Ovrednotenje novega na podatkovni gravitaciji temelječega klasifikatorja SDGC

Aljaž Heričko

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova ulica 17, 2000 Maribor
E-pošta: aljaz.hericko@student.um.si

Evaluation of a new data gravitation based classifier

Classification methods and techniques are used in various domains for categorizing objects into classes. The class of an object is determined depending on the attributes of the object and the relations between the uncategorized object and the objects for which classes are known. In the article we are proposing a new classification algorithm and a corresponding classifier named Simple Data Gravitation based Classifier (SDGC). The efficiency of the proposed classifier was evaluated using six standard data sets and compared to the results of selected well-known classification algorithms. Preliminary results of the conducted experimental study illustrate that the proposed classifier gives some promising results. Consequently, it makes sense to continue with research work in the area of gravity-based classifiers.

1 Uvod

Zaradi vse večje razširjenosti in intenzivnosti uporabe informacijskih tehnologij ter vseprisotnih senzorjev se srečujemo z vedno večjimi in obsežnejšimi zbirkami podatkov. Z analizo teh podatkov lahko generiramo in odkrivamo nova znanja ter pridobimo več informacij za odločanje, kot jih zagotavljajo posamezni nepovezani podatki. Prav zato je eden ključnih izzivov učinkovito iskanje povezav med objekti in njihovimi atributi ter uspešno združevanje ter klasificiranje sorodnih objektov v razrede.

Raziskovalno področje, ki naslavlja to problematiko, je podatkovno rudarjenje. Obstaja več tehnik in algoritmov obdelave podatkov, npr. klasifikacijski, regresijski, segmentacijski, asociacijski ipd., ki jih uporabljamo v odvisnosti od potreb in tipov problemov, ki jih rešujemo. V članku se bomo osredotočili na klasificiranje in predstavili nov klasifikator, imenovan SDGC.

Predlagani klasifikator SDGC uporablja enega izmed osnovnih principov, ki ga zasledimo v naravi - gravitacijo. Gravitacija predstavlja silo privlačnosti med dvema telesoma, ki imata določeno maso. Koncept sile gravitacije smo uporabili pri razvoju klasifikatorja z uporabo dejstva, da če določimo maso različnim množicam podatkov in te množice privlačijo nove elemente z različno veliko silo, lahko element uvrstimo v množico, ki element privlači z največjo silo. To silo poimenujemo podatkovna gravitacija (angl. data gravitation). Sama ideja o podatkovni gravitaciji ni nova, obstaja pa več različnih možnosti implementacije. V članku bomo uporabili eno izmed doslej še neuporabljenih možnosti.

Osnovni namen članka je torej predstaviti novo tehniko klasifikacije podatkov na osnovi gravitacije ter podati in analizirati rezultate preliminarne analize točnosti predlaganega klasifikatorja v primerjavi z nekaterimi drugimi, že uveljavljenimi in pogosto uporabljenimi tehnikami, kot so npr. Bayesov klasifikator, odločitvena drevesa in K najbližjih sosedov.

V drugem poglavju bomo predstavili osnovne definicije glede klasifikacije, povzeli osnovne značilnosti različnih tehnik klasifikacije. Predstavili bomo tudi osnovna merila oz. metrike, ki jih bomo uporabili za primerjavo točnosti klasifikacijskih tehnik. V tretjem poglavju bomo predstavili osnovne ideje klasifikacijskih tehnik, ki temeljijo na uporabi koncepta gravitacije.

V četrtem poglavju bomo predstavili koncept novega klasifikatorja na osnovi gravitacije, ki se od dosedanjih razlikuje po tem, da temelji masa na oddaljenosti od središč razredov, kar se nato uporabi za izračun gravitacije novih primerkov. V petem poglavju predstavimo rezultate vrednotenja uspešnosti predlagane klasifikacijske tehnike in sicer na šestih testnih množicah.

V zaključku povzamemo ključne ugotovitve ter podamo smernice za nadaljnje delo ter smiselnost uporabe klasifikatorjev na osnovi gravitacije.

2 Klasifikacijske tehnike in kriteriji vrednotenja

Namen klasifikacije je razvrstitev objektov v vnaprej definirane razrede. Razred določimo na podlagi atributov tega objekta ter povezav med atributi neuvrščenega objekta in atributi objektov, ki že imajo določen razred. Postopek klasifikacije poteka v dveh korakih. V prvem se iz učne množice (angl. training set) zgradi klasifikator, kar imenujemo učenje. Tako zgrajen klasifikator, pridobljen na osnovi že razvrščenih podatkov, nato v drugem koraku uporabimo za klasifikacijo primerkov iz testne množice (angl. test set) v ustrezne razrede.

Tehnike klasificiranja delimo na več skupin in sicer glede na njihovo osnovno idejo o razvrščanju primerkov. Med najbolj uporabljenimi in priznanimi

klasifikacijskimi tehnikami so npr. Bayesov klasifikator, odločitvena drevesa in K najbližjih sosedov. Bayesov klasifikator izhaja iz Bayesovega teorema, kjer se pripadnost izračuna z uporabo pogojnih verjetnosti. Najbolj znan primer Bayesovega klasifikatorja je naivni Bayes. Klasifikacijske tehnike z odločitvenimi drevesi oblikujejo drevesno strukturo iz vozlišč, skozi katere potuje neuvrščeni primerek, dokler ne pride do lista, ki določa razred objekta. Med bolj znanimi klasifikatorji, ki temeljijo na konceptu odločitvenih dreves, so ID3, C4.5 in CART. V eksperimentalnem delu smo uporabili implementacijo C4.5 imenovano J48, ki je eden izmed klasifikatorjev, vključenih v orodje WEKA [1]. Klasifikacijske tehnike K najbližjih sosedov določijo pripadnost razredu tako, da ustvarijo skupino vnaprej določene velikosti, v katere uvrstimo primerke oz. objekte iz učne množice, ki so najbližje novemu primerku. Nov primerek uvrstimo v razred, kateremu pripada najvišje število primerkov iz te skupine. Pri primerjavi rezultatov smo uporabili eno izmed implementacij, ki je podprta v orodju WEKA, in sicer klasifikator IBk.

Pri klasifikaciji lahko uspešnost in rezultate klasificiranja neznanih primerkov analiziramo in medsebojno primerjamo z uporabo različnih meril. Pri svojem delu smo uporabili dve merili: točnost klasifikacije in koeficient Cohen's kappa. Točnost (angl. accuracy) klasifikacije predstavlja delež pravilno klasificiranih primerkov glede na vse klasificirane primerke. Koeficient Cohen's kappa pa predstavlja natančnejše merilo za robustnost kot točnost klasifikatorja, saj se za izračun uporabi formula, ki vsebuje tako število uspešno kakor tudi število neuspešno klasificiranih primerkov. Za obe izbrani merili velja, da višji rezultat pomeni boljše delovanje klasifikatorja.

3 Klasifikacijske tehnike na osnovi podatkovne gravitacije

Gravitacija je eden izmed osnovnih naravnih konceptov, katere pravila in posledice je možno uporabiti na področju klasificiranja. Newton je že v 17. stoletju objavil delo, v katerem je zapisal, da med poljubnima objektoma obstaja določena sila[2]. Zakon je poimenoval gravitacijski zakon in zapisal formulo (1), po kateri se lahko izračuna sila F med tema dvema objektoma, ki sta na razdalji d in imata maso m_1 oz. m_2:

$$F = G\frac{m_1 * m_2}{d^2} \quad (1)$$

Klasifikatorji, ki uporabijo to formulo, in uporabljajo idejo gravitacije, delujejo tako, da primerjajo gravitacijo novega primerka (objekta, ki ga je potrebno razvrstiti) do razredov in ga klasificirajo v razred, do katerega ima primerek največjo gravitacijo. Ker je možnosti, na kakšen način izračunati gravitacijo, več, se v različnih virih pojavljajo različne ideje klasifikatorjev [3,4,5,6].

Eden izmed prvih člankov na področju klasificiranja z uporabo podatkovne gravitacije opisuje primer, kjer se gravitacija računa z uporabo središč razredov [3]. V tem delu so tudi definirani osnovni pojmi v povezavi med gravitacijo in klasifikacijo, ki so uporabljeni tudi v tem članku. Pristop, opisan v [3], deluje tako, da iz primerkov razreda sestavi množice ter jim določi maso glede na število primerkov v množici. Pripadnost klasifikatorju nato določi tako, da izračuna gravitacijo novega primerka do vseh množic vsakega razreda in ga uvrsti v razred, kjer je vsota gravitacij največja.

Kasnejša dela so idejo uporabe gravitacije nadalje razvijali z vpeljavo raznih uteži in novih algoritmov izračuna gravitacije. Ena izmed tehnik izračuna gravitacije je povezava klasifikatorja s podatkovno gravitacijo s klasifikacijsko tehniko K najbližjih sosedov [4]. Sodeč po rezultatih je algoritem predvsem primeren za zbirke, kjer je manjše število podatkov, hkrati pa so lahko razredi neuravnoteženi in je klasifikacija zahtevnejša, vendar dobljeni rezultati kažejo na uspešnost opisane tehnike v takšnih primerih.

Klasifikacijo podatkov s primerjavo gravitacije med razredi so raziskovali tudi [5], ki so se osredotočili na primerjavo klasifikacije standardnih in neuravnoteženih množic podatkov. Problematiko izračuna gravitacije, ki jo pogojuje različna relevanca posameznih atributov, so naslovili tako, da pri izračunu razdalj uporabljajo matriko uteži za opis pomembnosti posameznega atributa pri klasifikaciji vsakega razreda testnega vzorca. Učinkovitost klasifikacije so izboljšali z uporabo lokalnih in globalnih informacij o podatkih, še posebej pri mejnih odločitvah. Primerjava z drugimi uveljavljenimi klasifikacijskimi tehnikami je vključevala analizo s sedmimi algoritmi pri standardnih množicah ter osmimi algoritmi pri neuravnoteženih množicah. Primerjava točnosti klasifikacije ter koeficienta Cohen's kappa sta pokazala odlične rezultate predlaganega algoritma DGC+ na uporabljenih množicah.

Rezultati raziskav torej kažejo na primerljivost in v določenih primerih superiornost tehnik, ki temeljijo na podatkovni gravitaciji, vsekakor pa je področje klasifikatorjev na osnovi podatkovne gravitacije eno najaktualnejših področij [6, 7] ter so smiselne dodatne raziskave ter podrobne in celovite analize uspešnost in možnosti uporabe teh klasifikatorjev v različnih domenah ter na zbirkah podatkov z različnimi značilnostmi.

4 Predlog novega klasifikatorja SDGC

Pri razvoju lastnega klasifikatorja na osnovi gravitacije smo kot osnovo uporabili formulo za gravitacijo (1) in nato pridobili vse potrebne podatke ter z njimi izračunali pripadnost novega primerka določenemu razredu po postopku, ki je natančneje opisan v naslednjih odstavkih.

Kot je razvidno iz enačbe (1), potrebujemo za izračun gravitacije razdaljo med objektoma in maso obeh objektov, katerih gravitacijo računamo, ter gravitacijsko konstanto G. Razdaljo med primerkom iz testne in učne

množice izračunamo z uporabo Evklidove formule za izračun razdalje med dvema točkama. Točna enačba je zapisana pod številko (2), pri čemer je x_{1i} vrednost atributa številka i v objektu iz učne množice, x_{2i} pa vrednost atributa i objekta, ki ga je potrebno razvrstiti.

$$d = \sqrt{\sum_1^n (x_{1i} - x_{2i})^2} \quad (2)$$

Maso posameznega primerka iz učne množice izračunamo z uporabo enačbe (3). V enačbi predstavlja *d* oddaljenost primerka do središča njegovega razreda, d_{max} pa predstavlja največjo oddaljenost primerka znotraj istega razreda do središča tega razred. Torej dobimo za maso *m* vrednost med 1 in 2. Večja kot je oddaljenost primerka od središča razreda, manjšo maso ima.

$$m = 2 - \frac{d}{d_{max}} \quad (3)$$

Pri določanju mase za primerek iz testne množice, kateremu določamo pripadnost razredu, maso izračunamo z uporabo identične enačbe kot za maso primerka iz učne množice, le da za d_{max} namesto maksimalne oddaljenosti enega od primerkov razreda do središča razreda uporabimo največjo izmed razdalj med novim primerkom in središč razredov. Spremenljivka *d* predstavlja oddaljenost do središča posameznega razreda, saj je masa neuvrščenega primerka različna za vsak razred.

Ko pridobimo vse potrebne podatke, izračunamo gravitacijo novega primerka z vsakim posameznim primerkom iz učne množice po formuli (4).

$$F = \frac{m_1 * m_2}{d^2} \quad (4)$$

Formula se od osnovne formule za gravitacijo razlikuje le v tem, da pri uporabi enačbe za potrebe klasificiranja izpustimo konstanto *G*, saj ta nima vpliva na končen rezultat klasifikacije. Konstanta G bi le sorazmerno enako povečala vrednost gravitacije vsakega primerka in bi tako zgolj podaljšala čas izvajanja.

Za končno določitev razreda izračunamo vsoto vseh gravitacijskih sil do primerkov učne množice za vsak razred posebej ter vsoto gravitacij posameznega razreda delimo s številom primerkov razreda. Iz tega za vsak razred izračunamo povprečno gravitacijo vsakega izmed primerkov. Z uporabo povprečja se izognemo problemom zaradi neuravnoteženih množic. V zadnjem koraku primerek iz testne množice klasificiramo v razred, kjer je njegova povprečna gravitacija do primerkov razreda najvišja.

Naš klasifikator oz. klasifikacijska funkcija se razlikuje od predlogov ostalih avtorjev po tem, da pri izračunu mase, ki je uporabljena v izračunu gravitacije, upošteva središča posameznih razredov, vendar kljub izračunu središč izračuna gravitacijo do vsakega posameznega primerka razreda. S tem dosežemo, da ima vsak izmed primerkov vpliv na končni rezultat, hkrati pa imajo primerki, ki so bližje središču razreda, večjo maso in posledično večji vpliv na končni rezultat. Prav tako se od drugih loči zaradi svoje preprostosti, saj so uporabljene formule enostavnejše od formul ostalih avtorjev, zato smo klasifikator poimenovali preprost klasifikator na osnovi podatkovne gravitacije (angl. Simple Data Gravity Classifier) oz. krajše SDGC.

5 Rezultati vrednotenja klasifikatorja SDGC

Opisano klasifikacijsko tehniko smo implementirali v orodju WEKA. To ogrodje nudi funkcije za predobdelavo podatkov, izvajanje podatkovnega rudarjenja, vizualni prikaz vhodnih podatkov in avtomatsko analizo rezultatov učenja. V orodju imamo možnost uporabiti že vključene implementacije najbolj razširjenih in pogostih klasifikacijskih tehnik, kot so npr. naivni Bayes, J48, ki je odprtokodna implementacija odločitvenih dreves algoritma C4.5, in klasifikator IBk, ki implementira klasifikacijsko tehniko K najbližjih sosedov. Te tri algoritme bomo uporabili za primerjavo rezultatov klasifikatorja nastalega iz naše ideje z že obstoječimi algoritmi. Ena izmed novejših raziskav [9] je pokazala, da je orodje WEKA po več kriterijih (zmogljivosti, apliciranje različnih klasifikatorjev) eno najboljših orodij za klasifikacijo.

Uporaba orodja WEKA nam je omogočila sorazmerno enostavno primerjavo točnosti naše klasifikacijske tehnike s prej omenjenimi pristopi. Teste smo izvedli na šestih različnih standardnih podatkovnih množicah in sicer: Ionosphere, Iris, Segment, Vehicle, Diabetes in Sonar. Dostopne so na spletni strani *http://www.ics.uci.edu/~mlearn/* .

Osnovne informacije in značilnosti podatkovnih množic, uporabljenih pri eksperimentalni študiji, so predstavljene v tabeli I. Uporabili smo zgolj množice podatkov, kjer so vrednosti atributov predstavljene s celoštevilsko ali realno vrednostjo.

Tabela I: Uporabljene podatkovne množice

Naziv	Št. primerkov	Št. atributov	Št. razredov
Ionosphere	351	34	2
Iris	150	4	3
Segment	2310	19	7
Vehicle	846	18	4
Diabetes	768	8	2
Sonar	208	60	2

Pri vrednotenju klasifikatorja smo uporabili metodo razdelitve, kjer množico naključno razdelimo na dva dela. V učno množico, iz katere se ustvari klasifikator, smo naključno dodali dve tretjini vseh primerkov podatkovne zbirke. V testno množico smo uvrstili preostalo tretjino. Pri izvajanju testov smo vsak algoritem pognali desetkrat, vsakič nad drugačno učno in testno množico iz istega nabora podatkov. To smo storili za vsako podatkovno zbirko. Povzetek rezultatov je podan v tabelah II in III. V tabeli II je podano povprečje točnosti klasifikatorja v desetih zagonih. V tabeli III so zapisane pripadajoče povprečne vrednosti koeficienta Cohen's kappa v istih desetih zagonih.

Tabela II: Rezultati točnosti klasifikatorjev

	SDGC	NB	J48	IBk
Ionosphere	85,56	82,71	**86,22**	83,94
Iris	**96,20**	95,80	92,60	95,40
Segment	92,14	80,83	95,93	**96,52**
Vehicle	59,68	44,36	**69,43**	68,92
Diabetes	71,17	**74,49**	73,95	69,34
Sonar	76,71	66,47	69,12	**84,56**

Ugotovimo lahko, da je pri vseh šestih podatkovnih zbirkah predlagani klasifikator SDGC dosegel rezultate, ki so razmeroma blizu in v primeru ene podatkovne zbirke celo boljši od ostalih klasifikatorjev. Iz rezultatov je razvidno, da je nihanje točnosti med posameznimi podatkovnimi zbirkami dokaj veliko. Če analiziramo, v katerih primerih pride do odstopanj, ugotovimo, da je klasifikator SDGC trenutno uspešnejši pri preprostejših zbirkah podatkov, kjer je algoritem med bolj točnimi in uspešnimi, medtem ko točnost klasifikacije pade pri zahtevnejših množicah, kjer je število atributov večje.

Tabela III: Rezultati – koeficient Cohen's kappa

	SDGC	NB	J48	IBk
Ionosphere	0,66	0,64	**0,69**	0,62
Iris	**0,94**	**0,94**	0,89	0,93
Segment	0,90	0,78	0,95	**0,96**
Vehicle	0,46	0,26	**0,59**	**0,59**
Diabetes	0,38	**0,42**	0,41	0,32
Sonar	0,53	0,33	0,38	**0,69**

Iz tabele III lahko pridemo do podobnih zaključkov in dodatne potrditve ugotovitve iz prejšnjega odstavka, da je klasifikator primernejši za klasifikacijo enostavnejših zbirk podatkov z manjšim številom atributov. Rezultati so v določeni meri skladni z ostalimi raziskavami, ki so predlagale in primerjale tehnike na osnovi gravitacije [4, 6, 8, 10] z ostalimi klasifikacijskimi tehnikami na osnovi podobnega nabora standardnih zbirk testnih podatkov. Avtorji le teh so namreč podobno ugotavljali, da so rezultati v večini podatkovnih zbirk ugodni ter da obstaja možnost izboljšanja teh rezultatov z optimizacijo uporabljenih tehnik na osnovi podatkovne gravitacije.

6 Sklep

Razvoj novih in učinkovitejših klasifikacijskih tehnik je ključnega pomena za uspešno obdelavo in izkoriščanje potenciala, ki ga v smislu odkrivanja znanja predstavljajo obsežne zbirke podatkov. V prispevku predlagan klasifikator SDGC sodi v kategorijo pristopov, ki temeljijo na matematičnih oz. fizikalnih osnovah gravitacije. Rezultati dosedanjih raziskav na tem področju, vključno z našo, kažejo na možnost doseganja primerljivih rezultatov z že uveljavljenimi in dlje časa uporabljenimi tehnikami, kot sta npr. K najbližjih sosedov in odločitvena drevesa. V preliminarni raziskavi smo se omejili na manjše število zbirk podatkov. Ker gre za sorazmerno majhne zbirke, primerjava časa, potrebnega za učenje in analizo, ne omogoča sprejetja zanesljivih sklepov o učinkovitosti v smislu časa procesiranja.

V nadaljevanju raziskav bomo v okolju WEKA implementirali še algoritme klasifikacijskih tehnik na osnovi gravitacije, ki so jih predlagali drugi avtorji. Sledila bo celovita primerjava učinkovitosti vseh obravnavanih pristopov tako s stališča uspešnosti klasifikacije kot tudi zahtevnosti obdelave. Hkrati bo celovitejša primerjalna študija obsegala tudi primerjavo z različnimi tehnikami testiranja algoritmov, kot je navzkrižna validacija. Tako bomo prispevali k boljšemu razumevanju delovanja in uporabnosti klasifikatorjev na osnovi gravitacije.

Literatura

[1] S. Drazin and M. Montag, "Decision Tree Analysis using Weka," pp. 1–3, 2012.

[2] E. Verlinde, "On the origin of gravity and the laws of Newton," J. High Energy Phys., vol. 2011, no. 4, p. 29, Apr. 2011.

[3] L. Peng, Y. Chen, and B. Yang, "A novel classification method based on data gravitation," Neural Networks Brain, 2005. …, pp. 667–672, 2005.

[4] G. Wen, J. Wei, J. Wang, T. Zhou, and L. Chen, "Cognitive gravitation model for classification on small noisy data," Neurocomputing, vol. 118, pp. 245–252, Oct. 2013.

[5] A. Cano, A. Zafra, and S. Ventura, "Weighted Data Gravitation Classification for Standard and Imbalanced Data," Cybern. IEEE Trans., vol. 43, no. 6, pp. 1672–1687, 2013.

[6] P. Shafigh, S. Y. Hadi, and E. Sohrab, "Gravitation based classification," Inf. Sci. (Ny)., vol. 220, pp. 319–330, Jan. 2013.

[7] G. Tomasz, "A variant of gravitational classification," Bioemtrical Letters, vol. 51, no. 1, pp. 1–12, 2014.

[8] L. Peng, B. Yang, Y. Chen, and a Abraham, "Data gravitation based classification," Inf. Sci. (Ny)., vol. 179, no. 6, pp. 809–819, Mar. 2009.

[9] A. H. Wahbeh, Q. A. Al-radaideh, M. N. Al-kabi, and E. M. Al-shawakfa, "A Comparison Study between Data Mining Tools over some Classification Methods," IJACSA, vol. 2 (8), no. Special Issue on Artificial Intelligence, pp. 18–26, 2010.

[10] L. Junlin, "Data classification based on supporting data gravity," 2009 IEEE Int. Conf. Intell. Comput. Intell. Syst., pp. 22–28, Nov. 2009.