A Probabilistic Learning Reinforcement Model for the Performance Analysis of Multimedia Indexing and Packet Switching

Clement H. C. Leung
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen, China
clementleung@cuhk.edu.cn

Yao Tong
School of Data Science
Chinese University of Hong Kong
Shenzhen, China
yaotong@link.cuhk.edu.cn

Abstract — **A stochastic model of binary classification in the presence of noise is considered where classification outcomes are non-deterministic. To ensure the correctness of a particular classification decision, repeated reinforcements need to be acquired. By accumulating sufficient reinforcements, one would learn to predict the class label. In this study, we develop a probabilistic learning reinforcement classification model and apply it to multimedia information indexing and to noisy network transmission. Three learning strategies are analyzed. The first one requires the accumulation of a total of a given number of pre-specified positive labels, while the second one builds from the first and requires additionally that such reinforcements occur consecutively in the observation sequence. The third strategy views the classification process from a multi-agent stochastic game perspective, with the labelling decision determined by which class label attaining a given threshold first. The model characteristics are studied for the three different strategies and key measures of performance are obtained. The model is applied to fault-tolerant network communications over a noisy channel, with learning success corresponding to error-free transmission of data packets, and to multimedia indexing where learning success correspond to the successful automatic installation of an index term to a particular data object. The present learning paradigm will be useful in allowing the effectiveness and performance of these systems and similar ones to be meaningfully quantified and evaluated.**

Keywords – computer networks; packet switching; multimedia information indexing; reinforcement learning; multi-agent; naïve Bayes classifiers; stochastic game

I. INTRODUCTION AND RELATIONSHIP WITH OTHER WORK

Sequential classification problems are ubiquitous and many important decision problems can often be reduced to a classification problem. Among the variety of classification problems, binary classification problems consisting of two class labels are particularly common. However, classification verdicts returned by different classifiers over time are often non-deterministic, causing uncertainty in the classification decision. In such a situation, repeated reinforcements are necessary to ensure the reliability of the decision.

To be concrete, we shall use the multimedia indexing scenario [1] to explain the key concepts. Later we shall establish a correspondence between multimedia search and fault-tolerant network communications. In effective information and document retrieval, it is often necessary to involve the users in the search process so as to improve the overall return results [18][19][22] [29]. In addition, affective indexing of multimedia content combines emotional responses generated by the users is sometimes employed, e.g., the psycho-physiological signals, galvanic skin response, face tracking [19][27].

In [20], it is proposed that a reinforcement learning approach is suitable for users exposing to raw and high-dimensional information, whereas instant rewards of the agents is generally able to impart significant improvements in the searching process [21]. In [23], it is shown that using Markov decision process improves the efficiency of locating video frames in a video, and in [24], the distribution of visual words of multimedia data is found to be probabilistic in relation to the concept relationship formed [24]. Users often allocate the results based on some form of scoring metrics; for example, a linear combination of posterior probability is employed to refine the search results [25]. In reinforcement learning, an agent learns through the interaction with the dynamic environment to maximize its long-term rewards, in order to act optimally. Most of the time, when modeling real-world problems, the environment involved is non-stationary and noisy [3][4][6]. More precisely, the next state results from taking the same action in a specific state may not necessarily be the same but appears to be stochastic [2][7][31][32][33][34][35]. And the exploration strategies adopted in different categories of reinforcement learning algorithms provide different levels of control to the exploration of unknown factors, which in turn give various possibilities to the learning outcomes. Hence, the observed rewards and punishments are often non-deterministic. For example, when one is trying to find a video for performing a particular task, a shortening of the searching time with respect to some anticipated norm may be regarded as a reward, while a lengthening of the same may be viewed as punishment. Likewise, when one is exploring a new advertising channel, a resultant significant increase in sales may be viewed as a reward, while failure to do so may be regarded as punishment. In situations like these, there are stochastic elements governing the underlying environment. In the new route to work example, whether one receives rewards or punishments depends on a variety of chance factors, such as weather condition, day of the week, and whether there happens to be traffic delays or road works.

The effect of noise in multimedia data is generally numerous and cannot be known or enumerated in a practical

sense, and these tend to mask the underlying pattern. Indeed, if stochastic elements are absent, the learning problems involved could be greatly simplified and their presence has motivated early research in the area. As early as 1990s, mainstream research in reinforcement learning, such as the survey assessing existing methods carried out by Kaelbling *et al.* [2], adopts the common assumption of a stationary environment within a reinforcement learning framework. Later on, with further advances in reinforcement learning, theoretical analyses addressing the concern of non-stationary environment attracted great interests. One of the works by Brafman and Tennenholtz introduces a model-based RL algorithm R-Max to deal with stochastic games [5]. Such stochastic elements can notably increase the complexity in multi-agent systems and multi-agent tasks, where agents learn to cooperate and compete simultaneously [6][10]. Autonomous agents are required to learn new behaviors online and predict the behaviors of other agents in multi-agent systems. As other agents adapt and actively adjust their policies, the best strategy for each agent would evolve dynamically, giving rise to non-stationarity [8][9].

For most of the aforementioned scenarios, the cost of a trial or observation to receive either a reward or punishment can be significant, and preferably, one would like to arrive at the correct conclusion by incurring minimum cost. In the case of the advertising example, the cost of advertising can be considerable and one would therefore like to minimize it while acquiring the knowledge whether such advertising channel is effective. Similarly, in reinforcement learning algorithms, we are always in the hope to rapidly converge to an optimal strategy with least volumes of data, calculations, learning iterations, and minimal degree of complexity [11][12]. To do so, one should explicitly define the stopping rules for specifying the conditions under which learning should terminate and a conclusion drawn as to whether the learning has been successful or not based on the observations so far.

The issue of finding termination conditions, or stopping rules, is an intensive research topic in reinforcement learning, which is closely linked to the problems of optimal policies and policy convergence [13]. Traditional reinforcement learning algorithms mainly aim for relatively small-scale problems with finite states and actions. The stopping rules involved are well-defined for each category of algorithms, such as utilizing Bellman Equation in *Q*-learning [14]. To deal with continuous action spaces or state spaces, new algorithms, such as the Cacla algorithm [15] and CMA-ES algorithm [16], are developed with specific stopping criteria. Still, most studies on stopping criteria are algorithm-oriented and do not have a unified measurement for comparison purposes.

In this study, we present an approach to reinforcement learning by using a naïve Bayes classification framework, which explicitly incorporates the stochastic aspects of the environment in multimedia information search and retrieval. Applying naïve Bayes methods for classification problems are often employed in a variety of contexts [26][36], such as crowdsourcing and police surveillance. Here, we shall also learn and estimate the underlying stochastic structure of the environment by making use of the random classification labels gathered in the course of the learning process.

The structure of this paper is organized a s follows. Section II presents a unified framework in the probabilistic classification of binary outcomes, and key measures of performance are derived, while the estimation of parameters is described in Section III. Section IV views competing classification outcomes as a stochastic game involving multi-agents. Sections V and VI respectively apply the results to the performance analysis of network communications and multimedia search.

II. A Probabilistic Framework and Fundamental Strategies

We consider a binary classification problem with two class labels, 1 or 0, where for convenience the former is referred to as a success, and the latter, a failure. A success yields a positive outcome and may be referred to as a positive classification, while a failure may be referred to as a negative classification. We are interested in determining whether the sequential classifications indicate overall success or failure in the classification process. Evidently, if the number of 1-labels gathered is much greater than the number of 0-labels, then the conclusion drawn should be success, while if the opposite is true, then the corresponding conclusion should be failure. In order to proceed with the analysis, we first let p and q (with $p + q = 1$) denote the probabilities of receiving a 1-label or 0-label respectively for a given classification. Furthermore, we shall make use of the naïve Bayes property that different classifications are independent of each other. Later on, we shall derive estimates for p and q, which capture the stochastic structure of the environment. For example, if $p > q$, then clearly the final conclusion should be success. An error often committed is that when the first few observations are all 0, one would terminate prematurely and return a verdict of failure. Let us consider the following termination strategy; such a strategy is also studied in [26][36] and is called majority voting.

***Strategy A**: On accumulating a total of r labels all belonging to either 1 or 0, the process terminates and a decision is made in accordance with the accepted margin of the majority of voting of the classifiers.*

Here, we let the random variable $\boldsymbol{X}$ represent the number of classification labeling preceding the first positive classification; i.e., $\boldsymbol{X}$ may be viewed as the waiting time to the first positive classification,

$$\Pr[\boldsymbol{X} = k] = pq^k, \quad k = 0, 1, 2, 3, \ldots \tag{1}$$

The probability generating function $G(z)$ of $\boldsymbol{X}$ is given by

$$G(z) = \sum_{k=0}^{\infty} \Pr[\boldsymbol{X} = k]\, z^k$$

$$= p \sum_{k=0}^{\infty} q^k z^k = \frac{p}{(1-qz)}. \quad (2)$$

Note this is a regenerative process in that after the occurrence of the first positive classification, the process probabilistically repeats itself again, so that we have for the waiting time $\boldsymbol{W_r}$ of the *r*th positive classification

$$\boldsymbol{W}_r = \sum_{k=1}^{r} \boldsymbol{X}_k, \quad (3)$$

where each $\boldsymbol{X_k}$ has the same distributional characteristics as $\boldsymbol{X}$. From [17], the probability generating function of $G_r(z)$ corresponding to $\boldsymbol{W_r}$ may be obtained

$$G_r(z) = G_1(z)^r$$

$$= [\frac{p}{(1-qz)}]^r . \quad (4)$$

To gain a better understanding of behavior specified above, it is useful to obtain the average waiting time $\boldsymbol{W_r}$ and its variance when *r* positive labels are attained. From (4), the mean and variance of $\boldsymbol{W_r}$ can be derived

$$\mathrm{E}[\boldsymbol{W_r}] = G_r'(1) = \frac{rq}{p} , \quad (5)$$

$$\mathrm{Var}[\boldsymbol{W_r}] = G_r''(1) + G_r'(1) - G_r'(1)^2 = \frac{rq}{p^2} . \quad (6)$$

Moreover, the probabilities $\Pr[\boldsymbol{W_r} = k]$ may be readily obtained from the expansion of (4) so as to study the probabilities for various waiting time,

$$\Pr[\boldsymbol{W_r} = k] = \binom{-r}{k} p^r (-q)^k, \quad k = 0, 1, 2, 3, \ldots \quad (7)$$

We note that, while $-r$ appears as a negative integer in the binomial coefficient, the entire expression is actually non-negative [18]. Since $\boldsymbol{W_r}$ is the sum of independent identically distributed random variables, when *r* is appreciable, it may be approximated by the normal distribution [17]

$$\boldsymbol{W_r} \sim N\left(\frac{rq}{p}, \frac{rq}{p^2}\right), \quad (8)$$

whence we have, denoting by Φ the standard normal distribution,

$$\Pr[\boldsymbol{W_r} > b] = \int_{\frac{bp-rq}{\sqrt{rq}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

$$= 1 - \Phi\left(\frac{bp-rq}{\sqrt{rq}}\right). \quad (9)$$

We next consider a more stringent strategy whereby consecutive occurrence of labels is required. As we shall see, it would take much longer to complete the process in Strategy B than in Strategy A.

***Strategy B**: On the occurrence of m consecutive labels all belonging to either 1 or 0, the process terminates and a decision is made in accordance with the accepted margin of the majority of voting of the classifiers.*

To establish the results for this second case, we shall first derive the probability of occurrence of the event corresponding to Learning Strategy B *for the first time*. Let b_n be the probability that *m* consecutive positive rewards occurs at trial *n*, with $n \geq m$, not necessarily for the first time, and we denote by $B(z)$ be the corresponding probability generating function. From [17], this probability generating function can be obtained as

$$B(z) = \frac{1 - z + qp^m z^{m+1}}{(1-z)(1-p^m z^m)} . \quad (10)$$

Since we need to obtain the corresponding generating function for the probability that the associated event occurs for the first time, we need to consider the relationship between the two events. We shall use the random variable $\boldsymbol{V_m}$ to denote the number of plays preceding and including the receiving of the first set of *m* consecutive positive rewards, and we let a_n be the probability

$$a_n = \Pr[\boldsymbol{V_m} = n], \qquad n = m, m+1, \ldots . \quad (11)$$

We denote by $A(z)$ the probability generating function for the event that the accumulation of *m* positive rewards occurs for the first time. It can be shown in [17] that the generating function $A(z)$ is related to $B(z)$ by

$$A(z) = \frac{B(z) - 1}{B(z)} . \quad (12)$$

From this, we obtain, after simplification,

$$A(z) = \frac{p^m z^m}{1 - q^m \sum_{k=0}^{m-1} p^k z^k} . \quad (13)$$

From this, the mean and variance of $\boldsymbol{V_m}$ can be readily obtained after simplification,

$$\mathrm{E}[\boldsymbol{V_m}] = A'(1) = \frac{1-p^m}{qp^m} , \quad (14)$$

$$\mathrm{Var}[\boldsymbol{V_m}] = A''(1) + A'(1) - A'(1)^2$$

$$= \frac{1}{q^2 p^{2m}} - \frac{2m+1}{qp^m} - \frac{p}{q^2}\,, \qquad (15)$$

It is interesting to compare Strategies A and B. It is evident that Strategy B is more stringent than Strategy A, since for $m=r$, obtaining m consecutive labels necessarily implies obtaining m total labels. Acquiring consecutive labels implies for example that, once a 1-label is acquired, no 0-label from that point is tolerated until all m 1-labels are accumulated. Thus, when a 0-label arises, it may be interpreted as having the effect of cancelling out any previous 1-label, and the same applies to the commencement of the 0-label.

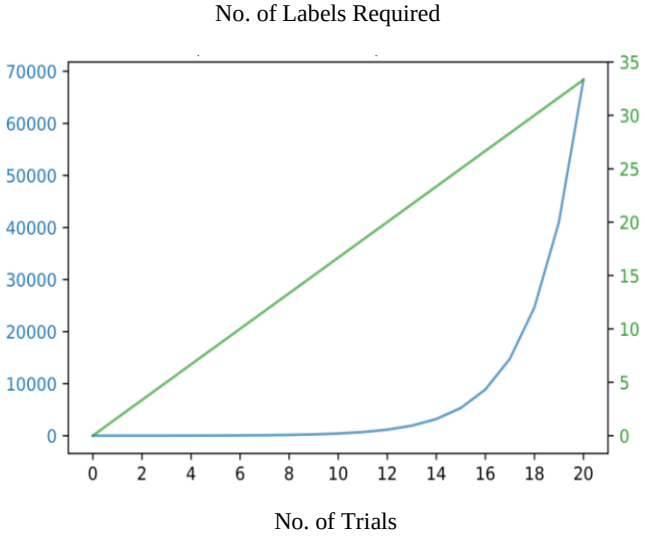

Fig. 1. Cost Comparison of Strategy A and Strategy B (p = 0.6).

Figure 1 compares the average cost of play of Strategy A and Strategy B. Here, the left vertical axis is used for $E(\boldsymbol{V_m})$ with an appropriate scale, while, the right vertical axis is used for $E(\boldsymbol{W_r})$. We see that the stringency of Strategy B is manifested in a steep climb in the number of trials as m increases, as opposed to a relatively moderate increase in Strategy A.

Depending on the outcomes of labelling, since Strategies A and B govern the time when the observation process terminates and accompanied by a decision being made, the underlying process may be regarded as a learning episode whereby a labelling is learned from the observations. Therefore, Strategies A and B may also be understood as learning strategies, and a resultant label of 1 would be referred to as learning success, whereas a resultant label of 0 would be referred to as learning failure.

III. The Ratio of Class Labels and Learning Success

We denote by ρ the ratio of the average number of negative labels to the number of positive labels; thus

$$\rho(p) = \frac{E[\boldsymbol{W_r}]}{r} = \frac{q}{p}\,. \qquad (16)$$

From the above relationship, we can determine the inherent stochastic structure of the environment by estimating p from actual observed labels ratio W/r, where W is the sample mean of $\boldsymbol{W_r}$. We can then form our estimator from the above just by solving for p. We shall estimate the probability P_b that the learning cost for this component exceeding this bound. From (7) above, this is given by

$$P_b = 1 - \sum_{k=0}^{b} \Pr[\boldsymbol{W_r} = k]$$

$$= 1 - \sum_{k=0}^{b} \binom{-r}{k} p^r (-q)^k. \qquad (17)$$

Here, the normal approximation can be invoked. In many reinforcement learning episodes, r tends to be under 100, as a lengthy iteration time is not feasible and most learning algorithms aim to converge in minimum time.

Clearly, the selection of the maximum cost weight b will have a significant impact on P_b. Very often, it is more meaningful to relate b to $\mathrm{E}[\boldsymbol{W_r}]$ either additively or multiplicatively. Table I tabulates the values of P_b for different values of b. The first part of Table I considers b by adding a fixed value d, with d = 5 and d = 10, while the second part considers b by multiplying by a fixed multiple α, with α = 1.2 and α = 1.5; here, b is rounded to the nearest integer. In the first part of Table I, we see that for either value of r, when p is appreciably greater than q, the probability of exceeding cost bounds tends to be acceptably small, and this is especially so for r = 20. The reason is that, since d is a fixed value, its relative contribution to b increases as p increases, produces a relatively large cost bound weight compared to the average one, and accordingly lowers the probability of exceeding the bound. However, in the second part of Table I, the difference between $\mathrm{E}[\boldsymbol{W_r}]$ and b decreases as $\mathrm{E}[\boldsymbol{W_r}]$ decreases, so that P_b tends to be large for higher values of p.

TABLE I. Analysis of Probabilities of Exceeding Cost Bounds

b Formula	r	p	q	$\mathrm{E}[W_r]$	b	P_b	P_b'	Err
$b = E[W_r] + d$ (d = 5)	20	0.5	0.5	20.00	25	0.215	0.186	0.029
		0.8	0.2	5.00	10	0.023	0.026	0.003
		0.9	0.1	2.22	7	0.001	0.004	0.003
	50	0.5	0.5	50.00	55	0.309	0.279	0.030
		0.8	0.2	12.50	17	0.127	0.108	0.019
		0.9	0.1	05.56	11	0.014	0.017	0.003
b =	20	0.5	0.5	20.00	30	0.057	0.059	0.002

b Formula	r	p	q	E[W_r]	b	P_b	P_b'	Err
$E[W_r] + d$ ($d = 10$)		0.8	0.2	5.00	15	0.000	0.001	0.001
		0.9	0.1	2.22	12	0.000	0.000	0.000
	50	0.5	0.5	50.00	60	0.159	0.147	0.012
		0.8	0.2	12.50	22	0.008	0.011	0.003
		0.9	0.1	05.56	16	0.000	0.000	0.000
$b = \alpha E[W_r]$ ($\alpha = 1.2$)	20	0.5	0.5	20.00	24	0.264	0.226	0.038
		0.8	0.2	5.00	6	0.345	0.253	0.092
		0.9	0.1	2.22	2	0.556	0.380	0.176
	50	0.5	0.5	50.00	50	0.159	0.147	0.012
		0.8	0.2	12.50	15	0.264	0.215	0.049
		0.9	0.1	05.56	7	0.280	0.207	0.073
$b = \alpha E[W_r]$ ($\alpha = 1.5$)	20	0.5	0.5	20.00	30	0.057	0.059	0.002
		0.8	0.2	5.00	7	0.212	0.156	0.056
		0.9	0.1	2.22	3	0.310	0.193	0.117
	50	0.5	0.5	50.00	75	0.006	0.010	0.004
		0.8	0.2	12.50	19	0.050	0.048	0.002
		0.9	0.1	05.56	8	0.163	0.121	0.042

In Table I, column P_b' gives the exact calculation using (17), while column P_b employs the normal approximation using (9). The absolute error between the exact calculation and the normal approximation is given by column *Err*. We see that the normal approximation is quite acceptable in most cases with absolute error less than 0.1. Note that no matter whether having b additively or multiplicatively related to E[$\boldsymbol{W_r}$], a higher value of d or α always gives smaller absolute error. We therefore suggest that the approximation should only be used when r, d and α are sufficiently large.

IV. A Learning Framework based on Competing Multi-agents

In Learning Strategies A and B above, the termination of a learning episode is triggered whenever a fixed number of positive labels r is obtained, irrespective of the number of negative labels accumulated in the process of doing so. Sometimes, however, this may not be desirable, especially when an inordinate number of negative labels have been accumulated, in which case, termination should take place earlier along with the conclusion of learning failure. Therefore, one is comparing the number of positive labels gathered against the number of negative labels, and the learning is concluded as success or failure according to which of these achieve the majority.

More precisely, this may be viewed as a multi-agent tournament with two competing agents A_1 and A_2, in which A_1 is responsible for giving out the positive labels, while A_2, the negative labels (respectively the 1 and 0 labels). This framework is not unlike the game theoretic approach in statistical decision theory, where both the statistician and nature are regarded as players in the game of estimation, and also this may be regarded as a kind of stochastic game involving agents [5][28][30]. While we shall focus on the agents A_1 and A_2, we note that there is a further agent, the learner, so that three agents exist in this situation. Here, when a classification results in a positive labels, then A_1 would gain a score of one, while when an observation results in a negative labels, then A_2 would gain a score of one. When either 1 or 0 label first reaches a given threshold h, then this will trigger a termination and the learning episode is concluded as success or failure according to which agent attains the threshold score first. Therefore, we have the following Learning Strategy.

Strategy C*: The learning process terminates when either agent, A_1 or A_2, first reach the threshold of either accumulating h labels of 1, or accumulating h labels of 0, which can then be concluded as a success or a failure according to which agent attains the threshold first.*

Here, without loss of generality, we shall let $h = 2m+1$ be odd, where m is an integer, and similar to Section II, we let p and q, with $p + q = 1$, signify the probabilities of receiving a positive labels, and negative labels, respectively for a particular classification. In other words, for a given classification, agent A_1 wins with probability p, while agent A_2 wins with probability q. In order to attain h for either agent, the number of classifications Ω will fall within the range

$$2m + 1 \leq \Omega \leq 4m + 1 .$$

If f_k represents the probability that A_1 wins at classifications number $4m+1-k$, which occurs if and only if A_1 scored $2m$ successes in the first $4m-k$ observations, and subsequently score a final success, then f_k is given by

$$f_k = \binom{4m-k}{2m} p^{2m+1} q^{2m-k} .$$

The probability that A_1 reaches the threshold first, irrespective of the classification number, is therefore given by

$$P_m = \sum_{k=0}^{2m} f_k = \sum_{k=0}^{2m} \binom{4m-k}{2m} p^{2m+1} q^{2m-k} .$$

That is, P_m gives the probability that the learning is successful (i.e., agent A_1 wins) according to Learning Strategy C.

Table II computes P_m for different values of p, q, and m for this tournament scenario. We see that, as expected, when $p = q = 1/2$, $P_m = 1/2$, since neither A_1 nor A_2 has any advantage over its opponent. As p increases, however, P_m will increase, reaching almost certainty as p increases beyond 0.8. If we regard p as a measure of A_1's winning ability per trial, then when $p >> q$, most trials will be scored by A_1, so that winning the entire game (i.e., reaching h first) is almost a certainty, and this is especially so for higher values of h. It is interesting to see that when h or m is sufficiently high (e.g., m=10), a moderate advantage for A_1 (e.g., $p = 0.6$) is enough to almost

guarantee success. On the other hand, $1-P_m$ gives the probability that agent A_2 wins, where the measure of A_2's winning probability per trial is given by q. For instance, when q=0.4, then A_2 stands a chance of around 27% of winning the game when m=2, and a chance of winning of around 10% when m=10.

TABLE II. PROBABILITIES OF LEARNING SUCCESS FOR THE TOURNAMENT SCENARIO

m	p	q	P_m	m	p	q	P_m
1	0.5	0.5	0.5000	5	0.5	0.5	0.5000
	0.6	0.4	0.6826		0.6	0.4	0.8256
	0.7	0.3	0.8369		0.7	0.3	0.9736
	0.8	0.2	0.9421		0.8	0.2	0.9990
	0.9	0.1	0.9914		0.9	0.1	1.0000
2	0.5	0.5	0.5000	10	0.5	0.5	0.5000
	0.6	0.4	0.7334		0.6	0.4	0.9035
	0.7	0.3	0.9012		0.7	0.3	0.9964
	0.8	0.2	0.9804		0.8	0.2	1.0000
	0.9	0.1	0.9991		0.9	0.1	1.0000

Returning to the estimation problem, by observing P_m, i.e., by computing the observed proportion of time that agent A wins, it is possible to infer the underlying probability p. While unlike in Section II, where an explicit formula exists linking directly the observations to the estimate, such explicit relationship is not available here. Nevertheless, as can be observed from Table II, useful estimation bounds can be drawn to determine whether $p > ½$ or $p < ½$. We see that it is quite reasonable to estimate $\hat{p} > ½$ whenever $P_m > ½$, and for most practical purposes, this would seem to be adequate.

In what follows, we shall apply the above analysis to network communications and the indexing of multimedia objects. While there are many situations that conform to the above framework such as those mentioned in the introduction, these particular applications are chosen partly because of their importance and partly because of their relevance to information processing in the present day big data era.

V. APPLICATION TO PACKET SWITCHING

In packet switching, suppose we wish to transmit a number of data packets over a noisy channel, where a successful error-free transmission occurs with probability p, and an erroneous transmission occurs with probability $q = 1 - p$. An erroneous transmission may, for example, be detected from the error-detection mechanisms when a packet is corrupted by random noise. Where the error-correction mechanism is able to correct the error despite the noise, the packet is regarded as a success, and this is incorporated into the probability p. For a message D consisting of r packets, we measure communications performance by examining the total number of transmissions required to achieve successful transmission of the entire message D. Let $\boldsymbol{T_1}$ be the time taken to successfully transmit a message consisting of r packets. An obvious analogy exists between the present situation and the multimedia information indexing situation above with respect to Learning Strategy A: we need a total of r classification of label 1 in order to achieve success. Consequently, we have the following results. Given a message consisting of r data packets, the total number of (error-free and erroneous) transmissions required in order to achieve a successful transmission of the entire message has mean and variance

$$\mathrm{E}[\boldsymbol{T_1}] = \frac{rq}{p}\ , \tag{18}$$

$$\mathrm{Var}[\boldsymbol{T_1}] = \frac{rq}{p^2}\ . \tag{19}$$

Moreover, the probabilities $\Pr[\boldsymbol{T_1} = k]$ is given by,

$$\Pr[\boldsymbol{T_1} = k] = \binom{-r}{k} p^r (-q)^k, \quad k = 0, 1, 2, 3, \ldots \tag{20}$$

This may be determined approximately by the normal distribution [17]

$$\boldsymbol{T_1} \sim N\left(\frac{rq}{p}, \frac{rq}{p^2}\right), \tag{21}$$

and,

$$\Pr[\boldsymbol{T_1} > t\ ] = \int_{\frac{tp-rq}{\sqrt{rq}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

$$= 1 - \Phi\left(\frac{tp - rq}{\sqrt{rq}}\right). \tag{22}$$

Next, suppose we have a message D broken up into m packets, and perhaps due to the error detection/correction or other requirements, it needs m consecutive successful transmissions to complete the entire transmission. In this situation, we again measure performance by examining the total number of transmissions required to achieve successful transmission of the entire message D. Let $\boldsymbol{T_2}$ be the time taken to successfully transmit the entire message consisting of m packets. Again, an analogy exists between the present situation and the multimedia information indexing situation above with respect to Learning Strategy B. Consequently, we have the following results.

$$\mathrm{E}[\boldsymbol{T_2}] = \frac{1 - p^m}{qp^m}\ , \tag{23}$$

$$\mathrm{Var}[\boldsymbol{T_2}] = \frac{1}{q^2 p^{2m}} - \frac{2m+1}{qp^m} - \frac{p}{q^2} . \quad (24)$$

VI. Application to Multimedia Search and Indexing

Many large enterprises rely on multimedia document repositories for their effective operation. However, unlike text-oriented objects, the retrieval of multimedia objects is often limited in their search and discovery mechanisms, since they do not readily lend themselves to automatic processing or indexing. The basic framework of the adaptive search mechanism is to capture human judgment in the course of normal usage from user queries in order to develop semantic indexes which link search terms to media objects semantics. This approach is particularly effective for the retrieval of such multimedia objects as images, sounds, and videos, where a direct analysis of the object features does not allow them to be linked to search terms, such as non-textual/icon-based search, deep semantic search, or when search terms are unknown at the time the media repository is built. The above model is able to represent such an adaptive search mechanism by making use of naïve Bayes classification approach based on the three learning strategies indicated. This approach allows for the efficient creation and updating of media indexes, which is able to instill and propagate deep knowledge relating to the enterprise functions into the media management system concerning the advanced search and usage of multimedia resources.

Thus, the above positive and negative classifications may be viewed as a learning sequence in multimedia indexing acquired from user interaction. Here, we are concerned with the status of the association of given search terms to particular multimedia objects. Through the interaction with users, positive and negative labels are handed out probabilistically. In the case of search terms to multimedia objects association, learning success would mean that the association in question is sound and should be incorporated as proper index, while failure would mean that the search term-object association cannot be established.

Similar to the previous application, the time to install an index term based on Strategy A would take a time of $\boldsymbol{I_1}$, with mean and variance given by

$$\mathrm{E}[\boldsymbol{I_1}] = \frac{rq}{p} , \quad (25)$$

$$\mathrm{Var}[\boldsymbol{I_1}] = \frac{rq}{p^2} . \quad (26)$$

In addition, we have,

$$\Pr[\boldsymbol{I_1} = k] = \binom{-r}{k} p^r (-q)^k, \quad k = 0, 1, 2, 3, \ldots \quad (27)$$

Similarly, the time $\boldsymbol{I_2}$ to install an index term based on Strategy B has mean and variance

$$\mathrm{E}[\boldsymbol{I_2}] = \frac{1-p^m}{qp^m} , \quad (28)$$

$$\mathrm{Var}[\boldsymbol{I_2}] = \frac{1}{q^2 p^{2m}} - \frac{2m+1}{qp^m} - \frac{p}{q^2} . \quad (29)$$

Furthermore, from Section IV, under Learning Strategy C with threshold *h*, the probability of successful installation of an index term is given by

$$\sum_{j=0}^{h-1} \binom{2h-j-2}{h-1} p^h q^{h-j-1} .$$

VII. Conclusion and Future Extension

We have presented a model of binary classification, operating in a stationary stochastic environment in the presence of noise. Stochastic methods are essential because various operating environments are often noisy and seldom static nor deterministic, and the use of probabilistic methods is therefore an unavoidable necessity. Indeed, if stochastic elements are absent, the same outcome will always occur, and repeated observations, and hence repeated reinforcements, are unnecessary. A unified probabilistic framework is developed for such a classification scenario. We first consider a situation where the cumulative number of classifications is pre-specified and fixed, which constitute the criterion for terminating the learning process. Two variations of this process are considered, one requires non-consecutive reinforcements and the other requires consecutive reinforcements. By observing the random positive to negative labels ratio, a meaningful estimation of either learning success or failure may be arrived at. In most practical situations, the cost of securing a classification can be significant, and this has been incorporated into our model, and we have obtained the probabilities of exceeding the classifications cost bounds.

A multi-agent framework where the handing out of positive and negative labels are viewed as being performed by agents have also been considered. Thus, the final learning outcome is determined by a kind of stochastic game with the agents competing against each other. The termination criterion here is determined by when and how the game is won. The respective probabilities of learning success and failure are also explicitly derived. Closed-form expressions of other relevant measures of interest are obtained. A procedure for estimating the underlying stochastic structure from the observed random agent winning frequencies is also developed.

The above results and algorithms have been applied to study the performance of human-assisted semi-automatic multimedia information indexing as well as to quantify communications network transmission performance operating in a noisy channel.

In this paper, we have employed the naïve Bayes assumption and assumed that positive labels and negative labels occur statistically independently. In the future, it may be more general to relax this assumption and incorporate different forms of dependencies into the model, such as single-step or multi-step Markov Chain conditional dependency.

REFERENCES

[1] C. H. C. Leung and Y. Tong, "Application of learning reinforcement classification to netowork communications and multimedia search", *Proc. International Conference on Advanced Engineering Computing and Applications in Sciences, ADVCOMP 2021*, Spain, 3-7 October, 2021.

[2] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," In *Int. Conf. on Machine Learning*, 1998.

[3] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey" *Journal of artificial intelligence research*, vol. 4, pp. 237-285, 1996.

[4] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum Entropy Inverse Reinforcement Learning," *Proc. Twenty-Third AAAI Conference on Artificial Intelligence (AAAI 08)*, vol. 8, pp. 1433-1438, 2008.

[5] H. Santana, G. Ramalho, V. Corruble, and B. Ratitch, "Multi-agent patrolling with reinforcement learning," *Proc. Third International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 3, pp. 1122-1129, IEEE Computer Society, 2004.

[6] R. I. Brafman and M. Tennenholtz, "R-max-a general polynomial time algorithm for near-optimal reinforcement learning," *Journal of Machine Learning Research*, vol.3, pp. 213-231, 2002.

[7] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous agents and multi-agent systems*, vol. 11, no. 3, pp. 387-434, 2005.

[8] E. Ipek, O. Mutlu, J. F. Martínez, and R. Caruana, "Self-optimizing memory controllers: A reinforcement learning approach," *ACM SIGARCH Computer Architecture News*, vol. 36, no. 3, IEEE Computer Society, 2008.

[9] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, And Cybernetics-Part C: Applications and Reviews*, vol. 38, no. 2 2, 2008.

[10] S. V. Albrecht and P. Stone, "Autonomous agents modelling other agents: A comprehensive survey and open problems," *Artificial Intelligence* 258, pp. 66-95, 2018.

[11] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PloS one*, vol. 12, no. 4: e0172395, 2017.

[12] A.W. Moore and C.G. Atkeson, "Prioritized sweeping: Reinforcement learning with less data and less time," *Machine learning*, vol. 13, no.1, pp. 103-130, 1993.

[13] E. Brochu, V. M. Cora, and N. De Freitas, "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning,", unpublished.

[14] Q. Wei, F. L. Lewis, Q. Sun, P. Yan, and R. Song, "Discrete-time deterministic Q-learning: A novel convergence analysis," *IEEE Transactions on Cybernetics*, vol. 47, no. 5, pp. 1224-1237, 2017.

[15] C. J. Watkins and P. Dayan, "Q-learning," Machine learning 8.3-4 pp. 279-292, 1992.

[16] H. Van Hasselt and M.A. Wiering, "Using continuous action spaces to solve discrete problems," *Proc. International Joint Conference on Neural Networks (IJCNN 09)*, pp. 1149-1156. IEEE, 2009.

[17] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)," *Evolutionary computation*, vol. 11, no. 1 pp. 1-18, 2003.

[18] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd Edition, Wiley & Sons, 1968.

[19] Q. Huang, A. Puri, and Z. Liu. "Multimedia search and retrieval: new concepts, system implementation, and application". *IEEE Transactions on circuits and systems for video technology*, 2000, 10.5: 679-692.

[20] R.Gupta, M. Khomami Abadi, J. A.Cárdenes Cabré, F.Morreale, T. H. Falk, and N. Sebe, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing". *Proc. ACM International Conference on Multimedia Retrieval*. ACM, p. 317-320, 2016.

[21] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, 34(6), 26-38, 2017.

[22] X. Yao, J. Du, N. Zhou, and C. Chen, "Microblog Search Based on Deep Reinforcement Learning," *In Proceedings of 2018 Chinese Intelligent Systems Conference* (pp. 23-32). Springer, Singapore, 2019.

[23] Y.C. Wu, T. H.Lin, Y. D. Chen, H. Y Lee, and L. S. Lee, "Interactive spoken content retrieval by deep reinforcement learning". arXiv preprint arXiv:1609.05234, 2016.

[24] S. Lan, R. Panda, Q. Zhu, and A. K. Roy-Chowdhury, "FFNet: Video fast-forwarding via reinforcement learning", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6771-6780, 2018.

[25] R. Hong, Y. Yang, M. Wang, and X. S. Hua, "Learning visual semantic relationships for efficient visual retrieval", *IEEE Transactions on Big Data*, 1(4), 152-161, 2015.

[26] R. Yan, A. Hauptmann, and R.Jin, "Multimedia search with pseudo-relevance feedback. In International Conference on Image and Video Retrieval" (pp. 238-247). *Springer*, Berlin, Heidelberg, 2003.

[27] E. Manino, L. Tran-Thanh, and N. R. Jennings. "On the Efficiency of Data Collection for Multiple Naïve Bayes Classifiers," Artificial Intelligence, 275: 356–378, 2019.

[28] J. Deng and C. H. C. Leung, "Dynamic Time Warping for Music Retrieval Using Time Series Modeling of Musical Emotions," *IEEE Transactions on Affective Computing*, Vol. 6, No. 2, pp. 137-151, 2015.

[29] H. L. Zhang, C. H. C. Leung, G. K. Raikundalia, "Topological analysis of AOCD-based agent networks and experimental results." *Journal of Computer and System Sciences*, pp. 255–278, 2008.

[30] I. Azzam, C. H. C. Leung, J. Horwood, "Implicit concept-based image indexing and retrieval." In *Proceedings of the IEEE International Conference on Multi-media Modeling*, pp. 354-359, Brisbane, Australia 2004.

[31] H. Zhang, C. H. C. Leung and G. K. Raikundalia, "Classification of intelligent agent network topologies and a new topological description language for agent networks." In *Proceedings of the 4th International Conference on Intelligent Information Processing*, Adelaide, Australia, pp. 21-31 2006.

[32] N. L. J. Kuang, C. H. C. Leung, and V. Sung, "Stochastic Reinforcement Learning." In *Proc. IEEE International Conference on Artificial Intelligence and Knowledge Engineering*, pp. 244-248, California, USA 2018.

[33] N. L. J. Kuang and C. H. C. Leung, "Performance Dynamics and Termination Errors in Reinforcement Learning – A Unifying Perspective," In *Proc. IEEE International Conference on Artificial Intelligence and Knowledge Engineering*, pp. 129-133, California, USA 2018.

[34] V. Sutton *et.al*., "Multi-armed bandit algorithms and empirical evaluation," In *Proc. 16th European Conference on Machine Learning*, Springer, Porto, Portugal, pp. 437–448, 2005.

[35] N. L. J. Kuang and C. H. C. Leung, "Analysis of Evolutionary Behavior in Self-Learning Media Search Engines," in *Proceedings of the 2019 IEEE International Conference on Big Data*, Los Angeles, USA, 2019.

[36] N. L. J. Kuang and C. H. C. Leung, "Performance Effectiveness of Multimedia Information Search Using the Epsilon-Greedy Algorithm," in *Proceedings of the 2019 IEEE International Conference on Machine Learning and Applications*, pp. 929-936, Florida, USA 2019.

[37] E. Manino, L. Tran-Thanh, and N. R. Jennings. "On the Efficiency of Data Collection for Crowdsourced Classification." In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 1568-1575, 2018.