CHAPTER I

INTRODUCTION

Since its inception in the 1950s (Fix and Hodges, 1951, 1989), k-nearest neighbor (k-NN) still receives regular interest among researchers; both in the theoretical aspect and the practical aspect. Its discrimination procedure is simple but powerful and needs virtually no modification to handle multi-class problems, i.e. it just obeys the majority vote for the classes among the k nearest neighbors of the sample being considered. k-NN decision rules gained theoretical acceptance since its early age of development, Fix and Hodges (1951) developed their notions of *consistencies* between, sequences of decision functions and showed that a formulation of k-NN is consistent with a reference decision rule. Many notable points are worth mentioning in their work. They initiated the field of *nonparametric classification*, the distribution generating the examples need not be assumed to be Gaussian or any other parametric distributions. The reference decision rule mentioned in their work as the "likelihood ratio procedure" (Welch, 1939) is closely related to what is known today as the Bayes classifier. The Bayes classifier is the best classifier that will yield the lowest possible expected misclassification given that we know the distribution of the data; it will be discussed in detail later. They established that whenever the number n of available examples approaches infinity and k_n are dependent of n such that $k_n \rightarrow \infty$ and $k_n/n \rightarrow \infty$, the decision of the k_n-NN will get arbitrarily closer to that of the *likelihood ratio procedure* with high probability. For example, one may choose k_n to be $\lceil \log n/n \rceil$. Later, Cover and Hart (1967) showed that simpler rules also posses good asymptotic properties; for a fixed k the error probability of k-NN will be at most twice that of the Bayes classifier in the limit as the number of examples grows to infinity. The link between k-NN error and Bayes error provides ways to estimate the theoretical limit one can achieve. Lower bounds of the Bayes error relative to errors of modified versions of k-NN rules are also studied.

More recently, various attempts to learn a good metric to use in k-NN classification have been proposed (Goldberger et al., 2005; Weinberger et al., 2006;

Zhang et al., 2007; Chatpatanasiri et al., 2008). Most of them learn the so called "Mahalanobis distance", which can be perceived as a Euclidean distance in a linear transformation of the original vector space of examples. Several objectives had been proposed and optimized in order to find the best linear transformation, and most of the proposed objectives are formulated to be able to be solved by convex optimization or the spectral method, where the optimum is guaranteed to be global. Some of them are optimized for local minimum by gradient descent algorithms or other non-convex optimization techniques. The common goal, however, is to optimize a quantity that are related to classification performance of the k-NN; the learned metric is used in k-NN. In their experiments, k-NN with the learned metric even outperforms the current state of the art learners such as support vector machines for some datasets (Weinberger et al., 2006).

The naïve version of the k-NN algorithm is easy to implement by computing the distances from the test sample to all stored vectors, but it is computationally intensive, especially when the size of the training set is large. From the practical point of view, large scale k-NN classification scenarios face the problem of speed. Several distance measures are ideated to augment existing well known basic distances such as the Euclidean distance and the ℓ^p distances and in many cases the new distance measures outperformed existing ones in terms of classification accuracy. But more accurate distances come with their price, they usually need more time to compute. A well known example of such event is the DTW distance whose running time grows like the square of time series lengths, while the ℓ^p distance takes linear time of time series lengths. The distance measures in use nowadays may be classified exclusively into two different kinds, namely

1. subadditive distance measures: by definition a distance d is subadditive if $d(x,z) \leq d(x,y) + d(y,z)$ for every x, y, z, and the inequality is called the triangle inequality or the triangle law,

2. non-subadditive distance measures, which is the complement of the first kind.

Subadditivity is useful in avoiding the need to compute every distance when the nearest neighbor is to be searched. A simple technique (Barros et al., 1996) to prune unnecessary computation of distance between some pair of items when doing nearest neighbor queries is to select an item from the pool of candidates which will be used as the *reference* item. The distance between the reference item and each of the candidates will be computed and stored in advance. Together with the distance between the reference item and the query item, those stored values can be used to lower bound the distance of the query item from each candidate item in constant time. If the lower bound distance from the query item to candidate x is greater than the *closest so far* distance, then x can be safely abandoned without having to compute its distance from the query item. Various works that take the advantage of this fact exists (Roussopoulos et al., 1995; Barros et al., 1996; Ciaccia et al., 1997; Dohnal et al., 2003); most of the work were done by database researchers and can be used instantly if only the distance measure we use is subadditive.

Although the DTW distance cannot be lower bounded using the triangle inequality, one can compute the lower bound of the DTW distance between each pair of time series instead and such bounds can be similarly used to prune out futile computations of actual DTW distances. The best known strategy to lower bound the DTW distance is due to Keogh and Ratanamahatana (Keogh and Ratanamahatana, 2005). Their lower bound can be computed in linear time.

A few questions arise naturally. Are the theoretical results regarding the asymptotic properties of k-NN applied for every distance in use today? Which of the widely used distances is of the first kind and which is not? What is a good distance for doing k-NN classification?

It may be unfair to the first question but we will answer the second first. Some widely used distances are not subadditive; examples are DTW and Shape Context Distance (Belongie and Malik, 2000). The distances that are of the first kind are the well known Euclidean and ℓ^p metrics, and instances of the less commonly known ones are Levenshtein distance or edit distance (Levenshtein, 1966) and Edit Distance with Real Penalty (ERP) (Chen and Ng, 2004), for example.

The answer to the first question is, unfortunately, negative. All of the nice asymptotic results for k-NN require that the distance measure be either the Euclidean metric (Fix and Hodges, 1951), a norm metric (Devroye et al., 1996, chap. 5) or a metric with some assumptions (Cover and Hart, 1967). Perhaps the least restrictive result, when considering only the conditions imposed on the distance used by k-NN, is in the work of Cover and Hart (1967), where the distance has to be a metric in a separable metric space, but since a non-subadditive distance fails to be a metric in the first place, k-NN with the second kind distances does not enjoy the existing results. Whether the results can be extended to cover non-metrics is still unknown. Although this does not necessarily imply that extensions of these nice results to non-metrics are impossible, it does indicate that more work has to be done in order to justify non-metrics k-NN theoretically. More precise statement regarding these asymptotic results will be formally given in Chapter 2.

As common sense and the formalized concept of "no free lunch" suggest (Devroye et al., 1996, chap. 7), a good distance is inevitably dependent on the problem at hand. For the last question we will not try to give a clear cut answer. Instead, we give a partial answer by a list of desirable properties. For a given set of examples, if a distance measure has the following properties,

1. it is a pseudometric,
2. it gives good accuracy for the particular set of examples,

then we say that it is a good distance for doing k-NN, with respect to the examples. The exact definition of pseudometric will be given in Chapter 2. It is briefly a symmetric subadditive distance measure. The first property has twofold advantages. First, it ensures us, up to some assumptions, that our classifier has the potential to perform incrementally better when we have more observed examples in the future (a pseudometric can be regarded as a metric in a technically adjusted space). Second, subadditivity is useful to hasten nearest neighbor searches and we can plug a pseudometric into existing systems that take advantage of the triangle law if we want k-NN to be faster. So pseudometrics are both theoretically and practically salient. The existing asymptotic results, at least the work by Cover and Hart (1967), still hold for a pseudometric given that the underlying space is separable. The second property is vital in its own right.

Generalizing from Euclidean and ℓ^p spaces to metric and pseudometric spaces is somehow a sensible next step of development since metric spaces bear some relationship with ℓ^p spaces. Metric spaces are well studied. For example, it is well known that metric spaces are Hausdorff, implying that every convergent sequence has a unique limit, and any metric spaces can be embedded isometrically into a Banach space (von Luxburg and Bousquet, 2004). Several fixed point theories for metric spaces are in the mathematical literature (Espinola and Khamsi, 2001). Other than speed gains for the nearest neighbor algorithm, more interesting results may be discovered for pseudometrics k-NN or related algorithms as well.

This work is restricted to pseudometrics for univariate time series, although it will be seen that some results in our work hold for more abstract structures than just time series. We attempt to study pseudometrics for time series first because time series are slightly different from vectors. We will be well equipped with tools and structures in linear spaces to work with.

1.1 Objectives

1. To provide a framework for designing subadditive dissimilarity measure for time series so that specific distance measure can be designed to suit each particular time series data.
2. To realize a pseudometric whose average classification accuracy is comparable to DTW.

1.2 Scope

1. We are only interested in distance functions for univariate time series.

2. Performance comparison will be compared to other distances by classification tasks using the nearest neighbor algorithm.
3. The datasets used in the experiments will be at least the 20 datasets from the UCR (Keogh et al., 2006).

1.3 Procedure

1. Propose the framework for designing subadditive distance measures.
2. Develop distance measures under the framework.
3. Design an algorithm that computes the distance measures.
4. Conduct experiments of classification tasks using the distance measure.

1.4 Contributions

1. We offer more choice of subadditive distance functions which can be directly plugged into machine learning or database systems that use the triangle inequality to improve the speed. They are theoretically guaranteed, under common statistical learning assumptions, to give asymptotically good classification performance.
2. Our framework could lead to discovery of more distance functions which will be suitable for some particular problems.

1.5 Organization of the Thesis

In the remaining chapters, we will make the problem setting more precise after the introduction of notations used throughout our expositions, followed by the main work corroborated with the experiments. Sufficient backgrounds and pointers to relevant references are in Chapter 2, one should be familiar with in order to follow the development. In Chapter 3, we introduced a concept called condensation to be used as a guideline for designing new distances. As a by product, we discover an alternative characterization of the DTW distance.

The second distances construction guideline called "shortcut distance" will also be discussed in Chapter 3 and we will demonstrate how it can be used to fine tune distances to yield better empirical classification performance. Numerical results are in Chapter 4. Conclusion and future work are given in Chapter 5.

1.6 Publication

As of the time of writing (April 2008), most of the work in Chapter chap:work will be published as the paper titled "Pseudometrics for Time Series Classification by Nearest Neighbor" in the Proceedings of the 2008 International Joint Conference on Neural Networks (IJCNN 2008), part of the 2008 IEEE World Congress on Computational Intelligence (WCCI 2008), which will be held in Hong Kong in June 2008.