ОБЪЕКТЛАРНИНГ МУҲИМЛИЛИК ДАРАЖАЛАРИГА МОС КЛАСТЕРЛАШ ВА СИНФЛАШТИРИШ АЛГОРИТМЛАРИ

Олламберганов Файзулла Фарход ўғли

Муҳаммад-ал Хоразмий номидаги Тошкент ахборот технологиялари университети, Ўзбекистон, Тошкент,
E-mail: fayzulla0804@gmail.com

K E Y W O R D S

Кластерлаш, синфлаштириш, объектларни муҳимлилик даражаси, объектни синфларга қўшган ҳиссасини баҳолаш мезони.

A B S T R A C T

Мақолада белгилар қиймати аниқ сонлардан иборат бўлган объектларни кластерлаш ва синфлаштириш масаласи муфассал тадқиқ этилган. Ушбу тадқиқотда берилган барча объектлар ихтиёрий равишда r-та синфга ажратилади. Бу синфлар мажмуаси ўқув танланма деб аталади. Ўқув танланманинг йўл элементлари объектлар бўлиб, устун элементлари эса белгилар сифатида қаралади. Бизнинг ҳолатимизда белгилар аниқ сонлар орқали берилади.

Ўқув танланмани дастлабки ишлов бериш жараёнида объект белгиларини нормаллаштириш, аномал маълумотларни олиб ташлаш ва улар ўрнида нормал маълумотларни қўйиш, ҳамда ўқув танланмадан тушиб қолган маълумотларни тиклаш амалга оширилади. Мақолада ҳар бир объект тадқиқ этилади ва танланмадаги объектнинг муҳимлиги мезон орқали баҳоланади. Бу мезон объектнинг тадқиқ этилаётган синф шаклланишига қўшган ҳиссасини баҳолашга ёрдам беради. Биринчи навбатда объектнинг ўз синфига қўшган ҳиссаси, сўнгра қолган синфларни шаклланишига қўшган ҳиссаси баҳоланади. Агар тадқиқ этилаётган объектнинг қандайдир синф шаклланишига қўшган ҳиссаси юқори бўлса, ушбу объект ўша синфга ўтказилади. Бу жараён барча синф объектлари учун кетма-кет тўлиқ равишда бир неча марта амалга оширилади. Жараён объектлар жойи ўзгармасдан ва ўхшашлик даражаси керакли фоиздан ошганда тўхтатилади.

I-Кириш

Объектларни кластерлаш ва синфлаштириш назарий ва амалий масалаларни ечимини аниқлашда муҳим ўрин тутади. Таклиф қилинаётган объектларни муҳимлигини инобатга олган алгоритмларнинг долзарблигини бир нечта асосий ҳолатлар орқали тушунтириш мумкин. Замонавий дунёда маълумотлар ҳажми кескин ошмоқда. Бизнинг алгоритм объектларни тўғри кластерлаш ва синфлаштиришда ёрдам бериши, кенг кўламдаги маълумотлар билан самарали ишлаш имконини яратади. Бу, кўпчилик соҳаларда, масалан, молия, тиббиёт ва маркетингда тўғри қарорлар қабул қилиш учун муҳимдир.

Кластерлаш ва синфлаштириш алгоритмлари маълумотларни тезда тўғри кластерларга ажратиш ва синфлаштиришга имкон беради. Бу, қарор қабул қилиш жараёнини тезлаштиришга ёрдам беради ва натижаларни аниқроқ қилиб кўрсатади. Аномал маълумотларни тўғри аниқлаш ва нормал маълумотлар билан алмаштириш жараёни алгоритмнинг ишлаш самарадорлигини оширади. Аномал маълумотларнинг таҳлили ва тузатилиши маълумотларнинг ишончлилигини ва аниқлигини таъминлайди. Кластерлаш ва синфлаштириш алгоритмлари турли

© Олламберганов Файзулла Фарход ўғли

соҳаларда, масалан, тиббий диагностика, молия соҳасидаги хатарларни баҳолаш, ва интернетдаги маълумотларни классификация қилишда қўлланилади. Бундай алгоритмларнинг долзарблиги уларнинг кўплаб амалий муаммоларни ҳал қилиш имконияти билан таъминланган.

Машинали ўқитиш ва сунъий интеллектнинг ривожланиши билан кластерлаш ва синфлаштириш усулларига янги талаблар пайдо бўлмоқда. Масалан, катта маълумотлар (big data) ва реал вақтда маълумотларни таҳлил қилиш учун самарали ва ишончли алгоритмлар зарур. Ушбу алгоритмлар янги технологиялар билан мувофиқлаштирилиши мумкин. Машина ўрганиш ва сунъий интеллектнинг ўрганиш жараёнларида кластерлаш ва синфлаштириш муҳим аҳамиятга эга. Моделларнинг самарадорлигини ошириш учун, объектларнинг белгиларга тегишлилик даражасига асосланган кластерлаш ва синфлаштириш алгоритмлари тўғри ва самарали ўрганишни таъминлайди.

Объектларнинг сифатига тегишлилик даражасига боғлиқ кластерлаш ва синфлаштириш алгоритми, маълумотларни самарали таҳлил қилиш ва улардан қимматли маълумотлар олиш имконини беради. Ушбу алгоритмлар замонавий маълумотлар тўпламларини аниқ ва тез кластерлаш ва синфлаштириш, шунингдек, аномал маълумотларни тузатиш учун аҳамиятга эга. Ушбу мақсадда алгоритмларнинг долзарблиги, уларнинг кенг кўламдаги қўлланиш имкониятлари ва янги технологиялар билан мувофиқлаштириш имкониятлари билан ҳам асосланади.

K-means алгоритми [1] кластерлаш соҳасида энг самарали усуллардан бири ҳисобланади. Бу метод объектларни белгилар қийматига асосланиб кластерларга ажратишда ва кластер марказларини аниқлашда кенг қўлланилади. K-means алгоритмида объектлар кластер марказларига энг яқин масофада жойлашади, бу эса уларнинг кластерларга тўғри ажратилганлигини таъминлайди.

Hierarchical кластерлаш алгоритми [2] объектларни иерархик тарзда кластерлашни амалга оширади. Ушбу метод объектлар ва кластерлар ўртасидаги ўхшашлик даражасини баҳолаш ва кластерларни босқичма-босқич шакллантиришни назарда тутади. Hierarchical кластерлашда объектлар нисбатан яқин кластерларга ёки узоқ кластерларга қўшилиши мумкин, ва бу жараён даражалар асосида тўлиқ амалга оширилади.

Гаусс-микси моделлари [3] статистик моделлаштиришни ишлатиб, объектларнинг белгиларини кластерлашда қўлланилади. Бу метод объектларнинг белгиларининг тақсимотини ва уларнинг кластерларга бўлишини моделлаштиришга асосланган бўлиб, статистик йўсинда кластерларни аниқлашга ёрдам беради.

Синфлаштириш алгоритмлари [4] объектларни олдиндан белгиланган синфларга ажратиш учун ишлатилади. Бу усуллар объектларни белгилар бўйича синфлаштиришда турли техникаларни, масалан, қарор дарахтлари, наив Байес классификатори ва нейрон тармоқларини қўллайди. Синфлаштириш алгоритмлари маълумотларни аниқ синфларга ажратиш ва уларни самарали таҳлил қилиш учун муҳимдир.

Маълумотларни нормализация қилиш ва аномал маълумотларни ёқтириш [5] кластерлаш ва синфлаштириш жараёнларида муҳим аҳамиятга эга. Нормализация объект белгиларини стандартлаштириш, аномал маълумотларни аниқлаш ва уларни тузатиш натижасида, кластерлаш ва синфлаштиришнинг ишончлилиги ва самарадорлигини оширишга ёрдам беради.

Dunnнинг тадқиқоти [6] объектларни кластерлардаги тегишлилик даражасини баҳолаш ва кластер шаклланишини аниқлашда қўлланиладиган асосий усулларни тақдим этади. Ушбу тадқиқот объектларнинг кластерлардаги ҳиссасини баҳолаш ва самарали синфлаштириш усулларини ишлаб чиқишга йўналтирилган.

Кластерлаш ва синфлаштириш алгоритмлари объектларни белгилар қиймати асосида аниқ ва самарали кластерларга ажратиш ва синфлаштиришни таъминлайди.

(cc) BY © Олламберганов Файзулла Фарход ўғли

K-means ва Hierarchical кластерлаш методлари объектларни кластерларга ажратишда кенг қўлланилади, Гаусс-микси моделлари статистик моделлаштиришга асосланган кластерлашни амалга оширади, синфлаштириш алгоритмлари объектларни олдиндан белгиланган синфларга ажратишда ёрдам беради. Нормализация ва аномал маълумотларни ёқотиш жараёнлари эса кластерлаш ва синфлаштиришнинг самарадорлигига катта таъсир кўрсатади.

II-Кластерлаш масаласининг ифодаланиши

Фараз қилайлик, N-ўлчамли қийматли белгилар фазосида Q_p объектлар берилган бўлсин. Ҳар бир объект $x_i = (x_i^1, x_i^2, \dots, x_i^N)$, $i = \overline{1, M}$, N ўлчамли қийматли белгилар фазосида ифодаланган. Улар X тўпламидаги объектлар деб аталади ва умумий танланма объектлари ҳисобланади. Шунингдек, улар ўзлари мансуб бўлган X тўплами орқали белгиланади.

Масала: Дастлаб берилган умумий танланма $x_i \in X$, $i = \overline{1, M}$, объектларини шундай гуруҳларга ажратиш талаб этиладики, янги ҳосил бўлган синф $x_{p1}, x_{p2}, \dots, x_{pm_p} \in X_p, p = \overline{1, r}$ объектлари ўрта объект $\bar{x}_p$-га нисбатан синфлаштирилганда, ўз синфини топа олмаган объектлар сони Q_p минимал қийматни қабул қилсин.

Бу ерда $X = \bigcup_{p=1}^{r} X_p$, $\bar{x}_p=(\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N)$ — ўрта объект. Ўрта объект белгилари қиймати қуйидаги формула орқали ҳисобланади: $\bar{x}_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} x_{pi}^j, p = \overline{1, r}, j = \overline{1, N}$.

Бу масала адабиётларда [7-16] объектларни кластерлаш масаласи деб аталади. Масаланинг ечими учун қуйидаги алгоритм таклиф этилган.

Масаланинг математик модели

Бу масала кластерлаш (гуруҳлаш) масаласи бўлиб, мақсад объектлар тўпламини бир нечта кластерларга ажратишдан иборат. Масаланинг математик ифодасини аниқлаш учун қуйидаги тушунчаларни эътиборга оламиз:

Умумий танланма объекти x_i- белгилар фазосида берилган N ўлчамли векторлар тўплами бўлиб, улар X тўпламни ташкил қилади.

Кластерлар ёки синфлар X_p - умумий танланмадан ажратиб олинган кичик тўпламлардир, ва уларнинг ҳар бири ўрта объектга (кластернинг марказига) нисбатан яқин объектлардан ташкил топган.

Масаланинг мақсади — объектларни шундай гуруҳларга ажратишки, ҳар бир кластердаги объектлар ўз синфининг ўрта объектига нисбатан яқин бўлсин ва шунга мос ҳолда "ноаниқ" ёки "ўз кластерини топа олмаган" объектлар сони минимал бўлсин.

Математик модел:

Объектлар $x_i \in X$, $i = \overline{1, M}$ берилган.

Ҳар бир объект $x_i = (x_i^1, x_i^2, \dots, x_i^N)$, $i = \overline{1, M}$, N кўринишида берилган.

Кластерлар $x_{p1}, x_{p2}, \dots, x_{pm_p} \in X_p, p = \overline{1, r}$ тўпламларига ажратилади.

Ўрта объект $\bar{x}_p=(\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N)$ компоненталарини ҳисоблаш**:**

Ҳар бир кластернинг ўрта объекти (кластер маркази) белгилари қуйидаги формула орқали ҳисобланади:

$$\bar{x}_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} x_{pi}^j, p = \overline{1, r}, j = \overline{1, N}.$$

Бу ерда:

m_p — X_p кластердаги объектлар сони;

x_{pi}^j— p синфдаги i-чи объектнингj-чи белгиси.

Алгоритм:

Масаланинг ечими учун таклиф этилган алгоритм қуйидаги қадамларни ўз ичига олиши мумкин:

© Олламберганов Файзулла Фарход ўғли

1. Дастлабки гуруҳлаш: Умумий танланмани қандайдир дастлабки метод (кўпинча тасодифий) орқали синфларга бўлиш.

2. Ўрта объектларни ҳисоблаш: Ҳар бир кластернинг ўрта объектини юқоридаги формула орқали топиш.

3. Қайта гуруҳлаш: Ҳар бир объектни унинг энг яқин бўлган ўрта объектига қайта тайинлаш.

4. Такрорлаш: Қадамлар 2 ва 3ни такрорлаш, то кластерлар ўзгаришдан тўхтагунча ёки қўйилган мезонларга эришилгунча.

5. Қийматни минималлаштириш: Кластерлаш натижасидаги умумий "хатолар" (кластердан четлашган объектлар сони) минималлаштирилади.

Бу алгоритмда объектларни гуруҳларга бўлиш жараёни давомида ҳар бир объект ўзининг энг яқин кластер марказига ёки ўрта объектига "яқинлик" бўйича синфланади. Тажриба-синов ишлари шуни кўрсатдики жараён 4-5 такрорланишда ўз ниҳоясига етади.

Алгоритмнинг тўлиқроқ кўринишини қуйидагича ифодалаш мумкин:

Алгоритм бешта банддан иборат бўлиб, кластерлаш масалаларининг фақатгина алоҳида олинган синф объектлари учун қўлланса мақсадга мувофиқ бўлади.

Биринчи қадам. Умумий ўқув танланма объектлари маълумотлар базасига киритилади. Бошланғич маълумотлар базаси барча $x_i \in X$, $i = \overline{1,M}$, объектлар кесимида шакллантирилади;

Иккинчи қадам. Маълумотларга дастлабки ишлов бериш жараёни амалга оширилади. Бунда маълумотларни тозалаш (Cleaning) яъни: тўпланган маълумотларда хатолар, йўқолган қийматлар, дубликатлар, ёки ноаниқ маълумотлар бўлиши мумкин. Булар тўғриланади ва маълумотлар нормировка қилинади (Normalization).

Учинчи қадам. Умумий танланма тасодифий равишда $x_{p1}, x_{p2}, \dots, x_{pm_p} \in X_p, p = \overline{1,r}$ синфларга ажратилади.

Тўртинчи қадам. Шакллантирилган ҳар бир синф учун ўрта объект $\bar{x}_p$=($\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N$), $\bar{x}_p^j = \frac{1}{m_p}\sum_{i=1}^{m_p} x_{pi}^j, p = \overline{1,r}, j = \overline{1,N}.$ аниқланади.

Бешинчи қадам. Умумий танланмадаги барча объектлар билан бирин кетин ўрта объект орасидаги масофа ҳисобланади. Агар $\min_p |\bar{x}_p - x_i|$ бажарилса, у ҳолда $x_i \in X_p$. Натижада янги синфлар пайдо бўлади. Сўнгра ҳосил бўлган ҳар бир синф учун янги ўрта объект аниқланади ва олдинги жараён қайтадан амалга оширилади. Жараён ўрта объектлар ва янги синфдаги объектларнинг ўз жойида қолгунига қадар давом этади.

III-Синфлаштириш масаласининг ифодаланиши

Берилган объектларни синфлаштириш масаласини математик ифодасини соддалаштириш ва самарали ечим топишга йўналтириш мумкин. Қуйида масаланинг асосий қисмларини кўриб чиқамиз:

Масала:

Объектлар маълум белгиларга асосан рақамли ифодаланган. Синфлаштириш қоидаси қуйидагича ифодаланган:

$$R_\lambda(x) = X_p <=> \frac{|x - \bar{x}_p|_\lambda}{\frac{1}{m_p}\sum_{i=1}^{m_p}\left(x_{pi}^j - \bar{x}_p^j\right)^2} = min_{q=\overline{1,r}} \frac{|x - \bar{x}_q|_\lambda}{\frac{1}{m_q}\sum_{i=1}^{m_q}\left(x_{qi}^j - \bar{x}_q^j\right)^2},$$

Бу ерда:

$r = 1, \dots, r;$ синфлар сони.

$\Lambda^\ell = \{\lambda: \sum_{j=1}^N \lambda^j = \ell;\ \lambda^j = \{0,1\}, j = 1, \dots r, \ell \ll N\}$ информатив белгилар мажмуаси.

$\bar{x}_p$: р-синфнинг ўртача объекти.

(cc) BY © Олламберганов Файзулла Фарход ўғли

m_p: p-синфга тегишли объектлар сони.

Оптималлаштириш:

Белгиларни танлаш: Λ^{ℓ} маълумотларнинг информативлигини билдирувчи белгилар мажмуаси. Бу ерда $\ell \ll N$, яъни $\sum_{j=1}^{N} \lambda^j = \ell$ бўлган ҳолатлар фақат информатив белгиларни танлашни ифода этади. Шундай қилиб, оптимал λ ни танлаш – белгилар танлаш муаммосини ҳал қилиш билан боғлиқ.

Метриканинг танланиши: Қоидадаги $|x - \bar{x}_p|_\lambda$ ўзгарувчиси маълумотларнинг хусусиятларига қараб турли хил метрикаларни ўз ичига олиши мумкин. Энг мақбул метрикани танлаш алгоритмни янада самарали қилади.

Тақсимотни таҳлил қилиш: Қоидада ҳар бир синф учун дисперсия ҳисобланади, бу эса объектларнинг синфларга бўлинишини ўлчовчи индикатор ҳисобланади. Дисперсия асосида синфлаштириш масаласининг аниқлиги ошиши мумкин.

Бу ерда маълумотларни ўқувчи ва оптимал алгоритм ёрдамида қайта ишлаш йўли билан синфлаштириш алгоритмини оптималлаштириш мумкин.

1. Маълумотларни тайёрлаш:

Кириш маълумотларини олиш: белгилар матрицаси $x_i \in X_p$. ва синфлар объекти $x_i = (x_i^1, x_i^2, \dots, x_i^N)$,.

Ҳар бир синф учун ўртача объект $\bar{x}_p$=$(\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N)$, -ни ҳисоблаш қуйидаги формула орқали амалга оширилади:

$$\bar{x}_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} x_{pi}^j, p = \overline{1,r}, j = \overline{1,N}$$

Бу ерда x_{pi}^j — p-синфга тегишли i объектнинг j- белгиси.

2. Информатив белгиларни танлаш:

Информатив белги λ, Λ^{ℓ} тўпламдан танланади.

$$\Lambda^{\ell} = \left\{ \lambda: \sum_{j=1}^{N} \lambda^j = \ell;\ \lambda^j = \{0,1\}, j = 1, \dots r, \ell \ll N \right\}$$

Танланган λ қайси белгиларнинг ҳисоблаш ишларида қатнашиши англатади.

3. Метрикани ҳисоблаш:

Объект ва ўртача объектлар ўртасидаги масофа $|x - \bar{x}_p|_\lambda$ қуйидаги формула орқали ҳисобланади.

Ҳар бир синф учун дисперсия қуйидаги формула орқали ҳисобланади:

$$\sigma_p^2 = \left(\frac{1}{m_p}\right) \sum_{i=1}^{m_p} \lambda^j \left(x_{pi}^j - \bar{x}_p^j\right)^2$$

4. Тенгликни текшириш:

Ҳар бир синф учун қуйидаги қиймат аниқланади ва текширилади:

$$\frac{|x - \bar{x}_p|_\lambda}{\frac{1}{m_p} \sum_{i=1}^{m_p} \left(x_{pi}^j - \bar{x}_p^j\right)^2}$$

Ҳар бир синф учун юқоридаги қийматни минимал қиймат билан солиштириш амалга оширилади:

$$R_\lambda(x) = X_p$$

тенглиги минимум қийматга интилиши керак.

5. Объектни синфлаштириш:

Агар қуйидаги шартлар бажарилса:

$$\frac{|x - \bar{x}_p|_\lambda}{\sigma_p^2} = min_{q=\overline{1,r}} \frac{|x - \bar{x}_q|_\lambda}{\sigma_q^2}$$

У ҳолда x объекти p-синфга мансуб деб қабул қилинади:

$$R_\lambda(x) = X_p$$

(cc) BY © Олламберганов Файзулла Фарход ўғли

6. Натижани чиқариш:

Ҳар бир объектнинг мансублиги натижаси чиқарилади.

1. Маълумотларни киритиш

$x_i \in X_p$ ва синфлар белгиси $x_i = (x_i^1, x_i^2, \dots, x_i^N)$.

$x = (x^1, x^2, \dots, x^N)$ - текширилаётган объект.

2. Синфнинг ўртача қийматини ҳисоблаш

$$\bar{x}_p = (\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N),$$

$$\bar{x}_p^j = \frac{1}{m_p} \sum_{i=1}^{m_p} x_{pi}^j, p = \overline{1,r}, j = \overline{1,N}$$

3. Дисперсияни ҳисоблаш

$$\sigma_p^2 = \left(\frac{1}{m_p}\right) \sum_{i=1}^{m_p} \left(x_{pi}^j - \bar{x}_p^j\right)^2$$

4. Масофани ва тенгликни ҳисоблаш

$$\frac{|x - \bar{x}_p|_\lambda}{\frac{1}{m_p} \sum_{i=1}^{m_p} \left(x_{pi}^j - \bar{x}_p^j\right)^2}$$

5. Ҳар бир синф учун солиштириш ва минимални топиш

$$\frac{|x - \bar{x}_p|_\lambda}{\sigma_p^2} = min_{q=\overline{1,r}} \frac{|x - \bar{x}_q|_\lambda}{\sigma_q^2}$$

Бу қисмни барча синфлар учун бажариш лозим ва минимал қийматни топиш керак.

Бу алгоритм тенгликни текшириш орқали объектни тўғри синфга мансублигини аниқлайди.

IV. Синфлаштиришда кенг қўлланиладиган алгоритмлари

K Nearest Neighbor (KNN) — оддий ва кучли классификация алгоритми бўлиб, жуда кўп машинани ўрганиш (*machine learning*) ва маълумотларни таҳлил қилиш соҳаларида қўлланилади. Бу алгоритм ўқитувчили ўрганиш (supervised learning) турига киради ва классификация ҳамда регрессия масалаларида қўлланилади. KNN асосий фалсафаси шундан иборатки, ҳар бир нуқта яқин атрофидаги бир нечта қўшни нуқталарнинг синфига қараб таснифланади.

KNN алгоритмининг умумий қадамлари:

1. **Маълумотларни сақлаш:** Алгоритм ҳеч қандай модел яратмайди ва маълумотларни тўлиқ сақлаб туради. Тренинг босқичида фақат маълумотлар сақланади, ҳеч қандай ҳисоблашлар бажарилмайди.

2. **Евклид масофасини ҳисоблаш:** Янги нуқта бериладиган бўлса, у тренингдаги ҳар бир мавжуд маълумот нуқтасига бўлган масофа ҳисобланади. Масофаларни ҳисоблаш учун кўпинча Евклид масофаси формуласи ишлатилади:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

бу ерда *x* ва *y* - икки таянч нуқта, x_i ва y_i -еса ушбу объектларнинг *i*- mos белгилари.

3. **К-сонидаги энг яқин қўшниларни танлаш:** Масофалар аниқлангандан сўнг, энг яқин ***K*** сонидаги қўшни нуқталар танланади. ***K*** гиперпараметри (одатда кичик бўлган бутун сон) алгоритмнинг нозиклигига таъсир қилади. *K* қиймати катта бўлса, классификация умумийлаштирилади, кичик бўлса, аниқлиги ошади.

4. **Қўшниларнинг синфини аниқлаш:** Энг яқин қўшнилар синфи аниқлангандан сўнг, бу синфлар орасида энг кўп учрайдиган синф янги нуқтага тегишли деб топилади. Бунинг учун одатда "кўпчилик овози" (majority vote) усули қўлланилади.

5. **Таснифлаш:** Янги объект ***K*** қўшнининг энг кўп учрайдиган синфига таснифланади.

(cc) BY © Олламберганов Файзулла Фарход ўғли

Random Forest алгоритми - бу йиғма ўрганиш (ensemble learning) усулига асосланган алгоритм бўлиб, тасодифий равишда танланган бир неча қарор дарахтлари (decision trees)ни бирлаштириб таснифлаш ёки регрессия вазифаларини бажаради. Алгоритмнинг асосий фалсафаси шундан иборатки, бир нечта алоҳида (ва тасодифий) қарор дарахтлари томонидан кирувчи маълумотлардан синф ёки регрессия чиқарилади ва кейин уларнинг натижалари бирлаштирилади. Бу усул умумийлаштириш қувватини оширади ва бир дарахтнинг камчиликларини ёпишга ёрдам беради.

Random Forest алгоритмининг умумий қадамлари:

1. **Маълумотларни бутстрап (bootstrap) ёрдамида танлаш:**

Random Forest алгоритми тренинг маълумотлари орасидан кўплаб (тасодифий равишда) кичик тўпламлар (subset) яратади. Бу бутстрап усули деб аталади. Ҳар бир дарахтни ўқитиш учун ушбу тўпламлардан фойдаланилади. Бунда маълумотлар бир неча бор танланиши мумкин.

2. Тасодифий хусусиятларни танлаш:

Ҳар бир қарор дарахти фақат маълумотлар тўпламининг тасодифий танланган кичик қисми билан ўқитилади. Ҳар бир дарахт ҳар қандай қадамда тасодифий танланган бир нечта хусусиятлар (features) орқали синфлашни амалга оширади, бу эса дарахтларнинг ўзаро бир-биридан мустақил ва турли хил бўлишини таъминлайди.

3. Қарор дарахтларини ўқитиш:

Ҳар бир тўплам учун турли хил қарор дарахтлари (decision trees) ўқитилади. Бу дарахтлар маълумотнинг тасодифий қисмлари билан ўқитилади ва ёндошлигига қараб синф ёки регрессияни башорат қилади. Қарор дарахтлари учун ҳар бир тугунда (node) хусусиятларнинг бир нечта тасодифий вариантидан бири танланади.

4. **Синфлаштириш ёки регрессия қилиш:**

Синфлаштириш вазифасида Random Forest ҳар бир дарахтдан натижаларни олади ва энг кўп учраган синфни танлайди. Регрессия вазифасида еса дарахтлар чиқарган регрессияларнинг ўртачаси олинади.

Naive Bayes - эҳтимолга асосланган классификация алгоритми бўлиб, унинг асосий қонуни **Байес теоремаси**га асосланади. Бу алгоритм шубҳали ҳолатларга қарамай тез, содда ва самарали синфлаш натижаларини бера олади. У “наив” (oddiy) деб аталишининг сабаби - алгоритм ҳар бир хусусиятнинг натижага таъсирини бир-биридан мустақил деб қарашидадир. Ҳақиқатда, кўпгина хусусиятлар ўзаро боғлиқ бўлиши мумкин, лекин бу тахминга қарамай, Naive Bayes кўплаб амалиётларда аниқ натижалар беради. Naive Bayes алгоритми синфлашда **Байес теоремаси**дан фойдаланади. Байес теоремаси маълумотларга асосланиб, айрим синфга тегишли бўлиш эҳтимолини ҳисоблайди.

Naive Bayes алгоритмининг қадамлари:

1. **Маълумотларни тайёрлаш:**

Маълумотлар тўпламидаги ҳар бир синф ва хусусиятлар орасидаги эҳтимолий боғлиқликлар аниқланади. Бу учун тренинг маълумотлари орқали хусусиятлар ва синфлар ўртасидаги алоқаларни баҳолаш талаб қилинади.

2. **Синфлар эҳтимолини ҳисоблаш:**

Байес теоремасига кўра, ҳар бир синф учун (P(C|X)) эҳтимоллиги ҳисобланади. Алгоритм учун синфнинг шартли эҳтимоли муҳим ҳисобланади.

3. **Натижаларни чиқариш:**

Барча синфлар учун (P(C|X)) ҳисоблангандан сўнг, энг юқори эҳтимолга эга бўлган синф танланади.

Decision Tree (қарор дарахти) - бу синфлаш ва регрессия вазифаларини бажариш

(cc) BY © Олламберганов Файзулла Фарход ўғли

учун қўлланиладиган ўқитувчили ўрганиш алгоритми. У дарахт кўринишидаги иерархик тузилмадан иборат бўлиб, ҳар бир тугун маълумотларни муайян хусусиятларга кўра шохлантирилади ва маълумотларни синфларга ёки қийматларга ажратади. **Decision Tree** асосий ғояси - қарор қабул қилиш жараёнини оддий ва тушунарли тарзда визуаллаштириш ва бу орқали ҳар бир қарори маълумотларнинг хусусиятларига асосланган бўлишидир.

Decision Tree алгоритмининг ишлаш жараёни:

1. **Илдиз тугунни танлаш:**

Илдиз тугун учун қайси хусусиятни танлаш масаласи муҳимдир. Бу хусусият маълумотларни энг яхши ажратиш учун танланади. Бунда асосий мезонлардан бири **энергетик ютуқ** ёки **Gini индекс** бўлиши мумкин.

2. **Маълумотларни тақсимлаш:**

Ҳар бир тугунда маълумотлар танланган хусусиятга кўра иккита ёки ундан кўпроқ қисмларга ажратилади. Бу ажратиш маълумотларнинг шунга мос хусусият қийматларига қараб бўлинади.

3. **Дарахтни чуқурлаштириш:**

Ҳар бир тугунда тақсимот амалга оширилганидан сўнг, янги тугунлар яратилади ва бу жараён маълумотларни тўлиқ синфлашгача давом этади.

4. **Баргга етиб келиш:**

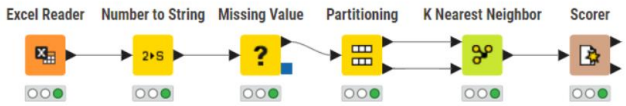
Барг тугунига етиб келганимизда, синфлаш ёки регрессия амалга оширилади. Агар синфлаш вазифаси бўлса, маълумотлар айнан ўша синфга тегишли деб белгиланади. Агар регрессия вазифаси бўлса, ўртача қиймат ёки бошқа регрессия усули ишлатилади.

5. **Дарахтни оптималлаштириш:**

Дарахтни чексиз чуқурлаштириш эҳтимоли мавжуд бўлгани сабабли, **pruning** жараёни амалга оширилади. Бу жараёнда ортиқча шохларни кесиб ташлаш орқали умумийлаштириш қувватини оширади.

V. Объектларни муҳимлилик даражаларига мос синфлаштириш ва KNIME муҳитидаги алгоритмларнинг амалий натижалар таҳлили

Юқорида таърифланган алгоритмлари нейрон тармоқларни scratch-blok шаклида қуришга мўлжалланган KNIME муҳитида тажриба-синовдан ўтказилди. KNIME муҳитидаги KNN алгоритмининг модели 1-расмда келтирилган.

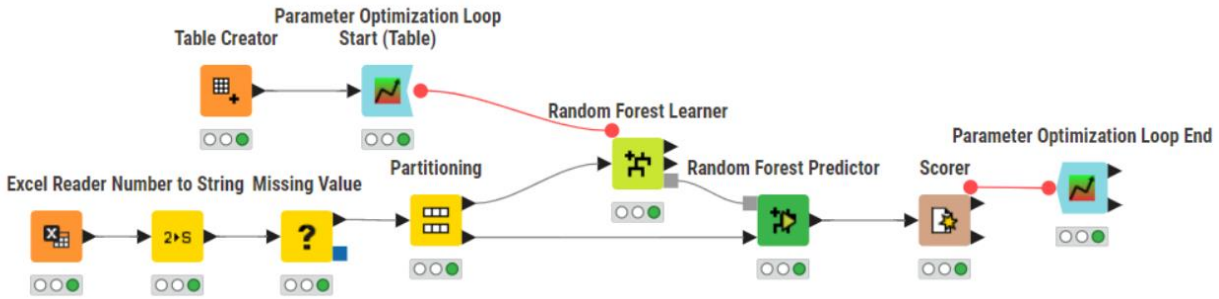


*1-расм. KNN алгоритмининг **KNIME** платформасидаги модели*

KNN моделидаги ҳар бир **scratch-blok** ларнинг вазифалари:

1. **Excel Reader** - Excel файлдан маълумотларни ўқиш ва юклаб олиш.
2. **Number to String** - Сон маълумотларини сатр маълумотларига айлантириш.
3. **SMOTE** - Ҳар хил типдаги маълумотларни қуйидаги усуллардан фойдаланиб кўпайтириш: "Synthetic Minority Over-sampling Technique".
4. **Missing Value** - Мавжуд бўлмаган маълумотларни тўлдириш ёки ёқотиш.
5. **Partitioning** - Маълумотларни тест ва ўқув танланмаларига бўлиш.
6. **K Nearest Neighbor** - KNN алгоритми билан маълумотларни классификациялаш ёки регрессиялаш.
7. **Scorer** - Модел аниқлиги метрик бахоларининг умумий натижаларини чоп этиш.

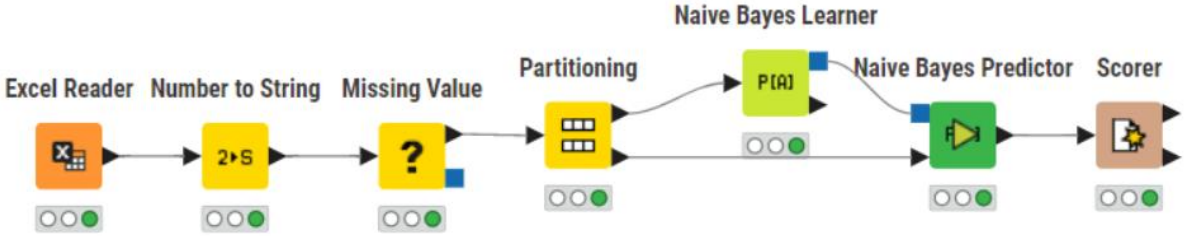
(cc) BY © Олламберганов Файзулла Фарход ўғли



*2-расм. Random Forest алгоритмининг **KNIME** платформасидаги модели*

Random Forest моделидаги **scratch-blok** ларнинг вазифалари:

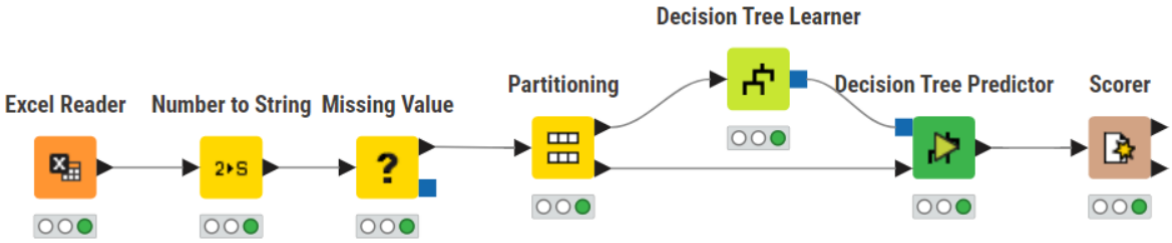
1. **Parameter Optimization Loop Start (Table)** - параметрларни оптимизация қилиш. Бунда турли параметр қийматлари текширилади ва оптимал параметр танланади.
2. **Random Forest Learner** - Random Forest алгоритмини ўқув танланма орқали ўқитиш.
3. **Random Forest Predictor** - Random Forest моделига асосланган ҳолда янги маълумотларни синфлаштириш.
4. **Parameter Optimization Loop End** - параметрлар оптимизациясини якунлаш.



*3-расм. Naive Bayes алгоритмининг **KNIME** платформасидаги модели*

Naive Bayes моделидаги **scratch-blok** ларнинг вазифалари:

1. **Naive Bayes Learner** - Naive Bayes алгоритми асосида синлаштириш моделини ўқув танланма орқали ўқитиш.
2. **Naive Bayes Predictor** - Naive Bayes моделини орқали янги маълумотларни синфлаштириш.



*4-расм. Decision Tree алгоритмининг **KNIME** платформасидаги модели*

Decision Tree моделидаги **scratch-blok** ларнинг вазифалари:

1. **Decision Tree Learner** - Decision Tree алгоритмини ўқув танланма орқали ўқитиш.

(cc) BY © Олламберганов Файзулла Фарход ўғли

2. **Decision Tree Predictor** - Decision Tree моделига асосланган ҳолда янги маълумотларни синфлаштириш.

Таҳлилда кластерлаш масаласи тадқиқ этилган бўлиб, бунда 1116 та қон босимига таъсир этувчи ҳар хил турдаги дори-воситалари жадвал кўринишда олинган. Бунда битта номинал белги ва тўққизта миқдорий белги тадқиқ этилган. Улар 1- жадвалда қисман берилган:

Жадвал-1

Қон босимига (Гипертензияга қарши) таъсир этувчи дори воситалари ҳақида маълумот											
№	**Дори воситаси номи**	**Фаол модда улуши мг**	**Эрув-чанлик %**	**Ранги**	**Ишлаб чиқариш**	**Қондаги концен-**	**Таркалиш**	**Тайёрланган таркибларнинг технологик кўрсаткичлари**			
								Сочилувчанлик, 10^{-3} кг/с	**Сочилувчанзичлик, кг/м^3**	**Жигардаги мотеболизми %**	**Парчаланиши**
1	x_1	0,10	96	сариқ	10	64	3,50	4,51	20,72	90	7,41
2	x_2	0,10	85	оқ	20	78	4,20	7,85	14,36	75	8,56
3	x_3	0,50	88	оч сариқ	25	20	4	3,87	25,48	95	5,69
4	x_4	0,03	89	оқ	50	72	0,20	16,87	15,64	75	7,96
5	x_5	0,20	91	оқ	15	60	3,20	19,20	25,63	85	5,98
6	x_6	0,50	93	оқ	25	57	3,90	4,89	48,65	89	6,85
.....											
1113	x_{1113}	0,5	76	оқ	25	86	4,01	7,41	13,25	78	9,65
1114	x_{1114}	0,25	74	кулранг	14	93	4,5	3,41	10,26	60	12,48
1115	x_{1115}	0,5	72	оқ	28	94	3,6	4,85	12,48	85	9,85

Кластерлаш ва синфлаштириш масаласини ечимини топиш учун қуйидаги алгоритм таклиф этилган.

Алгоритм қуйидаги бандлардан иборат:

Биринчи қадам. Дори воситалар маълумотлари $x_i = (x_i^1, x_i^2, \dots, x_i^N) \in X, \quad i = \overline{1, M}$ га дастлабки ишлов берилади. Бунда тушиб қолган маълумотлар тўлдирилади, аномал маълумотлар ўрта қийматга алмаштирилади ва миқдорий белгилар нормировка қилинади.

Иккинчи қадам. Ўқув танланма объектлари маълумотлар базасига киритилади. Бошланғич маълумотлар базаси барча X_p, $p = \overline{1, r}$ синф объектлари кесимида шакллантирилади;

Учинчи қадам. Умумий танланма тасодифий равишда $x_{p1}, x_{p2}, \dots, x_{pm_p} \in X_p, p = \overline{1, r}$ синфларга ажратилади.

Тўртинчи қадам. Шакллантирилган ҳар бир синф учун ўрта объект $\bar{x}_p$=$(\bar{x}_p^1, \bar{x}_p^2, \dots, \bar{x}_p^N)$, $\bar{x}_p^j = \frac{1}{m_p}\sum_{i=1}^{m_p} x_{pi}^j, p = \overline{1, r}, j = \overline{1, N}$.
аниқланади.

(cc) BY © Олламберганов Файзулла Фарход ўғли

Бешинчи қадам. Умумий танланмадаги барча объектлар билан бирин кетин ўрта объект орасидаги масофа ҳисобланади. Агар $\min_p |\bar{x}_p - x_i|$ бажарилса, у ҳолда $x_i \in X_p$. Натижада янги синфлар пайдо бўлади. Сўнгра ҳосил бўлган ҳар бир синф учун янги ўрта объект аниқланади ва олдинги жараён қайтадан амалга оширилади. Жараён ўрта объектлар ва янги синфдаги объектларнинг ўз жойида қолгунига қадар давом этади.

Таклиф этилган назарий тадқиқотлар, алгоритмлар асосида юқорида баён этилган масала ечилган бўлиб қон босимига (Гипертензияга қарши) таъсир этувчи дори воситаларини тўғри синфлаштириш бўйича ҳар бир синф кесимида ва алгоритмлар кесимида 2- жадвалда ўз аксини топган.

Жадвал-2

Алгоритмлар тахлилидан олинган натижалар

№	K Nearest Neighbor алгоритми	Random Forest алгоритми	Naive Bayes алгоритми	Decision Tree алгоритми	Таклиф этилаётган алгоритм
1-синф	93%	93%	100%	87%	100%
2-синф	81%	100%	88%	81%	100%
3-синф	67%	83%	67%	50%	100%
4-синф	92%	85%	100%	69%	100%
4-синф	83%	92%	75%	83%	100%
5-синф	93%	60%	93%	87%	100%
6-синф	86%	64%	86%	79%	100%
7-синф	100%	100%	78%	78%	100%
8-синф	88%	88%	100%	88%	100%
9-синф	100%	75%	50%	100%	100%
10-синф	93%	93%	100%	87%	100%

VI. Хулоса

Мақолада объектларнинг аниқ сонлар орқали берилган белгилар қийматлари асосида уларни кластерлаш ва синфлаштириш масаласи тадқиқ қилинди. Тадқиқот давомида объектлар ўқув танланма орқали бир неча синфларга ажратилган ва бу синфларни аниқлашда алгоритмлардан фойдаланилган. Алгоритмлар объектларни муҳимлик даражаларига қараб баҳолашда қўлланилди.

Аномал маълумотларни аниқлаш ва тузатиш жараёни алоҳида ўрганилган. Бу жараён маълумотларнинг ишончлилиги ва кластерлаш жараёнининг самарадорлигини оширишга ёрдам берди. Таклиф этилган алгоритмлар турли синфлаштириш ва кластерлаш масалаларида самарали натижаларни кўрсатган. KNIME муҳитидаги тажрибаларда, таклиф этилган алгоритм бошқа машҳур алгоритмларга нисбатан юқори аниқлик кўрсатди.

Таклиф этилган алгоритмнинг турли соҳаларда, хусусан, тиббиёт, молия ва маркетинг каби соҳалардаги амалиётларда қўлланилиши ва натижалари таҳлил қилинди. Таклиф этилган алгоритм объектларни синфлаштириш жараёнини самарали бошқариш имкониятини яратади ва турли соҳалардаги муаммоларни ҳал қилишда қўлланилиши мумкин.

Турли синфлаштириш алгоритмлари таҳлил қилинган, жумладан, KNN (K Nearest Neighbor), Random Forest, Naive Bayes ва Decision Tree алгоритмлари. Мақсад бу алгоритмларнинг самарадорлигини текшириш ва таклиф этилган алгоритм билан таққослашдан иборат бўлди. Таҳлиллар натижасида таклиф этилган алгоритм барча синфларда энг юқори аниқликни кўрсатди.

(cc) BY © Олламберганов Файзулла Фарход ўғли

Тадқиқ этилган алгоритмлар KNIME муҳитида синовдан ўтказилган. Турли хил объектларни, жумладан, қон босимига таъсир этувчи дори воситаларини синфлаштириш тажрибалари ўтказилди ва натижалар жадвал шаклида келтирилди. Ушбу алгоритмлар асосида амалга оширилган тажрибалар натижалари жадвал ва диаграммалар кўринишида тақдим этилди. Улар ҳар бир алгоритмнинг қўлланилишидаги аниқлик даражаларини кўрсатди.

Алгоритмларнинг асосий афзалликлари ва камчиликлари таҳлил қилинди. Таклиф этилган алгоритмлар катта ҳажмдаги маълумотларни кластерлаш ва синфлаштириш жараёнида энг самарали усул эканлиги кўрсатилди. K-means, Hierarchical кластерлаш ва Гаусс микси каби алгоритмлар билан таққосланганда, мақолада таклиф этилган алгоритм юқори аниқлик ва самарадорлик бериши кўрсатилди.

Якуний хулосада, таклиф этилган алгоритм объектларни белгиларнинг аниқ қийматлари ва синфлар шаклланишига қўшган ҳиссасига асосланган ҳолда синфлаштиришда самарали ечим эканлиги таъкидланди. Тадқиқот натижалари турли соҳаларда кенг қўлланилиши мумкинлиги тавсия қилинди. Келгусида ушбу алгоритмларни янада такомиллаштириш, уларни сунъий интеллект ва катта маълумотлар таҳлили билан интеграция қилиш орқали турли соҳалардаги самарадорликни ошириш бўйича тавсиялар берилди.

Хулоса қилиб айтганда, мақолада таклиф этилган алгоритм объектларнинг муҳимлик даражалари ва синфлар шаклланишига қўшган ҳиссасига асосланиб уларни кластерлаш ва синфлаштириш жараёнини самарали бошқариш имконини беради.

VII. Фойдаланилган адабиётлар

1. MacQueen, J. (1967). "Some Methods for Classification and Analysis of Multivariate Observations." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, pp. 281-297.
2. Johnson, S.C. (1967). "Hierarchical Clustering Schemes." Psychometrika, Volume 32, Issue 3, pp. 241-254. DOI: 10.1007/BF02289588
3. Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer. ISBN: 978-0387310732.
4. Mitchell, T.M. (1997). Machine Learning. McGraw-Hill. ISBN: 978-0070428072.
5. Iglewicz, B. & Hoaglin, D.C. (1993). How to Detect and Handle Outliers. Springer. ISBN: 978-0387978591.
6. Dunn, J.C. (1973). "A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters." Journal of Cybernetics, Volume 3, Issue 3, pp. 32-57. DOI: 10.1080/01969727308546059
7. Atoev, S., Nishanov, A., & Abdirazakov, F. (2021, November). Object Tracking Method Based on Kalman Filter and Camshift Algorithm for UAV Applications. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.
8. Нишанов, А. Х., & Холбоев, Б. М. (2022). Разработка и исследование методов определения информативных наборов признаков при распознавании одного типа явлений. in техника и технологии: пути инновационного развития (pp. 227-230).
9. Nishanov, A. K., Allamov, O. T., Ruzibaev, O. B., Abdullaev, A. S., & Allamova, S. T. (2021, November). An approach to finding the most optimal route in a dynamic graph. In 2021 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 01-05). IEEE.
10. Nishanov, A. K., Akbaraliev, B. B., & Djurayev, G. P. (2020, November). A symptom selection Algorithm based on classification errors. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.
11. Rakhmanov, A. T., Nishanov, A. K., Ruzibaev, O. B., & Shaazizova, M. E. (2020, November). On one method for solving the multi-class classification problem. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-4). IEEE.

© Олламберганов Файзулла Фарход ўғли

12. Nishanov, A. H., Akbaraliev, B. B., & Tajibaev, S. K. (2020, October). About one feature selection algorithm in pattern recognition. In World Conference Intelligent System for Industrial Automation (pp. 103-112). Cham: Springer International Publishing.
13. Nishanov, A. K., Ruzibaev, O. B., Chedjou, J. C., Kyamakya, K., Abhiram, K., De Silva, P., ... & Khasanova, M. A. (2020). Algorithm for the selection of informative symptoms in the classification of medical data. In Developments of Artificial Intelligence Technologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020) (pp. 647-658).
14. Nishanov, A. H., & Samandarov, B. S. (2015). Assessment model of monitoring and defining the completeness of course elements of information systems. Europaische Fachhochschule, (5), 56-58.
15. Nishanov. A. Kh, (1999). Turakulov Kh. A., Turakhanov Kh. V. A decision rule for identification of eye pathologies. Biomedical Engineering, 33(4), 178-179.
16. Камилов, М. М., Фазылов, Ш. Х., & Нишанов, А. Х. (1992). Метод выбора признаков с использованием критерия информативности фишеровского типа. Узбекский журнал «Проблемы информатики и энергетики», (2), 9.

© Олламберганов Файзулла Фарход ўғли