The use of clustering to understand disease progression in Rheumatoid Arthritis

1st Beatriz de la Iglesia
School of Computing Sciences
University of East Anglia)
Norwich, UK
b.iglesia@uea.ac.uk

2nd Kathapet Nawongs
School of Computing Sciences)
University of East Anglia)
Norwich, UK
k.nawongs@uea.ac.uk

3rd Jack Dainty
Norwich Medical School
University of East Anglia
Norwich, UK
Jack.Dainty@uea.ac.uk

4th Alexander Macgregor
Norwich Medical School
University of East Anglia
Norwich, UK
A.Macgregor@uea.ac.uk

***Index Terms*—clustering, unsupervised learning, rheumatoid arthritis, sequence distance metrics**

***Abstract*—In this paper we examine data representing patients with Rheumatoid Arthritis (RA). This is an important medical conditions that affects a proportion of the adult population and is very disabling. The data contains some demographics as well as follow up for up to 20 years where objective measures of 'joint involvement', e.g. counts of how many tender or swollen joints are present in a given follow up year, are recorded.**

To date the patterns of disease progression and joint involvement have not been investigated in detail for RA. We propose a clustering approach to extract patterns of joint involvement in disease progression for groups of patients. For this, we investigate how to measure distance for the type of data we analyse which consists of multiple attributes each corresponding to years of follow up measuring a particular objective measure. We settle for an aggregate Dynamic Time Warping measure of distance between patients and use it in combination with K-means clustering to cluster our patient trajectories. Our preliminary results, with some interpretation, show that it is possible to cluster such complex data to extract some meaningful patterns of joint involvement in disease progression.

I. Introduction

Rheumatoid arthritis (RA) is a chronic disabling condition that affects around 1% of the adult population. It can start at any age, and can progress to involve multiple joints and other organ systems. The peak age of incidence in the UK for both men and women is in the 70s [4]. The disability associated with RA increases with a patient's age. The rate of progression in individuals varies considerably, and while the disease can respond to treatment, individual responses also vary. RA can be detrimental to the patient's mental as well as physical health. Smolen et al. [16] describe "musculoskeletal deficits", "decline in physical function and quality of life and cumulative morbidity risk" as potential consequences. There is also a cost attached to it as it may render people unable to work or perform other physical activities such as sports.

Disease activity in RA is often measured as a composite summary score (known as the DAS28 [18] which includes data from a limited number of joints affected at one particular time. This score does not include information on the pattern of joint involvement (for example whether the disease is predominantly confined to small joints, or if larger joints or the spine are involved). To date the patterns of disease progression at individual joint sites over time in people with RA has not been described. Some authors [3] have looked at the patterns of inflammatory joint involvement, including symmetry patterns where joints on both sides are involved or ray patterns, where all joints in a particular digit (finger or toe) are involved.

In this paper, we apply a machine learning methodology to examine these patterns of disease progression in data that has been recorded in the Norfolk Arthritis Register [13]. This is the largest long term disease register of its type in the world, and includes serial data from over 5,000 patients with inflammatory arthritis assessed at up to 20 time points (years 0-5,7,8,10,12,15,18,20) for over 25 years. The data include objective measures from 44 joints at each time point. The analysis will enable the exploration and characterisation of the patterns of joint involvement over time (e.g. symmetry), and the extent to which they are modified by individual patient characteristics and treatments.

The data available consists of multiple attributes per patient (89), each patient is associated with a number of years of follow up which can be interpreted as a Time Series (TS) or a Sequence per attribute. Clustering such data requires methods for clustering multi-variate sequences or TS which are not well developed.

Clustering of complex medical data is now emerging as an area of application of machine learning including clustering of genetic signatures to identify leukaemia [12]; clustering of RA patients to investigate patient characteristics and drug effects [10]; and clustering of of Juvenile Rheumatoid Arthritis patient data to find disease progression based on early patterns of joint involvement [1].

We demonstrate how a machine learning clustering approach using a distance measure adapted specifically for our complex data can be used to group similar patterns of disease progression over time, which involve multiple joints presenting similar characteristics. Our contribution involves the novel clustering of multi-TS data as well as the application to RA data and further interpretation of the clustering results which should eventually enable a new understanding of disease

We acknowledge support from Grant Number ES/L011859/1, from The Business and Local Government Data Research Centre, funded by the Economic and Social Research Council to provide economic, scientific and social researchers and business analysts with secure data services

progression for specific patient groups.

Our paper is structured as follows: section II presents the data; section III present our method and IV presents the results with a final section for our conclusions and further work.

II. Data

For each patient we have both year of birth and gender plus a number of "joint variables" which either count someone having 'tender' or 'swollen' joints (or both). For an illustration of the joints of a human hand which the data refers to see Fig. 1

As previously mentioned, there are up to 20 years of follow-up recorded per patient in up to 12 visits in specific years, though for some patients there are a lot less, with many missed years of follow up. See Fig. 2 for a visualisation of the records available by gender for each year of follow up. Females are more prominent in the dataset for all follow up years. Hence, for each patient if we consider a specific attribute, e.g. swollen metacarpals (MCP) on the left, this is a sequence of values which can either be Boolean (taking values 0 for false and 1 for true) or an integer representing a count (e.g. number of swollen Proximal Phalanges (PIPs) which can take values up to 10).Clustering such data requires methods for clustering multi-variate sequence data/TS data which are not well developed.

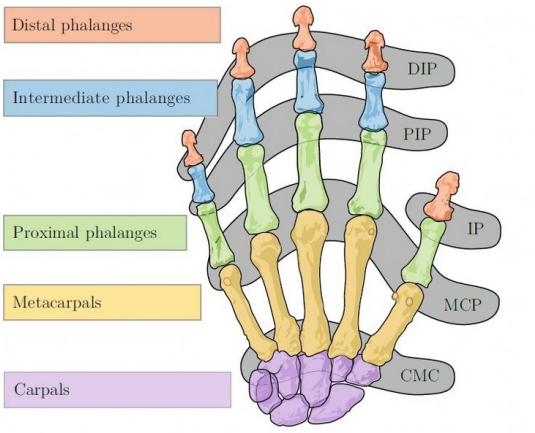

Fig. 1. Bones and joints of a human hand. Diagram is from [17]

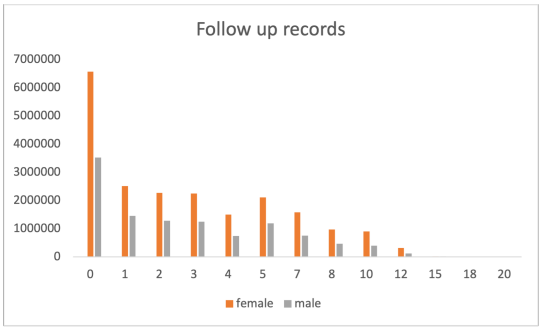

Fig. 2. Number of records available for each follow up year represented in the x-axis and broken down by gender

The attributes available for the analysis are described in Table I.

III. Methods

To cluster our data successfully, first we need a distance measure that can be used with multi-variate sequence or TS data. Our first approached used the Levenshtein distance [11] which represents the minimum number of single-character edits (insertions, deletions or substitutions) required to change one sequence into another. This was a basis for our own distance metric which we adapted to take account of the multi-sequence data (Boolean and numeric) so that we compute a distance for each pair of patient records. We then use the distance matrix generated to cluster the dataset using a hierarchical agglomerative algorithm [9]. We find 5 distinct clusters that correspond to different disease projection trajectories. However most of the clusters were found to be dependent on the length of follow up so this was not a particularly helpful segmentation and results are not presented here.

Our second approach looks at the data as TS. For numeric (not binary) attributes, we normalised the data as a pre-processing step to ensure that attributes with larger values do not dominate the distance calculations. We then use the Dynamic Time Warping (DTW) distance measure for two time series [5] to compute the distance between two patients at each TS attribute level. DTW is a distance measure that is insensitive to local compression and stretches; the warping optimally deforms one of he two input series onto the other. As we can see in Fig. 3 a single point "observation" in the first TS is aligned to one or more points in the second TS.

Consider two TS to compare X=x_1,x_2,x_3,...,x_n and Y=y_1,y_2,y_3,...,y_m. To compare X and Y, a point-wise distance matrix $M(n*m)$ is created, where every element in this matrix corresponds to the distance between two points $i \in X$, and $j \in Y$, as follows:

$$M_{i,j} = (x_i - y_j)^2 \tag{1}$$

To find the optimal alignment between X and Y, a warping path

$$W = w_1, w_2, w_3, \ldots, w_k,$$

in matrix M_i is constructed, where $w_k = (i,j)_k$ indicates the alignment and matching relationship between i and j.

The DTW distance between X and Y is calculated as follow:

$$DTW(X,Y) = min\left\{\frac{1}{K}\sqrt{\sum_{k=1}^{K} W_k}\right\}. \tag{2}$$

For our specific application, the overall distance between two patients is calculated as the average distance between all the attributes in the dataset. The aggregate distance matrix contains an entry for each pair of patients. This is, the aggregated distance is calculated as the simple average distance of all attributes (A_1, A_2, A_3, ..., A_n). For example for patients P and Q, the aggregate distance between these patients is:

Column index	Attribute name	Type	Description
0	regno	Numerical	Patient's unique ID number
1	fupno	Numerical	Follow-up year (from 0 to 20)
2	dobyear	Numerical	Date of birth year
3	gender	Categorical	Male or female
4–16	swwrst-swankl	Binary	If a specified joint or part of body is swollen or not
17–26	numpip-numjoin	Numerical	Count how many are swollen for the given attribute
27	swollen_28jt	Numerical	Sum of all swollen joints out of 28
28	numsym	Numerical	Number of swollen ARA joints for which symmetrical joint is affected
29	num_non	Numerical	Number of swollen ARA joints for which symmetrical joint is **not** affected
30	joiswlg	Numerical	Number of swollen large joints
31	lgejoint	Binary	If they have any swollen large joints or not
32	both_51jt	Numerical	Sum up the number of tender and swollen joints out of 51
33-48	baxmcpr-bankl	Binary	Whether a specified joint or body part is swollen and tender
49-59	bnmpip-bjoislg	Numerical	Counts how many are swollen and tender for an attribute
60	blgejt	Binary	If a large joint is swollen and tender
61	both_28jt	Numerical	Count how many joints are swollen and tender out of 28
62	tend_51jt	Numerical	Count only the number of tender joints out of 51
63-75	temcpr-tknee	Binary	If the specified joint/body points are swollen or not
76-86	tnmpip-tlgejt	Numerical	How many of each specified joint/body part is swollen
87	tlgejt	Binary	If the patient has a tender large joint or not
88	tend_28jt	Numerical	The number of tender joints out of 28 joints

TABLE I
SUMMARY DESCRIPTION OF THE ATTRIBUTES IN THE DATA SET

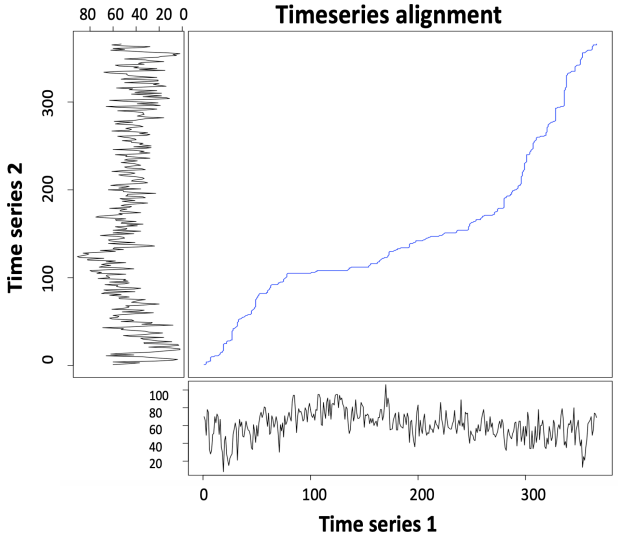

Fig. 3. Visual comparison of optimal alignment between two time series based on DTW.

$$AggDTW(A,B) = \frac{\sum_{i=1}^{n} DTW(P_{A_i}, Q_{A_i})}{n}. \quad (3)$$

where n is the number of attributes, P_{Ai} represents the values of follow up for patient P and attribute A_i and Q_{Ai} represents the values of follow up for patient Q in the same attribute and DTW is the distance between those values.

A number of clustering algorithms [8] were applied to the data once distances were obtained, including k-means [6], k-medoids [14] and an agglomerative algorithm [7].

To work out the best number of clusters k to report, we used the silhouette visualisations [15] which compare the tightness and separation of objects in particular clusters into one visual plot. The average silhouette width provides an evaluation of clustering validity and can be used to select an 'appropriate' number of clusters. The silhouette visualisations use Principal Component Analysis (PCA) which is able to reduce the dimensionality of data points [2] by keeping the variables that produce a large proportion of the variation in the data and are not correlated with each other. This enables visualisation in two dimensions.

IV. RESULTS

The best results were obtained using k-means with an aggregated DTW distance metric and we chose a value of k=5 clusters (see fig. 4), although still the silhouette coefficient value was low and we expect to carry out further experimentation in the future to improve on this. Nevertheless, as a first exercise it enable us to proceed a groping of patients that we can attempt to further interpret.

The resulting clusters enable us to group patients that show a similar trajectory of disease activity and analysing those clusters in terms of their characteristics. For example, in Table II we show the (normalised) average values for the groups of patients in a cluster for a specific attribute and year of follow up. We pick some representative attributes but others may also be interesting to compare.

We can see that the top cluster in Fig. 4, cluster 4, has an average of 0.6 (1 being the highest value) in Table II for 'any tender large joints', which is higher than other clusters in year 1 and stays high through follow up. They also have a high value at cluster 4 for 'Any Swollen MCPs on left' and 'Any tender' PIP's on right. This may represent a group of patients who have high levels of RA progression through the follow up years and in the visualisation it is positioned as cluster of high values for the PCA components.

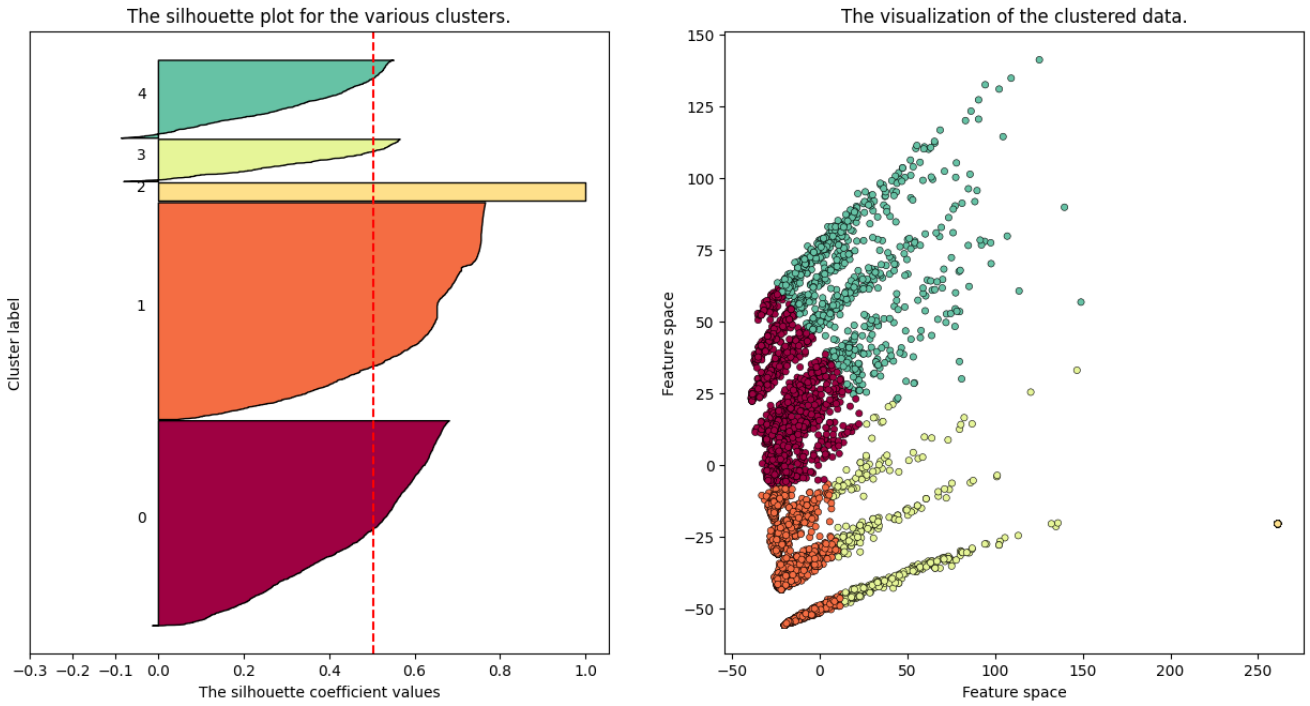

Fig. 4. Silhouette analysis for k-means with k=5 clusters

On the other hand, cluster 0 represents a cluster which over the follow up years has high values of 'any both (s and t) MPCs and perhaps other attributes.

Cluster 2 contains a group of outliers which cluster together very closely but separate from the other data points. This is observable in the visualisation although they overlap to appear as one point to the right of the diagram. The outliers may give the cluster a high silhouette coefficient but may overall deteriorate the clustering results. The cluster corresponds to 143 data points which are very closed together so one possibility may be to remove those and redo the clustering. It has, for example, low values for 'any large tender joints' in the initial years. It has also low values in the early years for most attributes we are examining in Table II.

Cluster 3 has low values for most of the later follow up years and for most of the attributes we are examining but not the lowest in year 0. Perhaps it represents patients who present improvements over time. Cluster 1 may display more middle values for most attributes and may be difficult to differentiate from other clusters, in particular from cluster 3.

V. Conclusions and Further Work

So far we demonstrated how a machine learning clustering approach using a distance measure adapted specifically for these complex data can be used to group similar patterns of disease progression over time, which involve multiple joints presenting similar characteristics. Our contribution involves the novel clustering of multi-sequence data using an aggregation of DTW as a distance metric, as well as the application to RA data and further interpretation of the clustering results which may enable new understanding of disease progression for specific patient groups. Our results are very preliminary and will require further interpretation. For example, we may be able to use a decision tree to extract a classification in terms of different attributes that denote disease progression for each of the clusters. This may enable further understanding/interpretation of results. As future work, we need to experiment more with clustering algorithms and we need to involve the experts in interpretation of results from a medical and disease progression point of view but we believe that the results are encouraging so far.

References

[1] Mohammad J Al-Matar, Ross E Petty, Lori B Tucker, Peter N Malleson, Maria-Louise Schroeder, and David A Cabral. The early pattern of joint involvement predicts disease progression in children with oligoarticular (pauciarticular) juvenile rheumatoid arthritis. *Arthritis & Rheumatism*, 46(10):2708–2715, 2002.

[2] D.J. Bartholomew. Principal components analysis. In Penelope Peterson, Eva Baker, and Barry McGaw, editors, *International Encyclopedia of Education (Third Edition)*, pages 374–377. Elsevier, Oxford, third edition edition, 2010.

[3] Vinod Chandran, Lynne Stecher, Vern Farewell, and Dafna D. Gladman. Patterns of peripheral joint involvement in psoriatic arthritis—symmetric, ray and/or row? *Seminars in Arthritis and Rheumatism*, 48(3):430–435, 2018.

[4] National Institute for Health and Care Excellence (NICE). Rheumatoid Arthritis: How common is it? https://cks.nice.org.uk/topics/rheumatoid-arthritis/background-information/prevalence-incidence/, 2022. [Online; accessed 16 March 2022].

[5] Toni Giorgino. Computing and visualizing dynamic time warping alignments in r: The dtw package. *Journal of Statistical Software*, 31(7):1–24, 2009.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Any tender large joints					
year 0	0.55	0.54	*0.42*	0.5	**0.6**
year 1	0.42	0.36	*0.32*	**0.5**	0.43
year 2	0.37	*0.24*	0.27	0.29	**0.4**
year 3	0.35	*0.17*	0.19	0.19	**0.36**
year 10	0.14	*0.06*	0.09	*0.06*	**0.19**
year 15	0.06	0.03	0.05	0.02	**0.1**
Any swollen MCPs on left					
year 0	0.38	0.35	*0.3*	0.34	**0.43**
year 1	**0.25**	0.2	*0.15*	*0.15*	**0.25**
year 2	0.2	0.11	0.14	*0.08*	**0.22**
year 3	0.17	0.09	0.1	*0.06*	**0.19**
year 10	0.07	0.03	0.03	*0.02*	**0.08**
year 15	0.02	0.01	0.02	*0.01*	**0.04**
Any tender PIPs on right					
year 0	0.39	0.37	*0.32*	0.36	**0.42**
year 1	**0.25**	0.2	0.19	*0.16*	0.24
year 2	0.21	0.14	0.14	*0.11*	**0.23**
year 3	0.2	0.08	0.11	*0.07*	**0.19**
year 10	0.07	0.03	0.03	*0.02*	**0.1**
year 15	0.03	0.02	0.02	*0.01*	**0.05**
Any both (s and t) MCPs					
year 0	0.34	0.31	*0.26*	0.32	**0.35**
year 1	**0.21**	0.16	0.16	*0.12*	0.16
year 2	**0.16**	0.1	0.09	*0.07*	0.11
year 3	**0.15**	0.09	0.07	*0.05*	0.1
year 10	**0.05**	0.03	*0.02*	0.07	*0.02*
year 15	**0.02**	*0.01*	*0.01*	0.03	*0.01*

TABLE II
AVERAGE (NORMALISED) VALUES FOR EACH CLUSTER AND SPECIFIC ATTRIBUTES, WITH LARGEST VALUES FOR ANY ROW IN BOLD AND SMALLEST VALUE IN ITALIC

[6] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[7] Henry Hexmoor. Chapter 6 - diffusion and contagion. In Henry Hexmoor, editor, *Computational Network Science*, Emerging Trends in Computer Science and Applied Computing, pages 45–64. Morgan Kaufmann, Boston, 2015.

[8] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[9] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[10] Seung Min Jung, Kyung-Su Park, and Ki-Jo Kim. Clinical phenotype with high risk for initiation of biologic therapy in rheumatoid arthritis: a data-driven cluster analysis. *Clinical and Experimental Rheumatology*, 39:1282–1290, 2021.

[11] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk SSSR*, 163:845–848, 1965.

[12] Christian Lopez, Scott Tucker, Tarik Salameh, and Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. *Journal of Biomedical Informatics*, 85:30–39, 2018.

[13] Alex Macgregor. Norfolk Arthritis Register (NOAR). https://noar.uea.ac.uk/, 2022. [Online; accessed 16 March 2022].

[14] Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 36(2, Part 2):3336–3341, 2009.

[15] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[16] Josef S. Smolen, Daniel Aletaha, and Iain B. McInnes. Rheumatoid arthritis. *The Lancet*, 388:2023–2038, 2016.

[17] Mahmoud Tavakoli, Rafael Batista, and Lucio Sgrigna. The uc softhand: Light weight adaptive bionic hand with a compact twisted string actuation system. *Actuators*, 5:1, 12 2015.

[18] DM Van der Heijde, Martin A van't Hof, PL Van Riel, LA Theunisse, Evelien W Lubberts, Miek A van Leeuwen, Martin H van Rijswijk, and LB Van de Putte. Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. *Annals of the rheumatic diseases*, 49(11):916–920, 1990.