Milestone 4

Andres Calderon
acald013@ucr.edu

March 19, 2016

1 Introduction

This report describes the wrap up of the final project in the course. The main goal of the project is to perform a reliability analysis of a machine learning algorithm (kNN). This last report aims to elaborate on the final thoughts in the analysis of the impact of error injection during the distance calculation of kNN. The main goal of the report is to discuss a second version of the formula proposed in milestone 3 to describe the behavior of the error rate.

2 Formal reasoning.

In order to understand in a better way how error rate affect the accuracy of kNN we face the following situation: if we have a training set with N points and a p percentage of them have been affected by error injection during the distance calculation, how probable will we have a misclassification? To answer this question I think we have to deal with two aspect for each given new point x_i to be classified: (1) Which is the probability that the closest point x_j is affected by error injection and (2) Which is the probability that x_j is the same class than x_i and we can fail to classify it correctly.

2.1 Closest point probability

In [1], Song et al define informativeness as a function of $Pr(x_j|Q = x_i)$ where Q denotes the query point x_i. They states that this term can be interpreted as the likelihood that point x_j is close to the Q. Then, they follow to explain that in order to achieve higher probability when two points are close each other, Pr should be a function inverse to the distance between them. In our case, we are using Euclidean distance, so we can use equation 1 to refer to this probability.

$$Pr(x_j|Q = x_i) = exp(-\sqrt{(x_i - x_j)^2}) \tag{1}$$

2.2 Same class probability

As it was stated in milestone 3, the probability of two points x_j and x_i belong to the same class depends on the number of classes (C) and the size of the training set (N) and simply find $\frac{N}{C}$. However, in a more general sense, we have to consider that the number of points in each class is not uniformly distributed. Given x_j, the probability of x_i to belong to the same class is defined in equation 2 as:

$$\frac{N_{x_j}}{N} \tag{2}$$

where N_{x_j} represents the number of points in the same class of x_j. Indeed, this notion is also used in [1] as a balancing factor.

3 Conclusion

In overall, we can model the impact of error injection during the distance calculation as function of p (probability of error injection) combining equations 1 and 2.

$$Err(p) = \frac{N_{x_j}}{N} * p * exp(-\sqrt{(x_i - x_j)^2}) \tag{3}$$

References

[1] Yang Song, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles. IKNN: Informative K-Nearest Neighbor Classification. In *PKDD 2007*. Springer Verlag, September 2007.