BMC Bioinformatics

RESEARCH ARTICLE Open Access

Hellinger distance-based stable sparse feature selection for high-dimensional class-imbalanced data

Guang-Hui Fu[1*], Yuan-Jiao Wu[1], Min-Jie Zong[1] and Jianxin Pan[2]

Abstract

Background: Feature selection in class-imbalance learning has gained increasing attention in recent years due to the massive growth of high-dimensional class-imbalanced data across many scientific fields. In addition to reducing model complexity and discovering key biomarkers, feature selection is also an effective method of combating overlapping which may arise in such data and become a crucial aspect for determining classification performance. However, ordinary feature selection techniques for classification can not be simply used for addressing class-imbalanced data without any adjustment. Thus, more efficient feature selection technique must be developed for complicated class-imbalanced data, especially in the context of high-dimensionality.

Results: We proposed an algorithm called sssHD to achieve stable sparse feature selection applied it to complicated class-imbalanced data. sssHD is based on the Hellinger distance (HD) coupled with sparse regularization techniques. We stated that Hellinger distance is not only class-insensitive but also translation-invariant. Simulation result indicates that HD-based selection algorithm is effective in recognizing key features and control false discoveries for class-imbalance learning. Five gene expression datasets are also employed to test the performance of the sssHD algorithm, and a comparison with several existing selection procedures is performed. The result shows that sssHD is highly competitive in terms of five assessment metrics. In addition, sssHD presents limited differences between performing and not performing re-balance preprocessing.

Conclusions: sssHD is a practical feature selection method for high-dimensional class-imbalanced data, which is simple and can be an alternative for performing feature selection in class-imbalanced data. sssHD can be easily extended by connecting it with different re-balance preprocessing, different sparse regularization structures as well as different classifiers. As such, the algorithm is extremely general and has a wide range of applicability.

Keywords: Hellinger distance, Class-imbalance learning, Feature selection, Sparse regularization

Background

Feature selection has recently gained considerable attention in class-imbalance learning due to the high-dimensionality of class-imbalanced data across many scientific disciplines [1–3]. To date, a variety of feature selection methods have been proposed to address high-dimensional data. However, only a small number of them are technically designed to handle the problem of class distribution under a class-imbalance setting [4–7]. Thus, performing feature selection from class-imbalanced data remains a challenging task due to the inherent complex characteristics of such data, and a new understanding or principle is required to efficiently transform vast amounts of raw data into information and knowledge representation [8].

*Correspondence: guanghuifu@kust.edu.cn;ghuifu@126.com
[1] School of Science, Kunming University of Science and Technology, Kunming 650500, People's Republic of China
Full list of author information is available at the end of the article

Feature selection can simplify data by eliminating uninformative predictors as well as selecting the key biomarkers for a certain task. In addition, feature selection is an effective strategy to alleviate the overlap caused by the interaction of high-dimensionality and class-imbalance [5, 9, 10]. In fact, standard classifiers can still produce good discrimination for some highly class-imbalanced data sets if the data (or two class distributions) at hand can be well separated, regardless of the class-imbalanced ratio and the lack of data. However, overlapping (non-separability) usually occurs in the settings of high-dimensionality and class-imbalance. An instance from class C belongs to an overlapping region if out of its k nearest neighbors, more than h (such as $h\ =[k/2]$) belong to a class other than C. Overlapping happens when a similar amount of training data for each class is mixed in the overlapping region. When overlapping arises, it is very difficult or impossible to separate a class from others. Some findings have shown that overlap can play an even larger role in determining classifier performance than class-imbalance [11]. As far as high-dimensional and class-imbalanced data is concerned, it is worthwhile to investigate the way to alleviate the overlap effectively with feature selection.

There are three categories of feature selection in the context of classification, depending on how these feature selection searches combine with the construction of the classification model: filtering, wrapping and embedding [12]. The filtering method assesses the relevance of features by looking only at the intrinsic properties of the data, and selects high-ranking features based on a statistical or information measure, such as information gain and gain ratio [13]. There are two drawbacks of filter-based selection: first, the filtering selection is independent of the classifier, which may lead to reduced classification accuracy with a certain kind of classifier; and second, it ignores the dependencies among features. Such dependency information should be considered in performing variable selection as strongly related features are often similar and should be aggregated. The related features may play an important role in performing feature selection, especially in high-dimensional settings. The wrapping method, such as a genetic algorithm [14], wraps a search algorithm around the classification model to search the space of all feature subsets. However, an obvious drawback of the wrapping method is that it is computationally intensive, as the number of subsets from the feature space grows exponentially as the number of features increases. The embedding method screens out key features while considering the construction of a classifier, such as LASSO-based feature selection and classification [15]. It is integrated in the modelling process and is classifier-dependent [12].

In class-imbalance scenarios, filtering and wrapping feature selection methods were the most frequently built and were used to solve real-world problems, such as disease diagnosis, textual sentiment analysis, and fraud detection [16]. Dozens of metrics and their variants are employed for building filter-based feature selection algorithms in many research studies [7, 17–19], such as odds-ratio, chi-squared, Relief, ReliefF, information gain, gain ratio, Gini index, $F - measure$, $G - mean$, signal-to-noise ratio, and area under receiver operating characteristics (ROC) graph. To effectively reduce the computation cost, wrapping feature selection techniques usually utilize ad hoc search strategies, such as a heuristic search [20, 21] and stochastic search [22]. To the best of our knowledge, embedding feature selection methods are less investigated than filtering and wrapping methods. One of the related studies is from the reference [23], which considers sparse logistic regression with stable selection in handling Alzheimer's disease neuroimaging initiative dataset and stated that it achieved competitive performance compared with several filter-based selection methods based on their experimental results.

Considering the drawbacks of filtering and wrapping algorithms, in this study, we focus on the embedding-based selection algorithm by bringing in sparse regularization [24] . Common embedding feature selection methods for classification can not be used simply for addressing class-imbalanced data without any adjustment because of the following issues:

(a) These standard selection algorithms are generally based on the assumption of balanced class distributions, and the selection results are affected due to the class-imbalanced ratio between classes and consequently produce highly biased classification prediction towards the majority class.

(b) The classifier's continuous output, to some extent, may shift due to the domination of the majority class [25, 26]. Figure 1 exhibits such shifting based on a support vector machine (SVM) classifier. It can be seen that the decision score varies greatly as the class-imbalanced ratio changes. When the threshold in decision function still keeps the default value (e.g., 0 in SVM and 0.5 in logistic regression), the decision boundary (or the separating hyperplane) must be shifted towards the minority. From the view of geometry, the separating hyperplane would be shifted towards the minority class due to the domination of the majority class. One of the attempts for handling this question is threshold adjustment[25, 27–29] via moving the decision threshold towards the majority examples so that the minority class examples become harder to misclassify. However, as was pointed out in the references [25, 26], it is difficult to decide how far the separating hyperplane should be moved towards the majority class, and such adjustment may over-correct the decision boundary towards the majority class, which leads to increasing error on the majority class.

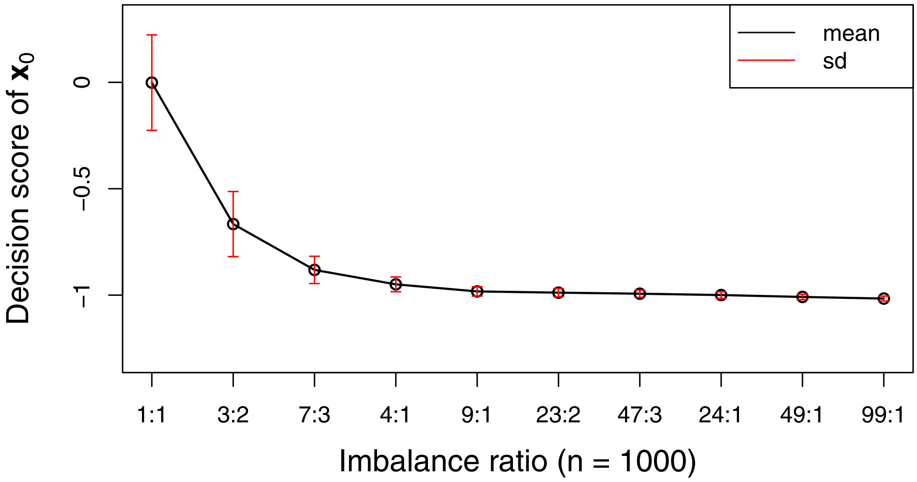

Fig. 1 Decision score of the SVM classifier varies greatly as the class-imbalanced ratio changes. In this simulation, the *X*-matrix (1000 × 3) is randomly produced then fixed while the class label $y = \pm 1$ is randomly changed at each iteration under the fixed class-imbalanced ratio. The prediction point is set be $\mathbf{x}_0 = (-5, 9, -1)^T$. The black line and the red vertical short segments are, respectively, the mean and standard deviation of prediction decision score with the class-imbalanced ratio changing from 1 to 99 based on 2000 iterations

(c) Common feature selection measurements are not suitable in class-imbalance learning. Traditionally, feature selection techniques were developed to maximize the total classification accuracy of a classifier. As is well-known, the majority class is more influential than the minority class in performing feature selection.

In this study, the Hellinger distance is employed as an assessment measurement on the assumption of binormal distributions to combat class-imbalance and output-shifting in class-imbalance learning.

Methods

Hellinger distance under binormal assumption

The Hellinger distance is a measure of the distributional divergence [30]. Let P and Q be two probability measures that are absolutely continuous with respect to a third probability measure λ. The square of the Hellinger distance can be defined as follows:

$$D_H^2(P, Q) = \int \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}} \right)^2 d\lambda \tag{1}$$

Here, λ is set be the Lebesgue measure, so that $\frac{dP}{d\lambda}$ and $\frac{dQ}{d\lambda}$ are two probability density functions. Based on the binormal assumption, P and Q are two normal distributions, and

$$\begin{cases} \frac{dP}{d\lambda} = f_1(x) \sim N\left(\mu_1, \sigma_1^2\right), \\ \frac{dQ}{d\lambda} = f_0(x) \sim N\left(\mu_0, \sigma_0^2\right) \end{cases} \tag{2}$$

Thus Eq. (1) can be rewritten as

$$\begin{aligned} D_H^2(P, Q) &= \int \left(\sqrt{f_1(x)} - \sqrt{f_0(x)} \right)^2 dx \\ &= 2 - 2 \int \sqrt{f_1(x) f_0(x)} dx \\ &= 2 - 2 \sqrt{\frac{2\sigma_1\sigma_0}{\sigma_1^2 + \sigma_0^2}} Exp \left\{ -\frac{(\mu_1 - \mu_0)^2}{4\left(\sigma_1^2 + \sigma_0^2\right)} \right\} \end{aligned} \tag{3}$$

In practice, the parameters μ_1, σ_1^2, μ_0, and σ_0^2 can be replaced by the corresponding sample statistics $\bar{X}_1$, S_1^2, $\bar{X}_0$, and S_0^2, respectively.

Without a loss of generality, let $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ be binary categorical response and $y_i = 1$ if it belongs to the minority class (positive) and $y_i = -1$ if it belongs to the majority class (negative); let $\boldsymbol{x}_j$ be the j^{th} feature ($j = 1, 2, \cdots, p$) and $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p)$; and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ be the vector of estimate coefficients. The normality assumption here is on a linear combination of the predictor matrix $\mathbf{X}$ rather than each single feature, namely

$$P = \mathbf{X}\boldsymbol{\beta} \Big| (\mathbf{y} = 1) = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_p \boldsymbol{x}_p \Big| (\mathbf{y} = 1), \tag{4}$$

$$Q = \mathbf{X}\boldsymbol{\beta} \Big| (\mathbf{y} = -1) = \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \cdots + \beta_p \boldsymbol{x}_p \Big| (\mathbf{y} = -1) \tag{5}$$

Obviously, the binormal assumption on a linear combination would be more likely to hold than a single variable, particularly for moderate to large size of features p according to the central limit theorem (CLT).

Definition 1. A quantity is called skew-insensitive if it is not influenced by the class priors.

Definition 2. Let $t_i (i = 1, 2, \cdots, n)$ be the original score of each observation and c be a constant $(c \neq 0)$. A quantity is called translation-invariant if it remains unchanged when each score moves to $t_i + c, (i = 1, 2, \cdots, n)$.

Property 1. The Hellinger distance is skew-insensitive under binormal assumption.

Equation (3) shows that the computation of the Hellinger distance is not influenced by the class-imbalanced ratio. It is just relevant with the expectations μ_0, μ_1 as well as variances σ_0^2, σ_1^2 of P and Q. The law of large numbers tells us that these four numerical characteristics are approached by their corresponding sample statistics if the sample size is large enough. They are independent of the class-imbalanced ratio. An example is given in Fig. 2 to exhibit this skew-insensitivity by means of calculating the magnitude of the Hellinger distance on two normal distributions. It can be seen from Fig. 2 that the value of the Hellinger distance stays consistent when the class-imbalanced ratio changes from 1 to 99, and such consistency tends to become increasingly true as the sample size increases. Namely, the magnitude of the Hellinger distance is not influenced by the class-imbalanced ratio. In fact, such skew-insensitivity has also been shown in the reference [31] in terms of comparing isometrics and giving a synthetic example.

Property 2. Hellinger distance is translation-invariant under binormal assumption.

Considering that two variances σ_0^2 and σ_1^2 as well as the difference $\mu_1 - \mu_0$ keep invariant as each score moves, the Hellinger distance will stay the same according to Eq. (3).

As mentioned above, Hellinger distance essentially captures the divergence between the feature value distributions of different classes and is not influenced by the class ratios under binormal assumption. This is the motivation why Hellinger distance is utilized for class-imbalanced data in this study. In addition, its translation-invariant is very useful to combat the output-shifting arisen when a standard classifier is used to distinguish class-imbalanced data. Unlike the usage of Hellinger distance in the previous work [5, 31], where the feature attributes should be discrete or to discretize the continuous features for the calculation of Hellinger distance, Hellinger distance in this study can be calculated directly based on continuous variables without discretization.

Hellinger distance-based stable sparse selection (sssHD) and its algorithm

Considering the above questions from class-imbalance learning and being motivated by the properties of the Hellinger distance, we proposed a Hellinger distance-based stable sparse selection (sssHD) approach to perform feature selection when the category data is class-

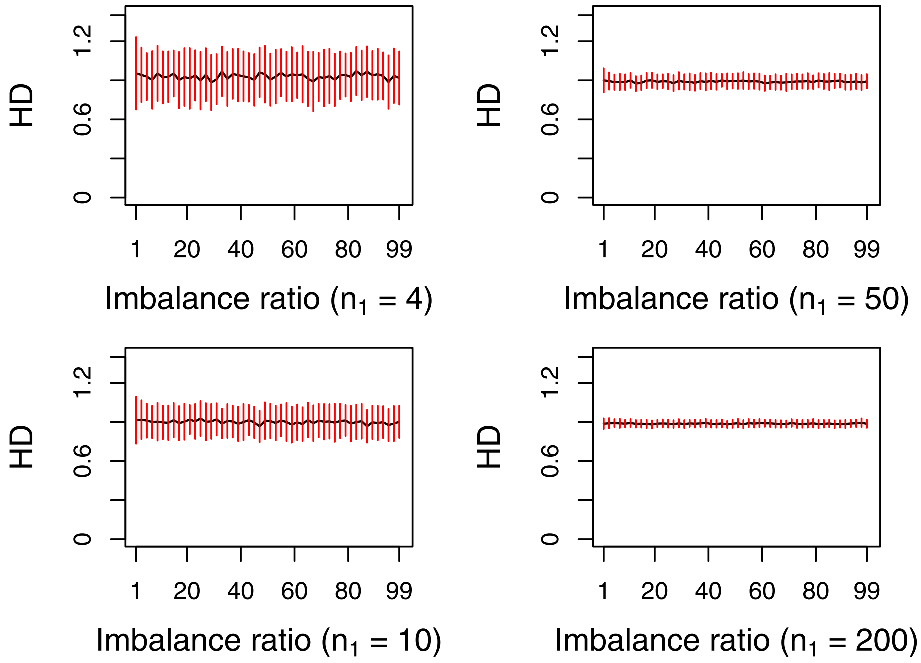

Fig. 2 Skew-insensitivity of the Hellinger distance on simulated normal distributions with four scenarios corresponding to different sizes of the minority. The black line and red vertical short segments are, respectively, the mean and standard deviation of the Hellinger distance with the ratio changing from 1 to 99 based on 100 iterations

imbalanced. An ordinary classifier generally can not perform feature selection automatically, but a kind of sparse penalty coupled with the Hellinger distance metric can be embedded into the classifier to achieve such a task. For convenience, a linear SVM classifier is employed as an example to establish our sssHD algorithm.

SVM [32–35] has shown promising capability in solving many classification problems. It performs two-classification task by constructing a hyperplane in the multidimensional space to differentiate two classes with a maximal margin. Let $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^T$ be the i^{th} instance and its class label $y_i = 1$ or -1 $(i = 1, 2, \cdots, n)$. The decision function of SVM can be expressed as follows:

$$f(\mathbf{x}_i) = \begin{cases} 1, & \boldsymbol{\beta}^T\mathbf{x}_i + \beta_0 > 0, \\ -1, & \boldsymbol{\beta}^T\mathbf{x}_i + \beta_0 \le 0 \end{cases} \tag{6}$$

where β_0 is the constant coefficient, and $\boldsymbol{\beta}^T\mathbf{x}_i+\beta_0$ is called the decision score in this study. The soft margin support vector classifier can be estimated by solving the following quadratic optimization problem:

$$\min \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i, \tag{7}$$

$$s.t. \quad \begin{cases} y_i(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0) \ge 1 - \xi_i, \\ \xi_i \ge 0, \quad (i = 1, 2, \cdots, n), \end{cases} \tag{8}$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_n)^T$ are slack variables which are associated with the misclassified individuals. Formula (7) with constraint (8) can be rewritten as

$$\min \quad Loss + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{9}$$

where $Loss = \sum_{i=1}^{n} max(0, 1 - y_i(\boldsymbol{\beta}^T\mathbf{x}_i + \beta_0))$ is hinge loss, and $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p}\beta_j^2$ is the ridge penalty. The ridge penalty shrinks the estimation coefficients towards zero and, hence, possibly improves the model's prediction accuracy; however, it can not perform feature selection automatically. Therefore, the ridge penalty should be replaced by a sparse regularization penalty to induce the sparsity for achieving feature selection. Sparse selection is a very popular technique to perform variable selection for high-dimensional data [24, 36–39]. Taking elastic-net [38] as an example, it is defined as follows:

$$C_\alpha(\boldsymbol{\beta}) = \frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1, \tag{10}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ is LASSO penalty[24], and $\alpha \in [0, 1]$. Actually, the elastic-net penalty is a combination of ridge and LASSO penalties, which is particularly useful and effective for feature selection, especially when the data is strongly correlated and high-dimensional. Sparse support vector machine with elastic-net penalty can be expressed as

$$\min \quad Loss + \lambda C_\alpha(\boldsymbol{\beta}), \tag{11}$$

where λ is the tuning parameter that controls the tradeoff between loss and penalty.

The optimal estimation $(\hat{\boldsymbol{\beta}}, \hat{\beta}_0)$ in objective (11) is the function of λ and α. Consequently, the decision score $\hat{\mathbf{t}} = (\hat{t}_1, \hat{t}_2, \cdots, \hat{t}_n)^T = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\beta}_0$ is also influenced by λ and α. Denoted $\hat{\mathbf{t}}$ by $\hat{\mathbf{t}}(\lambda, \alpha) = (\hat{\mathbf{t}}_0(\lambda, \alpha), \hat{\mathbf{t}}_1(\lambda, \alpha))^T$, where $\hat{\mathbf{t}}_0(\lambda, \alpha) = \{\hat{t}_i | y_i = -1\}$ and $\hat{\mathbf{t}}_1(\lambda, \alpha) = \{\hat{t}_i | y_i = 1\}$, the objective of sparse selection with Hellinger distance can be defined as

$$\hat{\boldsymbol{\beta}} = \max \quad D_H\left(\hat{\mathbf{t}}_0(\lambda, \alpha), \hat{\mathbf{t}}_1(\lambda, \alpha)\right) \tag{12}$$

A potential question of sparse feature selection is its instability caused by the variation from the training data [40, 41]. Class-imbalance is going to exacerbate this drawback. A decent strategy to overcome such disadvantage is to combine sparse selection with subsampling. Meinshausen et al. [40] pointed out that such marriage yields finite sample family-wise error control and significantly improves selection methods. In this study, objective (12) is conducted many times with subsampling to achieve stable selection. Denoted $\hat{\boldsymbol{\beta}}$ from objective (12) by $\hat{\boldsymbol{\beta}}^{(k)}$ in the k^{th} subsampling $(k = 1, 2, \cdots, K)$. The importance of the features is measured by the inclusion frequency, which is denoted by $\mathbf{f} = (f_1, f_2, \cdots, f_p)^T$, and is defined as follows:

$$f_j = \frac{1}{K}\sum_{k=1}^{K} g\left(\hat{\beta}_j^{(k)}\right), \qquad j = 1, 2, \cdots, p, \tag{13}$$

where $g\left(\hat{\beta}_j^{(k)}\right) = 1$ if $\hat{\beta}_j^{(k)} \neq 0$, otherwise $g\left(\hat{\beta}_j^{(k)}\right) = 0$. All the features are ranked with their inclusion frequencies, and the feature with maximal inclusion frequency is the most important. More details of the sssHD algorithm is given in Algorithm 1. The ratios of subsampling from the majority (r_0) and minority (r_1) are set to be equal in this study to keep the class-imbalance ratio of the subset the same as the original data. sssHD is extremely general and can be easily extended; for example, sparse SVM can be placed by sparse logistic regression [42] or Fisher linear discriminant [43]; re-balance methods such as oversampling [44] or under-sampling [45] could be connected if necessary; sparse regularization (Eq. (10)) also has many alternatives, such as SCAD [36], adaptive LASSO [39], group LASSO [46], and group bridge penalty [47].

Assessment metrics and experimental methods

In class-imbalance learning, the majority and the minority are generally called as negative and positive, respectively. A binary classifier predicts all the instances as either positive or negative. Thus, it produces four types of outcome: true positive (*TP*), true negative (*TN*), false positive (*FP*) and false negative (*FN*). Several metrics can be defined

Algorithm 1 Hellinger distance based stable sparse selection (sssHD)

```
$(\mathbf{X}_0, \mathbf{y}_0)$: predictor matrix and class label;
$n_0, n_1$: the size of the majority, minority respectively;
$r_0, r_1$: the ratio of subsampling from the majority, minority respectively;
$p$: the number of features;
$T$: times of iteration.
Initial $\mathbf{f} \leftarrow (0, 0, \cdots, 0)^T \in R^p$;
for $k = 1$ to $K$ do
    Randomly picking out $n_0 r_0$, $n_1 r_1$ samples from the majority, minority respectively, then merging together, denoted by $(\mathbf{X}, \mathbf{y})$;
    Do feature selection under the objective (12), then $f_j \leftarrow f_j + 1$ if $\hat{\beta}_j \neq 0$ , where $\hat{\beta}_j$ is the estimate coefficient of the $j^{th}$ predictor for all $j \in \{1, 2, \cdots, p\}$;
end for
$\mathbf{f} \leftarrow \mathbf{f}/K$.
```

according to these outcomes, such as

$$TPR = recall = \frac{TP}{TP + FN} = \frac{TP}{n_1};$$
$$TNR = \frac{TN}{TN + FP} = \frac{TN}{n_0};$$
$$FPR = \frac{FP}{FP + TN} = \frac{FP}{n_0};$$
$$precision = \frac{TP}{TP + FP};$$
$$G - mean = \sqrt{TPR \times TNR};$$
$$F - measure = \frac{2(precision \times recall)}{precision + recall}$$

As shown in above, *precision* is the proportion of true positives among the positive predictions; *recall* (*TPR*) measures the proportion of positives that are correctly identified; $G - mean$ is the geometric mean of *TPR* and *TNR*, which measures the accuracy on both the majority class and the minority class; $F - measure$ is a kind of combinations of *precision* and *recall*, and is high when both of them are high.

ROC curve can be created by plotting *TPR* on the y-axis against *FPR* on the x-axis at various threshold settings. Let T, T_1 and T_0 denote respectively the continuous outputs for total, positive and negative examples by the binary classifier (such as SVM); T is the mixture of T_1 and T_0. Larger output values are associated with positive examples. So for a given threshold c ($-\infty < c < +\infty$), an example is predicted positive if its value is greater than c. Thus,

$$TPR(c) = P(T_1 > c) = P(T > c|y = 1),$$
$$FPR(c) = P(T_0 > c) = P(T > c|y = 0)$$

A ROC curve may be defined as the set of points:

$$ROC(\cdot) = \left\{ (TPR(c), FPR(c)) \,\middle|\, -\infty < c < +\infty) \right\}$$

The area under ROC (*AUCROC*) is calculated here by using trapezoidal rule [48], where the point with the minimal value at this *FPR* is linked to the point with the maximal value at the next *FPR* value when there is more than one value at the same *FPR* value. Let $F_1, F_2, \cdots, F_K$ be all the different *FPR* values satisfying $F_1 < F_2 < \cdots < F_K$, and T_k^{max} and T_k^{min} are the maximal and minimal *TPR* values corresponding to F_k ($k = 1, 2, \cdots, K$), respectively. The empirical AUCROC with the lower trapezoidal rule is

$$AUCROC = \sum_{k=1}^{K-1} \frac{1}{2} \left(T_k^{min} + T_{k+1}^{max} \right) (F_{k+1} - F_k) \qquad (14)$$

In this study, *TPR*, $G - mean$, $F - measure$, *AUCROC* and *precision* are employed as assessment metrics to perform real data experiments. Cross validation is performed for each real data in computing these measures. In order to keep the invariant of the imbalance ratio in each fold, stratified sampling is utilized. Namely each fold contains the same size of negative (and positive) instances and their class-imbalance ratios are equal to the ratio of original data set.

To evaluate sssHD algorithm with the above real data sets, we compare it with other four filter-based feature selection methods: Fisher score [49], Relief [50], area under receiver operating characteristic (AUCROCfilter) [51] and area under precision-recall curve (AUCPRCfilter) [52].

Fisher score: the Fisher score could strongly depend on the directions of the spread of the data by calculating the difference of each feature's mean values in two classes:

$$F_j = \frac{|\mu_{1j} - \mu_{0j}|}{\sigma_{1j}^2 + \sigma_{0j}^2}, \qquad j = 1, 2, \cdots, p, \qquad (15)$$

where μ_{0j}, μ_{1j}, σ_{0j}^2, and σ_{1j}^2 are respectively the mean and the variance of the j^{th} predictor of the majority and the minority. Attributes with a higher score are more important for separating the two classes. Fisher score has been utilized successfully in many classification issues [53, 54].

Relief: the Relief is a randomized algorithm that attempts to give each predictor a weight indicating its level of relevance to the target. In each iteration, Relief first needs to search two nearest neighbors for any selected example point ($\mathbf{x}_i$): one from the same class ($\mathbf{nearhit}_i$), and one from the other class ($\mathbf{nearmiss}_i$); then, if Euclidean distance is employed, the weight vector $\mathbf{w} = (w_1, w_2, \cdots, w_p)^T$ is updated so that

$$w_j \longleftarrow w_j - (x_{ij} - nearhit_{ij})^2 + (x_{ij} - nearmiss_{ij})^2, \quad j = 1, 2, \cdots, p, \tag{16}$$

where x_{ij}, $nearhit_{ij}$ and $nearmiss_{ij}$ correspond to the j^{th} component of $\mathbf{x}_i$, $\mathbf{nearhit}_i$ and $\mathbf{nearmiss}_i$, respectively. Attributes with a larger weight are more relevant with the response. The Relief method can be applied to a variety of complicated situations and now has several generalizations, such as ReliefF [55].

AUCROCfilter: as stated above, ROC can be used as a metric to evaluate the final model. In addition, ROC and its area could be used as a filter feature selection method when just considering a single predictor each time. To obtain the predicted class labels, ROC should be combined with a classifier. Obviously, an attribute with larger area under the curve is more important for classifying the target. An ROC-based filter feature selection strategy has been used for high-dimensional class-imbalanced data [51].

AUCPRCfilter: as an alternative of ROC, the precision-recall curve (PRC) has gained increased attention recently in class-imbalance learning [56]. The PRC curve is created by plotting the recall on the x-axis against precision on the y-axis at various threshold settings. The area under PRC (AUCPRC) can be seen as the average of the precision weighted by the probability of a given threshold and is utilized to define how a classifier performs over the whole space. Similarly, AUCPRC coupled with a classifier can be used individually as a filter-based feature selection method for each attribute. Attributes with larger AUCPRC are more significant for separating classes.

Results and discussion

Simulation study

In this section, we test our HD-based method with simulation data in a range of settings, comparing it to another two embedded feature selection methods: classification accuracy (ACC)-based and ROC-based sparse selection techniques.

Simulation data

The $\mathbf{X}$-matrix corresponding to two classes are separately generated via multivariate normal distributions. Namely, $\mathbf{X}\big|(y = 0) \sim N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $\mathbf{X}\big|(y = 1) \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. Here $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$, namely, the predictors in two classes have the same covariance but different mean value. The first ten variables are set to be key features and the rest are null variables. The difference between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ is zero for null features, but two for key features. $\boldsymbol{\Sigma}$ is a blocked correlation matrix, with off-diagonal elements of $\rho^{|i-j|}$ for all $i, j = 1, 2, \cdots, 10$ (key features), and $\rho^{|i-j|}$ for all $i, j = 11, 12, \cdots, p$ (irrelevant features). Between-block correlation is zero. The size of total samples is fixed (namely $n_0 + n_1 = 960$), whereas the number of predictors p is set be, on one hand, 100 to evaluate an over-determined case ($n > p$), and on the other hand, 2000 to assess an under-determined situation ($n < p$). The majority to the minority ratio here is set be 1 : 1, 3 : 1, 9 : 1 and 15 : 1, respectively. Low ($\rho = 0$), moderate ($\rho = 0.4$) and high ($\rho = 0.92$) correlation structures are simulated under the above settings. The R code for generating this simulation data can be found in the supplementary information.

Computation and results

SVM with sparse penalty (10) is employed here to perform feature selection. Three measurements, namely ACC, ROC and HD, are utilized to search the optimal parameters in sparse SVM. To be fair, subsampling is not involved in computation with HD and all the model parameters are set to be same for three algorithms. The mean of the number of correctly (C) and incorrectly (IC) selected predictors are calculated based on 500 trials and the results are reported in Table 1. C corresponds to the number of selected variable from 10 key features. It can be seen that most of key features are correctly selected by using HD, especially when the correlation structures among predictors are not too high. In addition, HD performs best compared with ACC and ROC in terms of C in most situations, which means that the statistical power of HD-based technique is extremely competitive in comparison with other two assessments. IC is actually the number of selected features from the null variables. Table 1 shows that IC from HD is quite low. Considering a large number of null features, the false discoveries from HD are much less than that from both ACC and ROC in almost all the cases. Therefore HD-based selection is suitable to recognize key features and control the false discoveries.

Figure 3 shows the false discovery rate (FDR) derived from ACC, ROC and HD under different class-imbalance ratios. It can be easily seen that HD-based selection outperforms ACC-based and ROC-based selections in terms of FDR in most situations. In addition, with the increase of class-imbalance ratio, FDR from HD varies very slowly and this trend is much weaker than that from ACC or ROC. This is consistent with the properties of HD.

Real data study and discussion

Data sets and software

The following five gene expression datasets are employed to test the performance of the sssHD algorithm.

DLBCL dataset

The DLBCL dataset contains 58 diffuse large B-cell lymphomas (DLBCL) and 19 follicular lymphoma (FL) instances [57]. The original data includes the expression profiles of 7129 genes, and 6285 of them are retained by adopting the data preprocessing method [58]. The class-imbalanced ratio of this set is 3.05, and

Table 1 The feature selection result on simulation data in which 10 key biomakers are included. Mean reported based on 500 replications

				C			*IC*		
p/*r*	Ratio	n_0	n_1	ACC	ROC	HD	ACC	ROC	HD
A: $p = 100, r = 0$	1 : 1	480	480	8.43	2.76	9.93	0.10	0.09	0.00
	3 : 1	720	240	10.00	2.95	10.00	3.79	0.77	0.04
	9 : 1	864	96	10.00	6.28	10.00	19.64	1.33	0.42
	15 : 1	900	60	10.00	7.95	10.00	60.36	0.85	1.77
B: $p = 100, r = 0.4^{\|i-j\|}$	1 : 1	480	480	9.96	9.86	9.96	0.25	0.13	0.01
	3 : 1	720	240	10.00	9.96	9.99	6.67	0.90	0.14
	9 : 1	864	96	10.00	9.98	10.00	57.64	3.09	1.14
	15 : 1	900	60	3.80	9.91	10.00	31.01	3.23	1.85
C: $p = 100, r = 0.92^{\|i-j\|}$	1 : 1	480	480	9.98	10.00	10.00	0.51	0.63	0.57
	3 : 1	720	240	10.00	9.84	9.90	1.58	0.59	0.47
	9 : 1	864	96	0.69	8.01	8.01	5.87	0.67	0.59
	15 : 1	900	60	0.00	6.85	6.85	0.00	0.84	0.71
D: $p = 2000, r = 0$	1 : 1	480	480	7.89	2.79	9.94	0.10	0.04	0.00
	3 : 1	720	240	10.00	3.20	10.00	5.08	1.54	0.03
	9 : 1	864	96	10.00	5.83	10.00	98.24	1.83	0.20
	15 : 1	900	60	10.00	7.38	10.00	404.96	5.94	2.13
E: $p = 2000, r = 0.4^{\|i-j\|}$	1 : 1	480	480	9.97	9.91	10.00	0.28	0.18	0.01
	3 : 1	720	240	10.00	9.91	10.00	13.06	1.06	0.16
	9 : 1	864	96	10.00	9.95	10.00	293.50	4.30	0.57
	15 : 1	900	60	10.00	9.93	9.98	711.05	6.34	2.71
F: $p = 2000, r = 0.92^{\|i-j\|}$	1 : 1	480	480	9.95	10.00	10.00	0.76	0.54	0.53
	3 : 1	720	240	10.00	9.73	9.72	4.05	0.28	0.25
	9 : 1	864	96	3.15	7.95	7.95	161.82	0.68	0.56
	15 : 1	900	60	1.07	6.62	6.62	107.61	0.69	0.58

the selected top q predictors are employed to compare our results via several assessment methods, where $q = 1$ *to* $10, 20, 30, 40, 50, 100, 200, 3142$ and 6285.

SRBCT dataset

The SRBCT dataset [58, 59] includes 83 instances in total described by 2308 genes in four classes: the Ewing family of tumors (EWS), Burkitt lymphoma (BL), neuroblastoma (NB) and rhabdomyosarcoma (RMS), which have 29, 11, 18, and 25 cases, respectively. To adapt binary classification and follow the partition performed in reference [6], we investigate BL versus the rest, where their sizes are 11 and 72, respectively. The class-imbalanced ratio of this set is 6.55, and the selected top q predictors are employed to compare our results via several assessment methods, where the $q = 1$ *to* $10, 20, 30, 40, 50, 100, 200, 1154$ and 2308.

CAR dataset

The CAR dataset [60] contains in total 174 instances described by 12533 genes in eleven classes: prostate, bladder/ureter, breast, colorectal, gastroesophagus, kidney, liver, ovary, pancreas, lung adenocarcinomas, and lung squamous cell carcinoma. 9182 of 12533 features are left after doing the data preprocessing [58]. To adapt binary classification and follow the partition performed in the reference [6], we consider class kidney versus the rest, where their sizes are 11 and 163, respectively. Thus the class-imbalanced ratio of this set is 14.82, and the selected top q predictors are employed to compared

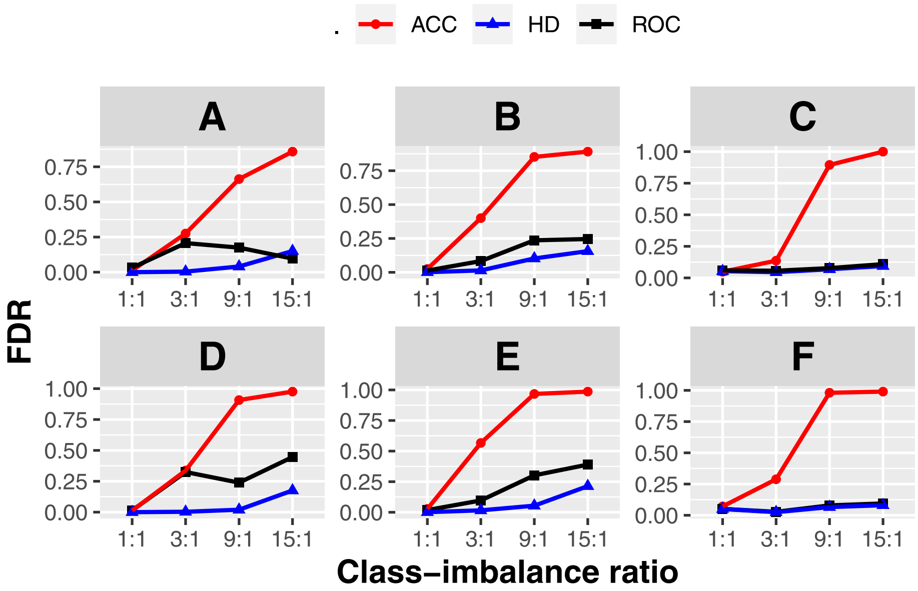

Fig. 3 The *FDR* under different class-imbalance ratios. Six subgraphs correspond to six cases shown in Table 1

our results via several assessment methods, where $q = 1$ *to* 10, 20, 30, 40, 50, 100, 200, 4591 and 9182.

GLIOMA dataset

The GLIOMA dataset [61] contains in total 50 instances described by 12625 genes in four classes: cancer glioblastomas (CG), non-cancer glioblastomas (NG), cancer oligodendrogliomas (CO) and non-cancer oligodendrogliomas (NO). Among the 50 instances, 14, 14, 7, and 15 cases belong to classes CG, NG, CO and NO, respectively. Among the 12625 genes, 4434[1] of them remain after data preprocessing [58]. To adapt binary classification and to follow the partition performed in the reference [6], we study the class CO versus the rest, where the numbers of two classes are 7 and 43, respectively. The class-imbalanced ratio of this set is 6.14, and the selected top q predictors are employed to compared our results via several assessment methods, where $q = 1$ *to* 10, 20, 30, 40, 50, 100, 200, 2217 and 4434.

LUNG dataset

The LUNG dataset [58] includes five classes, two of them are adenocarcinomas and squamous cell lung carcinomas with sample size of 21 and 20 respectively, and are used in our study. Three of all the 3312 predictors are removed beforehand as they are constant or nearly constant, leaving 3309 predictors after preprocessing. The selected top q predictors are utilized to compared our results via several assessment methods, where $q = 1$ *to* 10, 20, 30, 40, 50, 100, 200, 1654 and 3309. The class ratio of this set is 1.05, and it is employed to mainly exhibit the performance of our proposed algorithm on balanced dataset.

The summary of five data sets is shown in Table 2, and the data is given in the additional files.

Software

The sssHD algorithm and the related methods or procedures are performed with R language [62], building on packages sparseSVM (https://CRAN.R-project.org/package=sparseSVM), e1071 (https://CRAN.Rproject.org/package=e1071), precrec [48], and ggplot2 [63]. The R code, including the sssHD algorithm and other related procedures, is given in the additional files.

Experimental results and discussion

Our algorithm is compared with four methods: Fisher score, Relief, AUCROCfilter and AUCPRCfilter. All methods are performed in two situations: no resampling and resampling with SMOTE [44] over all of the training data sets. SMOTE is an intelligent oversampling approach, which adds new, artificial minority

Table 2 The number of instances, features, majority, and minority as well as the class-imbalanced ratio (CIR) of five datasets used in this study

Datasets	Instances	Features	Majority	Minority	CIR
DLBCL	77	6285	58	19	3.05
SRBCT	83	2308	72	11	6.55
CAR	174	9182	163	11	14.82
GLIOMA	50	4434	43	7	6.14
LUNG	41	3309	21	20	1.05

[1] 4433 was given in [58], but a value of 4434 is obtained via the data from [58].

examples by interpolating between pre-existing minority instances rather than simply duplicating original examples. The minority class is over-sampled by taking each minority class point and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, the neighbors from the k nearest neighbors are randomly chosen.

Classification results under no resampling

In this section, we show the efficacy of the proposed sssHD approach on five gene expression datasets, and compare it with four other filter-based feature selection methods by assessing several performance measurements with no resampling preprocessing. The SVM classifier is employed to finish the classification task. The prediction result is obtained by performing leave-one-out cross validation rather than k-fold cross validation, as it is quite likely that, in the case of class-imbalance, the distribution within each fold varies widely with the uniformly sampling for creating the folds. The results on the first set (DLBCL) are shown in Fig. 4, while the results on last four sets (SRBCT, CAR, GLIOMA and LUNG) are given in the supporting information as additional files to save space. Figure 4 includes five subgraphs that evaluate five feature selection methods with TPR, $G-mean$, $F-measure$, $AUCROC$ and $precision$, respectively. It can be seen that sssHD gains satisfactory classification performance with just several top-ranked features, regardless of the metric used. The sssHD approach is competitive with other four feature selection methods, especially when the number of top-ranked predictors used is not too large. This finding indicates that the top-ranked features recognized by sssHD actually have the most relevance with the target. Let q be the number of top-ranked features that used to firstly reach the maximal value with five metrics. q is 3 by sssHD, which is less or equal to the value identified by the other four methods for dataset DLBCL with $G-mean$, $F-measure$, $AUCROC$ and $precision$ measurements. Thus sssHD algorithm achieves the best

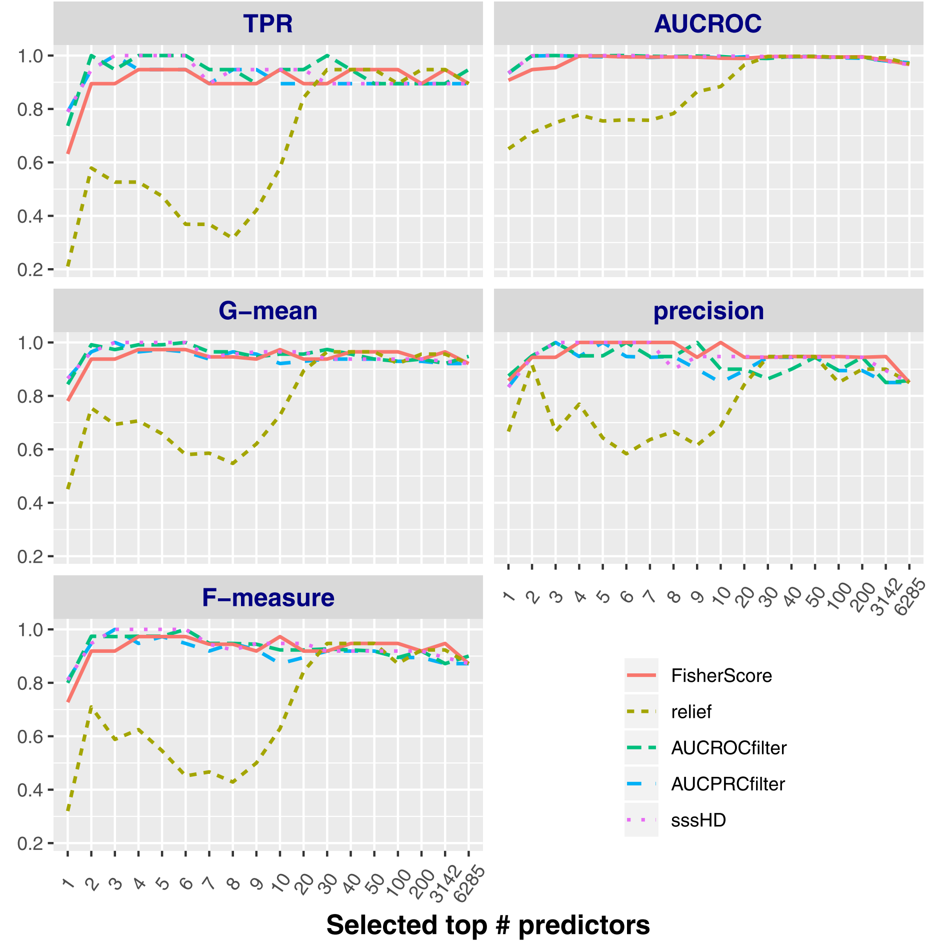

Fig. 4 The performance of five methods on DLBCL with no resampling. The five subgraphs report TPR, $G-mean$, $F-measure$, $AUCROC$ and $precision$, respectively

performance with the smallest number of features ranked at the top in the case of no resampling the data. Therefore, sssHD outperforms the other four methods in the original class-imbalanced data. This is also consistent with the property of skew-insensitivity of the Hellinger distance. A similar consequence can be obtained on SRBCT, CAR and GLIOMA datasets; for more details, see SI-Figs. 1, 3, 5 and 7 in the supporting information as additional files. In addition, sssHD achieves competitive performance and performs similarly well on balanced dataset LUNG, which further demostrates the skew-insensitivity of HD.

Classification results with SMOTE resampling

In this section, we consider the result by implementing re-balancing the data with SMOTE oversampling. The class ratio is approximately 1 : 1 after doing that. The SVM classifier is still utilized while coupled with stratified 5-fold cross validation due to the balance of the two classes herein. The performance of five methods on the DLBCL data set is shown in Fig. 5. Compared with Figs. 4, 5 shows that the performance with oversampling preprocessing is better than that with no resampling in most cases. Except Relief, the other four methods can obtain satisfactory accuracies, and they have almost no difference. sssHD is less affected by oversampling in comparison with other four methods. This result agrees with two properties of the Hellinger distance: skew-insensitivity and translation-invariant. An interesting discovery is that the optimal number of key features selected under original class-imbalanced data is less than that under SMOTE oversampling. We guess that such an oversampling strategy may lead to over-selection in choosing relevant variables. It also indirectly demostrates that it is necessary to develop feature selection technique designed for original class-imbalanced data rather than re-balanced data. A similar result can be obtained for SRBCT, CAR and GLIOMA datasets by SMOTE preprocessing (LUNG dataset is not performed here due to its balance); for more details, see SI-Figs. 2, 4 and 6 in the supporting information as additional files.

It should be pointed out that oversampling with SMOTE is applied before stratified cross validation to keep the

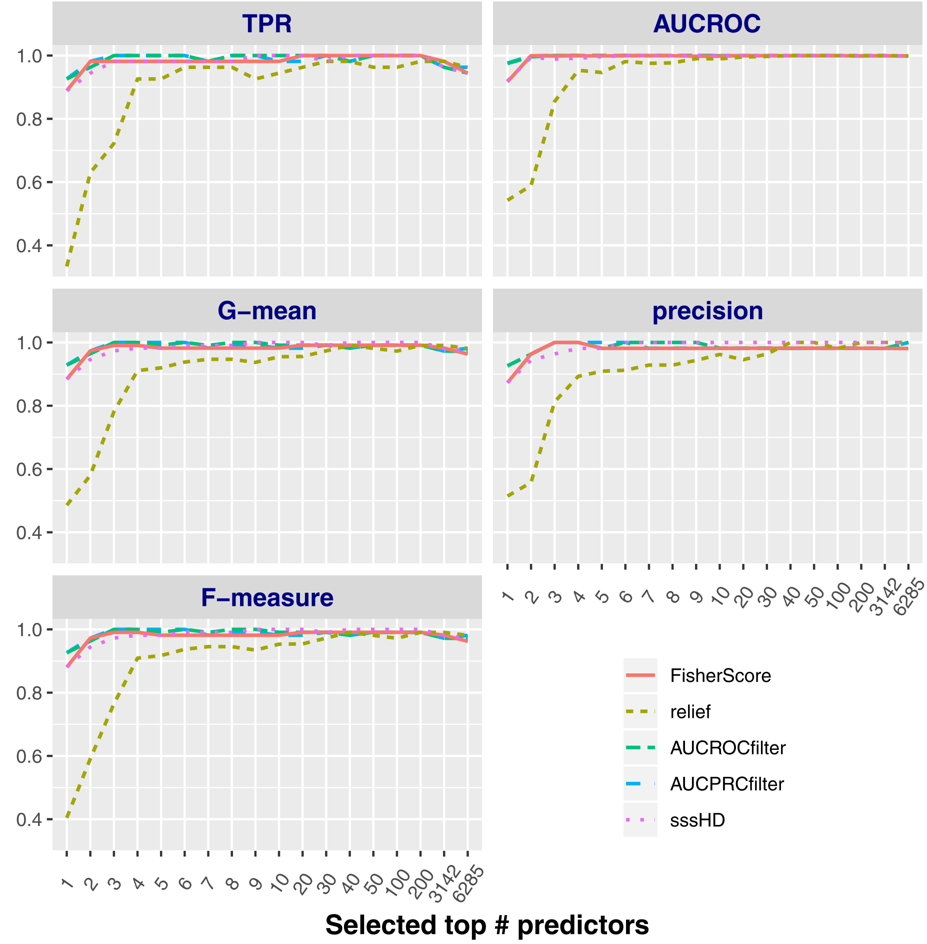

Fig. 5 The performance of five methods on DLBCL with SMOTE over-resampling. The five subgraphs report *TPR*, *G − mean*, *F − measure*, *AUCROC* and *precision*, respectively

class ratios consistent between the training set and the testing set. It is well known that SMOTE is not to simply replicate the original minority instances. In other words, the generated samples are different from the original data and also different from each other. Therefore, the points in testing set are generally not same with those in training set. However, if randomly resampling is used, where new samples are randomly duplicated from the minority class, the instances in testing set are likely to be similar to those in training set consequently leading to enhanced performance.

Conclusion

In this paper, we proposed a feature selection approach (sssHD) based on the Hellinger distance. Due to the properties of Hellinger distance, the sssHD algorithm is well suited to perform feature selection in class-imbalance learning. We have shown that sssHD can obtain high performance and is extremely competitive against several existing selection methods by means of several assessment measures. sssHD is extremely general as it can be easily extended from at least three aspects: 1) combining with different re-balance samplings such as under-sampling, over-sampling, SMOTE and so on; 2) changing the sparse regularization structure according to the characteristic of the predictor matrix, such as group LASSO [46], if the predictors possess some kind of group structure; and 3) the SVM classifier used in sssHD could be replaced by other classifiers, if necessary, such as discriminant analysis, naive Bayes, logistic regression, random forest, etc.. Therefore, many generalization algorithms can be derived from sssHD. In addition to discovering features that are truly associated with the response, controlling the FDR is also important in performing variable selection [64], so the Hellinger distance coupled with 'model-X' knockoffs [65], a useful technique to limit the FDR, would be a feasible choice to recognize true relevant feature and reduce the FDR as much as possible in high-dimensional class-imbalance learning. In addition, the Hellinger distance presents advantages for performing class-imbalanced data, so a worthy attempt may be to directly establish a Hellinger distance feature selection algorithm that does not depend on any classifiers.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3411-3.

Additional file 1: Supporting information is provided in a PDF file, in which the results on SRBCT, CAR and GLIOMA datasets are reposited.

Additional file 2: Five datasets used in this study are given as a .txt file.

Additional file 3: The R code for implementing the sssHD algorithm and related calculations is available.

Abbreviations

ACC: accuracy; AUCPRC: area under PRC; AUCROC: area under ROC; C: correctly; CLT: central limit theorem; FDR: false discovery rate; FN: false negative; FP: false positive; HD: Hellinger distance; IC: incorrectly; PRC: precision recall curve; ROC: receiver operating characteristics; sssHD: Hellinger distance based stable sparse selection; SVM: sopport vector machine; TN: true negative; TP: true positive.

Acknowledgements

The authors sincerely thank two Associate Editors and three anonymous reviewers for their constructive comments that led to the current improved version of the paper. We would like to thank the National Natural Science Foundation of China for its financial support (Grant Nos. 11761041 and 21775058).

Authors' contributions

GF and JP conceived the study. GF designed the algorithm and wrote the manuscript. GF, YW and MZ performed the computation and analysis. All authors read and approved the final manuscript.

Funding

This work is supported by the National Natural Science Foundation of China (Grant Nos. 11761041 and 21775058). The funder played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The supporting information, data as well as the R code for implementing the sssHD algorithm and related procedures are available in the additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1] School of Science, Kunming University of Science and Technology, Kunming 650500, People's Republic of China. [2] School of Mathematics, The University of Manchester, Manchester M13 9PL, UK.

Received: 5 April 2019 Accepted: 12 February 2020
Published online: 23 March 2020

References

1. Mamitsuka H. Selecting features in microarray classification using roc curves. Pattern Recog. 2006;39(12):2393–404.
2. Liu Z, Tan M. Roc-based utility function maximization for feature selection and classification with applications to high-dimensional protease data. Biometrics. 2008;64(4):1155–61.
3. Zhou P, Hu X, Li P, Wu X. Online feature selection for high-dimensional class-imbalanced data. Knowl-Based Syst. 2017;136:187–99.
4. Ma S, Huang J. Regularized roc method for disease classification and biomarker selection with microarray data. Bioinformatics. 2005;21(24): 4356–62.
5. Yin L, Ge Y, Xiao K, Wang X, Quan X. Feature selection for high-dimensional imbalanced data. Neurocomputing. 2013;105:3–11.
6. Maldonado S, Weber R, Famili F. Feature selection for high-dimensional class-imbalanced data sets using support vector machines. Inf Sci. 2014;286:228–46.
7. Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. ACM Sigkdd Explor Newsl. 2004;6(1):80–9.
8. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84.
9. Denil M, Trappenberg T. In: Farzindar A, Kešelj V, editors. Overlap versus Imbalance. Berlin, Heidelberg: Springer; 2010, pp. 220–31.
10. Alibeigi M, Hashemi S, Hamzeh A. Dbfs: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets. Data Knowl Eng. 2012;81-82:67–103.

11. García V, Sánchez J, Mollineda R. An empirical study of the behavior of classifiers on imbalanced and overlapped data sets. In: Rueda L, Mery D, Kittler J, editors. Progress in Pattern Recognition, Image Analysis and Applications. Berlin, Heidelberg: Springer; 2007. p. 397–406.
12. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
13. Kent JT. Information gain and a general measure of correlation. Biometrika. 1983;70(1):163–73.
14. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. BMC Bioinformatics. 2005;6(1):148.
15. Ghosh D, Chinnaiyan AM. Classification and selection of biomarkers in genomic data using lasso. BioMed Res Int. 2005;2005(2):147–54.
16. Guo H, Li Y, Shang J, Gu M, Huang Y, Gong B. Learning from class-imbalanced data: Review of methods and applications. Expert Syst Appl. 2017;73:220–39.
17. Wasikowski M, Chen XW. Combating the small sample class imbalance problem using feature selection. IEEE Trans Knowl Data Eng. 2010;22(10): 1388–400.
18. Ogura H, Amano H, Kondo M. Comparison of metrics for feature selection in imbalanced text classification. Expert Syst Appl. 2011;38(5):4978–89.
19. Blagus R, Lusa L. Class prediction for high-dimensional class-imbalanced data. BMC Bioinformatics. 2010;11(1):523.
20. Chen X-w, Wasikowski M. Fast: A roc-based feature selection metric for small samples and imbalanced data classification problems. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '08. New York: ACM; 2008. p. 124–32.
21. Maldonado S, López J. Imbalanced data classification using second-order cone programming support vector machines. Pattern Recog. 2014;47(5): 2070–9.
22. Li Y, Guo H, Liu X, Li Y, Li J. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowl-Based Syst. 2016;94:88–104.
23. Dubey R, Zhou J, Wang Y, Thompson PM, Ye J. Analysis of sampling techniques for imbalanced data: An n=648 adni study. NeuroImage. 2014;87:220–41.
24. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol. 1996;58(1):267–88.
25. Yu H, Mu C, Sun C, Yang W, Yang X, Xin Z. Support vector machine-based optimized decision threshold adjustment strategy for classifying imbalanced data. Knowl-Based Syst. 2015;76(1):67–78.
26. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32.
27. Zhou ZH, Liu XY. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Trans Knowl Data Eng. 2006;18(1):63–77.
28. Lin W-J, Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform. 2013;14(1):13–26.
29. Yu H, Sun C, Yang X, Yang W, Shen J, Qi Y. Odoc-elm: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data. Knowl-Based Syst. 2016;92:55–70.
30. Kailath T. The divergence and bhattacharyya distance measures in signal selection. IEEE Trans Commun Technol. 1967;15(1):52–60.
31. Cieslak DA, Hoens TR, Chawla NV, Kegelmeyer WP. Hellinger distance decision trees are robust and skew-insensitive. Data Min Knowl Disc. 2012;24(1):136–58.
32. Vapnik VN. The Nature of Statistical Learning Theory. Berlin: Springer; 2000.
33. Schölkopf B, Smola AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Massachusetts: MIT press; 2002.
34. Zhang T. An introduction to support vector machines and other kernel-based learning methods. AI Mag. 2001;22(2):103.
35. Fu GH, Cao DS, Xu QS, Li HD, Liang YZ. Combination of kernel pca and linear support vector machine for modeling a nonlinear relationship between bioactivity and molecular descriptors. J Chemom. 2011;25(2): 92–9.
36. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–60.
37. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. Ann Stat. 2004;32(2):407–99.
38. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–20.
39. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.
40. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B Stat Methodol. 2010;72(4):417–73.
41. Huang J, Breheny P, Ma S. A selective review of group selection in high-dimensional models. Stat Sci. 2012;27(4):481–99.
42. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
43. Witten DM, Tibshirani R. Penalized classification using fisher's linear discriminant. J R Stat Soc Ser B Stat Methodol. 2011;73(5):753–72.
44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
45. Lin W-C, Tsai C-F, Hu Y-H, Jhang J-S. Clustering-based undersampling in class-imbalanced data. Inf Sci. 2017;409–10:17–26.
46. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol. 2006;68(1):49–67.
47. Huang J, Ma S, Xie H, Zhang CH. A group bridge approach for variable selection. Biometrika. 2009;96(2):339–55.
48. Saito T, Rehmsmeier M. Precrec: fast and accurate precision–recall and roc curve calculations in r. Bioinformatics. 2017;33(1):145–7.
49. Guyon I, Gunn S, Nikravesh M, Zadeh LA. Feature Extraction: Foundations and Applications vol. 207. Berlin: Springer; 2008.
50. Kira K, Rendell LA. The feature selection problem: Traditional methods and a new algorithm. In: AAAI, vol. 2; 1992. p. 129–34.
51. Hulse JV, Khoshgoftaar TM, Napolitano A, Wald R. Feature selection with high-dimensional imbalanced data. In: 2009 IEEE International Conference on Data Mining Workshops; 2009. p. 507–14. https://doi.org/10.1109/ICDMW.2009.35.
52. Fu G-H, Yi L-Z, Pan J. Tuning model parameters in class-imbalanced learning with precision-recall curve. Biom J. 2018;0(0):. https://doi.org/10.1002/bimj.201800148.
53. Chowdhury S, Sing JK, Basu DK, Nasipuri M. Face recognition by generalized two-dimensional fld method and multi-class support vector machines. Appl Soft Comput. 2011;11(7):4282–92.
54. Wang S, Li D, Wei Y, Li H. A feature selection method based on fisher's discriminant ratio for text sentiment classification. Expert Syst Appl. 2009;38(7):8696–702.
55. Robnik-Šikonja M, Kononenko I. Theoretical and empirical analysis of relieff and rrelieff. Mach Learn. 2003;53(1):23–69.
56. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. Plos ONE. 2015;10(3):1–21.
57. Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RCT, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC, Golub TR. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8:68.
58. Yang K, Cai Z, Li J, Lin G. A stable gene selection in microarray data analysis. BMC Bioinformatics. 2006;7(1):228.
59. Khan J, Wei J, Ringner M, Saal L, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu C, Peterson C. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med. 2001;7(6):673–9.
60. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, Schultz PG, Powell SM, Moskaluk CA, Frierson HF. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res. 2001;61(20):7388–93.
61. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, von Deimling A, Pomeroy SL, Golub TR, Louis DN. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res. 2003;63(7):1602–7.
62. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5(3):299–314.
63. Wickham H. Ggplot2: Elegant Graphics for Data Analysis. Berlin: Springer; 2016.
64. Barber RF, Candès EJ, et al. Controlling the false discovery rate via knockoffs. Ann Stat. 2015;43(5):2055–85.

65. Candes E, Fan Y, Janson L, Lv J. Panning for gold:’model-x’knockoffs for high dimensional controlled variable selection. J R Stat Soc Ser B Stat Methodol. 2018;80(3):551–77.

Publisher’s Note

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