(\boldsymbol{R}_{ij}, s_1, s_2, \mu)^{\boldsymbol{X}_{ij}} (1 - f(\boldsymbol{R}_{ij}, s_1, s_2, \mu))^{1-\boldsymbol{X}_{ij}},$$

ahol az f „dudorfüggvény” definíciója

$$f(x, s_1, s_2, \mu) = \begin{cases} 1 & \text{if } |x-\mu| < s_1, \\ 0 & \text{if } |x-\mu| \geq s_2, \\ \sigma\left(-\frac{s_1^2+s_2^2-2(x-\mu)^2}{((x-\mu)^2-s_1^2)\cdot((x-\mu)^2-s_2^2)}\right) & \text{otherwise,} \end{cases}$$

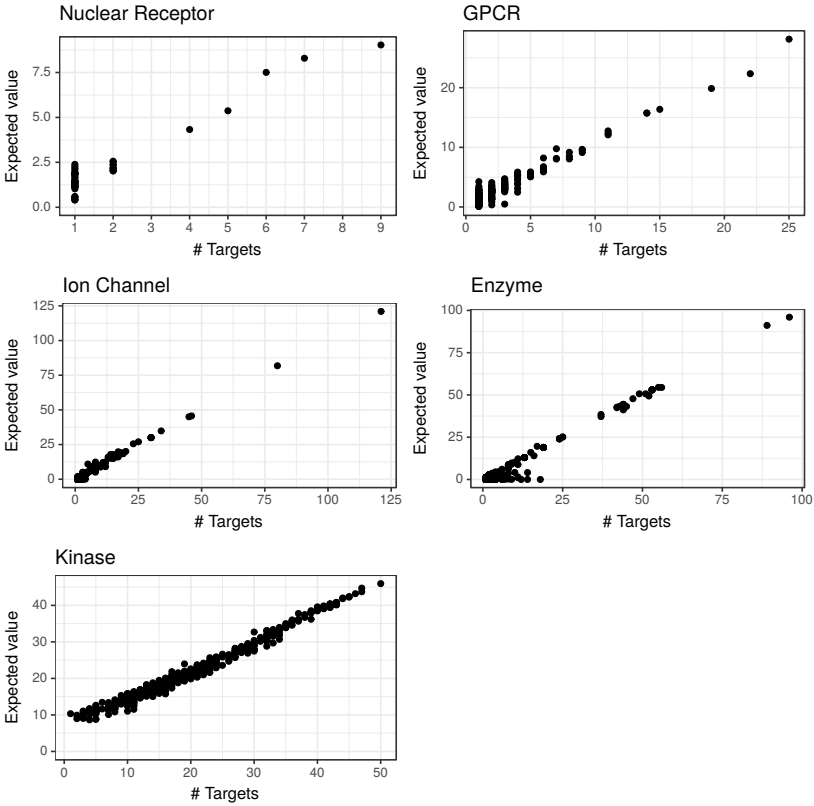
ahol μ az átlagérték, s_1, s_2 pedig az „átmeneti” régiók szélességét szabályozza. Ha $\boldsymbol{R}_{ij}$ kívül esik f tartóján, 1 valószínűséggel hiányzónak tekintjük. Ez lehetőséget nyújt arra, hogy $\boldsymbol{R}_{ij}$ ismeretlen értékeit megfelelő tartományokba kényszerítsük és kihasználjuk a hiányzó értékek által hordozott információt.



6. ábra. AUPRC értékek a benchmark adathalmazokon a látens faktorok számának növelésével. Az értékek 10 dimenzió után telítődnek, ami utal ezen dimenzionalitás elégségességére a látens tulajdonságok megragadására.

A konjugált priorok választásának köszönhetően a feltételes eloszlások egy része analitikusan számolható és a következtetés Gibbs-mintavételezéssel egyértelmű; nem triviális viszont faktorok, a kernel súlyok és $\boldsymbol{R}$ hiányzó értékeinek mintavételezése. A Gauss-folyamat priorok használatával az előbbi két esetben csak kisebb módosítások szükségesek a feltételes eloszlások szokásos levezetéséhez képest, ám a hiányzó adat-modell nem konjugált, így a hiányzó értékek mintavételezésére egy slice sampling alapú sémát adunk a Gibbs-mintavételbe ágyazva.

A modellek prediktív teljesítményét Yamanishi és mtsainak benchmark adatbázisán [Yam+08] demonstráljuk. Megmutatjuk, hogy mindkét modell szignifikánsan felülmúlja a korábbi egykerneles és többkerneles módszerek teljesítményét keresztkiértékeléses vizsgálatban. Igazoljuk, hogy korábbi eredményekkel ellentétben a látens reprezentációk szükséges dimenzionalitása meglepően alacsony ($L \sim 10$), ami kompatibilis egy kötőhely-alapú interpretációval (6. ábra). A priorok prediktív teljesítményre vonatkozó, az adatmérettel növekedésével eltűnő hatásának vizsgálatával kvantifikáljuk a háttérinformáció beépítésének előnyeit, és rámutatunk „kis mintaméret”-régiók létezésére, ahol a priorok használata megalapozott. Bizonyítjuk továbbá, hogy a modell valószínűségi kimenete felhasználható az interakciók várható számának megbecsülésére az egyes gyógyszerek esetében, amely a gyógyszer-promiszkuitás jóslását teszi elérhetővé (7. ábra).



7. ábra. Gyógyszer-promiszkuitás jóslása a célpontok várható számával minden gyógyszerre a benchmark adathalmazokban, amely erős korrelációt mutat az interakciók valódi számával.

Az esetalapú, prospektív kiértékelés során megvizsgáljuk a korábbi adatok alapján legmagasabb valószínűséggel jósolt interakciók jelenlétét a DrugBank adatbázis friss verziójában. Igazoljuk, hogy a jósolt interakciók legnagyobb része valóban szerepel az adatbázisban.

Végül megvizsgáljuk a bináris mátrixfaktorizációs algoritmus számítási hatékonyságát, különös tekintettel a GPU-alapú implementációra. Megmutatjuk, hogy egy 200×200 mátrixfaktorizációs feladatban két nagyságrendnyi gyorsulást jelent a CPU-alapú implementációhoz képest.

III.1. tézis. Kidolgoztuk egy korábbi logisztikus mátrixfaktorizációs metodológia bayesi többkerneles adaptációját, amely egyesíti a többkerneles tanulás, súlyozott megfigyelések, Laplace-típusú regularizáció, valamint a bináris gyógyszer–célpont interakciós valószínűségek explicit modellezésének előnyeit.

III.2. tézis. Kiterjesztettük a bayesi mátrixfaktorizációs metodológiát szakterület-specifikus, tudásintenzív alkalmazások felé, lehetőséget nyújtva az entitásszintű és interakció-szintű háttértudás integrációjára Gauss-folyamat priorok segítségével, valamint kidolgoztunk egy explicit nem-véletlenszerű hiányzási modellt.

III.3. tézis A hatékony következtetésre variációs approximáció és Gibbs-mintavételezés-alapú implementációkat fejlesztettünk, valamint megmutattuk, hogy mindkét modell felülmúlja korábbi módszerek prediktív teljesítményét standard benchmark feladatokban

III.4. tézis Megmutattuk, hogy az algoritmusok képesek farmakológiai szempontból releváns tulajdonságok, például gyógyszer-promiszkuitás jóslására.

A tézisek a következő publikációkon alapulnak: [4], [6].

3. Publikációs lista

3.1. A tézisekhez kapcsolódó publikációk

Folyóiratcikk

[1] B. Bolgár, Á. Arany, G. Temesi, B. Balogh, P. Antal, and P. Mátyus. In: *Current Topics in Medicinal Chemistry* 13.18 (2013), pp. 2337–2363. DOI: 10.2174/15680266113136660164

[2] A. Arany, B. Bolgar, B. Balogh, P. Antal, and P. Matyus. "Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources". In: *Curr. Med. Chem.* 20.1 (2013), pp. 95–107

[3] G. Temesi, B. Bolgár, Á. Arany, C. Szalai, P. Antal, and P. Mátyus. "Early repositioning through compound set enrichment analysis: a knowledge-recycling strategy". In: *Future medicinal chemistry* 6.5 (2014), pp. 563–575

[4] B. Bolgár and P. Antal. "VB-MK-LMF: Fusion of drugs, targets and interactions using Variational Bayesian Multiple Kernel Logistic Matrix Factorization". In: *BMC Bioinformatics* (2017, in press)

Nemzetközi konferencia

[5] B. Bolgár and P. Antal. "Towards Multipurpose Drug Repositioning: Fusion of Multiple Kernels and Partial Equivalence Relations Using GPU-accelerated Metric Learning". In: *First European Biomedical Engineering Conference for Young Investigators: ENCY2015, Budapest, May 28 - 30, 2015*. Ed. by Á. Jobbágy. Singapore: Springer Singapore, 2015, pp. 36–39. DOI: 10.1007/978-981-287-573-0_9. URL: https://doi.org/10.1007/978-981-287-573-0_9

[6] B. Bolgár and P. Antal. "Bayesian Matrix Factorization with Non-Random Missing Data using Informative Gaussian Process Priors and Soft Evidences". In: *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*. Ed. by A. Antonucci, G. Corani, and C. P. Campos. 2016, pp. 25–36

3.2. További publikációk

Folyóiratcikk

[7] A. Gezsi, B. Bolgar, P. Marx, P. Sarkozy, C. Szalai, and P. Antal. "VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering". In: *BMC Genomics* 16 (Oct. 2015), p. 875

[8] P. Marx, P. Antal, B. Bolgar, G. Bagdy, B. Deakin, and G. Juhasz. "Comorbidities in the diseasome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression". In: *PLoS Comput. Biol.* 13.6 (June 2017), e1005487

Fejezet szerkesztett könyvben

[9] G. Hullam, A. Gezsi, A. Millinghoffer, P. Sarkozy, B. Bolgar, S. K. Srivastava, Z. Pal, E. I. Buzas, and P. Antal. "Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis". In: *Methods Mol. Biol.* 1142 (2014), pp. 143–176

[10] B. Bolgár. "Network biology". In: *Bioinformatics*. Ed. by P. Antal. Budapest: Typotex Kiadó, 2014, pp. 129–140

[11] B. Bolgár. "Analysis of heterogeneous biomedical data through information fusion". In: *Bioinformatics*. Ed. by P. Antal. Budapest: Typotex Kiadó, 2014, pp. 206–216

[12] B. Bolgár. "Text mining methods in bioinformatics". In: *Bioinformatics*. Ed. by P. Antal. Budapest: Typotex Kiadó, 2014, pp. 168–181

Nemzetközi konferencia

[13] P. Marx, B. Bolgár, A. Gezsi, A. Gulyás-Kovács, and P. Antal. "MicroRNA Prioritization based on Target Profile Similarities". In: *BIOINFORMATICS*. Angers, France, 2014, pp. 278–285

Helyi konferencia

[14] B. Bolgár and P. Antal. "QDF2: A Kernel-based Data Fusion Framework for Drug Repositioning". In: *Proceedings of the 20th PhD Minisymposium.* Ed. by B. Pataki. Budapest, Hungary, Feb. 2013, pp. 24–25

[15] B. Bolgár and P. Antal. "Multiple Kernel Learning of Distance Metrics with Pairwise Equivalence and Inequivalence Constraints". In: *Proceedings of the 21st PhD Minisymposium.* Ed. by B. Pataki. Budapest, Hungary, Mar. 2014, pp. 12–15

References

[Agr+07] D. K. Agrafiotis et al. "Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world". In: *Journal of Chemical Information and Modeling* 47.6 (2007), pp. 1999–2014. DOI: 10.1021/ci700267w.

[Ara+13] A. Arany, B. Bolgar, B. Balogh, P. Antal, and P. Matyus. "Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources". In: *Curr. Med. Chem.* 20.1 (2013), pp. 95–107.

[Bur+01] R. Burbidge, M. Trotter, B. Buxton, and S. Holden. "Drug design by machine learning: support vector machines for pharmaceutical data analysis". In: *Computers & chemistry* 26.1 (2001), pp. 5–14.

[Dav+07] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. "Information-theoretic Metric Learning". In: *Proceedings of the 24th International Conference on Machine Learning.* ICML '07. 2007, pp. 209–216. DOI: 10.1145/1273496.1273523. URL: http://doi.acm.org/10.1145/1273496.1273523.

[DGL97] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* corrected 2nd. Vol. 31. Applications of Mathematics. missing. Springer, 1997.

[Gez+15] A. Gezsi, B. Bolgar, P. Marx, P. Sarkozy, C. Szalai, and P. Antal. "VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering". In: *BMC Genomics* 16 (2015), p. 875.

[GKK13] M. Gönen, S. Khan, and S. Kaski. "Kernelized Bayesian matrix factorization". In: *International Conference on Machine Learning*. 2013, pp. 864–872.

[HHG14] J. M. Hernandez-Lobato, N. Houlsby, and Z. Ghahramani. "Stochastic Inference for Scalable Probabilistic Modeling of Binary Matrices". In: *Proceedings of the 31st International Conference on Machine Learning (ICML)* 32 (2014), pp. 1–6.

[Hul+14] G. Hullam, A. Gezsi, A. Millinghoffer, P. Sarkozy, B. Bolgar, S. K. Srivastava, Z. Pal, E. I. Buzas, and P. Antal. "Bayesian systems-based genetic association analysis with effect strength estimation and omic wide interpretation: a case study in rheumatoid arthritis". In: *Methods Mol. Biol.* 1142 (2014), pp. 143–176.

[JJ00] T. S. Jaakkola and M. I. Jordan. "Bayesian parameter estimation via variational methods". In: *Statistics and Computing* 10.1 (2000), pp. 25–37. DOI: 10.1023/A:1008932416310. URL: http://dx.doi.org/10.1023/A:1008932416310.

[Kit+04] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. "Docking and scoring in virtual screening for drug discovery: methods and applications". In: *Nature reviews Drug discovery* 3.11 (2004), pp. 935–949.

[Köv+99] I. Kövesdi, M. F. Dominguez-Rodriguez, L. Ôrfi, G. Náray-Szabó, A. Varró, J. G. Papp, and P. Matyus. "Application of neural networks in structure–activity relationships". In: *Medicinal research reviews* 19.3 (1999), pp. 249–269.

[Liu+16] Y. Liu, M. Wu, C. Miao, P. Zhao, and X. L. Li. "Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction". In: *PLoS Computational Biology* 12.2 (2016), pp. 1–26. DOI: 10.1371/journal.pcbi.1004760. URL: http://dx.doi.org/10.1371/journal.pcbi.1004760.

[Mar+17] P. Marx, P. Antal, B. Bolgar, G. Bagdy, B. Deakin, and G. Juhasz. “Comorbidities in the diseasome are more apparent than real: What Bayesian filtering reveals about the comorbidities of depression”. In: *PLoS Comput. Biol.* 13.6 (2017), e1005487.

[MT12] Y. Moreau and L.-C. Tranchevent. “Computational tools for prioritizing candidate genes: boosting disease gene discovery”. In: *Nature Reviews Genetics* 13.8 (2012), pp. 523–536.

[Pav+02] P. Pavlidis, J. Weston, J. Cai, and W. S. Noble. “Learning gene functional classifications from multiple data types”. In: *J. Comput. Biol.* 9.2 (2002), pp. 401–411.

[SM08a] R. Salakhutdinov and A. Mnih. “Bayesian Probabilistic Matrix Factorization Using Markov Chain Monte Carlo”. In: *Proceedings of the 25th International Conference on Machine Learning*. ICML ’08. 2008, pp. 880–887.

[SM08b] R. Salakhutdinov and A. Mnih. “Probabilistic Matrix Factorization”. In: *Advances in Neural Information Processing Systems*. Vol. 20. 2008.

[SS01] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[Tob09] E. L. Tobinick. “The value of drug repositioning in the current pharmaceutical market”. In: *Drug News Perspect.* 22.2 (2009), pp. 119–125.

[Vis+10] S. V. N. Vishwanathan, Z. sun, N. Ampornpunt, and M. Varma. “Multiple Kernel Learning and the SMO Algorithm.” In: *NIPS*. 2010, pp. 2361–2369. URL: http://dblp.uni-trier.de/db/conf/nips/nips2010.html#VishwanathansAV10.

[Wan+11] S. Wang, Q. Huang, S. Jiang, and Q. Tian. “Efficient Lp-norm Multiple Feature Metric Learning for Image Categorization”. In: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. CIKM ’11. 2011, pp. 2077–2080. DOI: 10.1145/2063576.2063894. URL: http://doi.acm.org/10.1145/2063576.2063894.

[WBD98] P. Willett, J. M. Barnard, and G. M. Downs. “Chemical similarity searching”. In: *Journal of chemical information and computer sciences* 38.6 (1998), pp. 983–996.

[Wil+12] A. J. Williams et al. “Open PHACTS: semantic interoperability for drug discovery”. In: *Drug Discov. Today* 17.21-22 (2012), pp. 1188–1198.

[WS09] K. Q. Weinberger and L. K. Saul. “Distance Metric Learning for Large Margin Nearest Neighbor Classification”. In: *J. Mach. Learn. Res.* 10 (2009), pp. 207–244. URL: http://dl.acm.org/citation.cfm?id=1577069.1577078.

[WSN07] C. L. Waller, A. Shah, and M. Nolte. “Strategies to support drug discovery through integration of systems and data”. In: *Drug discovery today* 12.15 (2007), pp. 634–639.

[Yam+08] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. “Prediction of drug-target interaction networks from the integration of chemical and genomic spaces”. In: *Bioinformatics* 24.13 (2008), pp. i232–240.

[Yu+10] S. Yu, T. Falck, A. Daemen, L. C. Tranchevent, J. A. Suykens, B. De Moor, and Y. Moreau. “L2-norm multiple kernel learning and its application to biomedical data fusion”. In: *BMC Bioinformatics* 11 (2010), p. 309.

[Zhe+13] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu. “Collaborative matrix factorization with multiple similarities for predicting drug-target interactions”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’13* (2013), p. 1025. DOI: 10.1145/2487575.2487670. URL: http://dl.acm.org/citation.cfm?id=2487670%7B%%7D5Cnhttp://dl.acm.org/citation.cfm?doid=2487575.2487670.