Pavle Ćirić

Prepoznavanje vrste biljke na osnovu oblika njenog lista korišćenjem nasumične šume odlučivanja

U ovom radu je ispitivana metoda za prepoznavanje vrsta biljaka na osnovu oblika njihovih listova. Za rešavanje ovog problema korišćen je algoritam nasumične šume odlučivanja. Prepoznavanje se vrši na osnovu 10 različitih karakteristika koje opisuju oblik lista. Za obučavanje algoritma i ispitivanje tačnosti korišćen je skup podataka Flavia (Wu et al. *2007), koji u sebi sadrzi fotografije listova 32 vrste biljaka. Ovim algoritmom postignuta je tačnost od 69%. Dobijeni rezultati su upoređeni sa rezultatima radova iz literature.*

Uvod

U ovom radu je predstavljena nova metoda za prepoznavanja vrste biljaka. Motivacija za ovaj rad je da se olakša proces prepoznavanja vrste biljaka za nestručna lica, zbog velike složenosti bioloških metoda za rešavanje istog problema. Rad se fokusira isključivo na prepoznavanje vrste biljka na osnovu oblika (konture) njenog lista.

Postoji više različitih metoda za rešavanje ovog problema. Neke od već postojećih metoda su: klasifikator maksimalne margine (Priya *et al*. 2012), neuronske mreže (Wu *et al*. 2007), kombinacija neuronske mreže sa prenosnom funkcijom radijalne baze, i prepoznavanje pomoću Zernikeovih momenata (Kulkarni *et al*. 2013). U ovom radu se koristi algoritam nasumične šume odlučivanja (random decision forest, RDF; Ho 1998).

Cilj rada je bio da se proveri koliko je RDF efikasan u klasifikaciji biljaka na osnovu oblika njenog lista koje su jednostavne za računanje, i da se uporedi sa metodama u literaturi.

Metod

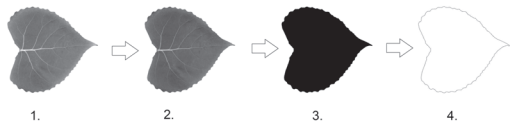
Prva faza rada bila je da se odaberu parametri na osnovu kojih se vrši klasiifkacia. Parametri su birani shodno mogućnostima odabranih alata, tako da budu što jednostavniji za računanje. Mogu podeliti u dve kategorije: bazne i glavne.

Prva kategorija su bazni parametri, i oni opisuju geometrijske karakteristike lista. U ovu kategoriju spadaju visina (H) i širina (L) lista, obim (O_O) i površinu (P_O) opisanog kruga, obim (O_P) i površina (P_P) opisanog pravougaonika najmanje površine, kao i obim i površina lista.

Početni korak u određivanju baznih parametara jednog lista je da se odredi njegova kontura i binarizovana slika lista. One su bitne zato što funkcije pomoću kojih se računaju bazni parametri primaju konturu i/ili binarizovanu sliku kao argumenate funkcije. Kontura i binarizovana slika se pronalaze tako što se prvo originalna fotografija pretvori u jednokanalnu fotografiju (grayscale). Zatim se jednokanalna fotografija binarizuje, i iz binarizovane slike se nalazi kontura (slika 1), korišćenjem Kenijevog algoritma za detekciju ivica (Canny 1986). Za ovo predprocesiranje se koriste funkcije iz biblioteke OpenCv.

Pavle Ćirić (2000), Pirot, učenik 4. razreda Tehničke škole u Pirotu

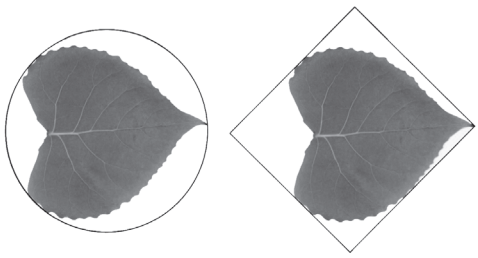
MENTOR: Stefan Nožinić, Continental Automotive Serbia



Slika 1. Preprocesiranje fotografije lista
1 – originalna fotografija u boji
2 – crno-bela slika
3 – binarizovana slika
4 – kontura

Figure 1. Image preprocessing
1 – Original image
2 – Grayscale image
3 – Threshold
4 – Contour

Kada se pronađu kontura i binarizovana slika, određuju se opisani krug i opisani pravougaonik najemanje površine (slika 2). Oni se određuju pomoću funkcija iz OpenCv biblioteke. Povratna vrednost funkcije koja određuje opisani pravougaonik najmanje površine je uređena četvorka koja sadrži kordinate centra pravougaonika, visinu, širinu i rotaciju pravougaonika. Od svih ovih vrednosti, bitne su samo širina (L) i visina (H) lista, pomoću kojih se računaju i obim (O_P) i površina (P_P) najmanjeg opisanog pravougaonika. Od koristi su i kordinate centra tog pravougaonika koje se kasnije koriste za računanje glavnih parametara.



Slika 2. Opisani krug oko lista (levo) i opisani pravougaonik oko lista (desno)

Figure 2. Enclosing circle around a leaf (left) and enclosing rectangle around a leaf (right).

Funkija koja računa opisani krug vraća uređeni par koji sadrži kordinate centra kruga i njegov poluprečnik, pomoću kojih se računaju obim (O_O) i površinu (P_O) opisanog kruga. Koordinate centra ovog kruga takođe se koriste za određivanje glavnih karakteristika.

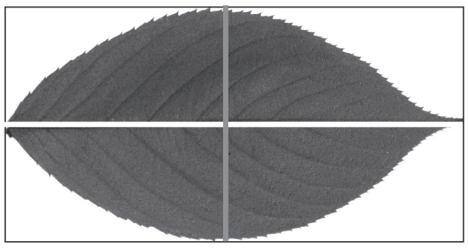
Svi bazni parametri su izraženi u broju piksela, samim tim se ne mogu koristiti za klasifikaciju, jer zavise od rezolucije fotografije.

Druga kategorija parametra su glavni parametri. Oni se računaju pomoću baznih parametra, tako da budu nezavisni od rezolucije fotografije lista. Glavni parametri su: odnos širine i visine lista, simetrija lista, sličnost lista sa krugom, sličnost lista sa pravougaonikom, odnos obima opisanog kruga i rastojanja između centara opisanog kruga i opisanog pravougaonika lista, prosečno rastojanje između centra lista i ekvidistantno raspoređenih tačaka po konturi lista, odnos obima opisanog kruga oko lista i površine lista, odnos obima opisanog kruga oko lista i obima lista, odnos obima opisanog kruga oko lista i visine lista i odnos obima opisanog kruga oko lista i širine lista.

Odnos širine i visine lista, $R = \frac{L}{H}$. Ukoliko je vrednost ovog odnosa $R < 1$, uzima se recipročna vrednost, te parametar ne silazi ispod 1.

Simetrija lista je definisana kao odnos površina sa jedne i druge strane simetrale. Određuje se kao proizvod uzdužne (s_H) i poprečne (s_V) simetrije lista:

$$S_I = s_H \cdot s_V$$



Slika 3. Uzdužna i poprečna simetrija lista

Figure 3. Longitudinal and transversal symmetry of a leaf

pri čemu se uzdužna simetrija simetrija se odnosi na uzdužnu, a poprečna simetrija na poprečnu simetralu (slika 3).

Sličnost sa krugom. Sličnost lista sa krugom određuje se kao količnik površine opisanog kruga oko lista i površine samog lista:

$$S_{\mathrm{K}} = \frac{P_{\mathrm{O}}}{P_L}$$

Sličnost sa pravougaonikom. Sličnost lista sa pravougaonikom predstavlja količnik površine najmanjeg opisanog pravougaonika sa površinom lista:

$$S_{\mathrm{P}} = \frac{P_{\mathrm{P}}}{P_L}$$

Odnos obima opisanog kruga i rastojanja između centara opisanog kruga i opisanog pravougaonika lista. Već smo pronašli centar opisanog kruga sa koordinatama (x_0, y_0) i centar najmanjeg opisanog pravougaonika sa koordinatama $(x_{\mathrm{P}}, y_{\mathrm{P}})$. Pitagorinom teoremom pronalazimo rastojanje ovih tačaka:

$$d = \sqrt{(x_0 - x_{\mathrm{P}})^2 + (y_0 - y_{\mathrm{P}})^2}$$

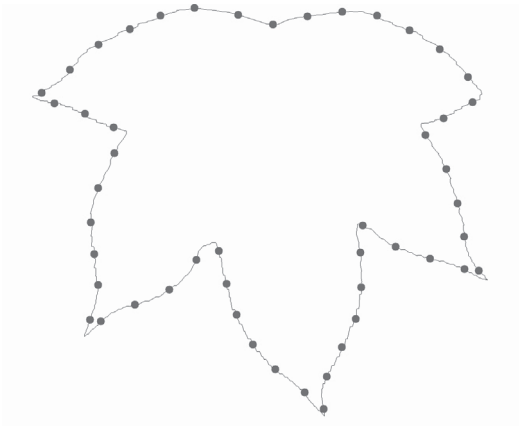
a zatim se računa količnik obima opisanog kruga oko lista i ovog rastojanja.

Faktor centra – prosečno rastojanje između težišta lista i ekvidistantno raspoređenih tačaka po konturi lista. Na konturi se postavlja n tačaka (P_1, P_2 … P_n) tako da svake dve tačke P_k i P_{k+1} (gde $k \in \mathbb{N}$ i $0 < k \leq n$) budu na jednakom rastojanju jedna od druge (slika 4).

Zatim se za svaku tačku na konturi lista P_k računa njeno rastojanje od težišta lista. Za težište lista se uzima presek dijagonala opisanog pravougaonika najmanje površine. Tako se dobija n rastojanja (d_1, d_2, ... d_n), iz kojih se računa prosečno rastojanje:

$$A = \frac{1}{n} \cdot \sum_{i=1}^{n} d_i \qquad (1)$$

Druga faza rada bila je da se ispita tačnost koja može da se postigne ovim parametrima, koristeći RDF algoritam. Glavni parametri se računaju za svaki list iz baze podataka Flavia (Wu *et al*. 2007; Flavia 2009), a dobijeni podaci su organizovani pomoću tabele u kojoj kolone predstavljaju parametre lista, a svaki red predstavlja jedan list. Vrsta biljke se upisuje u poslednjoj koloni.

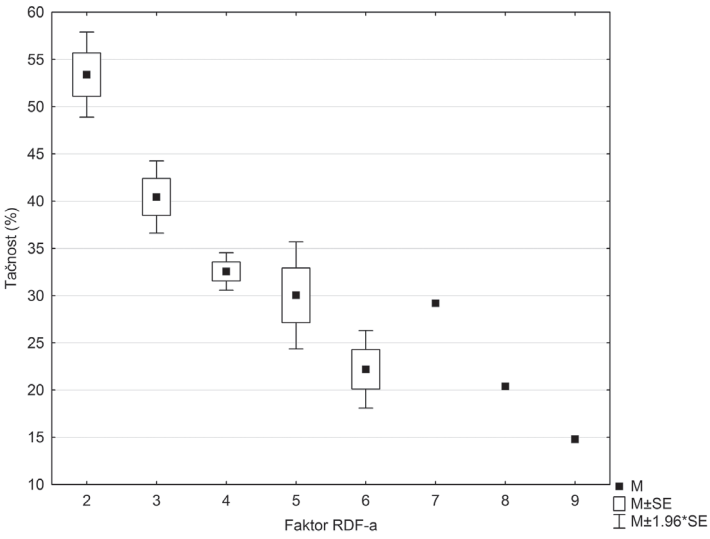


Slika 4. Ravnomerno raspoređene tačke na konturi lista (n = 50)

Figure 4. Evenly spaced points on the leaf's contour (n = 50)

Algoritam ima tri ulazna parametra: tabelu sa parametrima, dimenziju RDF-a i faktor RDF-a. Kao rezultat algoritma dobija se šuma odlučivanja, pomoću koje se vrši klasifikacija listova. Tačnost algoritma predstavlja tačnost klasifikacije rezultirajuće šume odlučivanja. Ona varira u zavisnosti od izabranih ulaznih podataka, pa je zato bitno da se izračuna za koje ulazne vrednosit se dobija najveća tačnost klasifikacije. U našem slučaju tabela sa parametrima je fiksna, pa samo ispitujem uticaj dimezije i faktora RDF-a.

Rekli smo da algoritam vraća šumu odlučivanja koja se sastoji od više stabala odlučivanja. Stablo odlučivanja se formira tako da se u korenu stabla nalazi parametar koji najviše pravi razliku među klasama, i tako rekurzivno dok se ne dođe do listova stabla. Kod običnih stabala odlučivanja se za svaki čvor uzimaju u obzir svi parametri koji se nalaze u tabeli da bi se našla karakteristika koja pravi najveću razliku među klasama. Stabla u RDF-u se razlikuju po tome što se ne uzimaju u obzir sve kolone tabele, već se bira n nasumičnih kolona od kojih će se tražiti kolona koja pravi najveću razliku među klasama. Prirodan broj n nazivamo faktor RDF-a.



Slika 5. Tačnost algoritma u zavisnosti od faktora RDF-a
M – srednja vrednost
SE – standardna greška srednje vrednosti

Figure 5. Relation between algorithm accuracy and RDF factor
M – mean
SE – standard error of the mean

Broj stabala u jednoj šumi odlučivanja nije fiksan. Ovaj broj utiče na tačnost algorima i nazivamo ga dimenzija RDF-a.

Tabela sa parapetrima se deli na dva tabele. Prvi, tzv. butstrepovana tabela (bootstrapped dataset), se formira tako što se nasumično uzimaju redovi iz originalne tabele i smeštaju u ovu. S obzirom da je biranje redova nasumično, u ovoj tabeli su dozvoljena ponavljanja. Ova tabela treba da ima isti broj redova kao i originalna tabela i koristi se za treniranje algoritma Druga, tzv. OOB tabela (out-of-bag dataset) se formira tako što se uzimaju svi redovi koji nisu izabrani za butstrepovanu tabelu i smeštaju u ovu. Ona se koristi za proveravanje tačnosti algoritma.

Rad je reazlizovan u programskom jeziku Python, a za obradu slika korišćena je biblioteka OpenCv.

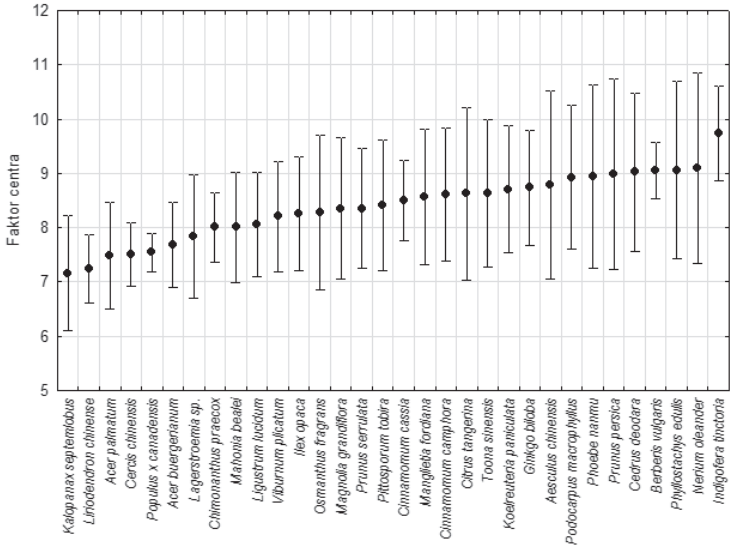
Rezultati i diskusija

Za ispitvanje algoritma korišćena je baza podataka Flavia koja se sastoji od 1907 fotografija listova (Flavia 2009). Među ovim fotografijama mogu se razlikovati 32 različite vrste biljaka. Sve fotografije su u rezoluciji 1600×1200, a listovi su fotografisani na beloj pozadini. Ova baza podataka se koristi za obučavanje algoritma, kao i za proveru njegove preciznosti.

Za početak je bilo potrebno da se izabere faktor RDF-a koji će dati najbolje razultate. Da bismo pronašli optimalnu vrednost ovog faktora, ispitano je za više dimenzija kako različiti faktori utiču na tačnost algoritma. Na slici 5 prikazana je dobijena zavisnost tačnosti od faktora RDF-a. Sa grafika se može videti da sa porastom faktora RDF-a, tačnost algoritma opada, pa su sva ostala merenja razmatrana sa faktorom 2.

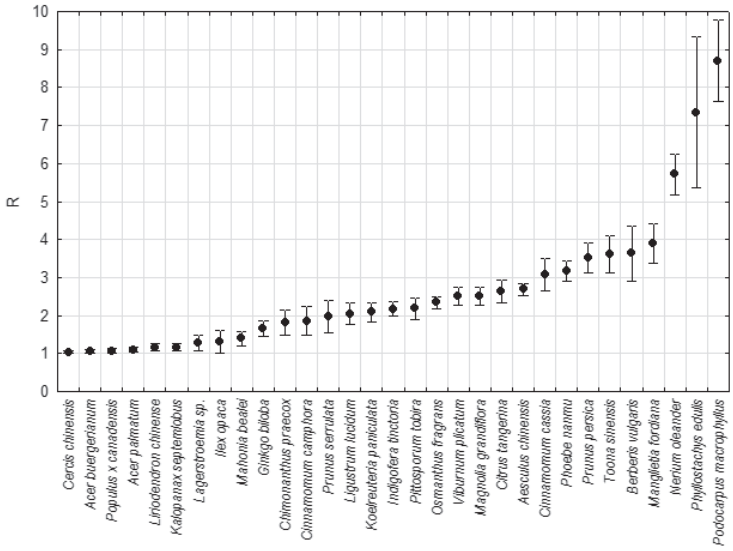
Što se tiče karakteristika na osnovu kojih se vrši klasifikacija, ispostavilo se da neke od predloženih karakteristika u ovom radu ne prave dovoljno veliku razliku između različitih klasa. Na primer, na slici 5 prikazane su srednje vrednosti faktora centra za svaku vrstu biljke sa standardnom divijacijom. Može se primetiti da ne postoji velika razlika između raličitih vrsta biljaka što se tiče faktora centra. Isti problem je konstatovan i kod još nekih parametara.

S druge strane, neke od karakteristika su više diskriminativne. Kao primer se može uzeti odnos visine i širine lista. Na slici 6 su prikazane srednje vrednosti ovog parametra sa standardnom devijacijom. Iz grafikona je izbačena jedna vrsta (himalajski kedar) zbog izuzetno visoke vrednosti ovog parametra u odnosu na ostale biljke. Ovo je urađeno zbog preglednosti grafikona, da bi se vizuelno istakla razlika između ostalih vrsta biljaka.



Slika 6. Vrednosti faktora centra (1) za ispitivane biljke. Tačkama je označena srednja vrednost, a vertikalnim linijama standardna devijacija.

Figure 6. Midpoint factor values for the leaves of plant species. Vertical bars denote the standard deviation.



Slika 7. Odnos širine i visine lista *R*. Tačkama je označena srednja vrednost, a vertikalnim linijama standardna devijacija. Zbog veličine ($R > 30$) nije prikazana vrednost za himalajski kedar (*Cedrus deodara*).

Figure 7. Length-width ratio values for the leaves of plant species. Vertical bars denote standard deviation. Due to its size ($R > 30$), deodara's (*Cedrus deodara*) value is not shown.

Tabela 1. Poređenje tačnosti rezultata predloženog algoritma sa drugim algoritmima za prepoznavanje vrste biljke.

Korišćena metoda	Tačnost (%)
Metoda potpornih vektora (Priya *et al*. 2012)	94.5
Neuronske mreže (Wu *et al*. 2007)	90.0
Kombinacija neuronske mreže sa prenosnom funkcijom radijalne baze i Zernikeovih momenata (Kulkarni *et al*. 2013)	93.8
Nasumična šuma odlučivanja (naš pristup)	68.7

Najveća postignuta tačnost iznosi 68.7%. Ovaj rezultat je postignut sa RDF-om sa 10 hiljada stabala odluke za faktor 2. Srednja tačnost pri ovim parametrima iznosi 66.0±1.5%. U tabeli 1 su prikazati rezultati ove metode klasifikacije u poređenju sa drugim metodama koje su predložene u literaturi. Merenja u navedenim radovima su takođe vršena nad bazom podataka Flavia.

Zaključak

Predloženom metodom klasifikacije nije postignuta zadovoljavajuća tačnost, s obzirom da već postoje druge metode sa boljim performansama. Neki od odabranih parametra nisu dovoljno diskriminirajući, pa možemo zakljuciti da je loš izbor parametra uzrok loših performansi ove metode klasifikacije. U poređenju sa metodama klasifikacije iz literature, predložena metoda je neupotrebljiva.

U daljem radu se može ispitati koje od izabranih karakteristika može da se izostave, one koje nisu dovoljno efektivne u razlikovanju različitih vrsta biljaka. Zatim, odabrati neke druge karakteristike koje bolje definišu list. Algoritam se može dodatno testirati metodom unakrsne validacije, da bi se dobila preciznija merenja.

Literatura

Canny J. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8** (6): 679.

Flavia 2009. Flavia plant leaf recognition system. http://flavia.sourceforge.net/

Ho T. K. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20** (8): 832.

Kulkarni A. H., Rai H. M., Jahagirdar K. A., Upparamani P. S. 2013. A leaf recognition technique for plant classification using RBPNN and Zernike moments. *International Journal of Advanced Research in Computer and Communication Engineering*, **2** (1): 984.

Priya C. A., Balasaravanan T., Thanamani A. S. 2012. An efficient leaf recognition algorithm for plant classification using support vector machine. U *International conference on pattern recognition, informatics and medical engineering (PRIME-2012)*. IEEE, str. 428-432.

Wu S. G., Bao F. S., Xu E. Y., Wang Y. X., Chang Y. F., Xiang Q. L. 2007. A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. U *2007 IEEE International Symposium on Signal Processing and Information Technology*. IEEE, str. 11-16.

Pavle Ćirić

Plant Recognition Technique Based on the Shape of Leaves Using Random Decision Forest

In this paper we have proposed a method of leaf recognition based on the shape of the leaves. The goal is to examine what results can be achieved by using Random Decision Forests (RDF) to solve this problem. The classification is based on 10 features that define the shape of a leaf. The Flavia dataset has been used to train the algorithm and to test its accuracy. The dataset contains high-quality images of leaves divided into 32 different types.

This method has yielded an accuracy rate of 69%. This result is compared with the accuracy of other already proposed methods. We can notice that the percentage rate yielded by this method is low in comparison. However, results have shown that this is not a RDF problem, but the consequence of bad feature design. To achieve better results, better features have to be chosen.