Unsupervised Multiple-Instance Learning for Functional Profiling of Genomic Data

Corneliu Henegar[1], Karine Clément[1,2,3], and Jean-Daniel Zucker[1,4]

[1] INSERM, UMR U-755 Nutriomique, Hôtel-Dieu, Paris, France
corneliu@henegar.info
[2] Université Paris VI, Faculté de Médecine Les Cordeliers, Paris, France
[3] AP-HP, Pitié-Salpêtrière, Service de Nutrition, Paris, France
[4] LIM&BIO EA3969, Université Paris Nord, Bobigny, France

Abstract. Multiple-instance learning (MIL) is a popular concept among the AI community to support supervised learning applications in situations where only incomplete knowledge is available. We propose an original reformulation of the MIL concept for the unsupervised context (UMIL), which can serve as a broader framework for clustering data objects adequately described by the multiple-instance representation. Three algorithmic solutions are suggested by derivation from available conventional methods: agglomerative or partition clustering and MIL's citation-kNN approach. Based on standard clustering quality measures, we evaluated these algorithms within a bioinformatic framework to perform a functional profiling of two genomic data sets, after relating expression data to biological annotations into an UMIL representation. Our analysis spotlighted meaningful interaction patterns relating biological processes and regulatory pathways into coherent functional modules, uncovering profound features of the biological model. These results indicate UMIL's usefulness in exploring hidden behavioral patterns from complex data.

1 Introduction

The conceptual frame of the multiple-instance learning (MIL) was proposed in 1997 by Dietterich [1], together with a first meaningful application to drug activity prediction. Since then, an important amount of research has dealt with the development of specific learning algorithms, adapted to MIL's particular context, and to comparative performance assessment in relation with different types of applications, as well as with various other conventional supervised learning approaches [2,3,4,5,6,7,8,9,10]. As a result, MIL's applicability has been tested in numerous domains, ranging from content-based image retrieval and classification [11], text categorization [6] and web mining [12], to protein sequence analysis, robot vision and stock market prediction [13,14]. Conventional MIL is a variation on supervised learning, fitting those situations in which the knowledge about the labels of training examples is incomplete. Under such circumstances MIL allows for modeling weaker assumptions about the labeling information by assigning labels to sets of instances (bags), instead of assigning them to each individual instance. Bags labels can be positive or negative in the Boolean case,

J. Fürnkranz, T. Scheffer, and M. Spiliopoulou (Eds.): ECML 2006, LNAI 4212, pp. 186–197, 2006.

or have a continuous real value in the real data MIL [15]. A bag is labeled as positive if *at least one* of its instances is positive (linearity constraint), and negative if *all* of its instances are negative. In generalized MIL, a variant of the conventional model, bags labels are determined by a non-disjunctive function over their instances, thus eliminating the linearity constraint in order to reduce noise level [9].

In this paper we propose an abstract reformulation of the conventional MIL paradigm, which preserves the general multiple-instance representation, while further weakening the supervised learning constraints into a fully unsupervised multiple-instance learning (UMIL) framework. The main motivation behind this reformulation resides in the usefulness of the multiple-instance schema, which allows to describe some difficult unsupervised learning problems through simple and yet robust representations. Such representations can provide a basis for solving intricate clustering problems, aiming at discovering hidden behavioral patterns from complex data objects described by multiple types of attributes (e.g. numerical, symbolic, etc.). Among other possible examples, such complex objects are found in genomic data sets in which RNA transcripts are sharing numerous descriptive features in relation to their various biological roles. Therefore, we relied on the functional genomics framework to illustrate the UMIL concept by relating RNA expression data to functional annotations to build multiple-instance representations. These representations were further used to perform a functional analysis of two genomic data sets, aiming at identifying context related biological interaction patterns involving cellular processes and regulatory pathways. Section two outlines the main characteristics of the UMIL paradigm. The third section suggests three algorithmic solutions, derived from existent unsupervised learning or conventional MIL approaches, adapted to the UMIL context. The fourth section details the experimental framework and results. Finally we indicate some potential directions for future work.

2 The Unsupervised Multiple-Instance Model

2.1 UMIL Definition

In order to allow for a maximum flexibility in building multiple-instance representations, we imagined the UMIL paradigm as an abstract generalization of the conventional multiple-instance schema. Let us consider a data set $\mathcal{D}$ composed of n objects $o_j \in \mathcal{D}$, sharing similar data structures, each of them being characterized by an ensemble of feature values $o_j = \{f_1 = v_{1j}, f_2 = v_{2j}, ..., f_i = v_{ij}, ...\}$, be it numerical, Boolean or set-valued attributes. Among the ensemble $\mathcal{F}$ of all features describing objects $o_j \in \mathcal{D}$, let $f_i \in \mathcal{F}$ be a *feature* whose domain contains m distinct values, $f_i = \{v_1, v_2, ..., v_k, ..., v_m\}$, each object $o_j \in \mathcal{D}$ being characterized by one or more values of f_i. Based on the feature f_i we derive the ensemble $\mathcal{B}$ of bags $b_k \in \mathcal{B}$, $(k \leq m)$, defining an UMIL model, where each bag b_k corresponds to the ensemble of objects $o_j \in \mathcal{D}$ sharing (at least) one common feature value $f_i = v_k$, which defines the bag b_k. As each of the objects $o_j \in \mathcal{D}$ can be characterized by one or more values of $f_i \in \mathcal{F}$, it follows that UMIL bags are

non disjoint (e.g. overlapping) sub-ensembles of $\mathcal{D}$, their distinctiveness being guaranteed by the common feature value $f_i = v_k$ of their instances. We propose that this multiple-instance abstraction may constitute a relevant framework for exploring complex relationships between multiple-instance objects in an unsupervised learning context. Under these circumstances, the UMIL problem can be stated formally as to find an optimum partition of $\mathcal{B}$ into $l < m$ disjoint classes of interrelated bags $C_1 \cup C_2 \cup ... \cup C_l$.

2.2 Multiple-Instance Representations of Genomic Data

In genomic data sets RNA transcripts are represented through complex data structures, which are regrouping heterogeneous information related to expression measurements (real value data), molecular structure, functional roles, regulatory mechanisms, etc. Biological roles of RNA transcripts are formally represented through functional annotations established in relation with available biological evidence. These representations are built through an annotation process which relates RNA transcripts to a taxonomic hierarchy of functional categories (set-valued attributes), allowing to represent biological knowledge about transcripts roles with various degrees of precision. In the most general case, the relations among transcripts and functional categories are of the many-to-many type, in which a transcript may be related to one or more biological processes, each of these processes involving one or more transcripts. Considered as a major challenge, the functional analysis, which aims at translating RNA expression data into relevant biological mechanisms, is an indispensable step for the comprehension of the underlying biological phenomena defining an experimental model. Besides assessing the individual dynamics of various biological processes, based upon the expression patterns of the transcripts known to be involved in those processes, the functional profiling aims also at characterizing intricate biological interactions involving cellular processes and regulatory pathways. These considerations suggest the relevance of the UMIL paradigm as a formal framework for assessing interactions between functional categories, represented as multiple-instance objects (e.g. *bags*) which regroup annotated transcripts (e.g. *instances*).

2.3 Similarity and Relationship Measures for UMIL Objects

As a consequence of definition (2.1) two types of measures seem relevant for comparing objects belonging to an UMIL representation. The first one will evaluate the similarity between individual instances, thus conditioning the second one which will assess the relationship between bags. In our context we selected the pairwise mutual information (MI) as the similarity metric for transcripts expression, based on its ability to recognize as proximal positively, negatively and nonlinearly correlated transcript profiles [16, 17]. MI computation is based on the notion of entropy of a random variable suggested by Shannon's theory of information. Thus for a discrete random variable X, whose probability

distribution is $P(X = x_i)$, $i = 1, ..., N_x$, where N_x is the number of possible values of X, the entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{N_x} P(X = x_i) \log_2 P(X = x_i) \ . \tag{1}$$

For the case of continuous random variables (e.g. expression profiles) a preliminary discretization, through a histogram technique, is necessary in order to compute their probability distribution. Based on (1) the pairwise mutual information of two random variables X, Y is defined as:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

where $H(X,Y)$ is their joint entropy. The normalized $\overline{MI}(X,Y)$ is a relative measure [17] which reduces the influence of the magnitudes of individual entropies:

$$\overline{MI}(X,Y) = \frac{MI(X,Y)}{\max\{H(X),\, H(Y)\}} \ . \tag{3}$$

From (3) it follows that $0 \leq \overline{MI}(X,Y) \leq 2$. Moreover, it is possible to estimate a threshold of significance T_{MI} for the pairwise mutual information through iterative random permutations over the matrix of expression measurements [16]. Given two possibly overlapping bags A and B, the strength of their relationship can be quantified separately, from each bag's perspective, through a non-disjunctive function over all instances belonging to that bag for which there is at least one similar (or identical) instance in the other bag, and vice versa. Let n_{ab} be the sub-ensemble of instances $a_i \in A$ for which there is at least one instance $b_j \in B$ satisfying the similarity constraint T_{MI}:

$$n_{ab} = \{a_i \in A \mid \exists\, b_j \in B,\, \overline{MI}(a_i, b_j) \geq T_{\mathrm{MI}}\} \ . \tag{4}$$

Consider $\bar{n}_{ab}$ the cardinality of n_{ab} and $\bar{n}_{ba}$ its equivalent for bag B. From (4) it follows that in the most general case $\bar{n}_{ab} \neq \bar{n}_{ba}$. Under these circumstances, the ratio $S_{A\to B} = \frac{\bar{n}_{ab}}{\bar{n}_A}$, where $\bar{n}_A$ is the cardinality of bag A, can be considered as an *asymmetrical measure* of the relationship between the two bags from bag A perspective, satisfying $0 \leq S_{A\to B} \leq 1$. In order to give a better account of the qualitative value of instances similarity we can further refine $S_{A\to B}$ by weighting it with the average of the maximal similarities of individual instances $a_i \in A$ satisfying (4) in relation to $b_j \in B$ and define an *asymmetrical measure* of the relationship of A with B as:

$$D_{A\to B} = 1 - S_{A\to B} \left[\frac{1}{2\bar{n}_{ab}} \sum_{i=1}^{\bar{n}_{ab}} \max_{j=1}^{\bar{n}_B} \overline{MI}(a_i, b_j) \right] \tag{5}$$

From (5) it follows that $0 \leq D_{A\to B} \leq 1$ and also that $D_{A\to B} \neq D_{B\to A}$ in the most general case. Based on (5) a *symmetrical measure* of the relationship between two bags A and B can be defined as:

$$D_{AB} = \frac{1}{2} \left(D_{A\to B} + D_{B\to A} \right) \ . \tag{6}$$

3 Algorithmic Solutions

Two directions were explored in search for algorithmic solutions adapted to the UMIL context. The first one was to examine possible adaptations of existing unsupervised learning approaches. The second was to consider adaptations of supervised MIL approaches to the unsupervised context. Our analysis shows that some of the difficulties which need to be addressed are different in each of these two cases, while others are common.

3.1 Unsupervised Clustering Approaches for the UMIL Context

The proposed definition (2.1) of the UMIL paradigm suggests the idea of adapting conventional unsupervised clustering approaches for the UMIL context. For instance, one simple solution could be to initiate a conventional hierarchical agglomerative clustering algorithm with the partition of the instances in their corresponding bags (considered as "clusters" of instances). In these circumstances, the hierarchical clustering algorithm could presumably be used to identify classes of related bags by relying only on the similarity of their instances. However, some of the characteristics of the UMIL representation, like the possible overlapping between bags in the most general case, cannot be handled correctly by a conventional unsupervised clustering approach. A possible solution to this obstacle could be to reduce the multiple instance model to a simple instance one, by relying on the symmetrical measure of the relationship between bags (6) defined previously. This reductive approach allowed us to test two conventional unsupervised clustering techniques for the UMIL context: an hierarchical agglomerative algorithm [18] and a k-means partitioning algorithm [19], each of them combined with a standard quality measure for cluster partitions which allows to identify an optimal partition of bags into classes. The prediction of the correct number of clusters is a fundamental question in unsupervised classification problems [20]. Although there is no best approach to fit all situations, the computation of the Silhouette index [19] was shown to be a simple and yet robust strategy for the prediction of optimal clustering partitions from transcript expression data [21].

3.2 A Citation Approach for the UMIL Context

A conventional MIL solution that may be easily adapted for the unsupervised context is that proposed originally by Wang and Zucker [3], which combines k-nearest neighbor (kNN) lazy learning with the citation concept (citation-kNN) inspired from library and information science. In our context the concept of bibliographic citations is suggested by the asymmetrical aspect of the relationship between bags (5). This results in the fact that two bags can "refer" to each other with a different degree of confidence strength. Based on this observation we imagined an *unsupervised citation-kNN* (UC-kNN) solution whose main steps are illustrated by Algorithm 1. Let m be the number of individual bags $b_i \in \mathcal{B}$ contained in the UMIL representation $\mathcal{B}$. Considering (5) as the measure of relationship between bags, a bag $b_j \in \mathcal{B}$ can be presumed to be a good "reference"

for another bag $b_i \in \mathcal{B}\backslash b_j$ if bag b_j is ranked among the $k < m$ most closely related bags to bag b_i (considered therefore as its k nearest neighbors or kNN).

Algorithm 1. A sketch of the UC-kNN algorithm

Input: an UMIL representation $\mathcal{B} = \{b_1, ..., b_m\}$, containing m bags with their instances, and the similarity matrix for instances computed with (3)
Output: the optimal partition of the bags
Compute bags relationship matrix with (5)
For each k, $1 \leq k \leq m-1$ (e.g. the number of nearest neighbors) do:

Compute a ranked vector $\mathcal{R}$ of the bags reference scores $R_b = \sum_i rank(b, b_i)$, for each $b \in \mathcal{B}$, in relation to the rest of the bags $b_i \in \mathcal{B}\backslash b$ which satisfy $rank(b, b_i) \leq k$
For each p, $2 \leq p < m$, select the first p bags from $\mathcal{R}$ as cluster seeds, then do:
For each $m - p$ bags b_i, distinct from the p selected cluster seeds, do:
Find the k best references b_j for b_i then compute for each of the p cluster seeds s the value $V_{sb_i} = rank(s, b_i) + \frac{1}{k} \sum_{j=1}^{k} rank(s, b_j)$ and cluster b_i to the closest seed
Compute the Silhouette index for the resulting partition of bags and store results

Select the optimal partition of bags, among those computed for each possible combination of the values of k and p, which maximizes the Silhouette index

On this base a reference score R_b can be computed for each value of $k < m$ and for each bag $b \in \mathcal{B}$, in relation to the rest of bags $b_i \in \mathcal{B}\backslash b$, as the sum of b's ranking positions for all the situations where $rank(b, b_i) \leq k$ (see Algorithm 1). This suggests that, for a given value of k, it is possible to initiate an agglomerative clustering procedure by considering as seeds of the future classes (or clusters) the first p bags, $2 \leq p < m$, having the best reference scores (e.g. the most "cited" ones). Under these circumstances, a kNN clustering approach can group each of the rest of the bags to their most closest seed, by relying not only on the individual similarity between the bags and the seeds, but by considering also the similarity of their k nearest neighbors to these seeds, integrated into a *weighted voting procedure*. This is to say that for each bag b_i, distinct from the considered p seeds, we search the closest seed s minimizing the value of:

$$V_{sb_i} = rank(s, b_i) + \frac{1}{k} \sum_{j=1}^{k} rank(s, b_j) \tag{7}$$

where b_j, $1 \leq j \leq k$, belongs to the k nearest neighbors of bag b_i. Thus, for each couple of values (k, p), with $k, p < m$, the UC-kNN approach will build a partition $P_{(k,p)} = \{C_1 \cup ... \cup C_p\}$ of the ensemble of bags $\mathcal{B}$ into p distinct classes. As for the adaptation of the conventional unsupervised clustering approaches, an optimal partition of bags can be selected from the ensemble of computed

partitions by using a standard quality evaluation measure. For coherence and simplicity reasons we combined UC-kNN with the Silhouette technique [19].

4 Experimental Frame

The experimental context, which served to build multiple-instance representations and to test UMIL algorithmic solutions, belongs to functional genomics.

4.1 Adipose Tissue Data Sets

The potential benefit of the UMIL concept for the genomic functional analysis was assessed on two interrelated RNA expression measurements data sets. Both of them resulted from pangenomic cDNA microarray expression profiling of white adipose tissue in morbidly obese human subjects, and were extensively described in [22]. The first data set resulted from differential expression profiling of the two cellular fractions of human white adipose tissue: mature adipocytes and stroma-vascular fraction cells (SVF). The second one resulted from microarray expression profiling of whole white adipose tissue in morbidly obese human subjects, before/after undergoing a form of bariatric surgery. These two data sets were combined in order to constitute a coherent experimental model, designed to characterize the functional profiles of each of the two cellular fractions of the adipose tissue in obese human subjects, as well as their evolution after a significant weight loss induced by bariatric surgery.

4.2 Experimental Setup

The three proposed algorithmic solutions were implemented in the R environment for statistical computation (available at `http://www.r-project.org/`). As originally indicated [22], transcripts with significant expression changes were identified by using the significance analysis of microarrays (SAM) procedure (available at `http://www-stat.stanford.edu/tibs/SAM/`). Significant differential expression was assessed by imposing a 5% false discovery rate (FDR) threshold in the SAM selection procedure. Automated functional annotation of the differentially expressed transcripts, identified in the two data sets, relied on Gene Ontology Consortium (GO [available at `http://www.geneontology.org`]) and Kyoto Encyclopedia of Genes and Genomes (KEGG [available at `http://www.genome.ad.jp/kegg/`]) annotations. EntrezGene numbers (available at `http://www.ncbi.nlm.nih.gov/entrez`) were used as a standard transcript accession system to ensure a correct over-representation analysis, as they allow to map transcript identifiers to GO or KEGG categories in an unequivocal way. In order to minimize the false over-representation resulting from redundant annotation, the automated GO annotation procedure was restricted to directly annotated transcripts by each GO category. As originally indicated [22], the significance of the over-representation of each GO and KEGG category was assessed by using a Fisher's exact test. Afterwards, significantly over-represented GO

and KEGG categories were related to their annotated transcripts into an UMIL model, in which each category (GO or KEGG) was considered as a bag of individual instances represented by its annotated transcripts. A threshold T_{MI} for the normalized pairwise mutual information of transcripts expression was computed previously to applying unsupervised agglomerative or partitioning clustering and UC-kNN algorithms to the UMIL representation of genomic data. As previously suggested [16], T_{MI} estimation was based on the average $\overline{MI}$ distribution computed from 30 randomly permuted repetitions of RNA expression measurements. The significance threshold for the pairwise mutual information among transcripts was chosen to be $T_{\mathrm{MI}} = mean(\overline{MI}) + 2SD(\overline{MI})$, where $mean(\overline{MI})$ is the average of $\overline{MI}$ and $SD(\overline{MI})$ the standard deviation of the mean.

Table 1. Characteristics of the optimal partitions obtained by applying the agglomerative hierarchical clustering (HC), k-means partition clustering (K-means) and the unsupervised citation kNN (UC-kNN) algorithms to the two adipose tissue data sets

HC	Min	Max	Average
Clusters number	2	29	6.81 ± 6.64
Clusters length	1	35	4.18 ± 7.05
Clusters Silhouette	0	0.83	0.14 ± 0.17
Partitions Silhouette	0.05	0.52	0.14 ± 0.11
K-means	**Min**	**Max**	**Average**
Clusters number	2	53	16.31 ± 13.23
Clusters length	1	12	1.75 ± 1.80
Clusters Silhouette	0	1	0.06 ± 0.16
Partitions Silhouette	0.05	0.20	0.11 ± 0.04
UC-kNN	**Min**	**Max**	**Average**
Clusters number	2	6	3.44 ± 1.21
Clusters length	1	64	8.29 ± 13.1
Clusters Silhouette	0	1	0.35 ± 0.34
Partitions Silhouette	0.04	0.68	0.37 ± 0.15

4.3 Results

A few characteristics of the results produced by the three algorithmic approaches are summarized in Table 1. As it can be seen the Silhouette indexes of the partitions produced by the UC-kNN approach are much higher than those resulting from the two adaptations of unsupervised clustering approaches. Moreover, unsupervised clustering partitions were on average more sparse than those produced by the UC-kNN solution. For all these reasons, and also because of space restrictions, only a fraction of the UC-kNN clustering results are detailed hereafter and discussed in terms of biological relevance.

Table 2. Main UC-kNN clusters of KEGG categories specifically expressed in each of the two adipose tissue fractions: adipocytes and stroma-vascular fraction (SVF)

KEGG Category	Nb. Transcr.*	P-value**
Cluster 1 - Adipocytes	109	$2.84\ 10^{-12}$
Tryptophan metabolism	26	$9.58\ 10^{-3}$
Fatty acid metabolism	23	$1.35\ 10^{-5}$
Pyruvate metabolism	22	$2.05\ 10^{-6}$
Valine, leucine & isoleucine degrad.	22	$1.57\ 10^{-4}$
Basal transcription factors	10	$4.87\ 10^{-2}$
Other metabolic processes (9 terms)	64	—
Cluster 1 - SVF	186	$3.93\ 10^{-22}$
Cytokine-cytokine recept. interact.	65	$5.61\ 10^{-8}$
Hematopoietic cell lineage	37	$5.10\ 10^{-9}$
Ribosome	33	$2.93\ 10^{-9}$
Natural killer cell med. cytotox.	32	$5.15\ 10^{-4}$
Complement & coagulation cascades	23	$7.47\ 10^{-4}$
TGF-beta signaling pathway	22	$2.65\ 10^{-2}$

* number of annotated transcripts
** transcript enrichment p-value computed with Fisher's exact test

Table 2 shows one cluster (from a total of 4, with individual Silhouettes of 0.50 and 0.48 respectively, and a partition Silhouette of 0.31) grouping KEGG categories annotating adipocytes transcripts, and one cluster (from a total of 3, with an individual Silhouette of 0.33, and a partition Silhouette of 0.31) characterizing the stroma-vascular fraction (SVF) transcripts. Cluster 1 - Adipocytes (Table 2) is grouping 13 metabolic processes known to be highly interrelated and specific of mature adipocytes. It thus depicts the functional profile of mature adipocytes involving various metabolic processes (energetic, lipidic or protidic) [22]. An interesting aspect is that these metabolic processes were grouped together with a set of 10 transcription factors, which suggests a specific regulating role over these processes. Indeed, at least four of them (TAF6, TAF7, TAF10 and TAF12) are known to be pro-adipogenic factors, enhancing the action of C/EBPα and TBP/TFIIB which are key regulators of the adipogenesis [23, 24]. Cluster 1 - SVF (Table 2) illustrates the preponderant role of the SVF in the pathogenesis of local and systemic inflammatory processes accompanying the inflation of the adipose tissue in humans. The presence of the TGF-beta signaling pathway in this cluster has strong biological significance, since TGF-beta is known to stimulate the proliferation of pre-adipocytes while inhibiting adipogenesis [23]. These findings may corroborate with available evidence, indicating the conversion of pre-adipocytes into macrophages under particular circumstances [25], thus supporting the paradigm of a major role of local adipose tissue macrophages in the pathogenesis of inflammatory processes characterizing human obesity [22]. For

Table 3. Main UC-kNN cluster of Gene Ontology (Biological Process) categories significantly down-regulated in human adipose tissue after bariatric surgery.

Gene Ontology Category	Nb. Transcr.*	P-value**
Cluster 1	86	$2.18\ 10^{-3}$
Apoptosis	61	$3.14\ 10^{-2}$
Anti-apoptosis	25	$8.31\ 10^{-3}$
Acute phase response	8	$3.18\ 10^{-2}$
Induction of apoptosis / intracel. sign.	5	$2.03\ 10^{-2}$

* number of annotated transcripts
** transcript enrichment p-value computed with Fisher's exact test

all these reasons the two analyzed clusters can be considered as a convincing illustration of the complex dynamics of the adipogenic regulatory mechanisms, in which pro-adipogenic factors act concomitantly with anti-adipogenic ones, thus resulting into an ever changing network of complex interactions [23].

Table 3 present one Gene Ontology Biological Process cluster (from a total of 4, with an individual Silhouette of 0.36, and a partition Silhouette of 0.40), characterizing adipose tissue transcripts down-regulated after bariatric surgery. This cluster indicate a coherent deflation of inflammatory phenomena accompanying weight loss. Indeed, the reduction in local synthesis of the acute phase response molecules, together with a consecutive reduction of apoptotic processes corroborate with previously reported results [22].

4.4 Discussion

Except for some particular situations in which supplementary knowledge is available, the validation of the unsupervised clustering results remains a difficult issue. In spite of their relative value, cluster quality measures were shown to be useful indicators of the relevance of transcript data partitions [21]. In our experimental context, the UC-kNN solution yielded much higher Silhouette indexes than the hierarchical clustering approach. These findings seem coherent with previous observations suggesting a good adequacy of the local approaches for the multiple-instance context [3]. Subsequently, the results of the functional profiling of the adipose tissue expression data were discussed in terms of biological significance, in accord with available biological knowledge. Our assessment pointed out the biological relevance of the UMIL functional analysis which spotlighted significant biological regulatory mechanisms, thus illustrating the underlying modular structure of the transcriptional regulatory networks.

5 Conclusion and Future Work

This paper proposes a new framework for the unsupervised clustering of complex data objects adequately described by an abstract multiple-instance representation.

Three algorithmic solutions, adapted to the new framework, are suggested. The application of the UMIL concept to the functional analysis of genomic data illustrates its usefulness in exploring hidden behavioral patterns from complex data. The UMIL model shares common features with other unsupervised learning models. Among them, the concept of a variable size transaction, used in market basket data analysis, may be the closest one from that of an UMIL bag. Defined as a finite set of items from a common item universe, the transaction concept can be considered as a particularization of the bag concept for the case in which instances are all categorical data structures. Therefore investigating the possibility of adapting existent categorical data algorithms to the UMIL context might prove interesting, as this could result in useful solutions for sparse and high dimensional data, known to be less adapted to local approaches. Another research direction will be to examine the possibility of a Bayesian solution for the UMIL frame. Besides this, other potential applications of the UMIL framework need to be considered, especially in those domains in which conventional multiple-instance framework proved useful. One such domain could be the content-based image retrieval and classification problem. An obvious advantage of considering this problem, besides the evident interest of this application, lies in a presumably simpler and more objective assessment of clustering results.

Acknowledgments

This work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM) and the Assistance Publique - Hôpitaux de Paris.

References

1. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artif. Intell. **89**(1-2) (1997) 31–71
2. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: NIPS. (1997)
3. Wang, J., Zucker, J.D.: Solving the multiple-instance problem: A lazy learning approach. In: ICML. (2000) 1119–1126
4. Chevaleyre, Y., Zucker, J.D.: Solving multiple-instance and multiple-part learning problems with decision trees and rule sets. application to the mutagenesis problem. In: Canadian Conference on AI. (2001) 204–214
5. Zhang, Q., Goldman, S.A.: Em-dd: An improved multiple-instance learning technique. In: NIPS. (2001) 1073–1080
6. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS. (2002) 561–568
7. Goldman, S.A., Scott, S.D.: Multiple-instance learning of real-valued geometric patterns. Ann. Math. Artif. Intell. **39**(3) (2003) 259–290
8. Tao, Q., Scott, S., Vinodchandran, N.V., Osugi, T.T.: SVM-based generalized multiple-instance learning via approximate box counting. In: ICML. (2004)
9. Tao, Q., Scott, S.D.: A faster algorithm for generalized multiple-instance learning. In: FLAIRS Conference. (2004)

10. Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: ICML 2005 Conference. (2005)
11. Zhang, Q., Goldman, S.A., Yu, W., Fritts, J.: Content-based image retrieval using multiple-instance learning. In: ICML. (2002) 682–689
12. Zhou, Z.H., Jiang, K., Li, M.: Multi-instance learning based web mining. Appl. Intell **22**(2) (2005) 135–147
13. Brown, J., Zhang, J., Scott, S.: On generalized multiple-instance learning. Technical report, University of Nebraska (2003)
14. Yang, J.: Review of multi-instance learning and its applications. Technical report, School of Computer Science Carnegie Mellon University (2005)
15. Dooly, D.R., Zhang, Q., Goldman, S.A., Amar, R.A.: Multiple-instance learning of real-valued data. Journal of Machine Learning Research **3** (2002) 651–678
16. Butte, A., Kohane, I.: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pac Symp Biocomput (2000) 418–29
17. Zhou, X., Wang, X., Dougherty, E., Russ, D., Suh, E.: Gene clustering based on clusterwide mutual information. J Comput Biol **11**(1) (2004) 147–61
18. Murtagh, F.: Multidimensional clustering algorithms. In Physica-Verlag, V., ed.: COMPSTAT Lectures 4. (1985)
19. Kaufman, L., Rousseuw, P.: Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, Inc. (1990)
20. Berkhin, P.: Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA (2002)
21. Azuaje, F., Bolshakova, N.: Cluster validation techniques for genome expression data. Signal Processing **83**(4) (2003) 825–833
22. Cancello, R., Henegar, C., Viguerie, N., Taleb, S., Poitou, C., Rouault, C., Coupaye, M., Pelloux, V., Hugol, D., Bouillot, J., Bouloumie, A., Barbatelli, G., Cinti, S., Svensson, P., Barsh, G., Zucker, J., Basdevant, A., Langin, D., Clement, K.: Reduction of macrophage infiltration and chemoattractant gene expression changes in white adipose tissue of morbidly obese subjects after surgery-induced weight loss. Diabetes **54**(8) (2005) 2277–86
23. Feve, B.: Adipogenesis: cellular and molecular aspects. Best Pract Res Clin Endocrinol Metab **19**(4) (2005) 483–99
24. Pedersen, T., Kowenz-Leutz, E., Leutz, A., Nerlov, C.: Cooperation between C/EBPalpha TBP/TFIIB and SWI/SNF recruiting domains is required for adipocyte differentiation. Genes Dev **15**(23) (2001) 3208–16
25. Charriere, G., Cousin, B., Arnaud, E., Andre, M., Bacou, F., Penicaud, L., Casteilla, L.: Preadipocyte conversion to macrophage. Evidence of plasticity. J Biol Chem **278**(11) (2003) 9850–5