Applied Soft Computing 34 (2015) 606–612

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Evolving accuracy: A genetic algorithm to improve election night forecasts

Ronald Hochreiter[a], Christoph Waldhauser[b,*]

[a] Department of Finance, Accounting and Statistics, WU Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria
[b] K Data Science Solutions, Amerlingstraße 4/14, 1060 Vienna, Austria

ARTICLE INFO

Article history:
Received 25 October 2012
Received in revised form 19 May 2015
Accepted 20 May 2015
Available online 1 June 2015

Keywords:
Election night forecasting
Genetic algorithm
Ecological regression
Constituency clustering

ABSTRACT

In this paper, we apply genetic algorithms to the field of electoral studies. Forecasting election results is one of the most exciting and demanding tasks in the area of market research, especially due to the fact that decisions have to be made within seconds on live television. We show that the proposed method outperforms currently applied approaches and thereby provides an argument to tighten the intersection between computer science and social science, especially political science, further. We scrutinize the performance of our algorithm's runtime behavior to evaluate its applicability in the field. Numerical results with real data from a local election in the Austrian province of Styria from 2010 substantiate the applicability of the proposed approach.

1. Introduction

When the last ballots have been cast and the last polling station closes, the fruits of a stressful afternoon are brought to bear: the first election forecast is being broadcast over the air. Much of the work behind it, however, actually took place long before that, starting weeks before the election and culminating shortly after noon.

Forecasting elections is arguably the most demanding and stressful but also the most exciting task market researchers can perform [6]. The term election forecast can mean different things. Karandikar et al. [14] and Morton [21] give a summary of different meanings, and when we speak of election forecasting in this paper, we mean exclusively what they termed a *results-based forecast*, a forecast based on partially counted votes without any exterior information like polls or surveys. Contrast this e.g. with an approach that uses external, i.e. historical, data in [4]. In the traditional forecasting process, after weeks of preparation, decisions have to be made in split seconds, possibly on live television. The preparation in the weeks before the election is a tedious process that involves many person hours and is error prone. This paper improves the current situation of the industry by contributing solutions based on genetic algorithms[1] to the most expensive and fragile elements of the field.

The remainder of this paper is structured as follows. First we will give an introduction to the methodology that constitutes the foundation of industry standard election forecasting as it is practiced today. Then the elements of this forecasting process that are especially expensive and fault intolerant are identified. In Section 2 we describe how genetic algorithms can be used to find near optimal solutions to the problems identified above. The devised algorithm is described in detail and evaluated using a standard set of indicators and real data from the field. Results of this analysis are presented in Section 3. Finally, we offer some concluding remarks and suggestions for further developments.

1.1. Ecological regression

Forecasting elections is a business that depends on meticulous preparations and accurate knowledge of the political processes behind the scenes. In the beginning of televised live election night forecasting, sometimes disastrous miscalculations paved the way for numerous endeavors, that were undertaken to improve the status quo [22,20]. Today, election forecasting using a methodology termed ecological regression [8,3,12,15,16,9], engages in the

* Corresponding author. Tel.: +43 6605490623.
E-mail addresses: ronald.hochreiter@wu.ac.at (R. Hochreiter), chw@kdss.at (C. Waldhauser).

[1] Preliminary results of this research project were presented at ACM GECCO conference 2011 [10].

http://dx.doi.org/10.1016/j.asoc.2015.05.033
1568-4946/

daunting task of comparing polling stations or constituencies of a geographical entity from a past election to the present one that is meant to be forecast. To make things worse, past election results do not easily translate into new election results because of old people dying and young ones becoming eligible to vote. Assuming, admittedly somewhat naively, that new voters behave in general similar to old voters, this transition becomes merely an exercise in multiplying old vote shares with the number of new voters. Any deviance in voting behavior will be accounted for in the regression model introduced later.

This only works because (1) not all polling stations provide their results at the same time and (2) voters that go to one polling station will behave similar to voters at another one.

Other frameworks are used to forecast elections well before they take place, usually with the aim of only predicting the winner, and not producing precise estimates for vote shares [28,7,27,25].

In a multi party system for any given election there are multiple parties competing against each other for votes. Voters can cast these votes at polling stations which are usually located close to their homes. It is also clear, that at least for developed democracies, parties have a history of performances in past elections. Any election forecast uses (at least) two elections, one in the past and the current one with the overall aim of predicting the vote shares of the current one. Since voters have formed an opinion and elect a party accordingly, not all parties will end up with the same share of votes, when comparing two elections. For two competing theories of how this might happen, see [17,18].

Note that there are two different kinds of vote shares that can be used as performance metrics for parties: either the proportion of the total electorate voting for a party or the proportion of the constituency that actually did cast a valid vote, that voted for a party. In the following, these quantities are called %Elec and %Vald, respectively. Most clients will be interested in the latter one, as it constitutes the post-election political reality.

In the regression-based model the performance p of a party i at a current election is a linear combination of all j parties' performances at the reference election plus the proportion of nonvoters (NV), for all k polling stations. To simplify things, the nonvoters are considered to be just another ordinary party and are thus included in the j parties. So the following equation has to be estimated for all j parties to link the old election results from polling station k to the new election results at that polling station:

$$p_{i,k} = \sum_j x_{j,k} p_{j,k} + p_{\mathrm{NV},k} \tag{1}$$

The factor $x_{j,k}$ in the equation above is the quantity of interest in the election forecasting process. This quantity can be considered as a transition multiplier. For instance a value of $x_{j,k} = 0.6$ for two parties i, j means that in the current election party i could mobilize 60 percent of the last time voters of party j for its own cause at polling station k. If $i = j$, $x_{i,k}$ boils down to the proportion of traditional i voters the party could again re-win at the current election (at this polling station); for all $i \neq j$, the different $x_{j,k}$ sum up to the votes that were won by party i from competing parties. All x_{*k} of a polling station k together make up a matrix with as many rows as parties in the current election and columns as parties in the old election. This matrix projects the old election's vote shares into the space of the new election. In the most trivial example, the same, let us say 4 parties compete in both elections. This means that the equation from above needs to be estimated four times for each polling station, leading to a 4×4 projection matrix for each polling station.

Obviously, a projection matrix can only be established for polling stations that already reported their results. The polling stations that did not yet report their results are then to be forecast. As stated earlier, it is assumed that any trend visible from the already declared polling stations will also apply to the polling stations not yet counted. So the idea is now to use the already obtained projection matrices on the old election results from those polling stations still missing. When, and this is quite quickly happening during an election day, more than one polling station have their results reported multiple projection matrices will be available. Then a cell-based average function over the available projection matrices is used to obtain an overall matrix.

Unfortunately, not all polling stations will follow the general trend, or will follow it only to some extend. Therefore, care must be taken in choosing the projection matrices that are used as input in computing the overall matrix.

As stated above, this method relies on the assumption that voters will behave similarly. However, consider that on the math side of things, as the used regression models are unbounded,[2] this method is a linear approximation of the choices the electorate makes. Also, since regression can be considered as computing an average over a number of data points, $x_{j,k}$ can take extreme values if there are heterogeneous trends between polling stations. This poses a problem to the election forecasting model as percentages below 0 and above 100 cannot be accounted for by voter mobilization.

The solution to this problem lies in grouping polling stations together that will exhibit a similar trend in the transition from the reference election to the current election. By doing so, the coefficients of the model, remain within the 0, 100 interval and thus interpretable. In other words, the linear approximation works if and only if the polling stations in each group are homogeneous enough. So the mean of the individual projection matrices are computed only for a subset of the available matrices.

So to summarize, in the election forecasting process, the relationships between old and new party results are used to project the results for yet missing polling stations. By means of multiple regression models, the transition multipliers are estimated per polling station. The transition multipliers of similar polling stations are then combined into averages. These averages are then used to compute the votes the parties are likely to obtain in the missing locations.

Traditionally the grouping, the identifying of polling stations that will exhibit similar trends, is done by experienced senior researchers using K-means clustering (see [19]) and constant size binning techniques. This process is usually very time consuming (and thus expensive), as there are no fixed rules and many different possibilities have to be evaluated by hand. Additionally, there is no guarantee that the groupings found in such a way will actually be homogeneous. Given the small number of possible combinations that can be tried in manual assessment, they are even quite unlikely to be related at all. If the resulting groups of polling stations, however, are homogeneous enough, stable forecasts will be available at a very early state of the vote counting process.

To summarize, for any election forecasting endeavor using the aforementioned method, the grouping of polling stations into homogeneous clusters is crucial. The search for a perfect grouping is a tedious and time consuming process especially given the huge number of possible combinations.

2. Optimization process

In this section we will describe the genetic optimization procedure we used to improve the quality of the grouping solutions and

[2] We use here the term unbounded to address linear regression's inherent tendency to fit straight lines (as opposed to the familiar sigmoid curves of logistic regression).