Een Ziekte Inbedding leren met het gebruik van Veralgemeende Word2Vec Methoden

Milan van der Meer

***Samenvatting*—In de medische wereld is er een stijgend gebruik van Electronic Health Records (EHRs). Hierdoor is er meer medische data beschikbaar en is er de mogelijkheid om nieuwe verbanden te vinden. Deze verbanden kunnen onderhanden gebruikt worden om de werking van medicaties te testen, kosten te bepalen van bepaalde ziektebeelden, en het vinden van nieuwe relaties tussen verschillende ziekten.**
In dit nieuwe vakgebied van EHR analytics, is er een beperkt gebruik van recent machine learning technieken. Het doel van dit paper: het gebruiken van recente machine learning technieken om verbanden te vinden in EHRs.

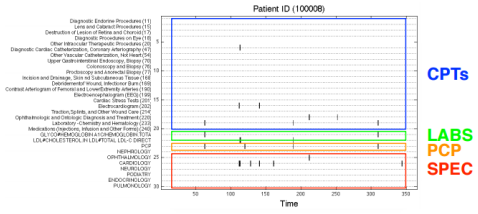
Om dit te kunnen doen, maken we het verband tussen zinnen van woorden en sequenties van EHR events. Met de analogie introduceren wij veralgemeende Word2Vec methoden. Deze breiden we uit met DeepWalk voor performantie redenen aangezien dit ons toelaat om kleinere datasets te genereren. Vervolgens lossen we een Word2Vec probleem op namelijk het niet kunnen verwerken van ongeziene instanties. Hiervoor gebruiken we een k-nearest neighbors methode in combinatie met Word2Vec.

We testen 504 verschillende parameter settings om onze modellen te bouwen. Deze modellen worden vervolgens getest en vergeleken met een paper over Deense EHRs. We vinden dat er een 60% match is tussen onze modellen en de Deense modellen. Hieruit concluderen we dat onze methodes belovend zijn zeker aangezien we verschillende benadering gebruiken die niet in de Deense EHRs aanwezig zijn. We concluderen ook dat doordat alle modellen hetzelfde scoren, de voorgestelde uitbreidingen op Word2Vec performant zijn.

***Index Terms*—Inbedding, clustering, Word2Vec, EHR analytics**

I. Electronic Health Records Analytics

IN de medische wereld is er een stijgend gebruik van medische hulp systemen. Deze systemen maken het mogelijk om medische data op een eenvoudige manier op te slaan. Voorbeelden hiervan zijn, dokters bezoeken, hospitalisatie, labo resultaten, en anderen, zie figuur 1. Deze gegevens worden samen opgeslagen in een Electronic Health Record (EHR) [19].



Figuur 1. Example of an EHR transformed into a matrix structure [21]

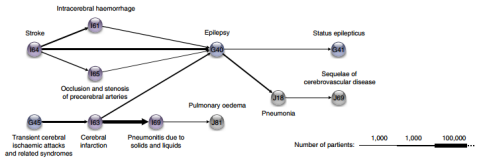
Door deze stijging in beschikbare medische gegevens, zijn er verschillende partijen zoals dokters, hospitalen, en overheden met interesse om nieuwe verbanden te vinden hierin. Deze verbanden kunnen onderhanden gebruikt worden om de werking van medicaties te testen, kosten te bepalen van bepaalde ziektebeelden, en het vinden van nieuwe relaties tussen verschillende ziekten.
Omdat er vele problemen zijn rond EHRs zoals privacy, grote hoeveelheden data, complexe structuren, en verschillende gebruikte ziekte coden (bv. MedDRA [1] of ICD-10 [1]) , zijn er recent verschillende onderzoeksgroepen gevormd die zich focusen op de analyse van EHRs. Deze groepen gebruiken van eenvoudige methoden tot zeer complexe methoden zoals querying, statistische analyses, data mining, en machine learning algoritmes [12] [10] [21] [6].

Voor dit artikel is een wel bepaald onderzoek zeer relevant namelijk het tot nu toe grootste EHR onderzoek [11]. Dit onderzoek is gedaan met behulp van een Deense dataset waarop men verschillende ziekte cluster zoekt met behulp van data mining methoden. Als resultaat hebben ze dus een overzicht van verschillende disease trajectory clusters, zie figuur 2.

June 3, 2016

II. Veralgemeende Word2Vec Methoden

In deze sectie leggen we de analogie tussen zinnen van woorden en sequenties van EHR events. Door deze



Figuur 2. Cerebrovascular disease trajectory cluster for the Danish population [11]

analogie, kunnen we de link leggen tussen Word2Vec [14] en medische data.

A. Data representatie

De medische geschiedenis van een patient kan gezien worden als een tijdserie van EHR events. We noteren een EHR event als een vector m_t^p, met p een patienten nummer en t een tijdstip. Dit betekent dat elke patient een sequentie heeft van vectors, namelijk $s^p = m_t^p, m_{t+1}^p, m_{t+3}^p, \ldots$ Een vector kan waarden bevatten zoals bloeddruk, leeftijd, diagnose, en anderen (zie figuur 1).

B. Veralgemeende Word2Vec

Word2Vec wordt typisch toegepast op grote tekst corpusen. Met deze methode kan een inbedding worden gevonden waar woorden worden geprojecteerd naar een nieuwe vector ruimte. In deze vector ruimte worden de relaties tussen de woorden weergegeven door afstand tussen elkaar.

Een medische dataset kan worden beschouwd als een grote tekst corpus. Het bevat verschillende patienten (of zinnen) die elke een sequentie van EHR events hebben (of woorden). Met deze analogie, kunnen we Word2Vec toepassen op medische data. Op deze manier leren we een inbedding voor verschillende EHR events.
Andere soorten vergelijkingen zijn alreeds gemaakt zoals met protein sequenties [4].

C. Knn Word2Vec

Vervolgend lossen we een Word2Vec probleem op namelijk het kunnen verwerken van ongeziene instanties. Hiervoor gebruiken we een k-nearest neighbors methode [7] in combinatie met Word2Vec.

Deze knn methode kan toegepast worden omdat we aan het werken zijn met EHR events. Op basis van de EHR events die zich in de lookup table van het getrainde Word2Vec model bevinden, kunnen we de knn vinden voor een ongeziene EHR event. Op basis van deze knn EHR events, nemen we het gewogen gemiddelde van hun representaties in de nieuwe vector ruimte. Dit gewogen gemiddelde is de nieuwe vector representatie voor het ongeziene EHR event.

D. Veralgemeende DeepWalk

DeepWalk [17] begint van een grafen-structuur en is daardoor niet direct toepasbaar op EHR data. Een voordeel van DeepWalk is dat het een methode aanbiedt om op basis van de originele dataset een kleinere dataset te maken. Dit doet het aan de hand van gewogen random walks. Met deze kleinere dataset, kunnen we een Word2Vec model sneller trainen.

Om deze random walks uit te kunnen voeren, moeten we eerst onze EHR data omzetten naar een grafen structuur. Dit doen we aan de hand van de sequenties in de originele dataset. Elke sequentie kan worden omgezet naar een directed graaf. Overeenkomstige vertices worden dan verbonden met elkaar en de gewichten van de edges worden aangepast op basis van de frequentie dat dezelfde vertices elkaar opvolgen.

III. Resultaten

A. Parameters

De performantie van een neuraal netwerk hangt sterk af van de gekozen waarden voor de parameters. Het tunen van deze parameters is een moeilijk probleem en complexe methoden zijn alreeds voorgesteld om dit op lossen [5]. We kozen 504 verschillende parameter settings om onze modellen te trainen. Merk op dat dit geen paper is over parameter tuning en daardoor het gebruik van de bovengenoemde methode mogelijk is in later werk.
In tabel I, zie je een overzicht van alle gekozen parameters. Alle mogelijke combinaties van deze parameters zijn getest.

Tot onze kennis, is er maar een enkele studie die zich toespitst op het tunen van Word2Vec parameters [13]. In deze studie nemen ze enkele random waarden en kijken vervolgens wat hun effecten zijn. Wij voeren een gelijkaardige methode uit door ook random waarden te nemen en kijken wat hun effecten zijn.

Parameter	Generalized Word2Vec	Knn Word2Vec	DeepWalk
Vectorlength	[50, 100]	[50, 100]	[50, 100]
Batch Size	500	500	500
Epoch	1	1	1
Window Size	[5, 10, 15]	[5, 10, 15]	[5, 10, 15]
Learning Rate	[0.025, 0.1]	[0.025, 0.1]	[0.025, 0.1]
Minimum Word Freq	[5, 10]	[5, 10]	[5, 10]
ClusterK	[100, 1000, 5000]	[100, 1000, 5000]	[100, 1000, 5000]
K	/	[10, 50, 100]	/
Walklength	/	/	[5, 10, 15]

Tabel I
DE GETESTE PARAMETERS VOOR ELKE VERALGEMEENDE WORD2VEC METHODE.

B. Vergelijking Methoden

We vergelijken onze 3 methoden met elkaar op basis van de gemiddelde en maximum matching percentage. Voor elke methode gebruiken we de parameter setting die de hoogste waarde geeft voor de gemiddelde matching percentage. Deze settings kan je terugvinden in tabel II.

Elke methode heeft een gelijkaardige matching percentage, zeker als je kijkt naar het maximum matching percentage. We concluderen dat beide onze knn Word2Vec en DeepWalk dezelfde performantie hebben dan de basis veralgemeende Word2Vec. Dit betekent dat we nu ongeziene EHR events aankunnen met behulp van onze knn Word2Vec en ook dat we kleinere dataset kunnen gebruiken met behulp van DeepWalk zonder performantie te verliezen.

Het blijft moeilijk om te quantificeren hoe goed een match percentage van 60% is. Maar we maken de assumptie dat dit goed genoeg is om het potentieel van onze methoden aan te tonen. Zeker indien we rekening houden met de gebruikte benaderings methoden zoals de code mapping, verschillende datasets, en de categorizatie.

IV. CONCLUSIE

In dit paper legde we het nieuwe vakgebied rond EHR analytics uit en waarom het interessant is om verbanden tussen vershillende ziekten te vinden hierin. Onderzoeksgroepen gebruiken hiervoor allerhande methoden maar een beperkt aantal machine learning methoden worden gebruikt.

Daardoor stelden wij onze veralgemeende Word2Vec methoden voor zodat Word2Vec kan worden toegepast op medische data. Deze breiden we uit met DeepWalk voor performantie redenen aangezien dit ons toelaat om kleinere datasets te genereren. Vervolgend lossen we een Word2Vec probleem op namelijk het kunnen verwerken van ongeziene instanties. Hiervoor gebruiken we een k-nearest neighbors methode in combinatie met Word2Vec.

We hebben 504 verschillende parameter settings getest om onze modellen te bouwen. Van deze parameter settings nemen wij diegene die voor elke model de hoogste gemiddelde matching percentage geeft tussen de Word2Vec clusters en de Deense clusters.

We vinden dat er een 60% match is tussen onze clusters en de Deense clusters. Hieruit concluderen we dat onze methodes belovend zijn zeker aangezien we verschillende benadering gebruiken die niet in de Deense EHRs aanwezig zijn.
We concluderen ook dat doordat alle modellen hetzelfde scoren, de voorgestelde uitbreidingen op Word2Vec performant zijn.

V. TOEKOMSTIG WERK

Een Word2Vec model kan gebruikt worden in combinatie met een neuraal netwerk. Dit neuraal netwerk maakt gebruik van de nieuwe vector representie om zijn voorspellingen te verbeteren. In de toekomst, zou het interessant zijn om ziektes van patienten te kunnen voorspellen hiermee en vervolgens te testen of het gebruik van onze Word2Vec methoden deze voorspellingen accurater maken.

REFERENTIES

[1] MedDRA. http://www.meddra.org/. (Accessed on 05/03/2016).
[2] The Speech Recognition Wiki. http://recognize-speech.com/. (Accessed on 05/16/2016).
[3] WHO — International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/. (Accessed on 04/27/2016).
[4] E. Asgari and M. R. K. Mofrad. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS ONE*, 10(11):1–15, 11 2015.
[5] M. Bashiri and A. F. Geranmayeh. Tuning the parameters of an artificial neural network using central composite design and genetic algorithm. *Scientia Iranica*, 18(6):1600 – 1608, 2011.
[6] C. Bennett and T. Doub. Data Mining and Electronic Health Records: Selecting Optimal Clinical Treatments in Practice. *CoRR*, abs/1112.1668, 2011.

Parameter	Generalized Word2Vec		Knn Word2Vec		DeepWalk	
	Exp 1	Exp 2	Exp 1	Exp 2	Exp 1	Exp 2
Vectorlength	100	50	100	50	50	100
Window Size	15	15	5	5	5	10
Learning Rate	0.025	0.025	0.025	0.025	0.025	0.025
Minimum Word Freq	10	5	5	5	10	5
ClusterK	100	5000	100	5000	100	5000
K	/	/	100	100	/	/
Walklength	/	/	/	/	5	15
Average Matching %	29	62	33	61	27	61
Maximum Matching %	56	69	69	69	61	69

Tabel II

BEST GEVONDEN PARAMETER SETTING VOOR ELKE METHODE.

[7] T. M. Cover. Nearest neighbor pattern classification. 1982.

[8] Google Brain Team. Vector Representations of Words. https://www.tensorflow.org/versions/0.6.0/tutorials/word2vec/index.html. (Accessed on 05/16/2016).

[9] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks. A Closer Look at Skip-gram Modelling.

[10] A. Hameurlain, J. Kung, R. Wagner, H. Decker, L. Lhotska, and S. Link, editors. *Transactions on Large-Scale Data- and Knowledge-Centered Systems IV - Special Issue on Database Systems for Biomedical Applications*, volume 8980 of *Lecture Notes in Computer Science*. Springer, 2011.

[11] A. B. Jensen, P. L. Moseley, T. I. Oprea, S. G. Ellesoe, R. Eriksson, H. Schmock, P. B. Jensen, L. J. Jensen, and S. Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun*, 5, Jun 2014. Article.

[12] C. K. R. Jimeng Sun. Big Data Analytics for Healthcare. 2013.

[13] O. Levy, Y. Goldberg, and I. Dagan. Improving Distributional Similarity with Lessons Learned from Word Embeddings. *TACL*, 3:211–225, 2015.

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. *CoRR*, abs/1310.4546, 2013.

[15] Observational Medical Outcomes Partnership. OSIM2 - Observational Medical Dataset Simulator Generation 2 — Observational Medical Outcomes Partnership. http://omop.org/OSIM2. (Accessed on 05/28/2016).

[16] C. Olah. Deep Learning, NLP, and Representations - colah's blog. http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/. (Accessed on 05/16/2016).

[17] B. Perozzi, R. Al-Rfou, and S. Skiena. DeepWalk: Online Learning of Social Representations. *CoRR*, abs/1403.6652, 2014.

[18] X. Rong. word2vec Parameter Learning Explained. *CoRR*, abs/1411.2738, 2014.

[19] C. P. Stone. A Glimpse at EHR Implementation Around the World: The Lessons the US Can Learn. 2014.

[20] J. Sun, F. Wang, J. Hu, and S. Edabollahi. Supervised Patient Similarity Measure of Heterogeneous Patient Records. *SIGKDD Explor. Newsl.*, 14(1):16–24, Dec. 2012.

[21] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi. Towards Heterogeneous Temporal Clinical Event Pattern Discovery: A Convolutional Approach. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 453–461, New York, NY, USA, 2012. ACM.