International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:8, No:1, 2014

Modified Naïve Bayes Based Prediction Modeling for Crop Yield Prediction

Kefaya Qaddoum

***Abstract*—**Most of greenhouse growers desire a determined amount of yields in order to accurately meet market requirements. The purpose of this paper is to model a simple but often satisfactory supervised classification method. The original naive Bayes have a serious weakness, which is producing redundant predictors. In this paper, utilized regularization technique was used to obtain a computationally efficient classifier based on naive Bayes. The suggested construction, utilized L1-penalty, is capable of clearing redundant predictors, where a modification of the LARS algorithm is devised to solve this problem, making this method applicable to a wide range of data. In the experimental section, a study conducted to examine the effect of redundant and irrelevant predictors, and test the method on WSG data set for tomato yields, where there are many more predictors than data, and the urge need to predict weekly yield is the goal of this approach. Finally, the modified approach is compared with several naive Bayes variants and other classification algorithms (SVM and kNN), and is shown to be fairly good.

***Keywords*—**Tomato yields prediction, naive Bayes, redundancy.

I. Introduction

FOR many years, reliable supplies of high quality food in agreed quantities are required by stakeholders such as the tomato growers, the supermarkets and various others such as the consumers. Growers have increased food quality and yield in many parts of the world through the use of greenhouses where the environmental conditions can be controlled, and by selecting better cultivars [2]. However, weekly yields can fluctuate and this can pose problems of both over-demand and over-production if the yield cannot be predicted accurately. In this respect growers and scientists are looking for ways to forecast tomato yield so as to be able to plan greenhouse operations, marketing etc. One of the dynamic and complex systems is tomato crop growth, and it has been studied through the development of mechanistic models. Two of the verified dynamic growth models are TOMGRO [13] and TOMSIM [14]. Both models are built on physiological processes and they model biomass dividing, growth, and yield as a function of several climate and physiological parameters. Their use is narrow, especially for practical application by growers, by their complexity and by the difficulty in obtaining the initial condition parameters required for implementation [12].

Bayesian network classifiers [1] are often used for classification problems. The model parameters are usually found by maximizing the joint likelihood. The naive Bayes model is a simple Bayesian network classifier [5] that assumes the predictors are independent given each class value. In spite of this strong assumption, this classifier has been proven to work satisfactorily in many domains [2], [3].

The lasso [10] is a popular regularization technique that imposes an L1-penalty on the usual least-squares linear regression, with the aim of reducing the variance of the estimates, preventing overfitting, performing simultaneously variable selection and, finally, improving the model interpretability. Depending on the chosen regularization parameter, some regression coefficients are set to exactly zero, and the corresponding predictors are discarded [10].

The L1-penalty [11] has been widely used in many classification paradigms, like logistic regression [8] with a minor modification, the LARS algorithm13 assesses the complete lasso regularization path, and that is, the whole set of regression coefficient estimates with regard to the regularization parameter. LARS is of particular interest because it solves the complete regularization path at the cost of an ordinary least squares fit. Besides least squares functions, the LARS algorithm can be used to efficiently minimize other loss functions subject to an L1-penalty provided these loss functions meet certain conditions [7]. In this paper, we introduce a supervised classification method that is inspired on naive Bayes and based on convex optimization. On the one hand, this formulation allows applying regularization techniques from linear regression that permit to discard both redundant and irrelevant predictors. Redundant predictors are known to be harmful for naive Bayes and variants, and also for our model [4]. On the other hand, like naïve Bayes, it can directly deal with both continuous and discrete predictors and can be directly used in multi-class problems. Thus, our method is applicable to a wide range of data sets [4].

The proposed method establishes a linear combination of the likelihood contributions of each predictor. This linear combination is chosen so that the result is maximized, assuming that the coefficients are somehow constrained. This will give priority to those variables whose likelihood contributions are higher. The applied constraint is an L1-penalty, which yields a sparse vector of coefficients, dropping the likelihood contribution of some predictors and, thus, enhancing the interpretability of the model. As results will show, this method can discard both redundant and irrelevant predictors (i.e. their respective likelihood contributions). The devised loss function also meets the requirements for applying a LARS type algorithm [7] This algorithm would efficiently compute the entire regularization path at one shot. This is beneficial in high dimensional settings on computational

Kefaya Qaddoum was with Warwick University, UK, (e-mail: k.s.qaddoum@gmail.com).

grounds. Finally, our method is applicable to a wide range of data [7].

The rest of the paper is organized as follows. Section II introduces the used scheme in detail. Section IV details the set of experiments used to test the algorithm. Section V discusses conclusions and future work.

II. Naïve Bayes and Variants

Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based is rarely true in real-world applications [4].

III. The Method

In this paper, each predictor builds a penalized linear expression whose minimization will yield a classifier that discards irrelevant and redundant predictors.

We first obtain the ML parameters "i, for i∈ _, and "μi and "2 where, following the Bayes' rule or Vector f = ("1, . . . "p) would be chosen to maximize (10), hence giving more weight to predictors that are more relevant for the classification. The rationale of this approach is that relevant predictors will have values P(Y =yr|Xi = xri, ˆi) closer to one than irrelevant predictors. Hence, when maximizing (10) across the data set, the coefficients _i of the relevant predictors are promoted to be higher [7].

Note also that, as long as ranges from 0 to 1, like a probability. This could be used as a basis for classifying future instances. Specifically, for a new instance given by xi, the class value j ∈ {1 . . . c} that maximizes Note that _i = 0 implies that predictor Xi is not selected. Likewise, higher values of _i would attach more importance to predictor Xi. Predictors that are considered to be relevant (i.e., with a high _i) are expected to have a higher probability P(Y = j|Xi = xi, ˆi) for the true class, as it was in the training data set.

To obtain _, we could devise a linear optimization problem that maximizes (10) for the data set. However, it will not drive any i to exactly zero, and, hence, will not perform variable selection. We alternatively propose an L1-constrained problem [9].

IV. Redundant Predictors and Irrelevant Predictors

Place Let Xi1 and Xi2 be two redundant predictors, for example, a predictor that appears twice.

First, if Xi1 and Xi2 are discrete and (5) is satisfied, the value of Xi2 can be determined if Xi1 is known and vice versa. Hence, there is a bisection between the i1-th and the i2-th columns of matrix X. Obviously, this means that P (Xi1 = xri1 |Y = yr, ˆ_i1) and P (Xi2 = xri2 |Y = yr, ˆ_i2) are equal, r = 1, . . . n. Therefore, it follows that P (Y = yr|Xi1 = xri1, ˆ_i1) and P (Y = yr|Xi2 = xri2, ˆ_i2), r = 1, . . . n, are equal too. Hence, if two predictors, Xi1 and Xi2, are highly correlated then vector P (Y = yr|Xi1 = xri1, ˆ_i1), r = 1, . . . n, and vector P (Y = yr|Xi2 = xri2, ˆ_i2), r = 1, . . . n, will also be highly correlated. Therefore, (12), which can be solved by LARS, would drop either Xi1 or Xi2 due to the lasso constraint properties (i.e., the ability of the L1-penalty to discard redundant predictors) [9].

If Xi1 and Xi2 are continuous and redundant, either Xi1 or Xi2 would also be discarded.

LARS starts with no predictors. Firstly, it includes the predictor that is most correlated with the response into the active set of predictors A. The response is regressed on this predictor, so that the coefficient of this predictor is moved towards the least squares solution until a new predictor reaches the same absolute correlation with the vector of residuals as that of A.

This new predictor is included in the active set A. Now, the vector of residuals is regressed on the predictors in A, moving their coefficients towards the joint least squares solution until a new predictor not in A reaches the same absolute correlation with such vector of residuals as that of A. When n ≥ p, this procedure is repeated until all predictors are into the model. Otherwise, after n − 1 steps, the residuals are zero and the algorithm terminates [7].

Now, to accomplish the condition _i ≥ 0, we compute as the minimum value such that some predictor i /∈ A reaches the same positive correlation with the vector of residuals as that of A. Thus, the difference is that the negative correlations with the residuals of predictors i /∈A are ignored for computing and deciding which predictor i /∈ A enters the model. This modification was presented in the paper by Efron et al. [6].

V. Experiments

A. Tomato Yield Prediction Data

Wide areas in WSG (Wight Salads Group) in the Isle of Wight, United Kingdom, are used for this study. These regions make major contribution in tomato production of UK. The inputs to the network are several parameters derived from the crop model [12]-[14], and including (temperature, CO_2, vapor pressure deficit (VPD), yield, and radiation), which originally were measured on weekly basis during each season.

We present some illustrative results on two different scenarios. First, we evaluate the effect of redundant and irrelevant predictors. Second, we test the proposed method on a high dimensional data set. Finally, we run the naive Bayes methods on a data set than combine numeric with categorical predictors.

In this section, we test the behavior of our method on WSG tomato yield data sets. We focus on the version with no missing values. This data set has n = 672 instances, p = 63 predictors and four classes, whose relative proportions are (0.21, 0.21, 0.21, 0.37).We have chosen this data because it is a well-structured data set, suitable for testing how sensitive the algorithm is to the above issues.

Based on the original data set, we built several new data sets by adding different numbers of irrelevant and redundant predictors. We tested all combinations of redundancy and irrelevance redundant. Predictors are randomly generated values that are highly correlated (0.8) with an existing

predictor, which is itself highly correlated to the class. We have tested a total of 9 x 9 = 81 data sets [10].

We compared the proposed method to ordinary naive Bayes [5] and selective naive Bayes. Three random predictors were introduced and sampled from a different distribution with five categories and equal probabilities for each category. Afterwards, those predictors whose mutual information is lower than one of the three random predictors have been discarded. For each data set we performed 5-fold cross-validation, so that 80% of the data is used for training at each fold. Model evaluation was based on the AIC statistic.

Fig. 1 indicates with a horizontal thick line that the difference of the L1-NB accuracy to the second best method is statistically significant with a significance level of 0.05. We do not show the number of correctly selected predictors because it is not clear which variables from the original set should really be selected. The total number of predictors in the data set is marked by the ordinary naive Bayes line.

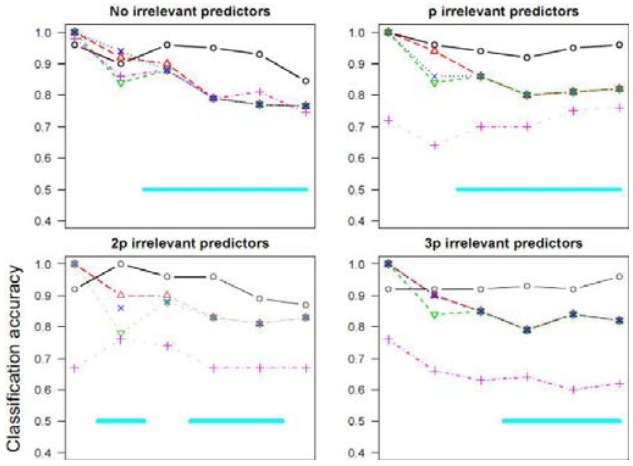

Fig. 1 Classification accuracy (Y-axis) for increasing redundant predictors (X-axis). The solid line plots L1-NB, the long-dashed line plots ordinary naïve Bayes, the short-dashed line plots naïve Bayes with prefiltering feature selection, the dotted line plots weighted naïve Bayes with prefiltering feature selection and he dashed-dotted-line plots selective naïve Bayes

As expected, irrelevant predictors do not affect the performance of the evaluated classifiers much, except for selective naive Bayes. Their accuracies do not greatly decrease as the number of irrelevant predictors grows. On the other hand, excepting this approach and selective naive Bayes, there is an increment of selected predictors for data sets containing more irrelevant predictors.

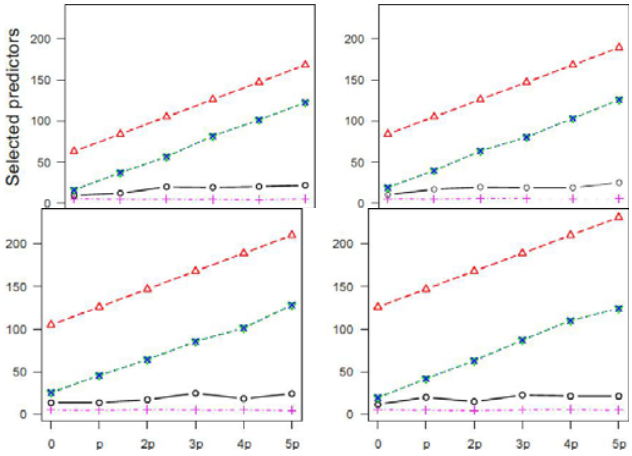

Fig. 2 Number of selected predictors (Y axis) against increasing redundant predictors (X axis). The solid- line plots L1-NB, the long-dashed-Δ line plots ordinary naïve Bayes, the short-dashed line plots naïve Bayes with prefiltering feature selection, the dotted line plots weighted naïve Bayes with prefiltering feature selection and the dashed-dotted line plots selective naïve Bayes

The effect of redundant predictors is stronger. As a general rule, selective naïve Bayes exhibits lower accuracy in the presence of redundant predictors. The more redundant predictors there are, the greater the number of selected predictors.

TABLE I
MEAN ACCURACY AND MEAN NUMBER OF SELECTED PREDICTORS

Method	Accuracy	predictors	Method	Accuracy	predictors
L1-NB	0.29(±0.09)	511(±130)	WNB	0.39(±0.01)	92.0(±0.0)
NB	0.37(±0.01)	100.0(±0.0)	SNB	0.31(±0.02)	60.5(±6.2)
SVM	0.39(±0.1)	100.0(±0.0)	SVM	0.31(±0.01)	81.0(±0.0)

On the other hand, the number of selected variables for L1-NB and selective naive Bayes barely fluctuates at around 3 predictors for all data sets, always selected from the original set of variables Fig. 2.

VI. Conclusion and Future Work

So far, the issue of irrelevant predictors and redundant predictors for the naive Bayes model has been discussed. The utilized model that, initially inspired by the naive Bayes scheme, deals reasonably well with these false predictors. This has been proved empirically on several data sets, where different numbers of irrelevant and redundant predictors have been added. The proposed method had been found much better than naive Bayes model, since the L1-penalty deals with redundancy then the redundant predictors could be discard. In the future, we plan to extend this or alternative formulations for exploring more complex predictor relations than redundancy.

References

[1] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian network classifiers Machine Learning 29 (1997) 131–163.

[2] P. Domingos and M. Pazzani, Beyond independence: Conditions for the optimality of the simple Bayesian classifier, Machine Learning 29 (1997) 103–130.
[3] D. J. Hand and K. Yu, Idiot's Bayes – Not so stupid after all? International Statistical Review 69 (2001) 385–398.
[4] J. T. A. S. Ferreira, D. G. T. Denison, and D. J. Hand, Data mining with products of trees, in Advances in Intelligent Data Analysis, Volume 2189 of Lecture Notes in Computer Science, (2001), pp. 167–176.
[5] P. Langley and S. Sage, Induction of Bayesian classifiers, in Proc. of the 10th Conf. on Uncertainty in Artificial Intelligence (1994), pp. 399–406.
[6] M. J. Pazzani, Searching for dependencies in Bayesian classifiers, in Learning from Data: Artificial Intelligence and Statistics V, (1996), pp. 239–248.
[7] M. Boull´e, Compression-based averaging of selective naive Bayes classifiers, Journal of Machine Learning Research 8 (2007) 1659–1685.
[8] R. Tibshirani, Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B 58 (1996) 267–288.
[9] P. Zhao and B. Yu, On model selection consistency of Lasso, Machine Learning Re- search 7 (2006) 2541–2567.
[10] D. Vidaurre, C. Bielza, and P. Larra˜naga, Forward stagewise naive Bayes, Progress in Artificial Intelligence 1 (2011) 57–69.
[11] Heuvelink, E. Growth, development, and yield of a tomato crop: Periodic destructive measurements in a greenhouse. Scientia Hort. 61(1-2): 77-99. 1995.
[12] Heuvelink, E. Tomato growth and yield: quantitative analysis and synthesis. PhD diss. Wageningen, the Netherlands: Wageningen Agricultural University. 1996.
[13] Heuvelink, E. Evaluation of a dynamic simulation model for tomato crop growth and development. Ann. Botany 83(4): 413-422. 1999.
[14] Heuvelink, E. Developmental process. In Tomatoes, 53-83. Crop Production Science in Horticulture Series. E. Heuvelink, ed. Wallingford, U.K.: CABI Publishing. 2005.