УДК 004.95

Бабомурадов О.Ж., Маматов Н С., Бобоев Л.Б., Отахонова Б.И.

Қарор дарахти алгоритмидан фойдаланиб матнларни таснифлаш

Аннотация. Таснифлаш объектни олдиндан маълум бўлган синфлардан бирига тегишлилигини аниқлайди. Матнларни таснифлаш компьютерли лингвистика масаласи бўлиб, бунда ҳужжатнинг мазмунига кўра уни олдиндан берилган бир неча рукнлардан бирига тегишлилигини аниқлаш амалга оширилади. Ҳозирги кунда матнларни таснифлашнинг кўплаб усуллари ишлаб чиқилган. Масалан, машинали ўқтиш, қарор дарахти, нейрон тўрлари, таянч векторлар ва бошқалар. Мазкур иш қарор дарахти усулидан фойдаланган ҳолда таснифлаш механизмини қуриш масаласи ечишга бағишланган бўлиб, унда ўзбек тилидаги матнлардан иборат ҳужжатларни таснифлаш учун қарор дарахти алгоритми атрибутлар қийматларини аниқлаш (*Information Gain, Gain Ratio, Gini index*), қарор дарахти алгоритмининг иш самарадорлигини ошириш ва тажрибавий тадқиқот натижалари келтирилган.

Калит сўзлар. белги, матн, ҳужжат, тўплам, қисм тўплам, қарор дарахти, усул, алгоритм, модел, атрибут, Джини индекси.

Кириш. Сўнгги йилларда матнли ахборотларнинг кескин ошиб бориши, маъноли ахборотларни ажратиб олиш, таҳлил қилиш ва таснифлаш каби масалаларни тезкор ечимини талаб қилади. Айниқса, глобал тармоқ контент базасининг ривожланиши гипер тезликка эришганлиги, ахборот истеъмолига бўлган эҳтиёжнинг инсон кундалик эҳтиёжлари таркибий элементлари орасида лидерлик мақомига эга бўлганлиги ва ахборот ресурслари маконидан энг зарурларини ажратиб олиш масаласининг долзарблигини оширмоқда.

Маълумки, матнларни таснифлаш матнли маълумотларни олдиндан берилган синфларга тегишлилигини назарда тутувчи маълумотларни тузилмалаштириш усулларидан бири ҳисобланади [1]. Матнли маълумотларни таснифлаш усуллари ахборотни қидириш ва матнларни ўқитиш билан боғлиқ. Мазкур икки ёндашувнинг умумий жиҳати ҳужжатни акс эттириш ва матнларни таснифлаш сифатини баҳолаш услубиятларида бўлиб, уларни фақат ўзига хос бўлган қидирув механизмига эга эканлиги уларнинг фарқли жиҳати ҳисобланади [2].

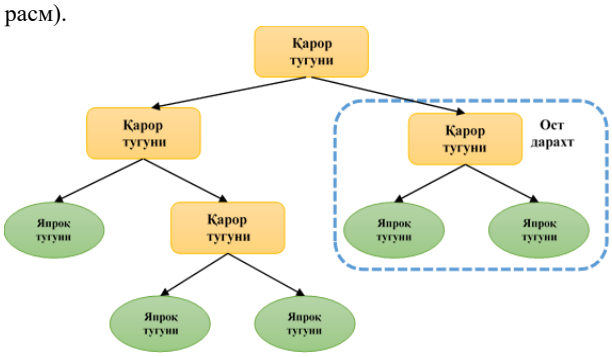
Матнларни таснифлаш масаласи ечиш бўйича бир нечта илмий гуруҳлар фаолият олиб боришаётганига қарамай, бу соҳадаги айрим саволлар ҳозиргача ўз ечимини топигани йўқ. Баъзи бир усулларнинг аниқлиги априор бўлиб, йўл қўйилиши мумкин бўлган хатоликлар миқдори ҳамда матнли маълумот тузилмаси (синфлар сони, синфларнинг ҳажми ва бир жинслилиги, синфлараро «чегара»нинг аниқлилиги)га боғлиқдир. Матнли маълумотларга ишлов беришда бир қанча муаммолар юзага келади [3]. Масалан, таснифлаш учун фойдали бўлган информатив белгилар миқдорини аниқлаш, матнли маълумотларни қайта ишлаш ва таҳлил қилишга кетадиган ҳисоблаш вақтини минималлаштириш кабилар. Бундан ташқари, ҳосил қилинадиган «ҳужжат-атама» матрицаси ўта соддалаштирилган, атамаларнинг бошқа ҳужжатларда учраш миқдори кўплиги ҳам турли муаммоларни келтириб чиқаради. Тартибланган ахборотлардан фарқли равишда тартибланмаган ахборот ягона матнли форматга эга бўлмайди, бу эса матнли маълумотга ишлов бериш ва таҳлил қилиш учун матнли ҳужжатларга ишлов беришнинг комплекс моделини ишлаб чиқишни талаб қилади [4-6]. Шунинг учун матнли ҳужжатларни таснифлаш масаласини ечишда комбинацияланган дастлабки ишлов бериш, ўқув танланма шакллантириш, синфларни самарали ташкил этиш ва таҳлил қилиш ёндашувларни биргаликда қўллаш мақсадга мувофиқ ҳисобланади.

Матнли ҳужжатларни таснифлаш масаласи. Таснифлаш алгоритми бирор-бир $D=\{d_i\}$ ҳужжатлар тўпламида амалга оширилаётган ва бу ҳужжатлар тўплами кесишмайдиган қисм тўплам синфларга ажратилган бўлсин, яъни

$$C=\{C_i\},\ \bigcup_{d\in C_i}^{\forall i} d=D,\ C_i\bigcap C_j=0,\ (i\neq j)$$

Таснифалаш масаласи кирувчи ҳужжатни мос синфга тегишлилигини аниқлаш ҳисобланади. Ҳар бир d элементга $d=\{X_i\}$ белгилар мажмуаси мос қўйилади ва таснифлаш алгоритми ёрдамида матнли ҳужжатни олдиндан берилган синфларнинг бирига тегишлилигини аниқланади. Қуйида қўйилган масалани ечишнинг қарор дарахти усулига асосланган алгоритм баён этилган.

Қарор дарахти алгоритми. Қарор дарахти бу блок схемага ўхшаш тузилишида бўлиб, унда белги (ёки арттрибут) ички тугунни, шох қарор қоидасини ва ҳар бир япроқ тугун натижасини ифодалайди [7]. Қарор дарахтидаги энг юқори тугун илдиз тугун деб аталади. У атрибут қийматлари асосида ажратишни амалга оширишда қўлланилади. Бунда рекурсив ажратиш деб номланган усул орқали маълумотлар рекурсив ажратилади. Бундай блок-схемага ўхшаш структура қарор қабул қилишда қўлланилади ва ушбу блок-схема диаграммаси визуализациялаш имконини бериб, у инсон даражасида фикрлашни осонлаштиради. Шунинг учун қарор дарахтлари тушунишда мураккабликюзага келмайди (1-расм).



1-расм. Қарор дарахти алгоритми блок-схемаси

Қарор дарахти алгоритми машинали ўқитишнинг оқ қути тури ҳисобланиб, у нейрон тўрлари каби қора қути алгоритмларидан фарқли равишда ички қарор қабул қилиш мантиғини ошкор қилади. Бунда ўқитиш вақти эса нейрон тўрларига нисбатан тезроқ амалга оширилади ва катта ўлчамли маълумотларни яхши аниқлик билан бошқариш имкониятига эгадир.

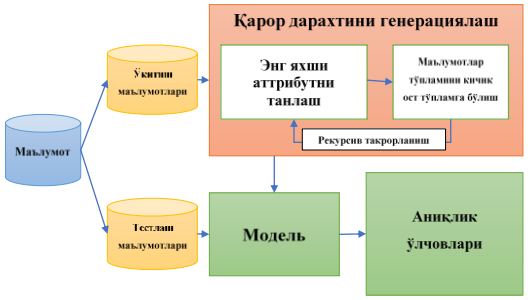
Қарор дарахти алгоритмида қуйидаги қуйидаги босқичларда бажарилади:

1. Атрибут танлаш. Бунда ўлчовларидан фойдаланиб энг яхши атрибутлар танланади.

2. Атрибутни қарор тугуни сифатида белгилаш ва маълумотлар тўпламини кичик қисм тўпламларга ажратиш.

3. Юқоридаги жараёнлар қуйидаги ҳолатлардан бирига келгунига қадар такрорлаш орқали дарахтни қуриш амалга оширилади:

- барча тўпламлар бир хил атрибутга тегишли.
- бошқа атрибутлар йўқ.
- бошқа мисоллар йўқ.



2-расм. Қарор дарахти алгоритми

Атрибут танлаш ўлчовлари. Атрибут танлаш ўлчовлари маълумотларни ажратиш мезонини танлаш учун муҳим ҳисобланади. Бундан ташқари, ажратиш қоидалари сифатида ҳам танилган, чунки у бизни муайян тугундаги тўпламларни аниқлашга ёрдам беради. Атрибут танлаш мезонлари ҳар бир белгига рейтинг беришни таъминлайди. Энг яхши рейтингли атрибут эса ажратиш атрибути сифатида олинади. Ҳозирги кундаги фойдаланилаётган энг машҳур танлаш ўлчовлари Information Gain, Gain Ratio ва Gini Index ҳисобланади.

Information Gain. Шаннон энтропия контсепциясини кашф этган, бу контсепция орқали кириш тўпламининг тозалик даражасини ўлчанади. Физика ва математикада энтропия тизимдаги тасодифлик ёки аралашма деб аталади. Ахборот назариясида у мисоллар гуруҳидаги тозалик даражасини ифодалайди. Information Gain - бу энтропиянинг пасайиши. Information Gain берилган атрибут қийматларида маълумотлар тўпламини ажратишдан олдин ва кейинги энтропия орасидаги фарқни ҳисоблайди. ID3 (Iterative Dichotomiser) қарор дарахти алгоритми Information Gain усулидан фойдаланади.

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

бу ерда p_i D тўпламнинг C_i синфда тегишлилик эҳтимоли.

$$Info_A(D) = \sum_{j=1}^{V} \frac{|D_j|}{D} Info(D_j)$$

$$Gain(A) = Info(D) - Info_A(D)$$

бу ерда, $Info(D)$ ифода D тўплам қисм тўплами синф ёрлиғини аниқлашдаги ахборотнинг ўртача миқдори, $|D_j|/|D|$ ифода j - қисм вазни, $Info_A(D)$ А томонидан ажратишга асосланган D тўплам қисм тўпламини таснифлаш учун талаб қилинадиган ва кутилган ахборот.

Энг юқори ахборот фойдасига эга атрибут A, яъни $Gain(A)$, N тугунда ажратиладиган атрибут сифатида олинади.

Gain Ratio. C4.5, ID3 тизимини такомиллаштирилгани бўлиб, Gain Ratio деб номланган. Одатда мазкур усулининг кенгайтирилган кўринишини қўлланилади. Gain ratio усулидан ахборотни ажратишда фойдаланиб, унда information gain усулини нормаллаштириш орқали юзага келадиган хатоликни аниқлайди.

$$SplitInfo_A(D) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{D} \right)$$

бу ерда, $|D_j|/|D|$ ифода j - қисм вазни, V A атрибутдаги дискрет қийматлар сони.

Gain ratio қуйидаги формула орқали ҳисобланади:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

Юқори gain ratio қийматига эга атрибут ажратиш атрибути сифатида танланади.

Gini index. Бошқа бир қарор дарахти алгоритми CART (Classification and Regression Tree) ажратиш нуқталарини яратиш учун Gini index усулидан фойдаланади.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

бу ерда D тўплам ичидаги қисм тўплам p_i - C_i синфга тегишлилик эҳтимоли.

Gini индекси ҳар бир атрибутнинг бинарликка тешириш орқали ҳар бир бўлакни ноаниқ оғирликлари йиғиндисини ҳисоблайди. Агар D маъумотнинг A атрибутдаги бинарлиги D_1 ва D_2 бўлса, у ҳолда D нинг Gini индекси қуйидагича ҳисобланади:

$$Gini_A(D) = \frac{|D_1|}{D} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

Агар дискрет қийматли атрибутлар бўлса, танланган атрибут минимал Gini индекс берувчи қисм тўпламни таъминласа, у ҳолда у ажратувчи атрибут сифатида олинади. Агар атрибутлар давомий қийматли бўлса, қуйидаги стратегия бўйиса ҳисобланади.

$$\Delta Gini(A) = Gini(D) - Gini_A(D)$$

Бунда минимал Gini индексга эга бўлган атрибут ажратувчи атрибут сифатида олинади ва қарор дарахти таснифлагичини қуриш қуйидаги босқичларда амалга оширилади.

- Маълумотларни юклаш.
- Белгиларни танлаш.
- Маълумотларни бўлаклаш.
- Моделни қуриш.
- Модел аниқлигини ўлчаш.
- Қарор дарахтини чизиш.

```
from sklearn.datasets import load_files
uza = load_files('.\\uza')
categories = uza.target_names
print(categories)
```

```
['Жамият', 'Иқтисодиёт', 'Маданият', 'Спорт', 'Технологиялар', 'Фан']
```

```
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer()
print(count_vect)
```

```
CountVectorizer(analyzer='word', binary=False, decode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_accents=None, token_pattern='(?u)\\b\\w\\w+\\b',
        tokenizer=None, vocabulary=None)
```

```
from sklearn import tree
from sklearn.model_selection import cross_val_score

X = count_vect.fit_transform(uza.data)
y = uza.target

clf = tree.DecisionTreeClassifier(criterion="gini")
scores = cross_val_score(clf, X, y, cv=5)

print("Aniqlik: %0.2f (+/- %0.2f)" % (scores.mean(), scores.std() * 2))
```

```
Aniqlik: 0.64 (+/- 0.16)
```

Хулоса

Тадқиқот натижалари баъзи усуллар аниқлигини априор эканлигини кўрсатди. Бу йўл қўйилиши мумкин бўлган хатоликлар миқдори ҳамда матнли маълумот тузилмаси (синфлар сони, синфларнинг ҳажми ва бир жинслилиги, синфлараро «чегара»нинг аниқлилиги)га боғлиқдир. Матнли маълумотларга ишлов беришда бир қанча муаммолар юзага келади ва бу муаммоларнинг айримлари қарор дарахти усулидан фойдаланган ҳолда таснифлаш механизмини қуриш масаласи ечиш орқали ҳал этилади. Ўтказилган тажрибаларда ўзбек тилидаги матнлардан иборат ҳужжатларни таснифлаш қарор дарахти алгоритми атрибутлар қийматларини аниқлаш (Information Gain, Gain Ratio, Gini index) орқали амалга оширилди. Бунда маълумотлар Ўзбекистон Миллий ахборот агентлиги давлат расмий ахборот манбаидан олинган олтита категорияга тегишли бўлган 600 та янгилик олинган бўлиб, қарор дарахти алгоритми асосланган дастур ишлаб чиқилди ва таснифлаш аниқлиги 64% ташкил этди.

Фойдаланилган адабиётлар

[1] W. Buntine. A theory of classification rules. 1992.

[2] S.Murthy. Automatic construction of decision trees from data: A Multi-disciplinary survey.1997.

[3] J. Ross Quinlan. C4.5: Programs for Machine learning. Morgan Kaufmann Publishers 1993.

[4] Machine Learning, Neural and Statistical Classification. Editors D. Mitchie et.al. 1994.

[5] К. Шеннон. Работы по теории информации и кибернетике. М. Иностранная Использованная литература, 1963.

[6] С.А. Айвазян, В.С Мхитарян Прикладная статистика и основы эконометрики, М. Юнити, 1998.

[7] Brett Lantz. Machine Learning with R. Packt Publishing, Birmingham - Mumbai, 2013.

Бабомурадов Озод Жўраевич - – техника фанлари доктори, Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети АТДТ кафедраси мудири.

Тел.: (+99897) 403-76-11 Факс: (0371) 237 62 48

E-mail: ozod_b_76@mail.ru

Маматов Нарзилло Солиджонович – техника фанлари доктори, Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети ҳузуридаги Ахборот-коммуникация технологиялари илмий-инновацион маркази етакчи илмий ходими.

Тел.: (+99897) 403-56-22 Факс: (0371) 237 62 48

E-mail: m__narzullo@mail.ru

Бобоев Лочинбек Боймуротович – Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети ҳузуридаги Ахборот-коммуникация технологиялари илмий-инновацион маркази таянч докторанти.

Тел.: (+99890) 115-10-10 Факс: (0371) 237 62 48

E-mail: replytolochin@gmail.com

Отахонова Бахрихон Ибрагимовна– Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети таянч докторанти.

Тел.: (+99891) 659-44-89 Факс: (0371) 238 65 07

E-mail: bahrixon@mail.ru

Babomuradov O.J., Mamatov N.S., Boboyev L.B., Otaxonova B.I.

Classification of texts using decision trees algorithms

Annotation. The classification determines whether the object belongs to one of the previously known classes. Classification of texts is a matter of computer linguistics, to determine whether the document refers to one of the few chapters previously given. Currently, many ways of classifying texts have been developed. For example, machine learning, decision tree, neural nets, base vectors, and more. The present work is devoted to solving the problem of constructing a classification mechanism using the decision tree method.

Keywords: character, text, document, collection, partition, decision tree, method, algorithm, model, attribute, Gini index.