Supervised machine learning algorithms for disease prediction using administrative claim data

Dr Shahadat Uddin

Senior lecturer

Complex Systems Research Group and John Grill Institute of Project,

Faculty of Engineering, The University of Sydney

New South Wales, Australia

Motivations

- Supervised machine learning algorithms already gained wide acceptance for developing predictive models in various contexts
- A large volume of healthcare data has been collected on a regular basis by different healthcare service providers

On the other side

- Chronic diseases are the leading causes of death worldwide.
- Diabetes is one of the major chronic diseases.
- About 422 million people worldwide have diabetes (WHO).
- According to Australian Institute of Health and Welfare (AIHW 2019)–
  - Diabetes contributes 11% of deaths in 2017
  - Type 2 diabetes accounts for over half of all diabetes deaths
  - An estimates 1.2 million (6%) Australian adults had diabetes in 2017-2018
- Diabetes (T2D) could lead to the development of other chronic diseases (e.g. CVD).

THE UNIVERSITY OF SYDNEY

Related other studies

- Rule-based scoring models including Charlson Comorbidity Index (Charlson et al., 1987)
  - to predict the 10-year mortality for a patient.

- A collaborative filtering method – CARE (Davis et al., 2010)
  - can predict future disease risk.
  - but it raises many false alarms to predict the future disease risks.

- Network-based approaches – (Khan et al., 2018)
  - to understand and represent the progression of T2D using graph analytics.
  - multiple chronic disease progression is not tested.

Research goal

- Employ supervised machine learning algorithms to develop predictive risk model for type 2 diabetes using only administrative claim data
  - Logistic regression
  - Support vector machine
  - Random forest
  - K-nearest neighbour
  - Artificial neural network

- Tuning of hyperparameter

Research methods

Data source

- Administrative claim data provided by CBHS (https://www.cbhs.com.au/)
- Total patients: 8000 (4000 diabetic and 4000 Non-diabetic)
- Use ICD codes to extract the records of diabetes patients

		Overall	Diabetic	Non-Diabetic
	Overall	8000	4000	4000
No of patients	Male	2618	1751	867
	Female	5382	2249	3133

Research methods (cont.…)

Variable selection

S/L	Comorbidity	S/L	Comorbidity
1	Congestive heart failure	16	Lymphoma
2	Cardiac arrhythmias	17	Metastatic cancer
3	Valvular disease	18	Solid tumour without metastasis
4	Pulmonary circulation disorders	19	Rheumatoid arthritis/collagen vascular diseases
5	Peripheral vascular disorders	20	Coagulopathy
6	Hypertension, uncomplicated	21	Obesity
7	Hypertension, complicated	22	Weight loss
8	Paralysis	23	Fluid and electrolyte disorders
9	Other neurological disorders	24	Blood loss anaemia
10	Chronic pulmonary disease	25	Deficiency anaemia
11	Hypothyroidism	26	Alcohol abuse
12	Renal failure	27	Drug abuse
13	Liver disease	28	Psychoses
14	Peptic ulcer disease excluding bleeding	29	Depression
15	AIDS/HIV		

Comorbidities and health conditions added to Elixhauser index			
30	Cataract	33	Macular degeneration
31	Anaemia, unspecified	34	Presence of coronary angioplasty implant and grafts
32	History of long-term medication, insulin	35	Presence of aortocoronary bypass graft

A Khan, S Uddin, U Srinivasan, (2019) *Chronic Disease Prediction Using Administrative Data and Graph Theory: The Case of Type 2 Diabetes*, Expert Systems with Applications

Results and Discussion

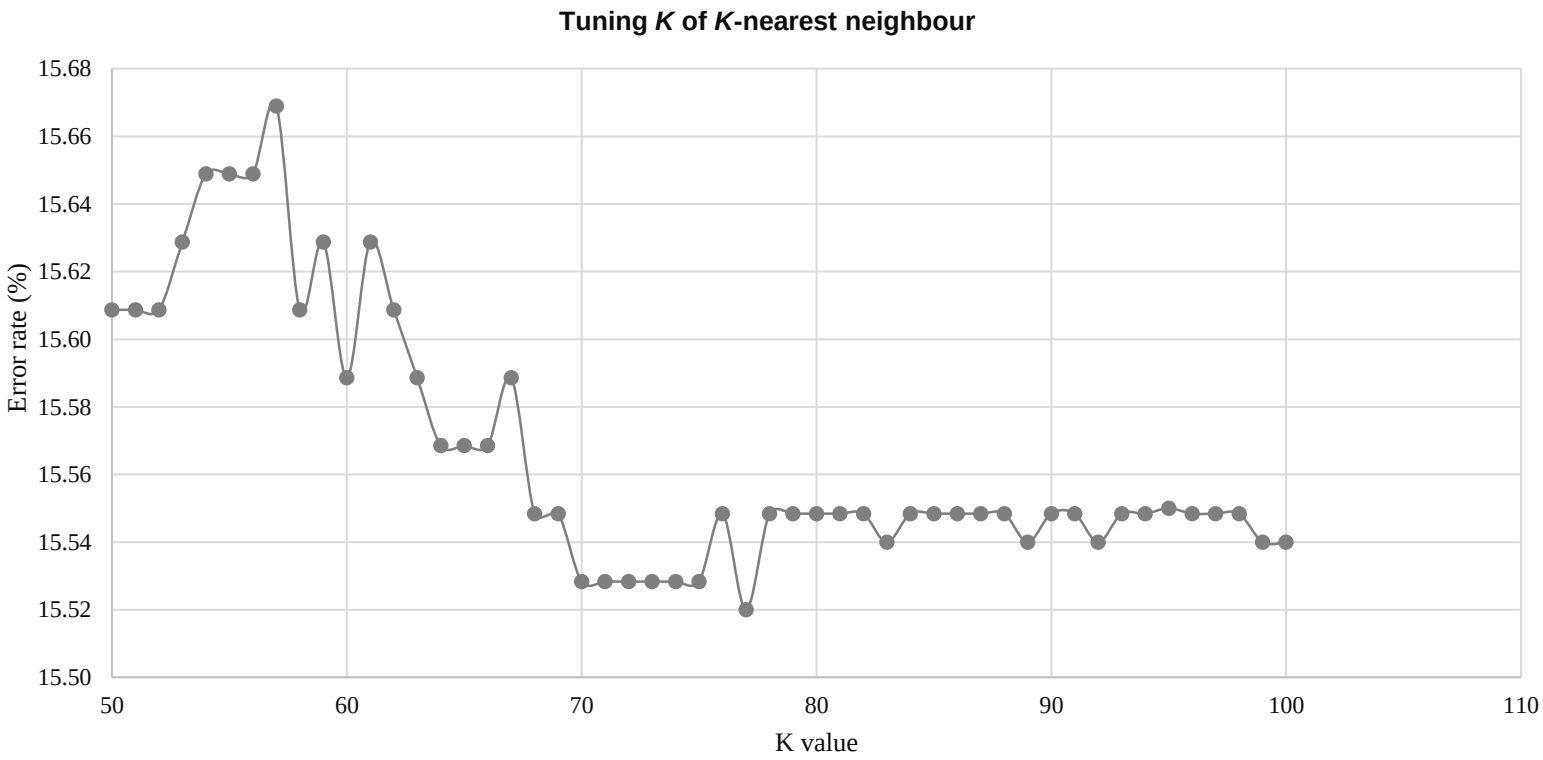
Comparison of performance (10 fold, 80/20 split, python SKlearn package)

$$Accuracy = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

ML (supervised) algorithms	Accuracy (%)
Logistic regression	77.56
Support vector machine	76.32
Random forest	81.95
K-nearest neighbour	**82.73**
Artificial neural network	80.42

Results and Discussion (cont…)

Tuning *k* value for KNN (10 fold and 80/20 split)



KNN improves its accuracy to 84.48%

Results and Discussion (cont…)

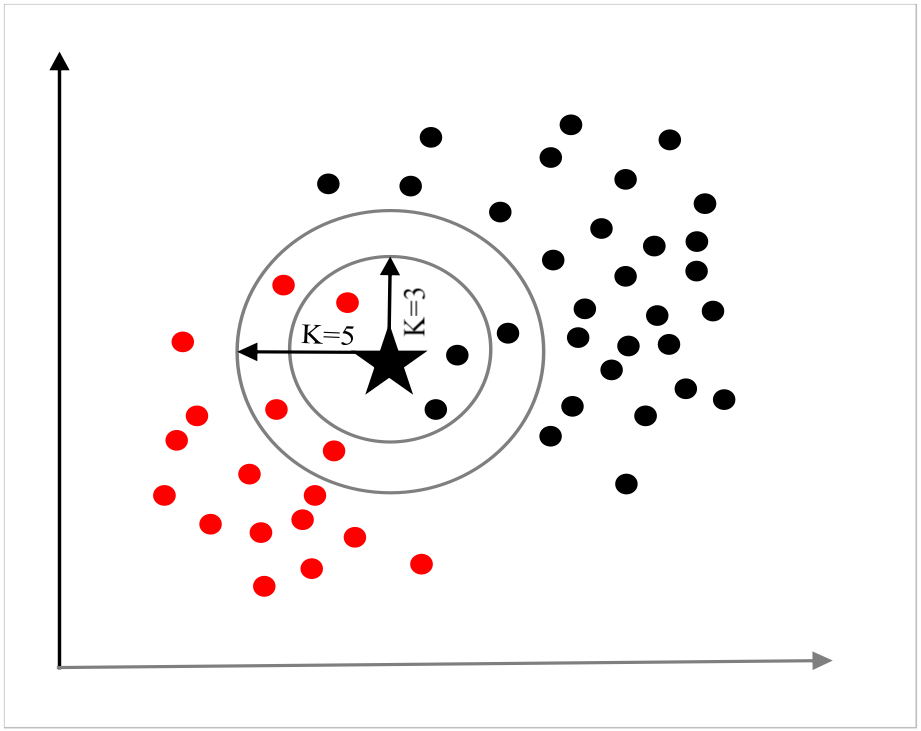
Further insight from KNN (group-wise performance)

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F1\ score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

Group	Precision	Recall	F1 socre
Non-diabetic	0.83	1.00	0.91
Diabetic	1.00	0.00	0.00

Results and Discussion (cont…)

Further insight from KNN – develop propensity model



Propensity model :– predict disease risk with an *p* value

Neighbour statistics is used to develop a propensity model

Received the IP rights from USyd of an integrated software tool (Database, SQL and Python)

THE UNIVERSITY OF SYDNEY

Summary

In a nutshell…

- Apply ML for disease risk prediction by using only administration claim data
- All variables considered in this study can be extracted from claim data
- The precision value for the diabetic patients indicates that this approach can be used for designing intervention program.

Future study…

- Similar experiment and study design for –
  - Other chronic diseases
  - Comorbidity of multiple chronic diseases

THE UNIVERSITY OF SYDNEY

References

- https://www.aihw.gov.au/reports/diabetes/diabetes-snapshot/contents/how-many-australians-have-diabetes/type-2-diabetes
- https://www.who.int/news-room/fact-sheets/detail/diabetes
- Charlson, M. E., et al. (1987). "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation." Journal of chronic diseases **40**(5): 373-383.
- Davis, D. A., et al. (2010). "Time to CARE: a collaborative engine for practical disease prediction." Data Mining and Knowledge Discovery **20**(3): 388-415.
- Khan, A., et al. (2018). "Comorbidity network for chronic disease: A novel approach to understand type 2 diabetes progression." International journal of medical informatics **115**: 1-9.


Q&A