Similarity Learning in Nearest Neighbor and Application to Information Retrieval

Ali Mustafa Qamar and Eric Gaussier
LIG (Laboratoire d'Informatique de Grenoble)
Université Joseph Fourier
ali-mustafa.qamar@imag.fr, eric.gaussier@imag.fr

Abstract

Many people have tried to learn Mahanalobis distance metric in kNN classification by considering the geometry of the space containing examples. However, similarity may have an edge specially while dealing with text e.g. Information Retrieval. We have proposed an online algorithm, SiLA (Similarity learning algorithm) where the aim is to learn a similarity metric (e.g. cosine measure, Dice and Jaccard coefficients) and its variation eSiLA where we project the matrix learnt onto the cone of positive, semi-definite matrices. Two incremental algorithms have been developed; one based on standard kNN rule while the other one is its symmetric version. SiLA can be used in Information Retrieval where the performance can be improved by using user feedback.

Keywords: Similarity learning, kNN, Information Retrieval, Machine Learning

1. INTRODUCTION

Many works have tried to improve the kNN algorithm by considering the geometry of the space containing examples. Most of these works learn Mahanalobis distance metric, a variation of Euclidean distance. The Mahanalobis distance between two objects \( x \) and \( y \) is given by:

\[
d_{A}(x, y) = \sqrt{(x - y)^T M (x - y)}
\]

However, similarity should be preferred over distance in many practical situations, e.g. text classification, information retrieval as was proved by our results on different datasets [4].

2. PROBLEM FORMULATION

The aim here, is to learn a similarity metric for kNN algorithm. Let \( x \) and \( y \) be two examples in \( \mathbb{R}^p \).

We consider similarity functions of the form:

\[
s_{A}(x, y) = \frac{x^T A y}{N(x, y)}
\]  

(1)

where \( A \) is a \( (p \times p) \) matrix (symmetric or asymmetric) and \( N(x, y) \) is a normalization which depends on \( x \) and \( y \). Equation 1 generalizes several different similarity functions (cosine measure (by replacing matrix \( A \) with the identity one), Dice coefficient, Jaccard coefficient)

3. SiLA (SIMILARITY LEARNING ALGORITHM) AND ESiLA

SiLA is based on voted perceptron developed by [3] and used by [2]. Figure 1 illustrates the notion of separability we are considering. In 1(a), the input point is separated, with \( k = 3 \), whereas it is not in 1(b) as it is closer both to points from the class it belongs as well as differently labeled examples. The separation does not need to take place in the original input space, but rather on the space induced by the metric defined by \( A \). 1(c) illustrates what we are aiming at: moving the target points closer to the input point, while pushing away differently labeled examples.

When an input example is not separated from examples belonging to different classes, the current \( A \) matrix is updated by the difference between the coordinates of the target neighbors and the
FIGURE 1: In (a) the input point is separated with \( k = 3 \), whereas it is not in (b). (c) illustrates the process we aim at: moving target points closer to the input point, while pushing away differently labeled examples.

closely differently labeled examples. This update corresponds to the standard perceptron update. However, in case of correct classification, the weight corresponding to the current \( A \) matrix is increased by 1. The weighted matrices are then used to classify unseen test examples. Two prediction rules have been developed: one is based on standard kNN (kNN-A) while the other one considers same number of examples in different classes (SkNN-A). It has been proved that the no of mistakes are bounded and the algorithms can generalize well beyond training examples.

In eSiLA, \( A \) matrix is orthogonally projected on the cone of positive semi-definite matrices inspired from POLA [5]. This projection guaranties convergence and generalization of the algorithm. However, eSiLA is similar to SiLA in all other aspects.

4. EXPERIMENTAL VALIDATION

SiLA and eSiLA were tested on eight standard test collections, namely Balance, Wine, Iris, Ionosphere, Soybean, Glass, Pima and 20-Newsgroups, where first seven were obtained from UCI [1]. 5-fold nested cross validation was used to learn the matrix \( A \) for the UCI datasets owing to their small size. We used two prediction rules for the experiments. In the first one, the classification is based on the \( k \) nearest neighbors (kNN rule) while the second one (SkNN) is based on the difference of similarity between \( k \) nearest neighbors from the same class and \( k \) from differently labeled classes. The results, given in table 1 demonstrate that similarity should be preferred over distance on non-textual collections also like Balance (gain of 7.6%), Wine (gain of 8%), Iris (gain of 0.9%), Ionosphere (gain of 1.7%) etc.

The results further show that the algorithm eSiLA performs better as compared to standard kNN on Wine (gain of 1.9% with SkNN-A), Ionosphere (gain of 1.9% with both kNN-A and SkNN-A) and Pima (gain of 0.8% with SkNN-A). All methods have comparative performance on Soybean and Glass (since base accuracy is already too high with just using cosine). SiLA improved the base results (with kNN-cos) on Balance, Wine, Ionosphere and News.

eSiLA performs better than SiLA on Wine (gain of 1.2%), while they are comparable on Ionosphere, Pima and Soybean.
5. SIMILARITY LEARNING AND INFORMATION RETRIEVAL

SiLA or eSiLA can be used in Information Retrieval, where the matrices can be tuned by incorporating user feedback. The similarity is calculated between a query $q$ and a document $d$. The basic theme rests the same: try to bring target documents (documents relevant to $q$) closer to $q$ while pushing away irrelevant documents which in turn yields matrix $A$. The top rated documents are presented to the user who can then change the order. This order is learnt by updating the weights in the same way as in SiLA and eSiLA.

6. CONCLUSION

In this paper, we have discussed an approach to apply similarity learning algorithms namely, SiLA and eSiLA, in the context of Information Retrieval. Both SiLA and eSiLA not only outperform standard euclidean distance on some collections, but also improve the base results using standard cosine. We have described a mechanism to incorporate user feedback in order to update the similarity matrix.

REFERENCES