Latent Patient Profile Modelling and Applications with Mixed-Variate Restricted Boltzmann Machine

Tu Dinh Nguyen, Truyen Tran, Dinh Phung, and Svetha Venkatesh

Center for Pattern Recognition and Data Analytics
School of Information Technology, Deakin University, Geelong, Australia
{ngtu,truyen.tran,dinh.phung,svetha.venkatesh}@deakin.edu.au

Abstract. Efficient management of chronic diseases is critical in modern health care. We consider *diabetes mellitus*, and our ongoing goal is to examine how machine learning can deliver information for clinical efficiency. The challenge is to aggregate highly heterogeneous sources including demographics, diagnoses, pathologies and treatments, and extract similar groups so that care plans can be designed. To this end, we extend our recent model, the mixed-variate restricted Boltzmann machine (MV.RBM), as it seamlessly integrates multiple data types for each patient aggregated over time and outputs a homogeneous representation called "latent profile" that can be used for patient clustering, visualisation, disease correlation analysis and prediction. We demonstrate that the method outperforms all baselines on these tasks - the primary characteristics of patients in the same groups are able to be identified and the good result can be achieved for the diagnosis codes prediction.

1 Introduction

Chronic diseases are rampant. Health care costs are increasingly related to such diseases. *Diabetes mellitus* is one such chronic disease from which 346 million people worldwide suffer, as estimated by The World Health Organization (WHO) [1]. Only about $5 - 10\%$ of them have Type I diabetes mellitus, whilst Type II comprises the rest. The people who suffer Type I are not able to produce insulin. In contrast, Type II diabetes means that there is an inability to absorb insulin. In 2004, 3.4 million people died from complications of high blood sugar. The incidence of diabetes mellitus is increasing, and being diagnosed in younger people. This leads to serious complications - deterioration in blood vessels, eyes, kidneys and nerves. It is a chronic, lifelong disease.

Escalating health costs are associated with such chronic diseases. To provide high quality healthcare, care plans are issued to patients to manage them within the community, taking steps in advance so that these people are not hospitalised. Thus, it is imperative to identify groups of patients with similar characteristics so that they can be covered by a coherent care plan. Additionally, if the hospital can predict the disease codes arising from escalating complication of chronic disease, it can adjust financial and manpower resources. Thus useful prediction of codes for chronic disease can lead to service efficiency.

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 123–135, 2013.

Clustering is a natural selection for this task. However, medical data is complex – it is mixed-type containing Boolean data (e.g., male/female), continuous quantities (e.g., age), single categories (e.g., regions), and repeated categories (e.g., disease codes). Traditional clustering methods cannot naturally integrate such data and we choose to extend the our recent model, the mixed-variate restricted Boltzmann machine (MV.RBM) [2]. The mixed-variate RBM uncovers *latent profile* factors, enabling subsequent clustering. Using a cohort of $6,931$ chronic diabetes patients with data from 2007 to 2011, we collect $3,178$ diagnosis codes (treated as repeated categories) and combine it with region-of-birth (as categories) and age (as Gaussian variables) to form our dataset. We show clustering results obtained from running affinity propagation (AP) [3], containing 10 clusters and qualitatively evaluate the disease codes of groups. We demonstrate that the mixed-variate RBM followed by AP outperforms all baseline methods – Bayesian mixture model and affinity propagation on the original diagnosis codes, and k-means and AP on latent profiles, discovered by just the plain RBM [4].

Predicting disease codes for future years enables hospitals to prepare finance, equipment and logistics for individual requirements of patients. Thus prediction of disease codes forms the next part of our study. Using the mixed-variate RBM and the dataset described above, we demonstrate that our method outperforms other methods, establishing the versatility of the latent profile discovery with mixed-variate RBM.

In short, our main contributions are: (i) Novel extension and application of a powerful data mining tool, mixed-variate RBM, to a complex hospital chronic disease dataset, for clustering and understanding of disease codes within subgroups; (ii) Disease code prediction, using the model; and (iii) Demonstration of the method and showing that it outperforms baseline models, in both clustering and prediction on this complex data.

The significance of our work is to build a framework that is able to support healthcare centres and clinicians delivering outcomes that can integrate with their operations to enhance clinical efficiencies. Using such systems, the management and supervision on diabetes patients in particular as well as other kinds of diseases patients in general would have the potential to improve.

The rest of paper is organized as follows. The next section presents our patient profile modelling framework. Next, we describe our implementation on the diabetes cohort and demonstrate efficiencies of our methods. Section 4 discusses related work and modelling choices as well as the other potentials of the proposed framework, followed by conclusions in Section 5.

2 Latent Patient Profiling

A patient profile in modern hospitals typically consists of multiple records including demographics, admissions, diagnoses, surgeries, pathologies and medication. Each record contains several fields, each of which is type-specific. For example, age can be considered as a continuous quantity but a diagnosis code is a discrete element. At the first approximation, each patient can be represented by

using a long vector of mixed types[1]. However, joint modelling of mixed types is known to be highly challenging even for a small set of variables [5,6]. Complete patient profiling, on the other hand, requires handling of thousands of variables. Of equal importance is that the profiling should readily support a variety of clinical analysis tasks such as patient clustering, visualisation and disease prediction. In what follows, we develop a representational and computational scheme to capture such heterogeneity in an efficient way. In particular, we extend the our recently introduced machinery known as mixed-variate restricted Boltzmann machine (MV.RBM) [2] for the task.

2.1 Mixed-Variate Restricted Boltzmann Machines

A restricted Boltzmann machine (RBM) is a bipartite undirected graphical model with two layers, where the input layer consists of visible units and the other layer the binary hidden units [7]. See, for example, Fig. 1 for an illustration. A mixed-variate RBM is a RBM with inhomogeneous input units, each of which has the own type. More formally, let $\boldsymbol{v}$ denote the joint set of visible variables: $\boldsymbol{v} = (v_1, v_2, ..., v_N)$, $\boldsymbol{h}$ the joint set of binary hidden units: $\boldsymbol{h} = (h_1, h_2, ..., h_K)$, where $h_k \in \{0, 1\}$ for all k. Each visible unit encodes type-specific information, and the hidden units capture the *latent factors* not presented in the observations. Thus the MV.RBM can be seen as a way to transform inhomogeneous observational record into a *homogeneous representation* of the patient profile. Another way to view this as a mixture model where there are 2^K mixture components. This capacity is arguably important to capture all factors of variation in the patient cohort.

The MV.RBM defines a Boltzmann distribution over all variables: $P(\boldsymbol{v}, \boldsymbol{h}; \psi) = e^{-E(\boldsymbol{v},\boldsymbol{h})}/Z(\psi)$, where $E(\boldsymbol{v}, \boldsymbol{h})$ is model energy, $Z(\psi)$ is the normalising constant and ψ is model parameter. In particular, the energy is defined as

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\left(\sum_i F_i(v_i) + \sum_i a_i v_i + \sum_k b_k h_k + \sum_{ik} h_k W_{ik} v_i\right) \quad (1)$$

where $\boldsymbol{a} = (a_1, a_2, .., a_N), \boldsymbol{b} = (b_1, b_2, ..., b_K)$ are biases of visible and hidden units, and $\boldsymbol{W} = [W_{ik}]$ represents the weights connecting hidden and visible units, and $F_i(v_i)$ are type-specific function. The bipartite structure allows conditional independence among intra-layer variables which lead to the following factorisations:

$$P(\boldsymbol{v} \mid \boldsymbol{h}) = \prod_{i=1}^{N} P(v_i \mid \boldsymbol{h}) \quad (2) \qquad P(\boldsymbol{h} \mid \boldsymbol{v}) = \prod_{k=1}^{K} P(h_k \mid \boldsymbol{v}) \quad (3)$$

The conditional separation of types in Eq. (2) is critical: This allows independent specification of type-specific data generative models and at the same

[1] Since each field may be repeated over time (e.g., diagnosis codes), we need an aggregation scheme to summarize the field. Here we use the simple counting for diagnosis codes.

time achieves higher-order dependencies through the "pooling" layer $\boldsymbol{h}$. For example, let $f_i(\boldsymbol{h}) = a_i + \sum_k W_{ik} h_k$, the *binary* units would be specified as: $P(v_i \mid \boldsymbol{h}) = 1/\left(1 + e^{-f_i(\boldsymbol{h})}\right)$ (i.e., $F_i(v_i) = 0$); the *Gaussian* units: $P(v_i \mid \boldsymbol{h}) = \mathcal{N}\left(\sigma_i^2 f_i(\boldsymbol{h}); \sigma_i\right)$ (i.e., $F_i(v_i) = -v_i^2/2\sigma_i^2$), and the *categorical* units: $P(v_i \mid \boldsymbol{h}) = e^{f_i(\boldsymbol{h})} / \sum_j e^{f_j(\boldsymbol{h})}$.

The model is typically estimated by maximising the data log-likelihood $\mathcal{L} = \log P(\boldsymbol{v}; \psi) = \log \sum_{\boldsymbol{h}} P(\boldsymbol{v}, \boldsymbol{h}; \psi)$. The parameters are updated in a gradient ascent fashion as follows:

$$\psi \leftarrow \psi + \nu \left(\mathbb{E}_{\boldsymbol{v},\boldsymbol{h}} \left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \psi} \right] - \mathbb{E}_{\boldsymbol{h}|\boldsymbol{v}} \left[\frac{\partial E(\boldsymbol{v}, \boldsymbol{h})}{\partial \psi} \right] \right) \tag{4}$$

for some learning rate $\nu > 0$. Here $\mathbb{E}_{\boldsymbol{v},\boldsymbol{h}}$ denotes the expectation with respect to the full model distribution $P(\boldsymbol{v}, \boldsymbol{h}; \psi)$, $\mathbb{E}_{\boldsymbol{h}|\boldsymbol{v}}$ the conditional distribution given the known $\boldsymbol{v}$. Whilst the conditional expectation can be compute efficiently, the full expectation is generally intractable. Thus we must resort to approximate methods, and in this paper, we choose a truncated MCMC-based method known as contrastive divergence (CD) [8] as it proves to be fast and accurate. A MCMC chain is obtained by alternating between $\hat{\boldsymbol{v}} \sim P\left(\boldsymbol{v} \mid \hat{\boldsymbol{h}}\right)$ and $\hat{\boldsymbol{h}} \sim P(\boldsymbol{h} \mid \hat{\boldsymbol{v}})$.

2.2 MV.RBM for Patient Profiling

The goal of patient profiling is to construct an effective *personal representation* from multiple hospital records. Here we focus mainly on patient demographics (e.g., *age*, *gender* and *region-of-birth*) and their existing health conditions (e.g., existing *diagnoses*). For simplicity, we consider a binary gender (male/female). Further, age can be considered as a continuous quantity and thus a Gaussian unit can be used[2]; and region-of-birth and diagnosis as categorical variables. However, since the same diagnosis can be repeated during the course of readmissions, it is better to include them all. In particular, we adopt the idea from the "replicated softmax" [4] where repeated diagnoses share the same parameters. In the end, we build one MV.RBM per patient due to the difference in the diagnosis sets. Further, to balance the contribution of the hidden units against the variation in input length, it is important to make a change to the energy model in Eq. (1) as follows: $D\boldsymbol{b} \leftarrow \boldsymbol{b}$ where D is the total number of input variables for each patient. We note that these parameter sharing and balancing are not readily present in the current MV.RBM of Truyen et al [2].

Once the model has been estimated, the *latent profiles* are generated by computing the posterior vector $\hat{\boldsymbol{h}} = \left(P\left(h_1^1 \mid \boldsymbol{v}\right), P\left(h_2^1 \mid \boldsymbol{v}\right), ..., P\left(h_K^1 \mid \boldsymbol{v}\right)\right)$, where $P\left(h_k^1 \mid \boldsymbol{v}\right)$is a shorthand for $P(h_k = 1 \mid \boldsymbol{v})$ – the probability that the k-th latent factor is activated given the demographic and clinical input $\boldsymbol{v}$:

[2] Although the distribution of ages for a particular disease is generally not Gaussian, our model is a mixture of exponentially many components (2^K, see Sec. 2.1 for detail), and thus can capture any distribution with high accuracy.

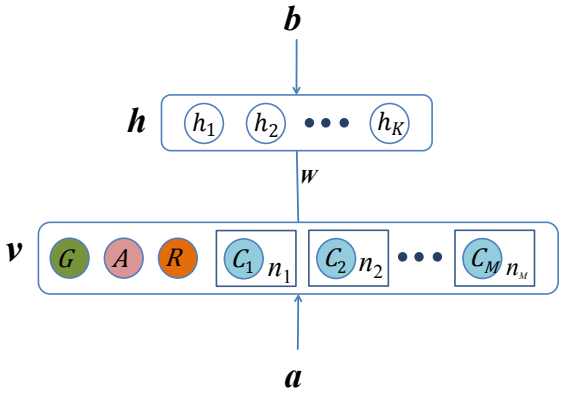


Fig. 1. Patient profiling using mixed-variate RBMs. The top layer represents stochastic binary units. The bottom layer encodes multiple type-specific inputs: A for age (continuous), G for gender (binary), R for region-of-birth, C_k for diagnosis codes. The circles within squares denote the replicated diagnosis codes (categorical) where the integers $\{n_k\}$ denotes the number of replications.

$$P\left(h_k^1 \mid \boldsymbol{v}\right) = \frac{1}{1 + \exp\left\{-Db_k - \sum_i W_{ik} v_i\right\}}.$$

As we will then demonstrate in Section 3, the latent profile can be used as input for a variety of analysis tasks such as patient clustering and visualisation.

Interestingly, the model also enables a certain degree of *disease prediction*, i.e., we want to guess which diagnoses will be positive for the patient in the future[3]. Although this may appear to be an impossible task, it is plausible statistically because some diseases are highly correlated or even causative, and there are certain pathways that a disease may progress. More specifically, subset of diagnoses at time $t+1$ can be predicted by searching for the mode of the following conditional distribution:

$$P\left(\boldsymbol{v}^{(t+1)} \mid \boldsymbol{v}^{(1:t)}\right) = \sum_{\boldsymbol{h}} P\left(\boldsymbol{v}^{(t+1)}, \boldsymbol{h} \mid \boldsymbol{v}^{(1:t)}\right).$$

Unfortunately the search is intractable as we need to traverse through the space of all possible disease combinations, which has the size of 2^M where M is the set of diagnosis codes. To simplify the task and to reuse of the readily discovered latent profile $\hat{\boldsymbol{h}}^{(1:t)}$, we assume that (i) the model distribution is not changed due to the "unseen" future, (ii) the latent profile at this point captures everything we can say about the state of the patient, and (iii) future diseases are conditionally

[3] Although this appears to resemble the traditional collaborative filtering, it is more complicated since diseases may be recurrent, and the strict temporal orders must be observed to make the model clinically plausible.

independent given the current latent profile. This leads to the following *mean-field* approximation[4]:

$$P\left(v_j^{(t+1)} \mid \boldsymbol{v}^{(1:t)}\right) \approx \frac{\exp\left\{a_j + \sum_k W_{jk} P\left(h_k^1 \mid \boldsymbol{v}^{(1:t)}\right)\right\}}{\sum_i \exp\left\{a_i + \sum_k W_{ik} P\left(h_k^1 \mid \boldsymbol{v}^{(1:t)}\right)\right\}}. \tag{5}$$

3 Implementation and Results

In this section we present the analysis of patient profiles using the data obtained from Barwon Health, Victoria, Australia[5], during the period of 2007 − 2011 using the extended MV.RBM described in Section 2. In particular, we evaluate the capacity of the MV.RBM for patient clustering and for predicting future diseases. For the former task, the MV.RBM is can be seen as a way to transform complex input data into a homogeneous vector from which post-processing steps (e.g., clustering and visualisation) can take place. For the prediction task, the MV.RBM acts as a classifier that map inputs into outputs.

Our main interest is in the *diabetes* cohort of 7,746 patients. There are two types of diabetes: Type I is typically present in younger patients who are not able to produce insulin; and Type II is more popular in the older group who, on the other hand, cannot adsorb insulin. One of the most important indicators of diabetes is the high blood sugar level compared to the general population. Diabetes are typically associated with multiple diseases and complications: The cohort contains 5,083 distinct diagnosis codes, many of which are related to other conditions and diseases such as obesity, tobacco use and heart problems. For robustness, we remove those rare diagnosis codes with less than 4 occurrences in the data. This results in a dataset of 6,931 patients who originally came from 102 regions and were diagnosed with totally 3,178 unique codes. The inclusion of age and gender into the model is obvious: they are not only related to and contributing to the diabetes types, they are also largely associated with other complications. Information about the regions-of-origin is also important for diabetes because it is strongly related to the social conditions and lifestyles, which are of critical importance to the proactive control of the blood sugar level, which is by far the most cost-effective method to mitigate diabetes-related consequences.

3.1 Implementation

Continuous variables are first normalised across the cohort so that the Gaussian inputs have zero-means and unit variances. We employ 1-step contrastive divergence (CD) [8] for learning. Learning rates vary from type to type and they are chosen so that reconstruction errors at each data sweep are gradually reduced.

[4] This result is obtained by first disconnecting the future diagnosis-codes from the latent units and then find the *suboptimal* factorised distribution $Q\left(\boldsymbol{v}^{(t+1)}, \boldsymbol{h} \mid \boldsymbol{v}^{(1:t)}\right) = \prod_j Q_j\left(v_j^{(t+1)} \mid \boldsymbol{v}^{(1:t)}\right) \prod_k P\left(h_k \mid \boldsymbol{v}^{(1:t)}\right)$ that minimises the Kullback-Leibner divergence from the original distribution $P\left(\boldsymbol{v}^{(t+1)}, \boldsymbol{h} \mid \boldsymbol{v}^{(1:t)}\right)$.

[5] Ethics approval 12/83.

Parameters are updated after each mini-batch of 100 patients, and learning is terminated after 100 data sweeps. The number of hidden units is determined empirically to be 200 since large size does not necessarily improve the clustering/prediction performance.

For *patient clustering*, once the model has been learned, the hidden posteriors that are computed using Eq. (3) can be used as the new representation of the data. To enable fast bitwise implementation (e.g., see [9]), we then convert the continuous posteriors into binary activation as follows: $\hat{h}_k = 1$ if $P(h_k^1 \mid \boldsymbol{v}) \geq \rho_1$ and $\hat{P}_k = 0$ otherwise for all $k = 1, 2.., K$ and some threshold $\rho_1 \in (0, 1)$. We then apply well-known clustering methods including affinity propagation (AP) [3], k-means and Bayesian mixture models (BMM). The AP is of particular interest for our exploratory analysis because it is capable of automatically determining the number of clusters. It requires the similarity measure between two patients, and in our binary profiles, a natural measure is the Jaccard coefficient:

$$J\left(p, q\right) = \frac{\mid S\left\{p\right\} \cap S\left\{q\right\} \mid}{\mid S\left\{p\right\} \cup S\left\{q\right\} \mid} \tag{6}$$

where $S\{p\}$ is the set of activated hidden units for patient p. Another hyper-parameter is the so-called 'preference' which we empirically set to the average of all pairwise similarities multiplied by -20. This setting gives a reasonable clustering.

The other two clustering methods require a prior number of clusters, and here we use the output from the AP. For the k-means, we use the the Hamming distance between activation vectors of the two patients[6]. The BMM is a Bayesian model with multinomial emission probability.

The task of *disease prediction* is translated into predicting diagnoses in the future for each patient. We split data into 2 subsets: The earlier subset, which contains those diagnoses in the period of $2007 - 2010$, is used to train the MV.RBM; and the later subset is used to evaluate the prediction performance. In the MV.RBM, we order the future diseases according to the probability that the disease occurs as in Eq. (5).

3.2 Patient Clustering

First we wish to validate that the latent profiles discovered by the MV.RBM are informative enough so that *clinically meaningful* clusters can be formed. Fig. 2 shows the 10 clusters returned by the AP and the similarity between every patient pair (depicted in colour, where the similarity increases with from blue to red). It is interesting that, out of 10 groups, we are able to discover a group whose conditions are mostly related to Type I diabetes (Figs. 3a and 3b), and another group associated with Type II diabetes (Figs. 4a and 4b). The grouping properties can be examined visually using a visualisation tool known as t-SNE [10] to project the latent profiles onto 2D. Fig. 5a depicts the distribution of patients, where the colours are based on the group indices assigned earlier by the AP.

[6] The centroid of each cluster is chosen according to the median elementwise.

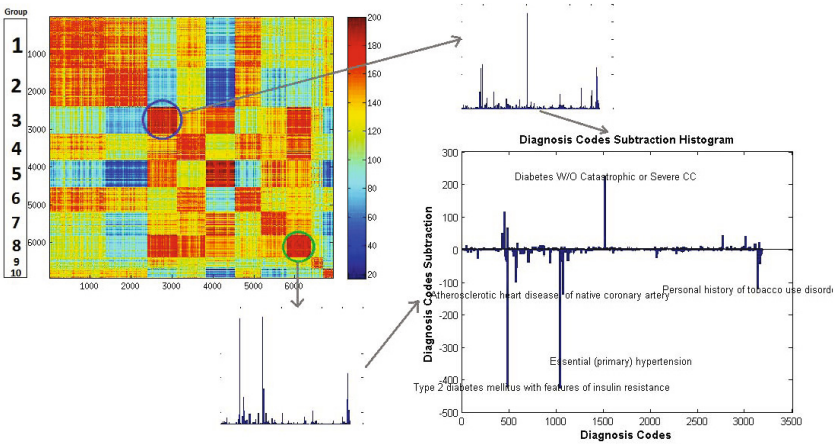
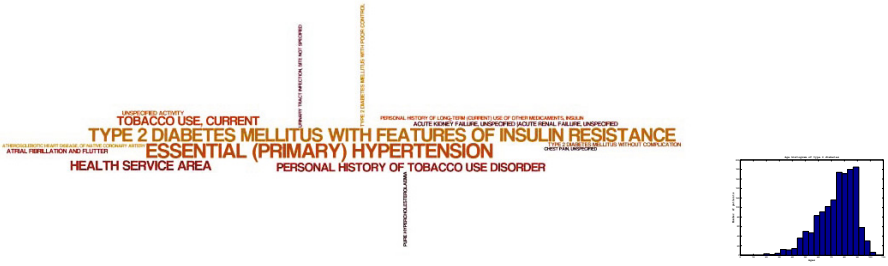


Fig. 2. Similarity matrix and diagnosis codes histograms. The matrix represents resemblances of pairwise patients while histograms show quantity of diagnoses. Group 3 and Group 8 look highly overlapping at the diagnosis level (top-left figure), but in fact, their clinical conditions are significantly different when we subtract the two histograms (lower-right figure). (Best viewed in colors).

(a) Tag cloud of diagnosis descriptions. (b) Age histogram.

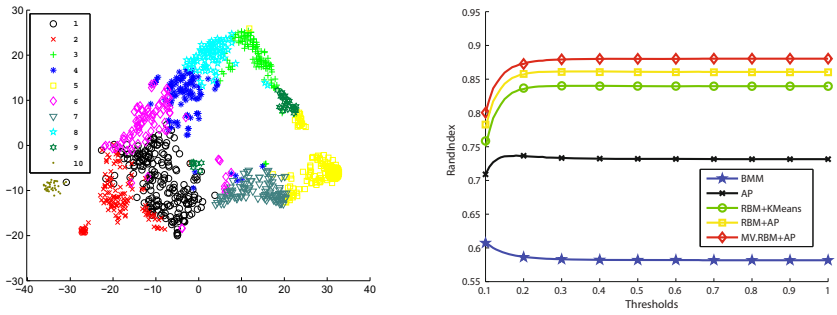
Fig. 3. Type I diabetes mellitus: Primary diagnoses and age distribution. Two figures confirms the existing knowledge that Type I diabetes mellitus often occurs in the younger population.



(a) Tag cloud of diagnosis descriptions. (b) Age histogram.

Fig. 4. Type II Diabetes mellitus: Primary diagnoses and age distribution. We can see that the age distribution is distinct from the Type I group.

For quantitative evaluation, we calculate the *Rand-index* [11] to assess the quality of resulting clusters, given that we do not have cluster labels. The Rand-index is the pairwise accuracy between any two patients. To judge whether two patients share the same cluster, we consult the diagnosis code hierarchy of the ICD-10 [12]. We use hierarchical assessment since a diagnosis code may have multiple levels. E11.12, for example, has two levels: E11 and E11.12. The lower level code specifies disease more clearly whilst the higher is more abstract. Therefore we have two ways for pairwise judgement: the Jaccard coefficient (Section 3.1) and code 'cluster' which is the grouping of codes that belong to the same disease class, as specified by the latest WHO standard ICD-10. At the lowest level, two patients are considered similar if the two corresponding code sets are sufficiently overlapping, i.e., their Jaccard coefficient is greater than a threshold $\rho_2 \in (0, 1)$. At higher level, on the other hand, we consider two patients to be clinically similar if they share higher level diabetes code of the same code 'cluster'. For instance, two patients with two codes E11.12 and E11.20 are similar at the level E11[7], but they are dissimilar at the lower level. Note that this hierarchical division is for evaluation only. We use codes at the lowest level as replicated softmax units in our model (Section 2.2).



(a) (b)

Fig. 5. Visualisation and quantitative assessment of clusters. (a) t-SNE [10] projection on 2,000 latent profiles. Groups are labelled by the outputs of the AP. (Best viewed in colors). (b) Rand-index curves in patient clustering. *AP*: affinity propagation, *BMM*: Bayesian mixture model, *RBM*: MV.RBM with diagnosis codes only.

Fig. 5b reports the Rank-indices with respect to the assessment at the lowest level in the ICD-10 hierarchy for clustering methods with and without MV.RBM pre-processing. At the next ICD-10 level, the MV.RBM/AP achieves a Rand-index of 0.6040, which is, again, the highest among all methods, e.g., using the RBM/AP yields the score of 0.5870, and using AP on diagnosis codes yields 0.5529. This clearly demonstrates that (i) MV.RBM latent profiles would lead to better clustering that those using diagnosis codes directly, and (ii) modelling mixed-types would be better than using just one input type (e.g., the diagnosis codes).

[7] This code group is for non-insulin-dependent *diabetes mellitus*.

3.3 Disease Prediction

The prediction results are summarised in Fig. 6, where the ROC curve of the MV.RBM is compared against that of the baseline using k-nearest neighbours (k-NN). The k-NN approach to disease ranking is as follows: For each patient, a neighbourhood of the 50 most similar patients is collected based on the Jaccard coefficient over sets of unique diagnoses. The diagnoses are then ranked according to their occurrence frequency within the neighbourhood. As can be seen from the figure, the latent profile approaches outperform the k-NN method. The MV.RBM with contextual information such as age, gender and region-of-birth proves to be useful. In particular the the areas under the ROC curve (AUC) of the MV.RBMs are 0.84 (with contextual information) and 0.82 (without contextual information). These are clearly better than the score obtained by k-NN, which is 0.77.

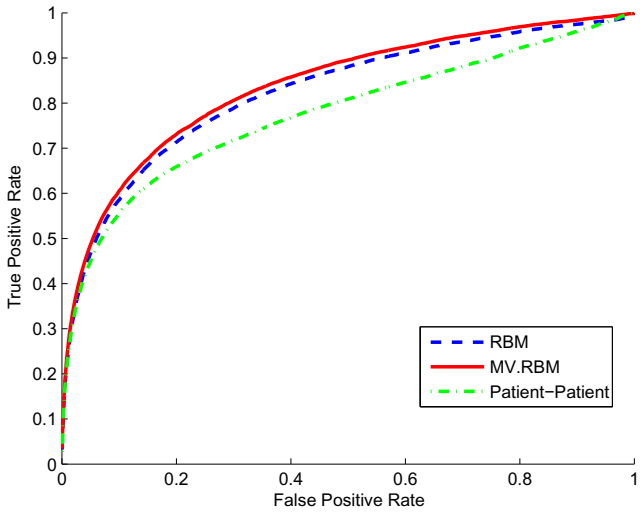


Fig. 6. ROC curves in disease prediction. *RBM* is MV.RBM with diagnosis codes only; *Patient-Patient* is the k-nearest neighbours method. Best viewed in colors.

4 Discussion and Related Work

This work is part of our ongoing effort to apply data mining and statistical techniques to understand the complex health databases in order to improve the services efficiency within health organisations and across coordinated networks. This line of research has recently attracted considerable interest in the data mining community (e.g., see [13,14]). Our focus on diabetes is motivated by the pressing demands to deliver personalised cares for the large population on an ongoing basis [15,16]. However, we wish to emphasize that the approach is quite general and could be applicable to other cohorts.

In terms of modelling, our work adds a few more flavours to the current mixed-variate analysis in biomedical domains [5,6,17]. The existing literature offers three approaches: The first is to specify the direct type-specific conditional relationship between two variables (e.g., see [5]), the second is to assume that each observable is generated from a latent variable (latent variables then encode the dependencies) (e.g., see [6]), and the third is to construct joint cumulative distributions using copula [18,17]. The drawback of the first approach is that it requires far more domain knowledge and statistical expertise to design a correct model even for a small number of variables. The second approach lifts the direct dependencies to the latent variables level. All approaches are, however, not very scalable to realistic setting of the hospital records. Our treatment using MV.RBM [2], along with the line of work using RBMs for representing complex data [19,20,21], offers the *fourth alternative*: Direct pairwise dependencies are substituted by indirect long-range dependencies. Not only this simplifies the model design, the inference is much more scalable: each MCMC sweep through all variables takes only linear time. Our most recent work in [21], while enjoying the similar computational efficiency, offers a better interpretation through the use of latent variables to capture the generative mechanism of data types.

Latent profiling could be important for other applications such as patient retrieval, i.e., we want to retrieve patients with clinically similar conditions to the patient under study. In this setting, using raw diagnosis codes may miss those whose codes are different from the present patient even if they share the same clinical conditions. The use of MV.RBM, on the other hand, would project these patients onto similar latent profiles. It is also of interest to ask whether it is justifiable for the choice of parameter sharing for repeated diagnoses. To get answer, we experimented with the "counting" treatments in which each code is considered as a Poisson variable, and our clustering/prediction results indicate that it is much better to employ the parameter sharing trick. This may due to the fact that under the Poisson treatment, diagnoses are assumed to "arrive" independently, while in reality diagnoses are generally correlated.

5 Conclusion

We have presented a latent profiling framework using our recently introduced architecture known as mixed-variate restricted Boltzmann machines (MV.RBM). The goal was to develop a representational and computational scheme that can handle complex, inhomogeneous data from real hospital settings. The MV.RBM was adapted to handle recurrent diagnoses by parameter sharing and variable balancing. We evaluated this scheme on a cohort of complex diseases such as diabetes, where there are many influential factors, and diagnoses are often repeated, correlated and causative. It is demonstrated that the chosen scheme is highly effective for exploratory tasks such as patient clustering and visualisation and predictive tasks such as 1-year diagnosis prognosis.

References

1. World Health Organization: Diabetes (2012), http://www.who.int/mediacentre/factsheets/fs312/en/index.html (accessed September 2012)
2. Tran, T., Phung, D.Q., Venkatesh, S.: Mixed-variate restricted Boltzmann machines. Journal of Machine Learning Research - Proceedings Track 20, 213–229 (2011)
3. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315, 972–976 (2007)
4. Salakhutdinov, R., Hinton, G.: Replicated softmax: an undirected topic model. Advances in Neural Information Processing Systems 22, 1607–1614 (2009)
5. McCulloch, C.: Joint modelling of mixed outcome types using latent variables. Statistical Methods in Medical Research 17(1), 53 (2008)
6. Dunson, D., Herring, A.: Bayesian latent variable models for mixed discrete outcomes. Biostatistics 6(1), 11 (2005)
7. Freund, Y., Haussler, D.: Unsupervised learning of distributions on binary vectors using two layer networks. Santa Cruz, CA, USA. Tech. Rep. (1994)
8. Hinton, G.E.: Training products of experts by minimizing contrastive divergence. Neural Computation 14(8), 1771–1800 (2002)
9. Salakhutdinov, R., Hinton, G.: Semantic hashing. In: SIGIR Workshop on Information Retrieval and Applications of Graphical Models, vol. 500(3). ACM Special Interest Group on Information Retrieva (2007)
10. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605 (2008)
11. Rand, W.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 846–850 (1971)
12. World Health Organization: ICD-10th (2010), http://apps.who.int/classifications/icd10/browse/2010/en (accessed September 2012)
13. Khosla, A., Cao, Y., Lin, C.C.-Y., Chiu, H.-K., Hu, J., Lee, H.: An integrated machine learning approach to stroke prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 183–192. ACM (2010)
14. Luo, D., Wang, F., Sun, J., Markatou, M., Hu, J., Ebadollahi, S.: Sor: Scalable orthogonal regression for non-redundant feature selection and its healthcare applications. In: SIAM Data Mining Conference (2012)
15. Ben-Hur, A., Iverson, T., Iyer, H.: Predicting the risk of type 2 diabetes using insurance claims data. In: Neural Information Processing System Foundation (2010)
16. Neuvirth, H., Ozery-Flato, M., Hu, J., Laserson, J., Kohn, M.S., Ebadollahi, S., Rosen-Zvi, M.: Toward personalized care management of patients at risk: the diabetes case study. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 395–403. ACM (2011)
17. de Leon, A.R., Wu, B.: Copula-based regression models for a bivariate mixed discrete and continuous outcome. Statistics in Medicine 30(2), 175–185 (2011)
18. Song, P.X.-K., Li, M., Yuan, Y.: Joint regression analysis of correlated data using gaussian copulas. Biometrics 65(1), 60–68 (2009)
19. Truyen, T., Phung, D., Venkatesh, S.: Ordinal Boltzmann machines for collaborative filtering. In: Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, Canada (June 2009)

20. Tran, T., Phung, D., Venkatesh, S.: Cumulative restricted Boltzmann machines for ordinal matrix data analysis. In: Proc. of 4th Asian Conference on Machine Learning (ACML), Singapore (2012)
21. Tran, T., Phung, D., Venkatesh, S.: Embedded Restricted Boltzmann Machines for fusion of mixed data type and applications in social measurements analysis. In: Proc. of the 15th International Conference on Information Fusion (FUSION), Singapore (2012)