Interactive Machine Learning Interface

Lilong Jiang(jiang.573)

1 Introduction & Motivation

Interactive visualization analysis is an importan area in information visualization, databases and HCI. Previous works and commercial products focus on data analysis area, like how to dynamically fitering datasets, zooming and panning, brushing and linking. Recently, more works focus on the machine learning and data mining algorithms. There several motivations for this trend. Firstly in many cases, there exists parameter tuning stage for machine learning algorithms and whether the point belongs to something (e.g. cluster, outlier) really depends on what metric to use, what threshold is set, etc. Secondly, interactive visualization provides much better usuability than traditional programming method. And non-experts can easily use this interfae. Thirdly, it is often common for users to verify results because of multiple reasons, e.g. overfitting and underfitting. In summary, it is necessary to provide an usable interactive interface allowing users to adjust machine learning algorithms on-line.

2 Work

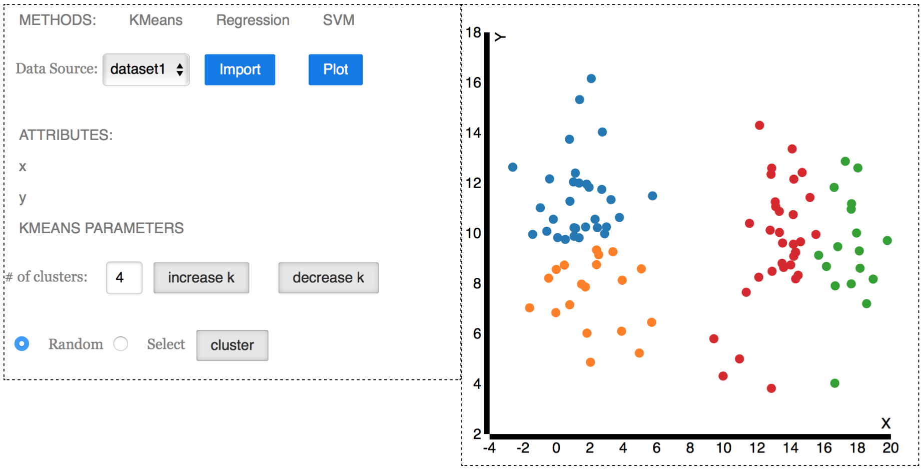
The overall space is divided into two parts: the left part shows what kind of machine learning method to run and parameters associated with the algorithm. Also users can choose different dataset and it will show the attributes of current manipulated dataset. In general, we implement three machine learning algorithms in the frontend using javascript: K-means (Figure **??**), linear regression (Figure **??**) and support vector machine (SVM) (Figure **??**). Table **??** summarizes operations that users can perform with our interface.

3 Performance Experiments

In this section, we focus on performance experiments since in the interactive environment, it is necessary to keep an interactive performance.
Computer Configuration:
Dataset: For K-means, we sample from 2 two-dimensional multi-variant datasets and run K-means for 100 interations. For SVM, we use Skin Segmentation Data Set from UCI [6]. There are totally 3 numerical attributes and 2 classes. We randomly sample from both classes.

algorithms	operations
k-means	(1) Select initial centers or generate centers randomly (2) Adjust number of clusters on-line (3) Colors shows different clusters
linear regression	(1) Adjust function's intercept by moving the line (2) Adjust function's slope and intercept by dragging end points (3) Dynamically show the function and square loss
svm	(1) Select kernel (2) Select which features to use by dragging into or out space (3) Dynamically show accuracy, precision, recall and confustion matrix

Table 1: Operations


Data Source: dataset1

# of clusters: 4

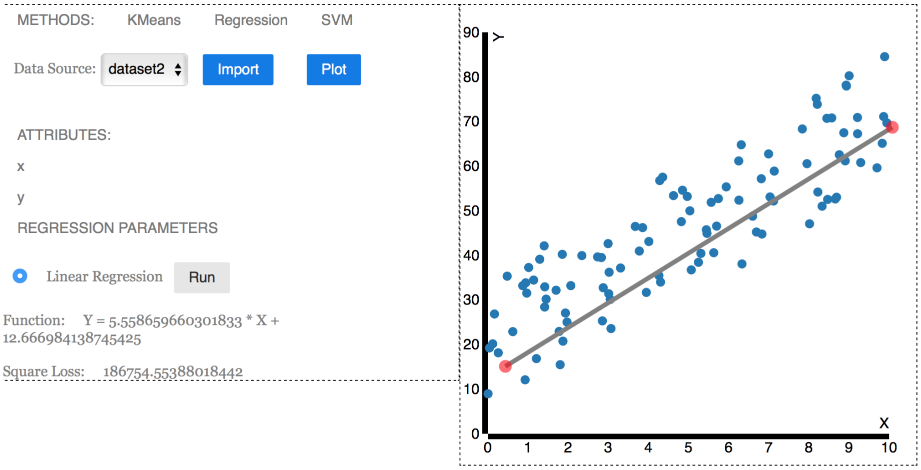
● Random ○ Select

Figure 1: K-means

Results: If we use 1ms as the interactive performance threshold, Figure **??** shows; Figure **??** shows that for SVM, it can only sustain less than 3 points. Optimization is necessary for maintain an interactive performance.

4 Limitations & Conclusions & Future Work

In this project, we focus on usability point. There are several limitations for this project. The first is interactive performance when the number of points is large. One method is to run machine learning algorithm in the backend and once it is done, then send it to frontend. However, users need to pay the internet cost. Another method is to use an incremental method, which update results incrementally as more data is processed. Works including [1] adopts this



Data Source: dataset2

● Linear Regression

Function: Y = 5.558659660301833 * X + 12.666984138745425

Square Loss: 186754.55388018442

Figure 2: Linear Regression

method.

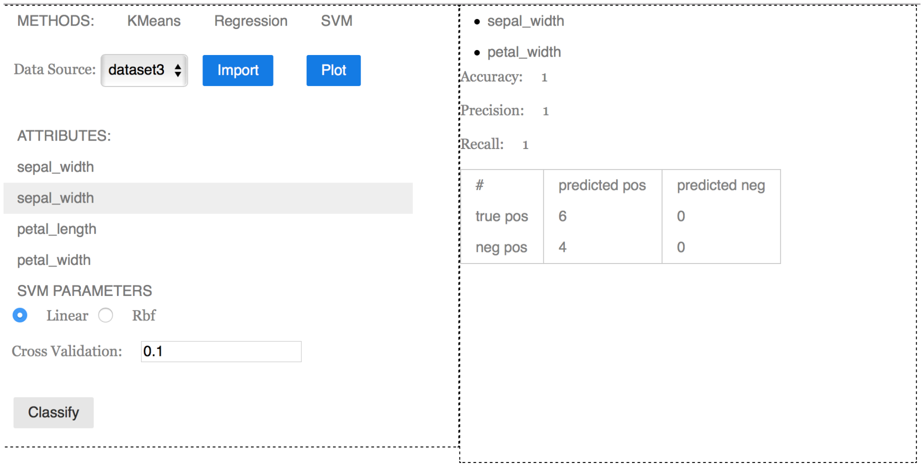
The second is project issue for high-dimensional dataset. *Projection* is a technique that maps high-dimensional dataset into a small set of dimensions. Some works use dimension reduction, that create latent dimensions that summarize dataset. For example, Principal Component Analysis (PCA) [4], Multidimensional Scaling (MDS) [5]. This method can reveal hidden variables. However, the generated dimensions are usually less intuitive to users. Another method is called feature selection, which selects subset of features to explore. Interactive feature selection is often used to identify relationship between features [3,8] and aid user remove redundant features and choose appropriate features. Recently, there is a work trying to allow users to craft their own projection function [2]. All these methods can be easily incorporated into our project.

In conclusion, we build an interactive machine learning interface which allows users to visualize dataset, run different kind of machine leanring algorithms, and compare results. Incorporating high-dimensional visualizations, expanding linear regression into other kind of regression types, and allowing supporting large datasets are ideal future work.

References

[1] Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. Vizdom: interactive analytics through pen and touch. *Proceedings of the VLDB Endowment*, 8(12):2024–2027, 2015.

[2] Michael Gleicher. Explainers: Expert explorations with crafted projections. *TVCG*, 2013.



Data Source: dataset3

ATTRIBUTES:

sepal_width

sepal_width

petal_length

petal_width

SVM PARAMETERS

● Linear ○ Rbf

Cross Validation: 0.1

- sepal_width
- petal_width

Accuracy: 1

Precision: 1

Recall: 1

#	predicted pos	predicted neg
true pos	6	0
neg pos	4	0

Figure 3: SVM

[3] Diansheng Guo. Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization*, 2003.

[4] Ian Jolliffe. *Principal component analysis.* 2002.

[5] A Mead. Review of the development of multidimensional scaling methods. *The Statistician*, 1992.

[6] Abhinav Dhall Rajen Bhatt. Skin segmentation dataset. *UCI Machine Learning Repository.*

[7] Jack W Smith, JE Everhart, WC Dickson, WC Knowler, and RS Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

[8] Jing Yang et al. Value and relation display for interactive exploration of high dimensional datasets. In *INFOVIS*, 2004.

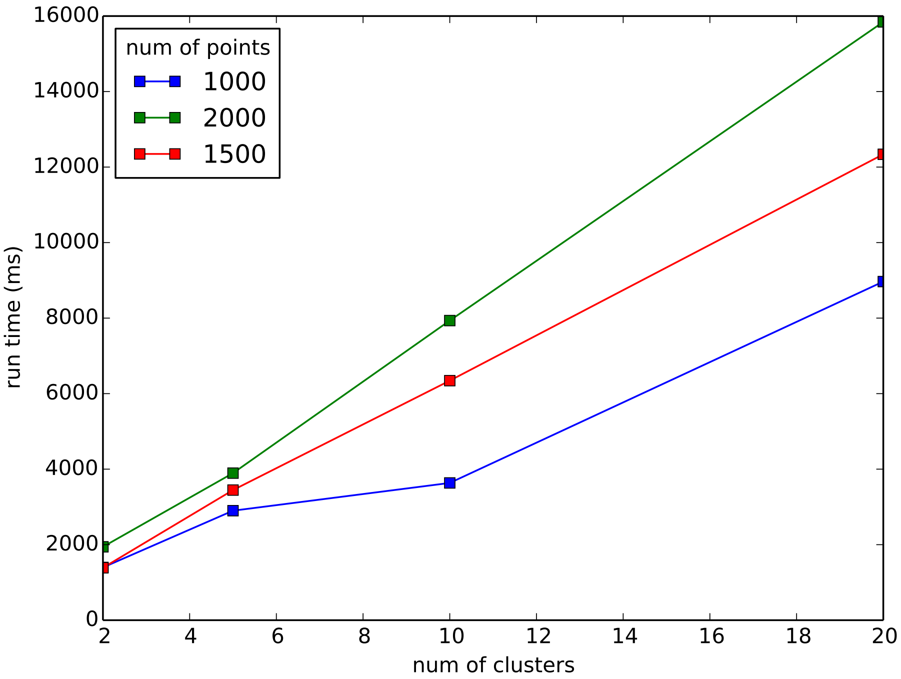


Figure 4: K-means Performance

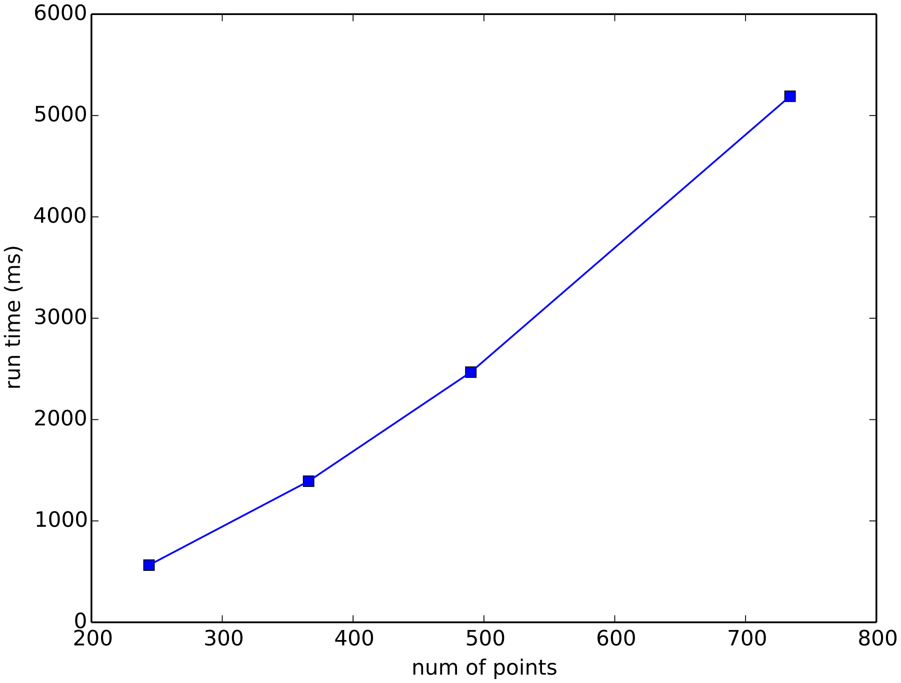


Figure 5: SVM Performance