On Enhancing the performance of nearest neighbour classifiers using Hassanat distance metric

Advanced Methods in Applied Statistics

Noah T. Bloss
sjb389@alumni.ku.dk
The University of Copenhagen
The Niels Bohr Institute

Frederik V. S. S. Hansen
cjb924@alumni.ku.dk
The University of Copenhagen
The Niels Bohr Institute

INTRODUCTION

The following paper is a summary of the article: "On enhancing the performance of nearest neighbour classifiers using Hassanat distance metric" by Alkasassbeh, M., et al. published in Canadian Journal of Pure and Applied Sciences (CJPAS), volume 9, issue 1, Feb 2015 [1].

The upcoming investigates the influence of different distance metrics on the accuracy of a classifier. The metrics used to determine the distance between two points are the Manhatten and the Hassanat distance metric. The K-Nearest Neighbour algorithm (KNN) is the classifier that is used. The traditional model is very simple and is additionally used with two enhanced variations (IINC & ENN) of the algorithm to compare the two metrics. The classifier is executed on the same data sets to conclude which metric provides the best accuracy for the algorithm.

K-NEAREST NEIGHBOUR CLASSIFIER

KNN is known as a simple however very effective Classifier. A classic KNN algorithm is a supervised algorithm meaning that it uses labelled data sets to train the algorithm to classify similar unknown data.

Generally a KNN algorithm tries to find the nearest data point or points also known as nearest neighbours noted as k in a data set according to an unlabeled data point. When the algorithm have located the nearest neighbours it can then classify which class the unlabelled data point belongs to. In the traditional model the class that is represented most in the k-nearest neighbours defines the tag of the query point. The choice of the number of nearest neighbours is already a question in the traditional model. The arising problem by selecting a k is that the accuracy of the algorithm can decline, i.e. if k is chosen too low.

Distance metrics

The way of defining the nearest neighbour in a KNN can vary depending on which distance metric one chooses to use. One of the most common distance metrics used for KNN is the euclidean distance metric which describes the shortest distance between two data points:

$$D_{Euclid}\left(x,x'\right) = \sqrt{\sum_{d} \left|x_d - x'_d\right|^2} \tag{1}$$

Another commonly used distance metric used for KNN's is the Manhattan distance metric which will be used to compare with the Hassanat distance metric later on.

$$D_{Manhattan}\left(x,x'\right) = \sum_{d} \left|x_d - x'_d\right|^2 \tag{2}$$

The Hassanat distance metric differs from most distance metrics as it is not affected by noise, outliers and different data scale. The major advantage is that it is bounded by the range $[0,1[$ for each feature in the vector of a given data set. The Hassanat distance can be written as:

$$D_{Hassanat}(A,B) = \sum_{i=1}^{m} \left(D\left(A_i,B_i\right)\right) \tag{3}$$

Where A and B describes two vectors with a size m consisting of the different data points. It then have the characteristics for the similarity function between any two points as

$$D\left(A_i,B_i\right) = \begin{cases} 1 - \frac{1+min(A_i,B_i)}{1+max(A_i,B_i)} & min\left(A_i,B_i\right) \geq 0 \\ 1 - \frac{1+min(A_i,B_i)+|min(A_i,B_i)|}{1+max(A_i,B_i)+|min(A_i,B_i)|} & min\left(A_i,B_i\right) < 0 \end{cases} \tag{4}$$

The Hassanat distance metric is bounded by the interval [0,1[as it only reaches 1 when the maximum value approaches infinity or minimum value approaches minus infinity as shown Equation 5 and Figure 1. Meaning that the closer to 0 the value is the more similar the points are and the closer to 1 the more dissimilar they are.

$$\lim_{max(A_i,B_i)\to\infty} \left(D\left(A_i,B_i\right)\right) = \lim_{min(A_i,B_i)\to-\infty} \left(D\left(A_i,B_i\right)\right) = 1 \tag{5}$$

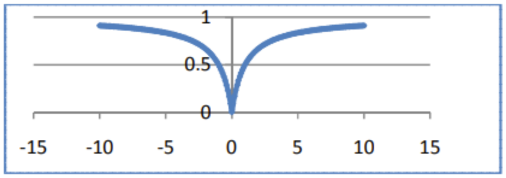

Figure 1: Plot of the Hassanat distance metric between the points 0 and n=[-10,10]. Figure taken from [1]

.

ENHANCED MODELS

Inverted Indexes of Neighbours Classifier (IINC)

With the IINC classifier vanishes the question, what the optimal k for the best result may be. This modification takes all training points into account and is in some circles denoted as the best choice for the number of nearest neighbours. In this type of KNN-classifier the distances from the point to the example points are playing a crucial role. The nearest training point has the biggest influence and the furthest away training point has the least influence on what class the point is assigned to. This implies, that the influence is proportional to the distance between the points. The first step in the algorithm is to determine the distances to each point from the query point and to sort them beginning with the smallest. With Equation 6 the summation of *the inverted indexes* is calculated for each class.

$$S_c = \sum_{i=1(c)}^{L_c} \frac{1}{i} \tag{6}$$

L_c represents the number of training points for each class c, and the index i represents the placement in the list of distances. The next step in this algorithm is to calculate the probability of each class, which is determined as follows:

$$P(x|c) = \frac{S_c}{S} \tag{7}$$

where $S = \sum_{i=1}^{N} \frac{1}{i}$ and N is the number of points in the training set. In the end the class with the highest probability is assigned to the query point.

Ensemble Nearest Neighbour Classifier (ENN)

In this algorithm the traditional KNN classifier is used, but here it is exploited several times in the process of classifying one point. The ENN executes the traditional method from $k = 1$ to $k = \sqrt{n}$, where n denotes the number of training points. Furthermore, the algorithm performs the process only with an odd number of k to increase the speed by avoiding that two different classes having the same number of votes. As in the IINC the first step is to determine the distances from the point to the training examples and order them starting with the smallest. The weight of each point is defined as:

$$w(k) = \frac{1}{log_2(1+k)} \tag{8}$$

with k representing the position in the ordered list of distances. Equation 8 implies, the further away the training point is located from the query point, the less influence does it have in the decision of the class. With these weights it is possible to calculate the weighted sum (WS), which is defined as:

$$WS_c = \sum_{k=1}^{\sqrt{n}} \sum_{i=1}^{k} \begin{cases} w(i), A_i = c \\ 0, \text{otherwise} \end{cases} \quad , k = k + 2 \tag{9}$$

where, A_i represents the list with the ordered example points. The inner sum in Equation 9 calculates the weights for each classifier and the outer sum represents the KNN classifier for every odd k. The point is included in the class with the highest WS:

$$class = \underset{cWS_c}{\text{argmax}} \tag{10}$$

COMPARISON OF THE DISTANCE METRICS

28 different data sets from the UCI Machine Learning Repository has been used for the evaluation of the efficiency of the Hassanat distance metric. The data sets has all been split up with 70% going to train the algorithms and 30% to test them. This has then been done 10 times to randomize the data examples as much as possible. All results presented further on will be the mean of the 10 tests.

From Table 1 which shows the average accuracy of different nearest neighbour classifiers using the Manhattan distance metric. It can be seen that the IINC and ENN classifiers on average performs better than the more classic nearest neigbour classifiers.

Algorithm	1NN	3NN	5NN	7NN	9NN	$\sqrt{n}$NN	IINC	ENN
Average	0.81	0.82	0.82	0.82	0.82	0.80	0.83	0.83

Table 1: Accuracy of nearest neighbour classifiers using Manhattan distance metric.

The nearest neighbour classifiers can then be further enhanced by using the Hassanat distance metric which can be seen in table Table 2.

Algorithm	1NN	3NN	5NN	7NN	9NN	$\sqrt{n}$NN	IINC	ENN
Average	0.84	0.85	0.86	0.85	0.85	0.84	0.87	0.87

Table 2: Accuracy of nearest neighbour classifiers using Hassanat distance metric.

From Table 2 it is clear that there is an increase in average accuracy of the different nearest neighbour classifiers using the Hassanat distance metric instead of the Manhattan classifier. This can also be seen in table Table 3 which shows the difference between the tests from table Table 1 and Table 2. This is used to confirm that the Hassanat distance metric on average enhances the performance of different nearest neighbour classifiers as the performance increases from 2.9% to 3.8%. It can also be confirmed that the INNC and ENN clasifiers performed well with the Hassanat distance metric as they saw an increase in performance of 3.1% and 3.3%, respectively.

Algorithm	1NN	3NN	5NN	7NN	9NN	$\sqrt{n}$NN	IINC	ENN
Average	0.03	0.03	0.04	0.04	0.04	0.04	0.03	0.03

Table 3: Increase in accuracy of each nearest neighbour classifiers after applying Hassanat distance metric instead of Manhattand distance metric.

In the article it is also noted that the performance of the different classifiers did not always improve. Some of the data sets see a decrease in performance where others see an increase. This is not something to worry about as it is well known that there is no perfect algorithm which can solve all problems. However from the produced results in the article it can be obtained, that the IINC and ENN in combination with the Hassanat distance metric can be used to enhance nearest neighbour classifiers.

CONCLUSION

Even if the Manhatten distance metric are commonly used in the KNN classifier, the Euclidean distance is also a very established metric to measure the distance. For an improved comparison, one could have wished that this metric also would have been included into the comparison of the precision of the KNN classifier. Nevertheless, it can still be concluded that there is an increase in precision by using the Hassanat instead of the Manhatten distance metric. In addition, the two enhanced models (IINC & ENN) exhibiting already an improvement in the accuracy of the traditional K-Nearest Neighbour classifier.

REFERENCES

[1] Alkasassbeh, M., et al. (2015). *On enhancing the performance of nearest neighbour classifiers using Hassanat distance metric.* Canadian Journal of Pure and Applied Sciences (CJPAS). volume 9, issue 1, Feb 2015. doi: https://arxiv.org/abs/1501.00687