Sports Analytics: Predicting Athletic Performance with a Genetic Algorithm

Full paper

Victor Cordes
Claremont Graduate University
victor.cordes@gmail.com

Lorne Olfman
Claremont Graduate University
Lorne.olfman@cgu.edu

Abstract

Existing predictive modeling in sports analytics often hinges on atheoretical assumptions winnowed from a large and diverse pool of game metrics. Feature subset selection by way of a genetic algorithm to identify and assess the combinatorial advantage for a group of metrics is a viable option to otherwise arbitrary model construction. However, this approach concedes similar arbitrariness as there is no general strategy or common practice design among the tightly coupled nucleus of genetic operators. The resulting dizzying ecosystem of choice is especially difficult to overcome and leaves a residual uncertainty regarding true strength of output, specifically for practical implementations. This study transposes ideas from extreme environmental change into a quasi-deterministic extension of standard GA functionality that seeks to punctuate converged populations with individuals from auxiliary metas. This strategy has the effect of challenging what might otherwise be considered shallow fitness, thereby promoting greater trust in output against innumerable alternatives.

Keywords

Genetic algorithm, feature subset selection, sports analytics

Introduction

The sparse and fragmented nature of the academic sports analytics domain has as a consequence athoeretical assumptions during model development. Stated differently, without existing theory from which to guide and justify model construction, assumptions are at best capricious and at worst tightly coupled with researcher bias. Yet, professional sporting franchises are increasingly seeking to establish clear quantitative demonstrations of the value for a data asset (Davenport, 2014). Said asset is typically of non-trivial dimensionality, containing numerous and disparate measures of athletic performance, game-day measurements, etc. However, without the benefit of theory, assumptions must be made that limit inclusion of noisy variables when developing inductive techniques in order to improve overall model accuracy. From a strategic perspective, reliance on conclusory assumptions as a consequence of a theory-thin landscape is problematic. Nevertheless, this standard does not intrinsically diminish the potential for knowledge discovery in sports analytics research.

Sports are big business, with an estimated total market value of hundreds of billions of dollars in the United States and ranking as one of the top-10 business markets globally (Fry and Ohlmann, 2012). An increasing number of sports organizations are implementing analytical approaches to decision-making, although its wholesale adoption remains remote (Davenport, 2014). Nevertheless, coupled with a recent surge in the level of scholarly interest, sports analytics offer a fertile research environment, however fragmented, as highlighted by lack of academic programs devoted to sports analytics (SA) with few and far between research forays into sports (Coleman, 2012).

According to Davenport (2014), who conducted a series of interviews with 25 sports teams and vendors in the US and Europe, analytical initiatives in the domain of professional sports boast impressive growth and activity, including increasing output channels for analytics, multiple analytical domains of interest, annual growth for one of its major conferences (Sloan Sports Analytics), etc. Among Davenport's conclusions is a need to move toward predictive and prescriptive analytics as current activities in professional sports continue to be mostly descriptive.

To that end, the research described herein applies iterative design science to develop and test a predictive model whose purpose is to improve athletic game-day performance forecasts for a skill position of a specific sport. Principles of feature subset selection are combined with a tailored genetic algorithm (GA), a flexible optimization heuristic based on population genetics. According to Yang and Honavar (1998) feature subset selection (FSS) refers to the task of identifying and selecting a useful subset of attributes to be used to represent patterns from a larger set of attributes. The pairing of FSS and GA is not often experimented with in academic sports analytics research, yet retains the advantage of minimizing the aforementioned assumption-making difficulties.

Literature Review

Sports Analytics

Academic sports analytics research remains fragmented and incoherent, comprising a mélange of techniques providing no consensus on general approaches. This fog is attributable to the competitive and secretive nature of professional sports (i.e., analytic results and methods are kept internal) in addition to the lack of continued pursuit by those authors that do engage the domain (Coleman, 2012). Historically, the research emphasis has focused on the efficiency of sports betting markets (Stekler, Sendor, and Verlander, 2010).

A cursory examination of machine learning (i.e., how computers can learn or improve their performance based on data, Han, Kamber, and Pei, 2011, p.24) and sports analytics literature highlights this fragmented nature. Neural networks application dates back decades, from NCAA college football objective ranking (Wilson, 1995), javelin flight prediction (Maier, Wank, and Bartonietz, 2000), NFL winner prediction as compared with media broadcasters (Purucker, 1996; Kahn, 2003), cricket performance prediction (Iyer and Sharda, 2009), etc.

Genetic algorithms are less popular, but still appear in a variety of applications, including optimal Formula One car performance (Wloch and Bentley, 2004), scheduling (Trick and Yildiz, 2012), NCAA college football ranking methodology (Cassady, Maillart, and Salman, 2005), cricket team composition (Ahmed, Deb, and Jindal, 2013), etc.

Ultimately, one could read every published article even tangentially related to sports analytics from the last 30 years and make no progress towards appropriate feature inclusion within their models. Song, Boulier, and Stekler (2007) happened upon this difficulty when they conducted an investigative survey comparing 70 expert forecasts vs. 31 statistical models predicting the outcome of American football games. While detailing their data collection method for the statistical models, they passingly mention "the variables that are used to construct the models differ widely" (p. 407), almost as if it were a given consequence of model construction.

Feature Subset Selection

Verleysen (2003) explains how the individual variable count in a vector describes its dimensionality, a characteristic that has important consequences in data mining. Further, he states that higher dimensionality datasets are more difficult to understand, both cognitively and from the perspective of extracting information to draw meaningful conclusions.

Real-world datasets can be characterized by high dimensionality with low sample sizes. Observations are made up of many variables, either due to the nature of the domain or simply because the lack of understanding begets collecting as much information about the data as possible prior to analysis. In either case, it stands to reason that the full complement of variables within the feature vector is not necessary to achieve successful data mining. Features may be irrelevant, redundant, or altogether noisy. In an effort to improve classification accuracy, an algorithm should only focus on the most relevant subset of data.

Feature subset selection is a domain-agnostic method. This field is of particular interest as its data deals with familiar problems of high dimensionality and limited sample size. Of the three FSS variants (filter, wrapper, embedded), the wrapper technique is tightly coupled to a chosen induction algorithm. Various feature subsets are selected and input into the learning algorithm. Said subsets are processed through an accuracy evaluation metric and ranked accordingly. In this way, the wrapper technique represents a more holistic approach to subset selection.

Genetic Algorithms

Genetic algorithms (GA) belong to the optimization class of machine learning algorithms and are loosely based on the concepts of population genetics. The technique first appeared in Holland (1975). The principles of evolution, natural selection, and survival of the fittest coalesce into a robust searching heuristic capable of identifying near-optimal solutions to a wide range of real-world problems, often characterized by large problem spaces (Beasley, Bull, and Martin, 1993).

In natural selection, individuals from a population compete with one another for survival. Fit individuals endure, mate, and pass on their genetic characteristics to their offspring. Successive generations are increasingly composed of good characteristics such that the species as a whole evolves relative to its environment. Genetic algorithms mimic these processes in data provided two prerequisites are met. First, the target problem must be suitably represented (e.g., binary encoding or bit string, also referred to as a chromosome). Second, encoded solutions require an appropriate fitness function to quantitatively measure their strength (i.e., a user-defined function returning a single figure of merit for any given solution). Genetic operators including selection, recombination (crossover), mutation, and replacement work together towards solution convergence.

Parameter setting research in the evolutionary computing domain has received substantial attention. All aspects of said domain are considered from theoretical and practical perspectives among deterministic, adaptive, and self-adaptive approaches. Karafotias, Hoogendoorn, and Eiben (2014) provide a comprehensive survey of 234 research articles covering parameter control techniques for core components including population, crossover, mutation, selection, fitness function, and more advanced approaches of parallel EAs that manage multiple subpopulations while governing their interactions.

A particularly interesting corollary of this fog is the equally broad research into premature convergence over the last two decades. Premature convergence is the general condition whereby population diversity is degraded before the fitness landscape is properly explored, resulting in redundant individuals concentrated around local optima (Friedrich, Oliveto, Sudholt, and Witt, 2009). Pandey, Chaudhary, and Mehrotra (2014) provide a comprehensive surveyed review of approaches to prevent premature convergence in GA between 1993 and 2013. All told, 168 articles are compiled producing 24 distinctive approaches, ranging from simple chromosome representation adjustment to heavily involved hybrid techniques implementing multiple crossover and mutation strategies.

Problem Definition

Evidently, parameter setting and premature convergence literature are symptomatic of the same chaotic environment that projects a plethora of tactics applied on many different kinds of problems. Likewise, Karafodias et al. note that "…no parameter control method or strategy has been widely adopted by the community or has become part of the common practice toolkit of evolutionary computing" (p.15). In effect, evolutionary computing—and by extension genetic algorithms—function as a supposed unity of parts awkwardly devoid of a unifying canon. Thus, the design of a heuristic artifact capable of outputting the ultimate correct solution is an illusion. It cannot be arrived at mathematically nor can it occur by process of elimination—there are too many design alternatives with respect to predictive modeling. Further, the secretive nature of professional sports, wherein statistical/predictive methods are internal and not published (Coleman, 2012), as well as the void of academic literature regarding evaluation of individual performance, makes for an unobvious status quo against which to compare a practitioner's model.

Sports analytics is a wicked organizational problem distinguished by pervasive uncertainty. This uncertainty lies beneath the surface of the algorithmic landscape, irrigating core problems of 1) identifying a satisfactory status-quo against which future techniques can be measured, 2) assuring future comparisons are fair, and 3) having confidence in the lasting quality of the output against an untold number of alternatives. Ultimately, from a practical perspective, the goal of maximal output yields to a sense of trust in output.

Remedial feature subset selection, by way of a genetic algorithm, exploits combinatorial advantages for a given observation. However, such an evolutionary strategy cannot escape a genetic infrastructure of tightly coupled operators whose joint effects are poorly understood; the resulting suboptimal mix further obstructs potential artifact effectiveness. An ecosystem with a multitude of choices undercuts ambitions for competitive testing among tailored algorithms in literature—to say nothing of commercial applications likely reinforced by concealed proprietary methods. Thus, the focus to find a solution shifts to productive investigation of performance elements responsible for algorithm effectiveness.

Solution

In the proposed implementation, problem representation is straightforward; binary cardinality suffices. Thus, a chromosome having length ℓ number of genes encapsulates a solution over the alphabet {0, 1} where a 0 excludes the feature from the vector to be trained by an induction algorithm. The standard genetic algorithm (SGA) randomly initializes a population of individuals. Assuming individual athletic performance is a function of 83 independent features of interest from a trusted data source (this will be shown later), then the expected value of activated features—allele frequency—at population initialization per chromosome will be 41.5 with a variance of 20.75 and standard deviation of 4.55 (see Figure 1).

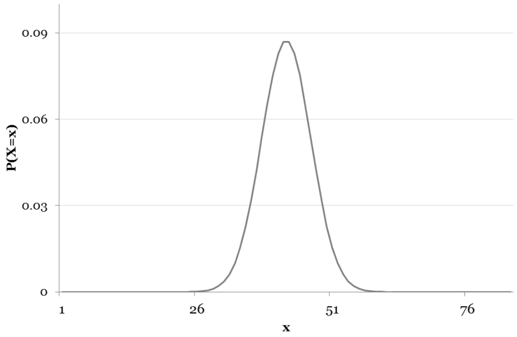

Figure 1. Discrete Distribution Population Allele Frequency

Basic probability theory straightforwardly handicaps the starting population of the standard genetic algorithm from the outset. Left to its own devices, this initial population will not be expected to include all allele frequencies, irrespective of the preset number of solutions. Worse yet, the probability that a vector will contain activated genes $P(X \geq 57) \vee P\ (X \leq 27)$ is effectively zero. There is neither logical premise nor foundation in literature for believing that the correct combination of features for predicting athletic performance is within the probabilistic range of Figure 1. From the perspective of the building block hypothesis, a converged SGA run whose evaluated chromosome set across all generations having allele frequencies mimicking the distribution of Figure 1 begs an obvious question: what guarantee is there that the solution was appropriately exposed to good building blocks? Stated differently, the genetic algorithm will not have access to the entire search space thereby reducing confidence its output is the true global optimum.

Extreme Environmental Change

An elementary understanding of organic evolution is typified as a process of slow and gradual change leading to improved form. Irrespective of said improvement, the hostility of natural environments continually threatens evolved organisms (e.g., volcanic eruptions, floods, etc.). Hoffmann and Parsons (1997) evaluate the ways environmental fluctuations influence. They describe how tropical species accustomed to warm temperatures died following short spells of 5°C (p. 3). Similarly, an El Niño event in the early 1980s causing elevated water temperatures and subsequent drop in nutrient availability resulted in high death rates among surrounding groups of organisms (p. 4).

Eldredge and Gould (1972) published a landmark paper proposing an alternate theory for the macroevolution of species known as punctuated equilibrium. At the time, the prevailing model of evolution accepted Darwinian principles of slow and gradual divergence. However, Eldredge and Gould found little empirical evidence in fossil records, and later concluded that once formed species will remain in a prolonged stasis that is episodically "punctuated" by a major shift leading to evolutionary change.

The consequences of rare stresses on organisms as characterized by Hoffmann and Parsons suggest that an evolved group may only be fit relative to a familiar environmental category defined by its underlying properties, hereafter referred to as its meta. If a meta undergoes dramatic shift, the organisms are necessarily faced with unfamiliar conditions for which they have not adapted to and can subsequently die from. Recalling the tropic species example, their evolution occurred according to an environmental meta characterized by warm temperatures but could not tolerate a meta-shift towards relatively cold temperatures. From this perspective, evolution as a whole was shallow; fitness did not extend beyond a strict category.

The proposed GA heuristic will metaphorically adopt the consequences of an environmental meta-shift together with ideas borrowed from punctuated equilibrium theory (see Figure 2). Upon GA convergence, the environment composed of supposed evolved solutions is considered in stasis; the average fitness of the population can no longer improve, or the best known solution is not beaten. Thereafter, a subroutine will initiate a meta-shift scheme that specifically punctuates said environment by introducing new solutions within a range of explicit allele frequency, thereby forcibly increasing variability in the gene pool. This altered population re-engages the evolutionary workflow undergoing selection, crossover, and mutation procedures. The iterative cycle continues until a user defined limit on meta-shifts is exceeded (e.g., number of shift without best solution improvement).

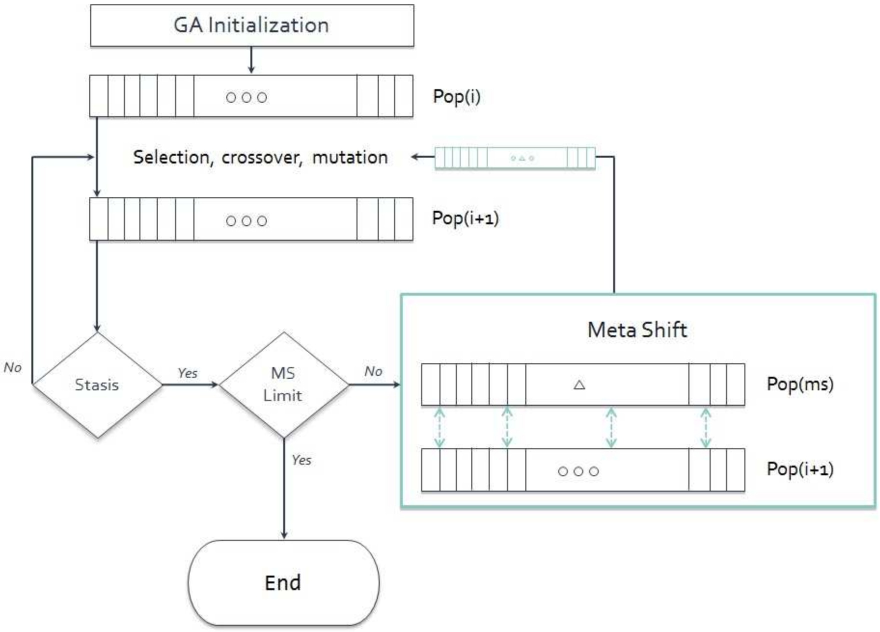

Figure 2. Meta-Shift Subroutine

Given properties of binomial random variables and discrete distributions, the new solutions will not be random but rather compensate for otherwise ignored pockets of the search space demarcated by intervals of standard deviation. Thus, the goal is to challenge the true fitness of the converged solution that might otherwise only be fit relative to a specific meta. The subsequent "biotic" interactions are anticipated to be intense with the resulting competition enhancing solution quality and robustness in concert with crossover and mutation schemes.

Data and Application

Relevant data was acquired by using web scrapping to collect multiple summary game statistics of a targeted sport (the National Football League) from a reputable sports media conglomerate (ESPN). Total individual rushing yardage production for a single game is the observation of interest. Altogether, season 2014 featuring 256 games over 17 weeks (playoff contests are excluded) was obtained. For over a century individual games have been summarized in a box score, a tabulated result containing statistics of various team and individual player performance. Roughly seventy of these descriptive statistics, for both the home and away team, were targeted for scraping and stored locally for processing. The primary groupings include passing, receiving, defense, punting, kicking, and team production. In short, all aspects of a single game are considered; none are arbitrarily discarded a priori. The data is retrieved and exported into a format suitable for consumption within the R statistical package.

To create a feature vector, the following rules are adhered to. A list of unique running back names is generated. For every name in said list, that player is searched throughout the season database document during the given week. The following repeats every week in the season, 1-17: once identified, obtain the running back's total yardage production for the game (dependent variable); for every week prior to the projected week, search the player's performance; for every game identified in which the player participated, gather box score statistics; separate the player's rushing production from the rest of his team, but retain both (i.e., a vector will include individual rushing performance and the remainder of the team total rushing performance); identify the projected week opponent and gather its defensive production for every week prior; combine the information from previous steps; compute the mean of the statistics, rounding to clean numbers; eliminate any vector with average total rushing carries below five; categorize the dependent variable (e.g., lowest, low, medium, high, highest), incremented by a fixed number of yards (25).

Accordingly, if player P has rushing production in week 8, the corresponding vector contains 83 independent variables representing the mean of player and game statistics said player participated in between weeks 1-7, together with the projected week opponent's average defensive production to that point in the season. In this way, the feature vector considers the player and the forthcoming opponent, which is generally how match-ups are conceptualized. A player may be effective at a position, but he is rarely discussed in a void; rather, analysis occurs in context relative to opponent capabilities.

The total number of independent variables (83) is arrived at as follows: team stats (21), team defensive (7), opposing defensive (7), team passing (10), team receiving (6), individual rushing (5), team rushing (5), team kick return (5), team punt return (5), team kick scoring (7), team punting (6), and next week opposing defensive (7), bringing the total number of features to 91. However, among them there are redundancies which are removed during data pre-processing. For instance, among team stats (21) there are several measurements which are repeated in team passing and team defensive, including total passing yards, completed passes, yards per pass, interceptions, etc. Altogether, 83 is the final number.

Evaluation

Encoded solutions require an appropriate fitness function to quantitatively measure their strength. This research incorporates a decision tree learning algorithm (C5.0); however, there are many alternatives to choose from (e.g., support vector machines, neural networks, etc.). In this implementation (wrapper FSS) the induction algorithm has no autonomy in the process of identifying candidate subsets, thus winnowing capabilities are disabled.

Observing the fitness of chromosomes guides the GA through the solution space. While critical to the evolutionary process, it nonetheless remains an intra evaluation with respect to the greater perspective that would emphasize overall artifact construction and evaluation. Although fit chromosomes may evolve and estimate accurately on resampled data, the artifact's overall utility remains unproven. Therefore, this research reinforce the intra C5.0 evaluation of solution candidates during cross-validated model training with an outer C5.0 evaluation procedure on the best evolved chromosome upon GA termination.

Accordingly, prediction of running back performance for the 12th week of an NFL season would first occur by deploying a genetic algorithm run against weeks 1-10. Vectors are evaluated using k-fold cross-validation. Upon convergence, the GA outputs a single feature subset with the greatest fitness. This subset is used to train a final and independent C5.0 model on week 11 data whose classification of player performance category is compared with actual week 12 production giving a final determination of overall artifact accuracy.

However, overall accuracy alone and presented without context is difficult to assess. In order to provide perspective, accuracy metrics from six sources are included. Having 83 independent variables in a binary encoded genetic algorithm assumes a guided search across a solution space in the septillions ($2^{83} = 9.67 \times 10^{24}$). Enumeration is not possible. Thus, the only reasonable single solution to consistently include is one in which all 83 features are activated—a chromosome with 83 trailing 1s. This evaluated solution has the added benefit of answering a basic question as to whether or not feature subsets are even required for enhanced accuracy. Either way, a standard genetic algorithm run is the second source for evaluation. Additionally, this research presents an extension of core SGA functionality based on a metaphor of extreme environmental change. Thus, a third source of evaluation is an executed meta-shift GA. The fourth source attempts to contextualize the research against some understanding of an industry standard. The same data source used to retrieve box score data (ESPN) also provides fantasy rushing projections for the 2014 season (note: the ESPN numeric projection must be categorized to align with research methodology). The fifth source further contextualizes the research against an alternate feature subset selection method. In an independent test using all activated features, the winnowing option of a decision tree is activated. In this way, the induction algorithm stands alone. It assumes complete responsibility for picking and choosing among the predictors used to construct and evaluate a subsequent tree. All five sources exist against the backdrop of the fifth and most basic requiring no model whatsoever, the no information rate (NIR).

Artifact performance is observed in stepwise fashion for the 2014 season, or week-to-week beginning with the 4th and ending with the 10th. Evaluation for a single week provides meager testing samples (even more so given the constraint of average carry limitation), therefore this perspective is at once a more challenging task but nevertheless more interesting as it closely approximates a real-life use case (e.g. fantasy sports industry). It is not uncommon for a single week to hover around 50 vectors—the minority class may have less than 10 instances.

Findings

The results of the various experiments, NIR included, appear in Table 1. Boosting was disabled for the first four of the seven-week batch (4 through 10). GA mutation rates are fixed at 0.1, crossover is of the single-point variant, and tournament selection is used. In this way, the meta-shift subroutine performance is compared against a canonical GA, the basis for theoretical GA research (Affenzeller, Wagner, Winkler, and Beham, 2009). Shaded meta-shift cells are due to inaccurate recordings for the particular experiment.

	Pop	Stasis	Meta Shifts	K-CV	Total Solutions	Test Accuracy
Week 4 (NIR = 0.3878)						
ESPN	▨	▨	▨	▨	▨	0.3469
All	▨	▨	▨	▨	▨	0.4286
All W	▨	▨	▨	▨	▨	0.4286
GA	15	▨	▨	5	1069	0.3673
GA MS	15	8	▨	5	1562	0.5510
Week 5 (NIR = 0.3636)						
ESPN	▨	▨	▨	▨	▨	0.3636
All	▨	▨	▨	▨	▨	0.2909
All W	▨	▨	▨	▨	▨	0.2909
GA	15	▨	▨	5	576	0.4000
GA MS	10	8	▨	3	1914	0.3818
Week 6 (NIR = 0.3529)						
ESPN	▨	▨	▨	▨	▨	0.3922
All	▨	▨	▨	▨	▨	0.4118

All W	▨	▨	▨	▨	▨	0.3725
GA	26	▨	▨	3	451	0.4510
GA MS	15	8	23	3	2772	0.4706
Week 7 (NIR = 0.3654)						
ESPN	▨	▨	▨	▨	▨	0.3462
All	▨	▨	▨	▨	▨	0.1731
All W	▨	▨	▨	▨	▨	0.3462
GA	26	▨	▨	3	578	0.3269
GA MS	50	8	▨	3	13020	0.3654
Week 8 (NIR = 0.3191)						
ESPN	▨	▨	▨	▨	▨	0.3404
All	▨	▨	▨	▨	▨	0.2979
All W	▨	▨	▨	▨	▨	0.4255
GA	26	▨	▨	3	960	0.4043
GA MS	26	7	20	3	1998	0.4468
Week 9 (NIR = 0.3953)						
ESPN	▨	▨	▨	▨	▨	0.3023
All	▨	▨	▨	▨	▨	0.3488
All W	▨	▨	▨	▨	▨	0.2791
GA	26	▨	▨	3	783	0.2791
GA MS	20	6	28	3	3395	0.3953
Week 10 (NIR = 0.4717)						
ESPN	▨	▨	▨	▨	▨	0.3396
All	▨	▨	▨	▨	▨	0.4340
All W	▨	▨	▨	▨	▨	0.3962
GA	26	▨	▨	3	432	0.4340
GA MS	20	6	22	3	2657	0.4906

Table 1. Results

Key: Pop=Population Size, Stasis=counter delineating number of generations without improvement to best solution before population is considered converged, Meta Shifts=number of times subroutine is called during the entirety of the run, K-CV=cross-validation partitions, "All" represents the vector with all 83 genes having bit value 1, "All W" executes a decision tree with its winnowing parameter activated

The meta-shift GA outperforms all other output (with the exception of week 5) while consistently overcoming the NIR where others cannot, notably in week 10 in which it was particularly elevated (0.4717). The MS GA individual performance gains range from 4.3% (week 6) to 28.5% (week 4) with an average of 9.3% over nearest competition from any of the remaining outputs across all seven weeks. Comparatively, the meta-shift approach improves on ESPN projections by an average of 28%, against SGA by 16%, on output that considers all features by 30%, and against an alternate feature subset selection method by 23%. The most interesting singularity occurs in the 4th week with testing accuracy of 0.551 (or 55%), encroaching on gambling market efficiency which is perhaps the objective gold standard of any predictive modeling technique in the domain of sports.

A bloated number of total solutions tested for the meta-shift approach is a natural consequence of re-engaging the search multiple times during the course of a single run. It is also a function of luck. If the first few shifts randomly settled upon weaker subsectors of space that nonetheless contained stronger individuals than the initially evolved population, then the resulting punctuating solutions would necessarily be defeated in time thereby resetting the MS convergence counter while accumulating an increased number of evaluated solutions. Conversely, if the first shifts happened upon the strongest pockets of the search space, ensuing shifts would fail to produce superior alternatives; thus, convergence criteria would be met much sooner and execution time would hasten. Of course, it is difficult to anticipate the appropriate search trajectory when operating against an unknown fitness landscape a priori.

An example of meta-shift allele frequency distribution is given in Figure 3 for the 10th week.

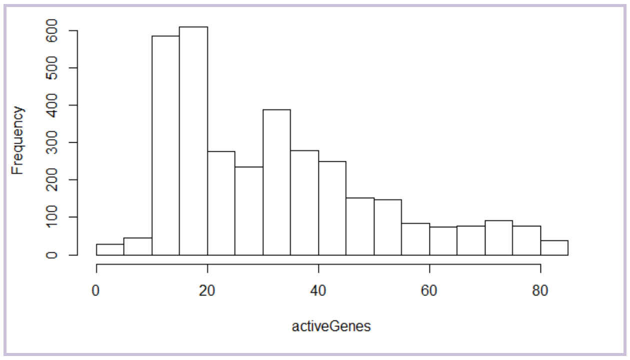

Figure 3. Allele Frequency Distribution Week 10

Its distribution is markedly different than the probabilistic confinement of random population initialization visualized in Figure 1. Cursory observation of Figure 3 reveals the search signature: primitively focused in the mid-tier and overcome by lower frequencies (settling on 15-20). Additionally, there appears to have been some temporary interest in the 70-75 range. Ultimately, the meta-shifted genetic algorithm output is reinforced by a certain robustness that cannot be attributed to its SGA sibling. Although a mere 2657 among a septillions of solutions were tested for week 10, characteristics of the search across the solution space reinforce latent concerns of trust in output as the final solution was necessarily exposed to and survived the genetically dissimilar makeup of the members with which the originating population of an SGA has little chance to interact.

Limitations

A principal limitation concerns the burden on computational resources of evaluating dozens/hundreds of solutions, over many generations, undergoing k-fold cross-validation, C5.0 boosting, etc. This research was not conducted on a data mining quality research computer; RAM is restricted to 12GB with an Intel Core i7-3770 CPU @ 3.40 GHz. These conditions do alleviate a few difficulties inherent to GA and evaluation parametrization; population size, cross-validation folds, and boosting trials have to remain low. Although there is no formal rule for the correct number of subsamples, it is understood that a lower number results in greater bias (Kuhn and Johnson, 2013).

Conclusion

Design science research favors pragmatic research to solve real-world problems (Hevner and Chatterjee, 2010). The latter materializes within an academic sports domain lacking guiding theory for determinant factors of athletic performance thereby muddling identification of an objectively correct approach towards the construction of a knowledge discovery artifact. The practical consequences of these circumstances implicitly limit enthusiasm for narrowly focused objectives of maximal output given whatever predictive apparatus may be constructed.

This research develops an artifact extending standard genetic algorithm functionality. Principles in nature of extreme environmental change and the macro evolutionary theory of punctuated equilibrium crossover and are transposed to challenge a population of solutions evolving in a singular meta that becomes stale over time. Shifting to alternate pockets of the search space and exposing the converged population to disparate solutions creates a more robust final output, ameliorating thematic concerns of trust.

The subroutine is not limited in application to sports. In an era of data science, where organizations (including government) increasingly seek quantitative demonstrations of the value for a data asset, coupled with increasing and more sophisticated measurement tools for said asset, feature subset selection is a natural fit. As such, genetic algorithms are always at least among available remedial options, carrying with it problems of sub-optimal parameter settings/control. Further, practitioners must always remain mindful of premature convergence; the meta-shift subroutine is a straightforward counter to this problem that rebuilds trust in output.

REFERENCES

Affenzeller, M., Wagner, S., Winkler, S., & Beham, A. (2009). Genetic algorithms and genetic programming: modern concepts and practical applications. Crc Press.

Ahmed, F., Deb, K., & Jindal, A. (2013). Multi-objective optimization and decision making approaches to cricket team selection. Applied Soft Computing,13(1), 402-414.

Cassady, C. R., Maillart, L. M., & Salman, S. (2005). Ranking sports teams: A customizable quadratic assignment approach. Interfaces, 35(6), 497-510.

Coleman, B. J. (2012). Identifying the "players" in sports analytics research. Interfaces, 42(2), 109-118.

Davenport, T. (2014). Analytics in Sports: The New Science of Winning. Retrieved June 1, 2015, from SAS Institute: http://www.sas.com/en_ca/whitepapers/iia-analytics-in-sports-106993.html

Friedrich, T., Oliveto, P. S., Sudholt, D., & Witt, C. (2009). Analysis of diversity-preserving mechanisms for global exploration*. Evolutionary Computation, 17(4), 455-476.

Fry, M. J., & Ohlmann, J. W. (2012). Introduction to the special issue on analytics in sports, Part I: General sports applications. Interfaces, 42(2), 105-108.

Gould, N. E. S. J. (1972). Punctuated equilibria: an alternative to phyletic gradualism.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

Hevner, A., & Chatterjee, S. (2010). Design research in information systems: theory and practice (Vol. 22). Springer Science & Business Media.

Holland, J. H. (1975). Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. U Michigan Press.

Iyer, S. R., & Sharda, R. (2009). Prediction of athletes performance using neural networks: An application in cricket team selection. Expert Systems with Applications, 36(3), 5510-5522.

Kahn, J. (2003). Neural network prediction of NFL football games. World Wide Web electronic publication, 9-15.

Karafotias, G., Hoogendoorn, M., & Eiben, A. E. (2014). Parameter control in evolutionary algorithms: Trends and challenges. IEEE Transactions on Evolutionary Computation, to appear.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (pp. 61-92). New York: Springer.

Maier, K. D., Wank, V., Bartonietz, K., & Blickhan, R. (2000). Neural network based models of javelin flight: prediction of flight distances and optimal release parameters. Sports Engineering, 3(1), 57-63.

Pandey, H. M., Chaudhary, A., & Mehrotra, D. (2014). A comparative review of approaches to prevent premature convergence in GA. Applied Soft Computing, 24, 1047-1077.

Parsons, P. A. (1997). Extreme environmental change and evolution. Cambridge University Press.

Purucker, M. C. (1996). Neural network quarterbacking. Potentials, IEEE, 15(3), 9-15.

Song, C., Boulier, B. L., & Stekler, H. O. (2007). The comparative accuracy of judgmental and model forecasts of American football games. International Journal of Forecasting, 23(3), 405-413.

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. International Journal of Forecasting, 26(3), 606-621.

Trick, M. A., & Yildiz, H. (2012). Locally optimized crossover for the traveling umpire problem. European Journal of Operational Research, 216(2), 286-292.

Verleysen, Michel. "Learning high-dimensional data." Nato Science Series Sub Series III Computer And Systems Sciences 186 (2003): 141-162.

Wilson, R. L. (1995). Ranking college football teams: A neural network approach. Interfaces, 25(4), 44-59.

Wloch, K., & Bentley, P. J. (2004, January). Optimising the performance of a formula one car using a genetic algorithm. In Parallel Problem Solving from Nature-PPSN VIII (pp. 702-711). Springer Berlin Heidelberg.

Yang, J., & Honavar, V. (1998). Feature subset selection using a genetic algorithm. In Feature extraction, construction and selection (pp. 117-136). Springer US.