Determination of regional variants in the versification of Estonian folksongs using an interpretable fuzzy rule-based classifier

Andri Riid[1] Mari Sarv[2]

[1]Laboratory of Proactive Technologies, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia
[2]Estonian Folklore Archives, Estonian Literary Museum, Vanemuise 42, Tartu 51003, Estonia

Abstract

In this paper, a method of hierarchical clustering and a selection of fuzzy classification algorithms are applied successively to the data set that contains measured characteristics of folk verses collected from 104 historical parishes of Estonia. The aim of the study is to detect the groups of parishes that are similar in terms of folk verse characteristics and to give us insight into the reasoning that the separation into these groups is based upon. The process of classification separates the initial groups into further subsets represented by fuzzy rules, which can be analyzed thanks to the interpretability of such rules. To emphasize the latter, most important features in individual rules are brought out by rule compression. The results of the analysis are backed by what is known from linguistic sciences.

Keywords: Fuzzy systems, classification, Estonian folksongs

1. Introduction

Over a decade, interpretability of fuzzy systems has been a topic of keen scientific interest [1, 2, 3]. However, most of this debate has focused on interpretability definitions, measures and novel algorithms that improve interpretability of fuzzy systems in a way or another. Only rarely, interpretability itself has been exploited for a specific practical purpose.

This paper, in which an interpretable fuzzy classifier is employed to explain the geographical variation of the metre of Estonian folksongs, however, provides one such example.

The data set on which the classifier is based, originates from an earlier study [4] where the occurrence of seven metrical features in folk-verse samples from each of 104 Estonian parishes was determined. The choropleth maps of individual features were geographically quite coherent in visual observation. The attempts to divide the whole area into a few "metrical regions" based on all features as to summarize the results, however, were not that successful. The isolines of the features observed did overlap in some cases but not always, and it was not clear enough, which ones of them should be of higher priority in making the division.

The procedure that is applied in this paper to accomplish the goal mentioned above is twofold. Firstly, we apply a method of hierarchical clustering to determine the clusters of parishes that exhibit similarities in terms of the verse metre. The cluster membership obtained this way is assigned to each parish. Secondly, we identify a data-driven fuzzy classifier to facilitate the analysis of the obtained clusters and to identify which features of the verse metre are critical in cluster assignment.

2. The data set

In [4], 500 folk-verse lines from each of 104 Estonian historical parishes were used as the source material, totalling in 51,382 lines (for some parishes a lesser number of lines was available). The metre of a poem is a generalization of rhythm regularities expressed by linguistic means; in Estonian language those are the number of syllables, the quantity and the stress.

In the case of Estonian archaic tradition, the common poetic form (including the metre) has formed a separate register used creatively in several occasions, thereby the metre is in dependence of the prosodic peculiarities of the language (dialect) in which it is used. During the times of most active collection (ca 1880-1920), Estonian folksongs were in transition stage from the quantity based metre to the stress based metre; linguistic changes that supposedly were behind that, took place mostly from 13th to 16th centuries.

All the lines in the study corpus were divided into four groups according to the metrical system used: the lines that are possible in both metrical systems (ca 50% of all the lines), lines characteristic to the stress based system, lines characteristic to the quantity based system and, finally, the lines which did not follow the rules of either of the two systems (a certain amount of exceptions is regular and expected in the case of oral tradition).

For each parish, the percentage of the lines belonging to each group was calculated and these figures were used as the values of four features that characterize each parish. Expectedly, there is a strong negative correlation between the second and third feature (the percentage of lines characteristic to quantity based system and the percentage of lines characteristic to stress based system, respectively).

In addition, the occurrence of the lines with specific syllabic structure was detected: the percentage of lines containing disyllabic verse positions (instead of regular one syllable corresponding to each of 8 verse positions), the percentage of heptapositional lines (instead of regular octapositional ones), and the percentage of the lines with a pre-beat (one or two additional syllables in the beginning of line). These figures were used as values of features 5, 6 and 7, respectively.

3. Creating the groups of parishes

To detect the groups of parishes that belong together, we opt for a well-known hard clustering method, known as hierarchical clustering [5], which is based on the intuitive idea of objects being more related to nearby objects than to objects farther away.

Hierarchical clustering is an agglomerative procedure in which the clusters are initially singletons (single-member clusters). At each stage, the individuals or groups of individuals that are closest according to the linkage criterion are joined to form a new, larger cluster, which, of course, leads to a single group consisting of all individuals, formed at the last stage. This process can be represented using a dendrogram, which can be cut to extract the desired level of partitioning and there is a whole family of hierarchical clustering methods that only differ by the way the linkage criterion is computed.

In complete linkage clustering [6], the distance between two clusters is computed as the maximum distance between a pair of objects, one in one cluster, and one in the other; as a result, in each step these two clusters are merged whose merger has the smallest diameter. Complete linkage clustering tends to find compact clusters of approximately equal diameters and is therefore applied to our problem.

Table 1: The dendrogram. R is the number of clusters and figures in table cells indicate the number of objects in given clusters.

R						
1	104					
2	8	96				
3	8	53		43		
4	8	53		18		25
5	8	37	16	18		25
6	8	37	16	10	8	25

For best results, the data has to be normalized into the unit interval prior clustering. Although the clustering algorithm is agglomerative by nature, we are going to interpret the process the other way round by turning the dendrogram upside down (Table 1). It can be seen that the first to separate is a small 8-parish cluster (south-east of Estonia). In next step, the large cluster is split in roughly two clusters (Figure 1). Although there are some loose parishes, e.g. Suure-Jaani and Nõo on both sides of the smallish lake Võrtsjärv in the middle of Estonia, there is a rather neat separation line between all three clusters. We stop at this level because further steps of clustering would only introduce geographical fragmentation to existing clusters.

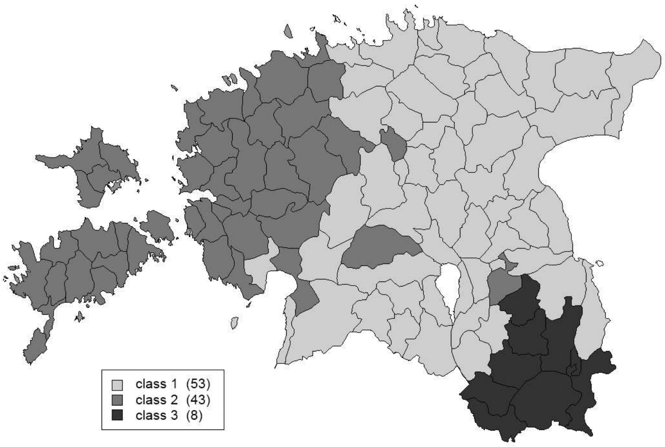

Figure 1: The parishes of Estonia divided into three groups by the hierarchical clustering method.

4. Building a classifier

A fuzzy classifier is a fuzzy rule-based system that utilizes fuzziness only in the reasoning mechanism and consists of rules in the following format

$$\begin{gathered}\text{IF } x_1 \text{ is } A_{1r} \text{ AND } x_2 \text{ is } A_{2r} \text{ AND ...} \\ \text{... AND } x_N \text{ is } A_{Nr} \\ \text{THEN } y \text{ belongs to class } c_r,\end{gathered} \tag{1}$$

where c_r is a class assigned to the r-th rule ($c_r \in (1, ..., T)$) and A_{ir} denote the linguistic labels of the i-th feature associated with the r-th rule ($i = 1, ..., N$).

Each A_{ir} has its representation in the numerical domain - a typically normal and convex membership function μ_{ir} such as a triangular membership function (MF), determined by three parameters a_{ir}, b_{ir} and c_{ir}:

$$\mu_{ir}(x_i) = \begin{cases} \frac{x_i - a_{ir}}{b_{ir} - a_{ir}}, & a_{ir} < x_i < b_{ir} \\ \frac{c_{ir} - x_i}{c_{ir} - b_{ir}}, & b_{ir} \leq x_i < c_{ir} \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

The reasoning mechanism of a fuzzy rule-based classifier is usually implemented by the single winner approach [2, 7, 8] that selects the class label

c_r, associated with the rule that provides the highest rule activation degree (τ_r) for the given feature values x_i.

$$y = c_r, \arg \max_{1 \leq r \leq R}(\tau_r), \tag{3}$$

where

$$\tau_r = \prod_{i=1}^{N} \mu_{ir}(x_i), \tag{4}$$

The goal in fuzzy rule-based classification is to obtain the maximum possible classification accuracy with as simple classifier as possible. Classification accuracy that a data driven fuzzy rule-based classifier is able to achieve, first and foremost depends on the properties of the data set. The class distributions that do not separate naturally in the product space typically need to be modeled with increased level of granularity, unless optimal or near-optimal decision border is provided by suitable rule placement.

The algorithm that we apply to construct the parish classifier consists of four steps: initialization, rule base expansion, rule base consolidation and rule compression, which are explained in the following subsections.

4.1. Classifier initialization

The simplest classifier possible is a minimal rule classifier (MRC) that specifies only one rule for each class. The training data set is divided into T subsets so that each subset contains only the samples belonging to one of T classes.

Given a subset of data S_j that contains K_j observations and its mean $\mathbf{m}_j = (m_{j1}, m_{j2}, ..., m_{jN})$ that is the geometric centroid of the data points in S_j

$$\mathbf{m}_j = \sum_{k \in S_j} \mathbf{x}_k / K_j, \tag{5}$$

where $\mathbf{x}_k = (x_1(k), x_2(k), ..., x_N(k))$ is a vector representing k-th observation, triangular MFs μ_{ij} given by parameters a_{ij}, b_{ij}, c_{ij} are created in all dimensions. For each i

$$a_{ir} = \min_{k \in S_j}(x_i(k)), c_{ir} = \max_{k \in S_j}(x_i(k)), b_{ir} = m_{ji}. \tag{6}$$

Note that the MFs are then slightly[1] enlarged so as to give nonzero membership values to the samples located at the very edges of the rule [9]. Following this, a rule of format (1) where c_r is the class that is assigned to the observations in subset S_j, is constructed[2].

Unless the classes are well separable in the product space, the MRC usually comes with a number of misclassified samples, which is also the case presently, as the MRC of the folk verse data set has 4 misclassifications (Figure 2).

4.2. Rule base expansion

To get rid of the classification error we do the obvious - increase the number of classification rules until the error disappears, using a procedure that is termed as rule base expansion. However, the expansion does not stop until what we obtain is a classifier, in which there is no overlap between the rules that represent different classes. This property becomes useful in later stages of the algorithm.

At each iteration of the procedure, a subset of data corresponding to the rule with the highest number of erroneous samples is split into two further subsets using the very same clustering method described in Section 3. If the classification error is already zero but the overlap between the rules representing different classes is still detected, we split the subset that corresponds to the rule with the maximum number of samples. Upon those two subsets, two new rules and corresponding MFs are built that ultimately replace the original one.

4.3. Rule base consolidation

The expanded parish classifier has 30 rules (the error itself disappears at 9th iteration). Obviously, so high number of rules is not acceptable and most likely, there are many rules which can be considered redundant and thus merged with neighboring rules.

The procedure for reducing this kind of redundancy is called rule base consolidation. During the consolidation, weaker rules (governing few samples) are constantly losing their samples to stronger rules (those governing many samples). Each such transfer is valid as long as accuracy is not compromised and the overlap between the rules representing different classes is not re-introduced. As a natural result, many of the weaker rules become obsolete.

The rules are ranked by their strength (the number of samples they govern) in ascending order $p \in \{1, ..., R\}$. The process starts from the lowest ranked rule ($p = 1$).

1. pick a rule R_r with the rank p
2. pick k-th sample ($k = 1, ..., K_r$) from the subset S_r governed by rule R_r.
3. transfer this sample from S_r to the subset S_q corresponding to next same class rule R_q ($c_r = c_q$) in the ranking.
4. update the MFs of both R_r and R_q on the basis of modified subsets S_r and S_q, respectively

Next we confirm if the merge can be actually executed. It depends on two conditions. The first one is that there is no accuracy loss (for which we need

[1] By 1 % of the feature domain.

[2] The rule generation/update procedure described here is used throughout the paper whenever a rule is built on a subset of data.

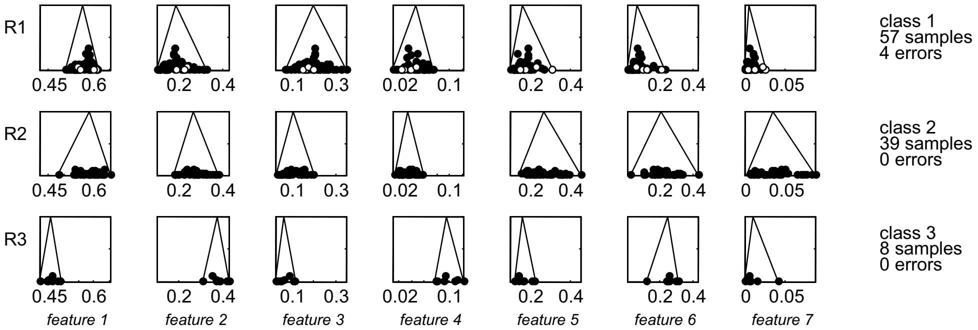

Figure 2: The minimum rule classifier of the folksong verse metre problem.

to evaluate the classifier). Secondly, we need to verify that the consolidated rule (R_q) is not overlapping with any of the rules representing classes other than c_r. Depending on if the merge has received a confirmation, there are a number of different scenarios on what to do next.

- if the merge is confirmed and $k < K_r$, increment k (select the next sample from S_r). If k, however, already equals K_r, delete rule R_r along with associated MFs, update the ranking, increment p and go back to step 1.
- if the merge is not confirmed, first discard the changes to the MFs of R_r and R_q, pick the next same class rule from the ranking and go back to step 3. If we already have reached the last same class rule in the ranking, select the next sample from subset S_r (increment k) and go to step 2. If k already equals K_r as well, increment p and return to step 1.

The process ends when we have reached the last rule in the ranking ($p = R$).

When applied to the classifier at hand, this algorithm reduces the number of rules from 30 to 8, in which three rules represent class no. 1, four rules class no. 2 and only one rule class no. 3 (Figure 3). Of those, rules no. 3, 7 and 8 are the major rules of classes 1, 2 and 3, respectively. Rules 2, 4 and 5 are minor rules describing smaller groups of parishes that are not compatible with the major rules. The classifier also contains two singleton rules (R_1 and R_6 corresponding to Torma and Mihkli parishes, respectively).

4.4. Rule compression

The last step of the classifier construction algorithm is the rule compression that is designed to improve classifier interpretability and removes the features/conditions from individual rules.

We apply the rule compression only to non-singleton rules (R_2,R_3,R_4,R_5,R_7 and R_8 in the classifier at hand). For this, we construct a table (Table 2) in which each entry in the given row contains the features in which the MFs of the given rule do not intersect with the MFs of the rule determined by the column.

	R_2	R_3	R_4	R_5	R_7	R_8
R_2	-	**6**	**7**	**4**	**7**	1,4
R_3	**6**	-	**7**	**6**	**6**	1,2,4,**6**
R_4	**7**	**7**	-	**6**,7	**6**	1-4,**6**
R_5	**4**	**6**	**6**,7	-	**7**	1,4
R_7	**7**	**6**	**6**	**7**	-	**4**
R_8	1,4	1,2,4,6	1-4,6	1,4	**4**	-

Table 2: The rule matrix that shows in which features the considered rules are not overlapping with each other.

The compression is based on the analysis of this table. For example R_3 can be compressed into features 6 and 7 because either of them is represented in all entries of the second row; the same applies to R_4. For R_2, R_5 and R_7 we need to add feature 4. R_8 only needs feature 4 for now obvious reasons. These preserved MFs are highlighted in Figure 3.

5. Analysis

Although we constrained the rule consolidation so as to ensure that the rules representing different classes would not overlap with each other, incidentally, there is no overlap even between these rules that represent the same class (this is evidenced by Table 2). This property becomes handy when interpreting the rules.

From Figure 3 it can be seen that class 3 (R_8) is characterized by a low percentage of lines characteristic to the quantity based system (feature 3) and by a high occurrence of lines characteristic to the stress based system (feature 2). These are, however, a low percentage of lines that are possible in both metrical systems (feature 1) and high percentage of exception lines (feature 4) in particular that set class 3 parishes apart from other classes. The percentage of dissyllabic verse positions (feature 5) is pretty low (similar to class 1 parishes), the percentage of heptapositional lines (feature 6) is about average and percentage of lines with pre-beat (feature 7) can be also considered low.

The parishes of class 2 are described by four rules

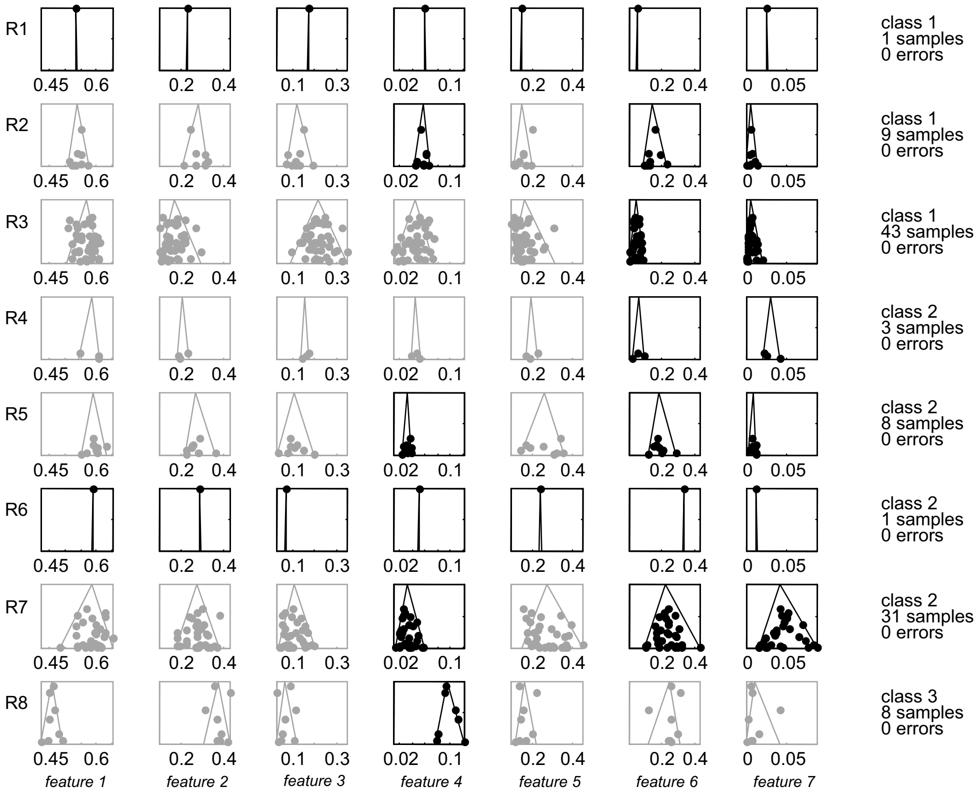

Figure 3: The rules of the parish classifier before and after rule compression. The highlighted MFs indicate the features preserved after rule compression in individual rules.

of which R_7 covers the largest part. It can be observed that there is quite a bit of variation in this group, as the MFs of the rule tend to extend over a large part of the domain. Despite that, in feature 1 the parishes remain on the higher side, in features 3 and 4 in the lower side and a clear distinction from class 1 rules can be made in features 7 and 8.

43 of 53 class 1 parishes are described by R_3 and specifically characterized by very low values in features 6 and 7. They also have lower values in features 2, 4 and 5 and those on the higher side in features 1 and 3.

R_2, R_4 and R_5 represent the deviations. For example, R_5 is every bit similar to R_7, except for the percentage of lines with pre-beat (which is very low). R_2 is quite similar to the major rule R_3 (note that the parishes in R_2 have a slightly higher occurence of lines characteristic to the stress based system and a lower percentage of lines characteristic to the quantity based system than the parishes in R_3) except for feature 6, in which it has more in common with R_7. The parishes in R_4, on the other hand, are comptaible both with R_3 or R_7 when it concerns features 1-5. By feature 6, however, they are similar to class 1 parishes in R_3 and by feature 7 to class 2 parishes in R_7. Because R_4 has as much in common with R_3 as it has with R_7, for geographical considerations alone it would make sense to reassign R_4 to class no. 1 (Figure 4).

Finally, we have two outliers, Torma and Mihkli parishes, which are described by singleton rules R_1 and R_6, respectively. Torma parish has a higher value of feature 7 than R_2 and R_3. Mihkli parish would be compatible with R_7, except for feature 7 in which it has a lower value (much like parishes in rules R_2, R_3 or R_5). The exceptionality of Mihkli parish can be explained by the earlier loss of tradition - the Brotherhood Congregation movement was very popular there and the "pagan" songs were strongly deprecated. Torma on the other hand, had mixed Estonian-Russian population and this can have had the influence to the song tradition as well.

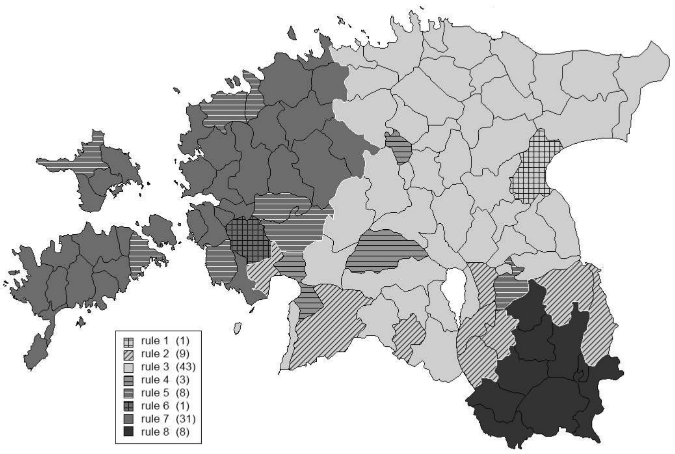

Figure 4: The groups of parishes corresponding to individual rules of the classifier.

Geographically speaking, we also have an outlier, a class 2 parish Nõo in R_5 that is situated at the border of class 3 parishes. Nõo remains in the transitional area between class 1 and class 3, but having no evidence of the main feature of class 3 (larger values of feature 4), it is similar to the other parishes in the transitional area between quantity-based and stress-based metrical systems, i.e. between classes 1 and 2. From numerical viewpoint, however, it has the second largest τ_r in R_5 (expressed by the distance of the point representing the parish from the x-axis in Figure 3), thus it is a rather typical parish for its rule and class.

6. Conclusions

The hierarchical clustering process applied on the data on the metre of Estonian folksongs succeeded to solve the initial problem - to divide the whole area into a few regions on the basis of seven metrical features. The groups obtained as a result of two first splits turned out to be geographically surprisingly clear-cut, with only a few exceptions. Considering the metrical developments, the central region (especially its northern part) represents the more conservative, quantity-based metre; the western and south-eastern regions represent the variants of stress-based metre. From the geographical placement of the classes, it is evident that the innovations in the metre must have been launched from two separate innovation centers - one in South-East of Estonia and another on the western islands. The nature of the innovations, though, has been similar by large and can be related to the systematic loss of the vowels in certain linguistic contexts in most of the Estonian dialects. (The only dialect that retained the vowels was the north-easternmost one.) The loss of vowels resulted as the loss of the syllables - the basic unit in the verse line - and brought about the differentiation of long and overlong syllables in the Estonian. The verse lines either preserved the archaic word forms or had to be restructured.

At the first sight, it might seem surprising that features 4, 6 and 7, not the features 2 and 3, which reflect the most ontological change from the viewpoint of metrical structure, are crucial for the classification. At the same time, it is clear that this would not have sufficed for differentiation of the western and south-eastern areas. The most clear rule of the classifier segregates class 3 in South-East of Estonia by abundance of the occurrence of feature 4 - exceptional lines that do not correspond to the regularities of neither system. At the closer look, it turned out that the new metrical rule had been emerged allowing overlong syllable to fill two positions in the verse line. Southeastern region is the only one where long and overlong syllables acquired separate functions in the metrical system reflecting the change in the prosodic system of language. Features 6 and 7, representing the rhythmical variants by and large concurrent to the stress-based metre, are responsible for the differentiation of classes 1 and 2. What considers the deviant rules, it is characteristic that geographically these represent the transitional areas.

In addition to the geographical placement (the western and south-eastern area separated by the central area), the classifier makes it clear once more that although the metrical innovations in two areas followed the same main pattern (from quantity-based system to the stress-based system), the developments probably took place separately as the different options were chosen to cope the loss of syllables - in the western area this was compensated by additional words or suffixes, in the south-eastern area it was accepted and the metrical structure was adjusted accordingly.

References

[1] J. Casillas, O. Cordon, F. Herrera and L. Magdalena, editors, *Interpretability Issues in Fuzzy Modeling (Studies in Fuzziness and Soft Computing, vol. 128)*, Springer-Verlag, Heidelberg, 2003.

[2] C. Mencar, C. Castiello, Cannone, R. and A. M. Fanelli, Interpretability assessment of fuzzy knowledge bases: A cointension based approach, Int. J. Approximate Reasoning, 52:501–508, 2009.

[3] J. M. Alonso and L. Magdalena, Special issue on interpretable fuzzy systems, *Inform. Sci.*, 181:4331–4339, 2011.

[4] M. Sarv, *Created for Creation: Verse Metre of Estonian Regilaul in the Tradition Process* (in Estonian, summary in English). Dissertationes folkloristicae Universitatis Tartuensis 11. University of Tartu Press, 2008.

[5] G.N. Lance and W.T. Williams, A General Theory of Classificatory Sorting Strategies. 1. Hierarchical Systems, *Computer Journal*, 9:373–380, 1967.

[6] T. Sorensen, A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, *Biologiske Skrifter*, 5:1–34, 1948.

[7] H. Ishibuchi, T. Nakashima and T. Murata, Three-Objective Genetics-Based Machine Learning for Linguistic Rule Extraction, *Inform. Sci.* 136:109–133, 2001.

[8] J. A. Roubos, M. Setnes and J. Abonyi, Learning Fuzzy Classification Rules from Labelled Data, *Inform. Sci.* 150:77–93, 2003.

[9] A. Riid and E. Rüstern, An Integrated Approach for the Identification of Compact, Interpretable and Accurate Fuzzy Rule-Based Classifiers from Data, In: *Proc. IEEE Int. Conf. Intelligent Eng. Syst.*, pages 101–107, Poprad, 2011.