k-Nearest Neighbor Classification Algorithm for Multiple Choice Sequential Sampling

Yung-Kyun Noh (nohyung@snu.ac.kr)
Frank Chongwoo Park (fcp@snu.ac.kr)
School of Mechanical and Aerospace Engineering, Seoul National University
Seoul 151-744, Korea

Daniel D. Lee (ddlee@seas.upenn.edu)
Department of Electrical and Systems Engineering, University of Pennsylvania
Philadelphia, PA 19104, USA

Abstract

Decision making from sequential sampling, especially when more than two alternative choices are possible, requires appropriate stopping criteria to maximize accuracy under time constraints. Optimal conditions for stopping have previously been investigated for modeling human decision making processes. In this work, we show how the *k*-nearest neighbor classification algorithm in machine learning can be utilized as a mathematical framework to derive a variety of novel sequential sampling models. We interpret these nearest neighbor models in the context of diffusion decision making (DDM) methods. We compare these nearest neighbor methods to exemplar-based models and accumulator models, such as Race and LCA. Computational experiments show that the new models demonstrate significantly higher accuracy given equivalent time constraints.

Keywords: sequential sampling; decision making; diffusion decision making model; *k*-nearest neighbor classification; evidence; sequential probability ratio test

Introduction

Whenever a faster decision is required to save time and resources, the decision making process should focus on choosing whether to proceed with a decision in light of the given information or to postpone the decision in order to collect more information for a higher confidence level. In many previous and recent psychology works, various computational models have been introduced seeking to explain the speed-accuracy tradeoff and to understand the decision making process in humans. However, apart from the understanding of individual models, there has been little systematic way of understanding these models in one mathematically unified framework. Moreover, multiple-choice problems were not discussed intensively in any of the methods.

The optimality in decision making with sequential sampling is discussed with the optimality in speed-accuracy tradeoff. In other words, the objective of the present work is to seek the fastest decision with the same average accuracy or the maximum accuracy if the same average decision time is used. Sequential sampling methods such as Race (Smith & Vickers, 1988; Vickers, 1970), diffusion decision making (DDM) (Ratcliff, 1978; Ratcliff & Rouder, 2000; Shadlen, Hanks, Churchland, Kiani, & Yang, 2006; Ratcliff & Mckoon, 2008), and leaky competing accumulator (LCA) (Usher & McClelland, 2001; Bogacz, Usher, Zhang, & McClelland, 2007) are all interested in explaining this optimality in the speed-accuracy tradeoff. In these methods, one or more variables are commonly introduced for accumulating sampled information, and a criterion is used to determine whether to continue collecting more information or to make a decision with given information. Here, we propose a common mathematical framework combining these methods and providing a systematic explanation for understanding different methods.

Our framework combining sequential sampling methods is the *k*-nearest neighbor (NN) classification in machine learning. The sequential sampling situation with multiple choices is explained as the multiway *k*-NN classification from the theoretical analysis on *k*-NNs in the asymptotic situation. Due to this connection, we can interpret all different types of sequential sampling methods as different methods of choosing *k* adaptively in *k*-NN classification. By further analyzing the strategy of choosing *k* in *k*-NN classification using the Sequential Probability Ratio Test (SPRT) (Wald & Wolfowitz, 1948) and Bayesian inference, we can obtain five different accumulating variable and stopping criteria for optimal tradeoff. Interestingly, all these five optimal methods are interpreted as different kinds of DDM strategies.

Our work is directly applied to a recently reported neuroscientific decision making mechanism. The proposed mechanism considers an output neuron which sends out a decision result. By collecting Poisson spike trains from different neurons, the output neuron makes a decision about which neuron gives Poisson spikes at the highest rate (Shadlen & Newsome, 1998; Ma, Beck, Latham, & Pouget, 2006; Beck et al., 2008; Zhang & Bogacz, 2010). The output neuron can achieve optimality by using our proposed strategies.

The proposed method can be compared with traditional exemplar models which explain memory retrieval using similarity weighted voting based on stored exemplars. Our work is different from this line of research by using majority voting of adaptively chosen *k* number of NNs. We discuss the advantages and disadvantages of our method when it is applied to the memory retrieval problem.

The rest of the paper is organized as follows. We introduce the sequential sampling problem in Section 2 especially from the point of view of multiple-choice. In Section 3, we introduce problems to which sequential sampling methods can be applied, and we show how *k*-NN classification can be naturally introduced as a common framework for explaining these problems. In Section 4, we derive the examples of two- and multiple-choice evidence for DDM in light of k-NN classification. After we explain the relationship between our method and other exemplar methods in Section 5, we present simulation experiments in Section 6. Finally, we conclude with discussion in Section 7.

Computational Methods in Sequential Sampling Problems

Sequential sampling methods consider decision making using incoming information over time. With unlimited time, the decision can be made late enough to increase the expected accuracy. However, if the decision should be made as soon as possible, there is a trade-off between the speed and accuracy of the decision. In order to address this tension, decision making strategies introduce criteria to determine whether or not to make a decision at a certain time.

Accumulator Model: One simple method of determining whether the accumulated information has reached a certain level of confidence is the accumulator model. This model considers one variable for each choice and accumulates information separately in favor of each choice. Once one of the accumulating variable reaches a predefined threshold, the decision is made immediately thereafter.

This simple model with no interaction between different choices is known as suboptimal. This method can be compared with the DDM strategy in the next section, where the accuracy of the accumulator model is always less than the accuracy of the DDM model (Zhang & Bogacz, 2010). This model of doing race between accumulators is also called the Race model.

Diffusion Decision Making (DDM) Model: In this model with two choices, one variable is introduced to collect information and diffuse toward one of the choice. This variable, also known as the evidence, represents the bias in the preference of accumulated information toward a choice. Finally, once the evidence reaches a pre-defined level of any choice, it stops diffusing and selects the choice.

A canonical method of determining the evidence variable and stopping criterion uses the sequential probability ratio test (SPRT) (Wald & Wolfowitz, 1948; Dragalin, Tertakovsky, & Veeravalli, 1999; Zhang & Bogacz, 2010). Previous work using this test has considered two incoming Poisson signals aiming to determine the signal with the higher Poisson rate from the accumulation of signals. In this case, the diffusing evidence is just the difference in the number of signals within a certain time, and the decision is made once this difference exceeds a threshold. This method is known to be optimal among sequential sampling methods such as Race and LCA (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Bogacz et al., 2007).

Leaky Competing Accumulator (LCA): LCA uses one variable for each choice similar to the accumulator model, but it considers the interaction between the variables. LCA considers the dynamics of activation with the decay of the activation as well as the inhibitory interaction between activation variables. This LCA dynamics is very flexible in that the strategy can be either similar to Race or DDM as its special case, but the maximum performance is known to be that of DDM (Bogacz et al., 2007).

Multiple-Choice Extension

Among the aforementioned sequential sampling models, the multiple-choice extension of the Race and LCA models is straightforward, by just increasing the number of accumulating variables. However, the extension of DDM is more complex. Fortunately, the Multiple SPRT (MSPRT) method was previously developed by extending the SPRT method using the number of signals (Dragalin et al., 1999; Zhang & Bogacz, 2010). In addition to this MSPRT result, we also provide different criteria for multiple-choice DDM using derivations from other approaches of MSPRT and Bayesian inference. Our result provides an evidence diffusing in a $C-1$ dimensional space for a C alternative choice problem.

Sequential Sampling Problems

Decision making problem using sequential sampling can be found in many examples. Here we introduce two exemplary problems. One example can be found in neuronal decision making as in the left figure of Fig. 1. When an output neuron tries to make a decision as to whether one incoming signal has a higher Poisson rate than the other has, the output neuron can collect signals until the accumulated information reaches a certain level.

Another example can be found in a Bayes classification problem where we only have data generated from unknown underlying density functions. Bayes classification selects the class having the highest underlying density, but the classifier in this case cannot directly access the underlying density information. A surrogate method of determining the class of highest density is through k-NN classification. By collecting more nearest neighbors, the confidence of choosing a class of the highest density is expected to increase to a targeted level.

Here, we show that the two problems are in fact exactly the same by explaining several theoretical results on k-NN classification in the asymptotic situation:

Majority Voting Rule in k-Nearest Neighbor Classification: When there are N number of training data with labels, $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$, where each datum $\mathbf{x}_i \in \mathbb{R}^D$ is represented as a D-dimensional vector, and the label has one of C labels, $y_i \in \{1, \ldots, C\}$, k-NN classification assigns class y to a class-unknown datum $\mathbf{x}$ according to the majority voting with k labels of nearest data in $\mathcal{D}$:

$$y = \arg\max_c \sum_{i=1}^{k} \mathbb{I}(y_{n(i)} = c) \tag{1}$$

with nearest neighbor index $n(i), i = 1, \ldots, k$. The theoretical study of this majority voting strategy originates from Cover

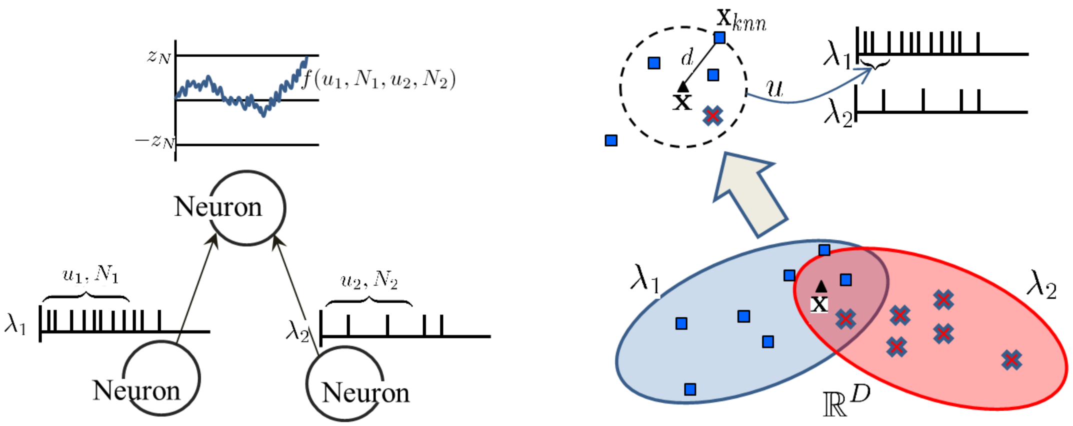

Figure 1: Diffusion decision making (DDM) in neurons determines the input neuron with the higher firing rate, and its analogous k-NN classification determines the larger class-conditional density function

and Hart (Cover & Hart, 1967; Cover, 1967), and in their work, once there are enough data, the expected error monotonically and asymptotically decreases with the number of k:

$$E(k=1) \geq E(k=2) \geq \ldots \geq E(k=\infty) \tag{2}$$

for $k \ll N$. This continuous decrease of error will encourage the use of more nearest neighbors, which explains the tradeoff between the number of nearest neighbors k and the classification accuracy.

Distance Comparison Rule of k'-th NNs of Each Class: If we consider a strategy of comparing two k'-th NN in class 1 and class 2, we can easily prove that this strategy is equivalent to the majority voting rule with $k = 2k' - 1$ NNs.

Proof: *Consider a comparison between k'-th NN of class 1 and another k'-th NN of class 2. If k'-th NN in class 1 is closer than k'-th NN in class 2, then we can say that the k'-th NN in class 2 can never be included in the closest $(2k'-1)$ NNs, because at least k' number of NNs in class 1 and $(k'-1)$ number of NNs in class 2 have less distance than the k'-th NN in class 2. Therefore, comparing strategies of k'-th NN in each class is the same as majority voting with $(2k'-1)$ nearest neighbors.*

Therefore, the monotonic increase of accuracy is also satisfied with the increase of k'.

Two Sequential Sampling Methods in k-NN Classification: From the monotonic increase of the accuracy with the increase of k (or k'), we can make two different sequential sampling methods showing the speed-accuracy tradeoff.

First, we can consider the majority voting strategy using number of NNs within a certain distance from the testing point. If we do not have enough accuracy with the current distance, we can increase it to use more resources. Another example can be designed by considering the distance to the same k'-th NN in each class and making a decision by comparing the distances.

The first design corresponds to the sequential sampling with continuous time and discrete accumulation of information, because the accumulation variable is the function of the number of NNs. In contrast, the second design uses the discrete time and continuous accumulation of information.

Distribution of the distances: Now, we show that k-NN classification is in fact equivalent to sequential sampling for determining the signal with the highest Poisson rate.

A recent study discussed the distribution of the distance to the NNs when there are enough data (Leonenko, Pronzato, & Savani, 2008). Instead of directly dealing with the distribution of distance, they changed the random variable to $u = NV$, with volume V of D dimensional hypersphere having the distance to the k-th NN as a diameter multiplied by the number of data N. Then the distribution of samples approaches the Erlang density function:

$$\rho(u|\lambda) = \frac{\lambda^k}{\Gamma(k)} \exp(-\lambda u) u^{k-1} \tag{3}$$

with a parameter λ, which is the probability density $p(\mathbf{x})$ at $\mathbf{x} \in \mathbb{R}^D$. Moreover, this special Erlang function implies the Poisson distribution of the number of NNs k within a specified volume of the hypersphere (Wasserman, 2003):

$$\rho(k|\lambda) = \frac{\lambda^k}{\Gamma(k+1)} \exp(-\lambda). \tag{4}$$

This equation shows that the number of NNs within a growing hypersphere at a constant rate in volume is a Poisson process.

Comparing this Poisson process interpretation with the aforementioned neuronal decision making, we can draw several corresponding analogies. The firing rate of the Poisson signal corresponds to the underlying density function in k-NN classification, the number of spikes corresponds to the number of NNs, the time within which spikes are counted corresponds to the volume of the hypersphere within which we count NNs, and as a consequence, determining a choice with the highest firing rate corresponds to the problem of determining the class with the highest underlying density function, which is also known as the Bayes classification.

The correspondence shows that these two very well-known methods from different disciplines can share optimal strategies as well as theoretical knowledge. However, the study of a method in one field is rarely investigated in another; the strength of the correspondence suggests that whenever a good strategy is found for DDM, a corresponding strategy should be examined for machine learning. Conversely, when a new strategy is provided for the k-NN method, its relevance to psychology should also be investigated.

Derivation of Stopping Criteria

In this section, we now derive stopping criteria from k-NN classification using MSPRT and Bayesian inference for a multiple-choice problem.

Multiple Sequential Probability Ratio Test

One simple statistical test for determining whether one of C different choices has the highest probability density is the MSPRT. MSPRT uses fixed parameters of densities λ_+ and λ_- where $\lambda_+ > \lambda_-$, it calculates the likelihood that the first data came from the density λ_+ and others from λ_-, and then compares those likelihoods.

Without loss of generality, we consider the likelihood that the highest density λ_+ is occupied by λ_1. In other words, $\lambda_1 = \lambda_+$, and $\lambda_c = \lambda_-$ for $c = 2, \ldots, C$. Because of the independence between classes,

$$\log P\left(k_1, \ldots, k_C, u_1, \ldots, u_C \middle| \lambda_1 = \lambda_+,\ \lambda_2 = \lambda_-,\ \ldots,\ \lambda_C = \lambda_-\right)$$

$$= \quad \log \rho(k_1, u_1|\lambda_+) + \sum_{c=2}^{C} \log \rho(k_c, u_c|\lambda_-). \tag{5}$$

The posterior P_1 that λ_1 occupies λ_+ is proportional to this likelihood Eq. (5). From the Poisson distribution in Eq. (4) with k_c, the number of NNs of class c within the same volume, we can obtain the log of posterior:

$$\log P_1 = g^* k_1 - \log\left(\sum_{c=1}^{C} \exp(g^* k_c)\right) \tag{6}$$

with a predetermined ratio $g^* = \log(\lambda_+/\lambda_-)$. If we consider the volume distribution for the same k-th NNs, the equation for the posterior also becomes

$$\log P_1 = -h^* u_1 - \log\left(\sum_{c=1}^{C} \exp(-h^* u_c)\right) \tag{7}$$

with $h^* = \lambda_+ - \lambda_-$. We call Eq. (6) "DN", which considers the difference in the number of NNs within a specific volume of the hypersphere and Eq. (7) "DV", which considers the difference in the volumes of the same k-th NNs. In order to make a decision with confidence, we can first increase the volume of hypersphere or increase the number of NNs until the criterion exceeds a pre-defined confidence level, then we can decide the choice. For two-choice problem ($C = 2$), comparing the MSPRT criteria with a certain value reduces to a simple comparison whether $g^*(k_1 - k_2)$ and $h^*(u_2 - u_1)$ is greater than a certain confidence threshold, for Eq. (6) and Eq. (7), respectively.

For DV, an additional conservative method can be considered. The decision can be made more carefully for the class of interest (here, class 1), by using the maximum possible volume containing k-th NN, in other words, the volume of the hypersphere of $(k+1)$-th NN of class 1 instead of the volume of k-th NN. We call this strategy "conservative DV" (CDV), and in CDV, an additional NN is always used to calculate the accumulated information.

Bayesian Inference

Another method of utilizing the Bayesian method is to use the prior density function for λ with parameters a and b:

$$p(\lambda) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-\lambda b). \tag{8}$$

With conjugacy relationship, we can calculate the posterior probability that the underlying density of choice 1, λ_1, is greater than the underlying densities of the other choices $\lambda_2, \ldots, \lambda_C$ with given condition D on nearest neighbor information. The calculation of $P(\lambda_1 > \lambda_2, \ldots, \lambda_C|D)$ is performed using the probability primitives such as $P(\lambda_1 < \lambda_2|D)$, $P(\lambda_1 < \lambda_3|D)$, ..., and $P(\lambda_1 < \lambda_2, \ldots, \lambda_C|D)$:

$$\begin{aligned} P(\lambda_1 > \lambda_2, \ldots, \lambda_C|D) &= \int_0^\infty d\lambda_1 p(\lambda_1|D) \left(1 - \int_{\lambda_1}^\infty d\lambda_2 p(\lambda_2|D)\right) \\ &\quad \cdots \left(1 - \int_{\lambda_1}^\infty d\lambda_C\, p(\lambda_C|D)\right) \qquad (9) \\ &= 1 - P(\lambda_1 < \lambda_2|D) \ldots - P(\lambda_1 < \lambda_C|D) + \\ &\quad \ldots + (-1)^{C-1} P(\lambda_1 < \lambda_2, \ldots, \lambda_C|D). \qquad (10) \end{aligned}$$

When the condition is on the number of nearest neighbors $k_1, \ldots, k_C$ within a certain volume, the general form of primitives is presented with multinomial coefficients:

$$P(\lambda_1 < \lambda_{j_2}, \ldots, \lambda_{j_L}|k_1, \ldots, k_C) = \tag{11}$$
$$\sum_{i_{j_2}=0}^{k_{j_2}} \cdots \sum_{i_{j_L}=0}^{k_{j_L}} \frac{1}{L^{(k_1+1+\sum_{c=2}^{L}(k_{j_c}-i_{j_c}))}} \begin{pmatrix} k_1 + \sum_{c=2}^{L}(k_{j_c} - i_{j_c}) \\ k_{j_2} - i_{j_2}\ ,\ \cdots\ ,\ k_{j_L} - i_{j_L} \end{pmatrix}$$

where L and $j_1, \ldots, j_L$ are determined according to the primitives in Eq. (10). In addition, when volume information $u_1, \ldots, u_C$ is given for $k_1, \cdots, k_C$-th NN in each class, respectively, the primitives are

$$P(\lambda_1 < \lambda_{j_2}, \ldots, \lambda_{j_L}|u_1, \ldots, u_C) = \tag{12}$$
$$\sum_{i_{j_2}=0}^{k_{j_2}} \cdots \sum_{i_{j_L}=0}^{k_{j_L}} \begin{pmatrix} k_1 + \sum_{c=2}^{L} i_{j_c} \\ i_{j_2}\ ,\ \cdots\ , i_{j_L} \end{pmatrix} \frac{u_1^{k_1+1} \prod_{c=2}^{L} u_{j_c}^{i_{j_c}}}{(u_1 + \sum_{c=2}^{L} u_{j_c})^{k_1+\sum_{c=2}^{L} i_{j_c}+1}}$$

for L and $j_1, \ldots, j_L$ determined from the primitive. Now, Eq. (10) with primitives in Eq. (11) can be considered as a

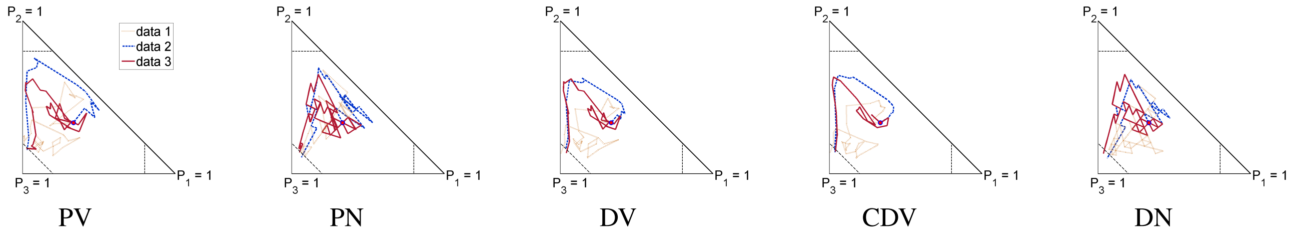

Figure 2: Examples of diffusion of evidence for three-choice decision making. The diffusion of posteriors, P_1 and P_2, are plotted on the horizontal and vertical axes. The threshold is set to .8.

criterion "PN" and Eq. (10) with primitives in Eq. (12) can be considered as a criterion "PV." Here, we used $a = 1$ and positive small value b.

For a two-choice problem, with k_c-th NNs in c class within the same hypersphere, the probability result becomes a very simple equation

$$P(\lambda_1 > \lambda_2 | u_1, u_2) = \sum_{m=0}^{k} \begin{pmatrix} 2k+1 \\ m \end{pmatrix} \frac{u_1^m u_2^{2k+1-m}}{(u_1+u_2)^{2k+1}} \qquad (13)$$

from Eq. (11). Similarly, with u_1 and u_2 of k-th NN in each class, Eq. (12) becomes

$$P(\lambda_1 > \lambda_2 | k_1, k_2) = \frac{1}{2^{k_1+k_2+1}} \sum_{m=0}^{k_1} \begin{pmatrix} k_1+k_2+1 \\ m \end{pmatrix}. \qquad (14)$$

Both Eq. (13) and Eq. (14) are the sums of binomial distributions which can be interpreted analogous to coin tossing problem with a biased and an unbiased coin. Eq. (13) corresponds to the probability of having heads less than or equal to k among $2k+1$ tosses of a biased coin, and Eq. (14) corresponds to the probability of having heads less than or equal to k_1 among k_1+k_2+1 tosses of an unbiased coin.

We can note that all derived stopping criteria have a posterior representation where the sum over classes equals one. Therefore, we can consider a $C-1$ dimensional simplex and the diffusion of the posterior within this simplex. Therefore, a vector with posterior elements for all candidate classes extending Eq. (10) can be considered as a diffusing evidence in a DDM model, and all criteria derived in this work can subsequently be considered as DDM models.

Relationship with other Exemplar Methods

One typical method of learning with exemplars is utilizing the similarity measures with exemplars (Nosofsky, 1986; Shepard, 1987). Recently, this model was connected to kernel learning methods in machine learning (Jäkel, Schölkopf, & Wichmann, 2008), which connected the similarity notion to an associated reproducing kernel Hilbert space as well as to Bayesian inference (Shi, Griffiths, Feldman, & Sanborn, 2010). These similarity-based methods utilizing exemplars are computationally well-integrated with various machine learning methods.

However, majority voting with equal weights, which is proposed in this work, is a completetly different approach of utilizing exemplars, where the theoretical explanation shows optimality in certain situations (Bailey & Jain, 1978). Our model is also different from the random walk model using conventional exemplar models (Nosofsky & Palmeri, 1997). The random walk is performed according to the random retrieval from already generated data, while our model directly considers the underlying density function and uses the generated data without any additional randomness. A severe problem in the memory retrieval of Nosofsky and Palmeri is that a repetitive retrieval of one very similar exemplar will affect the decision predominantly where a noise on this particular exemplar can severely affect the decision accuracy.

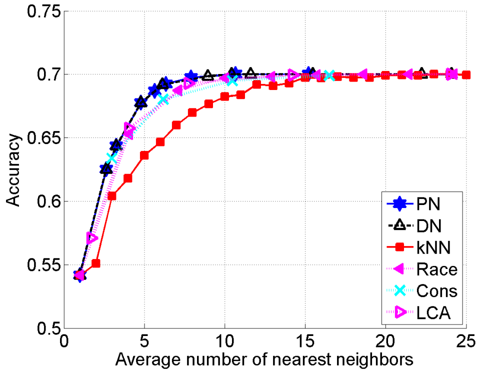

Figure 3: Performance of adaptive k-NN classification using PN, DN, Race, and a machine learning criteria, "Cons." Accuracy is plotted with an average number of NNs used for various thresholds of confidence. Cons makes a decision when the number of recent consecutive NNs of the same class exceeds a threshold (Ougiaroglou et al., 2007).

Experiments with Simulation Data

The examples of diffusion of the evidence for each criteria are shown in Fig. 2. In this experiment, the proposed five examples of evidence PV, PN, DV, CDV, and DN, diffuse with the same NN information. In the figure, all five examples diffuse differently, but they reach the same threshold. The parameters used are $\lambda_1 = .25$, $\lambda_2 = .35$, and $\lambda_3 = .4$, and the decision threshold is .8. Though they diffuse differently, CDV shows a smoother diffusion than the others, and PN and DN show more sampling-wise configuration.

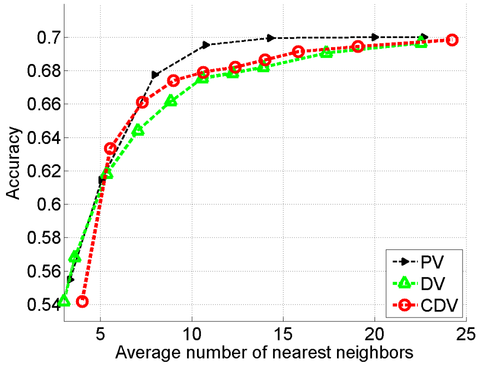

Figure 4: Performance using volume evidence. The CDV with smooth diffusion is slightly better than DV, while PV outperforms both CDV and DV with large margins.

The performance evaluation of the methods is shown using *k*-NN classification. We first generated data randomly from three uniform probability densities, $\lambda_1 = .2$, $\lambda_2 = .7$, and $\lambda_3 = .1$, and compared the adaptive *k*-NN classification method between the proposed criteria and other criteria from psychology and machine learning models. In Fig. 3, as expected in our analysis, the Race accumulator model without interaction does not outperform the criteria from statistical tests, PN and DN, although using a Race criterion does give a better performance than a simple majority voting method with fixed *k*. We also compared our results with a conventional machine learning method, which considers the number of recently appeared NNs belonging to the same class.

In Fig. 4, three criteria using volume information, PV, DV, and CDV, are compared. According to a few realizations in Fig. 1, the diffusion of CDV is in general much smoother than that of DV, and the CDV criterion shows a little better accuracy than DV. PV shows better performance than either DV or CDV.

Conclusion

In this work, a general framework integrating decision making with sequential sampling is proposed based on its relationship with the exemplar-type machine learning algorithm, *k*-NN classification. In contrast to previous research on suboptimal weighted voting, we have shown how *k*-NN majority voting can be used to better understand the sequential sampling decision making process. Using an adaptive *k*-NN classification framework, we also showed how the proposed five examples of optimal criteria are derived for multiple-choice decision making, minimizing the error for any given average resource that can be used. Our future work includes extending this relationship among decision making methods to form a scaffold of understanding within the mathematical framework of *k*-NN methods.

Acknowledgments

This research was supported in part by the AIM Center for Advanced Intelligent Manipulation and the ROSAEC-ERC.

References

Bailey, T., & Jain, A. K. (1978). A Note on Distance-Weighted *k*-Nearest Neighbor Rules. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-8*(4), 311–313.

Beck, J. M., Ma, W. J., Kiani, R., Hanks, T., Churchland, A. K., Roitman, J., et al. (2008, 26). Probabilistic population codes for Bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700–765.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society, Series B*.

Cover, T. (1967). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, *14*(1), 50–55.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Dragalin, V. P., Tertakovsky, A. G., & Veeravalli, V. V. (1999). Multihypothesis sequential probability ratio tests . part i: asymptotic optimality. *IEEETransactions on Information Technology*, *45*, 2448–61.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008, 4). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin and Review*, *15*(2), 256-271.

Leonenko, N., Pronzato, L., & Savani, V. (2008). A class of Rényi information estimators for multidimensional densities. *Annals of Statistics*, *36*, 2153–2182.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006, 22). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432–1438.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266–300.

Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A. N., Manolopoulos, Y., & Welzer-Druzovec, T. (2007). Adaptive k-nearest-neighbor classification using a dynamic number of nearest neighbors. In *Proceedings of the 11th east european conference on advances in databases and information systems* (pp. 66–82).

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & Mckoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two-choice letter identification. *Journal of Experimental Psychology Human Perception and Performance*, *26*(1), 127–140.

Shadlen, M. N., Hanks, A. K., Churchland, A. K., Kiani, R., & Yang, T. (2006). The speed and accuracy of a simple perceptual decision: a mathematical primer. *Bayesian brain: Probabilistic approaches to neural coding*.

Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *Journal of Neuroscience*, *18*, 3870–3896.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–464.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. *Journal of Mathematical Psychology*, *32*, 135–168.

Usher, M., & McClelland, J. L. (2001, July). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological review*, *108*(3), 550–592.

Vickers, D. (1970). Evidence for an accumulator model of psychophysical discrimination. *Ergonomics*, *13*, 37–58.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, *19*, 326–339.

Wasserman, L. (2003). *All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics)*. Springer. Hardcover.

Zhang, J., & Bogacz, R. (2010). Optimal decision making on the basis of evidence represented in spike trains. *Neural Computation*, *22*(5), 1113–1148.