Selecting Outstanding Patterns Based on Their Neighbourhood

Etienne Lehembre, Ronan Bureau, Bruno Crémilleux, Bertrand Cuissart, Jean-Luc Lamotte, Alban Lepailleur, Abdelkader Ouali, Albrecht Zimmermann

Selecting Outstanding Patterns Based on their Neighbourhood

Etienne Lehembre[1 [0000−0002−1374−5453]], Ronan Bureau[2 [0000−0001−9404−8117]], Bruno Cremilleux[1 [0000−0001−8294−9049]], Bertrand Cuissart[1 [0000−0003−4964−5427]], Jean-Luc Lamotte[2,3 [0000−0001−6493−1769]], Alban Lepailleur[2 [0000−0003−0202−1588]], Abdelkader Ouali[1 [0000−0003−0855−0181]], and Albrecht Zimmermann[1 [0000−0002−8319−7456]]

[1] GREYC, CNRS UMR 6072, UNICAEN, Normandy Univ. Caen, France
{etienne.lehembre,bruno.cremilleux,bertrand.cuissart,abdelkader.ouali, albrecht.zimmermann}@unicaen.fr
[2] CERMN, EA 4258 FR CNRS 3038 INC3M SF 4206 ICORE, UNICAEN, Normandy Univ. Caen, France
{ronan.bureau,jean-Luc.lamotte,alban.lepailleur}@unicaen.fr
[3] Sorbonne Université, UFR 919, 4 place Jussieu F-75252 Paris cedex 05

Abstract. The purpose of pattern mining is to help experts understand their data. Following the assumption that an analyst expects neighbouring patterns to show similar behavior, we investigate the interestingness of a pattern given its neighborhood. We define a new way of selecting *outstanding* patterns, based on an order relation between patterns and a quality score. An outstanding pattern shows only small syntactic variations compared to its neighbors but deviates strongly in quality. Using several supervised quality measures, we show experimentally that only very few patterns turn out to be outstanding. We also illustrate our approach with patterns mined from molecular data.

Keywords: Pattern selection · Structured pattern mining · Local deviation

1 Introduction

The purpose of data mining is to help experts to analyze their data by providing valuable results. When those results come in the form of patterns, whether conjunctions of attributes or items, sequences, trees, or graphs, a recurring problem is that there are simply too many of them for a human to work through. Once this problem was recognized, research first focused on reducing the output through the notion of *condensed representations* [11], a plethora of *quality measures* [13], and *pattern set mining* techniques [7] were designed, all of which fall short, however. Even when creating condensed representations, there are typically still hundreds or even thousands of patterns left, as is the case no matter which quality measure one uses. In addition, the latter lead to the question which measure to use for a given task. Pattern set mining, finally, works well

enough when the goal is to create a set of non-redundant patterns to be used as descriptors in downstream tasks such as classification or clustering, but less so when it comes to offering an expert an interpretable result set.

Here, we start from the assumption that an analyst expects that patterns which are neighbors in the pattern space show similar behavior. Hence, a pattern showing different behavior from what one expects given similar patterns deserves a second look. To find these patterns, which we will call *outstanding* going forward, we use a Hasse Diagram (HD), a directed acyclic graph (DAG), as a representation of the pattern space. This DAG encodes a partial order between patterns whose interestingness is quantified by a quality measure. Patterns that are scored very differently than the average of neighboring patterns are considered *outstanding*.

The main contribution of the paper is a new way of selecting outstanding patterns, given an order relation between patterns, and a quality measure on patterns. We formulate our idea in general terms since it can be applied for any pattern language (e.g., items, sequences, graphs). With items, we illustrate our approach by using the lattice of formal concepts derived from data as the encoding HD. We define the notion of a *selector*, a function that outputs the set of outstanding patterns given a HD and a quality measure. Outstanding patterns will then be those that show only small syntactic variations compared to their neighbors but deviate strongly in quality. Notably, this deviation is not necessarily positive: a pattern might be outstanding because it correlates much more weakly with a class label, for instance, than its neighbors. Using several supervised quality measures, we show experimentally that only very few patterns turn out to be outstanding and that the number varies depending on the measure. Our contribution is an outgrowth of the concept of activity cliffs [12] on molecular data, which define a noticeable modification of the biological activity for a small modification of the chemical structure. We therefore also illustrate our method on using patterns mined from molecular data, which are the main focus of our application interest.

The paper is organized as follows. In the next section, we discuss the literature related to our problem setting and proposal. In Section 3, we introduce necessary background knowledge. In Section 4, we present the selector. In Section 5, we report experimental results on transactional data derived from UCI data sets and on molecular data and discuss them. We conclude in Section 6.

2 Related work

Since the introduction of constraint-based pattern mining, an on-going theme has been how to help the experts identify the most valuable patterns from result sets containing thousands or even millions of them. A well established solution is to find a condensed representation of the patterns such as closed [11, 17] or free patterns [3], i.e., maximal or minimal patterns from the support-based equivalence classes. Since real data are often noisy, [3] proposed error-tolerant variants.

Another direction is to focus on the best patterns according to quality measures [13]. The survey [15] divides measures in two categories: absolute measures and advanced ones. Advanced measures are based on statistical models (independence model, partition models, MaxEnt models) having different complexities. However, there are numerous measures and it remains difficult to clearly identify the advantages and limitations of each one. The quality of a selected pattern can be assessed via syntactically linked patterns during computation [4], somewhat similar to our proposal.

Recent research has highlighted the benefits of the *unexpectedness* of a pattern when contrasted with given information depending either on the data or on prior knowledge of the analyst [2]. For instance, by sampling patterns fulfilling data-independent constraints under assumptions about the symbol distribution (i.e. null models), the authors of [1] derive a model of background noise, and identify thresholds expected to lead to interesting results, i.e. results that diverge from the expected support derived from super- and sub-patterns. Another approach combines sampling and isotonic regression in order to arrive at pattern frequency spectrum for frequent itemset mining [14]. By comparing those thresholds to ones derived from data where all items are independent, one can identify thresholds or which the result set is expected to contain interesting patterns. Self-sufficient itemsets, finally, are itemsets the support of which cannot be predicted from their sub-sets or super-sets [16]. However, these approach are limited to itemset data. Our method differs in that we do not make assumptions about syntactic relationships between patterns. In addition, we do not make an independence assumption w.r.t. pattern elements.

Also closely related to our work, in the context of web queries modeled according to the setting of the Formal Concept Analysis, [5] uses the siblings of a node to define the interestingness of a new query. However, the method does not take into account the whole set of siblings and it is linked to frequencies observed in the extents and intents of the concepts whereas our approach can use any quality measure defined on patterns.

3 Background

As usual in the pattern mining paradigm, let us consider $\mathcal{D}$ a dataset, $\mathcal{L}$ a pattern language and $\preceq$ a partial order relation on the patterns in $\mathcal{L}$. The support of a pattern p, $Supp(p)$, is the number of transactions containing p. The pattern space can be modelled by its Hasse diagram, a DAG whose set of vertices maps the set of patterns and whose edges depict the order relation: there is an edge (p, q) from a pattern p to a pattern q if $p \preceq q$ and if there is no other pattern r between p and q ($p \preceq r$ and $r \preceq q$). From an edge (p, q), we say that p is a *parent* of q, that q is a *child* of p. The *siblings* of a pattern is the set of patterns that share a common parent with it. Figure 1 depicts an example of these relationships: the siblings of the pattern S (in red) are S_i (in purple), the parents of S are P_i (in blue).

In the itemset setting, $\mathcal{D}$ is a set of transactions, each transaction containing one or more distinct literals called items I. A pattern X is an element of 2^I. The order relation on the patterns is the usual inclusion relation $\subseteq$. In the itemset setting and considering closed itemsets [11], the Hasse diagram is then a Galois lattice [6].

Many quality measures have been described in the literature [13, 15] and the interestingness of a pattern will be quantified by a measure $f : \mathcal{L} \times \mathcal{D} \mapsto \mathbb{R}$.

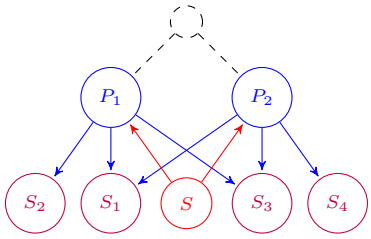

Fig. 1. Retrieving siblings (purple / S_i and red / S) from a source vertex (red) and its parents (blue / P_i).

4 Outstanding Pattern Selector: how to exploit siblings

To select outstanding patterns, the method is based on the principle that an analyst expects patterns that are neighbors in the pattern space to show similar behavior. Therefore a pattern showing different behavior from its neighbors according to a quality measure f deserves attention. The sibling patterns being structurally close, their quality should be similar. If a pattern is scored differently from its siblings, it is highly interesting as a outstanding sibling. Thus, we seek for local variations of interestingness. This phenomenon is not captured when f is applied to each pattern individually, as is usual in the frequent or association pattern setting. Concretely, we say that a pattern X is *outstanding* when its quality deviates from the mean quality of its siblings $\mathcal{S}(X)$. The *sibling mean* $\mu(\mathcal{S}(X), \mathcal{D})$ is:

$$\mu(\mathcal{S}(X), \mathcal{D}) = \frac{\sum\limits_{s \in \mathcal{S}(X)} f(s, \mathcal{D})}{|\mathcal{S}(X)|}$$

Then $\mu(\mathcal{S}(X), \mathcal{D})$ is compared to the standard deviation of the siblings:

$$\sigma(\mathcal{S}(X), \mathcal{D}) = \sqrt{\frac{\sum_{s \in \mathcal{S}(X)} (f(s, \mathcal{D}) - \mu(\mathcal{S}(X), \mathcal{D}))^2}{|\mathcal{S}(X)|}}$$

The selector is defined as:

$$OPS(\mathcal{L}, f, \mathcal{D}, \delta) = \{X \in \mathcal{L} : \ |f(X, \mathcal{D}) - \mu(\mathcal{S}(X), \mathcal{D})| \geq \delta * \sigma(\mathcal{S}(X), \mathcal{D})\}$$

Thus, X is outstanding if its quality deviates at least δ standard deviations from the mean of the qualities of its siblings, δ being a user-supplied parameter. We consider the quality measure as a random variable, which distribution varies locally, while staying normally distributed around a local mean. Moreover, the behavior of the quality measure will impact the selection. A homogeneous quality measure will lead the selector to select a few chosen ones while an heterogeneous quality measure will produce more outliers.

One of the appeals of using the standard deviation instead of a classic threshold is that the selector adjusts to its environment: if the siblings of a pattern are all relatively close to a particular support value, a small increase over this value can be interesting. Similarly, take the example of the *growth rate* [8] as the quality measure f. Let us assume furthermore that $\mathcal{D}$ is partitioned into two classes, and most of the siblings are *Jumping Emerging Pattern* (JEP) [8, 9], i.e. patterns that have a support of zero in the negative class. JEPs have a tendency to overfit; our selector, on the other hand, keeps a JEP only if it indicates a local deviation. Moreover, it can select interesting patterns that are not JEP.

In practice, as shown in the next section, the number of outstanding patterns is small, allowing a human domain expert to manually inspect them.

5 Experiments

In this section, we show experimental results illustrating the reduction in patterns, as well as the behavior of four quality measures. In the next section, we provide results on itemset data, and in Section 5.3 on graph data representing molecules. We use our experiments to answer several questions:

- Does selecting outstanding patterns reduce the size of the result set significantly?
- Does changing the quality measure change how many patterns are outstanding?
- Can outstanding patterns be easily characterized in terms of the score they receive from an interestingness measure?
- Do outstanding patterns from self-sufficient itemsets, another type of pattern that takes itemsets' neighborhoods into account, albeit syntactic ones?

5.1 Itemset data

The data we used are itemset data derived from UCI data sets, which we downloaded from the CP4IM repository[4]. The data have been binarized by the maintainers of the repository, the majority class named positive class, and minority classes merged into a single negative class.

We performed closed frequent set mining with minimum support thresholds (denoted by θ) of 10%, 15%, and 20%. In the resulting graph $G(\mathcal{V}, \mathcal{E})$ each vertex

[4] https://dtai.cs.kuleuven.be/CP4IM/datasets/

Table 1. Characteristics for selected UCI datasets and their number of self-sufficient itemsets.

Data set	Mushroom	Primary-tumor	Soybean	Splice-1
Transactions	8124	336	630	3190
Items	119	31	50	287
Density	18%	48%	32%	21%
Self-sufficient itemsets	69	16	55	38
for $\theta = 10\%/15\%/20\%$	69/68/53	13/11/8	49/33/30	30/9/3
Data set	Tic-tac-toe	Vote	Zoo-1	
Transactions	958	435	101	
Items	27	48	36	
Density	33%	33%	44%	
Self-sufficient itemsets	24	39	64	
for $\theta = 10\%/15\%/20\%$	24/24/0	39/39/39	62/60/48	

is labeled with a closed itemset. We tested four quality measures: χ^2, confidence, normalized Growth Rate (NGR)[5]:

$$\begin{cases} NGR(X,\mathcal{D}) = 1.0 & if\ GR(X,\mathcal{D}) = \infty \\ NGR(X,\mathcal{D}) = \frac{GR(X,\mathcal{D})}{1+GR(X,\mathcal{D})} & otherwise \end{cases}$$

and Weighted Relative Accuracy (WRAcc). For the latter three, we chose the positive class as target. For the OPS threshold, we chose $\delta = 2$ since 95% of all values of a normal distribution fall into the interval $[\mu - 2 \cdot \sigma, \mu + 2 \cdot \sigma]$.

As Figure 2 shows, only very few itemsets are outstanding compared to their siblings, with at most 3.052% selected by confidence and NGR on the *splice* data set for the 10% minimum support threshold. Notably, this is in addition to the reduction achieved by mining closed itemsets. We take this as evidence that selecting outstanding patterns results in small enough result sets that domain experts could inspect them (and their neighborhoods) manually to gain deeper insight into the underlying phenomena. We can also compare the behavior under different support thresholds, i.e. the results for a single data set and a single measure, and for different quality measures, i.e. the results in a single line.

While increasing the support threshold mostly reduces the number of outstanding patterns as well, this is not always the case, as can be seen for the *zoo-1* data set, for instance. Using confidence or GR as a quality measure leads to fewer outstanding patterns than using χ^2 and WRAcc does, with the exception of the *splice-1* (10%) and *vote* (10%, 15%) data sets. A particularly remarkable data set is the *tic-tac-toe* one where not a single pattern stands out.

Notably, using different quality measures lead to different sets of outstanding patterns to be selected. Figure 3 shows a heatmap representation of the Jaccard similarity between result sets for three example data sets. For the *primary-tumor*

[5] We normalize the growth rate because the unnormalized growth rate can have ∞ as a value, which prevents the calculation of mean and standard deviation.

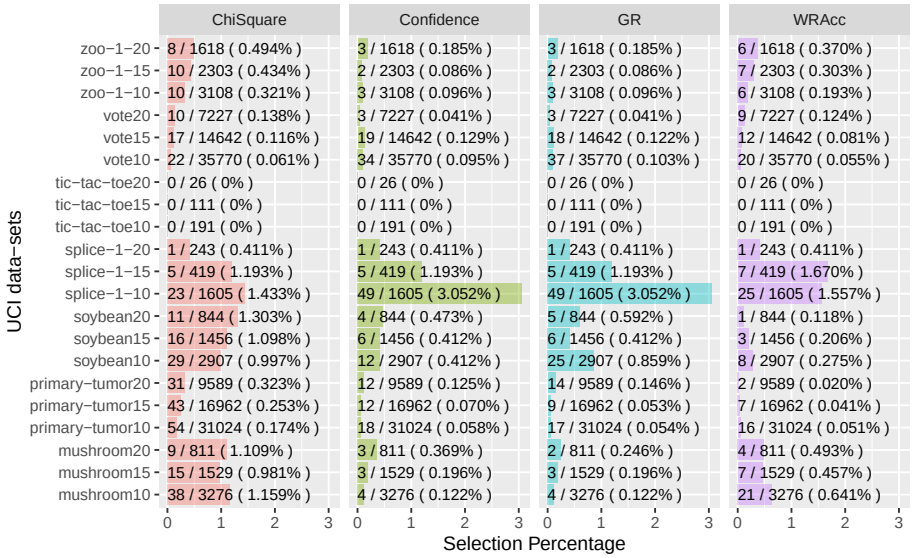

Fig. 2. Selection statistics (#outstanding patterns/#patterns/%) on UCI data-sets.

data set (middle), there is little similarity between the different results, for the *soybean* data set (right-most figure), *confidence* and *GR* give very similar results. The full set of figures can be found in the supplementary material.

As we mentioned in the introduction, outstanding patterns are not necessarily among the *best* patterns in terms of class correlation, for instance. This is shown by Fig. 4 on the primary tumor data set with the confidence measure: for all minimum support thresholds, also itemsets with low confidence are selected. Figures for other data sets and quality measures can be found in the supplementary material available at https://github.com/Etienne-Lehembre/Outstanding-Pattern-Selector.git.

5.2 Comparison to self-sufficient itemsets

A method that is close in spirit to our proposal are the *self-sufficient* itemsets proposed by Webb *et al.* [16]. Self-sufficient itemsets, can be considered independently from each other, as can outstanding patterns, which is not the case for patterns selected by pattern mining techniques. The full definition of self-sufficiency is too involved to reproduce here[6] but self-sufficiency includes the requirement that the probability of itemsets' occurrence cannot be inferred by the probability of subsets' and supersets' occurrences. This requirement translates into comparing itemsets to their predecessors and successors in a DAG where vertices are labeled with the full set of possible itemsets and edges indi-

[6] We direct the interested reader to the original publication.

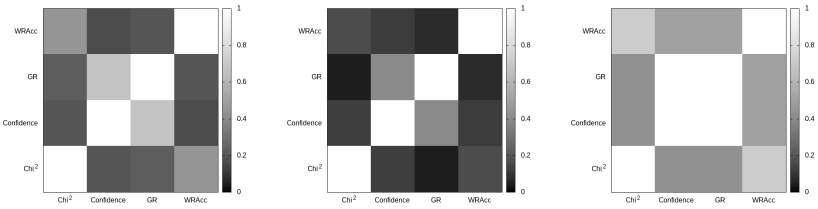

Fig. 3. Heatmap representation of Jaccard similarity for sets of outstanding patterns selected for different quality measures for *mushroom* ($\theta = 10\%$), *primary-tumor* ($\theta = 15\%$), *soybean* ($\theta = 15\%$).

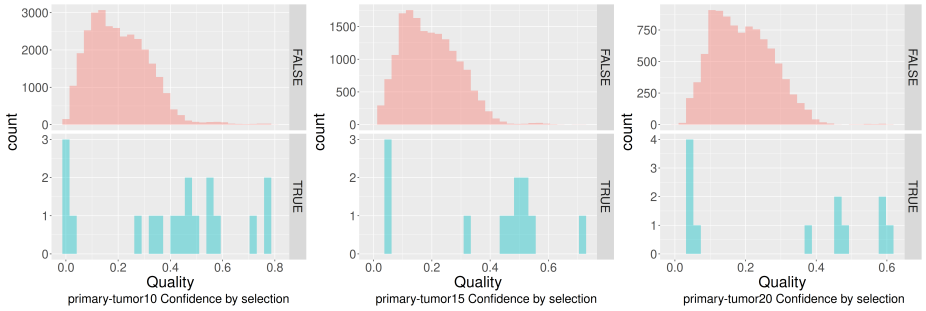

Fig. 4. Distribution of confidence values for the *primary tumor* data set, outstanding patterns in the bottom row, non-outstanding patterns on top. Results for minimum support 10% in the left-most column, 15% center, 20% right-most.

cating extension of itemsets with individual items. We therefore want to know how many of the outstanding itemsets we select are self-sufficient and vice versa.

We ran the OpusMiner implementation available at https://eda.mmci.uni-saarland.de/prj/selfsufs/ on the UCI data sets mentioned above. The lower part of Table 1 reports the number of self-sufficient itemsets and a comparison to Figure 2 shows that there is no obvious relationship between the number of outstanding and self-sufficient itemsets. Not all self-sufficient itemsets are frequent under the minimum support thresholds we use, and the bottom-most row of Table 1 shows their number for the three different support thresholds.

Self-sufficient itemsets also cannot expected to be closed itemsets. We therefore identified for each self-sufficient itemset the corresponding closed frequent itemset, and compared this set to the set of outstanding itemsets selected. Table 2 shows for each support and each quality measure which proportion of outstanding itemsets are also self-sufficient (left-hand column per quality measure) and which proportion of self-sufficient itemsets are also outstanding (right-hand column). Missing lines correspond to settings where all values are 0.0, which includes in particular the *tic-tac-toe* data set.

Table 2. Self-sufficient and outstanding itemsets for different minimum supports θ and different quality measures. For each measure, left-hand column shows the proportion of outstanding itemsets that are self-sufficient, right-hand column displays the proportion of self-sufficient that are outstanding.

Data set	θ	χ^2		Confidence		NGR		WRAcc	
mushroom	10	0.16	0.48	0.25	0.14	0.25	0.14	0.29	0.48
mushroom	15	0.13	0.15	0.33	0.13	0.33	0.13	0.14	0.13
primary-tumor	10	0.02	0.08	0.00	0.00	0.00	0.00	0.00	0.00
primary-tumor	15	0.02	0.09	0.00	0.00	0.00	0.00	0.00	0.00
primary-tumor	20	0.03	0.12	0.00	0.00	0.00	0.00	0.00	0.00
soybean	10	0.03	0.02	0.00	0.00	0.04	0.02	0.12	0.02
soybean	15	0.12	0.06	0.00	0.00	0.17	0.03	0.00	0.00
soybean	20	0.18	0.07	0.00	0.00	0.20	0.03	0.00	0.00
splice-1	10	0.13	0.10	0.06	0.10	0.06	0.10	0.12	0.10
vote	10	0.18	0.10	0.00	0.00	0.00	0.00	0.00	0.00
vote	15	0.12	0.05	0.00	0.00	0.00	0.00	0.00	0.00
vote	20	0.10	0.03	0.00	0.00	0.00	0.00	0.00	0.00
zoo-1	10	0.10	0.02	0.33	0.05	0.33	0.05	0.00	0.00
zoo-1	15	0.10	0.02	0.50	0.05	0.50	0.05	0.00	0.00
zoo-1	20	0.12	0.02	0.33	0.06	0.33	0.06	0.00	0.00

Generally speaking, we can remark that outstanding itemsets stand not to be self-sufficient and vice versa. W.r.t. individual data sets, we can observe some interesting phenomena. For *mushroom* at $\theta = 10\%$ and χ^2/WRAcc, only one of the four outstanding itemsets is self-sufficient but ten of the self-sufficient itemsets are represented by it, i.e. they are subsets that cover the same transactions. Once we increase the minimum support to 20%, there is no itemset left that is both self-sufficient and outstanding. For the *vote* data set, there is a certain correspondence between outstanding and self-sufficient itemsets for χ^2 but none whatsoever for the other quality measures.

5.3 Structured pattern selection

This section gives an experimental illustration of our method on graph-structured data. This experiment is motivated by the study of chemical and biological data *BCR-ABL* from *ChEMBL23* [7]. In the data, every molecule is labeled as active or inactive; their structure represented as graphs. Negative data is denoted by $\mathcal{D}^-$ in the following. From the 1485 graphs of the data set, we extract closed frequent sub-graphs with at most 7 nodes, and $\theta = 10$.

As in the case of the closed itemsets we considered above, edges in the resulting DAG connect two vertices u and v if u is labeled with a maximal predecessor of the closed graph labeling v. As before, we assess the behavior of different quality measures: χ^2, confidence, NGR, WRAcc.

[7] https://www.ebi.ac.uk/chembl/

Table 3. Selection statistics on graph data.

Quality measure	χ^2	Confidence	NGR	WRAcc
Selected	247	30	32	257
Percentage	1.589%	0.193%	0.205%	1.653%
Total # patterns	15,544			

As we can see in Table 3, we select at most 1.7% of mined patterns. We also notice different behaviors for different quality measures: whereas NGR and confidence select small sets, WRAcc and χ^2 select more than six times as many, a number of patterns that could be hard to process by a human domain expert.

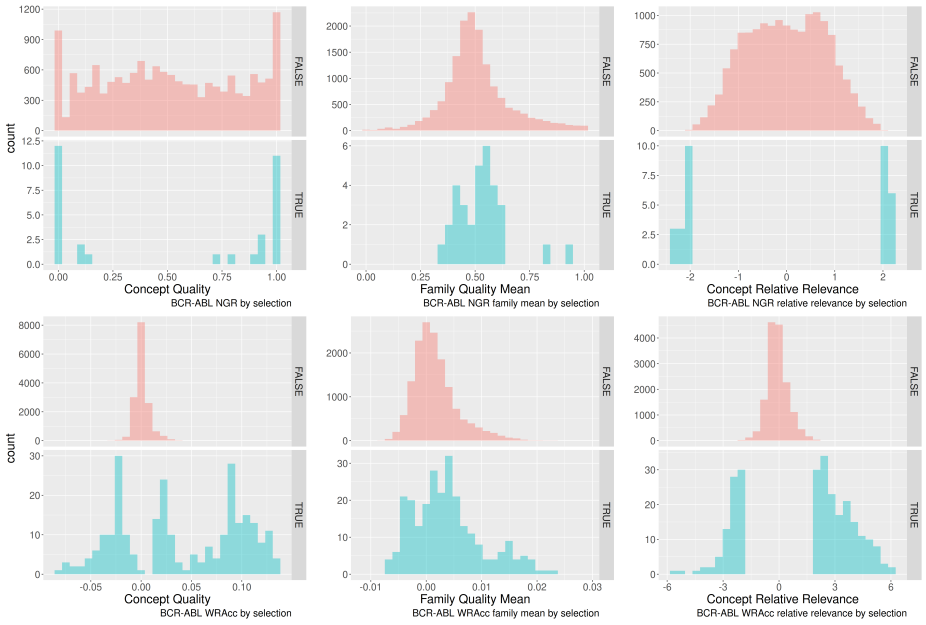

Fig. 5. Selection histograms for NGR (top) and WRAcc (bottom).

Figure 5, shows histograms of the scores for GR (top), and WRAcc (bottom). The left-most column shows the distribution of pattern scores, the center column the mean score for neighborhoods, and the right-most one the relative relevance, i.e. deviation of patterns from the mean of their neighborhood, normalized by the standard deviation. Blue histograms are for outstanding patterns, red for non-outstanding ones. As before, we see that outstanding patterns are not necessarily strongly correlated with the active class but might also be those that are unexpectedly weakly correlated (or even negatively correlated). We also see that

while the majority of outstanding patterns are two standard deviations off their neighborhood's mean score, there are patterns that deviate even more strongly.

5.4 Expert analysis upon an outstanding pattern and its family

The NGR used in the preceding section tends to discount jumping emerging patterns, which are however rather interesting in the context of activity analysis. In the following section we will therefore use another quality measure called GR_{max}, which avoids the ∞ problem but gives JEPs its due:

$$\begin{cases} GR_{max}(X,\mathcal{D}) = \quad |\mathcal{D}^-| \quad if\ GR(X,\mathcal{D}) = \infty \\ GR_{max}(X,\mathcal{D}) = GR(X,\mathcal{D}) \quad otherwise \end{cases}$$

We applied GR_{max} to the data-set BCR-ABL extracted from ChEMBL.

Table 4. Results on BCR-ABL using GR_{max}

	Order 1	*Order 2*	*Order 3*	*Order 4*	*Order 5*	*Order 6*	*Total*
Total	6	307	5 388	8 269	1 534	40	15 544
Selected	0	6	203	175	12	0	396
Percentage	0.00%	1.95%	3.77%	2.12%	0.78%	0.00%	2.55%

In Table 4, *order* indicates the number of nodes in the smallest free/generator sub-graph corresponding to closed graphs, allowing us to structure the graph into several layers. A closed graph together with its generator patterns induces an *equivalence class* of graph patterns covering the same data graphs. Each column correspond to a layer, numbered with its order. Rows *Total* indicates the number of equivalence classes in a layer, *Selected* the number of selected equivalence classes, and *Percentage* the percentage of selected equivalence classes. We observe that most of the outstanding patterns are found in the third and fourth layer. This is why the following analysis will be conducted on equivalence classes extracted from the third and fourth layers.

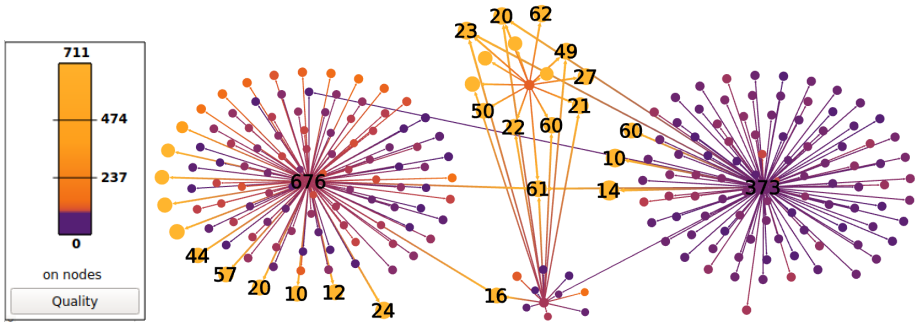

Fig. 6. Selected outstanding pattern (centered) along with its parents, and children of the parents. Labeled nodes are selected by OPS or important parent (373 and 676).

Starting from the outstanding pattern, an expert can expect to gain insight into structure-activity relationships (SAR) [10]. As an illustration, consider the center node of Figure 6. It shows an outstanding pattern appearing in 61 molecules, as well as its parents labelled as 676 and 373, colored according to their GR_{max} value with lighter colors corresponding to higher values. Larger nodes have a higher relative relevance. One of the parents, which differs only by one element syntactically, has significantly higher support values than the others. On both labelled parents, we see a large amount of children that include both patterns that do not correlate with the target class at all, and others that correlate very strongly. Furthermore we see that this family has several selected siblings (labelled nodes which are neither 676, nor 373). It implies that several subsets of molecules are outstanding regarding of each individual families for each pattern. It implies that the molecules' super-sets of our entry point contain cliffs [12] regarding the molecular activity. The outstanding pattern is therefore an entry point to visual analysis by the expert. Therefore, the outstanding pattern is an entry point for a visual analysis by the expert.

6 Conclusion

We have proposed a new way of selecting outstanding patterns by comparing them to neighboring patterns: a pattern is outstanding if it deviates clearly from the average of neighboring patterns w.r.t. the value of a quality measure. Our proposal is independent of the pattern language or the quality measure used. As experimentally shown, our selection patterns method leads to a strong reduction in the size of the result set, making the manual exploration by domain experts possible. Results differ significantly between different quality measures, i.e. the choice of quality measure becomes meaningful.

Finally, our selector puts an emphasis on the outstanding pattern's context. The selected pattern is interesting, but its parents, as well as its siblings, are also objects of interest. It can lead us to siblings linked to more than one outstanding pattern. Parents of such siblings become very interesting because outstanding pattern can have either positive or negative qualities, depending on the underlying data. Therefore, our selector offers a new way to study cleaving points inside the pattern language, and thus, the data, putting human in the loop.

Acknowledgements. This work was partially funded by the ANR project InvolvD (ANR-20-CE23-0023).

References

1. Besson, J., Rigotti, C., Mitasiunaite, I., Boulicaut, J.F.: Parameter tuning for differential mining of string patterns. In: ICDM Workshops. pp. 77–86. IEEE Computer Society (2008)
2. Bie, T.D.: Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Min. Knowl. Discov. **23**(3), 407–446 (2011)

3. Boulicaut, J.F., Bykowski, A., Rigotti, C.: Free-sets: a condensed representation of boolean data for the approximation of frequency queries. Data Mining and Knowledge Discovery **7**(1), 5–22 (2003)
4. Crémilleux, B., Giacometti, A., Soulet, A.: How your supporters and opponents define your interestingness. In: ECML PKDD. pp. 373–389. Springer (2018)
5. Dau, F., Ducrou, J., Eklund, P.: Concept similarity and related categories in searchsleuth. In: ICCS. pp. 255–268. Springer (2008)
6. Davey, B.A., Priestley, H.A.: Introduction to Lattices and Order, Second Edition. Cambridge University Press (2002)
7. De Raedt, L., Zimmermann, A.: Constraint-based pattern set mining. In: Proceedings of the Seventh SIAM International Conference on Data Mining. SIAM (2007)
8. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: Proceedings of the fifth ACM SIGKDD. pp. 43–52 (1999)
9. Kane, B., Cuissart, B., Cremilleux, B.: Minimal Jumping Emerging Patterns: Computation and Practical Assessment. In: PAKDD (2015)
10. Métivier, J.P., Cuissart, B., Bureau, R., Lepailleur, A.: The pharmacophore network: A computational method for exploring structure–activity relationships from a large chemical data set. Journal of Medicinal Chemistry **61**(8), 3551–3564 (2018)
11. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Discovering frequent closed itemsets for association rules. In: ICDT. pp. 398–416. Springer (1999)
12. Stumpfe, D., Hu, H., Bajorath, J.: Evolving concept of activity cliffs. ACS omega **4**(11), 14360–14368 (2019)
13. Tan, P., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Inf. Syst. **29**(4), 293–313 (2004)
14. Van Leeuwen, M., Ukkonen, A.: Fast estimation of the pattern frequency spectrum. In: ECML PKDD 2014. pp. 114–129 (2014)
15. Vreeken, J., Tatti, N.: Interesting patterns. In: Aggarwal, C.C., Han, J. (eds.) Frequent Pattern Mining, pp. 105–134. Springer (2014)
16. Webb, G.I.: Self-sufficient itemsets: An approach to screening potentially interesting associations between items. ACM Trans. Knowl. Discov. Data **4**(1), 3:1–3:20 (2010)
17. Yan, X., Han, J.: Closegraph: mining closed frequent graph patterns. In: Proceedings of the ninth ACM SIGKDD. pp. 286–295 (2003)