A review of k Nearest Neighbor Classifier

The Nearest Neighbor (k-NN) decision rule assigns to an unclassified sample point the classification of the nearest of a set of previously classified points (Cover and Hart 1967). It is an exemplar-based model of categorization that improves its performance adding new points to the database set and works once a query is made to the system so no learning time is required prior to use it.

Cover and Hart presented the Nearest Neighbor Pattern Classification paper in the *IEEE Transactions in Information Theory* in January 1967. This technique was developed from the need to perform analysis for unknown probability densities. Since then, it has been used for statistical estimation, pattern recognition, classification and interpretation of information in many different areas due to it simple implementation, good performance and no training time. It is an exemplar-based model of categorization – or "learning by examples" –, which identifies objects by their similarity to one or more of the stores examples (Fujinaga 1996). It is considered a *lazy learning* method because the calculation of the distance from a new *exemplar* to an *exemplar set* is delayed until a query is made to the system, thus it cannot simulate sophisticated logical relationships between features, but require essentially no training time (McKay 2004a).

(Moore 91) specifies in the following manner the nearest neighbor: given a multi-dimensional space (D,R), an *exemplar set* E and a target vector d, the *nearest neighbor* of d is any exemplar $(d',r') \in E$, such that *none-nearer*(E,d,d'). *None-nearer* is defined as

$$\textit{None nearer}(E,d,d') \Leftrightarrow \forall (d'',r'') \in E \; |d - d'| \leq |d - d''|$$

The Euclidean distance metric is

$$|d - d'| = \sqrt{\sum_{i=1}^{k_d} (d_i - d'_i)^2}$$

Where d_i is the *i*th component of vector d.

Thus, for each new *exemplar* putted into the multidimensional space it can be easily computed the distance to the known *exemplar set.* Comparing the values obtained and assigning a value for *k*, we can obtain the k-nearest neighbors. By this way, we can classify the queried *exemplar* to the class of its neighbors.

The accuracy in the classification of an *exemplar* into its real class could be improved adding new *trained* samples to the *exemplar set* database. In addition, the recognition could be enhanced transforming - stretching or widening – the feature space. However this last technique is difficult to implement in high-dimension environments (Fujinaga 1996).

Also, it is possible to solve regression problems using kNN. Given a set of independent variables, it is possible to predict the values of dependant variables finding its values using an increment in the factor of *k* and average their results. In other words, the *y* value of the query point *x* is taken to be the average of the outcomes of its k nearest neighbors.

In terms of performance, the trivial *linear* implementation of the algorithm (i.e. compare the vector of the exemplar to all the other vectors to determine its neighbors) is a slow process. Other methods are used in order to restructure the exemplar-set in other way or complement the learning process. The former could be achieved with *kd-tree*, which is a data structure for storing a finite set of points from a k-dimensional space, while for the latter *genetic algorithms* (Fujinaga 1996), *support vector machines* (Zhang 2006), *neural networks* (McKay 2004a) and other continuous-evolving methods to enhance its speed and accuracy (Younes 2008) (Jiansheng 2009)

Bibliography

Cover, T. and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 13(1): 21-7.

Cui, B., Shen, J., Cong, G., Shen, H. T., and Yu, C. 2006. Exploring composite acoustic features for efficient music similarity query. In *Proceedings of the 14th Annual ACM international Conference on Multimedia.*

Fujinaga, I. 1996. Exemplar-based learning in adaptive optical music recognition system. *Proceedings of the International Computer Music Conference*. 55–6.

Jiansheng, W. 2009. A Novel Artificial Neural Network Ensemble Model Based on K--Nearest Neighbor Nonparametric Estimation of Regression Function and Its Application for Rainfall Forecasting. *, International Joint Conference on Computational Sciences and Optimization*, 2:44-8.

McKay, C., and I. Fujinaga. 2004. Automatic Genre Classification Using Large High-Level Musical Feature Sets. *Proceedings of the International Conference on Music Information Retrieval.* 525-30.

Moore. A. 1991. An introductory tutorial on kd-trees. *Efficient Memory-based Learning for Robot Control.* 6-1 – 6-18.

Younes, Z., F. Aballah, and T. Denoeux (2008, August). Multi-label classification algorithm derived from k-nearest neighbor rule with label dependencies. In *Proceedings of the 16th European Signal Processing Conference*

Zhang, H., A. C. Berg, M. Maire, and J. Malik (2006). Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *In CVPR*. 2: 2126-36.