In Defense of Fuzzy Association Analysis

Eyke Hüllermeier and Yu Yi
Philipps-Universität Marburg
Department of Mathematics and Computer Science
Marburg, Germany
{eyke,yi}@informatik.uni-marburg.de

Draft of a paper that will appear in the IEEE Transactions on Systems, Man, and Cybernetics – Part B

Abstract

This short correspondence is a reply to the recently published paper "Fuzzy versus quantitative association rules: A fair data driven comparison" by H. Verlinde, M. De Cock, and R. Boute [10]. Even though we highly welcome the critical examination of the topic and definitely agree that fuzzy extensions of existing methods call for a thorough justification, the empirical comparison presented in the aforementioned paper is in our opinion not objective and extensive enough to fully warrant the conclusions drawn from the results. Apart from some general comments on the claims raised in the paper, we present empirical results based on an alternative experimental setup that lead to different conclusions.

1 Introduction

Association analysis is a widely applied data mining technique that has been studied intensively in recent years [4]. The goal in association analysis is to find "interesting" associations in a data set, that is, dependencies between so-called itemsets $\mathcal{A}$ and $\mathcal{B}$ expressed in terms of rules of the form $\mathcal{A} \rightharpoonup \mathcal{B}$. To illustrate, consider the well-known example where items are products and a data record (transaction) $\mathcal{I}$ is a shopping basket, such as $\{\texttt{butter}, \texttt{milk}, \texttt{bread}\}$. The intended meaning of an association $\mathcal{A} \rightharpoonup \mathcal{B}$ is that, if $\mathcal{A}$ is present in a transaction, then $\mathcal{B}$ is likely to be present as well. A standard problem in association analysis is to find all rules $\mathcal{A} \rightharpoonup \mathcal{B}$ the *support* (relative frequency of transactions $\mathcal{I}$ with $\mathcal{A} \cup \mathcal{B} \subseteq \mathcal{I}$) and *confidence* (relative frequency of transactions $\mathcal{I}$ with $\mathcal{B} \subseteq \mathcal{I}$ among those with $\mathcal{A} \subseteq \mathcal{I}$) of which reach user-defined thresholds `minsupp` and `minconf`, respectively.

In the above setting, a single item can be represented in terms of a binary (0/1-valued) attribute reflecting the presence or absence of the item. To make association analysis applicable to data sets involving numerical attributes, such attributes are typically discretized into intervals, and each interval is considered as a new binary attribute. For example, the attribute `temperature` might be replaced by two binary attributes `cold` and `warm`, where $\texttt{cold} = 1$ ($\texttt{warm} = 0$) if the temperature is below 10 degrees and $\texttt{warm} = 1$ ($\texttt{cold} = 0$) otherwise.

A further extension is to use fuzzy sets (fuzzy partitions) instead of intervals (interval partitions), and corresponding approaches to fuzzy association analysis have been proposed by several authors (see e.g. [2, 3] for recent overviews). There are different motivations for a fuzzy approach to association rule mining, notably the following: Firstly, by allowing for "soft" rather than crisp boundaries of intervals, fuzzy sets can avoid certain undesirable threshold or "boundary effects" [9]. Such effects are well-known, for instance, from *histograms* in statistics: A slight variation of the boundary points of the intervals can have a considerable effect on the histogram induced by a number of observations (it may even lead to qualitative changes, i.e., changes of the shape of the histogram) [8]. Likewise, the variation of a partition can strongly influence the evaluation of association rules [7]. Secondly, fuzzy association rules are very appealing from a knowledge representational point of view: The very idea of fuzzy sets is to act as an interface between a numerical scale and a symbolic scale which is usually composed of linguistic terms. Thus, the rules discovered in a database might be presented in a linguistic and hence comprehensible and user-friendly way.

2 General Comments

In their paper [10], Verlinde et al. raise the question whether or not a fuzzy extension of association analysis is actually useful. To this end, they compare the results produced by fuzzy and non-fuzzy rule mining for three different data sets. Since the results obtained appear to be quite similar, they conclude that "in real applications the net difference is very likely to be too small to really justify the fuzzy approach".

Restrictively, they add that their comparison is a purely *data-driven* one which ignores semantical issues and aspects like linguistic interpretability which, as mentioned above, might be cited as arguments in favor of a fuzzy approach.

We highly welcome a critical examination and discussion as initiated by Verlinde et al. In fact, it seems that, not only in data mining but also in many other fields, the "fuzzification" of existing methods is sometimes regarded as an end in itself, without critically investigating the need for an extension and, hence, complication of that kind. In the particular case of association rule mining, however, we believe that a fuzzy extension is legitimate. Moreover, despite the importance of the authors' contribution, we think that their experimental investigation is not extensive enough to warrant the conclusions drawn from the results.

A first critical remark concerns the decision to employ (fuzzy c-means) clustering for the purpose of discretizing numerical data. Apart from the fact that many alternative partitioning methods are conceivable (see e.g. [6]), one may wonder whether clustering is really advisable in this context. Verlinde et al. correctly point out that in many papers, artificial examples are constructed in such a way as to enforce the aforementioned boundary effect. On the other hand, by partitioning the data using a clustering approach, a possible boundary effect is automatically suppressed, since the data regions of high density will usually be located in the middle of an interval (a quite obvious observation, the authors seem to be well aware of). In our opinion, the second approach is hence hardly less biased than the first one. We shall come back to this point in section 3.3.

A second point concerns the significance of the experimental results. In their experiments, Verlinde et al. restrict themselves to the most simple type of association, namely rules with a single antecedent and a single consequent. Even though we agree that overly complex rules will hardly be understandable by a user, a restriction to antecedents of length 1 appears to be too restrictive. In our opinion, rules with antecedents of length 2 or even 3, such as $\{\texttt{bread}, \texttt{cheese}, \texttt{butter}\} \rightharpoonup \{\texttt{wine}\}$, are still understandable.

To quantify the difference between fuzzy and non-fuzzy association analysis, Verlinde et al. first order the *complete* set of potential association rules according to a quality measure (either support or confidence); considering the complete set of rules is possible because, by restricting the analysis to rules involving only two items and data sets with at most 6 attributes, the overall number of candidate rules is quite limited. Then, they compare the two rankings in terms of the Spearman rank correlation. For reasons to be detailed in section 3.4, we do not find this measure fully suitable in this context.

A related point concerns the small number of attributes in the data sets used for experimentation: For every data set, Verlinde et al. selected a subset of 5 or 6 attributes, which is a relatively small number in light of the existence of data sets involving up to hundreds of attributes. In fact, one should note that the number of potential rules critically depends on the number of attributes. Furthermore, one might speculate that, the higher the number of potential rules, the more strongly the rank of a rule might be affected by a small variation of the quality measure and, hence, the smaller the similarity between fuzzy and non-fuzzy mining might become.

3 Experimental Setup

In this section, we describe our experimental setup, in particular the differences between our setting and the one used in [10].

3.1 Data Sets

To make the results comparable, we used the same three data sets as in [10]: FAM95 (63,756 data records, 23 attributes),[1] HEMAT (42,915 records, 11 attributes),[2] and the ENTRY data from the STULONG project (1,417 records, 64 attributes).[3] Moreover, we reduced these data sets to the same subsets of (numerical) attributes (FAM95: number of persons in the family, number of children, family income, age of head of the family, educational level of head, and head's personal income; HEMAT: white blood cells count, hemoglobin count, hematocrit count, mean corpuscular volume, mean corpuscular hemoglobin, and mean corpuscular hemoglobin concentration; ENTRY: height, weight, systolic blood pressure, diastolic blood pressure, and cholesterol level), except for the experiments in section 4.4. See [10] for a more detailed description of the data sets.

[1] http://www.stat.ucla.edu/data/fpp

[2] http://lisp.vse.cz/challenge/ecmlpkdd2004

[3] http://euromise.vse.cz/challenge2004

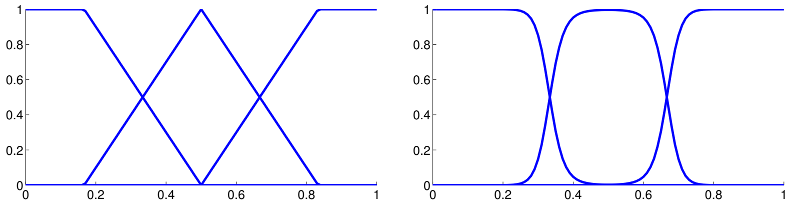

Figure 1: Fuzzification of an equi-width partition using piecewise linear (left) and Gaussian-like (right) membership functions.

3.2 Fuzzy Logical Operators

In fuzzy association analysis (see [4] for a systematic exposition), generalized operators have to be defined for set intersection and cardinality. Again, to assure comparability, we used the same operators as in [10], namely the min-operator as a generalized conjunction and the sum of membership degrees ("sigma-count") for set cardinality. In fact, these operators are frequently used by other authors as well.

3.3 Discretization

In [10], Verlinde et al. use fuzzy c-means clustering [1] with $c = 3$ (and a fuzzifier $m = 3$) in order to obtain a fuzzy partition for each attribute. That is, every cluster defines a (discretized) fuzzy attribute value. Moreover, a crisp partition is derived from the fuzzy one by assigning every data point to the nearest center of the fuzzy clusters.

As mentioned before, one might call the choice of a clustering method for the purpose of discretization into question, firstly because it is biased toward the "no effect by fuzzification" hypothesis, and secondly because it is not even the most simple approach. In fact, the arguably most simple (non-fuzzy) discretization methods are the well-known equi-width and equi-frequency partitioning. Correspondingly, since the simplest type of (one-dimensional) fuzzy set is a triangular membership function, one can argue that the simplest fuzzy version of these discretization methods is to replace the respective intervals by (overlapping) triangles or, in other words, crisp boundaries by soft boundaries with linear transition. More specifically, if a, b, c denote, respectively, the midpoints of the three intervals obtained by equi-width (equi-frequency) partitioning, our fuzzy partition consists of three fuzzy sets with cores $(-\infty, a]$, $\{b\}$, $[c, +\infty)$, and supports

$(-\infty, b)$, (a, c), $(b, +\infty)$, respectively, and a linear transition between core and support (that is, a left trapezoid, a triangle in the middle, and a right trapezoid, see Fig. 1).

3.4 Similarity Measures

In order to quantify the similarity between two lists (rankings) of association rules, A and B, Verlinde et al. employ the well-known Spearman rank correlation measure, which is a linear transformation of the sum of squared rank distances to the interval $[-1, +1]$. The latter is in turn given by

$$\sum_{\imath=1}^{N} (r_A(\imath) - r_B(\imath))^2,$$

where N is the total number of rules, $r_A(\imath)$ denotes the rank of the $\imath$-th rule in the list A and, likewise, $r_B(\imath)$ the corresponding position in list B.

We do not find the above measure very suitable in this context, mainly because it gives the same weight to every position. Thus, erroneously ranking the top-rule on position 51 is as bad as ranking a rule on position 111 instead of 161 (the rank distance is 50 in both cases). This is not very reasonable in association analysis, where the higher ranks are definitely more important than the lower ones; in fact, a user will typically only look at the top rules and ignore the rest. For a similar reason, measuring the similarity between the *complete* sets of association rules can be criticized. As Verlinde et al. only consider associations with an antecedent part of length 1, the number N of potential rules is quite limited (180, 270, and 270 for the three data sets). In real applications, however, where this restriction is not made, the number of potential rules will be huge.

There are two possibilities to put more weight on higher ranked rules: Either one restricts the lists A and B to the top-K rules, for a suitable K, or one weighs the positions, putting higher weight on higher positions (strictly speaking, the former can be seen as a special case of the latter). The question of how to define a distance between top-K lists in a reasonable way has recently been addressed in [5]. Among the measures proposed in this paper, we selected a modified version of the Spearman footrule, which fulfills the properties of a metric. Spearman's footrule is similar to his rank correlation, except that the squared rank differences $(r_A(\imath) - r_B(\imath))^2$ are replaced by the absolute differences $\|r_A(\imath) - r_B(\imath)\|$. The idea of the modification, which is suitable for measuring the distance between top-K lists, is to "lift" elements (rules) with a low rank to a fixed position ℓ, which effectively means reducing their distance to the top-elements and even ignoring their distance to other elements with a low rank. More specifically, consider a top-K list $r_A(\cdot)$ which is a ranking of a set $\mathcal{A}$ of elements (rules). Likewise, let $r_B(\cdot)$ be a ranking of a set $\mathcal{B}$. To compare $r_A(\cdot)$ and $r_B(\cdot)$, these lists are first transformed into extended lists $r'_A(\cdot)$ and $r'_B(\cdot)$, which are both rankings of $\mathcal{A} \cup \mathcal{B}$. For every $\imath \in \mathcal{A} \cup \mathcal{B}$, $r'_A(\imath) = r_A(\imath)$ if $\imath \in \mathcal{A}$ and $r'_A(\imath) = \ell$ otherwise; $r'_B(\cdot)$ is defined analogously. The distance between $r_A(\cdot)$ and $r_B(\cdot)$ is then given by the footrule distance between the respective transformations:

$$\sum_{\imath \in \mathcal{A} \cup \mathcal{B}} \| r'_A(\imath) - r'_B(\imath) \| . \tag{1}$$

The constant ℓ must of course fulfill $\ell > K$; a natural choice, which we used in our experiments, is $\ell = K + 1$. Furthermore, to guarantee comparability of results from different experiments, we normalized the measure (1) to the range $[0, 1]$: Two identical top-K lists $r_A(\cdot)$ and $r_B(\cdot)$ have distance 0, while the distance is 1 in case they do not share a single element ($\mathcal{A} \cap \mathcal{B} = \emptyset$). The measure thus obtained was used in the experiments in sections 4.2 and 4.4, where the large number of potential rules prevents from comparing complete rule lists. In the experiments in section 4.1, where comparing complete lists was still possible, we also employed the standard footrule, again normalized to the range $[0, 1]$.

4 Results

4.1 The Effect of Partitioning

In a first study, we exactly replicated the experiments in [10], except for using alternative (additional) partitioning methods and trying partition sizes of $c = 3$ and $c = 4$ instead of only $c = 3$. More specifically, we employed the following methods:

- Fuzzy c-means and the related crisp (k-means) partition, using a fuzzifier of $m = 2$ (FCM2, KM2) and $m = 3$ (FCM3, KM3). The latter methods (FCM3, KM3) are those used in [10].

- Equi-width partitioning (EW) and its fuzzy variant (FEW) as outlined in section 3.3.

- Equi-frequency partitioning (EF) and its fuzzy variant (FEF) as outlined in section 3.3.

The results are summarized in tables 1–2. Each entry in a table corresponds to the distance between the rule lists obtained by the non-fuzzy partitioning method associated with the respective column and its fuzzy variant.

A first interesting finding is that the distance between FCM2 and KM2 is always smaller than the distance between FCM3 and KM3. This can be taken as a first indication that fuzzification does indeed have an influence on the data mining results: The higher the fuzzifier m, the more fuzzy the partitions become, the greater the difference between fuzzy and non-fuzzy rule mining.

Another important observation concerns the effect of the partitioning method: The distances for the fuzzy and the non-fuzzy version of equi-width partitioning (EW) are still comparable to those for clustering (KM3), and often the former are even slightly smaller than the latter. As opposed to this, the distances become significantly larger for equi-frequency partitioning (EF). In fact, in the latter case, the distance values indicate that the rankings obtained, respectively, for fuzzy and non-fuzzy mining are far from similar, showing that the partitioning method definitely has an influence on the effect of fuzzification.

To give a possible explanation for these results, recall our hypothesis that the effect of fuzzification strongly depends on the density of the data around the interval boundaries. This hypothesis makes is quite plausible to obtain the highest distances for EF: While EW determines the interval boundaries without looking at the data and, hence, still has a good chance to hit a region of low density, EF will very likely place the boundaries in highly populated regions (the probability of passing a frequency threshold in a certain region increases with the density in that region).

The fact that the distances for EW are on average even slightly smaller than for KM3 may appear unexpected at first sight, since, as we already mentioned earlier, a clustering approach tends to put the interval centers in data regions of high density. The critical point to observe, however, is that this does not automatically imply that the interval boundaries are located in regions of low density. Actually, this effect only occurs if the distribution of the data is multi-modal (and the number of modes is not smaller than the number of clusters). In the case of unimodal data, however, all interval centers, and consequently the transitions between them, will be located in high density regions;

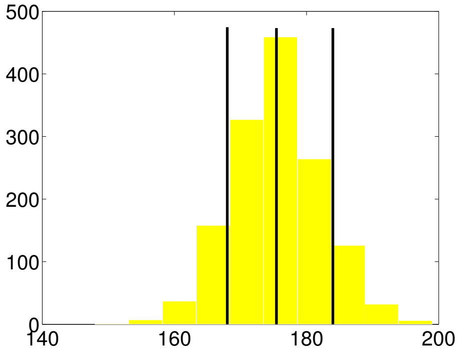

Figure 2: Distribution of the first attribute (height) in the ENTRY data and cluster centers (vertical lines) obtained by fuzzy c-means clustering for $c = 3$.

Fig. 2 shows a typical example for the first attribute of the ENTRY data. In such cases, there is obviously a high chance to have a lower density around the boundaries in EW.

Apart from that, however, it turned out that the results are also strongly influenced by other factors. For example, we derived the same distance measures for two variants of equi-width partitioning: In the first variant, called *random width* (RW), the interval boundaries are chosen at random (and results are averaged over 100 experiments). In the second variant, EWG, we used Gaussian-like membership functions instead of trapezoidal ones in order to fuzzify the equi-with partition; more specifically, these membership functions are of the form $x \mapsto G_i(x)/(G_1(x) + G_2(x) + G_3(x))$, $i = 1, 2, 3$, where the $G_i(\cdot)$ are Gaussian membership functions chosen such that the transition regions coincide with those in the case of EW (see Fig. 1). As can be seen in table 1, both variants lead to significantly higher distance values.

Regarding the size of the partition, the results for $c = 3$ and $c = 4$ are quite comparable and do not provide evidence for a possible influence of this factor.

4.2 The Effect of Rule Length

In a second experiment we investigated the effect of extending the length of association rules. More specifically, we considered rules of length $l = 3$ (two items in the antecedent part, one in the consequent part) and $l = 4$ (three items in the antecedent part). Since the total number of such rules now becomes exorbitantly large, we only mined rules exceeding a minimum support threshold of 0.1. Then, we sorted the rules according to confidence, considered the top-K rules (for $K = 50, 100$) and referred to (1) for measuring the distance between the different partitioning methods.

Even though there are a few exceptions, the results in tables 3–8 clearly show that by increasing the rule length, the distance between the rankings of rules becomes significantly larger. In fact, the rankings are quite dissimilar on average. Just to give an extreme example, the top-50 rules produced by FCM3 and KM3 on the ENTRY data do not have a single rule of length 4 in common. Besides, as already observed in the previous experiment, FCM2 and KM2 are always higher correlated than FCM3 and KM3.

4.3 The Effect of Focusing on Top-Rules

As mentioned earlier, considering a complete list of potential association rules does not appear very reasonable. When focusing on the top-K rules, the intuitive expectation is that the average distance between two mining results (fuzzy and non-fuzzy) will be higher for smaller K (imagine the extreme case $K = 1$, where the distance will almost always be 1). This expectation is confirmed by our results. In fact, the effect is quite obvious when passing from the complete lists to the top-100 lists. And even though the picture is not as clear when comparing the cases $K = 100$ and $K = 50$, the distance is larger for $K = 50$ most of the time.

4.4 The Effect of the Number of Attributes

In a final experiment, we investigated the influence of the number of attributes involved in the data set. To this end, we included all of the 11 numerical attributes of the FAM95 data instead of selecting only 6 of them as done by Verlinde et al. Likewise, we included all of the 8 (instead of only 5) numerical attributes in ENTRY and HEMAT.

The results are shown in tables 9–12. Comparing these results with the previous ones, no clear tendency can be spotted, and more often than not, the distance values are not changed very much. Thus, in contrast to our expectation, the number of attributes does not seem to be a determining factor for the similarity between fuzzy and non-fuzzy mining. Restrictively, however, one has to consider that even the extended attribute sets are still quite small for the three data sets in this study.

5 Conclusions

Using a quite particular experimental setting, the authors in [10] found that fuzzy and non-fuzzy rule mining produce very similar results. Even though these results are not incorrect, they have to be interpreted with much caution. In fact, our findings in this paper show that, by using alternative partitioning methods, considering more complex association rules, and focusing on the top-rules, the similarity between fuzzy and non-fuzzy rule mining becomes much smaller and in some cases completely disappears. Even though we do not regard these findings as a proof of the usefulness of fuzzy association analysis, we consider them as an invalidation of the opposite claim raised in [10].

Acknowledgements: The authors are grateful for the helpful comments and suggestions by three anonymous reviewers.

References

[1] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithm.* Plenum Press, New York, 1981.

[2] G. Chen, Q. Wei, E. Kerre, and G. Wets. Overview of fuzzy associations mining. In *Proc. ISIS–2003, 4th International Symposium on Advanced Intelligent Systems.* Jeju, Korea, September 2003.

[3] M. Delgado, N. Marin, D. Sanchez, and MA. Vila. Fuzzy association rules: general model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2):214–225, 2003.

[4] D. Dubois, E. Hüllermeier, and H. Prade. A systematic approach to the assessment of fuzzy association rules. *Data Mining and Knowledge Discovery*, 13(2):167.

[5] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM Journal of Discrete Mathematics*, 17(1):134–160, 2003.

[6] M. Kaya and R. Alhajj. Genetic algorithm based framework for mining fuzzy association rules. *Fuzzy Sets and Systems*, 152(3):587–601, 2005.

[7] C. Man Kuok, A. Fu, and M. Hon Wong. Mining fuzzy association rules in databases. *SIGMOD Record*, 27:41–46, 1998.

[8] O. Strauss, F. Comby, and MJ. Aldon. Rough histograms for robust statistics. In *ICPR–2000, 15th International Conference on Pattern Recognition*, pages 2684–2687. Barcelona, September 2000.

[9] T. Sudkamp. Examples, counterexamples, and measuring fuzzy associations. *Fuzzy Sets and Systems*, 149(1), 2005.

[10] H. Verlinde, M. De Cock, and R. Boute. Fuzzy versus quantitative association rules: A fair data driven comparison. *IEEE Transactions on Systems, Man and Cybernetics*, 36(3):679–684, 2006.

A Tables

	sorted by confidence					
	KM2	KM3	EF	EW	RW	EWG
ENTRY	0.10	0.21	0.63	0.12	0.30	0.24
FAM95	0.15	0.26	0.48	0.24	0.28	0.30
HEMAT	0.15	0.28	0.60	0.25	0.35	0.34
	sorted by support					
ENTRY	0.07	0.15	0.65	0.06	0.27	0.14
FAM95	0.14	0.21	0.47	0.17	0.21	0.26
HEMAT	0.07	0.12	0.66	0.21	0.30	0.31

Table 1: Distance matrix for partition size 3, rule length 2.

	sorted by confidence				sorted by support			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.15	0.29	0.53	0.12	0.13	0.21	0.52	0.08
FAM95	0.1	0.29	0.36	0.17	0.08	0.19	0.38	0.12
HEMAT	0.21	0.33	0.5	0.33	0.08	0.12	0.58	0.27

Table 2: Distance matrix for partition size 4, rule length 2.

	sorted by confidence				sorted by support			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.15	0.26	0.39	0.2	0.1	0.21	0.51	0.11
FAM95	0.28	0.37	0.34	0.33	0.16	0.23	0.39	0.26
HEMAT	0.31	0.4	0.46	0.37	0.15	0.19	0.56	0.32

Table 3: Distance matrix for partition size 3, rule length 3.

	sorted by confidence				sorted by support			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.19	0.29	0.36	0.28	0.15	0.28	0.41	0.19
FAM95	0.26	0.39	0.23	0.24	0.13	0.28	0.29	0.22
HEMAT	0.42	0.44	0.35	0.41	0.19	0.25	0.46	0.38

Table 4: Distance matrix for partition size 4, rule length 3.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.33	0.64	0.98	0.55	0.33	0.74	0.92	0.36
FAM95	0.76	0.49	0.67	0.27	0.43	0.48	0.78	0.17
HEMAT	0.08	0.31	0.93	0.3	0.27	0.37	0.93	0.3

Table 5: Distance matrix for partition size 3, rule length 3.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.57	0.99	0.95	0.54	0.6	0.98	0.87	0.56
FAM95	0.25	0.99	0.41	0.07	0.21	0.93	0.52	0.27
HEMAT	0.81	0.72	0.93	0.49	0.2	0.57	0.82	0.17

Table 6: Distance matrix for partition size 4, rule length 3.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.8	0.51	1	0.7	0.71	0.61	0.99	0.65
FAM95	0.68	0.75	0.85	0.42	0.46	0.63	0.82	0.24
HEMAT	0.65	0.28	0.92	0.76	0.38	0.31	1	0.34

Table 7: Distance matrix for partition size 3, rule length 4.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.78	1	0.99	0.56	0.79	1	0.99	0.56
FAM95	0.54	0.57	0.77	0.12	0.45	1	0.34	0.07
HEMAT	0.43	0.91	1	0.3	0.5	0.82	1	0.51

Table 8: Distance matrix for partition size 4, rule length 4.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.35	0.78	0.98	0.42	0.53	0.65	0.94	0.59
FAM95	0.4	0.41	0.79	0.35	0.6	0.31	0.78	0.52
HEMAT	0.11	0.33	0.96	0.31	0.1	0.26	0.92	0.11

Table 9: Distance matrix for partition size 3, rule length 3.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.5	0.68	0.98	0.37	0.46	1	1	0.36
FAM95	0.17	0.47	0.83	0.53	0.49	0.25	0.88	0.6
HEMAT	0.19	0.75	1	1	0.17	0.41	0.98	0.34

Table 10: Distance matrix for partition size 3, rule length 4.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.69	1	0.87	0.33	0.59	1	0.84	0.76
FAM95	0.09	0.18	0.72	0.27	0.22	0.37	0.54	0.32
HEMAT	0.21	0.53	1	0.35	0.8	0.82	0.97	0.52

Table 11: Distance matrix for partition size 4, rule length 3.

	top-50 ranking				top-100 ranking			
	KM2	KM3	EF	EW	KM2	KM3	EF	EW
ENTRY	0.53	0.98	1	0.44	0.63	0.98	0.92	0.37
FAM95	0.64	0.33	0.8	0.46	0.1	0.23	0.68	0.52
HEMAT	0.44	0.84	1	0.36	0.4	0.93	1	0.31

Table 12: Distance matrix for partition size 4, rule length 4.