VISUALIZING EMR OF DIABETIC PATIENTS

USING PAIRWISE SIMILARITY FOR EXPLAINABLE STRUCTURING

Joris Falip, Sarra Barraud, Frédéric Blanchard
joris.falip@utt.fr
www.datacrunch.sh

October 1st 2020

CReSTIC, University of Reims Champagne-Ardenne, France

Context

OVERVIEW

Collaboration with the University Hospital of Reims.

Type 1 diabetes

Incidence has been increasing worldwide, particularly in children.

Various complications affecting quality of life:

- Retinopathy
- Kidney disease
- Neuropathy
- Cardiovascular diseases

Main known marker for this risk is HbA1c.

OVERVIEW

HbA1c: average blood sugar level over 2-3 months.
Can't be used alone, other factors must be taken into account.

We helped medical experts understand factors favoring diabetes complications by studying the CARéDIAB database.

OVERVIEW

HbA1c: average blood sugar level over 2-3 months.
Can't be used alone, other factors must be taken into account.

We helped medical experts understand factors favoring diabetes complications by studying the CARéDIAB database.

CARéDIAB

Champagne **A**rdenne **Ré**seau **Diab**ète database.
Involves CReSTIC lab, Reims university hospital and ORNICARE health network.

Long term follow-up of type 1 diabetes patients in the Champagne-Ardenne area over 15 years, including each medical appointment and hospitalization.

GOALS

Aim of this work

Help domain experts carry out interactive exploratory data analysis by visualizing, manipulating and exploring their datasets.

We want a solution that:

- structures data around exemplars
- provides intuitive visual exploration
- offers links between data points to guide exploration

CONSTRAINTS

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensional data

CONSTRAINTS

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensional data

Curse of dimensionality

"Having more dimensions usually means data sets tend to be sparse and all distances between data points tend to become harder to distinguish." - N. Tomasev et al.

CONSTRAINTS

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensional data

Curse of dimensionality

"Having more dimensions usually means data sets tend to be sparse and all distances between data points tend to become harder to distinguish." - N. Tomasev et al.

3. No *a priori* knowledge

CONSTRAINTS

1. **Explainable** algorithm and **interpretable** results
2. Efficient with high-dimensional data

Curse of dimensionality

"Having more dimensions usually means data sets tend to be sparse and all distances between data points tend to become harder to distinguish." - N. Tomasev et al.

3. No *a priori* knowledge → no dimensionality reduction

Method

HUBNESS PHENOMENOM

Hubness

"High dimensionality causes some data points to be the nearest neighbor of many others points, thus becoming ′hubs′." - totally made up definition

Hubs exhibit interesting properties and can be used to improve traditional data mining approaches. They will be used as exemplars in the structuring process.

We want to compute the similarity of each data point to its exemplar.

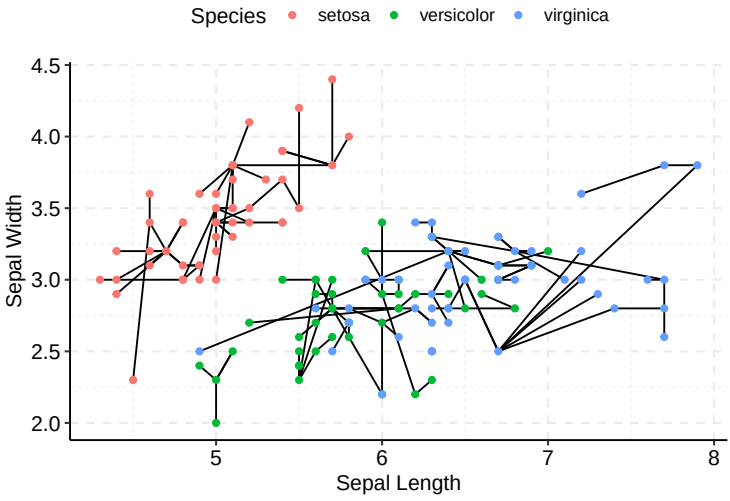
ALGORITHM

- Compute on each dimension then aggregate
- Use ranks to avoid sparsity and outliers' exclusion

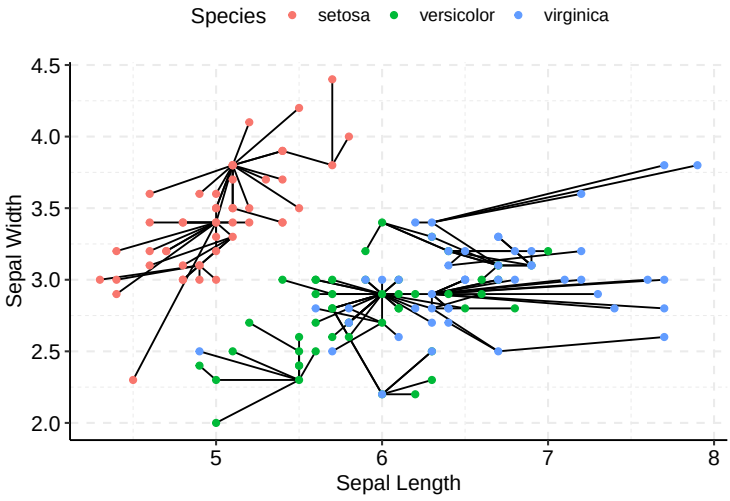
Steps

1. Rank individuals on each dimension
2. Aggregate the results into a score : the *Degree of Representativeness*
3. Link individuals to the element with the highest *DoR* in their k-neighborhood

IRIS - K = 25

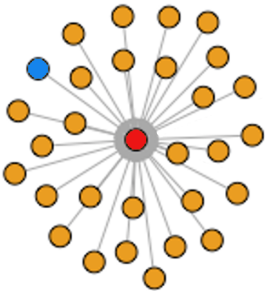

Species
setosa
versicolor
virginica
Sepal Width
Sepal Length
4.5
4.0
3.5
3.0
2.5
2.0
5
6
7
8

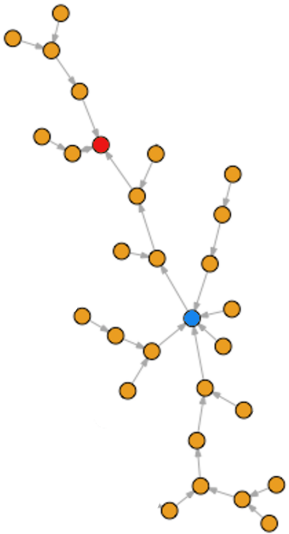
IRIS - K = 50


Species
setosa
versicolor
virginica
Sepal Width
Sepal Length
4.5
4.0
3.5
3.0
2.5
2.0
5
6
7
8

Experiments

EVALUATION

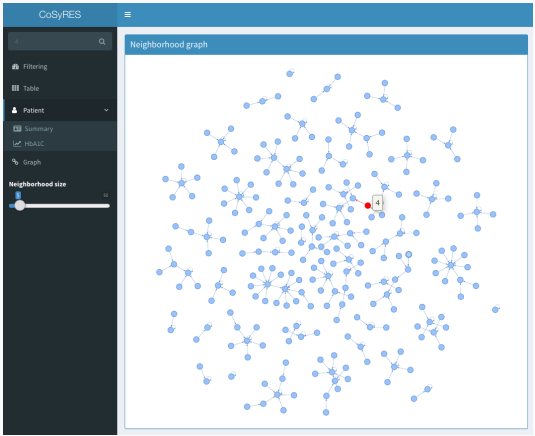


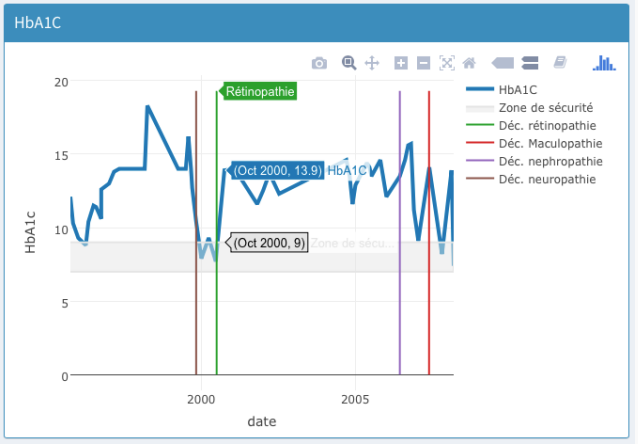


EXPERIMENTS - CONNECTED COMPONENTS

	exemplar structuring		nearest neighbor	
dataset	size	diameter	size	diameter
synthetic	16.7	5.2	13.6	2.8
residential	5.8	2.7	3.5	2.1
communities	8.7	3.2	5.5	2.5

CURRENT PROTOTYPE





Summary

CONCLUSION

In a nutshell

- Structuring of high-dimensional data
- Interpretable results for domain experts
- Graph structure suitable for exploration

CONCLUSION

In a nutshell

- Structuring of high-dimensional data
- Interpretable results for domain experts
- Graph structure suitable for exploration

Future work

- Automatically include user feedback
- Use all values of **K** to compute exemplars

VISUALIZING EMR OF DIABETIC PATIENTS

USING PAIRWISE SIMILARITY FOR EXPLAINABLE STRUCTURING

Joris Falip, Sarra Barraud, Frédéric Blanchard
joris.falip@utt.fr
www.datacrunch.sh

October 1st 2020

CReSTIC, University of Reims Champagne-Ardenne, France

REFERENCES

- "**Rapid Decision Making on the Fire Ground**"
  2010 - G. Klein et al.
- "**On the Surprising Behavior of Distance Metrics in High Dimensional Space**"
  2001 - C. C. Aggarwal et al.
- "**The Role of Hubness in Clustering High-Dimensional Data**"
  2014 - N. Tomasev et al.