Multi-Label Classification with k-Nearest Neighbor

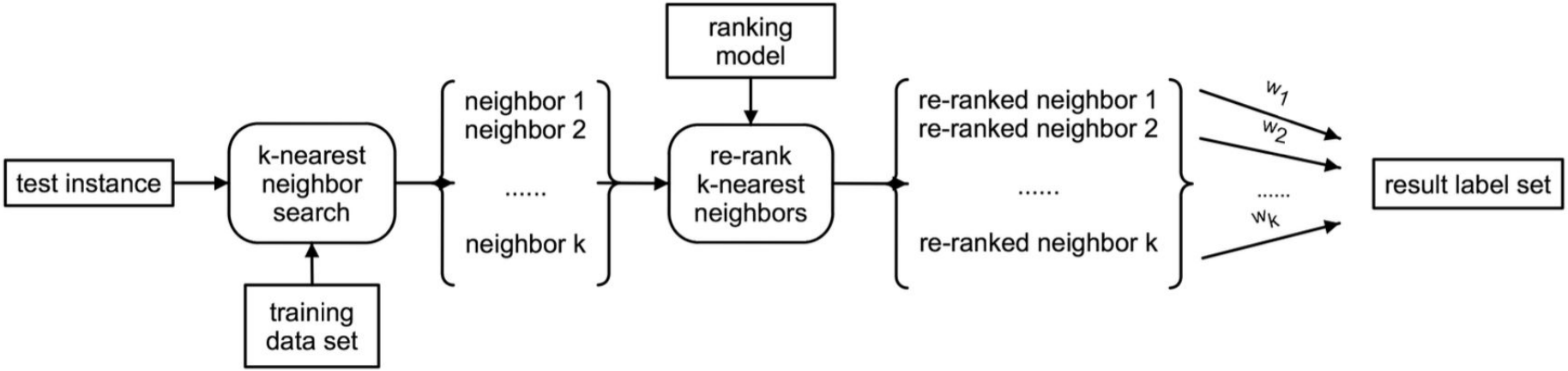
—

By David Liu, Branley Mmasi

How do we classify examples with multiple labels?

- Images
- Enzymes
- Songs
- Food

Chiang et al. 2012 paper



A ranking based kNN

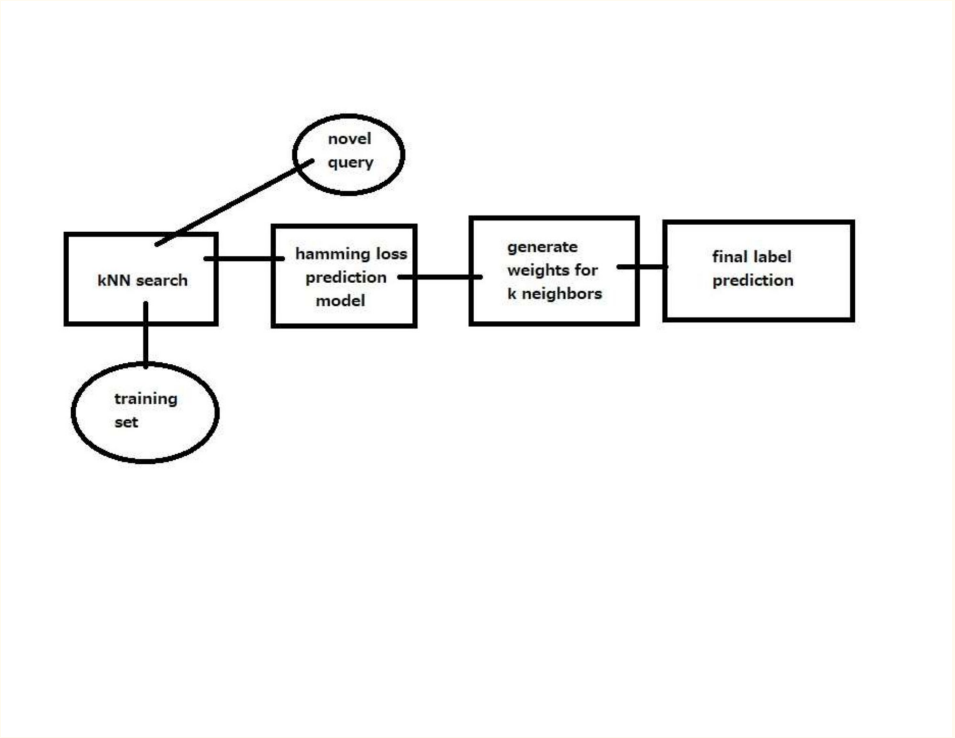
Limitations

- Ranking model
- Time

Methodology

Our Methodology

- Hamming Loss - % of labels different between two label sets
- Use LR model to predict the Hamming loss
- Weigh neighbors based on predicted Hamming Loss
- Weighted voting of final labels



Results & Analysis

Dataset

Multi-label classification of enzyme substrates

- Classifying substrates based on EC-classes
- Able to predict behaviour of enzymes on similarly structured molecules
- 1039 molecules, 196 features, 6 possible labels
- Pre-processing

Results

- 60/40 train-test split over multiple k
- kNN runtime scales very poorly with larger sets
- Runtime became extremely long

Table 1. Comparison of Hamming loss of traditional kNN and modified kNN

K	TRADITIONAL	MODIFIED
3	0.34	0.54
5	0.31	0.58
7	0.31	0.60
11	0.31	0.34
15	0.29	0.33
25	0.28	0.28

Analysis

- Poor results for lower k's
- Catches up to standard kNN for higher k
- Still poor performance

Table 1. Comparison of Hamming loss of traditional kNN and modified kNN

K	TRADITIONAL	MODIFIED
3	0.34	0.54
5	0.31	0.58
7	0.31	0.60
11	0.31	0.34
15	0.29	0.33
25	0.28	0.28

Discussions

- Why was our model performance bad?
- Hamming Loss as a direct weight strategy is not sophisticated enough to produce results better than standard kNN
- Small dataset with small label set
- Future work?

Thank you!

References

Chiang, Tsung-Hsien, Lo, Hung-Yi, and Lin, Shou-De. A ranking-based knn approach for multi-label classification. 2012.

Srivastava, Gopal N. Multi-label classification of enzyme substrates.