Machine Learning and Data Mining Methods in Diabetes Research (2017)

Kavakiotis et al.

1.0 My thoughts

I think this is a really good paper to use especially in our background. Depending on the direction we take for our algorithm, we may also cite this as one of the reasons why we made the decision we made. Also helps us brag about all the features that we have included to make our algorithm unique from others (ex. feature selection + SVMs for instance).

I definitely recommend reading this paper in more detail if you are interested in learning how to write review papers. It is very well laid out and the methodology section can teach you on how to search.

In terms of current work in live data analytics, from my numerous searches, I always come up with a ton more papers on diagnosing diabetes type 2 than on managing it.

The weakness of this paper is that it does not directly tackle diabetes management.
The strength is that it gives us a ton of information on what has been done with regards to diabetes and machine learning.

2.0 Goals

To survey the following four fields

1. Prediction and Diagnosis (most widely studied and researched)
2. Diabetic Complications
3. Genetic Background and Environment
4. Health Care and Management

3.0 General Info

- Definition of Machine Learning (ML) and Knowledge Discovery Databases (KDD)
    - **ML:** Machines that learn from experience
    - **KDD:** Making sense of the data and producing useful knowledge from them. Ultimate goal is finding patterns.

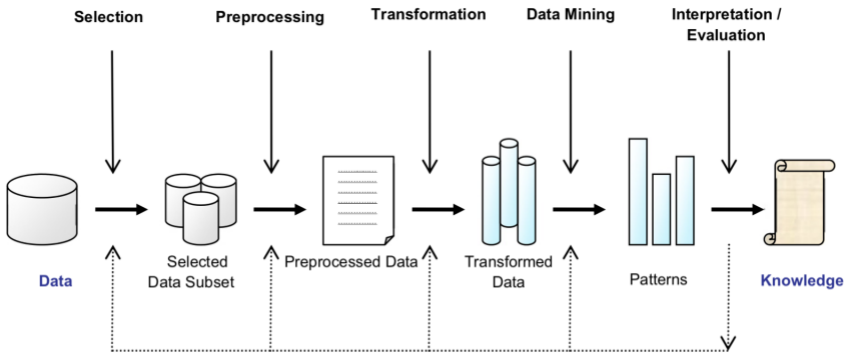

Fig. 1. The basic steps of the KDD process.

- Categories of ML
  - **Supervised**
    - coming up with an expression (target function) that describes that data by learning inductively given a training set.
    - classification and regression; some techniques include decision trees, artificial neural networks, and support vector machines.
  - **Unsupervised**
    - discovering relationships between data point without any label.
    - *Association Rule*: mostly used in biology and genetics
    - *Clustering:* Separation of data into clusters (idea is, **closer** data points are more **similar**. Reference: Type 2 diabetes mellitus prediction model based on data mining, Wu et al. 2018)
  - **Reinforcement**
    - Trial and error by cumulation of rewards (The robot navigation a matrix with some square on fire)
  - **Feature Selection** (a good way to increase savings in computational power cost + better prediction accuracy)
    - Only selecting features from a feature space that are most relevant to our goal
    - Advantages include shorter analysis time, better visualization and better prediction accuracy.
    - Either select subset by:
      - Assessment based on general characteristics of data, or
      - Machine learning algorithm that select subsets and tests them to find the optimal features (wrapper method).
- Def of Diabetes Mellitus (already covered in the past)

4.0 Methodology

Databases used: PubMed and DBLP.
The rest talks about how they searched and how many results each returned (Fig 2). (read if interested, otherwise skip)

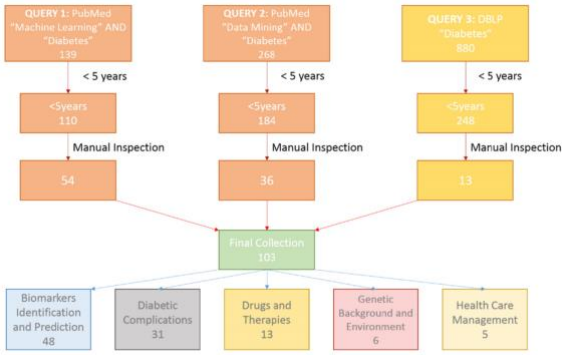

Fig. 2. Literature selection and classification process.

5.0 Results of the Survey

Biomarker identification and prediction + Prediction of DM

- What was found in this section was that using the wrapper method gives best results when identifying the predictive biomarkers
- What they also did was use the best result and add other features not included in the selected features to see if the results improved (ultimately, we can also modify our feature list to test the effects of the features we did not include in the first place)
- For high dimensional wrapping, Iterative Sure Independence Screening (ISIS) is recommended (have not had the time to look into this)

They also talk about complications such as nephropathy and cardiovascular disease when diabetes is not managed properly (wish they also talked about its management) which gives us all the more reason to want to manage diabetes.

In terms of personalized drug therapy, they found that the following points ML and data mining has been used effectively in:

- Recommendation and efficacy improvement of medication
- Prediction and suggestion of more personalized medications
- Design of more effective blood glucose lowering factors
- Improvements to insulin planning and dosage,
- Implementation of drug administration in a more specific manner.

Healthcare management

- Diabetes prevalence in 2010 → 2.8%
- Predicted to reach 4.4% by 2030
    - Around 330 million (self-management becomes very important)

5.1 Conclusion

1. Studies covered in this paper used 85% supervised learning and 15% unsupervised learning
2. Support vector machines (SVMs) have proved most successful in classification accuracy (in our case can be glucose is too high or too low, action is need or action is not needed, etc). Second and third goes to Artificial Neural Networks (ANN) and Decision Trees

(DT). **NOTE:** Although SVMs are most successful, there may be other methods that would fit our data sets… once we have one.

3. DM Treatment section is something that I will look into more as it will provide a lot more info with regards to the management/recommendation options