FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA

Kolozsvár, 1997, március 21-23

BESZÉDFELISMERÉS NEURONHÁLÓZATTAL

Buzogány Endre

Bevezetés

A Mesterséges Intelligencia kutatások egyik, jelentősen fejlődő területe az emberi agy magasabb szintű műveleteinek megvalósítása számítógéppel. Ilyen például képek felismerése vagy a beszédfelismerés és szintézis.

Az első digitális számítógépekkel építések már felvetődött a kérdés: mikor lesz képes arra, hogy az emberi teljesítményeket elérje? A mai számítógépek processzor-sebessége ugyan nagyságrendekkel túlszárnyalja az emberi agy neuronjainak működési sebességét, már ami az elemi műveletek végrehajtási sebességét illeti, mégis egyes esetekben teljesítményeik messze elmaradnak.

Természetes dolog, hogy a mai számítógép- és programtervezők felhasználják a biológiai modellt annak érdekében, hogy az alkalmazások teljesítménye minél jobb legyen. Az emberi neuron felépítéséből kiindulva olyan mesterséges neuron-modellt hozunk létre, melynek egyes tulajdonságai hasonlítanak az emberi neuronra: tanulási képesség, zaj- és hibatűrés. A mesterséges neuronháló (a továbbiakban MNH) jelentős alkalmazási területe az emberi beszéd felismerése. A dolgozat első részében a feldolgozás azon paramétereit mutatjuk be, amelyek a beszédelemzésben szerepet játszanak, a második rész pedig egy felismerő rendszer felépítését mutatja be.

A beszédhangok ábrázolása

Az ábrázolás fogalma a kommunikációs kutatások egyik alapvető problémája, mert minden esetben felmerül az ilyen típusú problémáknál. Megkülönböztetünk analóg és digitális ábrázolást. A számítógépes feldolgozás, tárolás, beszéd-szintézis esetén a digitális formában dolgozzuk fel. Az emberi hang felfogható úgy, mint egy valós, véges $s(t)$ jel, melynek energiája véges, és nem stacionárius. A szerkezete komplex, és időben változik: egyes esetekben periodikus (pontosabban kváziperiodikus), más esetekben véletlenszerűen változhat.

A feldolgozás célja az, hogy a lényeges összetevők megmaradása mellett, minél hatékonyabb formát kapjunk. Alapvetően két fajta feldolgozási módot különböztetünk meg: amplitúdó szerinti és frekvencia

szerinti feldolgozást. Az első csoportba tartozik a maximális amplitúdó vizsgálata, a null-átmenetek száma, a hang energiája. Ezek viszonylag kis számításigényűek, és általában a beszédegységek (fonémák, szavak) elkülönítésére használják.

A frekvencia doméniumban a leggyakrabban a *Diszkrét Fourier Transzformált* (DFT) használják. Ez megadja a jel összetevőit a frekvencia függvényében. Ha az intervallumot felosztjuk részintervallumokra, majd ezekre alkalmazzuk a transzformációt, megkapjuk a hangjelenség spektrogramját, mely az összetevő komponensek amplitúdóját mutatja a frekvencia függvényében.

A Fourier transzformált egy módszert ad arra, hogy az adott jel összetevőit meghatározzuk. Egy kiejtett szó átlagos hossza 0,3 és 0,5 másodperc között van. ezért 30 intervallumra osztottuk a jelet. Minden ilyen komponensre 26 frekvencia összetevőt vettünk figyelembe. Tehát a transzformációval egy jelből 780 komponensű valós vektort kaptunk, mely alkalmas a felismerés céljára.

Ekkor csak az a probléma merül fel, hogy az intervallum végpontjain a jelfüggvény szakadása miatt irreális frekvenciák is megjelennek a spektrumban. Ennek kiküszöbölésére a vizsgált jelet egy ablakfüggvénnyel megszorozzuk, majd az így kapott függvényt transzformáljuk. Az alkalmazás által használt ablakfüggvény a Hamming-ablak, melynek alakja:

$$h(n) = \begin{cases} 0,54 - 0,46 * \cos \frac{2\pi i}{N} , & \text{ha } 0 \leq i \leq N \\ 0, & \text{ellenkező esetben} \end{cases}$$

ahol N a minták száma az adott intervallumon.

Felismerés neuronhálózattal

Az emberi neuron mintájára egy *mesterséges neuron* modellt dolgoztak ki. Ez bizonyos számú bemeneti egységből, egy feldolgozó egységből és egy kimenetből áll (1.ábra)

1.ábra A mesterséges neuron

Az elemi neuronokat rétegekbe szervezzük, és ezt a rendszert nevezzük *neuronhálózatnak*. A neuronhálót tanítani kell. Erre a célra tanulópárnak nevezett kettősöket állítunk össze, mely a jelből és a

várt kimenetből áll. A felismerés céljára egy neuronháló- rendszert alkalmaztunk. Ez n darab elemi hálózatból áll, melyet egy kapunak nevezett hálózat irányít (2.ábra)

2.ábra. Moduláris neuronhálózat

Egy bemenő adatra minden elemi hálózat meghatározza a kimenetet, majd a kapu eldönti, hogy az n hálózat esetén melyiknek milyen része van a rendszer összkimenetének meghatározásánál.

A rendszer összkimenetét az $Y = \sum g_i O_i$ képlet adja, ahol O_i az i-edik expert kimenete, g_i pedig az i-edik hálózat eredményének súlyértéke: $g_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$, ahol x_i a rendszer súlyozott kimenetei. A g_i kapu-értékekre érvényes a következő képlet: $\sum_i g_i = 1$. Ebben az esetben a g_i értéket úgy is felfoghatjuk, mint az i-edik neuron kiválasztásának valószínűségét.

A felismerő alkalmazás felépítése

Az alkalmazott neuronhálók egy rejtett réteget tartalmazó, feedforward hálózatok. A kapu-hálózat Kohonen típusú önszervező háló. A tanuló és tesztadatok a számokat tartalmazzák 0-tól 9-ig. Az alkalmazás szerkezetét a 3. ábra mutatja.:

3.ábra Hangfelismerő rendszer felépítése

Mivel a cél egy beszélőtől független rendszer létrehozása volt, ezért az adatokat több személytől gyűjtöttük : 10 szó x 8 beszélő x 16 kiejtés. Egy másik teszteléskor egy adott beszélővel 10 szó x 40 kiejtést vettünk. Ekkor csak egyetlen neuronháló végezte a felismerést

A *tanuló és tesztadatokat* a következő algoritmussal választottuk ki. Adott az A halmaz, mely k adatot tartalmaz. Osszuk ezt két részre: n tanuló és $n-k$ tesztadatokra. Ezen halmazokkal tanítsuk a rendszert, amíg az összhibája egy lokális minimumot ér el. Ekkor vegyük az A egy másik $(k, n-k)$ felosztását addig, amíg a, amíg a rendszer hibája egy bizonyos küszöb alá nem csökken, vagy a lépések száma meghalad egy felső korlátot.

Megjegyzések

Az elért eredmények biztatóak, a 90 százalék fölötti teljesítmények nagyon jók, mert az adatok nagyon tág határok között változhatnak. A jobb teljesítmény szükséges feltétele a megfelelő számú, gondosan összeválogatott tanító és tesztadat. Távlati célként egy olyan rendszer létrehozását tűztük ki célul, amely képes nagy adatbázisok kezelésére is, ezenkívül lehetővé teszi a fonémák felismerését is. Így lehetőség nyílik arra, hogy folytonos beszédet is fel lehessen ismerni, valós időben.

Irodalomjegyzék

- [1] Li Min Fu, Neural Networks In Computer Intelligence , McGraw-Hill,1991, 197-222
- [2] G. Todorean et al. Retele Neuronale Artificiale, Microinformatica, Kolozsvár 1995
- [3] Timothy Masters, Signal and Image Processing with Neural Networks, John Wiley,NY, 1994

Buzogány Endre, IV.éves informatikus hallgató

Babes-Bolyai Tudományegyetem, Kolozsvár, str Hasdeu 90/96E-mail: be6573@scs.ubbcluj.ro