SCIENTIFIC REPORTS
nature research

OPEN

Likelihood contrasts: a machine learning algorithm for binary classification of longitudinal data

Riku Klén[1,2,3], Markku Karhunen[1,3] & Laura L. Elo[1*]

Machine learning methods have gained increased popularity in biomedical research during the recent years. However, very few of them support the analysis of longitudinal data, where several samples are collected from an individual over time. Additionally, most of the available longitudinal machine learning methods assume that the measurements are aligned in time, which is often not the case in real data. Here, we introduce a robust longitudinal machine learning method, named likelihood contrasts (LC), which supports study designs with unaligned time points. Our LC method is a binary classifier, which uses linear mixed models for modelling and log-likelihood for decision making. To demonstrate the benefits of our approach, we compared it with existing methods in four simulated and three real data sets. In each simulated data set, LC was the most accurate method, while the real data sets further supported the robust performance of the method. LC is also computationally efficient and easy to use.

Many biomedical studies consist of longitudinal data, i.e. data with multiple samples for each individual, taken at different time points. Here, we define longitudinal data so that the covariates are measured repeatedly, but not necessarily at even intervals or at the same time points for each individual. This type of data turns out to yield substantial modelling challenges. For example, the most widely used binary classifiers, such as Lasso[1], random forest[2] and artificial neural networks[3], are not designed for this type of data. Therefore, they cannot fully benefit from the repeated measurements. Moreover, those machine learning methods which support longitudinal data typically assume that the time points are aligned between the individuals[4].

Many statistical methods are available for longitudinal data, especially within the discipline of econometrics[5], but these methods typically also assume the time points to be aligned and evenly spaced. The only main exception suitable for biomedical data is the linear mixed-effects model (LME) and its modifications[6–9], which support data with non-aligned time points. However, the LME model is a regression model for a continuous response variable. Many different solutions to turn the model into binary classifier can be envisaged[10,11]. Here, we present one such solution: the method of likelihood contrasts (LC). We introduce this novel method because it exploits all longitudinal data in classification instead of a single time point or average. LC is fast and easy to calculate, and secondly, our results show its good performance in simulated and real data sets alike.

We take the univariate LME as the starting point and use it as a building block for our LC algorithm. Briefly, we fit LMEs using a standard software package (lme4 version 3.1–131.1)[12], and then use their maximised log-likelihood functions for inference. We assign each sample to the group where the log-likelihood changes most favourably. Thus, this method amounts to a binary classifier. However, contrary to many other machine learning methods, LC is computationally very efficient, easy to implement and the need to fine-tune parameters is minimal. We provide an open-source implementation of LC at https://elolab.utu.fi/software/.

In this paper, we demonstrate the performance of LC in four simulated and three publicly available real data sets. In each data set, we test the discriminatory power of LC regarding the case-control status of the study subjects and compare it to that of widely used machine learning and predictive algorithms, including Lasso[1], random forest (RF)[2], support vector machines (SVM)[13], neural networks (NN)[3], and LME regression models. Two of the real data sets derive from the book of Rizopoulos[14] on longitudinal models and are openly accessible[15]. These represent typical clinical data sets used for longitudinal modelling. The third real data set concerns molecular data on pediatric Type 1 Diabetes mellitus collected in the Finnish Type 1 Diabetes Prediction and Prevention (DIPP) study[16] and is also publicly available[17].

[1]Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland. [2]Turku PET Centre, University of Turku, Turku, Finland. [3]These authors contributed equally: Riku Klén and Markku Karhunen. *email: laura.elo@utu.fi

Methods

In this section, we introduce the method of likelihood contrasts (LC). We also describe other methods used in comparison.

Likelihood contrasts. Let z_i denote an observation of a new individual i, which may contain data from multiple time points, and response variables and covariates alike. The individuals are labelled as $\delta_i = 1$ (case) or $\delta_i = 0$ (control). We estimate two separate models and see which one gives z_i a better fit. To this end, let z^1_{-i} and z^0_{-i} denote training data from cases and controls, respectively, and let $\ell(z^1_{-i}|\theta^1)$ and $\ell(z^0_{-i}|\theta^0)$ denote the maximised log-likelihoods of the corresponding two separate models, M^1 and M^0, with the parameter estimates θ^1 and θ^0. We then calculate the likelihood contrasts as

$$\begin{aligned} d_1 &= \ell(z_i, z^1_{-i}|\theta^1) - \ell(z^1_{-i}|\theta^1), \\ d_0 &= \ell(z_i, z^0_{-i}|\theta^0) - \ell(z^0_{-i}|\theta^0) \end{aligned} \tag{1}$$

and assign z_i to the group where d_k is larger.

It is also possible to extend the method to yield probability scores. This can be justified by considering the log-likelihood difference $d_1 - d_0$. Quite intuitively, $d_1 - d_0 = 0$ can be used as a cut-off point in binary classification. In this respect, $d_1 - d_0$ is comparable to the linear predictor (i.e., the linear combination of covariates) found in logit models. In logit models, the linear predictor is mapped to probability score through the inverse of the logistic link function. Adopting this approach, we have

$$\hat{P}(\delta_i = 1) = \text{logit}^{-1}(d_1 - d_0) = \frac{\exp(d_1 - d_0)}{1 + \exp(d_1 - d_0)} = \frac{\exp(d_1)}{\exp(d_0) + \exp(d_1)}. \tag{2}$$

It is quite naturally possible to extend LC into a multinomial classifier. In that case,

$$\begin{aligned} d_k &= \ell(z_i, z^k_{-i}|\theta^k) - \ell(z^k_{-i}|\theta^k), \ k = 1, \ldots, L, \\ \hat{\delta}_i &= \underset{k}{\text{argmax}} d_k \end{aligned} \tag{3}$$

and

$$\hat{P}(\delta_i = k) = \frac{\exp(d_k)}{\sum_{l=1}^{L}\exp(d_l)}, \ k = 1, \ldots, L, \tag{4}$$

where L is the number of classes. However, we only use binary classification in this paper.

Implementation with mixed models. Above, z_i denotes any data, encompassing n_i measurements for individual i. Here, we use LC in combination with LMEs, which are versatile tools for modelling jointly the effect of covariates, confounders and sampling artefacts. In matrix notation, an LME can be defined as

$$\begin{aligned} &\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \\ &E(\epsilon|\mathbf{X}) = 0, \\ &Cov(\epsilon|\mathbf{X}) = \mathbf{A}(\mathbf{X}), \end{aligned} \tag{5}$$

where $\mathbf{y}$ is the vector of responses, $\mathbf{X}$ is the matrix of covariates, $\boldsymbol{\beta}$ is the vector of regression coefficients, and ϵ is the error term. $\mathbf{A}(\mathbf{X})$ is the covariance matrix of the measurement errors which may depend on $\mathbf{X}$. An LME differs from the usual linear model because $\mathbf{A}(\mathbf{X})$ is not a diagonal matrix. The off-diagonal terms of $\mathbf{A}(\mathbf{X})$ represent correlations between the samples, and are considered as the 'mixed effects'. They arise as a result of study design or spatio-temporal vicinity of the samples.

In practice, we assume that the data for each individual i involve a disease-specific longitudinal marker y_{ij} and a number of covariates denoted by $\boldsymbol{x}_{ij}$. In the sequel, we denote the j th measurement of a marker for individual i as y_{ij}, and we model the time course of this marker as

$$\begin{aligned} y_{ij} &= \boldsymbol{\beta}_1^{'}\boldsymbol{x}_{ij}t_{ij} + \gamma_1 t_{ij} + \boldsymbol{\theta}_1^{'}\boldsymbol{x}_{ij} + u_i + \epsilon_{ij}, \ \delta_i = 1, \\ y_{ij} &= \boldsymbol{\beta}_0^{'}\boldsymbol{x}_{ij}t_{ij} + \gamma_0 t_{ij} + \boldsymbol{\theta}_0^{'}\boldsymbol{x}_{ij} + u_i + \epsilon_{ij}, \ \delta_i = 0, \end{aligned} \tag{6}$$

where δi denotes the case-control status, t_{ij} denotes time, $u_i \sim NID(0, \sigma_i^2)$ denotes an individual-specific random effect and $\epsilon_{ij} \sim NID(0, \sigma^2)$ denotes a measurement error. Here, $NID(\mu, \sigma^2)$ means normal, independent and identically distributed with mean μ and variance σ^2. There are two models, as there are two groups of patients ($\delta_i = 1$ and $\delta_i = 0$). The purpose of LC is to distinguish between these groups. An LME-based LC algorithm is implemented in R and is publicly available at https://elolab.utu.fi/software/.

Connection to statistical paradigms. The decision rule of LC resembles the likelihood-ratio (LR) test and Bayesian posterior probabilities but differs from both. In this subsection, we discuss these similarities and differences. Generally speaking, as compared to LR test, our method is not based on nested models, and is thus more general. As compared to Bayesian inference, our method does not require numerical integration schemes,

and is thus more efficient. Moreover, our method is suited for situations where the likelihood contribution of each individual observation cannot be calculated, as it uses the change of log-likelihood as a proxy for likelihood contribution.

In more detail, the LR test statistic is defined as

$$LR = 2(\ell(y_i, y^1_{-i}, y^0_{-i}|\theta^1) - \ell(y_i, y^1_{-i}y^0_{-i}|\theta^0)), \tag{7}$$

This definition is very standard and can be found from statistics text books. It is based on the fact that the likelihood ratio thus defined has favourable distributional properties known as the Wilks' theorem. The LR test, however, is based on nested models. For our method, the implied test statistic is

$$d_1 - d_0 = \ell(y_i, y^1_{-i}|\theta^1) - \ell(y_i, y^0_{-i}|\theta^0) - \ell(y^1_{-i}|\theta^1) + \ell(y^0_{-i}|\theta^0) \tag{8}$$

and it can be calculated for disjoint models. In parallel with this, Bayesian posterior probabilities typically concern the whole data, as the purpose is to choose the optimal model. Contrary to this, LC concerns each observation separately. To see the connection to Bayesian inference, assume that the information in y^0_{-i} and y^1_{-i} is so great that it essentially fixes the values of θ^0 and θ^1. Following this,

$$\ell(y_i|\theta^k) = \log\pi(y_i|M^k), \quad k = 0, 1; \tag{9}$$

i.e. the maximised likelihood is the likelihood of the whole model. (In other cases, one would need to integrate over θ^k to get $\pi(y_i|M^k)$; see Gelman *et al.*[18].) Moreover, let us assume that the observation units are independent, and thus, the likelihood is separable as

$$\ell(y_i, y^k_{-i}|\theta^k) = \ell(y_i|\theta^k) + \ell(y^k_{-i}|\theta^k). \tag{10}$$

Thus, it follows that

$$\log\pi(y_i|M^k) = \ell(y_i|\theta^k) = \ell(y_i, y^k_{-i}|\theta^k) - \ell(y^k_{-i}|\theta^k) = d_k. \tag{11}$$

Now, if one gives equal prior weights for both models, i.e. $\pi(M^0)$, $\pi(M^1) = 1/2$, it follows that

$$\pi(M^1|y_i) = \frac{\pi(y_i|M^1)\pi(M^1)}{\pi(y_i|M^0)\pi(M^0) + \pi(y_i|M^1)\pi(M^1)} = \frac{\exp(d_1)}{\exp(d_0) + \exp(d_1)}, \tag{12}$$

i.e. the posterior probability of M^1 for y_i coincides with our probability score, see Eq. (2). Finally, someone might ask why we use the likelihood contrasts d_0 and d_1 to classify individual i, and not just the likelihood contributions $\ell(y_i|\theta^0)$ and $\ell(y_i|\theta^1)$. This is because in complex models, such as LME, it is not possible to calculate likelihood contributions as such. Thus, we use the likelihood contrast

$$d_k = \ell(y_i, y^1_{-i}|\theta^k) - \ell(y^1_{-i}|\theta^k) \tag{13}$$

as a proxy for $\ell(y_i|\theta^k)$. Note that LC does not produce a single model with fixed coefficients, but it creates new coefficients for each new individual.

Comparison with other methods. We compared the performance of LC to a number of statistical and machine learning methods, including LME, linear feature extraction (LF), logit mixed-effects regression (implemented as the function GLMER in the R package lme4[19]), Lasso, random forests (RF), support vector machines (SVM), and neural networks (NN). Among the compared methods, LME and GLMER represent statistical methods, while Lasso, RF, SVM and NN are widely used machine learning algorithms. LF uses a strategy to account for the longitudinal dimension, but relies on a standard statistical technique, logistic regression. Below, we briefly outline these methods using the notations given above. In all analyses, the task was to predict δ_i on the basis of the covariates $\boldsymbol{x}_{ij}$ and the longitudinal marker y_{ij}. We denote the averaged value of the longitudinal marker over time by $\overline{y}_i$ and averaged covariates by $\boldsymbol{x}_i$.

In LME, the marker was first modelled as

$$\hat{y}_{ij} = \boldsymbol{\beta}'\boldsymbol{x}_{ij}t_{ij} + \gamma t_{ij} + \boldsymbol{\theta}'\boldsymbol{x}_{ij} + u_i + \epsilon_{ij}, \tag{14}$$

where $u_i \sim NID(0, \sigma_1^2)$ denotes an individual-specific random effect, and $\epsilon_{ij} \sim NID(0, \sigma_2^2)$ denotes a measurement error. Then, we averaged the estimated values $\hat{y}_{ij}$ over time $j = 1, \ldots, n_i$ and ran a logit regression of δ_i on the averages. We fitted the LME models using the R package nlme version 3.1–131.1[12].

In LF, we first ran a linear regression of y_{ij} on t_{ij} within each individual i to obtain individual-specific slopes and intercepts, denoted by b_i and a_i, respectively,

$$\hat{y}_{ij} = b_i t_{ij} + a_i. \tag{15}$$

Subsequently, we ran a logit regression of δ_i on $(\boldsymbol{x}_i, b_i, a_i)$ to predict δ_i.

Logistic Lasso was fitted on $(\delta_i; \boldsymbol{x}_i, \overline{y}_i, \overline{t}_i)$ using the R package glmnet version 2.0–13[20] with ten-fold cross-validation. Here $\overline{t}_i$ is the averaged measurement time for each individual.

The GLMER model was constructed on $(\delta_i; \boldsymbol{x}_{ij}, y_{ij}, t_{ij})$ using the R package lme4 version 1.1–15[19].

The RF model was constructed on $(\delta_i; \boldsymbol{x}_i, \overline{y}_i, \bar{t}_i)$ using the R package randomForest version 4.6–12[21].

An SVM (more precisely, epsilon regression) was constructed on $(\delta_i; \boldsymbol{x}_i, \overline{y}_i, \bar{t}_i)$ using the R package e1071 version 1.6–8[22].

An artificial neural network was constructed with one hidden layer on $(\delta_i; \boldsymbol{x}_i, \overline{y}_i, \bar{t}_i)$ using the R package nnet version 7.3–12[23].

Methods LME, Lasso, RF, SVM and NN involved averaging of values before modelling. We also implemented the methods without averaging by considering each time point as a separate measurement. We denote these methods by LME2, Lasso2, RF2, SVM2 and NN2.

All machine learning models were built using default parameters. Internal cross validation was used to determine coeffiecients for the logistic model and the penalty factor in Lasso. RF implemented Breiman's random forest algorithm using 500 trees with sample replacement. In SVM, support vectors were defined using epsilon regression with $\varepsilon = 0.1$. NN used one hidden layer and the number of units in the hidden layer was determined to be half of the number of variables.

Note that methods LC, LF, LME, LME2, Lasso2, RF2, SVM2, NN2 and GLMER use information for each time point, while methods Lasso, RF, SVM and NN use information averaged over time points per subject. Out of the compared methods, only LC, LF and GLMER directly make prediction for a new subject, while the other methods create a prediction for a single time point. For these methods, we made predictions for each time point for each subject, and averaged the predictions to obtain a single prediction for each subject.

Model evaluations. We compared the performance of the different methods on the basis of their binary predictions for test data, using cross validations as explained in the sequel. We truncated the probability scores given by the different models into binary predictions by using 0.50 probability as the cut-off and then assessed the performance of the binary predictions by calculating sensitivity and specificity. We considered 0.50 as the baseline value of sensitivity and specificity, assuming that a completely uninformative classifier is equally likely to classify the subjects as cases or controls. We used Wilcoxon's rank sum test to compare the sensitivity and specificity obtained from each method to the baseline values. Different methods were compared using paired Wilcoxon's rank sum test. To account for multiple testing, we applied Benjamini-Hochberg false discovery rate (FDR) correction[24] to the Wilcoxon's rank sum test P-values.

Additionally, we also present the F1 scores, accuracies and receiver operating characteristic (ROC) curves for each method and data set in Supplementary material.

Materials

To evaluate LC along with existing predictive methods we used simulated and real data.

Simulated data. In the simulated data, we considered one static covariate, the 'treatment' denoted by x_i, and one longitudinal marker denoted by y_{ij}. We assumed here that the distributional form of the marker differed between cases and controls, and thus, y_{ij} was informative regarding the case-control label δ_i. Altogether, we considered four different scenarios described in detail below.

In each scenario, the individuals were equally likely to be cases or controls. We assumed four time points per individual ($n_i = 4$) and we assumed x_i to be Bernoulli distributed with parameter value 0.5 and t_{ij} to be uniformly distributed on the interval $(-1, 1)$, i.e. we assumed that the treatment was allocated randomly and the time was measured relative to the event. In each scenario, 1,000 replicate data sets were generated to control for the sampling variation.

In Scenario 1, we assumed that the cases and controls reacted differently to the treatment and also that the natural course of the marker was different between the groups. Thus, we specified the model as

$$\begin{aligned} y_{ij} &= x_i - t_{ij} + u_i + \varepsilon_{ij}, \ \delta_i = 0, \\ y_{ij} &= -x_i + t_{ij} + u_i + \varepsilon_{ij}, \ \delta_i = 1, \end{aligned} \tag{16}$$

where $u_i \sim NID(0, 0.25)$ is an individual-specific random effect and $\varepsilon_{ij} \sim NID(0, 0.25)$ is the measurement error. Here, as in Eq. (17) below, the coefficients were chosen to illustrate the biological phenomena explained in the text, simultaneously keeping the simulation model as simple and tangible as possible. In this scenario, we used $n_1 = 40$ samples as the training data and assessed the model performance in an independent test data of $n_2 = 20$ samples, repeating the process 1,000 times.

In Scenario 2, we assumed that the distribution of y_{ij} was more similar between the cases and controls than in Scenario 1. We assumed that the controls did not react to the treatment and the natural course of the marker was similar between the groups, albeit milder in controls. Thus, we specified the model as

$$\begin{aligned} y_{ij} &= 0.5t_{ij} + u_i + \varepsilon_{ij}, \ \delta_i = 0, \\ y_{ij} &= -x_i + t_{ij} + u_i + \varepsilon_{ij}, \ \delta_i = 1, \end{aligned} \tag{17}$$

where $u_i \sim NID(0, 0.5)$ is an individual-specific random effect and $\varepsilon_{ij} \sim NID(0, 0.5)$ is the measurement error. Also in this scenario, we used $n_1 = 40$ samples as the training data and assessed the model performance in an independent test data of $n_2 = 20$ samples, repeating the process 1,000 times.

Scenarios 3 and 4 were the same as Scenarios 1 and 2, respectively, but here we assumed larger training and test data sets with $n_1 = 160$ and $n_2 = 80$.

	Total number of samples (N)	Number of individuals (n)	Number of cases	Number of controls	Number of time points per individual ± SD
Pbc2	1,945	312	140	172	6.2 ± 3.8
Prothro	2,968	488	292	196	6.1 ± 3.5
DIPP, seroconverted	68	17	4	13	4.0 ± 0.0
DIPP, progressor	238	40	18	22	6.0 ± 1.9

Table 1. Sample sizes of the real data sets.

Real data. We used two clinical data sets and one high-throughput molecular data set. In each of these real data sets, we used 2/3 of the data set as training data and predicted the labels in the remaining 1/3 to assess the performance of the different methods. To control for sampling variation in model evaluation, we repeated the process 1,000 times which diminished the standard errors more than sufficiently (<0.01 for variables on a scale of 0–1).

Clinical data sets. The two clinical data sets (Pbc2 and Prothro) were chosen from Rizopoulos[14], distributed in the R package JM (version 1.4–7)[15]. The sample sizes of the real data sets are summarised in Table 1. In both clinical data sets, $\delta_i = 1$ means death and $\delta_i = 0$ staying alive.

The Pbc2 data set was from a study on primary biliary cirrhosis[25]. The longitudinal marker in this data set was logarithm of blood bilirubin (mg/dl) over time and we used the drug (placebo or D-penicillamine) as a static covariate. Time (years) and bilirubin values were scalar numbers and drug status had a binary value.

The Prothro data set was from a study of liver cirrhosis[26]. The longitudinal marker was prothropin level and we used the treatment (placebo or prednisone) as a static covariate. Time (years) and prothrombin levels were scalar numbers, and treatment status was a binary variable.

For Pbc2 and Prothro, we used the LME regression (in methods LC and LME) motivated by Rizopoulos[14] as

$$y_{ij} = \alpha + \beta x_i t_{ij} + \gamma t_{ij} + \theta x_i + u_i + \epsilon_{ij}, \tag{18}$$

where x_i denotes the medication status of individual i.

High-throughput molecular data set. The high-throughtput molecular data set was from the Finnish Type 1 Diabetes Prediction and Prevention (DIPP) study[16] and involved preprocessed mRNA expression levels measured on Affymetrix Human Genome U219 microarray[17]. There was a total of 49,386 probes corresponding to different genes in the data that were measured over time. The data were downloaded from The National Center for Biotechnology Information webpage (https://www.ncbi.nlm.nih.gov/) using Gene Expression Omnibus identifier GSE30211. Each probe was z-scored. The mRNA data consisted of two separate data sets: seroconverted children and progressors. The data set of seroconverted children contained samples from subjects close to the time when diabetes-related autoantibodies developed. Samples of the progressors' data set were concentrated close to the diagnosis of diabetes. Here, we focused on the clinical phenotype, i.e. progressors, and defined the case-control label as diagnosed ($\delta_i = 1$) or not diagnosed ($\delta_i = 0$) with Type 1 diabetes.

To select the probes to be used as the longitudinal covariates, we first used the data from seroconverted children. For each seroconverted child, the first four follow-up samples were selected. Probes with median expression lower than the median of median expressions (5.47) were excluded. The remaining 24,693 probes were ranked in two ways: 1., a ranking based on Wilcoxon's rank sum test P-value between cases and controls for all samples, and 2., a ranking based on Wilcoxon's rank sum test P-value between cases and controls for subjectwise median values. The two rankings were combined by taking the average rank and top five probes were selected. The selected probes were 11751509_a_at, 11723996_a_at, 11759536_a_at, 11733701_a_at and 11748922_x_at, and they mapped to genes *RCN1*, *GLCCI1*, *TTC17*, *FKBP11* and *NSMF*, respectively. For simplicity, we will refer to the probes by using their gene symbols. No static covariates were used.

To construct the predictive models we used the data set progressors. For methods LC and LME, we used age as marker y_{ij} and top 5 probes as longitudinal covariates $\boldsymbol{x}_{ij}$ as

$$y_{ij} = \alpha + \boldsymbol{\theta}' \boldsymbol{x}_{ij} + u_i + \epsilon_{ij}. \tag{19}$$

For the other methods, we trained the models by using averaged information from the 5 top probes and age.

Results

In this section, we represent the results for four scenarios of simulated data and three real data sets. In the simulated data, we considered one static covariate and one longitudinal marker. The four simulation scenarios differed in the distributions of the markers and in sample sizes. The real data sets contained two clinical data sets and one high-throughput molecular data set. We emphasise that we used multiple simulation replicates for simulated data, and exhaustive cross validation for real data.

We compared the performance of LC to a number of statistical and machine learning methods, including LF, LME, GLMER, Lasso, RF, SVM and NN, in terms of their sensitivity and specificity. Results for the methods LME2, Lasso2, RF2, SVM2 and NN2 are collected in the Supplementary material. Additionally, we present the F1 scores, accuracies and the receiver operating characteristic (ROC) curves for each method in the Supplementary material.

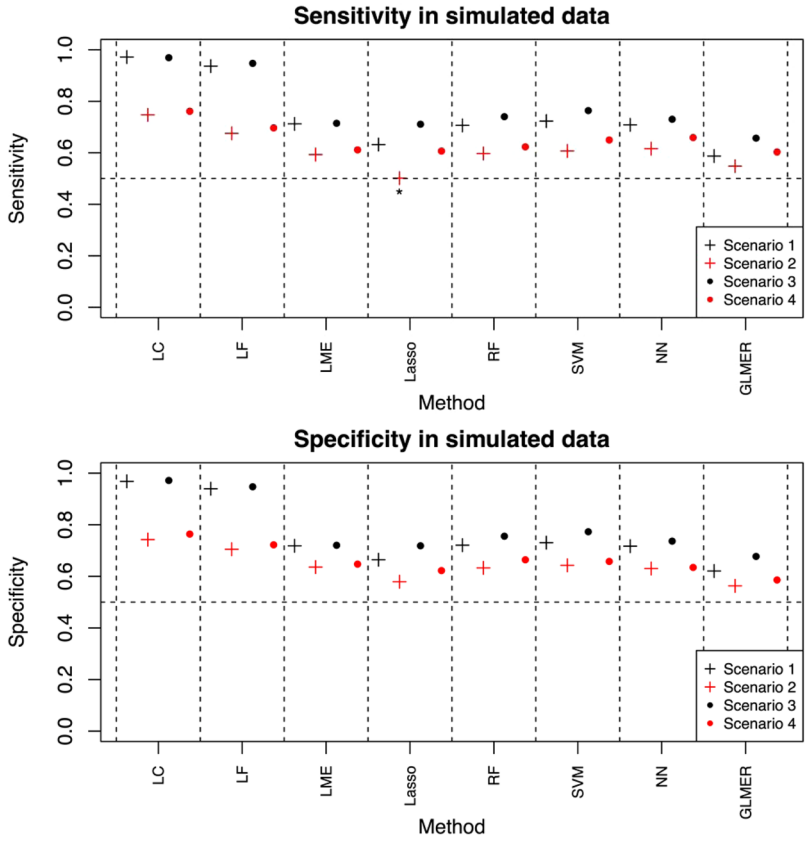

Figure 1. Model performance in simulated data. This figure represents sensitivity and specificity in simulated data. The Scenarios refer to data-generating process. The vertical lines around the dots represent standard error, when visible. Sensitivity and specificity have been calculated from 1,000 Monte Carlo replicates. Statistics *not* significant at the false discovery rate level of 0.05 have been indicated by asterisk (*).

Simulated data. In the simulated data, all methods had fairly good performance (Fig. 1, Supplementary Figs. 1–3, and Supplementary Table 1). All other combinations of methods and scenarios had highly significant sensitivity ($P < 1.0 \times 10^{-6}$) and specificity ($P < 1.0 \times 10^{-6}$) compared to the baseline value 0.5, except for sensitivity of Lasso in Scenario 2 ($P = 0.90$). Thus, it seems that Lasso was not able to distinguish between the cases and controls on the basis of the overlapping distributions of the temporal averages. In all scenarios, LC was the best method in terms of both sensitivity (in each pairwise comparison $P < 1.0 \times 10^{-6}$) and specificity (in each pairwise comparison $P < 1.0 \times 10^{-6}$), closely followed by LF. The F1 scores, accuracies and ROC curves supported similar conclusions (see Supplementary Table 1, and Supplementary Figs. 2 and 3).

The simulated scenarios differed from each other so that Scenarios 1 and 3 had greater distinction between cases and controls than Scenarios 2 and 4. On the other hand, Scenarios 3 and 4 had more samples than Scenarios 1 and 2. As expected, most methods achieved the best results in Scenarios 3 and 4, which had higher numbers of training samples (Fig. 1, Supplementary Fig. 1). However, for some methods (LC and LME), the difference between Scenarios 1 and 3, i.e. a difference attributable to sample size, was very small. Regarding the effect of the sample distributions, better results were achieved in Scenario 1 than in Scenario 2, as Scenario 2 had a smaller difference between the sample groups. A similar observation holds for Scenarios 3 and 4, as expected (Fig. 1, Supplementary Fig. 1). Methods LME2, Lasso2 and RF2 had similar performance compared to the corresponding averaged methods LME, Lasso and RF. In Scenarios 1 and 3, methods SVM2 and NN2 outperformed SVM and NN.

To conclude, the results obtained from the simulated data sets demonstrated that all methods could deliver meaningful results, and LC had a very good performance, as compared to the twelve other contemporary approaches tested.

Real data. Although all the methods tested here performed well in the simulated data sets, this pattern changed when we moved to the real data sets (Fig. 2, Supplementary Figs. 4–6, and Supplementary Table 2). While LC and RF had both specificity and sensitivity highly significantly over 0.50 in all three real data sets ($P < 10^{-6}$), this was not the case for any of the other methods. Instead, all the other methods had either sensitivity or specificity below 0.55 in at least one data set. The Prothro and DIPP data sets turned out to be the hardest to predict in terms of sensitivity and specificity. In the Prothro data, LC achieved sensitivity of 0.65 and specificity of 0.70, and in the DIPP data, sensitivity of 0.84 and specificity of 0.63. The relative difficulty of the real data sets was also seen in the F1 values, accuracies and ROC curves (Supplementary Table 2, Supplementary Figs. 5 and 6). LC was the only method that obtained accuracy and F1 value higher than 0.7 in all real data sets (Supplementary

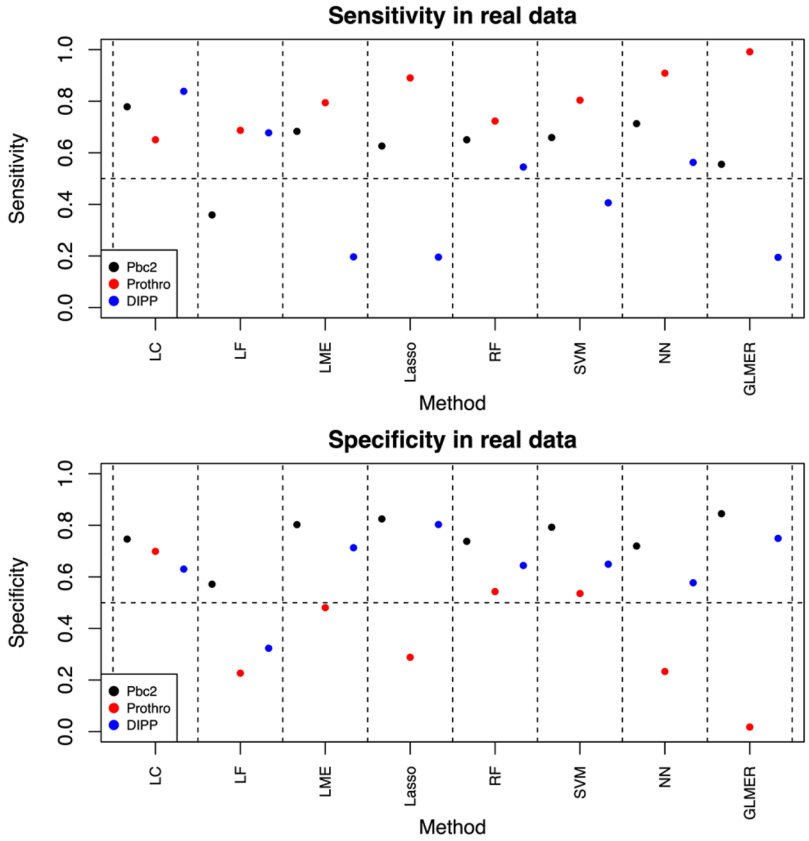

Figure 2. Model performance in real data. Pbc2 and Prothro are clinical data sets, whereas DIPP is a molecular data set from Type 1 Diabetes. The vertical lines around the dots represent standard error, when visible. Sensitivity and specificity have been calculated by using exhaustive cross validation, thus the small standard errors.

Table 2). In all real data sets, methods LME2, Lasso2, RF2, SVM2 and NN2 were slightly outperformed by the corresponding averaged methods.

Based on these results, one may conclude that LC was the only robust method among the tested thirteen, as it did not fail in any data set analysed in this study.

Discussion and Conclusions

In this study, we examined thirteen methods for binary classification of longitudinal data with non-aligned time points, which is a common scenario in biomedical studies. (Most of these methods needed to be adjusted on an ad-hoc basis to acknowledge for the longitudinal nature of the data, i.e. we used temporal averages of covariates. However, this was not the case for our method.) We introduced the method of likelihood contrasts (LC) and compared its performance to the twelve other approaches, using simulated data and three real data sets. In the simulated data sets, LC clearly outperformed the other methods in terms of sensitivity and specificity ($P < 1.0 \times 10^{-6}$). In the real data sets, the performance of all methods was lower than in the simulated data. However, unlike most of the other methods, LC provided reasonable classification performance also in all the real data sets (specificity and sensitivity significantly over 0.50, $P < 10^{-6}$), demonstrating its robustness over the other methods.

Another benefit of LC is that it can be generalised to any analysis scenario consisting of two models and a measure of model fit. For example, we used the difference of log-likelihood to measure the agreement between a new observation and the pre-existing data. In a non-parametric setting, one could use some other measure of model fit, such as the difference of mean squared errors.

Presently, we have used LC in combination with two LMEs. This derives from previous modelling tradition for longitudinal data[6,7], but could also be changed. For example, the log-likelihood could be derived from a non-linear regression with time. However, a benefit of the LME framework is that it is fairly general, and the theory of these models is well-known. Moreover, the LME framework can easily be extended to allow for a wider range of applications. For example, it is possible to use penalised random-effects models for automated model choice[4].

There are multiple studies comparing different binary classifiers for biomedical single time-point data. For example, Khondoker *et al.*[27] compared four classification methods using simulated and real data sets. They concluded that linear discriminant analysis gave the best results for small data sets and SVM for data sets with more than 20 samples. Johnson *et al.*[28] compared six classifiers in RNA-seq data. They found random forest to be the

best method, and transcript-level data to be better suited for classification than gene-level data. Babu *et al.*[29] studied the effect of feature selection for classification methods in cancer. They found that feature selection substantially improved the performance of classification, and in their study, SVM was one of the best methods together with Relief-F[30] and information gain[31].

Given the prevalence of longitudinal data sets in biomedicine, it is surprising that there are so few longitudinal binary classifiers. Longitudinal time-series experiments using DNA microarrays have already been performed for more than a decade[32–34]. In line with this, various statistical methods have been developed to detect the differentially expressed genes between experimental groups in the longitudinal data. For example, MaSigPro[32] can be used to analyse inter-group differences by fitting polynomials of various degrees to expression data. The moderated F-test in limma[35] can be used to discover differentially expressed genes by considering intergroup contrasts at different time points. Approaches relying on Bayesian statistics for detecting longitudinal differential gene expression have also been developed[33,34]. However, none of these methods directly addresses the question of classifying the longitudinal samples in two distinct groups, e.g. in patients and healthy controls. This is a shortcoming which we try to address by the proposed LC method.

Regarding binary classification of longitudinal data, there are some methods which operate on aligned time points[4]. However, a fully flexible model such as LC has not been developed before. Consequently, it is difficult to judge the performance of LC against a pre-existing baseline. Furthermore, regarding binary classification in static settings, earlier studies have not been able to highlight a single generally best method. For example, Pirooznia *et al.*[36] used eight machine learning algorithms in eight microarray gene expression data sets, and they found SVM to have the best performance. In line with this, Castillo *et al.*[37] analysed RNA-sequencing and microarray data sets, finding SVM to be more accurate than RF or nearest-neighbour classification. However, Bienkowska *et al.*[38] used SVM and RF in combination with iterated feature selection in gene expression data. They found RF to outperform SVM in three out of four cases, in terms of area under the ROC curve. Such comparisons are numerous and can be found in many application areas[39,40].

The two real clinical data sets used in our study were taken from Rizopoulos[14]. As our primary purpose was to compare the performance of the different classifiers, we used the same covariates and markers as the earlier studies[14]. In the Type 1 Diabetes data[17], the choice of the predictive markers was less obvious. In these high-throughput molecular data, there were initially 49,386 measured probes. Following observations from previous studies, we filtered these features. For example, Pirooznia *et al.*[36] have reported that up to 10% losses in predictive accuracy can be expected, if all features are used in place of an optimal feature set.

To conclude, results obtained in this study suggest that LC can be used as an accurate binary classifier in longitudinal data. LC outperformed the conventional machine learning methods in the simulated data. Although the three real data sets proved to be more difficult to predict correctly than the simulated data, LC was able to deliver statistically significant predictions in all data sets.

Data availability

The data sets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 13 June 2019; Accepted: 31 December 2019;
Published online: 23 January 2020

References

1. Tibshirani, R. Regression ShRegression Shrinkage and Selection via the Lassorinkage and Selection via the Lasso on JSTOR. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
2. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
3. Rosenblatt, F. & Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain. *Psychol. Rev.* 65–386 (1958).
4. Chen, S., Grant, E., Wu, T. T. & Bowman, F. D. Some recent statistical learning methods for longitudinal high-dimensional data. *Wiley Interdiscip. Rev. Comput. Stat.* **6**, 10–18 (2014).
5. Lütkepohl, H. *New introduction to multiple time series analysis*. (New York, 2005).
6. Fieuws, S., Verbeke, G. & Molenberghs, G. Random-effects models for multivariate repeated measures. *Stat. Methods Med. Res.* **16**, 387–397 (2007).
7. Bandyopadhyay, S., Ganguli, B. & Chatterjee, A. A review of multivariate longitudinal data analysis. *Stat. Methods Med. Res.* **20**, 299–330 (2011).
8. Verbeke, G., Fieuws, S., Molenberghs, G. & Davidian, M. The analysis of multivariate longitudinal data: a review. *Stat. Methods Med. Res.* **23**, 42–59 (2014).
9. Jensen, S. M. & Ritz, C. A comparison of approaches for simultaneous inference of fixed effects for multiple outcomes using linear mixed models. *Stat. Med.* **37**, 2474–2486 (2018).
10. Parzen, M. *et al.* A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *Ann. Appl. Stat.* **5**, 449–467 (2011).
11. Albert, P. S. A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Stat. Med.* **31**, 143–54 (2012).
12. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R_Core_Team. Linear and Nonlinear Mixed Effects Models [R package nlme version 3.1-137].
13. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J. & Scholkopf, B. Support vector machines. *IEEE Intell. Syst. their Appl.* **13**, 18–28 (1998).
14. Rizopoulos, D. *Joint models for longitudinal and time-to-event data: with applications in R*. (CRC Press, 2012).
15. Rizopoulos, D. JM: An R Package for the Joint Modelling of Longitudinal and Time-to-Event Data. *J. Stat. Softw.* **35**, 1–33 (2010).
16. Kupila, A. *et al.* Feasibility of genetic and immunological prediction of type I diabetes in a population-based birth cohort. *Diabetologia* **44**, 290–7 (2001).
17. Kallionpää, H. *et al.* Innate immune activity is detected prior to seroconversion in children with HLA-conferred type 1 diabetes susceptibility. *Diabetes* **63**, 2402–14 (2014).

18. Gelman, A. *et al. Bayesian data analysis*. (CRC Press, Boca Raton, FL, 2013).
19. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
20. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
21. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).
22. Meyer, D. *et al*. e1071: Misc Functions of the Department of Statistics, ProbabilityTheory Group (Formerly: E1071), TU Wien [R package e1071 version 1.7-0]. (2018).
23. Venables, W. N. (William N., Ripley, B. D. & Venables, W. N. (William N.). *Modern applied statistics with S*.
24. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
25. Murtaugh, P. A. *et al*. Primary biliary cirrhosis: Prediction of short-term survival based on repeated patient visits. *Hepatology*, https://doi.org/10.1002/hep.1840200120 (1994).
26. Andersen, P. K., Borgan, Ø., Gill, R. D. & Keiding, N. *Statistical Models Based on Counting Processes*, https://doi.org/10.1007/978-1-4612-4348-9 (Springer US, 1993).
27. Khondoker, M., Dobson, R., Skirrow, C., Simmons, A. & Stahl, D. A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Stat. Methods Med. Res.* **25**, 1804–1823 (2016).
28. Johnson, N. T., Dhroso, A., Hughes, K. J. & Korkin, D. Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers? *RNA* **24**, 1119–1132 (2018).
29. Babu, M. & Sarkar, K. A comparative study of gene selection methods for cancer classification using microarray data. In *Proceedings - 2016 2nd IEEE International Conference on Research in Computational Intelligence and Communication Networks, ICRCICN* **2016** 204–211, https://doi.org/10.1109/ICRCICN.2016.7813657 (Institute of Electrical and Electronics Engineers Inc., 2017).
30. Wall, M. E., Rechtsteiner, A. & Rocha, L. M. Singular Value Decomposition and Principal Component Analysis. In *A Practical Approach to Microarray Data Analysis* 91–109 (Kluwer Academic Publishers), https://doi.org/10.1007/0-306-47815-3_5.
31. Abusamra, H. A comparative study of feature selection and classification methods for gene expression data of glioma. In *Procedia Computer Science* **23**, 5–14 (Elsevier B.V., 2013).
32. Conesa, A., Nueda, M. J., Ferrer, A. & Talon, M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22**, 1096–1102 (2006).
33. Tai, Y. C. & Speed, T. P. A multivariate empirical Bayes statistic for replicated microarray time course data. *Ann. Stat.* **34**, 2387–2412 (2006).
34. Aryee, M. J., Gutiérrez-Pabello, J. A., Kramnik, I., Maiti, T. & Quackenbush, J. An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation). *BMC Bioinformatics* **10**, 409 (2009).
35. Ritchie, M. E. *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
36. Pirooznia, M., Yang, J. Y., Yang, M. Q. & Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* **9**, S13 (2008).
37. Castillo, D. *et al*. Integration of RNA-Seq data with heterogeneous microarray data for breast cancer profiling. *BMC Bioinformatics* **18**, 506 (2017).
38. Bienkowska, J. R. *et al*. Convergent random forest predictor: Methodology for predicting drug response from genome-scale data applied to anti-TNF response. *Genomics* **94**, 423–432 (2009).
39. Parmar, C., Grossmann, P., Bussink, J., Lambin, P. & Aerts, H. J. W. L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* **5**, 13087 (2015).
40. Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One* **12**, e0174944 (2017).

Acknowledgements

Dr. Elo reports grants from the European Research Council ERC (677943), European Union's Horizon 2020 research and innovation programme (675395), Academy of Finland (296801, 304995, 310561 and 313343), Juvenile Diabetes Research Foundation JDRF (2-2013-32), Tekes – the Finnish Funding Agency for Innovation (1877/31/2016) and Sigrid Juselius Foundation, during the conduct of the study. Thanks are due for Dr. Asta Laiho and MSc. Tommi Välikangas for commenting the manuscript.

Author contributions

R.K. invented the method, carried out the statistical analysis, prepared the figures and wrote the manuscript. M.K. implemented the method, carried out the statistical analysis, prepared the figures and wrote the manuscript. L.E. conceived the study, supervised the work and wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-57924-9.

Correspondence and requests for materials should be addressed to L.L.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.