Selection of Relevant and Non-Redundant Multivariate Ordinal Patterns for Time Series Classification - Supplementary Material

Arvind Kumar Shekar Marcus Pappik Patricia Iglesias Sanchez
Emmanuel Mueller

1 Experiments

1.1 Comparison of other information theoretic relevance measures In the work of *ordex*, we introduced a novel scoring method for ordinal patterns. We compare bivariate mutual information and multivariate KL-divergence based relevance measure [1] against our proposed scoring. We evaluate the run times and the test data accuracy on 18 UCR and 2 UCI datasets.

We observe that our relevance function, based on the Chebyshev-Inequality, performs better in comparison to the KL-divergence in the context of our algorithm. For 70% of the data sets, using our relevance measure yields a better accuracy. On average, the achieved accuracy was 2.77% better than with KLD. In all cases, our relevance measure outperforms the KLD in terms of run time. Even if both run times seem to grow quite similar with respect to the data set properties, the KLD needed 4.77% more time on average.

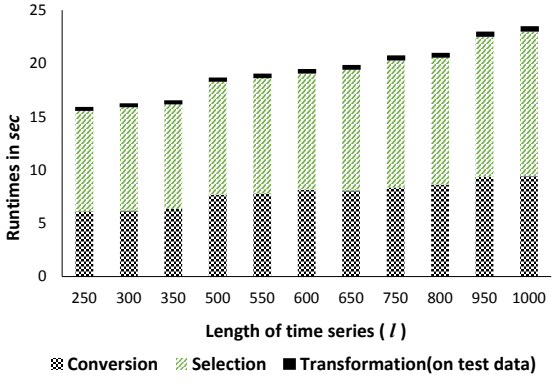

Figure 1: Scalability w.r.t. increasing l, where m=5, n=100, d=5, o=10 and I=200

By comparing our Chebyshev-based relevance function with Mutual Information, our relevance measure outperformed MI on 75% of the data sets in terms of accuracy and on 80% of the data sets in terms of standard deviation. Also our scoring outperformed on all data sets w.r.t. the run time. Using MI, the average run time was 18.09% higher for the selection phase.

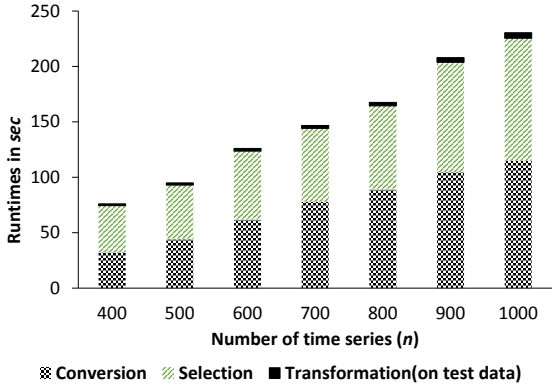

Figure 2: Scalability w.r.t increasing n, where m=5, l=600, d=5, o=10 and I=200

1.2 Scalability Experiments Figure 1 and Figure 2 show the scalability of *ordex* with increasing length of time series and number of samples in a dataset.

2 Upper bound of mis-classfication

Consider f as a feature extracted using a multivariate ordinal pattern set s to classify between c_a and c_b. We denote its distribution for class c_a as $f|c_a$. The expected value and the variance of the distribution are represented as $E[f|c_a]$ and $Var[f|c_a]$ respectively. Similarly, for class c_b, we define the distribution $f|c_b$, expected value $E[f|c_b]$ and variance $Var[f|c_b]$. Without loss of generality, we assume $E[f|c_a] < E[f|c_b]$. The upper bound of mis-classification for feature f with arbitrary distribution is strongly founded by the principles of Chebyshev-inequality,

Chebychev's inequality defines the upper bound of the fraction of samples that can lie beyond a threshold $a > 0$. For any feature f, no more than $1/a^2$ of the values can be more than a standard deviations away from the mean [2],

$$P(|f - E[f]| \geq a) \leq \frac{Var[f]}{a^2}, \tag{2.1}$$

where $a > 0$.

Given that the expected value $E[f]$ and variance

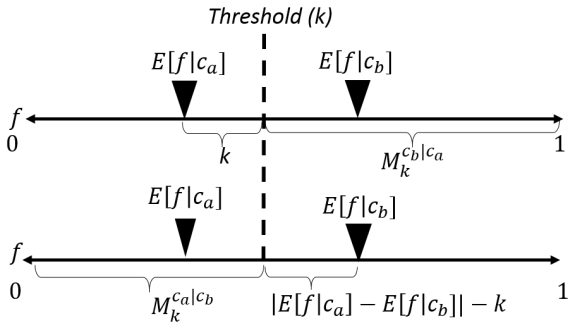

Figure 3: Example: A number line with limits [0,1]

$Var[f]$ of the feature, the inequality represents the probability of a sample that is greater than a. The approach is commonly used for finding outliers, i.e, instances with a high probability of being greater than $E[f] + a$ are outliers.

Applying the rule of Chebychev's inequality [2] for classification problems, an instance of feature f is classified as c_a or c_b based on the arbitrary threshold value $k \mid 0 < k < |E[f|c_b] - E[f|c_a]|$ (c.f. Figure 3). We denote $P(M_k^{c_a,c_b})$ as the probability that c_a is mis-classified as c_b or c_b is misclassified as c_a. Under the assumption that $f|c_a$ and $f|c_b$ are symmetrically distributed around their expected values, a sample is classified as c_b when its expected value is greater than $E[f|c_a] + k$. Hence, to estimate $P(M_k^{c_a,c_b})$ we need to quantify the maximum number of c_a that exceed the threshold and likewise for c_b. The upper bound of mis-classification is represented as,

$$P(M_k^{c_a,c_b}) \leq \frac{Var[f|c_a]}{2k^2} + \frac{Var[f|c_b]}{2(|E[f|c_b] - E[f|c_a]| - k)^2}.$$

Proof: For a classification task, using our threshold $E[f|c_a] + k$, we aim to estimate:
$M_k^{c_b|c_a}$: a sample with class c_a is misclassified as c_b based on this threshold.
$M_k^{c_a|c_b}$: a sample with class c_b is misclassified as c_a based on this threshold.
Figure 3 visualizes both cases on a simple number-line, by applying the Chebychev inequality,

$$\begin{aligned} P(M_k^{c_b|c_a}) &= P((f|c_a) \geq E[f|c_a] + k) \\ &= P((f|c_a) - E[f|c_a] \geq k). \end{aligned}$$

We assumed that $f|c_a$ is distributed symmetrically around its expected value. Thus $P(M_k^{c_b|c_a})$ is half the probability that the value of $f|c_a$ has at least a distance of k to its expected value,

$$P(M_k^{c_b|c_a}) = \frac{1}{2}P(|(f|c_a) - E[f|c_a]| \geq k).$$

Using the Chebyshev-Inequality and setting $a = k$ in Eq 2.1, we can estimate an upper bound of $P(M_k^{c_b|c_a})$ as,

$$P(M_k^{c_b|c_a}) \leq \frac{Var[f|c_a]}{2k^2}. \tag{2.2}$$

On applying the same symmetric assumption on $(f|c_b)$, $M_k^{c_a|c_b}$ is half of the probability that values of $(f|c_b)$ are at least $|E[f|c_b] - E[f|c_a]| - k$ away from their expected value

$$\begin{aligned} &P(M_k^{c_a|c_b}) = P((f|c_b) \leq E[f|c_a] + k) \\ &= P((f|c_b) - E[f|c_b] \leq E[f|c_a] - E[f|c_b] + k) \\ &= P(E[f|c_b] - (f|c_b) \geq E[f|c_b] - E[f|c_a] - k) \\ &= \frac{1}{2}P(|(f|c_b) - E[f|c_b]| \geq |E[f|c_b] - E[f|c_a]| - k) \end{aligned}$$

Comparing the above result with Eq 2.1, we derive, $a = |E[f|c_b] - E[f|c_a]| - k$. To estimate an upper bound

$$P(M_k^{c_a|c_b}) \leq \frac{Var[f|c_b]}{2(|E[f|c_b] - E[f|c_a]| - k)^2} \tag{2.3}$$

From Eq 2.2 and 2.3 this, we derive an upper bound for the total mis-classification probability for c_a and c_b as,

$$\begin{aligned} P(M_k^{c_a,c_b}) &= P(M_k^{c_b|c_a} \cup M_k^{c_a|c_b}) \\ &\leq P(M_k^{c_b|c_a}) + P(M_k^{c_a|c_b}) \\ &\leq \frac{Var[f|c_a]}{2k^2} + \frac{Var[f|c_b]}{2(|E[f|c_b] - E[f|c_a]| - k)^2}. \end{aligned}$$

This means, given an optimal value of the threshold $E[f|c_a] + k$, we can calculate an upper bound for the minimal misclassification of each pair of classes. However, we cannot assume that all classifiers find such an optimal k based on the data. Moreover, finding this bound costs additional computation time. In order to be independent of the classifier and have a better efficiency, we use the fact, that $0 < k < |E[f|c_b] - E[f|c_a]|$. Due to that, the upper bound of the mis-classification grows approximately as fast as

$$\frac{Var[f|c_a] + Var[f|c_b]}{2(|E[f|c_b] - E[f|c_a]|)^2}.$$

3 Parameters for real world dataset

Table 1: Real world data experiment parameter settings

Dataset	d	I	m'	α	o
EMG limb sen	5	200	3	0.5	10
EMG limb pie	5	300	3	0.1	15
EMG limb mar	5	200	3	0.3	20
Character	5	300	3	0.1	50
Activity recognition	5	100	2	0.3	20
User Movement	5	300	3	0.8	30
Occupancy	5	100	3	0.5	10
Bosch	5	200	3	0.1	10

4 Accuracy on real world datasets with different penalization technique

In this section we show that the results of *ordex* is not largely different for different penalization techniques. In the original paper we use the penalization technique where redundancy is subtracted from relevance. Here, we show that the use of harmonic mean between relevance and redundancy also have similar accuracy on real world datasets.

$$score(f) = \left[\frac{2 \cdot rel(f) \cdot (1 - red(f)}{rel(f) + (1 - red(f))}\right] \tag{4.4}$$

Table 2: Real world data experiment parameter settings

Dataset	Accuracy in %	Standard deviation in ± %
EMG limb sen	90	6.8
EMG limb pie	93.33	8.1
EMG limb mar	93.3	8.1
Character	72.3	1.5
Activity recognition	100	0
User Movement	56.98	3
Occupancy	92.1	0
Bosch	95	7

4.1 Time Complexity For an n sample, m-dimensional dataset of length l, we calculate $l-(d-1)$ ordinalities for each univariate time series. As computing each ordinality involves sorting the time series values of degree d, the total complexity for the conversion of a time series dataset into its ordinal representation is $\mathcal{O}(n \cdot m \cdot l \cdot d \cdot log(d))$.

The run time of the algorithm depends on the number of iterations I. In addition, extraction of s (c.f. Algorithm 1, line 3) depends on the maximum number of dimensions m' and maximum number of ordinalities in each dimension. Thus the complexity is represented as $\mathcal{O}(I \cdot m' \cdot d!)$.

The transformation of an extracted pattern into its numeric feature involves evaluation of its probability in each time series sample with a maximum of m' dimensions. As the relevance scoring is done for each pair of classes, the complexity of our scoring function for a classification problem with k classes is represented as $\mathcal{O}(n \cdot m' \cdot l + k^2)$.

For a set of o selected features, the complexity for computing the redundancy between two features using Spearman's correlation is represented as $\mathcal{O}(n \cdot log(n))$. However, as we compute the redundancy of all feature pairs, the time complexity for redundancy estimation is $\mathcal{O}(o^2 \cdot n \cdot log(n))$.

References

[1] A. K. Shekar, T. Bocklisch, C. N. Straehle, P. I. Sánchez, and E. Mller, "Including multi-feature interactions and redundancy for feature ranking in mixed datasets," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Macedonia, Skopje, September 18-22, 2017, Proceedings*, ser. Lecture Notes in Computer Science. Springer, 2017.

[2] S. Karlin and W. J. Studden, *Tchebycheff systems: With applications in analysis and statistics*. Interscience New York, 1966.