Mining the Interval Pattern of the Biomedical Clusters using Greedy Algorithms

Alexey V Galatenko, Stepan A Nersisyan, Vera V Pankratieva
Lomonosov Moscow State University
Leninskie gory 1, 119991 Moscow
Russia

ABSTRACT: *Interval pattern concepts are a particular case of pattern structures. They can be used to clusterize rows of a numerical formal context (data matrix): two rows are close to each other if their entries at the corresponding positions fall within a given interval.*

The problem of mining interval pattern concepts has much in common with the known problem related to computational geometry: given a nite set of points in the Euclidean space, position a box of a given size in such a way that it encloses as many points as possible. This problem and its variations have been thoroughly studied in the case of a plane; however, the authors are not aware of the existence of algorithms which in a reasonable time produce an exact solution in the space of an arbitrary dimension.

There exists an approximate greedy algorithm for solving this problem. It produces a solution with time which is linear in the number of points and polynomial in dimension. We apply a clustering approach based on that algorithm to the gene expression table from the dataset "The Cancer Cell Line Encyclopedia". The resulting partition well agrees with a priori known biological factors.

Keywords: Interval Pattern Concepts, Clustering, Greedy Algorithm

Received: 4. Received: 1 October 2019, Revised 4 January 2020, Accepted 9 January 2020

DOI: 10.6025/pca/2020/9/1/17-23

1. Introduction

In our days researchers frequently need to investigate various biological and medical data represented as numerical contexts (data tables). Rows of tables correspond to objects; columns correspond to attributes. It is often necessary to find clusters that are composed of objects featuring similar attributes. One of the most convenient tools that can be used for clustering this kind of data is Formal Concept Analysis.

Formal concept analysis (FCA) is a data analysis method based on applied lattice theory and order theory. Within the framework of this theory a formal concept is defined as a pair (extent, intent) obeying the Galois connection (see the monograph [1] by B. Ganter and R. Wille).

One of the variations of FCA is known as the theory of pattern structures, which was elaborated by B. Ganter and S. Kuznetsov in [2]. An important particular case of pattern structures is interval pattern structure with the operation of interval intersection, which allows one to apply cluster analysis to rows of numerical contexts [3]. In this case similarity means that all the differences between the values of the corresponding attributes fall into given intervals.

It is easily seen that the problem of detecting similar objects can be reformulated in geometrical terms, namely, as the problem of optimal positioning of a d-dimensional box with given edge lengths for the set P of points, i.e. finding a position of the box that maximizes the number of points of the set P enclosed by the box (here $d \in N$ is the number of attributes in the numerical context considered, P is the set generated by the rows of the numerical context).

In practice, biomedical data often involve thousands of entries, and each entry is described by hundreds of attributes. The existing algorithms that solve the problem of finding an optimal position of a box do not allow one to obtain an exact solution for high-dimensional data within a reasonable time. In [4] the authors introduced a fast approximate greedy algorithm for solving this problem and applied the corresponding clustering approach to the dataset of tactile images registered by the Medical Tactile Endosurgical Complex (MTEC, [5]). The experiment results demonstrated significant advantage of the proposed algorithm over the conventional k-means method in clustering quality.

In this paper we apply this clustering algorithm to the dataset "The Cancer Cell Line Encyclopedia" [6]. This dataset includes an expression table for about 20000 genes in 917 cancer cell lines. The cell lines were derived from tissues of 23 different organs. The aim of the study is to check if cancers from close organs have close gene expression values.

The rest of the paper is organised as follows. In Section 2 we introduce definitions from the formal concepts theory. In Section 3 we overview the clustering algorithm from [4]. In Section 4 we describe the procedure and present the results of application of the algorithm to the gene expression data, and in Section 5 we make concluding remarks.

2. Main Definitions

In this section we briefly recall the main definitions of the theory of formal concepts and give a geometrical interpretation of the problem of finding an interval pattern concept of maximum extent size.

Definition 1: *A semilattice operation on the partially ordered set* $(M, \leq)$ *is a binary operation* $\sqcap: M \times M$ *that features the following properties for a certain* $e \in M$ *and any elements* $x, y, z \in M$*:*

$x \sqcap x = x$ *(idempotency);*

– $x \sqcap y = y \sqcap x$ (*commutativity*);

– $(x \sqcap y) \sqcap z = x \sqcap (y \sqcap z)$ *(associativity);*

– $e \sqcap x = e$.

Definition 2: *Let* $(P, \leq P)$ *and* $(Q, \leq Q)$ *be partially ordered sets. A Galois connection between these sets is a pair of maps* $\varphi : P \to Q$ *and* $\psi : Q \to P$ *(each of* them is referred to as a Galois operator) such that the following relations hold *for any* $p_1, p_2 \in P$ *and* $q_1, q_2 \in Q$:

– $p_1 \leq_P p_2 \Rightarrow \varphi(p_1) \geq_Q \varphi(p_2)$ (*anti-isotone property*);

– $q_1 \leq_Q q_2 \Rightarrow \psi(q_1) \geq_P \psi(q_2)$ (*anti-isotone property);*

– $p_1 \leq_P \psi(\varphi(p_1))$ *and* $q_1 \leq_Q \varphi(\psi(q_1))$ *(isotone property).*

Applying the Galois operator twice, namely, $\psi(\varphi(p))$ *and* $\varphi(\psi(q))$*, defines a* closure operator.

Definition 3: *A closure operator* $\overline{(\cdot)}$ *on M is a map that assigns a closure* $\overline{X} \subseteq M$ *to each subset* $X \subseteq M$ *under the following conditions:*

– $X \leq Y \Rightarrow \overline{X} \leq \overline{Y}$ *(monotony);*

– $X \leq \overline{X}$ *(extensity);*

– $\overline{\overline{X}} = \overline{X}$ *(idempotency).*

Definition 4: *A pattern structure is a triple* $(G, (D, \sqcap), \delta)$*, where G is a set of objects,* $(D, \sqcap)$ *is a meet-semilattice of potential object descriptions, and* $\delta : G \to D$ *is a function that associates descriptions with objects.*

The Galois connection between the subsets of the set of objects and the set of descriptions for the pattern structure $(G, (D, \sqcap), \delta)$*, is defined as follows:*

$$A^{\square} := \sqcap_{g \in A} \delta(g), \qquad where\ A \subseteq G,$$
$$d^{\square} := \{g \in G \mid d \sqsubseteq \delta(g)\}, \qquad where\ A \subseteq G.$$

Definition 5: *A pattern concept of the pattern structure*$(G, (D, \sqcap), \delta)$ *is a pair* (A, d)*, where* $A \subseteq G$ *is a subset of the set of objects and* $d \in D$ *is one of the descriptions in the semilattice, such that* $A^{\square} = d$ *and* $d^{\square} = A$*; A is called the pattern extent of the concept and d is the pattern intent.*

A particular case of a pattern concept is the interval pattern concept. The set *D* consists of rows of a numerical context which are treated as tuples of intervals of zero length. An interval pattern concept is a pair (*A, d*), where *A* is a subset of the set of objects and *d* is a tuple of intervals with ends determined by the smallest and the largest values of the corresponding component in the descriptions of all objects in *A*.

Since interval pattern concepts are determined by objects that have similarly "distributed" attributes, these concepts are convenitent to use in data clustering. The interval width can be either the same for all components (in such case it is denoted by δ), or different for different components (in such case the widths are denoted by $\delta_1, \delta_2, \ldots, \delta_d$).

Let *P* be a set of *n* points in $\mathbb{R}^d \ (d \in \mathbb{N})$, $\delta_1, \delta_2, \ldots, \delta_d$ be positive real numbers.

Definition 6: A *d-orthotope* (also called a box) with center $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and edge lengths $\delta_1, \delta_2, \ldots, \delta_d$ is the Cartesian product of the intervals

$$\left[x_1 - \frac{\delta_1}{2}, x_1 + \frac{\delta_1}{2}\right] \times \ldots \times \left[x_d - \frac{\delta_d}{2}, x_d + \frac{\delta_d}{2}\right]$$

It can be easily seen that the problem of identification of a maximum interval concept can be reformulated in terms of nding an – optimal position of the box with the edge lengths $\delta_1, \delta_2, \ldots, \delta_d$, that is, maximizing the number of points of the set *P* enclosed by the box. This formulation can be generalized to the problem of finding an optimal position of a ball in an arbitrary metric space, since any box can be treated as a ball in the stretched L_∞ metric in which the distance $\rho(x, y)$ between the points $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$ is defined as

$$\rho(x, y) = \max_{1 \leq i \leq d} \delta_i |x_i - y_i|.$$

3. The Greedy Clustering Algorithm Based on Interval Pattern Concepts

In this section we briefly overview the greedy clustering algorithm which was introduced in [4]. Given the set $P = \{p_i\}_{i=1}^n \subset \mathbb{R}^d$, the algorithm splits it into mutually disjoint clusters $C_1, \ldots, C_k$. The splitting procedure is based on optimal box positioning and uses a standard greedy approach. Namely, at each step an optimal position D_i of the box for the set $P \backslash (C_1, \ldots, C_{i-1})$ is determined,

and C_i is assigned to be equal to $P\setminus(C_1, \ldots, C_{i-1})\cap D_i$. In order to avoid producing a big number of small clusters consisting of outliers, the algorithm uses a restriction on the number of points in the resulting clusters — they must include at least cmin objects. With this restriction some points can be considered unclustered.

The clustering procedure uses the approximate greedy iterative algorithm for solving the problem of an optimal box positioning. The parameters of that algorithm are the box edge lengths $\delta_1, \delta_2, \ldots, \delta_d$, the positive real numbers s, s_{min}, $\lambda < 1$ and the function $f: \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. The parameters s, s_{min}, and regulate the duration of one iteration, while the function f returns the number of iterations for the given values n and d. Greater number of iterations and greater duration of each iteration provide better approximation.

Now we will briefly describe the greedy algorithm for finding an approximately optimal position of a box. After a short preprocessing procedure the box with the edge lengths $\delta_1, \delta_2, \ldots, \delta_d$ is transformed into the d-dimensional unit cube, and the algorithm locates the *base unit cube*, i.e. the optimal unit cube with integer vertice coordinates. The main idea of the algorithm consists in constructing $f(n, d)$ sequences of unit cubes in such a way that each sequence starts from a random point in the base unit cube and satisfies the condition that the next cube contains more points than the previous one. After that the algorithm returns a locally optimal cube C. Each sequence is constucted iteratively. Suppose that m cubes from a sequence are already constructed. There are two possible cases.

1. If the current cube can be translated with the current step by one of the axes (the initial step size is equal to s) with an increase in the number of enclosed points, then the current cube is moved to this position.

2. Otherwise, the current step size is decreased by a factor of $\lambda < 1$. If the step size threshold s_{min} is reached then the procedure is terminated.

Under additional technical restrictions the authors of [4] proved the following precision and complexity bounds.

Theorem 1: *Let D_{alg} be an optimal cube produced by the algorithm and D_{opt} be a globally optimal cube. Then and this estimate is sharp.*

$$\frac{1}{2^d} \le \frac{D_{alg} \cap P}{D_{opt} \cap P} \le 1$$

Theorem 2: *The algorithm for finding an approximately optimal position of the box has*

$$O\left(dn\log(n) + \frac{d^3 n^{1-\frac{1}{d}}}{s_{min}} f(n,d)\right)$$

worst-case time complexity and $O(dn)$ space complexity.

Theorem 3: *The clustering algorithm has*

$$O\left(\left(dn\log(n) + \frac{d^3 n^{1-\frac{1}{d}}}{s_{\min}} f(n,d)\right) \cdot \frac{n}{c_{\min}}\right)$$

worst-case time complexity and $O(dn)$ space complexity.

4. Applying the Clustering Algorithm to "The Cancer Cell Line Encyclopedia"

We consulted biologists and selected 432 columns of the expression table associated with genes encoding receptors, channels and transcription factors. First, we applied the clustering algorithm to the whole table. Thus, in our notation we have n equal to 917 and d equal to 432. For tuning algorithm parameters we used the following procedure. Let D denote the maximal pairwise

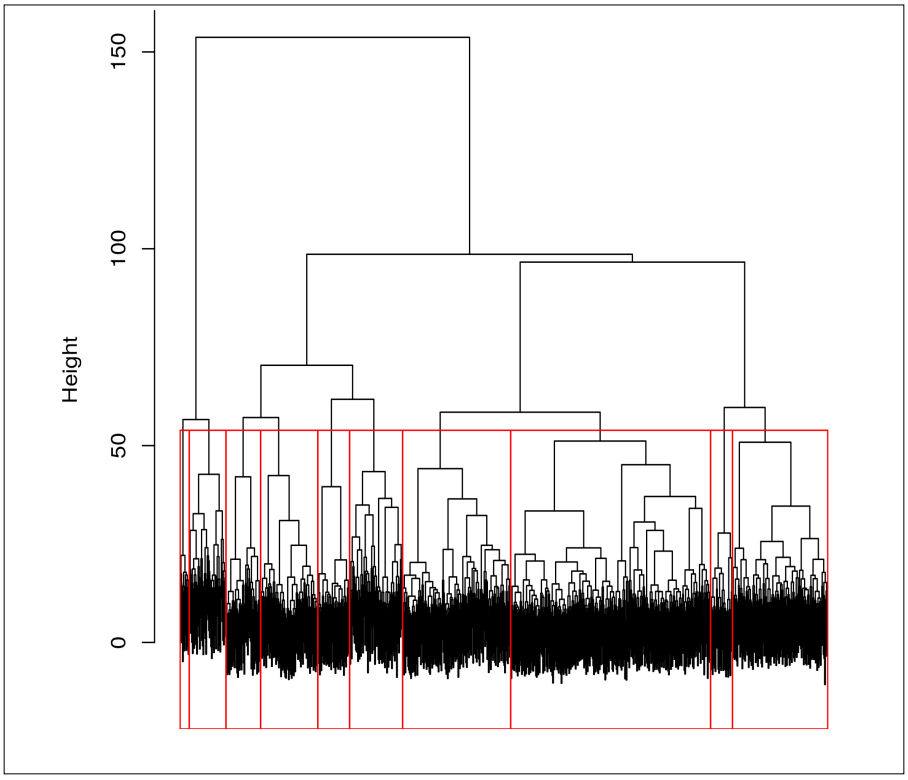

Figure 1. Dendrogram of the hierarchial clustering of features

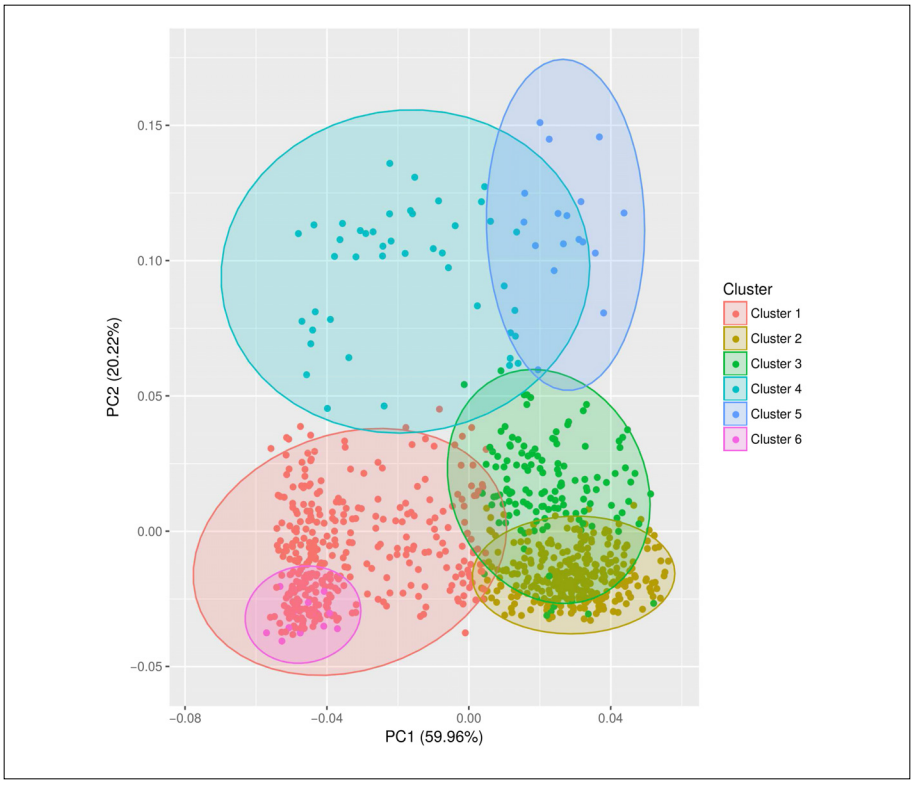

Figure 2. Plot of the first two principal components for the clusters; 29 outlying samples are removed from this figure

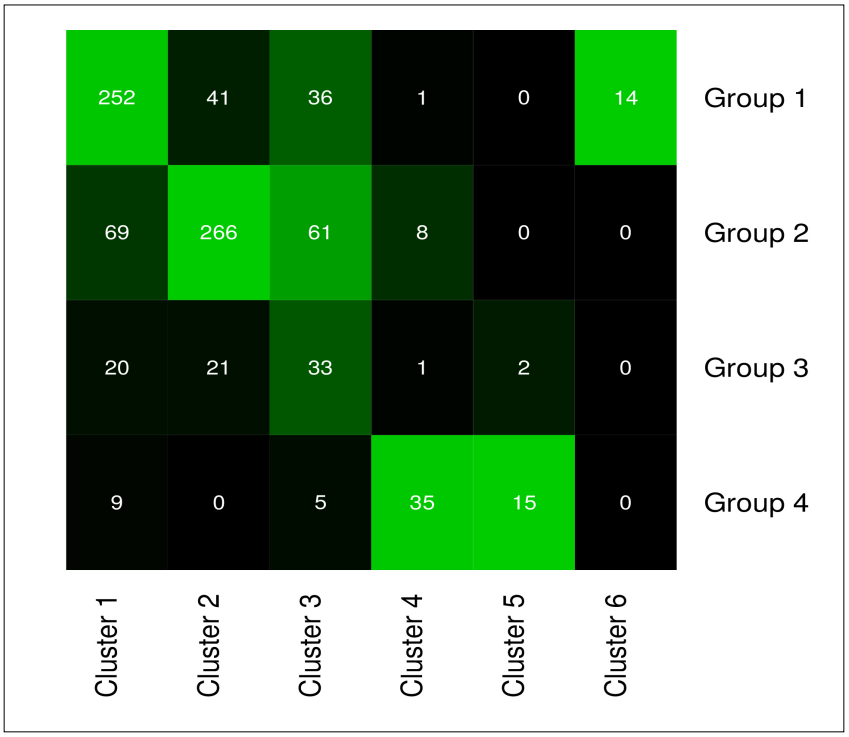

Figure 3. Mutual arrangment of the clusters and organ groups. The number at the intersection of the i^{th} row and the jth column indicates the number of samples which fall into both Group i and Cluster j; 29 outlying samples are excluded from consideration

distance between the points considered. By the Pythagoras theorem, all points can be placed in a cube with edge length $D\sqrt{d}$. Then, a simple grid search approach on the interval $(0, D\sqrt{d})$ was utilized for finding an acceptable cube edge length. The remaining parameters were manually tuned in order to reach acceptable (accuracy) / (running time) ratio.

We seleceted the cube edge length equal to 6.7 (i.e. $\delta_1 = \delta_2 = \ldots \delta_{432} = 6.7$); c_{min}, s, s_{min} and λ were set equal to 10, 0.5, 0.3, 0.9, respectively, and the function $f(n, d)$ was taken as $[log(dn)]$, where $[x]$ denotes the integer part of x. Despite an acceptable run time (several minutes) the results were unsatisfactory: the output of the algorithm included one huge 390-element cluster, two medium-sized 77- and 69-element clusters, and the remaining approximately optimal cubes contained less than 10 points each. This means that more than 40% of samples (381 out of 917) actually were not clusterized. Such behavior was the result of strictness of the relation "a point lies in a box" which means that each coordinate of a point must fall into a fixed range. Under this restriction, even one outlying coordinate of a point knocks it out of a cube. In high dimensional spaces single coordinate outliers are quite probable and inevitable, so before using the clustering algorithm it is reasonable to apply some dimension reduction and smoothing technique.

We applied Ward's method of hierarchial clustering to data features (R function hclust from the package stats [7] was used). The dendrogram produced (Figure 1) was cut at height 55, which corresponds to 10 clusters. Then the expression values in the clusters were averaged. The new feature space had dimension d equal to 10. The greedy clustering algorithm was run on the dataset with reduced dimension with the cube length equal to 3; the other parameter values were left unchanged. The number of outliers essentially decreased after moving to the new agglomerated feature space — their quantity varied in the range between 25 and 35. The resulting partition consisted of 6 groups (see Figure 2) and had an interesting biological interpretation. Namely, we calculated the number of samples in all intersections of clusters and organs. Based on this cardinalities we concluded that the clusters obtained were highly correlated with organ groups (see Figure 3):

– **Group 1:** Haematopoietic and lymphoid tissue, liver, skin, central nervous system, bone, soft tissue, pleura;

– **Group 2:** Salivary gland, upper aerodigestive tract, oesophagus, biliary tract, stomach, pancreas, small intestine, large intestine,

breast, thyroid, endometrium, urinary tract, lung (non-small cell cancer);

– Group 3: Kidney, ovary, prostate;

– Group 4: Autonomic ganglia and lung (small cell cancer).

It can be seen that major organ systems fall into different groups. Namely, Group 1 contains almost all non-solid organs, Group 2 contains organs from digestive system, Group 3 contains organs from genitourinary system, and Group 4 contains organs from autonomic nervous system and respiratory system. However Groups 2 and 3 seem to be dependent: Group 2 also contains some organs from genitourinary system. Thus the clusters differ by the organ systems they contain. Note that Figures 2 and 3 give ground to merge Cluster 6 with Cluster 1 and Cluster 5 with Cluster 4. The quality of the clusters can be additionally illustrated by more subtle arguments. For example, separation of small and non-small cell lung cancers seems to be reasonable, since there are some receptor coding genes which are differently expressed in these cancer types [8]. Note also that small cell lung cancer appear in the same cluster with the autonomic ganglia cancer (neuroblastoma), since their molecular mechanisms include some number of the same receptors [9].

5. Conclusions

In this paper we tested the applicability of the greedy clustering algorithm based on interval pattern concepts from the paper [4] to high-dimensional biomedical data. We showed that the clusters produced by the algorithm applied to "The Cancer Cell Line Encyclopedia" dataset were highly correlated with different organ groups and sophisticated molecular mechanismes of different cancer types.

References

[1] Ganter, B., Wille, R. (1999). Formal Concept Analysis. Springer-Verlag, Berlin.

[2] Ganter, B., Kuznetsov, S. (2000). Pattern Structures and Their Projections. Preprint MATH-AL-14-2000, Technische Universit at Dresden, Herausgeber, Der Rektor.

[3] Kaytoue, M., Duplessis, S., Kuznetsov, S.O., Napoli, A. (2009). Two FCA-Based Methods for Mining Gene Expression Data. In: S. Ferr´e and S. Rudolph (Eds.): ICFCA 2009, LNAI 5548, p 2511–266, 2009.

[4] Nersisyan, S.A., Pankratieva, V. V., Staroverov, V. M., Podolskii, V. E. (2017). A Greedy Clustering Algorithm Based on Interval Pattern Concepts and the Problem of Optimal Box Positioning. *Journal of Applied Mathematics*, Article ID 4323590 (2017)

[5] Barmin, V., Sadovnichy., Sokolov, M., Amiraliev, A., Pikin, O. (2014). An original device for intraoperative detection of small indeterminate nodules. *European Journal of Cardio-thoracic Surgery*, 46 (6), 1027–1031.

[6] Barretina, J., Caponigro, G., Stransky, N. et al.: The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483(7391), 603– 607.

[7] R Core Team: R. (2018). A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[8] Wistuba, I. I., Gazdar, A. F., Minna, J. D. (2001). Molecular genetics of small cell lung carcinoma. *Seminars in Oncology*, 28(2 Suppl 4), 3–13.

[9] Stone, J. P., Wagner, D. D. (1993). P-selectin mediates adhesion of platelets to neuroblastoma and small cell lung cancer. *The Journal of Clinical Investigation,* 92 (2).