Frequent Weighted Itemset Mining from Gene Expression Data

Elena Baralis, *Member, IEEE,* Luca Cagliero, Tania Cerquitelli,
Silvia Chiusano, *Member, IEEE,* and Paolo Garza

***Abstract*— Gene Expression Datasets (GEDs) usually consist of the expression values of thousands of genes within hundreds of samples. Frequent itemset and association rule mining algorithms have been applied to discover significant co-expressions among multiple genes from GEDs. To perform these data analyses, gene expression values are commonly discretized into a predefined number of bins. Such an expert-driven and not trivial preprocessing step could bias the quality of the mining result. This paper presents a novel approach to discovering gene correlations from GEDs which does not require data discretization. By representing per-sample gene expression values as item weights, frequent weighted itemsets can be extracted. The discovery of weighted itemsets instead of traditional (not weighted) ones prevents experts from discretizing GEDs before analyzing them and thus improves the effectiveness of the knowledge discovery process. Experiments performed on real GEDs demonstrate the effectiveness of the proposed approach.**

I. INTRODUCTION

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product [1]. These products are often proteins, but in non-protein coding genes such as rRNA genes or tRNA genes, the product is a functional RNA. In recent years, the rapid advance of molecular biology techniques (e.g., microarray analysis) has allowed biologists to generate thousands of gene expression measurements in a short time. Gene Expression Datasets (GEDs) usually collect the expression values of thousands of genes within hundreds of samples. Samples can relate to different organisms or tissues and can be acquired in different environmental conditions.

Data mining, which focuses on studying effective and efficient algorithms to transform large amounts of data into useful knowledge [2], may provide valuable insights into GEDs. Several works have exploited clustering algorithms to identify groups of genes that are strongly correlated with each other, but uncorrelated with those of other groups [3]–[5]. In [6] a step further towards the generation of 3D gene clusters has been made. The authors propose Par-TriCluster, an algorithm that discovers groups of genes behaving similarly across samples and time stamps. The research community also proposed effective classification techniques, i.e. supervised data analysis methods, to correlate gene expression patterns with given classification labels [7]–[9]. In the context of GED analysis, frequent itemset and association rule mining [10] have been exploited to (i) extract biologically relevant co-expressions among multiple genes [11]; (ii) discover correlations between environmental effects and gene expressions [12]; (iii) profile gene expressions according to a worthwhile subset of gene correlations [13]; (iv) determine biological data duplicates [14]. A parallel effort has also been devoted to developing novel itemset mining algorithms that are able to effectively cope with high-dimensional biological data (e.g. GEDs containing thousands of genes) [15], [16]. However, to perform itemset and rule mining, gene expression values are commonly discretized into a predefined number of bins. Specifically, experts are first asked to partition gene expression values into three discrete subsets (i.e., low-expressed, unexpressed, high-expressed). Then frequent itemsets, i.e. sets of co-regulated genes (items) that frequently co-occur in a GED, are extracted from discretized GEDs. The discretization step could bias the quality of the mining result because experts have to assume a reliable data distribution. Consequently, analysts often analyze and compare the results produced by different discretization methods [13], [17].

This paper presents a novel and more effective approach to discovering itemsets from GEDs while avoiding the discretization step. Rather than discretizing gene expression values before executing the itemset mining process, we represent per-sample gene expression values as item weights. In other words, we consider GEDs as weighted datasets [18] for which expression values are mapped to item (gene) occurrences within each sample. Then, *weighted itemsets* are extracted from weighted data. Since item weights can be continuous, discovering weighted itemsets instead of traditional (not weighted) ones prevents experts from discretizing GEDs before analyzing them. For this reason our approach improves the effectiveness of the knowledge discovery process. To the best of our knowledge, this work is the first attempt to discover weighted itemsets from GEDs.

Several weighted itemset mining algorithms (e.g., [18], [19]) have been proposed to consider item weights during the itemset extraction process. In this study we adopted the weighted itemset mining strategy that has recently been proposed in [20]. To demonstrate the effectiveness of our approach we analyzed many real GEDs. The results achieved show the applicability of the proposed approach and significance of the patterns discovered.

This paper is organized as follows. Section II thoroughly describes the weighted itemset mining process from GEDs.

Manuscript received July 25, 2013. This work was supported by the GenData2020 project grant, which is funded by the Italian Ministry of Research (MIUR).

Elena Baralis, Luca Cagliero, Tania Cerquitelli, Silvia Chiusano, and Paolo Garza are with the Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi, 24 10129 Turin, Italy {name}.{surname}@polito.it

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

Section III presents the performed experiments, while Section IV draws conclusions and presents future works.

II. Weighted Itemset Mining from Gene Expression Data

The weighted itemset mining process from Gene Expression Datasets (GEDs) entails the following steps:

(A) *Data preparation.* This step focuses on preparing GEDs to the subsequent itemset mining phase. To make data preparation as simple as possible, we applied the minimal amount of preprocessing steps. Notably, we prevent experts from discretizing gene expression values before executing the itemset mining algorithm.

(B) *Weighted itemset extraction.* The preprocessed GED data is analyzed to discover significant co-expressions among multiple genes.

(C) *Weighted itemset selection and ranking.* To allow experts to manually explore the extracted patterns, the mined itemsets are ranked and filtered according to their main quality measures.

In the following each step is thoroughly described.

A. Data preparation

A GED consists of a set of samples, where for each sample the expression values of a subset of genes is given. For our purposes, we model GEDs as weighted relational datasets. Let us consider a fixed subset G=$\{g_1, g_2, \ldots, g_m\}$ of m genes g_i. Genes will be also called *items* throughout the paper. A weighted relational dataset D=$\{r_1, r_2, \ldots, r_n\}$ is a set of n records r_i, one for each GED sample. A record r_i, $[1 \leq i \leq n]$, consists of a set of pairs $\{\langle g_1, ev_{1i}\rangle, \langle g_2, ev_{2i}\rangle, \ldots, \langle g_m, ev_{mi}\rangle\}$, where $g_j \in G$, $\forall\ 1 \leq j \leq m$. Gene occurrences in r_i are characterized by a weight, which indicates the gene expression value within the corresponding sample. We will denote as ev_{ji} the expression value (weight) of the j-th gene g_j in r_i throughout the paper.

Since the expression values of different genes in different samples are often spread across a relatively large value range, we normalized item weights using z-score normalization [2]. Normalization is commonly applied in GED analysis [13], [17]. Note that, unlike many data discretization methods, z-score normalization does not require experts to set appropriate threshold values.

Table I reports an example of normalized dataset with 3 samples and 4 genes. Gene occurrences within each sample are weighted by the corresponding expression value. For example, the normalized expression value of gene g_1 in sample r_1 is 0.61. Note that genes can take either negative, or null, or positive continuous normalized expression values.

B. Weighted itemset extraction

Frequent weighted itemsets are extracted from a weighted relational GED dataset D using a recently proposed weighted itemset mining strategy [20].

In the context of GED analysis, a k-itemset (i.e. an itemset of length k) is a set of k distinct genes in D. For example, $\{g_1,\ g_2\}$ is a 2-itemset that occurs in Table I. Traditional (not weighted) itemset mining algorithms (e.g. Apriori [21]) commonly generate and select itemsets based on the relative frequency of occurrence (i.e. the support [10]) in the analyzed data (disregarding item weights).

TABLE I

Example of weighted relational dataset

sample ID	**⟨Gene, expression value⟩ pairs**
r_1	$\langle g_1$,0.61$\rangle$, $\langle g_2$,-0.31$\rangle$, $\langle g_3$,-0.72$\rangle$, $\langle g_4$,-0.45$\rangle$
r_2	$\langle g_1$,0.52$\rangle$, $\langle g_2$,0.45$\rangle$, $\langle g_3$,0.28$\rangle$, $\langle g_4$,0.39$\rangle$
r_3	$\langle g_1$,0.51$\rangle$, $\langle g_2$,0.67$\rangle$, $\langle g_3$,0.45$\rangle$, $\langle g_4$,0.38$\rangle$

To consider item weights during itemset mining, the concept of weighted support has already been introduced [18]. The key idea is to weigh itemset occurrences in each record (sample) by the weight (expression value) of the corresponding items (genes). In [20] the occurrences of an arbitrary itemset I in D are weighted by the weight of the least weighted item in I within each sample.

Definition 1: **Weighted itemset support.** Let D be a weighted relational dataset, I a k-itemset, and $G(r_i)$ the subset of genes that are contained in an arbitrary record $r_i \in D$. The weighted support of I in D is defined as follows.

$$\text{wsup }(I, D) \quad = \quad \frac{\sum_{r_i \in D|\ I \subseteq G(r_i)} \min_{j|\ g_j \in I} ev_{ji}}{|D|}$$

In the context of GED analysis, itemsets represent gene combinations, while the weighted support measure indicates their relative frequency of occurrence in D weighted by the expression value of their least expressed gene within each sample. For example, the weighted support of $\{g_1,\ g_2\}$ in Table I is 0.22, because the least weighted gene expression values in r_1, r_2, and r_3 are -0.31, 0.45, and 0.51, respectively.

The frequent weighted itemset mining task entails extracting all the *frequent* weighted itemsets, i.e., the itemsets whose weighted support is equal to or above a given (analyst-provided) threshold wminsup.

However, the mined itemset set is often redundant, because frequent itemsets can represent partially overlapped information. Hence, the interestingness of part of the mining result can be limited. To address this issue, a relevant research effort has been devoted to discovering compact and not redundant frequent itemset subsets [22]–[24]. In this paper we target the extraction of two established itemset subsets, i.e. the maximal and closed itemsets [22], [23], because they have already been considered to be relevant itemset subsets in the context of GED analysis [15], [16].

Closed itemsets. Closed itemsets [22] are frequent itemsets for which none of their immediate supersets have their same support. Since the weighted support measure satisfies the anti-monotonicity property [20], it trivially follows that the immediate supersets of a closed itemset have support strictly less than those of the itemset itself. In the context of weighted itemset mining, I is closed if and only if (i) wsup(I,D) $\geq$ wminsup and (ii) for every $I_2|\ I \subset I_2$ wsup(I_2,D) $\leq$ wsup(I,D).

Recalling the previous example, if we set wminsup=0.10

then $\{g_1,\ g_2\}$ is closed because it is frequent and none of its immediate supersets (i.e., $\{g_1,\ g_2,\ g_3\}$ and $\{g_1,\ g_2,\ g_4\}$) have its same support value (0.22).

Maximal itemsets. Maximal itemsets [23] are frequent itemsets for which all of their immediate supersets are infrequent with respect to the given support threshold. In the context of weighted itemset mining, I is maximal if and only if (i) wsup(I,D) $\geq$ wminsup and (ii) for every $I_2 |\ I \subset I_2$ wsup(I,D) $<$ wminsup. Maximal itemsets are the subset of closed itemsets characterized by maximal length.

For example, if we set wminsup=0.20 then $\{g_1,\ g_2\}$ is maximal because all of its immediate supersets are infrequent with respect to the support threshold.

To extract closed and maximal weighted itemsets, we adapted the FP-Growth-like [25] weighted itemset mining algorithm implementation, which was first proposed in [20], to closed and maximal itemset mining.

C. Weighted itemset selection and ranking

The extracted itemsets are analyzed by domain experts to discover significant co-expressions among multiple genes. Since the number of extracted closed or maximal itemsets can be relatively high, analysts can select the top-K itemsets in order of decreasing weighted support (where K is an analyst-provided parameter). Top-K itemsets are the most frequent gene correlations that occur in a gene expression dataset. According to Definition 1, itemset occurrences are weighted by the expression value of the least expressed gene. Hence, high-support itemsets represent gene combinations for which *all* the genes are highly expressed within each sample. On the other hand, low-support itemsets could represent noisy or less relevant information. Note that if experts are interested in discovering valuable correlations among multiple genes, then the analysis of the 1-itemsets (i.e. itemsets that consist of a single gene) is meaningless. Therefore, in such context of analysis, 1-itemsets can be discarded early to further reduce the mined set cardinality.

III. EXPERIMENTS

We conducted experiments to analyze (i) the number and characteristics of the extracted itemsets and (ii) the biological significance of the mining result. We analyzed five publicly available GEDs, whose main characteristics are summarized in Columns (1)-(3) of Table II. Each GED contains a subset of genes that appear in all samples. BRC-ABL, T-ALL[1], and NeuroBlastoma[2] had already been analyzed in previous research works concerning traditional (not weighted) itemset mining (e.g. [15], [16]), whereas COLON[3] and SRBCT[4] had already been used to assess the performance of biological data classifiers (e.g. [9]).

A. Characteristics of the extracted itemsets

Table II reports the main characteristics of the weighted closed and maximal itemsets that were mined from the analyzed datasets. For each dataset Column (4) indicates the wminsup value enforced, while Columns (5) and (9) report the number of mined weighted closed and maximal itemsets, respectively. To demonstrate that our approach also discovers high-order gene correlations we report the per-length itemset cardinality. For 4 out of 5 datasets 3-length itemsets or longer (i.e. sets of co-expressed genes composed of at least three genes) were extracted. Such patterns are often not considered by previous approaches. Nevertheless to allow experts to manually explore the mining result the number of discovered itemsets should be limited. To achieve this goal without discarding potentially interesting gene co-expressions, experts could consider only sets of co-expressed genes (i.e., 2-length itemsets or longer). This pruning step yields a significant itemset set cardinality reduction (i.e. above 50%) for 4 out of 5 GEDs.

B. Result validation

We validated the significance of the results achieved on two representative gene expression datasets, i.e., BRC-ABL and T-ALL, which had previously been analyzed in [26] to perform classification, subtype discovery, and prediction of outcome in pediatric lymphoblastic leukemia. Specifically, both datasets relate the treatment of pediatric acute lymphoblastic leukemia (ALL). The research goal is to tailor the intensity of therapies to a patient's risk of relapse. Biologists used oligonucleotide microarrays to analyze the pattern of genes expressed in leukemic blasts from 360 pediatric ALL patients. High-density oligonucleotide arrays offer the opportunity to examine patterns of gene expression on a genome scale.

Each dataset consists of a set of expression profiles which are related to a specific prognostically important leukemia subtype, i.e., T-ALL, and BRC-ABL. Let us consider the T-ALL dataset first. Setting a minimum weighted support threshold wminsup=10, the itemset {RPSA RPS23}, with weighted support equal to 11.62, is extracted and ranked first in order of decreasing weighted support. This pattern represents an established correlation between two human ribosomal protein genes [27]. Similarly, the top-ranked itemset {BioB-3, SPECC1L, MAGED2} (wsup=0.01), which was extracted from the BRC dataset, represents a co-expression between the genes BioB-3, SPECC1L and MAGED2, which are targeted by the microarray probes. Gene co-expressions may provide important insights into the biology of the considered leukemia subgroups. Moreover, within each genetic subgroup the expression profiles that are highlighted by the patterns discovered could allow biologists to early identify those patients that would eventually fail therapies.

IV. CONCLUSIONS AND FUTURE WORKS

This paper presents a novel approach to itemset mining from Gene Expression Datasets (GEDs). The aim of this work is to ease GED preparation, which commonly requires

[1] http://www.stjuderesearch.org/data/

[2] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[3] http://genomics-pubs.princeton.edu/oncology/affydata/index.html

[4] http://research.nhgri.nih.gov/microarray/

TABLE II

GENE EXPRESSION DATASETS AND CHARACTERISTICS OF THE ITEMSETS EXTRACTED

Name	Num. of	Num. of	wminsup	Num. of closed				Num. of maximal			
	samples	genes		Total	Len.=1	Len.=2	Len.$\geq$3	Total	Len.=1	Len.=2	Len.$\geq$3
BRC-ABL	15	12625	0.007	10	7	1	2	3	1	0	2
T-ALL	42	12625	10	865	13	112	740	78	2	1	75
COLON	62	2000	-0.0002	1998	1993	2	3	1991	1988	0	3
NEUROBLASTOMA	14	22283	0.007	12	10	2	0	8	6	2	0
SRBCT	88	2308	-0.17	2405	2308	81	16	2314	2245	53	16

a not trivial and expert-driven data discretization step. Instead of discovering traditional itemsets from discretized GEDs, we propose to consider gene expression values as item weights, which indicate gene expression intensity within each sample, and apply a weighted itemset mining algorithm [20] directly to non-discretized GED. The experimental results show the applicability and usefulness of the proposed approach on real GEDs.

Currently, weighted itemset mining algorithms are not designed for efficiently coping with high-dimensional data. Since itemset mining from GEDs is challenged by the so-called "curse of dimensionality" [16], as future work we aim at studying novel weighted itemset extraction algorithms that rely on the vertical data format [15] and integrate taxonomy information (e.g., [28], [29]).

REFERENCES

[1] D. Clark and N. Pazdernik, *Molecular Biology: Understanding the Genetic Revolution*. Elsevier Science, 2012.

[2] P.-N. Tan, M. Steinbach, and V. Kumar, *Intoduction to Data Mining*. Addison Wesley, 2005.

[3] A. Ben-Dor and Z. Yakhini, "Clustering gene expression patterns," in *Proceedings of the third annual international conference on Computational molecular biology*, ser. RECOMB '99, 1999, pp. 33–42.

[4] Y. Cheng and G. M. Church, "Biclustering of expression data," in *ISMB*, 2000, pp. 93–103.

[5] W.-H. Au, K. C. C. Chan, A. K. C. Wong, and Y. Wang, "Attribute clustering for grouping, selection, and classification of gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 2, no. 2, pp. 83–101, Apr. 2005.

[6] R. B. Araújo, G. H. T. Ferreira, G. H. Orair, W. Meira, R. A. C. Ferreira, D. O. G. Neto, and M. J. Zaki, "The partricluster algorithm for gene expression analysis," *Int. J. Parallel Program.*, vol. 36, no. 2, pp. 226–249, Apr. 2008.

[7] Y. Lu and J. Han, "Cancer classification using gene expression data," *Inf. Syst.*, vol. 28, no. 4, pp. 243–268, Jun. 2003.

[8] M. Khashei, A. Zeinal Hamadani, and M. Bijari, "A fuzzy intelligent approach to the classification problem in gene expression data analysis," *Know.-Based Syst.*, vol. 27, pp. 465–474, Mar. 2012.

[9] M. A. Iwen, W. Lang, and J. M. Patel, "Scalable rule-based gene expression data classification," in *ICDE*, 2008, pp. 1062–1071.

[10] R. Agrawal, T. Imielinski, and Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD 1993*, 1993, pp. 207–216.

[11] C. Creighton and S. Hanash, "Mining gene expression databases for association rules." *Bioinformatics*, vol. 19, no. 1, pp. 79–86, 2003.

[12] R. Martinez, N. Pasquier, and C. Pasquier, "Computational intelligence methods for bioinformatics and biostatistics," F. Masulli, R. Tagliaferri, and G. M. Verkhivker, Eds., 2009, ch. Mining Association Rule Bases from Integrated Genomic Data and Annotations, pp. 78–90.

[13] P. C. Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. Carazo, and A. P. Montano, "Integrated analysis of gene expression by association rules discovery," *BMC Bioinformatics*, vol. 7, no. 1, pp. 54+, 2006.

[14] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2007.

[15] F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. J. Zaki, "Carpenter: finding closed patterns in long biological datasets," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '03. ACM, 2003, pp. 637–642.

[16] G. Cong, A. K. H. Tung, X. Xu, F. Pan, and J. Yang, "Farmer: finding interesting rule groups in microarray datasets," in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '04. ACM, 2004, pp. 143–154.

[17] V. Belcastro, V. Siciliano, F. Gregoretti, P. Mithbaokar, G. Dharmalingam, S. Berlingieri, F. Iorio, G. Oliva, R. Polishchuck, N. Brunetti-Pierri, and D. di Bernardo, "Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function." *Nucleic acids research*, vol. 39, no. 20, pp. 8677–8688, Nov. 2011.

[18] W. Wang, J. Yang, and P. S. Yu, "Efficient mining of weighted association rules (WAR)," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'00*, 2000, pp. 270–274.

[19] K. Sun and F. Bai, "Mining weighted association rules without preassigned weights," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 489 –495, 2008.

[20] L. Cagliero and P. Garza, "Infrequent weighted itemset mining using frequent pattern growth," *IEEE Transactions on Knowledge and Data Engineering*, vol. 99, no. PrePrints, p. 1, 2013.

[21] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, J. B. Bocca, M. Jarke, and C. Zaniolo, Eds. Morgan Kaufmann, 1994, pp. 487–499.

[22] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules," in *Proceedings of the 7th International Conference on Database Theory*, ser. ICDT '99. London, UK, UK: Springer-Verlag, 1999, pp. 398–416. [Online]. Available: http://dl.acm.org/citation.cfm?id=645503.656256

[23] J. Roberto and J. Bayardo, "Efficiently mining long patterns from databases," in *SIGMOD 1998*, L. M. Haas and A. Tiwary, Eds., 1998, pp. 85–93.

[24] M. Mampaey, N. Tatti, and J. Vreeken, "Tell me what I need to know: Succinctly summarizing data with itemsets," in *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2011.

[25] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 1–12.

[26] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, K. W. Williams, D. Patel, R. Mahfouz, F. G. Behm, S. C. Raimondi, M. V. Relling, A. Patel, and Cheng, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, no. 2, pp. 133–143, 2002.

[27] K. Ishii, T. Washio, T. Uechi, M. Yoshihama, N. Kenmochi, and M. Tomita, "Characteristics and clustering of human ribosomal protein genes," *BMC Genomics*, vol. 7, no. 1, pp. 1–16, 2006.

[28] L. Cagliero and P. Garza, "Itemset generalization with cardinality-based constraints," *Inf. Sci.*, vol. 244, pp. 161–174, 2013.

[29] E. Baralis, L. Cagliero, T. Cerquitelli, V. D'Elia, and P. Garza, "Support driven opportunistic aggregation for generalized itemset extraction," in *IEEE Conf. of Intelligent Systems*, 2010, pp. 102–107.