ECAI 2012
Luc De Raedt et al. (Eds.)

doi:10.3233/978-1-61499-098-7-97

Multirelational Consensus Clustering with Nonnegative Decompositions

Liviu Badea[1]

Abstract. Unsupervised multirelational learning (clustering) in non-sparse domains such as molecular biology is especially difficult as most clustering algorithms tend to produce distinct clusters in slightly different runs (either with different initializations or with slightly different training data).

In this paper we develop a *multirelational consensus clustering* algorithm based on nonnegative decompositions, which are known to produce sparser and more interpretable clusterings than other data-oriented algorithms.

We apply this algorithm to the joint analysis of the largest available gene expression datasets for leukemia and respectively normal hematopoiesis in order to develop a more comprehensive genomic characterization of the heterogeneity of leukemia in terms of 38 normal hematopoietic cell states. Surprisingly, we find unusually complex expression programs involving large numbers of transcription factors, whose further in-depth analysis may help develop personalized therapies.

1 Introduction

We are currently witnessing an explosion of multi-relational data in many real-life application domains, such as social network analysis, bioinformatics, Web mining, collaborative filtering and many more [16].

Learning is especially difficult in complex domains because the combinatorial explosion of hypotheses hugely exceeds the existing data that may discriminate between them. While this phenomenon already manifests itself in the single relation case, it is significantly more difficult to deal with in a multi-relational setting.

Therefore, there has been a significant recent increase in interest in learning from such multi-relational data [16], but most work has focused on supervised approaches (e.g. classification). However, unsupervised, discovery-based settings are of equal importance, despite having received relatively less attention due to the inherent difficulty in evaluating their results [9].

In this paper, we present an unsupervised data-oriented approach to multi-relational learning for discovery in leukemia biology.

Among the many different approaches that have been tried in the domain of multirelational learning, including probabilistic (Statistical Relational Learning), logical (Inductive Logic Programming) and data-oriented, we have concentrated on the last type of approach, as it can better deal with the *non-sparse numerical relations* that predominate in the field of high-throughput molecular biology.

Many relational domains, such as the link structure of the Web, or various collaborative filtering settings (e.g. movie recommendation), involve (relatively) *sparse* relations. On the other hand, the interaction networks encountered in molecular biology have the small-world property, so that almost any pair of nodes is connected via a relatively short chain of links. Moreover, many of these relations are best represented using numerical features rather than using a logical or probabilistic representation. For example, gene expression matrices in genomics are best represented as full (rather than sparse) numerical matrices encoding the expression levels of individual genes in specific samples of well-defined biological phenotypes.

While extracting a simple characterization of a set of *sparse* relations is complicated[2], inferring such a simple model for a set of almost "fully connected" relations is truly daunting.

A typical application in genomics of complex diseases is finding a molecular-level characterization of the disease and predicting its evolution using high-throughput data (such as gene expression microarrays) and related biological knowledge on gene/protein interactions and pathways. Unfortunately however, complex diseases (such as cancer) are quite heterogeneous at a molecular level, so that virtually every patient is essentially a unique case. Therefore, for many types of cancer it has been impossible to determine good predictors of disease evolution, despite numerous attempts with the best supervised (classification) techniques. Thus, if *direct* prediction of evolution is sometimes too difficult for the entire population of patients, it may be of interest to break down the problem by characterizing the most important disease subtypes in an unsupervised, or semi-supervised manner. Much less work has addressed this problem in a multi-relational setting, perhaps especially due to the extreme difficulty of validation, which involves in-depth expert knowledge and cannot simply rely on the traditional validation methods used in a supervised setting [9].

Leukemias are a very heterogeneous group of cancers of the hematopoietic system. Recent large-scale genomic studies, such as the Microarray Innovations in Leukemia (MILE) study [2] have made the expression profiles of 2096 patients publicly available and have shown that the current clinical subclassification (involving 18 subtypes) can be accurately recovered from the genomic profiles. Unfortunately however, achieving a detailed molecular-level understanding of the various leukemia subtypes is still a goal for the future, mostly because of the disease heterogeneity.

This heterogeneity can be explained by the very large genomic and transcriptomic variability of the normal hematopoietic cell compartment (which is comparable to the variability of the entire repertoire of human cell types [13]), given the fact that leukemias are diseases of the hematopoietic stem cells.

Some of the simplest data-oriented unsupervised learning methods involve dimensional reduction methods such as matrix factoriza-

[1] AI group, National Institute for Research in Informatics, Romania, email: badea@ici.ro

[2] Due to the combinatorics involved.

tions. *Nonnegative Matrix Factorization (NMF)* in particular tends to produce *sparse* and *domain-interpretable*[3] decompositions, within an extremely simple computational framework [8]. While a large number of gene expression studies employing matrix factorization in general and NMF in particular have been put forward (e.g. [1, 3]), only very few have been able to exploit the inherent multi-relational structure of the domain (e.g. [7, 5]).

Moreover, unsupervised learning (clustering) is prone to instability (or ambiguity) especially in multi-relational domains, where different runs of a given algorithm (either with different initializations or with slightly different training data) tend to produce distinct results (clusters). Preliminary investigations of the MILE study gene expression data with various clustering algorithms have emphasized clustering instability as the main obstacle toward determining a detailed *genomic* subclassification of leukemias.[4]

In this paper we introduce a *multi-relational consensus clustering* method that is able to deal with the inherent instability of multi-relational clustering and apply it to the problem of unsupervised leukemia subclassification.

Developing a consensus clustering algorithm for multi-relational decompositions is highly nontrivial. Typical consensus clustering systems [12] construct a square consensus matrix that records for each pair of items the frequency of their co-clustering. Unfortunately, this simple idea only works for unidimensional clustering, while multi-relational decompositions produce biclusters (two-way clusters).

In order to better understand the relationships of the leukemia subtypes with the normal hematopoietic cell types, we have performed a simultaneous clustering of the MILE leukemia dataset [2] (the largest transcriptomic dataset for leukemia) with the largest transcriptomic dataset of normal hematopoietic cell types [13] (which contains transcriptomic data for 211 samples of 38 distinct cell types, including hematopoietic *stem cells*).

More precisely, we are searching for gene expression modules that are shared between leukemia and certain normal hematopoietic cells, as well as for the specific differences between leukemia and normal hematopoiesis.

The paper is organized as follows. After a more formal introduction of multirelational nonnegative decompositions, we present a simple multiplicative update algorithm for inferring such factorizations. We then develop a consensus clustering algorithm based on a Positive Tensor Factorization [17] of several individual runs of the base algorithm. The consensus clustering algorithm is subsequently applied to leukemia subclassification. The paper concludes with a short discussion of the results as well as with a brief mention of related works.

2 Multirelational learning via Nonnegative Matrix Factorization (MNMF)

We start by presenting the framework of multirelational learning using nonnegative decompositions.

A *multirelational domain* involves a set of *entity types* $\{\mathcal{E}^{(n)}\}_n$ as well as a set of numerical relations $\{R^{(mn)}\}_{mn}$ between these entity types. An *entity type* $\mathcal{E}^{(n)}$ is a set of N_n related entities (such as genes, documents or movies). In our setting, the nonnegative real-valued relation matrices $R^{(mn)}_{ij}$ are weighted by means of *weight matrices* $W^{(mn)}_{ij}$, which allow us to represent unknown relation entries (i, j) (by setting $W^{(mn)}_{ij} = 0$), as well as to balance relations with widely disparate variation ranges.

As already amply demonstrated in the unirelational setting by Nonnegative Matrix Factorization (NMF) [8], the nonnegativity constraints are essential for obtaining sparse and easily interpretable decompositions. Problems featuring relations with negative values can usually be reformulated in a nonnegative framework, depending on their precise semantics (see e.g. [3] for an example).

A rank N_c *multirelational nonnegative decomposition* of a multirelational structure $\langle\{\mathcal{E}^{(n)}\}_n, \{R^{(mn)}\}_{mn}, W^{(mn)}\}_{mn}\}\rangle$ is an assignment of a nonnegative factor matrix $E^{(n)}$ of size $N_n \times N_c$ to each entity type $\mathcal{E}^{(n)}$, such that all relations $R^{(mn)}$ are approximated by the product of the corresponding entity type matrices

$$R^{(mn)} \approx E^{(m)} \cdot E^{(n)T}. \tag{1}$$

More formally, we are minimizing the following weighted squared error function

$$\begin{aligned} f &= \frac{1}{2}\sum_{s,d}\left\|R^{(sd)} - E^{(s)}\cdot E^{(d)T}\right\|^2_{W^{(sd)}} \\ &= \frac{1}{2}\sum_{s,d}\sum_{i,j} W^{(sd)}_{ij}\Big(R^{(sd)}_{ij} - \sum_{c=1}^{N_c} E^{(s)}_{ic}\cdot E^{(d)}_{jc}\Big)^2 \end{aligned} \tag{2}$$

subject to nonnegativity constraints for the entity matrices $E^{(n)} \geq 0$, where $\|\cdot\|_W$ is the W-weighted Frobenius norm $\|X\|^2_W = \sum_{i,j} W_{ij}X^2_{ij}$.

A simple algorithm solving the optimization problem (2) can be developed by generalizing the method employed by Lee and Seung for standard NMF [8]. Introducing the Lagrangean $\mathcal{L} = f - \sum_n \mu^{(n)} \circ E^{(n)}$, we obtain the following *Karush-Kuhn-Tucker conditions*:

$$\frac{\partial f}{\partial E^{(n)}} - \mu^{(n)} = 0 \tag{3}$$

$$\mu^{(n)} \circ E^{(n)} = 0 \tag{4}$$

$$\mu^{(n)} \geq 0 \tag{5}$$

Explicitly splitting the gradient of the error function $\frac{\partial f}{\partial E^{(n)}}$ into a positive and a negative part:

$$\frac{\partial f}{\partial E^{(n)}} = \left(\frac{\partial f}{\partial E^{(n)}}\right)_+ - \left(\frac{\partial f}{\partial E^{(n)}}\right)_- \tag{6}$$

with $\left(\frac{\partial f}{\partial E^{(n)}}\right)_\pm \geq 0$ and then using (6) and (3) in the complementarity conditions (4), we get the fixpoint equation

$$\left[\left(\frac{\partial f}{\partial E^{(n)}}\right)_+ - \left(\frac{\partial f}{\partial E^{(n)}}\right)_-\right] \circ E^{(n)} = 0 \tag{7}$$

which can be solved by the following multiplicative update rules for $E^{(n)}$:

$$E^{(n)} \leftarrow E^{(n)} \circ \frac{\left(\frac{\partial f}{\partial E^{(n)}}\right)_-}{\left(\frac{\partial f}{\partial E^{(n)}}\right)_+} \tag{8}$$

where '$\circ$' and '$-$' represent element-wise (Hadamard) multiplication and respectively division of matrices.

[3] As opposed to Principal Component Analysis, SVD or other factorization methods which tend to produce more "holistic" decompositions.

[4] We distinguish a detailed genomic subclassification from the above mentioned "clinical"/"histopathologic" subclassification involving 18 subtypes, each of which may be heterogeneous genomically.

The gradient of the error function (2) is given by (6) and the following:

$$\left(\frac{\partial f}{\partial E^{(n)}}\right)_+ = \sum_{(s,n)\in\mathcal{R}} \left[W^{(sn)} \circ \left(E^{(s)} \cdot E^{(n)T}\right)\right]^T \cdot E^{(s)} + \sum_{(n,d)\in\mathcal{R}} \left[W^{(nd)} \circ \left(E^{(n)} \cdot E^{(d)T}\right)\right] \cdot E^{(d)} \quad (9)$$

$$\left(\frac{\partial f}{\partial E^{(n)}}\right)_- = \sum_{(s,n)\in\mathcal{R}} \left[W^{(sn)} \circ R^{(sn)}\right]^T \cdot E^{(s)} + \sum_{(n,d)\in\mathcal{R}} \left[W^{(nd)} \circ R^{(nd)}\right] \cdot E^{(d)} \quad (10)$$

where $(m,n) \in \mathcal{R}$ denotes the existence of a relation between entity types $\mathcal{E}^{(m)}$ and $\mathcal{E}^{(n)}$.

We have thus arrived at a simple algorithm for multirelational NMF (MNMF) that randomly initializes the entity matrices $E^{(n)}$ and then iteratively applies the multiplicative update rules (8) with the gradient components given by (9) and (10).

An important condition for the convergence of the algorithm is ensured by the following theorem.

Theorem 1. *The weighted error function* (2) *is nonincreasing under the multiplicative update rules* (8)*.*

The proof relies on combining the following two lemmas.

Lemma 1. *A* weighted multirelational NMF (MNMF) *problem* (2) *can be reduced to an equivalent single relation* weighted symmetric NMF *problem* $R \approx E \cdot E^T$*, minimizing* $f = \frac{1}{2}\left\|R - E \cdot E^T\right\|_W^2$ *for a symmetric matrix* R*.*

The proof of the lemma involves constructing a single relation matrix R (as well as an associated weight matrix W) with a block structure, whose block rows and columns correspond to the entity types $\mathcal{E}^{(n)}$. For each relation $R^{(mn)}$, we set the corresponding (m,n) block of R to $R^{(mn)}$ and the (n,m) block to $R^{(nm)T}$. The remaining blocks are set to zero. (Similarly, we construct a weight matrix from $W^{(mn)}$.) Figure 1 illustrates the construction on a simple example.

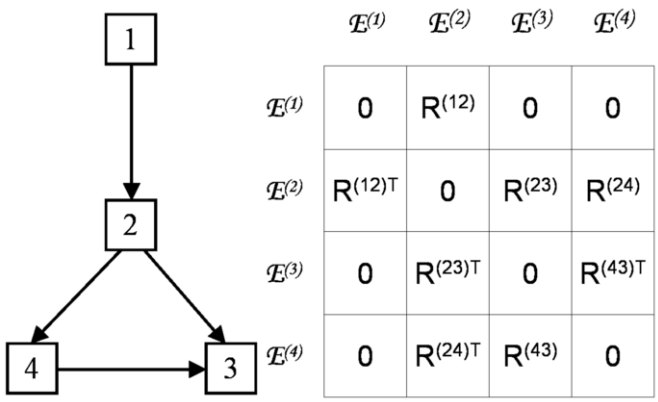

Figure 1. The symmetric matrix R associated with a multirelational structure

Lemma 2. *For a symmetric matrix* R*, the error function* $f = \frac{1}{2}\left\|R - E \cdot E^T\right\|_W^2$ *(corresponding to the weighted symmetric NMF decomposition* $R \approx E \cdot E^T$*) is nonincreasing under the update rule*

$$E \leftarrow E \circ \frac{(W \circ R) \cdot E}{[W \circ (E \cdot E^T)] \cdot E}. \quad (11)$$

The proof of this lemma closely follows the auxiliary function approach of Lee and Seung [8]. An easy analysis shows that in the case of the construction from Lemma 1, the update rule (11) decomposes into the update rules given by (8,9,10), thereby proving Theorem 1.[5]

3 Multirelational consensus clustering

Learning in domains with many variables but small sample sizes is notoriously difficult. Unsupervised learning (clustering) in such domains tends to produce *unstable* clusters, which vary from run to run, depending on slight changes in the training data or in the initialization of the algorithm. Such small sample sizes compared to the number of variables turn up in many domains. For example, gene expression data record the expression of virtually all genes in a given biological sample. However, the number of genes (around 20,000) significantly exceeds even the largest sample sizes (hundreds or at most a couple of thousand samples, in the case of the MILE study).

The situation only worsens in most real-life multirelational settings, where obtaining sufficiently large sample sizes (as compared to the number of variables) is complicated by the need to gather coherent data across the relevant relations. For example, if we intend to combine gene expression with mutation data for a certain disease, it is of crucial importance that the data comes from the same set of patients. But even in this case, the number of variables increases and the instability of the clustering algorithms worsens.

Consensus clustering refers to a family of approaches that tends to alleviate clustering instability by searching for items that cluster together in a significant number of runs. A typical consensus clustering approach [12] constructs a consensus matrix, which for each pair (i_1, i_2) of items records the percentage $C(i_1, i_2)$ of runs in which they have ended up in the same cluster.

Unfortunately, this simple approach designed for unidimensional clustering cannot be easily generalized to clustering methods based on matrix factorization, which produce two-way clusters (biclusters).

An elegant method of *consensus clustering of biclusters*, put forward in [6], uses *Positive Tensor Factorization* [17] for clustering the biclusters obtained in a number of different factorization runs.

In the following, we generalize this approach to the multirelational setting. We start with a number N_r of different runs of the multirelational MNMF algorithm, which is assumed to have produced N_r individual factorizations $\{E_r^{(n)}\}_{\substack{n=1,\ldots,N_e \\ r=1,\ldots,N_r}}$ (index n refers to the entity type, while r refers to the run). $E_r^{(n)}$ are entity matrices whose entries $E_{icr}^{(n)}$ denote the membership of entity i (having entity type n) to cluster c of run r.

A *consensus clustering* corresponds to

- a set of *consensus entity matrices* $e_{ik}^{(n)}$ (with i an entity and $k \in \{1,\ldots,N_c\}$ an index referring to a specific *consensus cluster*), together with
- a *cluster correspondence array* α_{crk} (which shows how the individual clusters c from run r are recomposed from consensus clusters k)

such that the biclusters obtained in the different runs can be recovered from the following Positive Tensor Factorization:

$$E_{icr}^{(s)} \cdot E_{jcr}^{(d)} \approx \sum_{k=1}^{N_c} \alpha_{crk} e_{ik}^{(s)} e_{jk}^{(d)}. \quad (12)$$

[5] Note that although formally useful, the above single relation representation of a multirelational domain is highly impractical due to its size.

More formally, (12) is rewritten as a minimization problem for the following error function:

$$F\left(\alpha, \{e^{(n)}\}_n\right) = \frac{1}{2} \sum_{\substack{(s,d)\in\mathcal{R}\\ c,r,i,j}} \left(E^{(s)}_{i(cr)} E^{(d)}_{j(cr)} - \sum_{k=1}^{N_c} \alpha_{(cr)k} e^{(s)}_{ik} e^{(d)}_{jk} \right)^2. \tag{13}$$

Note that in (13) we have grouped the (cr) indices in α and E in order to deal with matrices rather than 3-dimensional arrays.

The objective function (13) above aims at minimizing the Euclidean distance between the bicluster c from run r (given by $\left(E^{(s)}_{i(cr)} E^{(d)}_{j(cr)}\right)_{ij}$) and the cluster reconstructed from the consensus biclusters $\left(e^{(s)}_{ik} e^{(d)}_{jk}\right)_{ij}$ by means of the cluster correspondence matrix $\alpha_{(cr)k}$.

To obtain a multiplicative update algorithm for minimizing (13), we proceed in a similar way as in the case of MNMF (2). Introducing the Lagrangean $\mathcal{L} = F - \sum_n \mu^{(n)} \circ e^{(n)} - \nu \circ \alpha$, we obtain Karush-Kuhn-Tucker conditions which combined with a splitting of the gradient of F into positive and negative parts $\left(\frac{\partial F}{\partial e^{(n)}}\right)_{\pm}$, $\left(\frac{\partial F}{\partial \alpha}\right)_{\pm}$ lead to the following multiplicative update rules for $e^{(n)}$ and α:

$$e^{(n)} \leftarrow e^{(n)} \circ \frac{\left(\frac{\partial F}{\partial e^{(n)}}\right)_-}{\left(\frac{\partial F}{\partial e^{(n)}}\right)_+} \tag{14}$$

$$\alpha \leftarrow \alpha \circ \frac{\left(\frac{\partial F}{\partial \alpha}\right)_-}{\left(\frac{\partial F}{\partial \alpha}\right)_+} \tag{15}$$

Computing the gradient of F leads to the following explicit form of the update rules:

$$e^{(n)} \leftarrow e^{(n)} \circ \frac{E^{(n)} \cdot \left[\alpha \circ \sum_{\substack{(d,n)\in\mathcal{R} \text{ or}\\ (n,d)\in\mathcal{R}}} E^{(d)T} \cdot e^{(d)} \right]}{e^{(n)} \cdot \left[(\alpha^T \cdot \alpha) \circ \sum_{\substack{(d,n)\in\mathcal{R} \text{ or}\\ (n,d)\in\mathcal{R}}} e^{(d)T} \cdot e^{(d)} \right]} \tag{16}$$

$$\alpha \leftarrow \alpha \circ \frac{\sum_{(s,d)\in\mathcal{R}} \left(E^{(s)T} \cdot e^{(s)}\right) \left(E^{(d)T} \cdot e^{(d)}\right)}{\alpha \cdot \sum_{(s,d)\in\mathcal{R}} \left(e^{(s)T} \cdot e^{(s)}\right) \left(e^{(d)T} \cdot e^{(d)}\right)}. \tag{17}$$

Note that in any run r we have:

$$\begin{aligned} R^{(sd)}_{ij} &\approx \sum_{c=1}^{N_c} E^{(s)}_{i(cr)} E^{(d)}_{j(cr)} \approx \sum_{c=1}^{N_c} \sum_{k=1}^{N_c} \alpha_{(cr)k} e^{(s)}_{ik} e^{(d)}_{jk} \\ &= \sum_{k=1}^{N_c} \left(\sum_{c=1}^{N_c} \alpha_{(cr)k} \right) e^{(s)}_{ik} e^{(d)}_{jk} \end{aligned}$$

Therefore, in order to interpret $e^{(s)}$ as a consensus of $E^{(s)}_r$ in the different runs, we need to have $\sum_{c=1}^{N_c} \alpha_{(cr)k} \approx 1$ for each run r. Thus, we impose a normalization of α of the form

$$\sum_{c,r} \alpha_{(cr)k} = N_r. \tag{18}$$

Summing up, our *consensus clustering algorithm* runs MNMF N_r times, randomly initializes $\{e^{(n)}\}_n$ and α, then iteratively applies the update rules (16,17) until convergence and subsequently normalizes α using (18). Finally, the consensus clusters $\{e^{(n)}\}_n$ are used as initialization for a final MNMF run.

Note that the consensus clusters need not necessarily be highly recurring clusters across the different runs. They could form a "base" set of clusters out of which all the clusters could be reconstructed by means of linear combinations. This allows learning of frequently occurring *subclusters*, thereby alleviating the need for very large numbers of runs.

4 Evaluation on synthetic datasets

We have evaluated our algorithm on synthetic datasets of the form

$$R^{(mn)} = E^{(m)} \cdot E^{(n)} + \epsilon^{(mn)}$$

with $\epsilon^{(mn)}$ a noise term. The consensus clustering algorithm robustly recovered the original clusters, performing slightly better than the base level clustering algorithm.

Although important for algorithm validation, tests on synthetic datasets are rarely indicative of the performance on real-life gene expression data, as most genomic subtypes of cancer are still incompletely known. We therefore concentrate in the following on the most detailed genomic datasets of leukemia and respectively normal hematopoiesis.

5 A joint genomic analysis of leukemia and normal hematopoiesis

Leukemia is one of the most heterogeneous diseases. Its highest-level classification includes acute lymphoblastic leukemia (ALL), chronic lymphocytic leukemia (CLL), acute myelogenous leukemia (AML), chronic myelogenous leukemia (CML), with a myriad of subtypes and other rarer types. Since its main cause consists of genomic defects in the hematopoietic stem or progenitor cells and since the hematopoietic system is in its turn extremely complex[6], it seems an extremely important task to investigate the similarities and differences between leukemia subtypes and normal hematopoietic cells (stem cells, progenitors and differentiated cells).

In the following we briefly present such an analysis of the largest available gene expression datasets of leukemia and respectively normal hematopoiesis.

The Microarray Innovations in Leukemia (MILE) study [2] has obtained gene expression profiles of 2096 leukemia patients (with 17 clinical subtypes of leukemia) and normal subjects (74 persons) using Affymetrix U133 Plus 2.0 microarrays.

On the other hand, the study of Novershtern et al. [13] has produced gene expression measurements of 38 distinct types of purified hematopoietic cells (211 samples in all) employing a slightly different microarray platform (Affymetrix U133A).

We have reprocessed the raw Affymetrix CEL files using RMA normalization and retained only the common probesets between the two profiling platforms (U133A probesets are almost completely included on the U133 Plus 2.0 platform). We further filtered the probesets (genes) retaining only those with a significant expression (mean of the log_2-values $> log_2(100)$ and standard deviation of log_2 values > 0.8). We thus ended up with 7417 probesets.

Besides the gene expression matrices of the leukemia (X_L) and respectively hematopoiesis dataset (X_H), we employed the given subtype information, Y_L for leukemia and Y_H for hematopoiesis.

[6] Its transcriptome is comparable in variability with the entire set of human cell types.

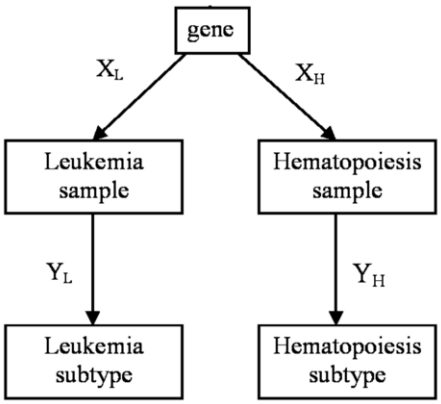

Figure 2. The multi-relational structure of the joint analysis of leukemia and normal hematopoiesis

We constructed a relational structure containing 5 entities and 4 relations as shown in Figure 2. We used relation weights to balance the Euclidean norms of the relations and subsequently reduced the weights of the subtype relations by $1/100$ to avoid any significant bias of the known subtype information on the inferred clusters. Note that such a very flexible form of *semi-supervised learning* can be easily adapted in our framework.

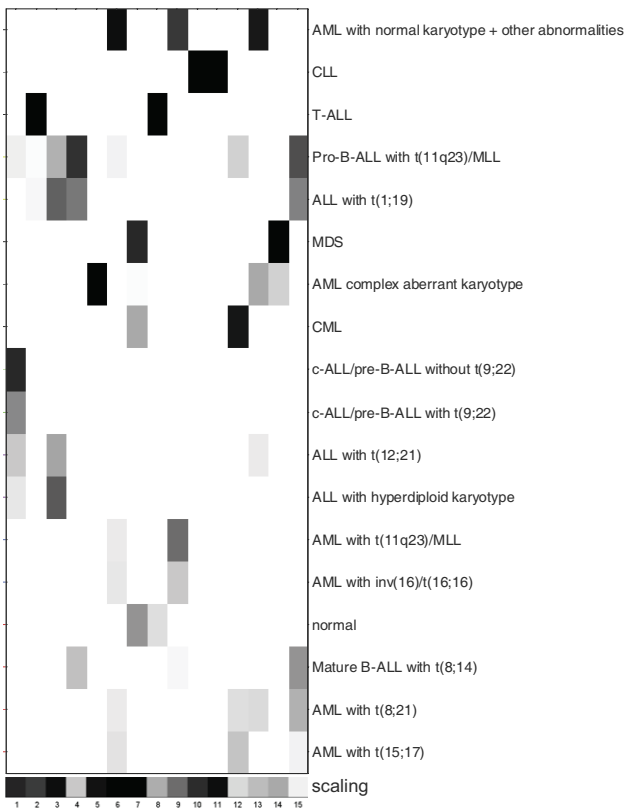

Figure 3. The entity matrix for the leukemia subtypes

Next we ran our multi-relational consensus clustering algorithm with $N_c = 15$ clusters and $N_r = 10$ runs. The number of clusters was chosen based on a series of runs of MNMF with progressively larger numbers of clusters, ranging from 2 to 50. To avoid overfitting, we performed a similar set of runs on the randomized entity matrices and compared the decrease of the error with N_c in the two cases. An N_c was chosen such that the error decrease on the real data was significantly larger than that on the randomized data [3].

We subsequently analyzed in more detail the clusters obtained. Note that the algorithm infers sample-specific *gene modules* (biclusters) rather than simple unidimensional sample clusters. Some modules may be involved both in disease and in normal cells, although certain modules are predominantly activated in leukemia while others – in normal hematopoietic cells. Figures 3 and 4 show the entity matrices for the leukemia and respectively hematopoiesis subtypes. (Rows correspond to subtypes, while columns correspond to clusters.[7]) These two matrices are especially informative since they establish a correspondence between the leukemia subtypes and the cell types of the normal hematopoietic system.

Remarkably, the algorithm has been able to link major leukemia types to their putative cells of origin in a completely unsupervised manner. For example, gene modules (clusters) 10 and 11 are mainly active in *chronic lymphocytic leukemia (CLL)* samples, but are also weakly activated, in the hematopoietic dataset, in *differentiated* (mature) B-cells.

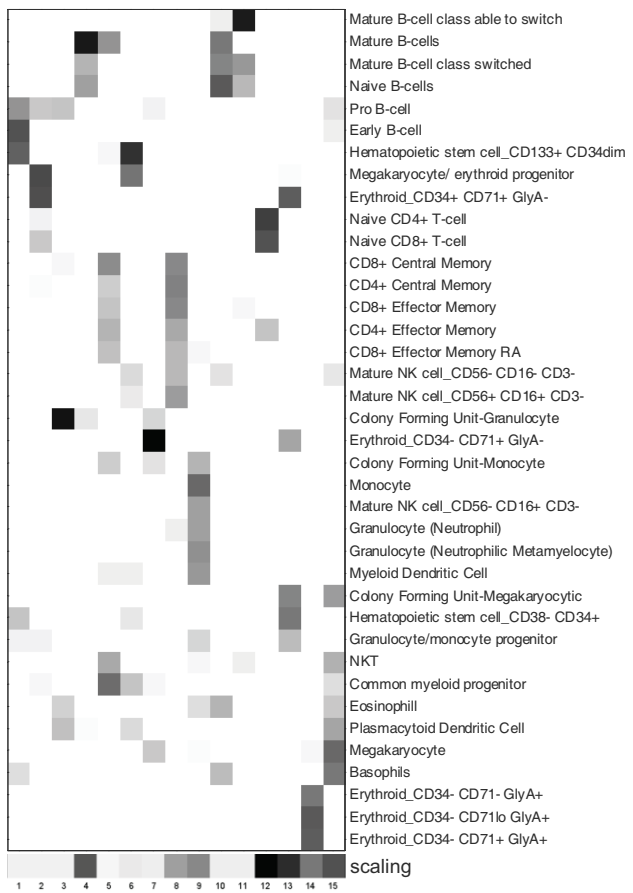

Figure 4. The entity matrix for the hematopoiesis subtypes

Conversely, module 4 is primarily activated in normal mature B-

[7] In the figures, the columns of the *subtype* clusters have been normalized to unit norm. Given that the *gene* clusters had also been normalized to unit norm, the corresponding scaling factors of the *sample* clusters (representing activation strengths) are shown in the last rows of the figures.

cells, but is also weakly involved in Pro-B ALL with t(11q23)/MLL and ALL with t(1;19).

On the other hand, cluster 1 predominantly involves *B precursor ALL* cases (c-ALL/pre-B-ALL, pro-B-ALL with t(11q23)/MLL, ALL with t(12;21) and ALL with hyperdiploid karyotype, but also affects *less differentiated B-cells* (such as early B-cells or pro B-cells), or even *hematopoietic stem cells* (either CD133+CD34dim or CD38-CD34+).

Gene module 2 covers the T-ALL cases, while its "normal" counterpart, module 8 is mainly active in normal T-cells and certain natural killer (NK) cells, with weaker activation in T-ALL.

Gene module 6 is dominant mainly in AML cases, but is also weakly active in hematopoietic stem cells (HSC CD133+CD34dim and CD38-CD34+), megakaryocyte/erythroid progeniors (MEP) and common myeloid progenitors (CMP). Its closest normal counterpart is gene module 13, which is primarily expressed in HSC, as well as in the least differentiated erythroid progenitors (CD34+CD71+GlyA- and CD34-CD71+GlyA-). It is remarkable that the highest level stem cell in the hematopoietic lineage (CD133+CD34dim) is primarily involved in *acute* leukemias (B precursor ALL in module 1 and respectively AML in module 6).

Gene module 14 covers normal differentiated erythroid cells and is only weakly active in myelodysplastic syndrome (MDS) cases.

Overall, it is impressive that the various leukemia subtypes have been matched, in an unsupervised manner, to the main hematopoietic cells affected by the disease.

A detailed "dissection" of each individual subtype and associated expression program is needed to understand them at a molecular level. An indication of the *unusual complexity of these expression programs* is given by the unusually large numbers of transcription factors involved. More precisely, using a relatively strict significance threshold for the normalized gene cluster matrix[8] $E^{(1)} > \frac{2}{\sqrt{N_1}}$, we obtain 273 transcription factors (TFs) significantly involved in the $N_c = 15$ clusters, many more than the TFs with a known role in leukemia or normal hematopoiesis. However, this is less surprising given the already known very large transcriptomic variability of the normal hematopoietic cell types [13].

As an example, the hematopoietic stem cell program for cluster 6 involves 35 transcription factors, among which SOX4, HOXA10, CEBPA, MYB, SATB1, CITED2, etc. A literature search has shown that many of these transcription factors have been previously linked to hematopoietic stem cells and/or leukemia. For instance, although the normal function of SOX4 in hematopoietic stem cells (HSCs) is not known, its over-expression in mouse HSCs has recently been shown to cause myeloid leukemia [14].

HOXA10 is a critical regulator of hematopoietic stem cells and erythroid/megakaryocyte development [11] (a fact consistent with its observed role in cluster 6, related to AML).

Also, CITED2 is known to be an essential regulator of adult hematopoietic stem cells [4].

The new perspective opened by our study is the large number of such transcription factors that probably control the various associated normal and leukemic gene expression programs (in a combinatorial manner). This insight should further help to develop personalized therapies, based on the specific genomic changes encountered in each patient.

[8] We have normalized the columns of the gene cluster matrix to unit Euclidean norm.

6 Conclusions

A comprehensive discussion (or even enumeration) of all approaches to multi-relational learning is impossible due to space limitations. Focusing on numerical data-oriented approaches only, the frameworks closest to our approach are Collective Matrix Factorization (CMF) [15], Multi Relational Matrix Factorization (MRMF) [10] and NMRF [5]. None of these approaches are able to deal with clustering instability, which as mentioned previously is one of the main problems facing multi-relational discovery systems.

Moreover, a simple data-oriented approach like the one presented in this paper avoids the combinatorics that tends to plague logical multirelational discovery systems (e.g. Inductive Logic Programming).

The results of the genomics application are also encouraging.

ACKNOWLEDGEMENTS

This research was partially supported by the project PN-II-ID-PCE-2011-3-0198. I am grateful to Andrei Halanay, Daniel Coriu and Jardan Dumitru for discussions.

REFERENCES

[1] J.P. Brunet, et al. Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. 101: 4164-4169, 2004.
[2] Haferlach T, et al. Global approach to the diagnosis of leukemia using gene expression profiling. Blood. 2005 Aug 15;106(4):1189-98.
[3] P. M. Kim and B. Tidor. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res*, **13**(7), 1706-18, 2003.
[4] Kranc KR, et al. Cited2 is an essential regulator of adult hematopoietic stem cells. Cell Stem Cell. 2009 Dec 4;5(6):659-65.
[5] L. Badea. Multi-relational factorizations for cancer subclassification, Proc. ICACTE-2010, V1-248-252, 2010.
[6] L. Badea. Clustering and Metaclustering with Nonnegative Matrix Decompositions. Proc. ECML-2005:10-22, 2005.
[7] Lee CH, et al. GSVD Comparison of Patient-Matched Normal and Tumor aCGH Profiles Reveals Global Copy-Number Alterations Predicting Glioblastoma Multiforme Survival. PLoS One. 2012;7(1):e30098.
[8] Lee DD and Seung HS. Algorithms for non-negative matrix factorization. in *NIPS*, pp. 556–562, 2000.
[9] S. Lin and H. Chalupsky. Issues of Verification for Unsupervised Discovery Systems, Proc. KDD04 Workshop Link Analysis and Group Detection, 2004.
[10] Lippert C, et al. Relation-Prediction in Multi-Relational Domains using Matrix-Factorization. in *NIPS 2008 Workshop: Structured Input-Structured Output*, 2008.
[11] Magnusson M, et al. HOXA10 is a critical regulator for hematopoietic stem cells and erythroid/megakaryocyte development. Blood. 2007 May 1;109(9):3687-96.
[12] S. Monti, et al. Consensus Clustering: A Resamlping Based Method for Class Discovery and Visualization of Gene Expression Microarray Data, Journal of Machine Learning, 52(1-2), 2003.
[13] Novershtern N, et al. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell. 2011 Jan 21;144(2):296-309.
[14] Richter K, et al. Global gene expression analyses of hematopoietic stem cell-like cell lines with inducible Lhx2 expression. BMC Genomics. 2006 Apr 6;7:75.
[15] Singh AP, Gordon GJ. Relational learning via collective matrix factorization. in *Proc. KDD '08*, pp. 650–658, ACM, 2008.
[16] L. Getoor, B. Taskar (eds.) Introduction to Statistical Relational Learning. *MIT Press*, 2007.
[17] Welling M., Weber M. Positive tensor factorization. Pattern Recognition Letters 22(12): 1255-1261 (2001).