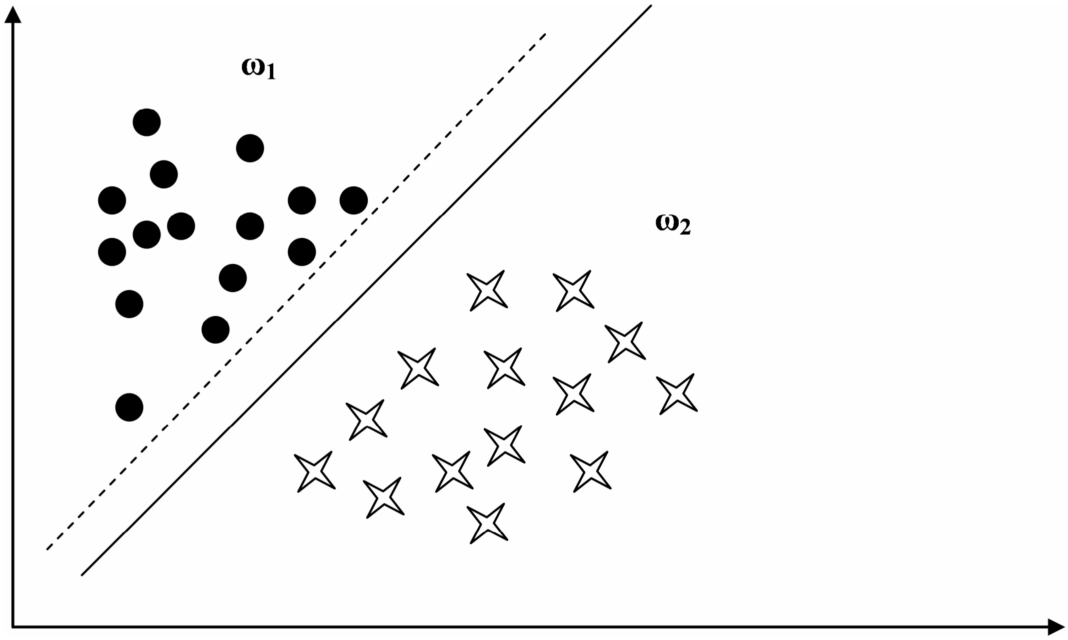
STROJ S POTPORNIM VEKTORIMA

Sadržaj

1. Razdvajanje dva razreda uzoraka 3
   1.1 Linearno separabilni razredi 3
   1.2 Linearno neseparabilni razredi 5
2. Klasifikacija više razreda uzoraka 8
   2.1 Jedan nasuprot svih metoda 8
   2.2 Klasifikacija temeljena na stablima odluke 9

1. Razdvajanje dva razreda uzoraka

Mnogi binarni problemi u raspoznavanju uzoraka se rješavaju pronalaskom decizijske funkcije (hiperravnina) koja razdvaja dva skupa uzoraka. Takvih decizijskih funkcija ima beskonačno mnogo (uz pretpostavku da su razredi linearno separabilni).



Slika 1.1 Linearno separabilni razredi

Dvije decizijske funkcije su prikazane na slici. Kao što se može uočiti obje pravilno razvrstavaju uzorke za učenje, no kada bi se moglo birati koju kasnije koristiti svi bi odabrali označenu punom linijom. Razlog tome je taj da funkcija označena punom linijom omogućuje i jednom i drugom razredu podjednake fluktuacije, tj. isprekidana linija ne omogućuje uzorcima iz razreda ω_1 nikakve pomake u desno. To svojstvo se naziva mogućnost generaliziranja i vrlo je važno svojstvo svakog klasifikatora jer ono određuje koliko će dobro klasifikator klasificirati uzorke izvan skupa za učenje. Cilj stroja s potpornim vektorima (*eng.support vector machine* ,*SVM*) je upravo pronaći takvu decizijsku funkciju kod koje je udaljenost do razreda maksimalna.

1.1 Linearno separabilni razredi

Neka je zadano N uzoraka x_i (i = 1,…,N) koji pripadaju razredu ω_1 ili ω_2. Budući da su razredi linearno separabilni decizijska funkcija je oblika

$$d(x) = w^T x + w_0 \qquad \textbf{(1.1)}$$

uz uvjet razvrstavanja

$$w^T x_i + w_0 \begin{cases} > 0 & x_i \in \omega_1 \\ < 0 & x_i \in \omega_2 \end{cases} \qquad \textbf{(1.2)}$$

Cilj je maksimizirati udaljenost najbližih elemenata iz razreda ω_1 i ω_2 do decizijske funkcije.
Poznato je da je udaljenost neke točke do decizijske funkcije d(x) jednaka

$$\frac{|d(x)|}{\|w\|} \tag{1.3}$$

Funkcija d(x) se može skalirati na taj način da u najbližoj točci razreda ω_1 poprimi vrijednost 1 ,a u najbližoj točci razreda ω_2 poprimi vrijednost -1.
Duljina margine nakon skaliranja iznosi

$$\frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \tag{1.4}$$

Sad se i uvjet razvrstavanja može zapisati u obliku:

$$w^T x_i + w_0 \begin{cases} \geq 1 & x_i \in \omega_1 \\ \leq -1 & x_i \in \omega_2 \end{cases} \tag{1.5}$$

Optimizacijski problem koji moramo riješiti je pronaći minimum funkcije (1.6) uz ograničenje (1.7).

$$\text{J(w)} = \frac{1}{2}\|w\|^2 \tag{1.6}$$

$$y_i(w^T x_i + w_0) \geq 1 \tag{1.7}$$

y_i je pomoćna varijabla koja ovisno o uzorku poprima vrijednost 1 ili -1 ($y_i = 1$ ako $x_i \in \omega_1$ odnosno $y_i = -1$ ako $x_i \in \omega_2$).
Ovo predstavlja kvadratni optimizacijski problem ograničen linearnom nejednadžbom.
Problem možemo zapisati u neograničenom obliku.

$$L(w, w_0, \lambda) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i \left[y_i (w^T x_i + w_0) - 1 \right] \tag{1.8}$$

Ovaj oblik se još naziva i Lagrangeova funkcija, a λ_i Lagrangeovi multiplikatori.
Optimum zadovoljava KKT(Karush-Kuhn-Tucker) uvjete dane jednadžbama (1.9) - (1.12).

$$\frac{\partial}{\partial w} L(w, w_0, \lambda) = 0 \tag{1.9}$$

$$\frac{\partial}{\partial w_0} L(w, w_0, \lambda) = 0 \tag{1.10}$$

$$\lambda_i \geq 0 \tag{1.11}$$

$$\lambda_i \left[y_i (w^T x_i + w_0) - 1 \right] = 0 \tag{1.12}$$

Uvrštavanjem (1.8) u (1.9) i (1.10) dobivamo jednadžbe (1.13) i (1.14).

$$w = \sum_{i=1}^{N} \lambda_i y_i x_i \qquad \textbf{(1.13)}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \qquad \textbf{(1.14)}$$

Iz jednadžbe (1.12) se može uočiti da je umnožak Lagrangeovog multiplikatora i iznosa decizisjke funkcije umanjenog za 1 pripadne točke uvijek jednak 0. Točke koje su udaljene za 1 od decizijske funkcije (te je time njihov iznos umanjen za 1 jednak 0) su upravo točke koje se nalaze na rubovima margina. U tim slučajevima pripadni Lagrangeov multiplikator može biti različit od 0. Te točke se nazivaju potporni vektori. Tu valja obratiti pozornost da nije nužno da sve točke koje se nalaze na rubu margina budu ujedno i potpornji vektori. Pripadni Lagrangeov multiplikator točaka koje se ne nalaze na rubu margine tj. za njih vrijedi d(x)>1 iznosi 0. To se može protumačiti na način da te točke nisu važne za određivanje decizijske funkcije i njihovim uklanjanjem se ne utječe na samu decizijsku funkciju.

Uvrštavanjem (1.13) i (1.14) u (1.8) dobiva se sljedeći optimizacijski problem uz uvjet (1.14).

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda i \lambda j y_i y_j x_i^T x_j \right) \qquad \textbf{(1.15)}$$

Normala ravnine je zadana jednadžbom (1.13) dok je slobodna konstanta(w_0) dana izrazom .

$$w_0 = y_i - w^T x_i \qquad \textbf{(1.16)}$$

Sa gledišta numeričke preciznosti bolje je uzeti izraz (1.17)

$$w_0 = \frac{1}{S} \sum_{i \in S} \left(y_i - w^T x_i \right) \qquad \textbf{(1.17)}$$

Sada se decizijska funkcija može zapisati u obliku

$$d(x) = \sum_{i \in S} \lambda_i y_i x_i^T x + w_0 \qquad \textbf{(1.18)}$$

sa uvjetom razvrstavanja

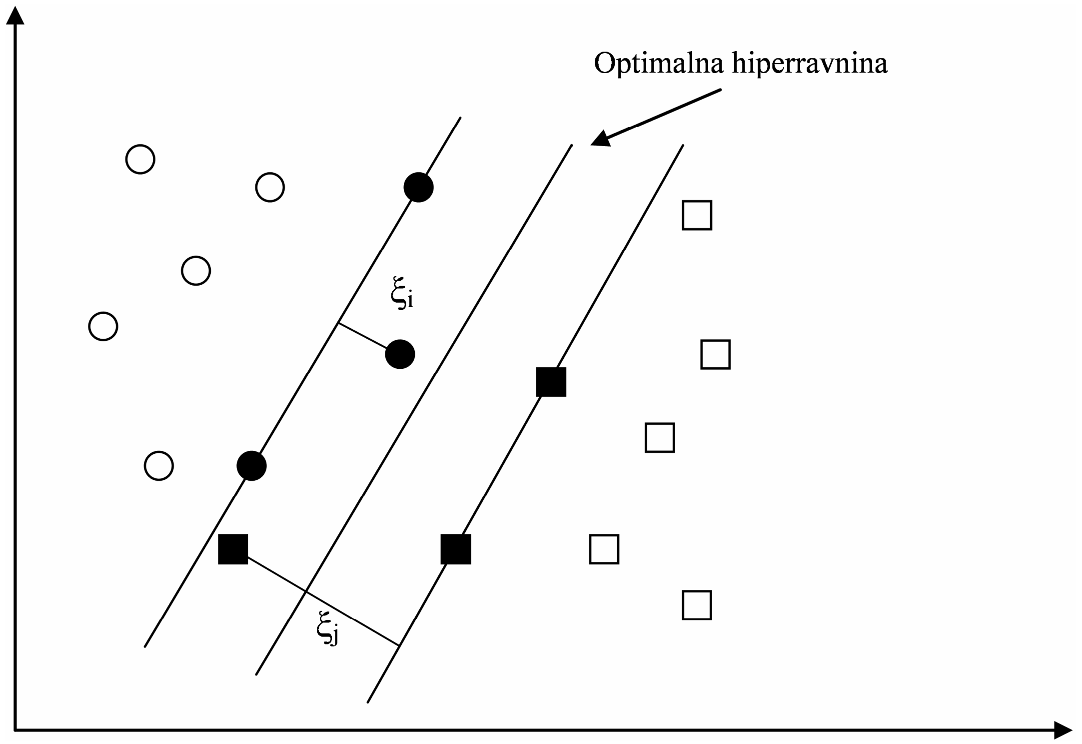
$$d(x_i) \begin{cases} > 0 & x_i \in \omega_1 \\ < 0 & x_i \in \omega_2 \end{cases} \qquad \textbf{(1.19)}$$

1.2 Linearno neseparabilni razredi

Najčešće su razredi linearno neseparabili i model predstavljen u prethodnom poglavlju ne može pronaći decizijsku funkciju.
Da bi se prilagodio postojeći model SVM-a linearno neseparabilnim razredima, uvodi se nova varijabla – nenegativna varijabla slabljenja ξ_i.

Uvjet zadan jednadžbom (1.7) postaje uvjet zadan jednadžbom (1.20)

$$y_i(w^T x_i + w_0) \geq 1 - \xi_i \qquad \textbf{(1.20)}$$



Slika 1.2 Linearno neseparabilni razredi

Uvođenjem varijable ξ_i postoji prihvatljiva decizijska funkcija. Ovisno o varijabli ξ_i postoje tri kategorije uzoraka:

1) $\xi_i=0$ uzorci koji pripadaju linearno separabilnom dijelu i obrađeni su u prethodnom poglavlju
2) $\xi_i > 0$ i $\xi_i \leq 1$ uzorci koji više nemaju maksimalnu marginu, ali još uvijek su točno klasificirani
3) $\xi_i > 1$ uzorci koji su pogrešno klasificirani

Sada je potrebno pronaći takvu decizijsku funkciju u kojoj je broj uzoraka koji nemaju maksimalnu marginu minimalan.
Potrebno je pronaći minimum funkcije

$$J(w) = \sum_{i=1}^{M} I(\xi_i) \qquad \textbf{(1.21)}$$

gdje je

$$I(\xi_i) = \begin{cases} 1 & \xi_i > 0 \\ 0 & \xi_i = 0 \end{cases} \qquad \textbf{(1.22)}$$

Ako se uzme u obzir i maksimizacija margine dobiva se sljedeća funkcija koju je potrebno minimizirati.

$$J(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} I(\xi_i) \qquad \textbf{(1.23)}$$

uz uvjete

$$y_i(w^T x_i + w_0) \geq 1 - \xi_i \tag{1.24}$$

$$\xi_i \geq 0 \tag{1.25}$$

C je parametar koji određuje važnost širine margine i broja uzoraka koji nemaju maksimalnu marginu.
Lagrangeov oblik danog problema dan je funkcijom

$$L(w, w_0, \xi, \lambda, \mu) = \frac{1}{2} \| w \|^2 + C\sum_{i=1}^{N} I(\xi_i) - \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} \lambda_i \left[y_i \left(w^T x_i + w_0 \right) - 1 + \xi_i \right] \tag{1.26}$$

KKT uvjeti su

$$\frac{\partial}{\partial w} L(w, w_0, \xi, \lambda, \mu) = 0 \tag{1.27}$$

$$\frac{\partial}{\partial w_0} L(w, w_0, \xi, \lambda, \mu) = 0 \tag{1.28}$$

$$\frac{\partial}{\partial \xi} L(w, w_0, \xi, \lambda, \mu) = 0 \tag{1.29}$$

$$\lambda_i \geq 0, \mu_i \geq 0 \tag{1.30}$$

$$\lambda_i \left[y_i (w^T x_i + w_0) - 1 + \xi_i \right] = 0 \tag{1.31}$$

$$\mu_i \xi_i = 0 \tag{1.32}$$

Iz (1.26) i (1.29) slijedi

$$C - \mu_i - \lambda_i = 0 \tag{1.33}$$

Dualni oblik je

$$\max_{\lambda} \left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda i \lambda j y_i y_j x_i^T x_j \right) \tag{1.34}$$

uz uvjete

$$0 \leq \lambda_i \leq C \tag{1.35}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{1.36}$$

2. Klasifikacija više razreda uzoraka

Iako možda SVM nije najpogodnija metoda za rješavanje problema raspoznavanja velikog broja razreda uzoraka(npr. 200 razreda) razvijene su mnoge metode koje daju zadovoljavajuće rezultate. Najveće probleme naravno, predstavlja samo učenje koje se zasniva na kvadratnom programiranju. Veliki broj razreda i veliki broj uzoraka za učenje može pretvoriti kvadratno programiranje u vrlo neefikasnu metodu koja zahtijeva veliku količinu memorije, a samo učenje traje vrlo dugo. Drugi problem koji se javlja kod klasifikacije više razreda su mogućnost pojavljivanja nedefiniranih regija. Za oba problema postoje efikasni pristupi koji ih rješavaju.
Neke od metoda su :

1. jedan nasuprot svih metoda
2. usporedba po parovima
3. svi odjednom metoda

Usporedba po parovima se temelji na izračunavanju $\frac{k(k-1)}{2}$ decizijskih funkcija, gdje je k broj razreda. Prilikom klasifikacije potrebno je izračunati vrijednosti $\frac{k(k-1)}{2}$ decizijskih funkcija i odlučiti se na temelju rezultata kojem razredu pripada nepoznati uzorak. Samo učenje decizijskih funkcija kao i kasnije testiranje i klasifikacija mogu biti neefikasni upravo zbog vrlo velikog broja samih decizijskih funkcija.
Za razliku od metode usporedbe po parovima, metoda svi odjednom računa samo k decizijskih funkcija. Funkcije računa tako da sve objedini u jedan velik optimizacijski problem (koji se i dalje rješava kvadratnim programiranjem). Iako se sada računa samo k decizijskih funkcija, optimizacijski problem je vrlo velik i zahtijeva veliku količinu resursa za veliki broj uzoraka.

2.1 Jedan nasuprot svih metoda

Ideja jedan nasuprot svih metode je odvajanje decizijskom funkcijom jednog razreda od svih preostalih. Pritom je potrebno izgraditi k decizijskih funkcija, gdje je k broj razreda.
Neka je formulom (2.1) zadana decizijska funkcija $d_i(x)$.

$$d_i(x) = w_i^T x + w_0 \qquad \textbf{(2.1)}$$

Tada će uzorak biti razvrstan u razred ω_i ako je zadovoljen uvjet (2.2)

$$d_i(x) > 0 \qquad \textbf{(2.2)}$$

Može se odmah uočiti jedan vrlo važan nedostatak ove metode, a to je pojava nedefiniranih regija. Tako klasifikator neće znati klasificirati slučaj kada je $d_1(x)>0$, $d_2(x)>0$, $d_3(x)<0$ ili slučaj kada je $d_1(x)<0$, $d_2(x)<0$, $d_3(x)<0$. Ovo je takozvana diskretna verzija decizijskih funkcija jer se uzima samo podatak je li iznos funkcije u danoj točci veći ili manji od 0.
Da bi se izbjeglo pojavljivanje nedefiniranih regija, uvodi se kontinuirani oblik decizijske funkcije dan izrazom (2.3).

$$\arg\max_{i=1,\dots,k} d_i(x) \qquad \textbf{(2.3)}$$

Sada će uzorak biti razvrstan u razred čija pripadna decizijska funkcija ima najveću vrijednost. Ovo je samo jedan od načina kojim se može riješiti problem nedefiniranih područja. Postoje još nekoliko metoda, a jedna su i stabla odluke.

2.2 Klasifikacija temeljena na stablima odluke

Stabla odluke su jedna od metoda kojom se može riješiti problem nedefiniranih područja koja se pojavljuju kod metode jedan nasuprot svih. Ideja je da se u svakom čvoru stabla provodi binarna klasifikacija, tj. da se odvoji jedna grupa razreda od druge. Potrebno je dati algoritam koji će po nekom pravilu grupirati razrede u posebne grupe. Upravo se prema tim algoritmima razlikuju nekoliko slučajeva, a ovdje će biti obrađen (i kasnije korišten) samo jedan.

Algoritam počinje grupiranjem svih razreda u dvije grupe. Za tako dobivene grupe određuje se decizijska funkcija koja ih odvaja. Postupak se rekurzivno nastavlja u čvorovima djece s određenim razredima.

Algoritam:
Pretpostavlja se da na početku razredi pripadaju različitim grupama, tj. svaki razred je jedna grupa.

1) Izračunaj centre grupa c_i prema formuli (2.4)

$$c_i = \frac{1}{|X_i|} \sum_{x \in X_i} x \qquad \textbf{(2.4)}$$

gdje je X_i grupa i.

2) Pronađi sve udaljenosti između centara l_{ij} prema formuli (2.5).

$$l_{ij} = \| c_i - c_j \| \qquad \textbf{(2.5)}$$

3) Grupiraj dvije najbliže grupe u jednu grupu.

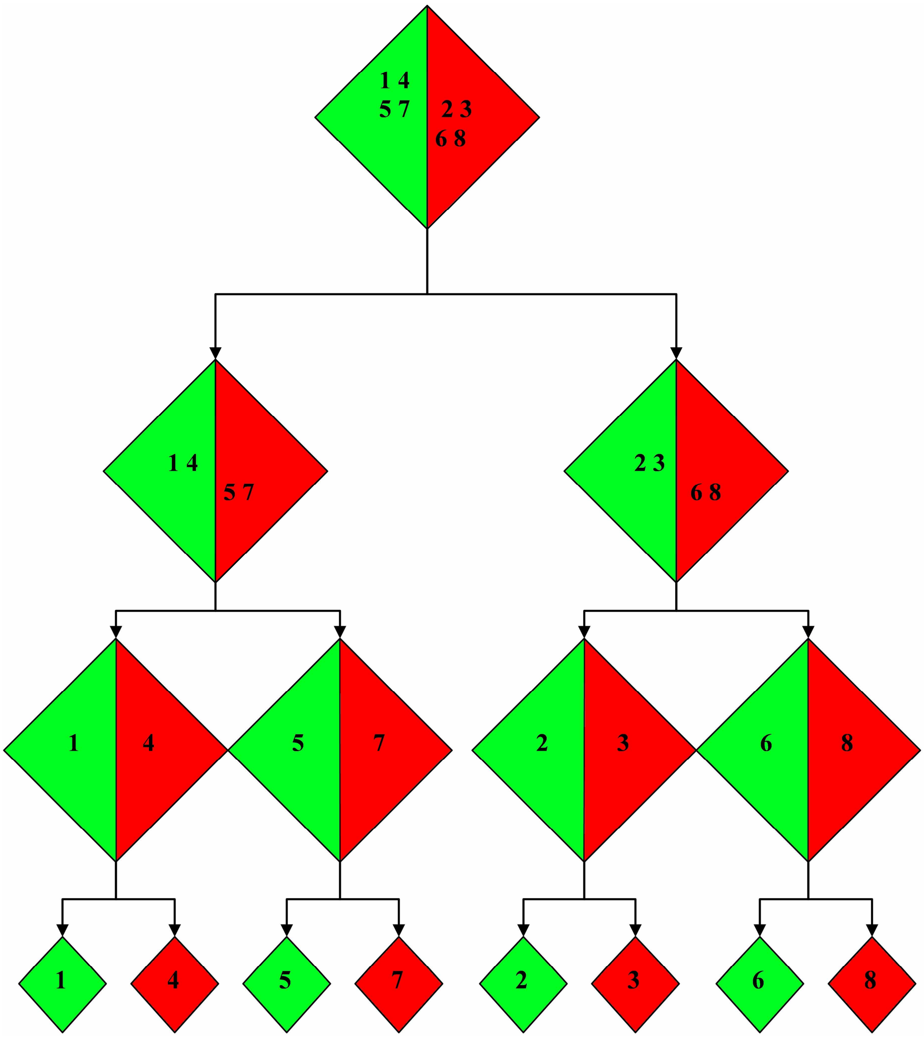
4) Ako je novi broj grupa veći od 2 idi na korak 1) inače idi na korak 5)

5) Odredi optimalnu decizijsku funkciju između dobivenih grupa.

6) Za svaku grupu stvori čvor dijete i ako je broj razreda u grupi veći od 1 idi na korak 1) inače prekid izvođenja.

Prilikom klasifikacije u čvoru se uzorak klasificira u jednu skupinu od grupa razreda. Klasifikacija se provodi sve dok se ne stigne do čvora koji sadrži jedan razred. Tada se nepoznati uzorak pridjeljuje u razred koji se nalazi u tom čvoru.
Neka je zadano 8 razreda i neka je sagrađeno stablo odluke kao na slici Slika 2.1. Tada bi se tijekom klasifikacije uzorka koji pripada razredu 5 prošlo kroz sljedeće korake.

1) u korijenu bi uzorak bio klasificiran u zelenu grupu
2) na prvoj razini uzorak bi bio klasificiran u crvenu grupu
3) na drugoj razini uzorak bi bio klasificiran u zelenu grupu
4) budući da se nalazimo u listu uzorak pripada razredu koji se nalazi u listu, a to je razred 5



Slika 2.1 Stablo odluke

Sada se mogu uočiti neka vrlo korisna svojstva ovakvog pristupa.
Prvo tijekom same klasifikacije nije potrebno računati k decizijskih funkcija, nego samo $\log_2 k$.
Također tijekom same izgradnje stabla izgradit će se samo k-1 decizijskih funkcija što je veliki napredak u usporedbi s $\frac{k(k-1)}{2}$. Također valja uočiti da je u korijenu optimizacijski problem jednak po složenosti metodama jedan nasuprot svih dok se u svakoj nižoj razini optimizacijski problem reducira i tako zahtijeva manje resursa.
Postignuto je značajnije napredovanje kako u brzini izgradnje samog klasifikatora tako i u postupku klasifikacije. Naravno jasno je da će sami rezultati biti ovisni o algoritmu grupiranja. Cilj samog algoritma grupiranja je da u svakom čvoru grupira razrede u podjednake (po broju razreda) grupe.