OSZTÁLYOZÁSI FELADATOK A KÉRDÉSGENERÁLÁSI MINTARENDSZERBEN

CLASSIFICATION TASKS IN THE SYSTEM OF QUESTION GENERATION MODEL

Bednarik László, dr. Kovács László***

ABSTRACT

Automatic question generation on text documents has become a more and more important field of the recent researches of artificial intelligence. Therefore analysing text documents and gaining useful information from them requires more and more effective methods. One stage in our results gained in this field is to classify elements of documents. In the concept described below we realised this task by building on the neural net based learning theory. The new idea in this method is to fit the general theory to the specific features of the task to be carried out. The correctness of the model was proven by an application implemented in Java language that supports to demonstrate teaching with involving specialist and results of classification in an interactive graphical interface.

1. BEVEZETÉS

A dokumentumok szavainak klaszterezése révén előálltak mindazok az objektív koordináták, melyek az emberi tényezőktől függetlenül leírják minden szó elhelyezkedését a definiált hét dimenziós objektív térben. A cél egy olyan szabályrendszer feltárása, mely képes az objektív térben adott szavakat (mint pontokat) automatikusan elhelyezni az igényeink szerint definiált szubjektív térben. Az osztályozási feladat magját egy előrecsatolt háromrétegű neurális-hálóval valósítottuk meg.

2. AZ OSZTÁLYOZÁS IRODALMI ÁTTEKINTÉSE

A szöveges adatok egyik jellemzője, hogy az információt strukturálatlan vagy gyengén strukturált formában tartalmazzák [8]. A strukturálatlan adatok kezelésére részben megoldást jelent a strukturált tárolásuk, azaz a dokumentumok hierarchikus rendszerbe (taxonómiába) való rendszerezése [1].

Az osztályozás célja, hogy az objektumok halmazán egy tulajdonsághoz osztálykódot rendelő függvényt alkossunk meg, amely a tanuló mintára adja vissza az ismert osztálykódot [4].

Dokumentumok osztályozására a legelterjedtebben alkalmazott módszereket a döntési fa, a legközelebbi szomszéd (k-NN), a Bayes háló, a szupportvektor gép, a *neurális háló*, valamint a *CPN háló* algoritmusainak feladatspecifikus továbbfejlesztései alkotják. Az említett módszerek közül kettő kerül bemutatásra.

2.1. Neurális hálózat alapú osztályozó

Az osztályozást végző modellt úgy terveztük meg, hogy képes legyen bármilyen tématerülethez tartozó szöveg kezelésére. Az egyik legelterjedtebben alkalmazott módszer a neurális hálóval végzett tanulás, melynek elvét McCulloch és Pitts publikálta 1943-ban. A neurális hálók irányított kapcsolatokkal összekötött egységekből (neuron) állnak. A j jelű neuronból az i jelű felé vezető kapcsolat hivatott a j jelű neuron aktivációs állapotát (jele: a_j) az i jelű neuronhoz továbbítani. Minden egyes kapcsolat rendelkezik egy hozzá társított $W_{j,i}$ numerikus súllyal, ami meghatározza a kapcsolat erősségét és előjelét. Minden egyes neuron először a bemeneteinek egy súlyozott összegét számítja ki az 2.1. összefüggés alapján.

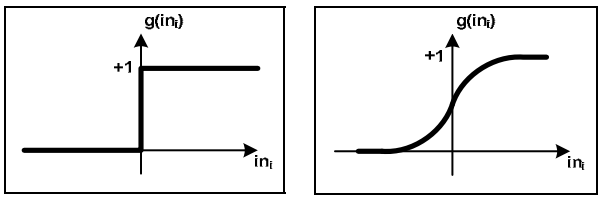
$$in_i = \sum_{j=0}^{n} W_{j,i} * a_j \tag{2.1}$$

Az i-edik neuron aktivációs állapota a bemeneteinek súlyozott összegére alkalmazott g jelű aktivációs függvénnyel kerül meghatározásra az 2.2. összefüggés szerint [5].

$$a_i = g(in_i) = g\left(\sum_{j=0}^{n} W_{j,i} * a_j\right) \tag{2.2}$$

A 2.1. ábra két lehetséges aktivációs függvényt mutat be, a küszöbérték függvényt, és a szigmoid függvényt [3].

Mindkét függvény küszöbpontja az $x = 0$ értéknél van. A megoldott feladat során ezt az értéket minden neuron esetén $x = 0$-ra állítottuk be ám lehetőséget biztosítottunk a későbbi módosíthatóságára.



2.1. ábra. Két lehetséges aktivációs függvény a neuron kimenetének meghatározásához

Az információk áramlási iránya alapján a neurális hálózatok két fő csoportba sorolhatóak. Ezek az előrecsatolt hálók (*feed-forward network*) [7] illetve a visszacsatolt hálók (*recurrent network*) [6].

Az előrecsatolt neurális hálózatok megkülönböztethetőek egymástól a rétegek száma, az egyes rétegekben lévő neuronok száma, valamint a neuronok közötti kapcsolatok alapján is. Általánosságban egy neurális hálózat három különböző feladatot ellátó rétegből épül fel. Ezek a bemeneti réteg, rejtett réteg, kimeneti réteg [9].

A CPN (*Counter-Propagation Network*) hálózat alapú osztályozók lényegében a neurális hálózatok egy feladatspecifikus alkalmazásai. A CPN hálózat bemeneti rétege itt is az osztályozandó objektumok leírását tartalmazza. A rejtett rétegben (Kohonen réteg) lévő neuronok ebben az esetben a klasztereket reprezentálják. A kimeneti réteg neuronjai pedig azokat az osztályokat írják le, amelyekbe a bemeneten lévő objektumok besorolásra kerülnek. A neuronok bemeneti függvénye a 2.3 összefüggés alapján kerül meghatározásra.

$$s = \sum w_i * s_i \tag{2.3}$$

Az s_i jelöli a bejövő jelet, w_i pedig az él súlyát. Az s érték megfelel a w és az s vektorok skaláris szorzatának, ami a közbezárt szög koszinuszával arányos. A tanulási folyamat során a győztes neuron mindig közelebb kerül a bemenetre kapcsolt objektumhoz.

3. AZ OSZTÁLYOZÁSI FELADAT MEGVALÓSÍTÁSA

3.1. Az algoritmus bemeneteinek és kimeneteinek megtervezése

A megvalósításra került osztályozó algoritmus a dokumentum szavainak objektív koordinátái illetve a tanító minta alapján állítja elő szavak szubjektív koordinátáit. Az objektív koordináták megválasztásánál a szavak olyan tulajdonságainak a megtalálása volt a célunk, melyekkel jól jellemezhetőek a vizsgált szavak, ugyanakkor meghatározásuk nem ró lényeges számítási többletet az osztályozást végző programmodulra. Ezek alapján minden szó objektív vetülete hét koordinátával került definiálásra. Az alkalmazott objektív koordináták [2] és lehetséges értékei: szófaj [-1, 0, 1, 2, 3,4, 5, 6, 7], ragozottság [-1, 0, 1], szó száma a dokumentumban [1 ... előford. szám], szó száma a mondatban [1 ... előford. szám], az adott szó helye a mondatban [1 ... mondat szavai], klaszter sorszáma [-1, +1] egész szám, szó távolságátlaga [0.0 ... 1.0] valós szám.

Az osztályozást végző algoritmus kimenetét a szavaknak az emberi tényezőktől függő (szubjektív) koordinátái alkotják. Az osztályozás eredményeként szolgáló szubjektív koordináták: relevancia, specialitás, értelmesség, nehézség és kérdésként kiemelésre kerüljön-e a szó.

3.2. Az alkalmazott neurális hálózat struktúrája

Az osztályozás egy előrecsatolt háromrétegű (egy belső réteg) neurális háló alkalmazásával került megvalósításra. A neuronok közötti kapcsolatok súlyértékei [-1.0, ..., 1.0] tartományba tartozó valós számok. A háló neuronjainak aktiváltsági állapota eltolás nélküli küszöbérték függvénnyel lett reprezentálva, mely a bemeneti jeleket súlyozottan összegző bemeneti függvény értékét a [0, 1] egész intervallumra képezi le az 3.1. összefüggés alapján.

$$a_i = 1, ha \sum_{j=0}^{n} W_{i,j} * a_j > 0.0$$
$$a_i = 0, egyébként \tag{3.1}$$

A háromrétegű neurális hálózat első rétegét a bemeneti neuronok alkotják, melyeknek feladata, hogy az objektív koordináták értékeit a hálózat belső (rejtett) rétegében lévő neuronok felé továbbítsák. A kimeneti neuronok az osztályozandó szavak szubjektív koordinátáira vannak leképezve. A neurális hálózatok építésének egyik kulcskérdését jelenti a belső rétegben lévő neuronok számának meghatározása. Az optimális értéktől alacsonyabb számú neuron alkalmazása esetén a

hálózat nem lesz képes a feladat megtanulásához szükséges mennyiségű információ tárolására.

Jelen feladat esetében a belső rétegben helyezendő neuronok számának pontos meghatározása továbbfejlesztési lehetőségként jelentkezik. Első közelítésként számukat azonosnak vettük a kimeneti rétegben lévő neuronok számával. Eredményként a vizsgált dokumentumok szinte mindegyikénél sikerült 0.01 pontossággal megközelíteni a tökéletes tudásszintnek számító 1.0 értéket. Ez tehát azt támasztja alá, hogy a közbenső rétegben lévő neuronok száma nem kevesebb az optimális értéktől.

3.3. Az alkalmazott neurális hálózat tanulási folyamata

A neurális hálózat tanítása a felügyelt tanítás szabályai szerint lett megvalósítva. A tanulás első lépéseként az előállított hálózat neuronjai közötti kapcsolatok súlyértékei véletlenszerűen beállításra kerülnek a [-1.0, ..., 1.0] intervallumból. A véletlen számokat generáló algoritmus az adott intervallumon egyenletes eloszlás szerint állítja elő a véletlen számokat.

A súlyok kezdeti értékének beállítását követően az algoritmus sorra veszi a tanítóminta szavait és egymás után a hálózat bemenetére kapcsolja azokat. Ezt egy speciális illesztőmodulon keresztül végzi el, mely a szó minden objektív koordinátájának értékéhez meghatározza a neurális hálózatnak azt a neuronját a dinamikusan felépített bemeneti rétegben, amely az adott koordináta értékét hivatott reprezentálni. Az illesztés eredményeként az adott szónak az objektív térben való elhelyezkedését megadó koordinátákat reprezentáló bemeneti neuronok aktív állapotba kerülnek. A rejtett rétegben lévő neuronok aktiváltsági állapotának beállítását követően kerül sor a hálózat kimeneti rétegében lévő neuronok aktiváltsági állapotának meghatározására. Ehhez a 2.3. összefüggés kerül alkalmazásra, de ebben az esetben a_i az i jelű külső rétegbeli neuron aktiváltsági állapotát, a_j pedig a j jelű rejtett rétegbeli neuronét jelöli. Összefoglalóan az i jelű – kimeneti rétegben lévő – neuron aktiváltsági állapota a bemeneti rétegben lévő neuronok aktiváltsági állapotának függvényében a 3.2. összefüggés alapján határozható meg.

$$a_i = g\left(\sum_{j=0}^{n} W_{j,i} * g\left(\sum_{k=0}^{m} W_{k,j} * a_k\right)\right) \qquad (3.2)$$

ahol

a_k: a k jelű bemeneti rétegbeli neuron aktiváltsági állapota,

a_i: az i jelű kimeneti rétegbeli neuron aktiváltsági állapota,

m: a bemeneti rétegben lévő neuronok száma,

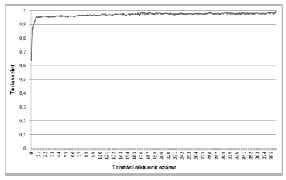
n: a rejtett rétegben lévő neuronok száma,

$W_{k,j}$: a k jelű bemeneti rétegbeli neuron és a j jelű rejtett rétegbeli neuron közötti kapcsolat erőssége,

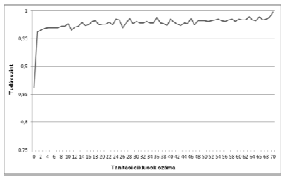
$W_{j,i}$: a j jelű rejtett rétegbeli neuron és az i jelű kimeneti rétegbeli neuron közötti kapcsolat erőssége,

g: aktivációs függvény, mely 1-es értéket reprezentál, ha a paramétereként kapott összeg pozitív, 0-át egyébként.

A kimeneti rétegben lévő neuronok aktiváltsági állapotai írják le a neurális hálónak a bemenetként kapott szó objektív koordinátáira adott válaszát a szubjektív térben. A tanulás akkor fejeződik be, ha a tanítómintában szereplő szavak osztályozásának átlagos hibája már nem haladja meg a megengedett mértéket, vagy a tanítóminta szavainak ismételt tanulása már eléri az engedélyezett ismétlési számot. A 3.1. ábra a tanulás *súlyegység* paraméterének különböző értékei esetén a háló tudásszintjében tapasztalt változást mutatja a tanítóminta ismétlési számának függvényében.



súlyegység: 0.01



súlyegység: 0.1

3.1. ábra: A tanulás súlyegység paraméterének különböző értékei

A 3.1. ábrán szemléltetett tesztek alapján megfigyelhető, hogy kisebb súlyegység esetén a tanulás sokkal pontosabban tart a célállapot felé. A neurális háló a betanulását követően a tanítómintában szereplő szavak esetén is megőrizte a paraméterként meghatározott elvárt pontosságot.

3.4. A tanulás eredményeinek újrahasznosítása

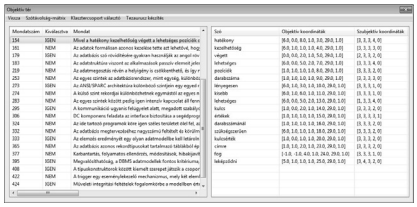
Az általunk alkalmazott neurális háló struktúrája mindig dinamikusan alkalmazkodik az éppen beolvasott dokumentum szerkezetéhez.

Az ismertetett módszer kiegészítéseként implementálásra került egy másik tudás-újrahasznosítási stratégia is. Ebben nem a betanult neurális hálózat adatait, hanem az osztályozás eredményéül szolgáló mondatokat menti el a rendszer. A megvalósított módszer lényege, hogy az osztályozott mondatok az

osztályozás eredményéül szolgáló szubjektív koordinátáikkal együtt egy adatbázisban kerülnek eltárolásra.

4. AZ ELKÉSZÍTETT ALKALMAZÁS BEMUTATÁSA

Az 3.1. pontokban ismertetett osztályozási feladatot megvalósító modell alapján elkészítettünk egy Java nyelven implementált alkalmazást, mely grafikus interaktív felhasználói felülete révén támogatást nyújt mind a szakértő bevonásával végzett tanítás, mind pedig az osztályozás eredményeinek demonstrálása területén. A 4.1. ábrán látható képernyőn a szakértő kiválaszthatja azokat a mondatokat, amelyek a neurális háló tanítására szolgáló tanítómintát fogják képezni.



4.1. ábra. A tanítóminta mondatainak kiválasztására szolgáló dialógusablak

A tanítóminta mondatainak beállítani kívánt szavain jobb egérgombbal kattintva megjelenik egy dialógusablak, melyben a szakértő megadhatja a kiválasztott szónak a szubjektív térben való elhelyezkedését leíró koordinátákat. A tanítóminta szavainak a szubjektív térben való elhelyezését követően a neurális háló tanítása a ***Tanítás*** menüpont kiválasztásával indítható el.

A tanulás befejezésével a neurális háló a megtanult hipotézisfüggvény alkalmazásával automatikusan kitölti a tanítómintába nem tartozó mondatok szavainak szubjektív koordinátáit.

5. AZ OSZTÁLYOZÁS EREDMÉNYEINEK ÖSSZEFOGLALÁSA

A dokumentum szavainak osztályozását végző algoritmus eredményeként megvalósításra került a szavak szubjektív vetületét reprezentáló koordináták automatizált feltárása, valamint megalapozott iránymutatást adtunk a szöveg mondataiból kérdésként kiemelendő szó meghatározására. Ezen lépések képezik alapját a kérdésként kiemelt szó lehetséges helyettesítésére felkínálandó szavak automatikus generálásának, valamint a kérdésre adott válasz értékelésének.

6. KÖSZÖNETNYÍLVÁNÍTÁS

A bemutatott kutató munka a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

7. IRODALOM

[1] BLUMBERG R., SHAKU A.: The problem with unstructured data, DM Review*,* 1 February. 2003b

[2] CSENDES D., CSIRIK J., GYIMÓTHY T.: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (LINC 2004) Geneva, Switzerland, 23-29 August, pp. 19-23. 2004

[3] DAGAN I., KAROV Y., ROTH D.: Mistake-driven learning in text categorization, In proc. of EMNLP-97, 2nd Conf. on Empirical Methods in Natural Language Processing, pp. 55-63, Providence, USA. 1997

[4] FABRIZIO S.: Machine Learning in Automated Text Categorization, ACM Computing Surveys*,* 34(1): 1-47. 2002

[5] MC.CULLOCH W. S., PITTS W.: A logical calculus of the ideas immanent in nervous activity, Bulletin of Mathematical Biophysics*,* 5:115-137. 1943

[6] RAMASUNDARAM R., VICTOR S. P.: Text Categorization by Backpropagation Network, International Journal of Computer Applications (0975 – 8887), Volume 8– No.6, 2010

[7] ROBERT H. N.: Theory of the Backpropagation Neural Network, Washington, DC, USA, 1989

[8] RUDIN, K., CRESSY D.: Will the Real Analytic Application Please Stand Up? DM Review, (13)3, pp. 30-34. 2003

[9] STUART R., PETER N.: Mesterséges intelligencia modern megközelítésben, ISBN 963 545 411 2, Panem könyvkiadó, Budapest, 2005

[10] TIKK D.: Szövegbányászat, Typotex, Budapest, 2007