ARSYETIMI BAZUAR NË KUJTESË NË PËRZGJEDHJEN E PUNONJËSVE TË RINJ

KRESHNIK VUKATANA[1], NEVILA BACI[1], REXHEP RADA[2]
[1]Universiteti i Tiranës, Fakulteti i Ekonomisë, Departamenti i Statistikës dhe Informatikës së Zbatuar
[2]Universiteti i Elbasanit, Fakulteti i Shkencave të Natyrës, Departamenti i Informatikës
e-mail: kreshnik.vukatana@unitir.edu.al

Përmbledhje

Modeli me arsyetim të bazuar në kujtesë është pjesë e atyre modeleve që përdoren në analizën parashikuese, në gërmimin e të dhënave. Këto modele përdorin formula të cilat përpunojnë rekorde historike të të dhënave, duke evidentuar rastet e suksesit dhe ato të dështimit dhe në bazë të tyre bëjnë parashikime mbi rekordet e reja që futen në model. Fushat e aplikimit të këtyre modeleve janë të gjera si për shembull zbulimi i mashtrimeve, klasifikimi i klientëve, trajtimet mjekësore, kreditë konsumatore, etj. Rasti studimor që trajtohet në këtë punim analizon parashikimin në përzgjedhjen e punonjësve të rinj, duke u bazuar në të dhënat historike të punonjësve ekzistues të një kompanie. Atributet që merren në konsideratë janë: mosha, diplomimi, shteti, drejtimi dhe përvoja. Ndërsa parashikimi merr vlerat: dobët, minimal, përshtatshëm dhe shkëlqyeshëm. Në këtë punim tregohet sesi mund të aplikohet modeli me arsyetim të bazuar në kujtesë, për të pasur një vendimmarrje për një kandidat të ri, duke parashikuar nga të dhënat historike të punonjësve ekzistues të ngjashëm me karakteristikat e kandidatit të ri.

Fjalëkyçe: Nxjerrja e të dhënave, arsyetimi bazuar në kujtesë, fqinji-K më i afërt, vendimmarrje.

Abstract

The memory-based reasoning model is part of the models used in predictive analysis, in data mining. These models use formulas that process historical records of data, identifying cases of success and failure, and based on them make predictions about new records that are inserted into the model. The areas of application of those models are wide, such as fraud detection, customer classification, medical treatments, consumer loans, etc. The case study presented in this paper analyzes the forecast in the selection process of new employees, based on the historical data of existing employees in the company. The attributes that are taken into consideration are age, degree, state, direction, and experience. While the expected forecast values for the given employee are poor, minimal, convenient, and excellent. This paper shows how the memory-based reasoning model can be applied to make a decision to hire a new candidate in the company, predicting from existing employee historical records, similar to the new candidate's characteristics.

Key words: Data mining, memory-based reasoning, k-nearest neighbors algorithm, decision-making.

Hyrje

Analiza parashikuese në gërmimin e të dhënave mundëson zhvillimin e modeleve matematikore duke u mbështetur në formula që krahasojnë rastet e suksesit dhe dështimeve në rekordet historike. Këto formula përdoren më tej për të parashikuar rezultatet për rekordet e reja. Nga teknikat më të përdorura në gërmimin e të dhënave për analizën parashikuese mund të përmendim: modelin e regresionit, atë të zgjedhjes, modelin me induksion, modelin me grupim, rrjetat nervore, arsyetimin e bazuar në kujtesë dhe pemët e vendimit. Teknikat e modelit me arsyetim të bazuar në kujtesë (ABK[1]), që do të trajtohen në këtë punim, konsistojnë në gjetjen e zgjidhjes së problemave të reja, bazuar në mënyrën se si janë zgjidhur situatat e mëparshme. Në këtë mënyrë, kur duhet të parashikojmë sjelljen e një rekordi të ri, konsultohet baza e të dhënave historike. ABK-ja, në ndryshim me teknikat e tjera të nxjerrjes së të dhënave, nuk kujdeset për formatin e regjistrimeve për sa kohë që përcaktohen dy veprimet: funksioni i distancës dhe funksioni i kombinimit (Olson & Delen, 2008).

Funksioni i distancës mund të njehsohet si një distancë midis dy të dhënave dhe funksioni i kombinimit përdoret për të bashkuar rezultatet nga fqinjët e shumtë, në mënyrë që të shfaqë një parashikim. Cilësia e këtij të fundit varet nga gjetja e distancës dhe kombinimit më të mirë (Park & Han, 2002). Këto funksione janë të përcaktuara për disa lloje rekordesh të tilla si rekorde me tipe komplekse e të pazakonta të të dhënave, duke përfshirë të dhëna si zonat gjeografike, imazhet apo tekst i pa strukturuar, të cilat në përgjithësi janë komplekse për t'u menaxhuar me teknika të tjera. Një avantazh tjetër i ABK-së është se ai përshtatet lehtësisht duke i marrë menjëherë për njehsim të dhënat e reja të ngarkuara në model. Ai prodhon rezultate të mira pa trajnim të gjatë dhe pa nevojën e ndryshimit të algoritmit. Këto përfitime vijnë me një kosto. ABK-ja mund të kthehet në një proces që konsumon një sasi të madhe burimesh pa finalizuar në një produkt domethënës. Kjo vjen nga karakteristika themelore e saj, që një sasi e madhe e të dhënave historike duhet të jetë lehtësisht e disponueshme për zbulimin e fqinjëve më të afërt me rekordin e ri. Pra, ABK-ja kërkon shumë të dhëna historike për të dhënë një parashikim sa më të besueshëm (Lan & Neagu, 2007).

Të dhënat e pakta mund të çojnë në një gabim të madh parashikimi sepse nuk ka rekorde të mjaftueshme që përputhen me modelin. Të gjitha këto të dhëna duhet të ruhen dhe ky proces duhet të merret parasysh kur punojmë me këto teknika. Klasifikimi i të dhënave të reja mund të kërkojë përpunimin e të gjitha të dhënave historike për të zbuluar fqinjët më të ngjashëm - një proces më i ngadalshëm sesa përdorimi i një rrjeti nervor tashmë të trajnuar ose një pemë vendimi. Ekziston edhe vështirësia e zbulimit të funksioneve të duhura, të distancës së mirë apo kombinimit, të cilat kanë nevojë për t'u testuar për gabime dhe parashikime të sakta. Nëse kemi vetëm një klasifikimin, atëherë duhet të

përpunohen të gjitha të dhënat historike, nëse ka më shumë se një, ato ndahen në nëngrupe bazuar në klasifikimet përkatëse (Stanfill & Waltz, 1986).

Teknikat e ABK-së aplikohen në një bashkësi të madhe fushash të kërkimit shkencor, ndër të cilat po paraqesim vetëm disa të marra nga rishikimi i literaturës, si:

Zbulimi i mashtrimeve - Rastet e reja të mashtrimit janë të njëjta me rastet e njohura. ABK-ja mund t'i zbulojë dhe të sinjalizojë ato raste që kanë nevojë për hetim. (Sadgali et al., 2019)

Klasifikimi i klientëve - Përdorimi i teknikave të ABK-së për të klasifikuar klientët në botën reale, duke parashikuar sjelljen blerëse të klientëve për një produkt specifik, duke përdorur karakteristikat e tyre demografike (Ahn et al., 2007).

Trajtimet mjekësore - Teknikat ABK-së përdoren në zgjidhjen e pikselëve konfliktual për të përmirësuar segmentet që ndodhen në imazherinë e veshkave të deformuara dhe segmenteve të nefroblastomës (Corbat et al., 2020).

Kreditë konsumatore - Objektivi i ABK-së këtu është të identifikojë sa më saktë aplikantët e aftë për kredi, të cilëve mund t'u jepet një e tillë, duke rritur kështu fitimet dhe duke i dalluar nga ata aplikantë jo kreditues, të cilëve do t'u refuzohej aplikimi i kredisë, për të ulur kështu humbjet (Zurada & Barker, 2007).

Rasti studimor i marrë në këtë punim trajton punësimin efektiv në një kompani. Qëllimi është që nëpërmjet teknikave të ABK-së të arrihet të vlerësohet një kandidat i ri nëse është i vlefshëm ose jo për kompaninë.

Materiale dhe metodat

Në rastin studimor të paraqitur në këtë punim, metodologjia e ndjekur fillon me zbatimin e algoritmit të K-fqinjit më të afërt[2] për gjetjen e rekordeve të ngjashme me rekordin e ri të ngarkuar në model. Pastaj, ky algoritëm përdoret për të klasifikuar ose vlerësuar rekordin e ri (López-Fernández et al., 2011).

Modeli ABK

Teknika më e lehtë për t'u zbatuar në këtë model është ajo me vetëm një interval fqinj. Çdo atribut vendoset vetëm në një interval të njohur, ndaj dhe kombinimi do të jetë më i lehtë. Modeli përdor grupin e të dhënave të trajnimit dhe një matje mbi ngjashmërinë. Kjo e fundit është më e lehtë nëse të gjitha atributet janë numerike në grupin e trajnimit dhe në këtë rast të dhënat mund të konsiderohen si pika në hapësirë. Në këtë mënyrë matet distanca Euklidiane, si rrënja katrore e shumës së katrorëve të diferencave të llogaritura nga distanca e rekordit të ri nga rekordi në grupin e trajnimit (Sidhu et al., 2011). Në modelet me

më shumë se një fqinj përdoren disa funksione për kombinim. Teknika më e thjeshtë në këtë rast është funksioni i mesatares mbi një numër arbitrar të fqinjëve në grup.

Metodologjia ABK

Performanca e teknikës ABK varet nga përzgjedhja e një grupi të duhur të trajnimit, i cili duhet të mbulojë të gjitha fushat e interesit dhe klasifikimi i atributeve duhet të jetë shumë i qartë. Për sa i përket pjesës vepruese duhet përzgjedhur funksioni më i përshtatshëm i distancës e i kombinimit, si dhe numri i fqinjëve sa më përfshirës. Çdo rekord i ri që ngarkohet duhet të përshtatet me atributet e rekordeve që janë në grupin e të dhënave të trajnimit. Metoda më e përdorur është gjetja e fqinjit më të afërt me distancën më të vogël nga rekordi i ri (Cunningham & Delany, 2007). Atributet e grupit të trajnimit mund të renditen dhe më pas të filtrohen vetëm atribute që na interesojnë për të përpunuar. Kjo do të ulë koston e kohës për një rekord të vetëm. Nëse kemi shumë rekorde, kjo kosto do të jetë shumë më e lartë.

Funksioni i distancës

Funksioni i distancës përdoret për të vërtetuar se sa të ngjashëm janë fqinjët. Ka disa funksione për të matur këtë distancë. Matja është më e lehtë kur matet distanca mbi atributin numerik, sesa atë kategorik (Cunningham & Delany, 2007).

Një hapësirë metrike përbëhet nga një çift (X, d), ku X është një bashkësi dhe d: $X \times X \rightarrow R$ është një funksion, i quajtur funksioni i distancës, i tillë që vlen për të gjitha ato x, y, z ∈ X. Ky funksion ka vetitë e mëposhtme:

1. Simetria: $d(x, y) = d(y, x)$.

2. Përcaktimi pozitiv: $d(x, y) \geq 0$ dhe $d(x, y) = 0$, në qoftë se dhe vetëm në qoftë se $x = y$.

3. Pabarazia e trekëndëshit: $d(x, z) \leq d(x, y) + d(y, z)$.

Hapësira metrike më e njohur është ajo Euklidiane e dimensionit n, me formulën standarde për distancën midis dy rekordeve dhe n atributeve:

$$d((x_1, \dots, x_n)(y_1, \dots, y_n)) = ((x_1 - y_1)^2 + \dots + (x_n - y_n)^2)^{1/2}$$ (Funksioni Euklidian)

Vlera përdoret në vlerë absolute. Atributet numerike kanë spektër të ndryshëm. Disa vlera mund të jenë më të mëdha se një numër i caktuar dhe të tjera shumë më të vogla. Ky numër më i madh i dalë nga këto formula matematikore mund të ndikojë në parashikim. Për të normalizuar këto vlera ekzistojnë disa funksione standardizimi. Më të përdorurat janë:

$$x^* = \frac{x - \min(x)}{range(x)} = \frac{x - \min(x)}{max(x) - \min(x)}$$ (Funksioni i normalizimit min–max)

$$x^* = \frac{x - mean(x)}{SD(x)}$$ (Funksioni i standardizimit të pikës Z)

Për variablat kategorike, zakonisht përdoret funksioni zero-një për dy atribute të fushës së vlerave, nëse ato janë vlera diskrete. Nëse janë vlera të vazhdueshme ato transformohen që të paraqiten në segmentin zero-një (Cunningham & Delany, 2007):

$$d(x_i, y_i) = \begin{cases} 0, & nëse\ x_i = y_i \\ 1, & përndryshe \end{cases}$$ (Funksioni i transformimit zero-një)

Funksioni i kombinimit

Veprimi tjetër i rëndësishëm i ABK-së është kombinimi i të dhënave të mbledhura nga fqinjët për të marrë zgjidhjen e duhur. Detyra këtu është të gjejmë K rekorde të ngjashme me rekordin e ri të ngarkuar në model. Pastaj përdorim rezultatin e këtyre k-fqinjëve më të afërt për të bërë parashikimin. Kur kemi vetëm një fqinj, kombinimi është shumë i thjeshtë. Rezultati do të jetë i njëjtë me fqinjin më të afërt, përndryshe kur k>1 mund të përdorim modelin e votimit, si funksion kombinimi (Cunningham & Delany, 2007), me dy variantet e tij të mëposhtme:

Votimi i papeshuar - Së pari vendoset numri i fqinjëve më të afërt me rekordin e ri. Ka disa teknika për të gjetur vlerën e duhur të K-së, por kjo del jashtë fokusit të këtij punimi. Zakonisht kjo vlerë duhet të jetë më e madhe se 3 dhe më e vogël se 6. Pasi vlera K është zgjedhur, përzgjidhen K-fqinjët me distancën më të vogël nga rekordi i ri. Në fund analizohet trendi i K-shembujve të përzgjedhur dhe i jepet rekordit të ri sjellja e shumicës.

Votimi i peshuar - Kjo metodë përdor faktin që fqinjët më të afërt duhet të kenë më shumë ndikim në parashikim. Zgjidhet një funksion kombinimi për të peshuar këta fqinjë. Më të përdorurat janë inversi i distancave ose katrori i kundërt i distancave. Në fazën e fundit për çdo sjellje bëhet shuma e peshës, duke gjetur sjelljen më optimale.

Rezultate dhe diskutime

Në rastin studimor të trajtuar, fokusi është përzgjedhja e stafit të ri për t'u punësuar në një kompani. Bashkësia e të dhënave gjendet *online* (Nargundkar, 2019) dhe përbëhet nga atributet mosha, shteti, diploma (lloji, p.sh., *Bachelor* apo *Master*), drejtimi (fusha e profesionit), përvoja (shprehur në vite), rezultati i

arritur ne kompani (dobët, minimal, përshtatshëm dhe shkëlqyer). Tabela 1 tregon një listë me 10 rekorde historike të cilat janë përdorur gjatë testimit të modelit. Rekordi i ri (i njëmbëdhjeti) që do i shtohet modelit dhe për të cilin duam një parashikim për rezultatin e pritshëm në kompani është si në vijim:

(11, 26, CA, MS, Sisteme Informacioni, 0 vite, ?) Rekordi i ri

Hapi i parë është konvertimi në vlera numerike duke përdorur funksionin zero-një. Në shembullin e paraqitur atributet "shteti" dhe "diploma" do të kenë vlerat diskrete 0 ose 1, ku nëse punonjësi vjen nga CA kjo vlerë është 1, përndryshe është 0. Ndërkohë nëse ai ka vetëm diplomë “Bachelor” atëherë vlera diskrete është 0, përndryshe kjo vlerë është 1. Atributi "drejtimi" është përshtatur në bazë të figurave profesionale që i nevojiten kompanisë, ku për secilin drejtim jepet një peshë me vlera numerike nga segmenti [0, 1]. Atributi "përvoja" është projeksioni në segmentin [0, 1], duke u mbështetur në shtrirjen fillestare të vlerave historike dhe duke i normalizuar ato me funksionet min-max.

Tabela 1. Rekordet historike të punonjësve me atributet përkatëse.

Nr.	Mosha	Shteti	Diploma	Drejtimi	Përvoja	Rezultati
1	27	CA	BS	Inxhinier	2 vite	Shkëlqyer
2	33	NV	MBA	Shofer	5 vite	Përshtatshëm
3	30	CA	MS	Shkenca Kompj.	0 vite	Përshtatshëm
4	22	CA	BS	Sist. Informacioni	0 vite	Dobët
5	28	CA	BS	Sist. Informacioni	2 vite	Minimal
6	26	CA	MS	Shofer	0 vite	Shkëlqyer
7	25	CA	BS	Inxhinier	3 vite	Përshtatshëm
8	28	OR	MS	Shkenca Kompj.	2 vite	Përshtatshëm
9	25	CA	BS	Sist. Informacioni	2 vite	Minimal
10	24	CA	BS	Sist. Informacioni	1 vit	Përshtatshëm

Tabela e re (Tab. 2) do të ketë vlera numerike të segmentit [0,1].

Tabela 2. Rekordet historike të punonjësve me vlerat e normalizuara.

Nr.	Mosha	Shteti	Diploma	Drejtimi	Përvoja	Rezultati
1	0.233	1	0	0.8	0.4	Shkëlqyer
2	0.433	0	1	0.6	1	Përshtatshëm
3	0.333	1	1	0.9	0	Përshtatshëm

4	0.067	1	0	1	0	Dobët
5	0.267	1	0	1	0.4	Minimal
6	0.2	1	1	0.6	0	Shkëlqyer
7	0.167	1	0	0.8	0.6	Përshtatshëm
8	0.267	0	1	0.9	0.4	Përshtatshëm
9	0.167	1	0	1	0.4	Minimal
10	0.133	1	0	1	0.2	Përshtatshëm

Ndërkohë, rekordi i ri që do i shtohet modelit normalizohet si në vijim:

(11, 0.533, 1, 1, 1, 0, ?) Rekordi i ri i normalizuar

Hapi pasardhës është të llogaritet distanca totale. Rezultatet janë të paraqitura në Tabelën 3. Nëse zgjidhet modeli me vetëm një fqinj, rekordi me numër 3, na jep parashikimin për rekordin e ri paraqitur në model. Nëse zgjidhet modeli me katër fqinjët më të afërt dhe me votim të papeshuar, do të kemi rezultatet: "Përshtatshëm", "Shkëlqyer", "Dobët", "Përshtatshëm". Rezultati i shumicës është "Përshtatshëm" sepse ka më shumë vota se të tjerët.

Tabela 3. Rekordet historike të punonjësve me distancat e llogaritura në diferencë me rekordin e ri.

Nr.	**Mosha**	**Shteti**	**Dipl.**	**Drejt.**	**Përv.**	**Rezultati**	**Totali**	**Invers**
1	0.3	0	1	0.2	0.4	Shkëlqyer	1.900	0.5
2	0.1	1	0	0.4	1	Përshtatshëm	2.500	0.4
3	0.2	0	0	0.1	0	Përshtatshëm	0.300	3.3
4	0.466	0	1	0	0	Dobët	1.466	0.7
5	0.266	0	1	0	0.4	Minimal	1.666	0.6
6	0.333	0	0	0.4	0	Shkëlqyer	0.733	1.4
7	0.266	0	1	0.2	0.6	Përshtatshëm	2.166	0.5
8	0.266	1	0	0.1	0.4	Përshtatshëm	1.766	0.6
9	0.366	0	1	0	0.4	Minimal	1.766	0.6
10	0.4	0	1	0	0.2	Përshtatshëm	1.600	0.6

Modeli me katër fqinjët më të afërt dhe me votim të peshuar, i jep çdo rezultati një peshë duke përdorur funksionin e distancës inverse, si më poshtë:

- "Përshtatshëm" = 3.3 + 0.6 = 3.9

- "Shkëlqyer" = 1.4
- "Dobët" = 0.7

Në bazë të vlerave të gjetura dhe të paraqitura në kolonën "Inversi" të Tabelës 3, rezultati është klasifikuar si i "Përshtatshëm". Ky është një rast ku me tre funksionet e ndryshme të paraqitura në punim, parashikimi doli i njëjtë, por mund të përzgjidhen shembuj të tjerë ku rezultati i parashikimit është i ndryshëm në varësi të funksionit të zbatuar.

Përfundime

Në këtë punim u zbatua modeli ABK duke i shtjelluar fazat e tij nëpërmjet një rasti studimor, ku fokusi ishte në parashikimin e një vendimmarrje të suksesshme për përzgjedhjen e punonjësve të rinj, duke u bazuar në të dhënat historike të punonjësve ekzistues të një kompanie.

Efikasiteti i modelit ABK rritet kur nuk përdoren baza të të dhënave të mëdha, dhe atributet hyrëse dhe dalëse janë në numër të vogël. Parashikimi varet nga mënyra se si zgjidhet funksioni i distancës apo funksioni i kombinimit, por edhe nga përzgjedhja e K-së. Të tre këto faktorë mund të ndikojë në saktësinë e rezultatit të parashikuar. Nga ana tjetër, aftësia përshtatëse e modelit ABK mund të quhet një pikë e fortë. Të dhënat e reja mund të bashkohen me ato historike duke riorganizuar kategoritë, por pa prishur modelin.

Matja e saktësisë së modelit llogaritet duke marrë të gjitha parashikimet e vërteta dhe duke i ndarë ato midis të gjitha vlerave të parashikuara. Në këtë drejtim, puna jonë e ardhshme do të fokusohet në rritjen e numrit të rekordeve historike dhe rritjen e numrit të testeve me rekorde të reja.

Literatura

Ahn, H., Kim, K. and Han, I. (2007). A case-based reasoning system with the two-dimensional reduction technique for customer classification. Expert Systems with Applications, 32(4), fq.1011-1019.

Corbat, L., Nauval, M., Henriet, J. and Lapayre, J. (2020). A fusion method based on Deep Learning and Case-Based Reasoning which improves the resulting medical image segmentations. Expert Systems with Applications, 147, fq.113-200.

Cunningham, P. & Delany, S. (2007). k-Nearest neighbour classifiers. Mult Classif Syst. 54. 10.1145/3459665

Sidhu, I., Lim, A., Max Shen, M. (2011). Quantitative Technology Methods that can Improve Business Operations, CET Technical Brief, Mars 2011

Lan, Y. & Neagu, D. (2007): A New Time Series Prediction Algorithm Based on Moving Average of nth-Order Difference. In: Proc. of Sixth International Conference on Machine Learning and Applications, fq. 248–253

López-Fernández, H., Fdez-Riverola, F., Reboiro-Jato, M., Glez-Peña, D., & Méndez, J. R. (2011). Using CBR as Design Methodology for Developing Adaptable Decision Support Systems. In (Ed.), Efficient Decision Support Systems - Practice and Challenges From Current to Future. IntechOpen. https://doi.org/10.5772/16923

Nargundkar, S. (2019). Characteristics of Past Job Applicants.

www.nargund.com/gsu/mgs8040/lecture/memory.xls. Aksesimi i fundit më 17.04.2022

Olson, D. and Delen, D. (2008). Advanced Data Mining Techniques. Verlag Berlin: Springer, fq.39-53.

Park, C. & Han, I. (2002): A Case-Based Reasoning with the Feature Weights Derived by Analytic Hierarchy Process for Bankruptcy Prediction. Expert Systems with Applications. 23, fq. 255-264. 10.1016/S0957-4174(02)00045-3.

Sadgali, I., Sael, N., & Benabbou, F. (2019): Performance of machine learning techniques in the detection of financial frauds. Procedia Computer Science, 148, fq. 45-54. https://doi.org/10.1016/j.procs.2019.01.007

Stanfill, C. & Waltz, D. (1986): Toward Memory-based Reasoning. Communications of the ACM 29 (12).

Zurada, J. & Barker, R. M. (2007): Using Memory-Based Reasoning For Predicting Default Rates On Consumer Loans. Review of Business Information Systems (RBIS), 11(1), fq. 1–16. https://doi.org/10.19030/rbis.v11i1.4426

---

[1]Shkurtimi për modelin "Memory-based reasoning"

[2]Algoritmi "K-nearest neighbor"