Discovering potential clinical profiles of Multiple Sclerosis from clinical and pathological free text data with Constrained Non-negative Matrix Factorization

Jacopo Acquarelli[1], The Netherlands Brain Bank[2], Monica Bianchini[3], Elena Marchiori[1]

[1] Institute for Computing and Information Sciences (ICIS), Radboud University, Nijmegen, The Netherlands
j.acquarelli@cs.ru.nl, elenam@cs.ru.nl
[2] Netherlands Institute for Neuroscience, Amsterdam, The Netherlands
[3] Dipartimento di Ingegneria dell'Informazione e Scienze Matematiche, Università degli Studi di Siena, Italy
monica@diism.unisi.it

Abstract. Constrained non-negative matrix factorization (CNMF) is an effective machine learning technique to cluster documents in the presence of class label constraints. In this work, we provide a novel application of this technique in research on neuro-degenerative diseases. Specifically, we consider a dataset of documents from the Netherlands Brain Bank containing free text describing clinical and pathological information about donors affected by Multiple Sclerosis. The goal is to use CNMF for identifying clinical profiles with pathological information as constraints. After pre-processing the documents by means of standard filtering techniques, a feature representation of the documents in terms of bi-grams is constructed. The high dimensional feature space is reduced by applying a trimming procedure. The resulting datasets of clinical and pathological bi-grams are then clustered using non-negative matrix factorization (NMF) and, next, clinical data are clustered using CNMF with constraints induced by the clustering of pathological data. Results indicate the presence of interesting clinical profiles, for instance related to vision or movement problems. In particular, the use of CNMF leads to the identification of a clinical profile related to diabetes mellitus. Pathological characteristics and duration of disease of the identified profiles are analysed. Although highly promising, results of this investigation should be interpreted with care due to the relatively small size of the considered datasets.

Introduction

Due to the high cost and large amount of under-utilized data, health care needs computational methods to boost the benefits of available data, in order to yield more efficient practice and research [1]. Application of such methods to data can

help to guide and facilitate research, leading to new knowledge and improved clinical standards [2].

The situation applies prominently to the Netherlands Brain Bank (NBB). The NBB supplies the international scientific community with clinical and neuropathological well-documented brain tissue from donors [3]. It contains also a host of yet un-utilized data, describing anonymous clinical information collected from the donors, including the medical and family history, the type and course of the diseases, and the clinical diagnosis. Moreover the NBB contains pathological post-mortem information about the donors including the true (pathological) diagnosis and various pathological hallmarks.

Pathological parameters collected post-mortem cannot be used as features for performing diagnosis, treatment and prognosis. Nevertheless, such valuable side information can be exploited for discovering clinical disease profiles characterized by few clinical features linked to pathological parameters of interest. These profiles could be used by domain experts as a guidance for performing targeted studies, in order to improve diagnosis, treatment and prognosis of neuro-degenerative diseases. In machine learning, data that can be used for both constructing and testing a model (such as the clinical data of the NBB) are also called *technical*, while data that cannot be used for testing (such as the pathological data of the NBB) are called *privileged* [4].

In this paper we investigate a methodology based on clustering with free text technical and privileged data from the NBB to identify clinical profiles of donors affected by Multiple Sclerosis (MS), a chronic inflammatory disease of the central nervous system.

In our analysis of NBB documents we do not take into account medical knowledge, but only use the text as it is, except for the application of standard lexical and statistical filtering methods. We perform cluster analysis of these data using Non-Negative Matrix Factorization (NMF), a state-of-the-art machine learning technique for clustering documents [5]. First, a clustering of the pathological data using NMF is computed. The resulting clusters induce a partition of the data in classes. Next, the clinical data are clustered using Constrained NMF (CNMF) [6], by integrating the class information of the donors as an additional constraint. In this way, donors from the same pathological cluster will be merged together in the new (clinical) representation space.

Results of this clustering methodology indicate the presence of clinical profiles describing subtypes of the MS disease. In particular, we show that the obtained clusters correspond to groups of donors with different duration of disease.

1 Related work

Medical documents, such as clinical or pathological annotations, contain valuable information about patients, such as their medical history (diseases, injuries, medical symptoms) and responses (diagnoses, prescriptions, and drugs) [7]. These data have a huge potential in order to build profiles for individual patients or classes of patients sharing a similar disease course, to discover disease correla-

tions [8], and enhance patient care [9]. Moreover, also in medical error detection a number of tools have emerged, from medical informatics and computer science — natural language processing, visualization, and machine learning tools — as well as methods for understanding cognitive processes, which can collect a large amount of important clinical information, that normally lie locked in narrative reports, unavailable to automated decision support systems [10]. Whatever the particular problem, large volumes of medical documents have been recently generated by electronic health record systems, whose nature is normally unstructured or semi–structured, making the information extraction procedure a very difficult task. Due to the intrinsic diversity among medical documents, it is just a challenge to discover the underlying patterns from a corpus. Thus, document clustering techniques, being an efficient way of navigating and summarizing documents, have been intensively investigated in biomedical research. As a dimension reduction method, non–negative matrix factorization [11] has been widely applied to medical document clustering [12,13]. By imposing non–negativity constraints in both basis and weight factorization matrices, NMF guarantees to preserve the local structure of the original data. Moreover, the resulting latent semantic space of the clustered documents produced by NMF may be explained in a very intuitive way. Specifically, each axis in the semantic space represents the basic topic of a particular cluster, whilst each document in a collection is viewed as the additive combination of the basic topics. Therefore, a particular document is grouped into the cluster where it has the largest projection value. Many extensions of the basic NMF method have also been explored for clustering biomedical documents. For instance, in [13], Multi–view NMF, which can integrate different data sources, was applied for clustering clinical document, based on medication/symptom names, whereas, in [12], ensemble NMF, able to achieve a consensus solution from a set of runs with different initial conditions, was tested on the TREC genomic 2004 track. Finally, also more complex techniques were recently introduced in order to cope with graph representations of medical documents [14]. Actually, Subgraph Augmented Non–negative Tensor Factorization (SANTF), in addition to relying on atomic features (e.g., words in clinical narrative text), automatically mines higher–order features by converting sentences into a graph representation and identifying important subgraphs. Latent groups of atomic features were shown to help in better correlate latent groups of higher–order features. Moreover, feature analysis also identified latent groups of higher–order features that lead to interesting medical insights.

2 Multiple Sclerosis

Multiple sclerosis (MS) is a complex, chronic inflammatory disease that affect the central nervous system and whose causes are largely unknown. The disease usually causes relapsing-remitting attacks of inflammation, demyelination and axonal damage, leading to various degrees and spectra of neurological symptoms and disability (see [15], [16]). The main types of MS[4] are listed below.

[4] See `http://www.multiplesclerosis.com/us/treatment.php`

- *Relapsing-Remitting* MS (RR-MS) is the most diffuse type. People affected by RR-MS have temporarily periods called relapse when the symptoms appear.
- *Secondarily Progressive* MS (SP-MS) is considered as an advanced state of RR-MS, with symptoms that go worse steadily without relapses.
- *Primarily Progressive* MS (PP-MS) is characterized by slowly worsening of the neurological functions without any remission from the beginning.
- *Progressive-Relapsing* MS (PR-MS) is a rare form of MS characterized by a steadily worsening disease state from the beginning, with acute relapses but no remissions, with or without recovery.

Main MS symptoms include numbness or weakness in the arms and legs, blurred or double vision, or pain during eye movement, partial or total vision loss, pain or tingling in different areas of the body, lack of coordination or unsteady walking, tremors, dizziness, and extreme fatigue.

3 Methods

We begin by describing the data and techniques used in our analysis.

3.1 Data pre-processing

We consider 149 records of MS donors, which are free text documents composed by three parts:

- *General information*: contains gender, age, etc.;
- *Clinical information*: contains the clinical history of the patient including prescriptions and diagnosis;
- *Pathological information*: contains information including the presence of diverse types of brain lesions and the pathological diagnosis.

The pathological and clinical parts of the records are used in our analysis as privileged and technical data, respectively.

A free text is a mixture of words and signs so the first pre-processing step is devoted to put it in the form of '*word space word*'. Given that it is a medical text, there are a lot of acronyms and nomenclatures typical of this area. Also a free text generally contains noise from uninformative words to be filtered out. We choose to consider only words that belong to the adjective, adverb, noun and verb classes, with a length greater than two.

A Python script was used for filtering, taking advantage of the WordNet library, a large lexical database, containing tools to process English text, including grammatical analysis and stemming, and stop-word removing procedures. Stemming transforms a generic word into its basic form (for example 'chronically' will be transformed into 'chronic'). It is an important operation because it serves to merge semantically similar words. Instead, stop-word removing discards uninformative words contained in the text.

After filtering, bi-grams are generated, yielding 64858 bi-gram features, whose values are represented using the term frequency-inverse document frequency (TF-IDF), a score estimating how important a bi-gram is to a document within a collection [17]. For a bi-gram t, in a record x of the (filtered) data X, term frequency and inverse document frequency are defined as follows.

$$TF(t,x) = 0.5 + 0.5f(t,x)/\max(\{f(d,x) \mid d \in X\}),$$

where $f(t,x)$ is the frequency of t in x;

$$IDF(t,X) = log(|X|/|\{x \in X \text{s.t. } t \text{ occurs in } x\}|),$$

where $|\ |$ denotes the size of a set.

The TF-IDF of a bi-gram t is the product of $TF(t,x)$ and $IDF(t,X)$.

In order to further reduce the input dimension of the data, *trimming* is applied to the distribution of TF-IDF values: 2-grams with TF-IDF value on the tail of the distribution are removed from the data. Specifically, we remove features that appear in more than 50% of the records, as well as those that appear in less than three documents. In this way a dataset with 1084 features is obtained.

3.2 Non-Negative Matrix Factorization

NMF approximates the input matrix X of dimension $m \times n$ (in our context $m = 1084$, $n = 149$) as the product of two matrices W and H, of dimension $m \times k$ and $k \times n$ respectively, with k much smaller than m and n.

The problem can be formulated as finding W and H that minimize the objective function

$$||X - WH||^2,$$

with the constraint that both W and H are non-negative. Here $||\ ||$ denotes the Frobenius norm. Since this problem is intractable [18], heuristic methods are used to find locally optimal solutions. Here, to find W and H that locally optimize the objective function, the following multiplicative update rules [19] are used:

$$H_{\alpha\mu} = H_{\alpha\mu} \frac{\sum_i W_{i\alpha} X_{i\mu}/(WH)_{i\mu}}{\sum_k W_{k\alpha}},$$

$$W_{\alpha\mu} = W_{\alpha\mu} \frac{\sum_i H_{i\alpha} X_{i\mu}/(WH)_{i\mu}}{\sum_k H_{k\alpha}}.$$

H represents a soft clustering, where the entry (i,j) can be interpreted as the membership degree with which record j belongs to cluster i. W provides information about the relevance of terms for the clusters: each element w_{ij} of W represents the degree with which term (in our context bi-gram) t_i belongs to cluster j.

Crisp clusters can be generated from H by assigning each donor to the cluster having the highest membership degree (where ties are broken randomly).

3.3 Constrained Non-Negative Matrix Factorization

The resulting clusters form disjoint classes of donors, which can be used as privileged information to cluster the clinical data by means of CNMF.

CNMF [6] is a kind of semi-supervised clustering method. It considers an input dataset for which the class label of some (l) elements is given. From these class labels, an $l \times k$ indicator matrix C is constructed, whose (i, j) entry contains a 1 if the i-th point of the dataset X is labelled with the j-th class. C is used to define the constraint matrix A:

$$A = \begin{bmatrix} C_{l \times k} & 0 \\ 0 & I_{n-l} \end{bmatrix},$$

where I_{n-l} is a $n-l \times n-l$ identity matrix. Then, CNMF approximates the input matrix X by the product WAZ, with the constraint that W and Z are both non-negative. This corresponds to find W and Z that minimize the objective

$$||X - WAZ||^2,$$

under the constraint that $W \geq 0$, $H \geq 0$. To find W and Z that locally optimize this function, the multiplicative update rules given in [6] are used.

4 Experiments

In our experiments we set the parameters, namely the number k of clusters and the number l of labelled donors, to the somewhat arbitrary values of 3 and 120 (about 80% of the donors randomly selected from the dataset). Therefore, we aim at identifying three main clinical MS profiles, using a large amount of privileged information provided by the clustering on the pathological data. The following experiments were carried out on the pre-processed data:

1. cluster pathological data using NMF;
2. cluster clinical data using NMF;
3. cluster clinical data using CNMF, using the clustering of pathological data as privileged knowledge.

A set matching algorithm [20] was used to align clusters from two clusterings of the donors, for instance the two clustering obtained by using the clinical and pathological data. For each cluster in one clustering C, a best match in the other clustering C' is found. This is done by processing the elements n_{ij} of the contingency table in decreasing order, where

$$n_{ij} = |C_i \cap C'_j|$$

is the number of donors in the intersection of the cluster i of the clustering set C and the cluster j in the clustering set C'. The largest number, say n_{ab}, entails a match between the two clusters $a \in C$ and $b \in C'$, whereas the second largest number entails the second match, and so on.

5 Results

We first analyze results of clustering pathological and clinical documents separately, and then by using CNMF. Specifically, for each cluster we report the 10 words considered as most important by the clustering algorithms, and analyze the composition of the clusters with respect to two external metrics: duration of disease and type of MS.

5.1 Clustering pathological data

The top 10 relevant bi-grams for the pathological clusters are shown in Table 1.

Table 1. Top 10 relevant bi-grams for clusters of the pre-processed pathological dataset clustered using NMF.

Cluster 1	Cluster 2	Cluster 3
inactive-plaque	cortex-white	spinal-marrow
chronically-inactive	matter-ependyma	abnormally-spinal
chronic-inactive	chronic-active	thoracic-spinal
lateral-string	sclerosis-plaque	brain-laminate
sclerosis-plaque	abnormality-spc	cord-section
posterior-string	waes-contains	abnormality-conclusion
small-plaque	section-contain	nucleus-putamen
axonal-density	matter-cortex	brain-stem
decrease-myeline	diagnosis-multiple	plaque-find
myelinated-axon	cord-abnormality	visible-brain

Clustering pathological data yields interesting results. As shown in Table 1, the top 10 relevant bi-grams reveal three pathological profiles characterized by different brain lesion types:

- (chronic-inactive): this profile contains as top terms 'chronic-inactive', but also terms such as 'small-plaque', 'decrease-myeline' and 'myelinated-axon' which provide further characteristics of brain lesions in this profile.
- (chronic-active): this profile contains as top terms 'chronic-active', but also terms such as 'abnormality-spc' and 'matter-ependyma' which provide further characteristics of brain lesions in this profile.
- (spinal-cord): this profile contains as top terms concerning spinal cord injuries, but also terms such as 'brain-laminate' and 'nucleus-putamen' which provide further characteristics of brain lesions in this profile.

5.2 Clustering clinical data

The top 10 relevant bi-grams for the clinical clusters are shown in Table 2.

The main characteristics of the three clinical profiles stemming from the clustering can be summarized as follows.

Table 2. Top 10 relevant bi-grams for the clusters of the pre-processed clinical dataset clustered using NMF.

Cluster 1	Cluster 2	Cluster 3
patient-suffer	right-arm	relapsive-progressive
patient-complain	lesion-visible	start-special
patient-underwent	arm-leg	phase-start
admit-hospital	focal-lesion	neuritis-subsequent
right-leg	left-arm	optic-neuritis
physical-examination	left-side	sympt-optic
complain-pain	right-side	subsequent-relapse
tension-mmhg	right-leg	eds-die
get-worse	raise-signal	iggindex-elevate
situation-get	periventricular-lesion	remark-prominent

- Cluster 1 appears to contain terms related to **pain and complaints** from the patient (patient-suffer, patient-complain, complain-pain, get-worse, situation-get) and other patient-focused outcomes (patient-underwent, admit-hospital, physical examination).
- Cluster 2 seems to have a preponderance of symptoms related to **movement disabilities**.
- Cluster 3 seems to emphasize disease progression and state (start-special, phase-start, relapsive-progressive, subsequent-relapse, eds-die, neuritis-subsequent) and contains symptoms related to **vision** ('optic-neuritis' and 'sympt-optic').

Table 3 shows the size of the crisp clusters obtained by clustering clinical and pathological data separately using NMF: there are two somewhat large clusters and a small one.

Table 3. Number of records in the aligned clusters obtained by applying NMF to the two pre-processed clinical and pathological datasets independently.

Cluster no.	Clinical Docs	Pathological Docs
1	50	40
2	88	97
3	11	12

5.3 Clustering clinical data with privileged pathological information

The three pathological clusters indicate the presence of pathological profiles of donors with diverse types of brain lesions. These clusters are used in the sequel as privileged information to identify profiles from the clinical data using CNMF.

The alignment between the clinical clustering generated by NMF and CNMF leads to same correspondence between clusters as that indirectly obtained from

the separate alignment of the NMF and CNMF clusterings with the pathological one. Therefore, for instance, NMF cluster 1 of the clinical data, CNMF cluster 1 of the clinical data and NMF cluster 1 of the pathological data are all aligned with each other.

Table 4 shows the top 10 bi-grams of these new clinical clusters.

Table 4. Top 10 relevant bi-grams for the clusters of the pre-processed clinical dataset clustered with CNMF.

Cluster 1	Cluster 2	Cluster 3
patient-suffer	right-arm	periventricular-lesion
patient-complain	arm-leg	lesion-visible
progressive-phase	left-arm	lateral-ventricle
phase-start	left-side	patient-suffer
right-leg	patient-underwent	signal-intensity
eds-die	muscle-strength	matter-lesion
optic-neuritis	patient-suffer	raise-signal
start-special	left-leg	white-matter
examination-reveal	right-leg	mellitus-type
right-side	paresis-right	diabetes-mellitus

From this table three clinical profiles emerge, whose properties can be summarized as follows.

- The top 10 most relevant bi-grams in cluster 1 represent symptoms about pain and complaints by the patient. Moreover the presence of the bi-gram 'progressive-phase' indicates the presence of a progressive phase of MS, and 'phase-start' could indicate that a new phase of MS is starting (relapse) after remitting. This could mean an SP type of MS. Cluster 1 contains also as top term 'optic-neuritis', a symptom involving vision problems.
- Cluster 2 top bi-grams refer mainly to symptoms related to movement disabilities, involving arms and legs, as well as paresis.
- Cluster 3 contains two bi-grams referring to diabetes mellitus. These could indicate a specific characteristic of a disease sub-type. Top bi-grams in cluster 3 refer also to brain lesions detected by magnetic resonance imaging (bi-grams: periventricular-lesion, lesion-visible, lateral-ventricle, matter-lesion).

The size of the clusters obtained by applying CNMF to the clinical data, and the size of the corresponding best match pathological clusters are shown in Table 5.

6 Interpretation of results

We now use an external cluster validation metric, namely the duration of disease (DOD), to analyze clinical clusters. DOD measures the number of years from

Table 5. Size of clusters of the clinical dataset partitioned with CNMF and of their best match pathological clusters.

Cluster no.	Clinical Docs	Pathological Docs
1	109	97
2	32	12
3	8	40

the moment a clinical diagnosis of the disease was performed to the death of the patient.

Table 6 shows the average DOD per type of MS and the corresponding number of records contained in the dataset.

Table 6. Average DOD for the different MS types and number of donors of that type in the dataset.

MS Type	# Records	Average DOD
RR	0	-
SP	33	22.21
PP	12	25.92
PR	0	-

In the clustering of pathological data the majority of donors with SP-MS belongs to cluster 1 and 3 while the majority of PP-MS donors is in cluster 2. Information about the type of MS is only partially available (see Table 7).

Table 7. Composition of pathological clusters wrt type of MS.

Cluster n.	Available/Total	# SP	# PP
1	26/97	20 (60.61%)	6 (50.00%)
2	5/12	2 (6.06%)	3 (25.00%)
3	14/40	11 (33.33%)	3 (25.00%)
Total		33 (100.00%)	12 (100.00%)

6.1 Clinical clusters generated by NMF

Table 8 shows the average DOD of the three clinical clusters produced by NMF.

In order to assess whether clusters have significantly different DOD, the Rank Sum Wilcoxon test is applied. Results of the test show that the three NMF clinical clusters differ significantly one from each other with respect to DOD. Composition of clinical clusters with respect to the type of MS is shown in Table 9.

Table 8. Average DOD values (Average DOD) and standard deviation (STD) of the clusters of the clinical data obtained using NMF.

Cluster n.	Available/Total	Average DOD	STD
1	66/88	27.97	±14.15
2	11/11	24.45	±11.80
3	38/50	27.21	±12.78

Table 9. Composition of clinical clusters generated by NMF with respect to the type of MS.

Cluster n.	Available/Total	# SP	# PP
1	27/88	18 (54.55%)	9 (75.00%)
2	7/11	7 (21.21%)	0 (0.00%)
3	11/50	8 (24.24%)	3 (25.00%)
Total		33 (100.00%)	12 (100.00%)

In summary, the clustering analysis using NMF identifies clinical profiles which can be summarized as follows:

- (pain-complaints-relative-long-DOD): this clinical profile involves symptoms involving pain and complaints. This cluster contains donors with a relatively long DOD.
- (movement-disabilities-short-DOD): this clinical profile involves symptoms concerning movement disabilities. This cluster contains donors with a relatively short DOD.
- (vision-MS-phases-SP/PP-relative-long-DOD): this clinical profile concerns symptoms involving vision and phases of the MS. This cluster contains donors with a relatively long DOD.

6.2 Clinical clusters generated by CNMF

Table 10 shows the average DOD of the three clinical clusters produced by CNMF.

Table 10. Average DOD values (Average DOD) and standard deviation (STD) of the clusters of the clinical data obtained using CNMF.

Cluster n.	Available/Total	Average DOD	STD
1	87/109	26.52	±12.89
2	25/32	30.24	±13.73
3	3/8	28.67	±23.30

In order to assess whether clusters have significantly different DOD, the Rank Sum Wilcoxon test is applied. Results of the test show that cluster 1 and 2 differ

significantly, as well as cluster 1 and 3. Cluster 3 has few available records so it is not possible to use its DOD value to link it to its corresponding pathological cluster.

The composition of the clusters with respect to the type of MS is shown in Table 11.

Table 11. Composition of clinical clusters generated by CNMF with respect to the type of MS.

Cluster n.	Available/Total	# SP	# PP
1	32/109	26 (78.79%)	6 (50.00%)
2	12/32	6 (18.18%)	6 (50.00%)
3	1/8	1 (3.03%)	0 (0.00%)
Total		33 (100.00%)	12 (100.00%)

The clustering analysis using CNMF identifies clinical profiles which can be summarized as follows:

- (pain-complaints-short-DOD): this clinical profile involves symptoms involving pain and vision complaints of the patient, and symptoms indicating advanced RR-MS. This cluster contains donors with a relatively short DOD.
- (movement-disabilities-long-DOD): this clinical profile involves symptoms involving movement disabilities of arms and legs, and paresis. This cluster contains donors with a long DOD and PP-MS type of disease.
- (diabetes-mellitus-long-DOD): this clinical profile is related to symptoms involving lesions and diabetes mellitus. This cluster contains donors with a relatively long DOD.

As expected, in this case the composition of clusters with respect to different MS types identified by the clinic CNMF clustering corresponds to that produced by clustering pathological data.

6.3 Comparison between NMF and CNMF clusters

Clinical records mainly report symptoms or treatments and describe the medical history of the patients. Thus the lexicon is quite different between the pathological and the clinic parts of the patient record: that explains the importance of bi-grams in the NMF clustering such as 'admit-hospital', 'complain-pain' or bi-grams describing parts of the patients' body. CNMF clusters contains less of these somewhat generic bi-grams which are replaced by more specific ones in terms of information from the medical literature, for example terms referring to diabetes. Also as expected CNMF clusters are more faithful to the pathological ones with respect to MS type composition.

The use of privileged information from the pathological clustering changes the rank of important bi-grams in the clinical clusters; in some cases bi-grams disappear or migrate from one cluster to another one.

Table 12 report words which appear in both the NMF and CNMF clusterings (column 'Stay'), which are in CNMF but not NMF (column 'In'), and which disappear (column 'Out').

From this table we can see that while for cluster 1 and 2 some of the 10 most important bi-grams of the NMF clustering remain most important also in the CNMF clustering (3/10 for cluster 1 and 5/10 for cluster 2), no bi-grams for NMF cluster 3 are kept in the corresponding CNMF cluster.

Table 12. Changes in top 10 relevant bi-grams from clinical clustering with NMF to that with CNMF.

Cluster	In	Out	Stay
1	progressive-phase phase-start eds-die optic-neuritis start-special esamination-reveal right-side	patient-underwent admit-hospital physical-examination complain-pain tension-mmhg get-worse situation-get	patient-suffer patient-complain right-leg
2	patient-underwent muscle-strength patient-suffer left-leg paresis-right	lesion-visible focal-lesion right-side raise-signal periventricular-lesion	right-arm arm-leg left-arm left-side right-leg
3	periventricular-lesion lesion-visible lateral-ventricle patient-suffer signal-intensity matter-lesion raise-signal white-matter mellitus-type diabetes-mellitus	relapsive-progressive start-special phase-start neuritis-subsequent optic-neuritis simpt-optic subsequent-relapse eds-die iggindex-elevate remark-prominent	

Bi-grams which appear in both the clusterings (column 'Stay') mainly refer to pain and complaints in parts of the patients' body. For cluster 1 top bi-grams which are in CNMF but not NMF (column 'In') are mainly associated with the phase of the disease, while for cluster 2 such bi-grams are related to symptoms related to movement, and for cluster 3 they mainly refer to lesions and diabetes.

In general, results indicate that the information provided by the pathological clustering affect the characteristics of clinical profiles by favoring the emergence of a small cluster associated to diabetes mellitus. The co-occurrence of MS and diabetes mellitus has been reported by a number of studies. In particular as mentioned in [21] for the combined effect of diabetes mellitus type 1 and type 2 there is evidence that this is associated with a worse progression of disability compared to MS patients without type 1 or type 2 diabetes. This is in accordance with clinical profile we identified using privileged information which indicates that the neuropathology associated with this form of diabetes might influence the disease course (symptoms related to lesions, spinal cord brain lesion) and contribute to the severity of MS (short DOD).

7 Conclusion

In this paper, we investigated a methodology to identify clinical MS profiles using pathological information as privileged data to guide the identification process. To this aim, data from the NBB, consisting of free text documents containing clinical and pathological records, were used. The data were first pre-processed and NMF was used to cluster both clinical and pathological data independently. CNMF was then employed to cluster clinical data, using constraints inferred from the pathological data. The obtained results indicate the presence of profiles with characteristics which reflect underlying neuropathological differences, and differing DOD outcomes. In particular, the use of privileged information lead to the identification of a clinical profile related to diabetes mellitus.

Although potentially interesting, these results should be interpreted with care because of the number of DOD missing values and the small size of the data. The proposed analysis applied to a larger dataset would provide deeper understanding and confidence on the potential relevance of these profiles.

It is an interesting and relevant matter of future research to provide stronger ties between the observed clusterings, in particular the final ones, and the neuropathology of MS from literature studies, as well as to use a corpus of documents from the literature as external prior knowledge to enhance the clustering process.

Acknowledgments

This work has been partially funded by the Netherlands Organization for Scientific Research (NWO) within the NWO project 612.001.119.

References

1. D. Urbach and J. H. Moore. Data mining and the evolution of biological complexity. *BioData mining*, 4, 2011.
2. D. Davis and N. V. Chawla. Exploring and exploiting disease interactions from multi–relational gene and phenotype networks. *PloS One*, 6(7), 2011. e22670.
3. J. E. Bell *et al.* Management of a twenty–first century brain bank: experience in the BrainNet Europe consortium. *Acta Neuropathol.*, 115(5):497–507, 2008.
4. V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5–6):544–557, 2009.
5. D. D. Lee and H. S. Seung. Learning the parts of objects by non–negative matrix factorization. *Nature*, 401(6755):788–791, 1999.
6. H. Wu and Z. Liu. Non–negative matrix factorization with constraints. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, pages 506–511, 2010.
7. K. Roberts and S. M. Harabagiu. A flexible framework for deriving assertions from electronic medical records. *J. Amer. Med. Informat. Assoc.*, 18(5):568–573, 2011.
8. F. S. Roque *et al.* Using electronic patient records to discover disease correlations and stratify patient cohorts. *PLoS Comput. Biol.*, 7(8):E1002141, 2011.
9. G. Hripcsak *et al.* Mining complex clinical data for patient safety research: A framework for event discovery. *J. Biomed. Informat.*, 36(1):120–130, 2003.

10. G. B. Melton and G. Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *J. Am. Med. Inform. Assoc.*, 12:448–457, 2005.
11. W. Xu, X. Liu, and Y. Gong. Document clustering based on non–negative matrix factorization. In *Proceedings of the 26th Ann. Int. ACM SIGIR Conference*, pages 267–273. ACM, 2003.
12. X. Huang, X. Zheng, W. Yuan, and S. Zhu. Enhanced clustering of biomedical documents using ensemble non–negative matrix factorization. *Information Sciences*, 181:2293–2302, 2012.
13. Y. Ling, X. Pan, G. Li, and X. Hu. Clinical documents clustering based on medication/symptom names using multi–view nonnegative matrix factorization. *IEEE Transactions on Nanobioscience*, 14(5):500–504, 2015.
14. Y. Luo *et al.* Subgraph augmented non–negative tensor factorization (SANTF) for modeling clinical narrative text. *J. Am. Med. Inform. Assoc.*, 22(5):1009–1019, 2015.
15. L. Bö, J. J. G. Geurts, S. J. Mörk, and P. Van der Valk. Grey matter pathology in multiple sclerosis. *Acta Neurologica Scandinavica*, 113:48–50, 2006.
16. P. Van der Valk and C. J. A. De Groot. Staging of multiple sclerosis (MS) lesions: pathology of the time frame of MS. *Neuropathology and applied neurobiology*, 26:2–10, 2000.
17. R. Feldman *et al.* Text mining at the term level. In *Principles of Data Mining and Knowledge Discovery*, pages 65–73. Springer, 1998.
18. S. A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.
19. D. D. Lee and H. S. Seung. Algorithms for non–negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.
20. M. Meilă and D. Heckerman. An experimental comparison of model–based clustering methods. *Machine learning*, 42(1–2):9–29, 2001.
21. P. Tettey, S. Simpson, B.V. Taylor, and I.A.F. van der Mei. The co-occurrence of multiple sclerosis and type 1 diabetes: Shared aetiologic features and clinical implication for ms aetiology. *Journal of the neurological sciences*, 348(1):126–131, 2015.