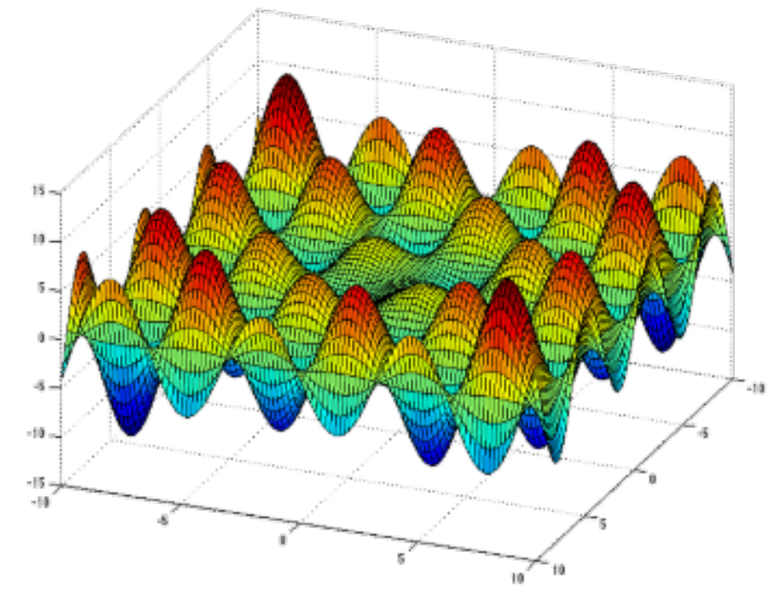
Shrinking Big Data with Swarm Intelligence

Big Data problem:

In machine learning there has been an immense increase in high-dimensional data with thousands of features. This applies to many types of data including gene expression, images, and web mining [1]. These high-dimensional datasets raise several challenges for classification tasks, reffered to as the "curse of dimensionality".

To mitigate the complexity of building classifiers for these datasets, a process of feature selection can be very effective. It is particularly relevent for datasets with high levels of dimensionality as they typically contain a large number of irrelevant or redundant features.

Unfortunately, many state of the art feature selection methods do not scale gracefully with these extremely high-dimensional datasets. Here a new method is proposed which demonstrates strong feature discrimination ability on datasets containing up to 10,000 features



A non-convex search space with only 3 dimensions

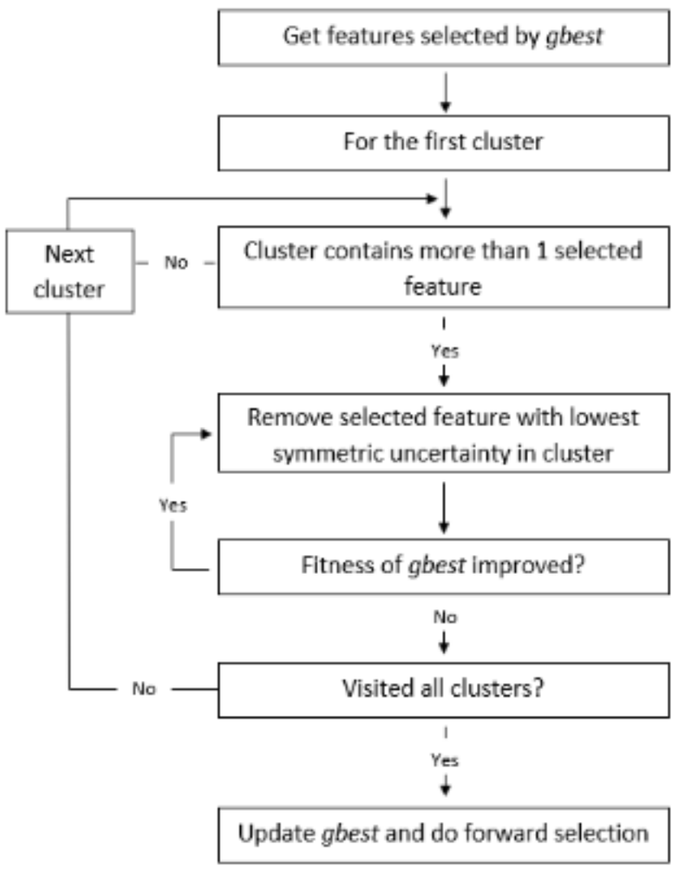
New Method:

Particle Swarm Optimisation (PSO) is an evolutionary computation technique which draws inspiration from the the social behaviours of animals in groups. In PSO, a swarm of candidate solutions are created and allowed to evolve. The evolution is guided by the best solution found by each individual particle (personal best/pbest) and the best solution found by the swarm (global best/gbest).

Recent research has shown that combining PSO with a statistical optimisation of these guiding values can be very powerful [2,3] . The new method seeks to utilise the demonstrated strength of these hybrid approaches. **At each PSO iteration, the new method applies a backward elimination and a forward selection process on the global best.** This helps to optimise the global best, refine the search, and guide the swarm.

Before the evolution begins, the algorithm performs two calculations which are used to determine features to add or remove during the evolution. Firstly features are clustered into similar groups according to their correlation levels. Then, the Symmetric uncertainty (SU) of each feature is calculated. SU is an entropy-based measure which is detailed in [1]. This information is relatively cheap to calculate, meaning it can be used efficiently on very high-dimensional datasets. **The clusters and SU values guide the backward elimination and forward selection towards good features to add/remove.**

Backward Elimination flowchart



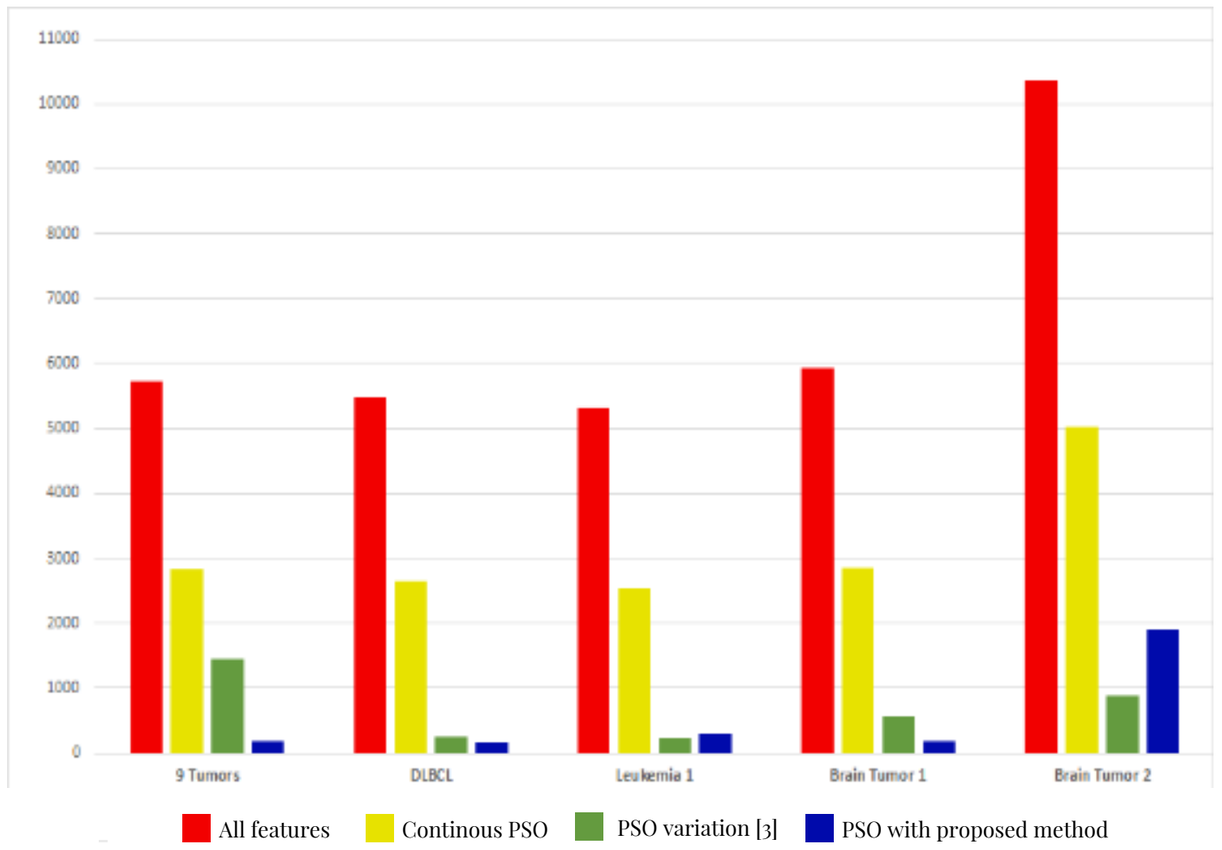
Results:

The new method is very competitive with other PSO algorithms. The graph displayed shows the feature reduction of the new method against all features and a normal PSO. It also includes a recently published PSO algorithm that utilises a similar iterative improvement methodology on the personal best and includes a gbest resetting mechanism. **The new method consistently finds greater reductions than PSO alone for all datasets.**

For other important elements of feature selection such as classification accuracy and run-time, the new method usually maintains but often slightly decreases classification accuracy relative to normal PSO. **The fitness of the generated solution is typically the same compared with regular PSO. It is, however, usually faster. A comparison of forward and backward elimation against backward elimination alone was also conducted. It found that using both processess resulted in similar fitness and classification accuracy, but smaller feature subsets and more variable run-times compared with only using backward elimination.** These results are heavily dependent on the correlation value used for clustering.

It is worth noting that the method becomes significantly less effective on the largest dataset, Brain Tumor 2. This invites further investigation into how this idea can be scaled for datasets greater than 10,000 features.

Feature reduction between methods



Poster created and research conducted by Luke Johnson under the supervision of Bing Xue and Mengjie Zhang (School of Engineering and Computer Science). Funded as part of the VUW summer research scholarship program 2018.

References:

[1] B. Tran, B. Xue, and M. Zhang, "Using feature clustering for gp-based feature construction on high-dimensional data," in European Conference on Genetic Programming, pp. 210–226, Springer, 2017
[2] H. B. Nguyen, B. Xue, I. Liu, and M. Zhang, "Filter based backward elimination in wrapper based pso for feature selection in classification," in Evolutionary Computation (CEC), 2014 IEEE Congress on, pp. 3111–3118, IEEE, 2014.
[3] B. Tran, B. Xue, M. Zhang, and S. Nguyen, "Investigation on particle swarm optimisation for feature selection on high-dimensional data: Local search and selection bias," Connection Science, vol. 28, no. 3, pp. 270–294, 2016.

TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON